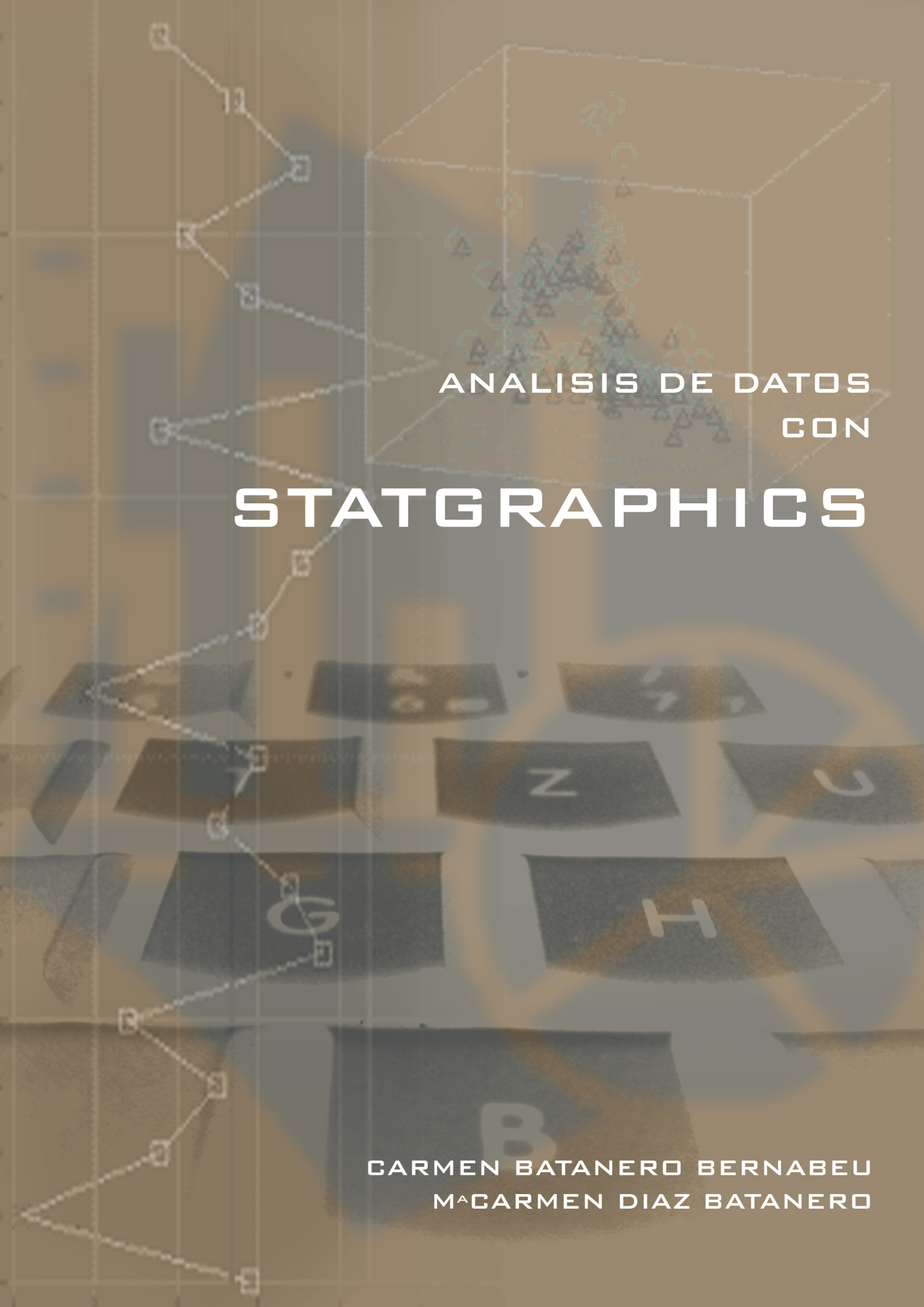


ANALISIS DE DATOS CON STATGRAPHICS

CARMEN BATANERO BERNABEU
MªCARMEN DIAZ BATANERO

ANÁLISIS DE DATOS CON STATGRAPHICS

CARMEN BATANERO BERNABEU

CARMEN DÍAZ BATANERO

GRANADA, 2008

ANÁLISIS DE DATOS CON STATGRAPHICS

© Las autoras
Departamento de Didáctica de la Matemática
Facultad de Ciencias de la Educación
Universidad de Granada
18071 Granada

ISBN: 978-84-691-4796-2
Depósito Legal: GR-1733-2008

Impresión:
La Gioconda, S. L.
Melchor Almagro, 16 18002

Publicación realizada en el marco del Proyecto SEJ2007-60110/EDUC, MEC- FEDER

PRESENTACIÓN

La estadística se incluye en la actualidad como asignatura instrumental en la mayor parte de las licenciaturas e ingenierías, debido a su utilidad en la modelización de situaciones reales, análisis de datos, realización de inferencias y toma de decisiones. Se sugiere también un cambio en la metodología de enseñanza: Una estadística basada en las aplicaciones y centrada en el análisis de datos, aprovechando las ventajas que proporciona la tecnología para posibilitar el trabajo con problemas más abiertos y realistas, así como para la exploración de conceptos y propiedades. La finalidad ha de ser el desarrollo del razonamiento estadístico, esencial en la sociedad moderna, que complementa y refuerza el currículo global del estudiante.

Siguiendo estos principios, presentamos en este texto un material para un curso básico de análisis de datos, abordando desde el análisis exploratorio de datos y la probabilidad, hasta la introducción de la inferencia con una y dos muestras, análisis de varianza, correlación y regresión. El curso está dirigido a estudiantes de ciencias sociales: Psicología, Educación, Sociología, etc., pero puede ser también útil en cursos de introducción a la estadística para estudiantes de otras especialidades.

Un peso importante lo constituyen los Proyectos y actividades, que tratan de motivar, contextualizar y dotar de sentido a las diferentes técnicas y conceptos, según van siendo introducidos. Los ejemplos y actividades complementarias permiten presentar aplicaciones en áreas variadas, así como reflexionar sobre el significado de los conceptos y propiedades, tratando de tener en cuenta los errores frecuentes en la aplicación e interpretación de la estadística. La incorporación de Statgraphics es gradual desde el comienzo del curso, destacando no sólo sus posibilidades de análisis, simulación y visualización, sino sobre todo, enfatizando la interpretación adecuada de los resultados y su puesta en relación con las preguntas que guían el análisis de los datos.

Estamos convencidas de que el curso será de gran utilidad para los estudiantes y profesores de análisis de datos.

INDICE

TEMA 1

LA ESTADÍSTICA, SUS APLICACIONES Y PROYECTOS DE ANÁLISIS DE DATOS

1.1. ¿QUÉ ES LA ESTADISTICA?

En lenguaje coloquial acostumbramos a llamar "estadísticas" a ciertas colecciones de datos, presentados usualmente en forma de tablas y gráficos. Así, es frecuente hablar de estadísticas de empleo, de emigración, de producción, de morbilidad, etc. Una definición de la estadística es la siguiente:

> *"La estadística estudia el comportamiento de los fenómenos llamados de colectivo. Está caracterizada por una información acerca de un colectivo o universo, lo que constituye su objeto material; un modo propio de razonamiento, el método estadístico, lo que constituye su objeto formal y unas previsiones de cara al futuro, lo que implica un ambiente de incertidumbre, que constituyen su objeto o causa final"* (Cabriá, 1994).

Como rama de las matemáticas, y utilizando el cálculo de probabilidades, la estadística estudia los fenómenos o experimentos aleatorios intentando deducir leyes sobre los mismos y aplicando dichas leyes para la predicción y toma de decisiones. Para aclarar este segundo significado, conviene precisar el concepto de fenómeno "aleatorio" o de azar.

Experimentos aleatorios y deterministas

Dentro de los diferentes hechos que pueden ser observados en la naturaleza, o de los experimentos que pueden ser realizados, distinguiremos

dos categorías. Llamaremos *experimento o fenómeno determinista* a aquél que siempre se produce en igual forma cuando se dan las mismas condiciones. Esto ocurre, por ejemplo, con el tiempo que tarda un móvil en recorrer un espacio dado con movimiento uniforme, a velocidad constante.

Por el contrario, con el término *"aleatorio"* se indica la posibilidad de que en idénticas condiciones puedan producirse resultados diferentes, que no son, por tanto, previstos de antemano. Tal ocurre, por ejemplo, al contar el número de semillas de una fruta, o al observar la duración de un televisor, o el tiempo transcurrido entre dos llamadas a una central telefónica. Igualmente, el resultado de cualquiera de los denominados juegos de azar, como lotería, dados, monedas es imprevisible de antemano. Sin embargo, si se hace una larga serie de una de tales experiencias, se observa una regularidad que es fundamental para el estudio de los fenómenos de azar y que se conoce como ley del azar o de estabilidad de las frecuencias: al repetir un mismo experimento aleatorio A una serie n de veces, el cociente n_A/n (llamado frecuencia relativa) entre las veces que aparece A (n_A) y el número total de realizaciones tiende a estabilizarse alrededor de un número que se conoce como *probabilidad* de dicho resultado.

Actividades

1.1. Recopila una lista de definiciones de la estadística a partir de textos de autores de prestigio y a partir de ella prepara una lista de las características que te parezcan más esenciales de la estadística.

1.2. Escribe algunos ejemplos de fenómenos aleatorios y no aleatorios.

Poblaciones, censos y muestras

Una *población (o universo)* es el conjunto total de objetos que son de interés para un problema dado. Los objetos pueden ser personas, animales, productos fabricados, etc. Cada uno de ellos recibe el nombre de *elemento (o individuo)* de la población. Generalmente, en un estudio estadístico, estamos interesados en analizar algún aspecto parcial de los individuos que componen la población; por ejemplo, si se trata de personas, puede que nos interese, la edad, profesión, nivel de estudios, el sueldo mensual que recibe, el número de personas de su familia, la opinión que le merece el partido que gobierna, etc. Estos aspectos parciales reciben el nombre de *caracteres* de los elementos de una población y son, por su naturaleza, variables, de

forma que en distintos individuos pueden tomar valores o modalidades diferentes.

El principal objetivo del análisis estadístico es conocer algunas de las propiedades de la población que interesa. Si la población es finita, el mejor procedimiento será la inspección de cada individuo (siempre que esto sea posible). Un estudio estadístico realizado sobre la totalidad de una población se denomina *censo.* Estudios de este tipo son realizados periódicamente por el Gobierno y otras instituciones.

Sin embargo, la mayoría de los problemas de interés, implican, bien poblaciones infinitas, o poblaciones finitas que son difíciles, costosas o imposibles de inspeccionar. Esto obliga a tener que seleccionar, por procedimientos adecuados, un subconjunto de *n* elementos de la población, que constituyen una muestra de tamaño *n*, examinar la característica que interesa y después generalizar estos resultados a la población. Esta generalización a la población se realiza por medio de la parte de la estadística que se conoce con el nombre de *inferencia estadística.* Para que estas conclusiones ofrezcan las debidas garantías es preciso comprobar que se cumple el requisito básico de que la muestra sea *representativa.*

Actividades

1.3. ¿Cuáles son los principales motivos de emplear el muestreo en un estudio estadístico, en lugar de usar una población completa?

1.4. Poner ejemplos de una población de personas y otra población de objetos y definir algunas posibles variables sobre las cuáles podríamos efectuar un estudio estadístico.

1.5. Al realizar una encuesta sobre preferencias de horarios, el 30 por ciento de los alumnos encuestados no devolvieron los cuestionarios. ¿Crees que este porcentaje de no respuestas puede afectar las conclusiones?

1.6. Supón que tienes que realizar una encuesta entre los alumnos de la Facultad de Educación para saber si eligieron sus estudios como primera opción o no. Piensa en algunas formas posibles de elegir una muestra representativa de 300 alumnos entre todos los de la Facultad.

1.7. ¿Sería adecuado hacer una encuesta sobre el número de hijos por familia en la ciudad de Granada a partir de una lista de teléfonos?

1.8. Pon ejemplos de algunos sesgos que pueden aparecer en una investigación por muestreo ¿Cómo se podrían controlar?

1.9. Buscar en la prensa alguna encuesta reciente. Identificar la población y la muestra, el tema de la encuesta, y analizar las variables estudiadas.

Orígenes de la estadística

Los orígenes de la estadística son muy antiguos, ya que se han encontrado pruebas de recogida de datos sobre población, bienes y producción en civilizaciones como la china (aproximadamente 1000 años a. c.), sumeria y egipcia. Incluso en la Biblia, en el libro de *Números* aparecen referencias al recuento de los israelitas en edad de servicio militar. No olvidemos que precisamente fue un censo lo que motivó del viaje de José y María a Belén, según el Evangelio. Los censos propiamente dichos eran ya una institución el siglo IV a.C. en el imperio romano.

Sin embargo sólo muy recientemente la estadística ha adquirido la categoría de ciencia. En el siglo XVII surge la aritmética política, desde la escuela alemana de Conring, quien imparte un curso son este título en la universidad de Helmsted. Posteriormente su discípulo Achenwall orienta su trabajo a la recogida y análisis de datos numéricos, con fines específicos y en base a los cuales se hacen estimaciones y conjeturas, es decir se observa ya los elementos básicos del método estadístico. Para los aritméticos políticos de los siglos XVII y XVIII la estadística era el arte de gobernar; su función era la de servir de ojos y oídos al gobierno.

La proliferación de tablas numéricas permitió observar la frecuencia de distintos sucesos y el descubrimiento de leyes estadísticas. Son ejemplos notables los estudios de Graunt sobre tablas de mortalidad y esperanza de vida a partir de los registros estadísticos de Londres desde 1592 a 1603 o los de Halley entre 1687 y 1691, para resolver el problema de las rentas vitalicias en las compañías de seguros. En el siglo XIX aparecen las leyes de los grandes números con Bernouilli y Poisson.

Otro problema que recibe gran interés por parte de los matemáticos de su tiempo, como Euler, Simpson, Lagrange, Laplace, Legendre y Gauss es el del ajuste de curvas a los datos. La estadística logra con estos descubrimientos una relevancia científica creciente, siendo reconocida por la British Association for the Advancement of Science, como una sección en 1834, naciendo así la Royal Statistical Society. En el momento de su fundación se definió la estadística como "conjunto de hechos, en relación con el hombre, susceptibles de ser expresados en números, y lo suficiente numerosos para ser representados por leyes".

Se crearon poco a poco sociedades estadísticas y oficinas estadísticas para organizar la recogida de datos estadísticos; la primera de ellas en Francia en 1800. Como consecuencia, fue posible comparar las estadísticas de cada país en relación con los demás, para determinar los factores determinantes del crecimiento económico y comenzaron los congresos internacionales, con el fin de homogeneizar los métodos usados. El primero de ellos fue organizado por Quetelet en Bruselas en 1853. Posteriormente, se decidió crear una sociedad estadística internacional, naciendo en 1885 el Instituto Internacional de Estadística (ISI) que, desde entonces celebra reuniones bianuales. Su finalidad especifica es conseguir uniformidad en los métodos de recopilación y abstracción de resultados e invitar a los gobiernos al uso correcto de la estadística en la solución de los problemas políticos y sociales. En la actualidad el ISI cuenta con 5 secciones, una de las cuales, la IASE, fundada en 1991, se dedica a la promoción de la educación estadística.

Corrientes en el análisis de datos

Aunque es difícil dividir la estadística en partes separadas, una división clásica hasta hace unos 30 años ha sido entre *estadística descriptiva y estadística inferencial.*

La e*stadística descriptiva*, se utiliza para describir los datos, resumirlos y presentarlos de forma que sean fáciles de interpretar. El interés se centra en el conjunto de datos dados y no se plantea el extender las conclusiones a otros datos diferentes. La *estadística inductiva o inferencia* trata de obtener conocimientos sobre ciertos conjuntos extensos o poblaciones, a partir de la información disponible de un subconjunto de tal población llamada muestra. Utiliza como herramienta matemática el cálculo de probabilidades.

Hasta 1900 la estadística se restringía a la estadística descriptiva, que, a pesar de sus limitaciones, hizo grandes aportaciones al desarrollo de la ciencia. A partir de esa época comenzaría la inferencia estadística, con los trabajos de Fisher, Pearson y sus colaboradores. Los avances del cálculo de probabilidades llevaron a la creación de la estadística teórica, que en cierto modo se alejó de las ideas primitivas, que se centraban en el análisis y recogida de datos. De este modo, en los años 60 la mayor parte de los libros de texto se ocupaban especialmente de los modelos inferenciales y hubo una tendencia a la matematización, junto con un descuido en la enseñanza de los aspectos prácticos del análisis de datos.

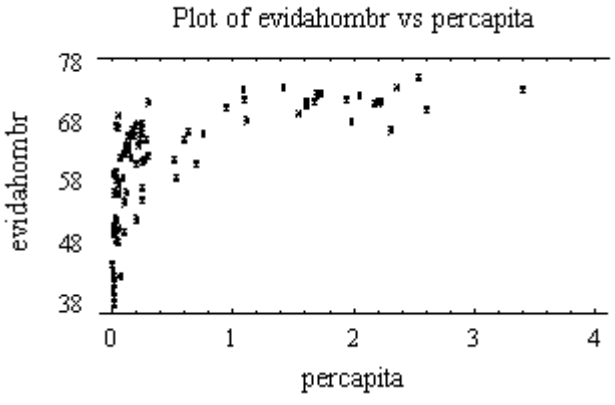
Con el desarrollo de la informática en la segunda mitad del siglo XX y la posibilidad de manejar rápidamente grandes masas de datos, se produjo, por un lado, una reacción ante tanta matematización, y por otro, disminuyó la importancia de los estudios muestrales. Puesto que era fácil analizar grandes muestras ya no había por qué limitarse a los métodos estadísticos basados en distribuciones conocidas, cuya principal aplicación eran las pequeñas muestras. Tampoco había por qué limitarse a analizar una o unas pocas variables, porque el tiempo de cálculo se había eliminado y era preferible aprovechar toda la información disponible.

Con todo ello suge una nueva filosofía en los estudios estadísticos: el *análisis exploratorio de datos*, introducido por Tukey, quien compara la labor del estadístico con la de un detective.

Anteriormente a este enfoque, el análisis de datos se basaba fundamentalmente en la estimación de parámetros (medias, o coeficientes de correlación en la población) y se disminuía la importancia de la representación de los datos. Además, se pensaba que para obtener conclusiones de los datos era preciso recurrir a la inferencia (modelo confirmatorio), donde el conjunto de valores observados se supone que se ajusta a un modelo preestablecido; por ejemplo, se supone que los datos se han obtenido de una población normal con media y desviación típica desconocidas.

Partiendo de esta hipótesis, que es previa a la recogida de datos, se calculan los estadísticos (media, coeficiente de correlación en la muestra) que servirán para aceptar o rechazar ciertas hipótesis establecidas de antemano. Al contemplar solamente dos alternativas, (confirmación o no de la hipótesis), los datos no se exploraban para extraer cualquier otra información que pueda deducirse de los mismos.

En el análisis exploratorio de datos, en lugar de imponer un modelo dado a las observaciones, se genera dicho modelo desde las mismas. Por ejemplo, cuando se estudian las relaciones entre dos variables, el investigador no solamente necesita ajustar los puntos a una línea recta, sino que estudia otros modelos distintos del lineal. En el gráfico adjunto relacionamos la renta per cápita con la esperanza de vida en 97 países.



Aunque los estadísticos calculados en este conjunto de datos presenten un valor estadísticamente significativo (el coeficiente de correlación sea significativamente distinto de cero), la relación entre las variables no se ajusta bien a una línea recta. En este ejemplo, al representar gráficamente los datos el investigador descubre algo importante: el modelo que mejor se ajusta a los datos no es una línea recta.

Actividades

1.10. El análisis de datos se basa en el método de elaboración de proyectos por parte de los estudiantes. Piensa algunos proyectos sencillos en los que se pueda recoger datos significativos y apropiados para el aprendizaje de conceptos elementales de análisis de datos.

1.2. APLICACIONES DE LA ESTADISTICA

La importancia que la estadística ha alcanzado en nuestros días, tanto como cultura básica, como en el trabajo profesional y en la investigación, es innegable. Ello es debido a la abundancia de información con la que el ciudadano debe enfrentarse en su trabajo diario. La mayor parte de las veces estas informaciones vienen expresadas en forma de tablas o gráficos estadísticos, por lo que un conocimiento básico de esta ciencia es necesario para la correcta interpretación de los mismos.

La principal razón que induce a incluir el estudio matemático de los fenómenos aleatorios en la educación primaria y secundaria es que las situaciones de tipo aleatorio tienen una fuerte presencia en nuestro entorno. Si queremos que el alumno valore el papel de la probabilidad y estadística, es importante que los ejemplos y aplicaciones que mostramos en la clase hagan ver de la forma más amplia posible esta fenomenología que analizamos a continuación.

Al final de la década de los 60 un comité de la American Statistical Association y del National Council of Teachers of Mathematics preparó un libro en el que se muestra la amplitud de las aplicaciones de la estadística. Este libro, editado por Tanur (1972) clasifica en cuatro grupos estas aplicaciones:

- el hombre en su mundo biológico
- el hombre en su mundo social

- el hombre en su mundo político
- el hombre en su mundo físico

A continuación hacemos un resumen de los problemas incluidos en cada una de estas categorías.

Nuestro mundo biológico

Dentro del campo biológico, puede hacerse notar al alumno que muchas de las características heredadas en el nacimiento no se pueden prever de antemano: el sexo, color de pelo, peso al nacer, etc. Algunos rasgos como la estatura, número de pulsaciones por minuto, recuento de hematíes, etc., dependen incluso del momento en que son medidas.

Otras aplicaciones se refieren al campo de la medicina. La posibilidad de contagio o no en una epidemia, la edad en que se sufre una enfermedad infantil, la duración de un cierto síntoma, o la posibilidad de un diagnóstico correcto cuando hay varias posibles enfermedades que presentan síntomas parecidos varían de uno a otro chico. El efecto posible de una vacuna, el riesgo de reacción a la misma, la posibilidad de heredar una cierta enfermedad o defecto, o el modo en que se determina el recuento de glóbulos rojos a partir de una muestra de sangre son ejemplos de situaciones aleatorias.

Cuando se hacen predicciones sobre la población mundial o en una región dada para el año 2050, por ejemplo, o sobre la posibilidad de extinción de las ballenas, se están usado estudios probabilísticos de modelos de crecimiento de poblaciones, de igual forma que cuando se hacen estimaciones de la extensión de una cierta enfermedad o de la esperanza de vida de un individuo.

En agricultura y zootecnia se utilizan estos modelos para prever el efecto del uso de fertilizantes o pesticidas, evaluar el rendimiento de una cosecha o las consecuencias de la extensión de una epidemia, nube tóxica, etc. Por último, y en el ámbito de la psicología, observamos el efecto del azar sobre el cociente intelectual o en la intensidad de respuesta a un estímulo, así como en los tipos diferentes de caracteres o capacidades de los individuos.

El mundo físico

Además del contexto biológico del propio individuo, nos hallamos inmersos en un medio físico variable. ¿Qué mejor fuente de ejemplos sobre fenómenos aleatorios que los meteorológicos? La duración, intensidad,

extensión de las lluvias, tormentas o granizos; las temperaturas máximas y mínimas, la intensidad y dirección del viento son variables aleatorias. También lo son las posibles consecuencias de estos fenómenos: el volumen de agua en un pantano, la magnitud de daños de una riada o granizo son ejemplos en los que se presenta la ocasión del estudio de la estadística y probabilidad.

También en nuestro mundo físico dependemos de ciertas materias primas como el petróleo, carbón y otros minerales; la estimación de estas necesidades, localización de fuentes de energía, el precio, etc., están sujetos a variaciones de un claro carácter aleatorio.

Otra fuente de variabilidad aleatoria es la medida de magnitudes. Cuando pesamos, medimos tiempo, longitudes, etc., cometemos errores aleatorios. Uno de los problemas que se puede plantear es la estimación del error del instrumento y asignar una estimación lo más precisa posible de la medida. Por último, citamos los problemas de fiabilidad y control de la calidad de los aparatos y dispositivos que usamos: coche, televisor, etc.

El mundo social

El hombre no vive aislado: vivimos en sociedad; la familia, la escuela, el trabajo, el ocio están llenos de situaciones en las que predomina la incertidumbre: El número de hijos de la familia, la edad de los padres al contraer matrimonio, el tipo de trabajo, las creencias o aficiones de los miembros varían de una familia a otra.

En la escuela, ¿podemos prever las preguntas del próximo examen?; ¿quién ganará el próximo partido? Para desplazarnos de casa a la escuela, o para ir de vacaciones, dependemos del transporte público que puede sufrir retrasos. ¿Cuantos viajeros usarán el autobús? ¿Cuantos clientes habrá en la caja del supermercado el viernes a las 7 de la tarde?

En nuestros ratos de ocio practicamos juegos de azar tales como quinielas o loterías. Acudimos a encuentros deportivos cuyos resultados son inciertos y en los que tendremos que hacer cola para conseguir las entradas. Cuando hacemos una póliza de seguros no sabemos si la cobraremos o por el contrario perderemos el dinero pagado; cuando compramos acciones en bolsa estamos expuestos a la variación en las cotizaciones,...

El mundo político

El Gobierno, a cualquier nivel, local, nacional o de organismos internacionales, necesita tomar múltiples decisiones que dependen de

fenómenos inciertos y sobre los cuales necesita información. Por este motivo la administración precisa de la elaboración de censos y encuestas diversas. Desde los resultados electorales hasta los censos de población hay muchas estadísticas cuyos resultados afectan las decisiones de gobierno y todas estas estadísticas se refieren a distintas variables aleatorias relativas a un cierto colectivo. Entre las más importantes citaremos: el índice de precios al consumo, las tasas de población activa, emigración – inmigración, estadísticas demográficas, producción de los distintos bienes, comercio, etc., de las que diariamente escuchamos sus valores en las noticias.

1.3. ENSEÑANZA DE LA ESTADÍSTICA BASADA EN PROYECTOS DE ANÁLISIS DE DATOS

En asignaturas como física, química o biología, en los niveles de enseñanza secundaria y primeros cursos universitarios es tradicional alternar las clases teóricas y de resolución de problemas con las prácticas en laboratorio. Sin embargo, en la enseñanza de la estadística, hasta hace poco tiempo, las clases prácticas se han reducido, en general, a la resolución de problemas típicos, que, con frecuencia, se han alejados de las aplicaciones reales. Esto es debido a la dificultad de realizar el análisis de un volumen relativamente grande datos con la mera ayuda de calculadoras de bolsillo. Con esta metodología tradicional el alumno se siente poco motivado hacia el estudio de esta materia y encuentra dificultades para aplicar los conocimientos teóricos a la resolución de casos prácticos.

Ahora bien, la mayor disponibilidad, en la actualidad, tanto de equipos informáticos de bajo coste, como de programas de ordenador para el análisis de datos permite la organización de clases prácticas complementarias con la filosofía didáctica del "laboratorio". Por otra parte, el análisis de datos estadísticos se realiza en la actualidad utilizando medios informáticos, por la considerable ventaja que suponen en rapidez y fiabilidad. Por tanto, el aprendizaje del manejo de esta herramienta debe formar parte del currículo para preparar al estudiante para un uso adecuado de estos medios.

Pero además de este uso de tipo instrumental las capacidades de simulación y representación gráfica de los ordenadores actuales facilitan su uso como recurso didáctico en la formación de conceptos y el aprendizaje constructivista. En un ordenador pueden simularse fenómenos cuya observación en la vida real sería costosa o larga. Desde la obtención de números aleatorios a la simulación de procesos estocásticos hay un gran número de temas en los cuales los ordenadores pueden desempeñar una

ayuda valiosa: teoremas de límite, distribuciones en el muestreo, caminatas al azar, etc. En síntesis podemos decir que el uso de los ordenadores en la enseñanza de la estadística permite al estudiante:

- Estudiar datos procedentes de casos prácticos reales, incorporándose el "método de proyectos";
- Adquirir destrezas en el manejo de la herramienta informática;
- La comprensión de conceptos y técnicas estadísticas a través de simulaciones y el proceso de análisis de los datos.

1.4. ALGUNOS PROYECTOS INICIALES

El análisis de datos es sólo una parte (aunque importante) en el proceso de investigación. Este proceso comienza con la definición de un problema, el estudio de la bibliografía relacionada y el diseño del trabajo de campo, en el cual recogeremos datos para el estudio, mediante encuestas, observación o mediciones. Una vez recogidos los datos y planteadas las preguntas de investigación el análisis de datos permitirá contestar estas preguntas si están bien planteadas y se han recogido los datos necesarios. Finalmente será necesario escribir un informe.

En la enseñanza de la estadística podemos plantear a los alumnos pequeñas investigaciones que contextualicen el aprendizaje y les sirva para llegar a comprender el papel de la estadística en el proceso más amplio de investigación. Plantearemos algunos de estos proyectos a lo largo del curso, comenzando en esta unidad por dos ejemplos.

Proyecto 1. Diferencias demográficas en países desarrollados y en vías de desarrollo

La actividad se desarrolla en torno a un fichero que contiene datos de 97 países y que ha sido adaptado del preparado por Rouncenfield (1995) y ha sido tomado de Internet, del servidor de la revista Journal of Statistical Education (http://www2.ncsu.edu/ncsu/pams/stat/info/jse/homepage.html). Contiene las siguientes variables, que se refieren al año 1990:

- *Tasa de natalidad*: Niños nacidos vivos en el año por cada 1000 habitantes;
- *Tasa de mortalidad*: Número de muertes en el año por cada 1000

habitantes;

- *Mortalidad infantil*: Número de muertes en el por cada 1000 niños de menos de 1 año;
- *Esperanza de vida* al nacer para hombres y mujeres;
- *Producto Nacional Bruto* per cápita en dólares (USA);
- *Grupo*: Clasificación de países en función de la zona geográfica y situación económica, en las siguientes categorías: 1 = Europa Oriental: 2 = Ibero América; 3 = Europa Occidental, Norte América, Japón, Australia, Nueva Zelanda; 4 = Oriente Medio; 5 = Asia; 6 = África.

Hemos añadido el número de habitantes en 1995 en miles de personas (*Población*), tomado del anuario publicado por el periódico español "El País". A continuación listamos los datos correspondientes a los 10 primeros países en el fichero.

El objetivo de este proyecto es estudiar las tendencias y variabilidad de las diversas variables, analizar las diferencias demográficas en los diferentes grupos de países, y cómo dependen del PNB y estudiar la interrelación entre las diferentes variables del fichero.

País	Grupo	Tasa natalidad	Tasa mortalidad	Mortalidad infantil	Esperanza vida hombre	Esperanza vida mujer	PNB	Población (miles)
Afganistán	5	40,4	18,7	181,6	41,0	42,0	168	16000
Albania	1	24,7	5,7	30,8	69,6	75,5	600	3204
Alemania (Oeste)	3	11,4	11,2	7,4	71,8	78,4	22320	16691
Alemania Este	1	12,0	12,4	7,6	69,8	75,9		61337
Algeria	6	35,5	8,3	74,0	61,6	63,3	2060	24453
Angola	6	47,2	20,2	137,0	42,9	46,1	610	9694
Arabia Saudí	4	42,1	7,6	71,0	61,7	65,2	7050	13562
Argentina	2	20,7	8,4	25,7	65,5	72,7	2370	31883
Austria	3	14,9	7,4	8,0	73,3	79,6	17000	7598
Bahrein	4	28,4	3,8	16,0	66,8	69,4	6340	459

Proyecto 2. Actitudes hacia la estadística

Se trata de recoger datos en clase sobre la actitud de los estudiantes hacia la estadística, utilizando como instrumento de recogida de datos, la escala de actitudes presentada a continuación.

Se recogerán también datos sobre el sexo del alumno, especialidad que cursa y si tiene o no estudios previos de estadística. El objetivo del proyecto es analizar los componentes de las actitudes, así como la actitud global hacia la estadística y comparar según sexos, especialidades y estudios previos del tema.

Para cada una de las siguiente preguntas indica en la escala 1 a 5 tu grado de acuerdo, según el siguiente convenio

1	2	3	4	5
Fuertemente en desacuerdo	No estoy de acuerdo	Indiferente	De acuerdo	Fuertemente de acuerdo

1. Uso a menudo la información estadística para formar mis opiniones o tomar decisiones
1 2 3 4 5
2. Es necesario conocer algo de estadística para ser un consumidor inteligente.
1 2 3 4 5
3. Ya que es fácil mentir con la estadística, no me fío de ella en absoluto.
1 2 3 4 5
4. La estadística es cada vez más importante en nuestra sociedad y saber estadística será tan necesario como saber leer y escribir.
1 2 3 4 5
5. Me gustaría aprender más estadística si tuviese oportunidad.
1 2 3 4 5
6. Debes ser bueno en matemáticas para comprender los conceptos estadísticos básicos.
1 2 3 4 5
7. Cuando te compras un coche nuevo es preferible preguntar a los amigos que consultar una encuesta sobre la satisfacción de usuarios de distintas marcas, en una revista de información al consumidor.
1 2 3 4 5
8. Me parecen muy claras las frases que se refieren a la probabilidad, como, por ejemplo, las probabilidades de ganar una lotería.
1 2 3 4 5
9. Entiendo casi todos los términos estadísticos que encuentro en los periódicos o noticias.
1 2 3 4 5
10. Podría explicar a otra persona como funciona una encuesta de opinión.
1 2 3 4 5

1.5. TIPOS DE DATOS Y ESCALAS DE MEDIDA

Como resultado de nuestras medidas sobre individuos o unidades experimentales de la población bajo estudio, obtenemos un conjunto de datos, o resultados del experimento estadístico. Para facilitar el análisis asignaremos unos valores a cada unidad experimental de acuerdo con

ciertas reglas; así, podemos asignar el número 1 a los varones y el 2 a las hembras, o bien los símbolos "V" y "H".

Pueden observarse muchas características diferentes para un mismo individuo. Estas características, dependiendo del tipo de valores que originan, pueden medirse con cuatro tipos distintos de *escalas de medida:* escala nominal, ordinal, de intervalo y de razón. Vamos a analizar las características de cada una.

Escala nominal

La forma más simple de observación es la clasificación de individuos en clases que simplemente pueden distinguirse entre si pero no compararse ni realizar entre ellas operaciones aritméticas. En este tipo se incluyen características tales como la profesión, nacionalidad o grupo sanguíneo.

Escala ordinal

A veces, las categorías obtenidas pueden ser ordenadas, aunque diferencias numéricas iguales a lo largo de la escala numérica utilizada para medir dichas clases no correspondan a incrementos iguales en la propiedad que se mide. Por ejemplo, puede asignarse un número de orden de nacimiento a un grupo de hermanos, sin que la diferencia de edad entre el 1º y el 2º de ellos sea la misma que la del 2º al 3º.

Escala de intervalo

Esta escala, además de clasificar y ordenar a los individuos, cuantifica la diferencia entre dos clases, es decir, puede indicar cuanto más significa una categoría que otra. Para ello es necesario que se defina una unidad de medida y un origen, que es por su naturaleza arbitrario. Tal ocurre con la temperatura y también con la escala cronológica.

Escala de razón

Es idéntica a la anterior, pero además existe un cero absoluto. En el apartado anterior hemos incluido el caso del tiempo, ya que no puede medirse con una escala de razón. En efecto, si consideramos las fechas 2000 DC y 1000 DC, aunque 2000 es el doble que 1000 no quiere decirse que el tiempo desde el origen del hombre sea el doble en un caso que en otro, pues hasta el año 0 DC han transcurrido un número de años

desconocido. Ejemplos de características que pueden ser medidas a nivel de razón son el cociente intelectual, grado de depresión o puntuación en un cuestionario.

El nivel elegido para medir una característica condiciona el resto del análisis estadístico, pues las técnicas utilizadas deben tener en cuenta la escala que se ha empleado. En general cuanto mayor sea el nivel utilizado, mayor número de técnicas podrán aplicarse y mayor precisión se logrará, por lo que se recomienda usar la escala de intervalo o la de razón siempre que sea posible.

Actividades

1.11. Poner un ejemplo de características estadísticas en las siguientes escalas de medida: Nominal, ordinal, de intervalo, de razón.

1.12. Hemos realizado una encuesta a un grupo de alumnos. Clasifica las siguientes características, según su escala de medida y tipo de variable: Peso, religión, número de hermanos, orden de nacimiento respecto a sus hermanos, tiempo que tarda en completar la encuesta, deporte preferido.

1.13. ¿Por qué no podemos decir que una temperatura de 100 grados Fahrenheit indica doble calor que una temperatura de 50 grados Fahrenheit?

1.14. Agrupamos a los niños de la clase en altos, medianos y bajos. ¿Qué tipo de escala de medida usamos? ¿Y si los ordenamos por estatura?

1.15. ¿Cuál es la escala de medida de cada una de las variables de los proyectos 1 y 2?

Variables estadísticas

Para representar los distintos tipos de datos empleamos variables. Una variable es un símbolo que puede tomar valores diferentes. Cuando estos valores son los resultados de un experimento estadístico, la llamamos *variable estadística,* y representa generalmente un cierto carácter de los individuos de una población.

Usualmente, las variables estadísticas se clasifican en *cualitativas* y *cuantitativas,* según que las modalidades del carácter que representan sean o no numéricas. (Algunos autores no consideran las variables cualitativas, puesto que puede asignarse un número diferente a cada una de las modalidades de una variable cualitativa).

Dentro de las variables cuantitativas se distingue entre variables *discretas* y *continuas,* siendo discretas aquellas que por su naturaleza sólo pueden tomar valores aislados –generalmente números enteros– y continuas las que pueden tomar todos los valores de un cierto intervalo.

Así, los experimentos que consisten en el recuento de objetos, como pueden ser: número de miembros de una familia, número de empleados de una empresa, etc., dan lugar a variables discretas, mientras que al medir magnitudes tales como el peso, el tiempo, capacidad, longitud, etc. se obtienen variables continuas.

Hay que tener en cuenta que, a veces, la naturaleza de la variable utilizada depende del tipo y necesidades de la investigación. Así, los datos nominales y ordinales son necesariamente cualitativos y discretos mientras que los de intervalo y razón pueden ser discretos o continuos. Por ejemplo, las magnitudes monetarias, temperatura, etc.

Actividades

1.16. Para cada una de las siguientes variables, indica si es mejor considerarla discreta o continua: a) Tiempo para completar una tarea; b) Número de años de escolaridad; c) Número de sillas en una habitación

1.17. Clasifica las variables de los proyectos 1 y 2 en cualitativas y cuantitativas, discretas y continuas.

1.18. En una encuesta codifico la provincia de nacimiento con un número de 1 a 50. ¿Qué topo de variable estadística es la provincia de nacimiento, cualitativa o cuantitativa?

1.19. Para codificar la edad de una persona un alumno sugiere usar el siguiente criterio: De 0 a 10 años: codificar como 1; de 10 a 20 años codificar como 2, de 20 a 30 años codificar como 3, etc. El alumno propone este sistema de codificación para tener un menor número de códigos. ¿Crees que es acertada la propuesta del alumno? ¿En qué casos estaría justificada?

1.6. CARACTERÍSTICAS GENERALES Y ESTRUCTURA DE STATGRAPHICS

Statgraphics es un paquete estadístico profesional, es decir, proporciona los tipos de análisis estadísticos más comunes, y proporciona otros instrumentos necesarios en el análisis de los datos, tales como editor de datos, utilidades para manejar los ficheros de datos, opciones para

cambiar parámetros del sistema y ayudas. *Statgraphics* usa varios tipos de ficheros, entre ellos, los siguientes:

1. *Ficheros de dato:* En ellos introducimos los datos a analizar. Para realizar un análisis estadístico es necesario que haya un fichero de datos abierto.
2. *Ficheros Statfolio:* Estos son ficheros que podemos grabar con los resultados de nuestros análisis, para tenerlos disponibles en el futuro. El *Statfolio* incluye también el fichero de datos, así como todas las ventanas de resultados que no se hubieran cerrado al grabar el *Statfolio.*
3. *Statreporter:* Sirve para escribir el informe a medida que se analizan los datos.

Este programa está estructurado mediante una serie de *menús* encadenados, cada uno de los cuáles tiene diversas opciones que podemos usar para cambiar los resultados de los análisis o para pedir nuevos análisis. El funcionamiento de todos los programas es muy parecido, de modo que lo que se aprende en un curso inicial servirá para avanzar posteriormente o incluso aprender el uso de otros programas, ya que la mayoría tiene una estructura y filosofía similar.

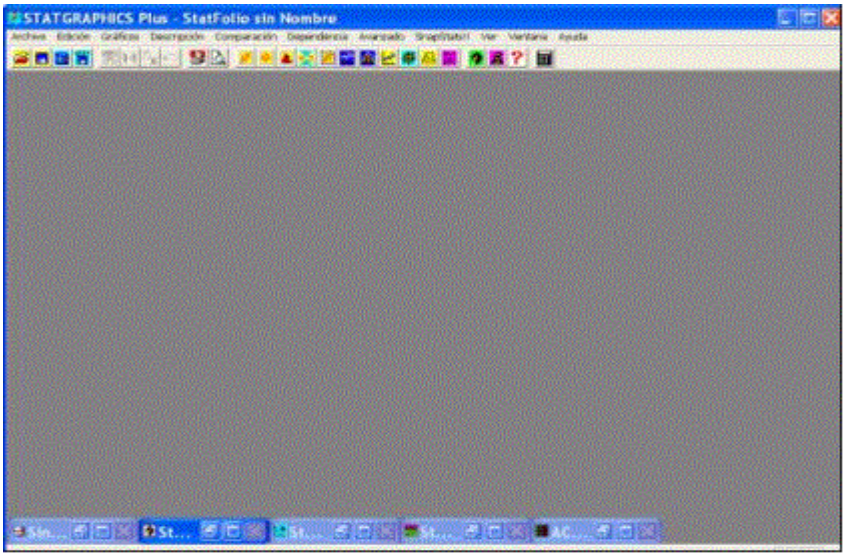
El ratón del ordenador nos sirve para pulsar en diversos iconos o palabras que nos permiten acceder a diversos menús, introducir variables o cambiar parámetros en el análisis de los datos.

Generalmente tenemos varias ventanas activas en la pantalla: La ventana de editor de datos, la ventana de resultados del análisis, la ventana de ayuda, etc. Podemos pasar de una a otra ventana pulsando con el ratón la palabra VENTANAS. Cuando iniciamos el funcionamiento de *Statgraphics* entramos en el *menú principal*, que describimos a continuación.

1.7. MENÚ PRINCIPAL

En esta sección describiremos los elementos principales de las ventanas y menús de STATGRAPHICS, su aplicación y utilidad. En el *menú principal* aparecen distintos elementos que pueden verse en la figura 1.1.

Figura 1.1. Menú principal



La barra de título es la barra azul que aparece en la parte superior de la ventana en la que se observa el nombre del programa (es decir, *Statgraphics Plus*) seguido por el nombre del archivo *StatFolio*, en caso de que haya alguno abierto. Si no hay ninguno abierto aparece el letrero: StatFolio sin Nombre.

La *barra de menú principal* contiene todas las opciones generales que puede ejecutar el programa:

1. *Archivo* es la opción que maneja los ficheros de datos, abre y cierra los ficheros, puede juntar varios o separar un fichero en partes.
2. *Edición* es el editor de ficheros que sirve para grabar datos nuevos, modificar los existentes o transformar las variables.
3. *Gráficos* proporciona diversos gráficos.
4. *Descripción*, *Comparación*, *Dependencia*, *Avanzado* y *SnapStats* remiten a una serie de procedimientos estadísticos.
5. *Ver* controla lo que vemos en la pantalla.
6. *Ventana* permite pasar de una a otra ventana o modificar las ventanas.
7. *Ayuda* proporciona ayuda de diverso tipo.

La *barra de herramientas* está compuesta por diferentes iconos que permiten acceder rápidamente a las opciones más comunes en el trabajo sin necesidad de acudir al menú general. El significado de cada icono puede verse si se apoya el indicador del ratón sobre el propio icono. Si se observa en la barra en la figura 1.1, de izquierda a derecha, encontraremos una serie de iconos que permiten:

- Abrir StatFolio; Guardar StatFolio
- Abrir fichero de datos; Guardar fichero de datos
- Cortar; Pega, Deshacer, Imprimir
- Vista preliminar
- Gráfico de dispersión; Gráfico de caja; Histograma
- Resumen estadístico; Regresión múltiple
- Gráficos X-bar y R; Análisis de capacidad; Predicción
- Abrir archivo de diseño
- Análisis cluster
- Modelos lineales generales
- StatAdvisor (Ayuda en la interpretación del análisis estadístico)
- StatWizard; Ayuda

La barra de ventanas aparece en la parte inferior figura 1.1 y presenta cinco iconos, que se utilizan para abrir y cerrar diferentes ventanas activas (de izquierda a derecha):

1. El primero <Sin nombre Comentarios> es similar a un bloc de notas donde escribir comentarios sobre el StatFolio sobre el que estamos trabajando.
2. El segundo icono, StatAdvisor, se utiliza para abrir una ventana en la que siempre aparece una ayuda sobre la interpretación de los resultados obtenidos en cada análisis realizado.
3. El tercer icono, StatGallery, se utiliza para activar presentaciones gráficas.
4. El cuarto StatReporter, es la ventana de un editor de texto en el que

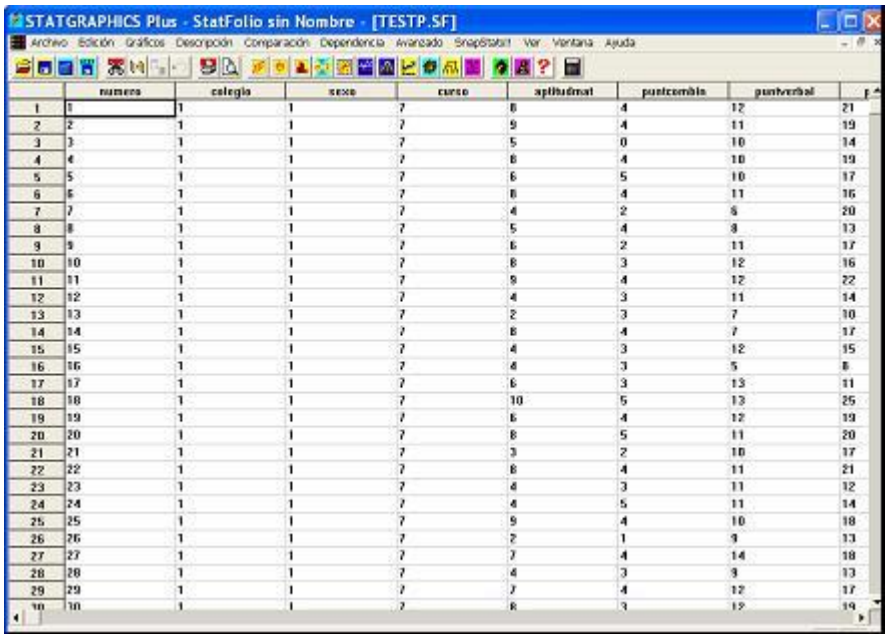
pueden copiarse los análisis realizados y editarlos para luego copiarlo a un procesador de textos.

5. Por último, el quinto icono proporciona acceso al editor de datos, informando sobre el conjunto de datos que se está utilizando.

1.8. EDITOR DE DATOS

Los datos se introducen a través de la ventana de editor de datos, que es similar a una hoja de cálculo, aunque se diferencia de ellas en que, directamente de esta ventana no se pueden producir gráficos. Se accede a la misma pulsando sobre el icono <Sin nombre> y se obtiene la ventana de la figura 1.2 (con las celdas en blanco).

Figura 1.2. Editor de datos con fichero de datos abierto



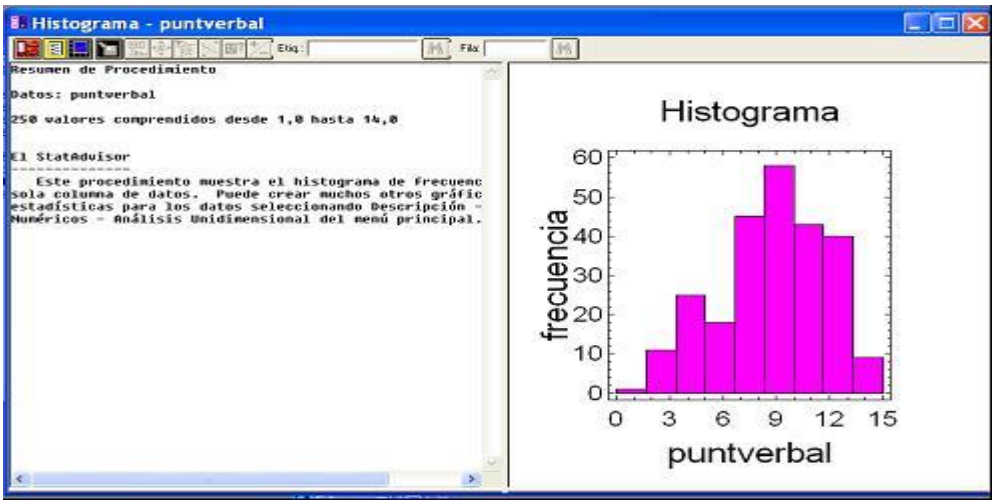
En la tabla de datos las columnas van a ser las variables que se utilizarán en los análisis estadísticos y que habrá que definir por un nombre y especificando sus características. Cada fila representa una unidad estadística. En el caso de que ya tuviéramos un fichero de datos abierto, como TESTP el icono aparecería con el nombre del fichero y al pulsarlo aparecería la ventana de la figura 1.2.

1.9. VENTANA DE RESULTADOS DEL ANÁLISIS

Después de ejecutar un procedimiento estadístico cualquiera, Statgraphics presenta una ventana de resultados del procedimiento que permite interactuar pidiendo nuevos resultados y ofrece un marco de trabajo similar para todos los procedimientos. Por ejemplo, si realizamos un histograma de frecuencias para la variable *puntverbal* del fichero TESTP (puntuación de aptitud verbal en el cuestionario de probabilidad), se obtiene la ventana de análisis que aparece en la figura 1.3. En esta ventana de análisis se observan tres zonas esenciales:

- *La barra de título del análisis* presenta el nombre del procedimiento estadístico cuyos resultados se muestran en la ventana de análisis (en este caso se trata del procedimiento *Histograma*). A continuación aparece el nombre de la variable analizada (en nuestro caso se trata de la variable *puntverbal)*.

Figura 1.3. Ventana de resultados del análisis



- *La barra de herramientas de análisis* presenta una sucesión de iconos que van a posibilitar las diferentes opciones de trabajo en el análisis actual. De izquierda a derecha, el primer icono lleva a la ventana de entrada de variables el segundo icono, etiquetado como *Opciones tabulares*, se utiliza para presentar todas las posibles subopciones con resultados numéricos que permite el procedimiento que estemos utilizando. El tercer icono, etiquetado como *Opciones gráficas*, se utiliza para presentar todas las posibles subopciones con resultados

gráficos que permite el procedimiento que se esté utilizando. El cuarto icono, etiquetado como *Guardar resultados*, se utiliza para guardar los resultados numéricos del análisis estadístico en variables que indicaremos en la pantalla correspondiente.

- *La salida de resultados* se sitúa debajo de la barra de herramientas de análisis y se divide en dos zonas (ver figura 1.3).
- La zona de la izquierda (zona de texto) presenta los resultados numéricos del análisis estadístico y la zona de la derecha (zona de gráficos) ofrece los resultados gráficos. Una vez obtenidas las salidas gráficas, si se pulsa dos veces con el ratón sobre cualquiera de las ventanas en que está dividida la pantalla, ésta se maximizará y ocupará toda la pantalla. Se regresa a la situación anterior volviendo a pulsar dos veces con el ratón en cualquier parte de la pantalla maximizada.

1.10. GRABACIÓN DE DATOS Y OPERACIONES CON FICHEROS

En este programa podemos analizar dos tipos de ficheros: a) Ficheros previamente grabados por el profesor que os proporcionará estos ficheros, justo con su descripción, así como del objetivo de la investigación en que se recogieron los datos; b) Ficheros grabados por vosotros mismos.

Abrir un fichero ya grabado

Cuando queremos trabajar con *un fichero de datos que ha sido grabado previamente*, lo primero que se deberá hacer antes de realizar cualquier acción es abrir el fichero con el que deseamos trabajar, es decir, cargarlo desde el disco a la memoria del ordenador.

Para *cargar en memoria el fichero* debemos realizar los siguientes pasos: Ir al menú Archivo (menú de manejo de ficheros), seleccionar la opción Abrir Datos (abrir un fichero de datos).

Figura 1.4. Pantalla de selección del fichero de datos

Aparece en la pantalla la lista de los ficheros disponibles, como se muestra en la figura 1.4. Debemos seleccionar el archivo deseado y pulsar sobre el botón que dice Abrir. Veremos que el fichero se carga en la memoria. Podemos comprobarlo pasando al editor de datos a partir de la opción Ventana en el menú principal, donde ahora aparece el nombre del fichero como título de la ventana de editor de datos. Podemos explorar el fichero, ver qué variables contiene y el significado de los códigos.

Grabar un fichero

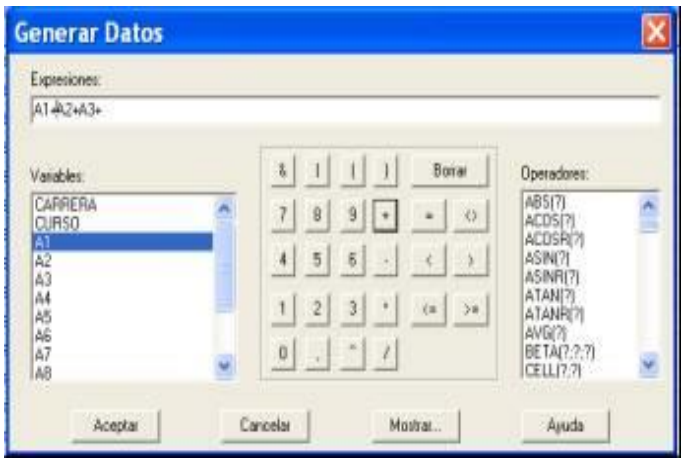
Para grabar un *nuevo fichero de datos* hay que usar el editor de datos, donde cómo hemos dicho aparece una cuadrícula parecida a la que se usa en una hoja electrónica. Cada fila representa una unidad estadística y cada columna una variable. Podemos grabar un nuevo fichero simplemente introduciendo datos en la cuadrícula y grabando al finalizar el fichero producido mediante la opción Guardar Datos como, dando un nombre al fichero de datos, con un proceso similar al anterior. Es necesario también dar un nombre significativo a las variables, ya que si no damos nombres el programa por defecto les asigna el nombre Col_1, Col_2, etc. Para ello se selecciona la columna y luego se pincha sobre ella con el botón derecho del ratón, apareciendo el menú de la figura 1.5; en él seleccionar la opción Modificar columna que nos permite dar un nombre a la variable y definir su tipo: numérica, carácter, entera o bien con un número fijo de decimales.

Cálculo de nuevas variables

A veces queremos generar una variable nueva a partir de las grabadas, por ejemplo, supongamos que en el fichero ACTITUD queremos sumar las tres primeras puntuaciones. Para ello se debe definir el nombre de la variable y en consecuencia, de la columna en la que se ubicará dicha variable, para ello se selecciona la columna y luego se hace clic sobre ella con el botón derecho del ratón, volverá a aparecer el menú de la figura 1.6, en él seleccionar la opción Modificar Columna entonces aparecerá un cuadro de diálogo como el de la figura 1.5.

Figura 1.5. Modificar una variable

Figura 1.6. Generar datos



En ese menú, en el campo Nombre, colocar el nombre de la nueva variable, en este caso TOTAL. En el campo Tipo, se puede seleccionar el tipo de variable con la que se desea trabajar y el número de dígitos que se desea utilizar, si es una variable numérica. Una vez que se definen las características de la nueva variable, para *generar los datos*, se vuelve a seleccionar la columna y pulsando con el botón derecho, seleccionar la opción *Generar Datos*, aparecerá un cuadro de diálogo como el de la figura 1.6, que funciona como una calculadora, donde podemos seleccionar diversas expresiones algebraicas, como podéis ver en las ventanas.

En el campo Expresiones, debemos escribir la fórmula por la que se generarán los nuevos datos, en nuestro ejemplo, la variable nueva es la suma de las tres primeras puntuaciones. Por último, se graba el fichero de datos de la misma forma que en la sección anterior, para que no se pierdan

los cálculos.

1.11. RECODIFICACIÓN DE DATOS

Cuando introducimos una variable en el fichero usamos con frecuencia *códigos*. Por ejemplo, para representar el sexo de un alumno, podemos codificar los varones con un "1" y las mujeres con un "2". A veces, sucede que es necesario recodificar variables ya definidas de un fichero, por ejemplo, podríamos estar interesados en representar los valore 1 a 5 mediante el código 1 y los 6 a 10 mediante el código 2.

Figura 1.7. Selección de la variable

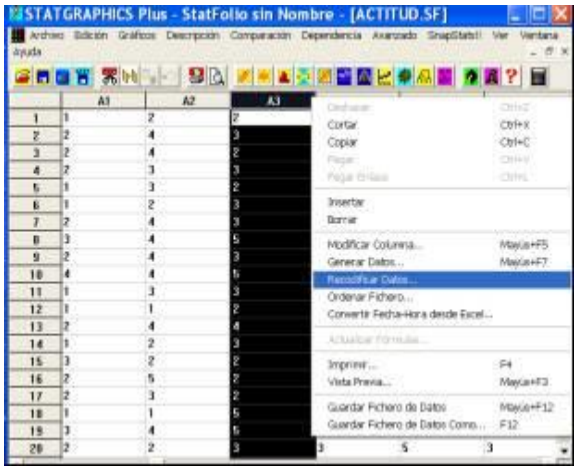


Figura 1.8. Ingresar los nuevos valores

Para recodificar la variable se selecciona la columna correspondiente colocando el indicador del ratón sobre el título de la variable y pulsando sobre ella. Pulsando el botón derecho del ratón, aparecerá un menú desplegable como el de la figura 1.7. Seleccionando la opción Recodificar Datos, aparecerá un menú como el de la figura 1.8, donde se debe escribir los nuevos límites y valores por medio de la cual se realizará la recodificación, también puede elegirse el tipo de intervalo.

TEMA 2

TABLAS DE FRECUENCIAS Y GRAFICOS

2.1. VARIABLES ESTADÍSTICAS CUALITATIVAS. FRECUENCIAS

Cuando se comienza a analizar una nueva variable estamos interesados es saber los valores que puede tomar, el número total de datos y cuántas veces aparecen los diferentes valores. La distribución de una variable nos proporciona esta información.

Ejemplo 2.1. El censo Estadístico de 1980 para la provincia de Jaén presenta la tabla 2.1., que tiene datos simplificados de población activa, clasificada por su relación laboral con la empresa en que trabaja:

Tabla 2.1. Población activa de Jaén (1980) según relación laboral

Relación laboral	Frecuencia
Patronos	4,548
Trabajadores autónomos	17,423
Cooperativistas	2,406
Empleados fijos.	61,935
Trabajadores eventuales	47,358
Trabaja en empresa familiar	3,580
Otros	,98
Total	138,248

Ya hemos indicado que las variables estadísticas cualitativas son aquellas que estudian características de una población o muestra que esencialmente no son numéricas. También sabemos que estas variables se pueden medir con una escala nominal. Ejemplos de variables estadísticas cualitativas son: el sexo, profesión, estado civil, etc, de los habitantes de una ciudad o la relación laboral de los componentes de la población activa.

Frecuencias absolutas

Para poder operar con los datos de la tabla 2.1 o referirnos a ellos, podemos representar la característica a observar (la relación laboral) mediante la variable X y a la modalidad número i de dicha variable con la notación x_i; f_i representará el número de individuos que presentan esa modalidad, que se llama *frecuencia absoluta.*

Frecuencias relativas

Los datos de la tabla 2.1 proporcionan exactamente el número de personas que pertenecen a un determinado sector profesional. Pero decir que en la provincia de Jaén existen 4.548 patronos, nos proporciona poca información sobre si el número de patronos es muy significativo, respecto al total de la población ocupada. Para valorar la representatividad de cada categoría respecto al total de datos se calcula la *frecuencia relativa* h_i, dividiendo la frecuencia absoluta f_i por el número total de observaciones (N), es decir:

(2.1) $$h_i = f_i/N$$

En la tabla 2.2 podemos observar que la suma de las frecuencias relativas es uno y que la frecuencia relativa de la modalidad patronos es 0.033, lo que significa que de cada 1000 personas ocupadas en Jaén en 1980 33 eran patronos.

Tabla 2.2. Frecuencias relativas y porcentajes del tipo de relación laboral en la población activa de Jaén (1980)

X_i	f_i	h_i	%
Patronos	4,548	0,033	3,3
Trabajadores autónomos	17,423	0,126	12,6
Cooperativistas	2,406	0,017	17,0
Empleados fijos.	61,935	0,448	44,8
Trabajadores eventuales	47,358	0,343	34,3
Trabaja en empresa familiar	3,580	0,026	2,6
Otros	998	0,007	0,7
Total	l38248	1,000	100

Porcentajes

En lugar de utilizar frecuencias relativas, usualmente se utilizan los porcentajes, que se calculan multiplicando la frecuencia relativa por 100.

Actividades

2.1. ¿Cuáles son los motivos para construir una tabla de frecuencias en lugar de usar el listado de los datos tal y como se recogen?

2.2. Supongamos que en una muestra de n elementos la frecuencia absoluta de la categoría A es n_A. ¿Cuál será el valor de la nueva frecuencia absoluta y relativa si añadimos a la muestra un nuevo sujeto que pertenezca a la categoría A?

2.3. En una muestra de 6000 estudiantes el 35% practica regularmente algún deporte. ¿Cuál es la frecuencia absoluta y relativa de estudiantes que practica algún deporte?

2.2. DIAGRAMA DE BARRAS Y GRÁFICOS DE SECTORES

Aunque una tabla de frecuencias nos proporciona un resumen de los datos, en la práctica hay que observar, generalmente, más de un conjunto de datos, compararlos, conseguir una apreciación global y rápida de los mismos. Esto se ve facilitado mediante una adecuada representación gráfica. Los gráficos más usuales para variables cualitativas y discretas son: diagramas de barras y gráficos de sectores.

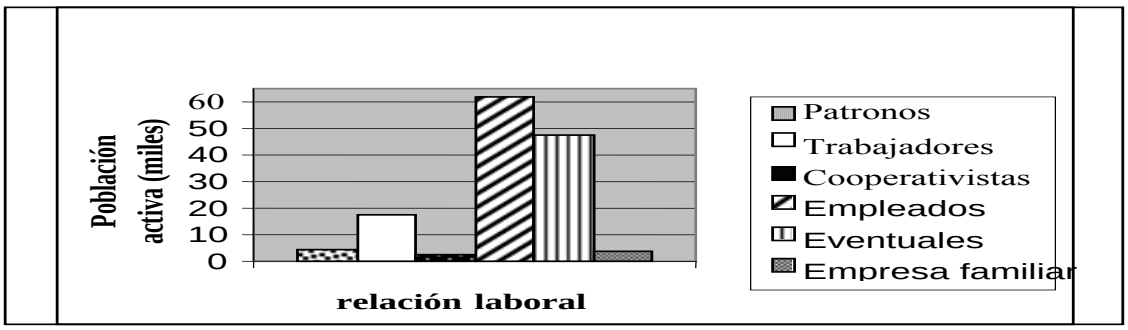


Diagrama de barras

Es una representación gráfica en la que cada una de las modalidades del carácter se representa mediante una barra. En este gráfico se suelen disponer los datos en el primer cuadrante de unos ejes de coordenadas, levantando sobre el eje de abscisas un bloque o barra para cada modalidad

de la variable observada. La altura de la barra ha de ser proporcional a la frecuencia absoluta o relativa, que se representará en el eje de ordenadas. En la figura 2.1 podemos observar los diagramas de barras correspondientes a la tabla 2.1.

Gráfico de sectores

Si lo que nos interesa es información sobre el "peso" que una de las modalidades observadas tiene en relación con el total y al mismo tiempo con las demás, podemos representar los datos en un diagrama de sectores, que consiste en representar cada modalidad por un sector circular, cuyo ángulo central y, por lo tanto también su área, es proporcional a la frecuencia. Una forma sencilla de construirlo es multiplicando la frecuencia relativa por 360; así obtendremos la amplitud del ángulo central que tendrá cada una de las modalidades observadas. El gráfico de sectores correspondiente a la tabla se muestra en la figura 2.2.

Figura 2.2. Gráfico de sectores

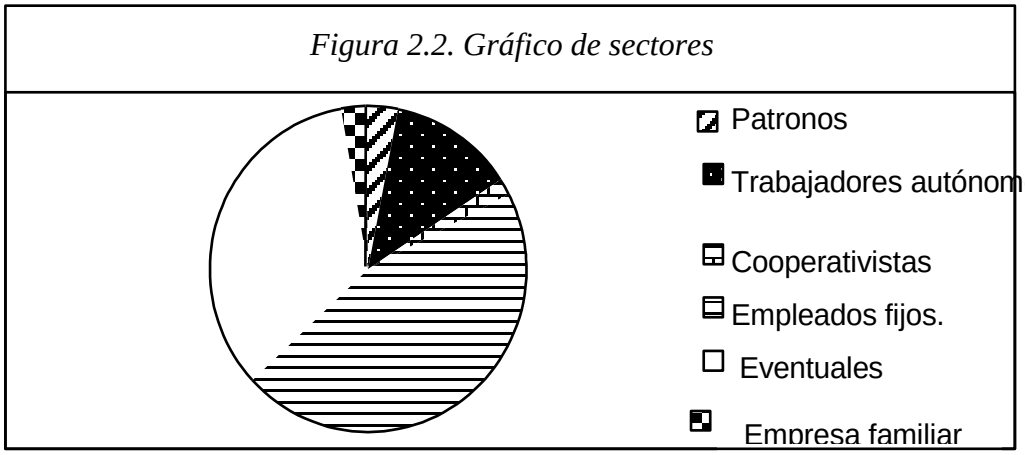


Figura 2.2. Gráfico de sectores

2.3. OBTENCIÓN DE TABLAS DE VARIABLES CATEGÓRICAS CON STATGRAPHICS

Para preparar una *tabla de frecuencias de datos cualitativos,* o de *variables discretas* con pocos valores, elegimos el menú Descripción, y dentro de él Datos Cualitativos. Se selecciona Tabulación y en la ventana de diálogo que aparece, se selecciona la variable que se desea. Por ejemplo, en la figura 2.3, se ha seleccionado la variable A2. Luego aparece una ventana de análisis como la de la figura 3, en la cual seleccionando el botón Opciones Tabulares, se podrá seleccionar la Tabla de frecuencias como se observa en la figura 2.4. Para representar un *diagrama de barras o un diagrama de sectores*, en la ventana de análisis se selecciona el botón

Opciones Gráficas, y allí aparecerá un cuadro de diálogo, en el que podrán seleccionarse ambos diagramas, diagrama de barras y diagrama de sectores. Si seleccionamos los dos gráficos y la tabla de frecuencias se obtendrá una pantalla similar a la de la figura 2.5.

Figura 2.3. Selección de una variable

Figura 2.4. Selección de gráficos

En la tabla 2.3 presentamos la tabla de frecuencias de la variable A2 en el fichero ACTITUDES (Proyecto 2), es decir, la puntuación otorgada por los alumnos en la pregunta 2: ¿Es necesario conocer algo de estadística para ser un consumidor inteligente?

Figura 2.5. Tabla de frecuencias y gráficos de barras y de sectores

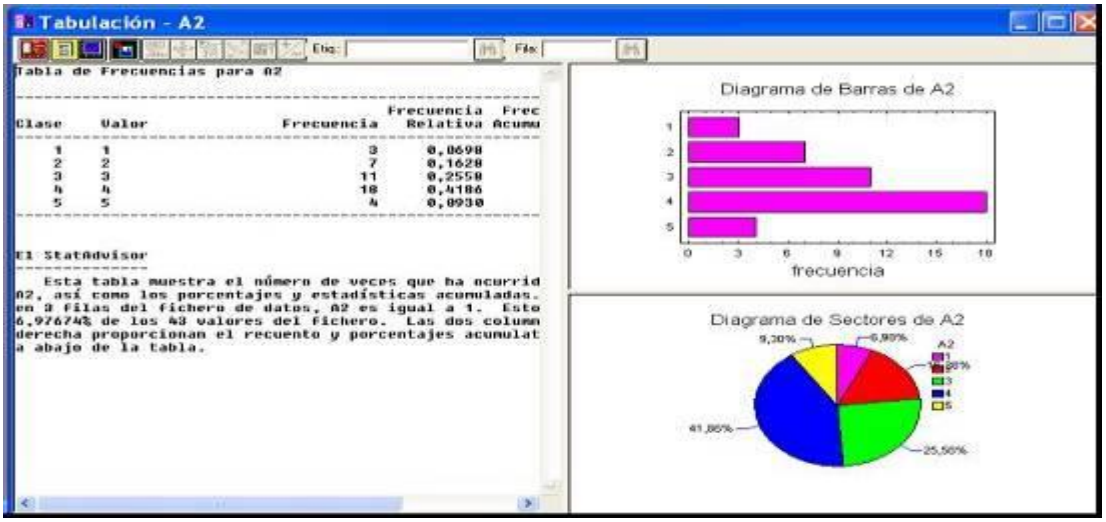
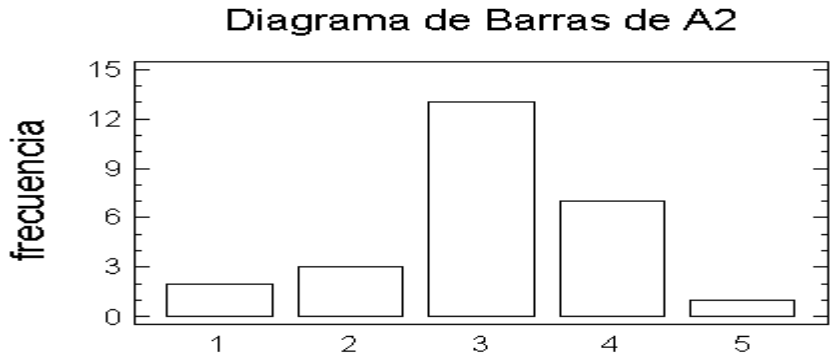


Tabla 2.3. Tabla de frecuencias para A2

Clase	Valor	Frecuencia	Frecuencia Relativa	Frecuencia Acumulativa	Frecuencia Acum.Rel.
1	1	2	0,0769	2	0,0769
2	2	3	0,1154	5	0,1923
3	3	13	0,5000	18	0,6923
4	4	7	0,2692	25	0,9615
5	5	1	0,0385	26	1,0000

También dentro de DATOS CUALITATIVOS, en OPCIONES GRÁFICAS se pueden realizar el diagrama de barras y el diagrama de sectores. Cada uno de ellos tiene diversas opciones que pueden explorarse pulsando la tecla OPCIONES DE VENTANA. Como ejemplo, en la figura 2.6 representamos el diagrama de barras correspondiente a la tabla 2.3.

Figura 2.6. Diagrama de barras de puntuación en utilidad de la estadística



Modificación de las opciones por defecto

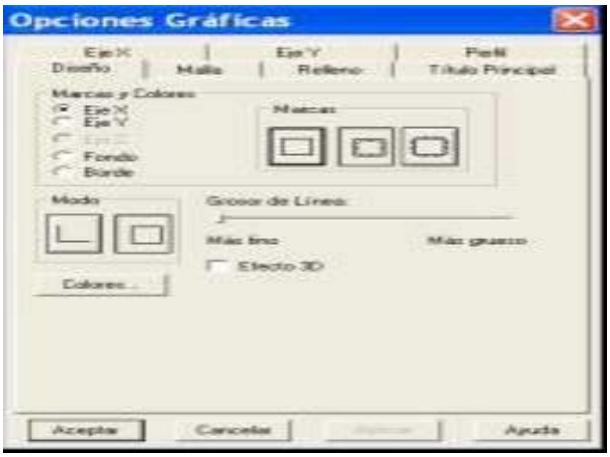
Generalmente, en todos los menús y opciones del programa, hay características que están definidas por defecto, como por ejemplo el color de los gráficos, o el tipo de resúmenes estadísticos que se calculan, pero que pueden modificarse. Un ejemplo es el aspecto de los gráficos. A continuación describimos la forma en que pueden cambiarse. Otras opciones del programa se cambian en forma similar, de modo que el ejemplo de modificación de los gráficos puede ser útil para aprender a modificar otras opciones por defecto. Os animamos a que en los diferentes programas que vamos a manejar exploréis las diversas opciones.

Cambio de las características gráficas

Los gráficos están definidos con un fondo blanco, el gráfico propiamente dicho en rosa y el texto en negro. Es interesante cambiar estas características para poder realizar una buena impresión en blanco y negro. Para ello deberemos realizar los pasos que detallamos a continuación, teniendo el gráfico en pantalla.

El color del marco y del fondo se cambia pulsando el borde del gráfico con el botón derecho del ratón. Aparecerá un menú, y pulsando en Opciones Gráficas aparecerá un cuadro de diálogo, como en la figura 2.7.

Figura 2. 7. Cuadro de selección de características gráficas



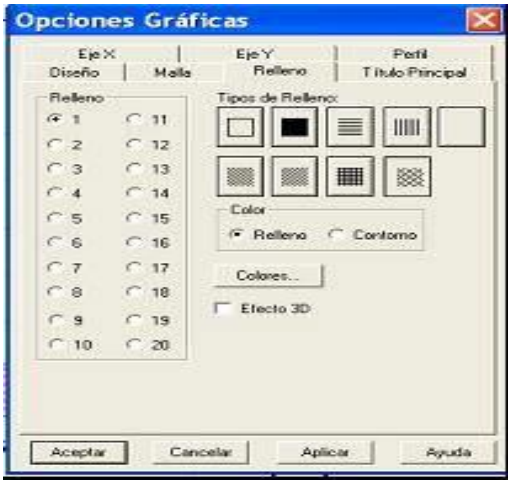
Para cambiar el color de fondo, se pulsa en la pestaña Diseño y luego en la opción Fondo. Si pulsamos en Colores aparecerá un cuadro de diálogo donde podemos seleccionar el color para el fondo.. Se volverá al cuadro de la figura 2.8. Para que se aplique el cambio tenemos que pulsar el botón Aplicar. En este cuadro, y pulsando Borde, podemos modificar el color del borde en la misma forma o el de los Ejes X e Y pulsando sus respectivas opciones.

Figura 2.8. Cuadro de selección de color

El menú mostrado en la figura 2.9 también nos permite cambiar las escalas del gráfico. Pulsando las pestañas Eje X y Eje Y aparece un menú donde podemos variar el origen y extremo de cada uno de los ejes y el número de divisiones mostradas. También podemos ponerle título a los ejes, cambiar el color de las fuentes e incluso la orientación del texto. La pestaña Relleno nos permite cambiar el tipo de relleno del gráfico y el contorno. Cuando pulsamos en esta pestaña aparecerá el menú de la figura 2.9.

Figura 2.9. Selección del relleno

Figura 2.10.Características de texto



En el caso de que se deseen las barras rayadas, se debe primero seleccionar cualquiera de los botones que se ven en la figura 2.10 y luego entrar al botón Colores.. y seleccionar el negro. Podemos poner un título al gráfico en la pestaña Título Principal. En la casilla Título escribiremos el título que queremos poner a nuestro gráfico. Para cambiar *las características del texto*, pulsaremos en el botón Línea 1 Fuente... Aparecerá un menú donde podemos cambiar la fuente, el tamaño, y el color del texto En el caso de que se desee cambiar la orientación del texto, se pulsa la opción Vertical.

2.4. VARIABLES CUANTITATIVAS: FRECUENCIAS ACUMULADAS

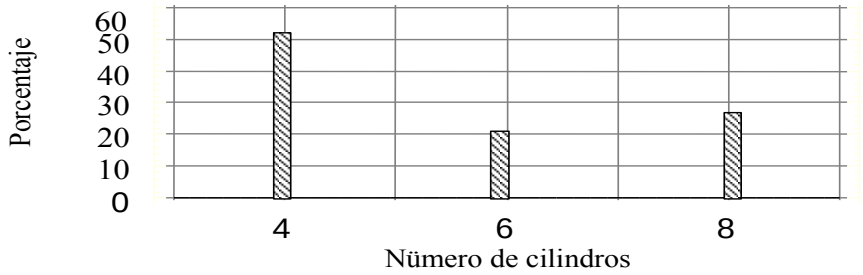
Las tablas estadísticas para variables cuantitativas discretas son similares a las anteriores, aunque, en este caso, la variable aparece ordenada.

Ejemplo 2.2. En la tabla 2.4 presentamos la distribución de frecuencias del número de cilindros de un conjunto de 398 tipos de automóviles de diferentes marcas y modelos, fabricados en Europa, Japón y Estados Unidos y en la figura 2.11 representamos el diagrama de barras correspondiente.

Tabla 2.4. Distribución del número de cilindros en coches de diferentes modelos

Número de cilindros	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia relativa acumulada
4	207	0,5201	207	0,5201
6	84	0,2111	291	0,7312
8	107	0,2688	398	1,0000
Total	398	1,0000		

Figura 2.11. Distribución del número de cilindros enautomóviles



Muchas veces la característica a observar toma valores numéricos aislados (generalmente números enteros); es el caso, por ejemplo, del número de monedas que una persona lleva en el bolsillo o el número de hijos de una familia. Nótese que algunas variables como el número de glóbulos rojos por mm3 que tiene una persona, que son esencialmente discretas pueden, por conveniencia, ser tratadas como continua.

Frecuencias acumuladas

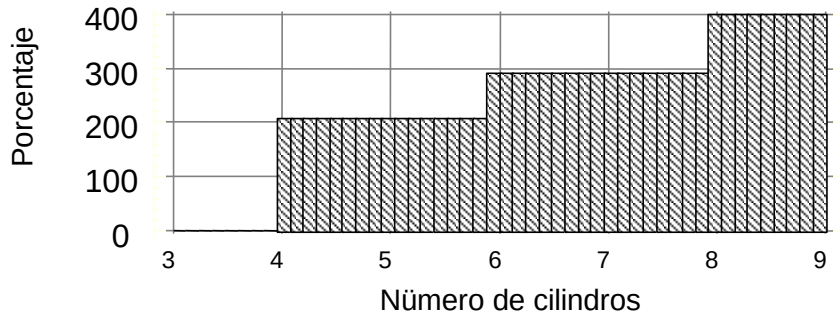
Algunas veces, es interesante conocer el número de valores de una variable estadística que son menores que un valor dado. Para conseguir esto, se calculan las *frecuencias absolutas acumuladas,* que se obtienen sumando a la frecuencia absoluta de un valor todas las anteriores. De igual forma se calculan las *frecuencias relativas acumuladas.*

En la tabla 2.4 podemos interesarnos por conocer cuántos coches en la muestra tienen x o menos cilindros. Esto se puede observar en la cuarta columna de la tabla 2.4, donde observamos que 291 (73%) de los coches tienen 6 o menos cilindros. Todas estas observaciones serán más rápidas si

tenemos una representación gráfica de las frecuencias absolutas acumuladas y de las frecuencias relativas acumuladas. Para ello basta dibujar un *diagrama de frecuencias acumuladas.*

Para construirlo, representamos en el eje de abscisas los valores de la variable. Para cada uno de estos valores, levantamos sobre el eje de abscisas una línea de altura proporcional a la frecuencia acumulada. Trazando desde el extremo de cada línea una paralela al eje X, que corte a la línea siguiente, se completa el diagrama, como se muestra en la figura 2.12. En esta gráfica podemos ver cómo las frecuencias acumuladas experimentan un aumento en cada valor de la variable.

Figura 2.12. Distribución del número de cilindros en automóviles



Actividades

2.4. Sabiendo que la frecuencia absoluta de alumnos que tiene 3 hermanos es 30 y que la frecuencia acumulada de alumnos que tiene hasta 3 hermanos es 80. ¿Cuántos alumnos tienen 2 hermanos o menos?

2.5. ¿Por qué la representación gráfica de la frecuencia acumulada nunca puede ser decreciente?

2.6. Pensar en algunas situaciones en que interese la frecuencia acumulada para una variable.

2.4. VARIABLES AGRUPADAS: INTERVALOS DE CLASE

Algunas variables cuantitativas toman valores aislados -variable discreta- (número de hijos en un matrimonio). Otras veces la variable puede tomar cualquier valor dentro de un intervalo, en cuyo caso se *llama variable continua*; por ejemplo, el peso, la temperatura corporal, la velocidad de un coche. Tanto estas variables como las variables discretas con un número grande de valores (habitantes de un país, número de hojas en un árbol) se suelen agrupar en intervalos al elaborar las tablas de frecuencia.

Intervalos y marcas de clase

Ejemplo 2.3. En la tabla 2.5 presentamos las alturas de un grupo de alumnos universitarios. Al haber 26 valores diferentes obtendríamos una tabla de frecuencias poco apropiada, si estudiásemos la frecuencia de cada uno de los valores aislados, ya que estamos trabajando con una muestra de solo 60 individuos.

Tabla 2.5. Datos de alturas de un grupo de universitarios

150	160	161	160	160	172	162	160	172	151	161	172	160
169	169	176	160	173	183	172	160	170	153	167	167	175
166	173	169	162	178	170	179	175	174	174	160	149	162
161	168	170	173	156	159	154	156	160	166	170	169	164
168	171	178	179	164	176	164	182					

Para resumir la información y adquirir una visión global y sintética de la misma, agruparemos los datos en intervalos. No obstante, esta operación implica una pérdida de información que será preciso tener en cuenta en la interpretación de las tablas, gráficos y estadísticos de datos agrupados.

2.5. OBTENCIÓN DE TABLAS DE FRECUENCIAS AGRUPADAS CON STATGRAPHICS

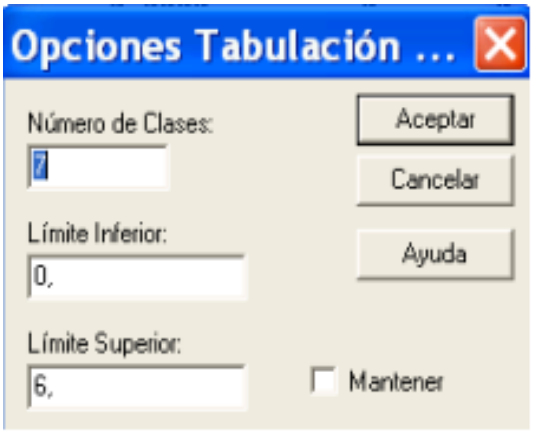
El menú Descripción, en la opción Datos Numéricos – Análisis Unidimensional contiene diferentes procedimientos numéricos y gráficos para hacer un estudio descriptivo de una variable numérica.

Tablas de frecuencias agrupadas en intervalos

Para construir *una tabla de frecuencias para datos cuantitativos,* entramos en el menú Descripción, en la opción Datos Numéricos – Análisis Unidimensional. Aparecerá un cuadro en el que se debe seleccionar la variable que se desea estudiar y luego se observará una ventana de análisis y en ella entrar al botón Opciones Tabulares, aparecerá un menú como el de la figura 2.13, allí seleccionar la tabla de frecuencias.

Figura 2.13. Selección de opciones tabulares

Figura 2.14. Opciones en intervalos



Teniendo la tabla de frecuencias en pantalla, si se desea, se puede cambiar la cantidad de intervalos y los límites superior e inferior, para ello hacer clic con el botón derecho sobre la tabla, se observará un menú, seleccionar Opciones de Ventana, y aparecerá un cuadro como el de la figura 2.14, allí se podrá modificar el número de intervalos en el campo Número de clases y los límites inferior y superior en los campos.La tabla 2.6 contiene una distribución de frecuencias de la variable "altura" en el conjunto de alumnos dado, agrupada en intervalos de amplitud 10, obtenida con el programa Statgraphics.

Tabla 2.6. Tabla de Frecuencias para altura

Tabla 2.6. Distribución de frecuencias de la variable altura

Clase	Límite Inferior	Límite Superior	Marca	Frecuencia	Frecuencia Relativa	Frecuencia Acumulativa	Frecuencia Acum.Rel.
menor o igual		150,0		0	0,0000	0	0,0000
1	150,0	155,0	152,5	2	0,0333	2	0,0333
2	155,0	160,0	157,5	7	0,1167	9	0,1500
3	160,0	165,0	162,5	19	0,3167	28	0,4667
4	165,0	170,0	167,5	11	0,1833	39	0,6500
5	170,0	175,0	172,5	10	0,1667	49	0,8167
6	175,0	180,0	177,5	5	0,0833	54	0,9000
7	180,0	185,0	182,5	5	0,0833	59	0,9833
8	185,0	190,0	187,5	0	0,0000	59	0,9833
9	190,0	195,0	192,5	1	0,0167	60	1,0000
10	195,0	200,0	197,5	0	0,0000	60	1,0000
mayor	200,0			0	0,0000	60	1,0000

Media = 168,467 Desviación típica = 8,40594

La primera decisión que hay que tomar para agrupar una variable es el número de intervalos en que se debe dividir. No existe una regla fija, y

en última instancia será un compromiso entre la pérdida de la información que supone el agrupamiento y la visión global y sintética que se persigue. Una regla que se utiliza a menudo es tomar un entero próximo a la raíz cuadrada del número de datos como número de intervalos. Para proceder a la construcción de una distribución de frecuencias con datos agrupados es preciso tener en cuenta las siguientes nociones:

- *Máximo:* Se llama máximo de una variable estadística continua al mayor valor que toma la variable en toda la serie estadística. Ejemplo: En el caso de la talla de los alumnos el máximo es 183.
- *Mínimo* de una variable estadística es el menor valor que toma la variable en toda la serie estadística. Ejemplo: En el caso de la talla de los alumnos el mínimo es 149.
- *Recorrido*: es la diferencia entre el máximo y el mínimo en una serie estadística. En nuestro caso será, 183 - 149 = 34 cm.
- *Clase*: Se llama clase a cada uno de los intervalos en que podemos dividir el recorrido de la variable estadística. Ejemplo: cada uno de los intervalos de la tabla anterior (145.0-150.0; 150.0-155.0;...). Los intervalos pueden ser o no de la misma amplitud.
- *Extremo superior de clase:* Es el máximo valor de dicha clase; lo representaremos por E_{i+1}.
- *Extremo inferior de clase:* Es el mínimo valor del intervalo; lo representaremos por E_i.
- *Marca de clase*: Es el punto medio de cada clase; se representa por x_i y es la media de los extremos de la clase.

Según que el extremo superior de cada clase coincida o no con el extremo inferior de la clase siguiente, podemos distinguir dos tipos de tablas de frecuencias. Si el extremo superior de cada clase coincide con el inferior de la siguiente, los intervalos se suponen semiabiertos por la derecha. Es decir, en cada clase se incluyen los valores de la variable que sean mayores o iguales que el extremo inferior del intervalo, pero estrictamente menores que el extremo superior. Los programas utilizados en los apuntes hacen este convenio.

Algunos autores no hacen coincidir el extremo superior de una clase con el extremo inferior de la siguiente. Por ejemplo, para construir una

tabla de frecuencias para los datos de las alturas se podrían utilizar intervalos del tipo: 145-149, 150-154,..... En este caso, quedará la duda de cual es el intervalo al que habría de asignarse un supuesto valor 149.3.

Para resolver esta indeterminación teórica (ya que en la práctica, tales valores no pueden aparecer, al no haber tomado cifras decimales), y para una construcción más adecuada de los gráficos, se definen en este caso los *extremo reales de clase*: se determinan añadiendo al extremo superior de cada clase la mitad de la diferencia que lo separa de la contigua, y la otra mitad de la diferencia se resta al extremo inferior de la clase contigua. De este modo se trabaja como si en la tabla original se hubieran ampliado los intervalos, aunque se conserva la marca de clase. Estos extremos reales de clase serán los utilizados para efectuar el resto del análisis.

Actividades

2.7. Indica algunos aspectos positivos y negativos de la agrupación de datos en intervalos de clase.

2.8. ¿Cuándo se pierde más información sobre los datos originales, al tomar intervalos de clase grandes o pequeños?

2.9. Indica algunos criterios para elegir el número de intervalos en una tabla de frecuencia.

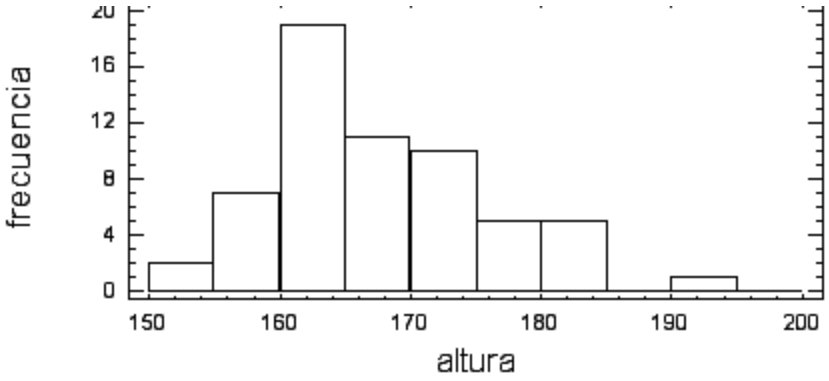
2.6. HISTOGRAMAS Y POLÍGONOS DE FRECUENCIAS

La información numérica proporcionada por una tabla de frecuencias se puede representar gráficamente de una forma más sintética. En el caso de las variables agrupadas las representaciones que se utilizan frecuentemente son los *histogramas* y los *polígonos de frecuencias.*

Un histograma se obtiene construyendo sobre unos ejes cartesianos unos rectángulos cuyas áreas son proporcionales a las frecuencias de cada intervalo. Para ello, las bases de los rectángulos, colocadas sobre el eje de abscisas, serán los intervalos de clase y las alturas serán las necesarias para obtener un área proporcional a la frecuencia de cada clase.

En la figura 2.15 se presenta el histograma de frecuencias de la variable altura del conjunto de datos ALUMNOS, correspondientes a la tabla 2.6.

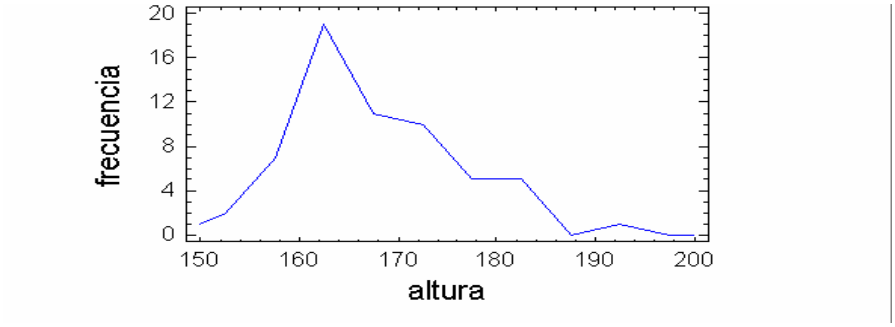
Figura 2.15. Histograma para la variable altura



Polígono de frecuencias

Otra forma de representar los datos es el polígono de frecuencias, que es la línea que resulta de unir los puntos medios de las bases superiores de los rectángulos de un histograma de frecuencias. En la figura 2.16 se representa un polígono de frecuencias de la variable altura del conjunto de los datos de la tabla 2.6.

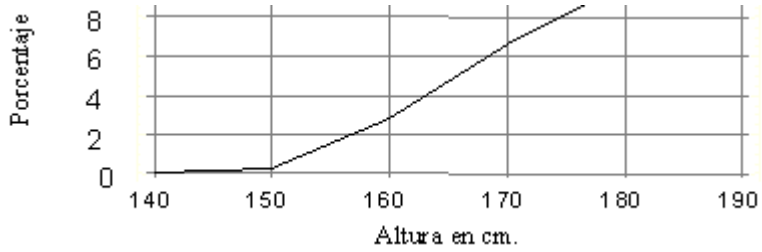
Figura 2.16. Polígono de frecuencias para la variable altura



Polígono acumulativo de frecuencias

La figura 2.17 representa el polígono acumulativo de frecuencias para la variable altura del conjunto de ALUMNOS. Se obtiene uniendo los puntos cuyas coordenadas son: la abscisa corresponde al extremo superior de cada clase y la ordenada a la frecuencia (absoluta o relativa) acumulada hasta dicha clase.

Figura 2.17. Polígono acumulativo de frecuencias

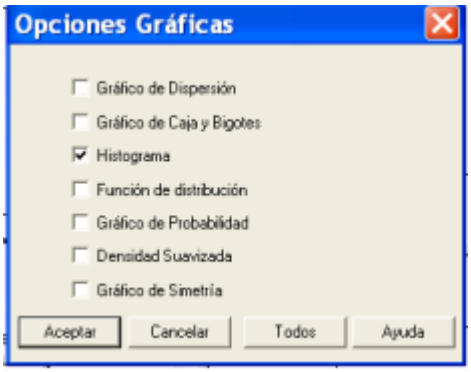


Para realizar el histograma y polígonos con STATGRAPHICS

Si se desea obtener un *histograma de frecuencias*, se deben seguir los siguientes pasos:

1. Elegir el menú Descripción, en la opción Datos Numéricos – Análisis Unidimensional, seleccionando la variable a representar. Aparecerá una ventana de análisis.
2. Pulsar el icono Opciones Gráficas, en el cuadro de diálogo (figura 2.18), seleccionando la opción Histograma.

Figura 2.18. Opciones gráficas

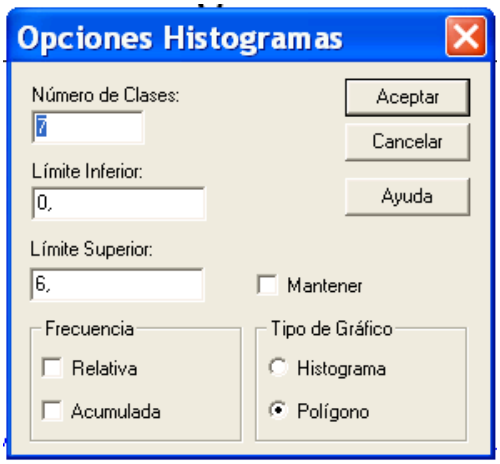


Para modificar la cantidad de intervalos, o los extremos del recorrido, bien en la tabla o en el histograma se debe elegir Opciones de Ventana del menú que aparece cuando se aprieta el botón derecho del ratón, al ingresar en dicha opción aparecerá la ventana de la figura 20. Al modificar la cantidad de intervalos, automáticamente se modifican los extremos de cada intervalo.

Polígono de frecuencias

Sobre la ventana en la que aparece el histograma, hacer clic con el botón derecho del ratón, aparecerá un menú desplegable, allí hacer clic sobre la opción Opciones de Ventana, aparecerá un cuadro de diálogo en el que aparece seleccionado por defecto el histograma, seleccionar el polígono y si lo que se desea es trabajar con las frecuencias relativas, seleccionar la opción Relativa del mismo cuadro (ver figura 2.19). Allí también puede cambiarse la cantidad de intervalos. Haciendo clic en el botón Aceptar, se obtiene el polígono de frecuencias relativas.

Figura 2.19. Cuadro de diálogo para obtener el polígono de frecuencias



- Ir al menú DESCRIPCIÓN. Elegir la opción DATOS NUMÉRICOS y luego ANÁLISIS UNIDIMENSIONAL. Seleccionar la variable con la que se desea trabajar.
- Situados en la ventana de opción de gráficos, se selecciona el histograma de frecuencias (HISTOGRAMA).
- Para realizar el polígono de frecuencias: Sobre la ventana en la que aparece el histograma, hacer clic con el botón derecho del ratón (OPCIONES DE VENTANA), aparecerá un cuadro de diálogo en el que aparece seleccionado por defecto el histograma, seleccionar el polígono (POLÍGONO) y si lo que se desea es trabajar con las frecuencias relativas, seleccionar la opción RELATIVA.

Actividades

2.10. En las figuras 2.20 y 2.21 representamos los datos sobre esperanza de vida en hombres y mujeres tomados del proyecto 1. Escribir un informe de media

página razonando en base a esos gráficos si es verdad que las mujeres tienen una esperanza de vida mayor que los hombres.

Figura 2.20. Distribución de la esperanza de vida en hombres y mujeres

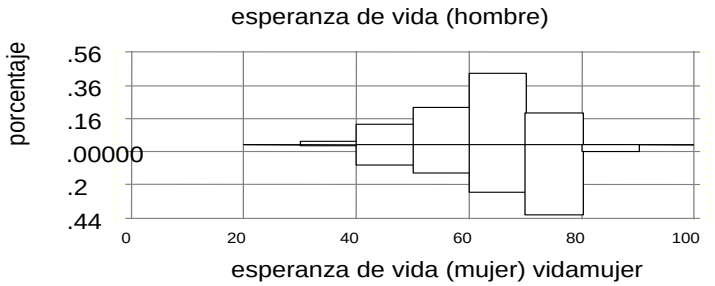
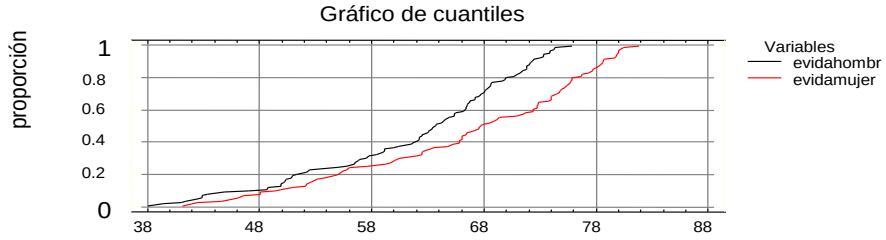


Figura 2.21. Distribución acumulativa de esperanza de vida en hombres y



2.6. DIAGRAMA DE TALLO Y HOJAS

En las representaciones gráficas descritas hasta ahora, rápidamente podemos observar la distribución que sigue el conjunto de datos y, por tanto, en qué valores o intervalos se agrupan con más o menos frecuencia. Pero algunas veces es interesante conocer simultáneamente el valor individual de cada uno de las observaciones. Con las técnicas descritas hasta ahora podemos conseguir lo anterior estudiando simultáneamente la tabla de datos original y el histograma. ¿Hay alguna forma de conseguirlo con una sola técnica? La respuesta la encontramos en el gráfico denominado gráfico del tallo y hojas descrito en Tukey (l977). Para realizar este gráfico, basta seguir los siguientes pasos:

1. Primero se ordenan los datos, por ejemplo de menor a mayor.
2. Se apartan uno o más dígitos de cada dato, según el número de filas que se desea obtener - en general no más de 12 ó 15 - empezando por la izquierda. Cada valor diferente de estos dígitos apartados, se lista uno debajo del otro, trazando a la derecha de los mismos una línea vertical. Este es el tronco.

3. Para cada dato original se busca la línea en la que aparece su "tallo". Los dígitos que nos quedaban los vamos escribiendo en la fila correspondiente de forma ordenada.

Tabla 2.7. Talla de alumnos

149 150 151 153 154 156 156 159 160 160
160 160 160 160 160 160 160 161 161 161
162 162 162 164 164 164 166 166 167 167
168 168 169 169 169 169 170 170 170 170
171 172 172 172 172 173 173 173 174 174
175 175 176 176 178 178 179 179 182 183

Por ejemplo, tomaremos el conjunto de datos que representan la talla de los alumnos de la tabla 2.4. Hacemos los pasos siguientes:

1. Los ordenamos de menor a mayor en la tabla 2.7:
2. Observamos que en todos los datos los dos dígitos de la izquierda son uno de estos números: 14, 15, 16, 17, 18. Listamos estos números de arriba -abajo y dibujamos una línea vertical a la derecha.
3. A continuación, para cada dato original, vamos escribiendo, ordenadamente, el dígito que nos quedaba en su fila correspondiente. Si alguno se repite se escribe tantas veces como lo esté. La representación obtenida se presenta en la figura 2.22

Figura 2.22. Diagrama de tallo y hojas

14 | 9
15 | 0134669
16 | 0000000001112224446677889999
17 | 0000122223334455668899
18 | 23

Como se observa, el resultado es un histograma que conserva ordenados todos los valores observados de los datos; sin embargo, al

mismo tiempo nos proporciona un diagrama que expresa la forma de la distribución.

En algunas tablas de datos, con valores de muchos dígitos, se redondean a dos o tres cifras para construir el tallo y las hojas. Esta representación puede ser ampliada o condensada, aumentando o disminuyendo el número de filas, subdividiendo o fundiendo dos o más filas adyacentes.

Figura 2.23. Diagrama de tallo y hojas extendido

```
14*|
14. | 9
15*| 0134
15. | 669
16*| 000000000111222444
16. | 6677889999
17*| 00001222233344
17. | 55668899
18*| 23
```

El gráfico de la figura 2.22 está bastante condensado, casi la mitad de los datos (28) están en la tercera fila. Para extender el gráfico, cada fila la podemos subdividir en dos de la siguiente forma: marcamos con "*" las filas cuyos dígitos de la derecha van de cero hasta cuatro y, con un "." las filas cuyos dígitos de la derecha van de cinco a nueve el resultado simplificado y extendido se muestra en la figura 2.23. Las ventajas de este gráfico sobre el histograma son las siguientes:

a) Su fácil construcción.
b) Se puede observar con más detalle que el histograma, porque los rectángulos del histograma pueden ocultar distancias entre valores de los datos. Sin embargo, estas lagunas se pueden detectar en la representación del tronco, porque retienen los valores numéricos de los datos.

La desventaja respecto al histograma, es que la escala vertical es una imposición del sistema de numeración, más que una división en intervalos del recorrido de la variable apropiadamente elegida, como se hace en el histograma. El uso del sistema de numeración hace muy fácil su construcción manual, pero puede inducir inadvertidamente comparaciones

inapropiadas. Dos distribuciones con distinta escala vertical son difíciles de comparar.

Para construir el gráfico de tronco y hojas con STATGRAPHICS: ingresar al menú Descripción – Datos Numéricos – Análisis Unidimensional. Una vez obtenida la ventana de análisis, ingresar al botón Opciones Tabulares y allí seleccionar la opción Diagrama de Tallo y Hojas.

Actividades

2.11. En la figura 2.24 representamos las edades de un grupo de personas que se encontraban en un supermercado y en una discoteca. a) Asigna cada diagrama al lugar que le corresponde, razonando la respuesta b) ¿Cuál es en cada caso el promedio (media, mediana o moda) que mejor representa los datos? c) ¿Es en algún caso la edad promedio de los hombres y mujeres diferente?

Figura 2.24 Edades en hombres y mujeres en un supermercado y una discoteca

mujeres		hombres
9998887	1	78888999
9987654200	2	0033468
431	3	23
0	4	5
	5	1

mujeres		hombres
2	0	11
999	1	
998766	2	79
9877662	3	568
86553	4	157
7443	5	2
32	6	5

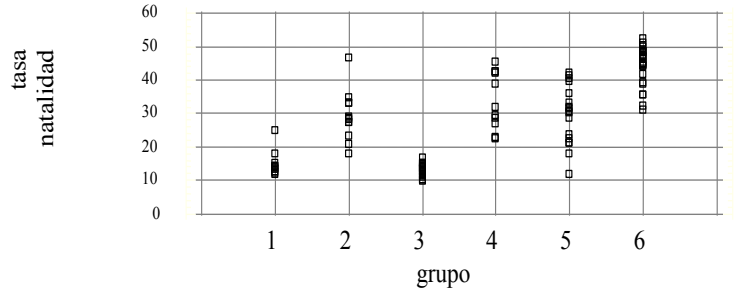
2.12. En la figura 2.25 representamos la distribución de las tasas de natalidad del conjunto de datos relativo al Proyecto 1. ¿Por qué en este caso conviene agrupar en intervalos? ¿Cuántos intervalos conviene usar en la tabla de frecuencias? Construye un histograma de frecuencias. Compara con tus compañeros cómo cambia la forma al variar el número de intervalos.

Figura 2.25. Diagrama de tallo y hojas: Tasa de natalidad

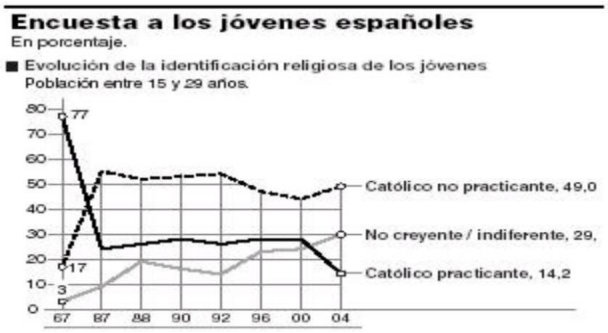
```
0|99
1|001111222223333333444444
1|556778
2|011222334
2|677888899
3|00111122234
3|5568899
4|011122224444
4|55566677788888
5|0012
```

2.13. En la figura 2.26 usamos una nueva representación (diagrama de puntos) de la tasa de natalidad, en este caso diferenciando los grupos de países. Comenta las principales diferencias observadas en los distintos grupos.

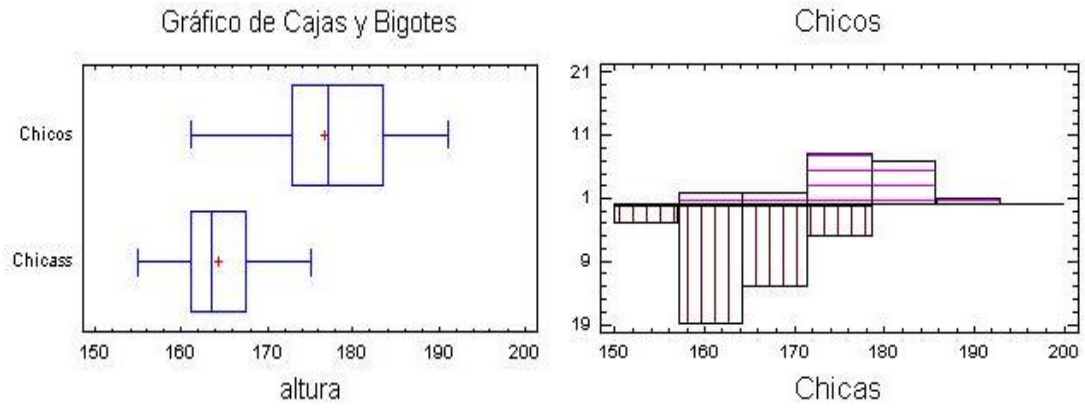
Figura 2.26: Diagrama de puntos



2.14. El gráfico anexo fue publicado hace unos meses en la prensa española. A) ¿Crees que la escala en que se representa los datos es adecuada? ¿Por qué? B) Haz un nuevo gráfico para estos datos con una escala más adecuada.



2.15. Supongamos que queremos estudiar si existe o no diferencia en el tiempo que la altura de chicos y chicas en una clase. Analiza los dos gráficos que reproducimos a continuación. ¿Qué conocimientos estadísticos y sobre el gráfico necesitan los alumnos en cada caso para resolver el problema a partir del gráfico?



TEMA 3

RESÚMENES ESTADÍSTICOS DE UNA DISTRIBUCIÓN DE FRECUENCIAS

3.1. INTRODUCCIÓN

Una vez realizadas algunas representaciones gráficas de las expuestas en el tema anterior, el siguiente paso del análisis de datos es el cálculo de una serie de valores, llamados *estadísticos*, que nos proporcionan un resumen acerca de cómo se distribuyen los datos. Estos estadísticos o características las podemos clasificar de la siguiente forma:

a) *Características de posición o tendencia central:* Son los valores alrededor de los cuales se agrupan los datos. Dentro de esta clase se incluye a la media, mediana y la moda.

b) *Características de dispersión***:** Nos proporcionan una medida de la desviación de los datos con respecto a los valores de tendencia central (recorrido, varianza, ...).

c) *Características de forma:* Nos proporcionan una medida de la forma gráfica de la distribución (simetría, asimetría, etc...).

Estos resúmenes nos serán útiles para resolver problemas como los que te planteamos a continuación.

Actividades

3.1. Como parte de un proyecto los estudiantes de una clase miden cada uno su número de calzado, obteniéndose los siguientes datos:

26 26 26 27 27 27 27 28 28 28 28 28 28 29
29 29 29 29 30 30 30 30 30 30 30 31 32 32 33

Si te preguntan cuál sería el mejor número para representar este conjunto de datos, ¿Qué número o números elegirías? Explícanos por qué has elegido ese (esos) número(s).

3.2. Al medir la altura en cm. que pueden saltar un grupo de escolares, antes y después de haber efectuado un cierto entrenamiento deportivo, se obtuvieron los valores siguientes. ¿Piensas que el entrenamiento es efectivo?

Altura saltada en cm.										
Alumno	Ana	Bea	Carol	Diana	Elena	Fanny	Gilda	Hilda	Inés	Juana
Antes del entrenamiento	115	112	107	119	115	138	126	105	104	115
Después del entrenamiento	128	115	106	128	122	145	132	109	102	117

3.3. Un objeto pequeño se pesa con un mismo instrumento por ocho estudiantes de una clase, obteniéndose los siguientes valores en gramos: 6,2, 6,0, 6,0, 6,3, 6,1, 6,23, 6,15, 6,2 ¿Cuál sería la mejor estimación del peso real del objeto?

3.2. CARACTERISTICAS DE POSICION CENTRAL: LA MEDIA

La principal medida de tendencia central es la *media aritmética*. La media de una muestra se representa por $\bar{x}$ y se calcula mediante la expresión (3.1)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{N} \qquad (3.1)$$

donde N es el número de valores observados, x_i cada uno de los valores observados y f_i la frecuencia con que se presenta el valor x_i

Para calcular la media basta aplicar la definición (3.1). En caso de que los datos se presenten en una tabla de valores agrupados en intervalos, se aplica la misma fórmula, siendo x_i los valores de las marcas de clase. Como se indicó en el tema 2, la agrupación de los valores de la variable implica una pérdida de información sobre dichos valores. Esto se traduce en el hecho de que los estadísticos calculados a partir de valores agrupados están afectados por el error de agrupamiento. Por este motivo, y siempre que sea posible han de calcularse los estadísticos a partir de los datos originales, utilizando las fórmulas para datos no agrupados. Este es el método seguido en los diferentes programas de cálculo utilizados en este curso.

No obstante, puede suceder a veces, que no tengamos los valores individuales de las observaciones sino, por el contrario, dispongamos tan

solo de una tabla de frecuencias. En este caso conviene recordar que los valores obtenidos son sólo aproximados.

Actividades

3.4. Unos niños llevan a clase caramelos. Andrés lleva 5, María 8, José 6, Carmen 1 y Daniel no lleva ninguno. ¿Cómo repartir los caramelos de forma equitativa?

3.5. Un anuncio de cajas de cerillas indica que el número medio de cerillas por caja es 35. Representa una gráfica de una posible distribución del número de cerillas en 100 cajas, de modo que la media sea igual a 35.

3.6. La edad media de un grupo de niños es 5,6 años. ¿Cuál será el tiempo medio si expresamos los datos en meses? ¿Cuál será la edad media de los niños dentro de 3 años?

3.7. La altura media de los alumnos de un colegio es 1,40. Si extraemos una muestra aleatoria de 5 estudiantes y resulta que la altura de los 4 primeros es de 1,38, 1,42, 1,60, 1,40. ¿Cuál sería la altura más probable del quinto estudiante?

Propiedades de la media

Cada una de las actividades 3.4 a 3.7 remite a una propiedad de la media. A continuación describimos estas y otras propiedades, para que identifiques cuál de ellas corresponde a cada actividad.

1. La media aritmética es el centro de gravedad de la distribución de la variable, es decir, la suma de las desviaciones de los valores con respecto a ella es igual a cero, o sea.

$$\Sigma(x_i - x)f_i = 0$$

2. La media aritmética del producto de una constante, a, por una variable, X, es igual al producto de la constante por la media aritmética de la variable dada, o sea:

$$\frac{\sum_{i=1}^{n} a x_i f_i}{N} = \frac{a\sum_{i=1}^{n} x_i f_i}{N} = a\,\bar{x}$$

Esta propiedad implica que, al efectuar un cambio de unidad de medida a los datos (por ejemplo al pasar de metros a centímetros), la media queda afectada por dicho cambio de escala.

3. La media aritmética de la suma de dos variables, *X* e *Y*, es igual a la suma de las medias aritméticas de cada una de las variables:

$$\overline{x+y} = \bar{x} + \bar{y}$$

y también, en general se cumple para cualquier número de variables.

$$\overline{x+y+\ldots z} = \bar{x} + \bar{y} + \ldots \bar{z}$$

4. La media aritmética de la suma de una constante entera, *a*, con una variable, *X*, es igual a la suma de la constante, a, con la media aritmética de la variable dada, es decir:

$$\frac{\sum_{i=1}^{n}(a+x_i)\,f_i}{N} = \frac{na + \sum_{i=1}^{n} x_i\,f_i}{N} = a + \bar{x}$$

Esta propiedad implica que, al efectuar un cambio en el origen desde el que se han medido los datos, la media queda afectada por dicho cambio de origen.

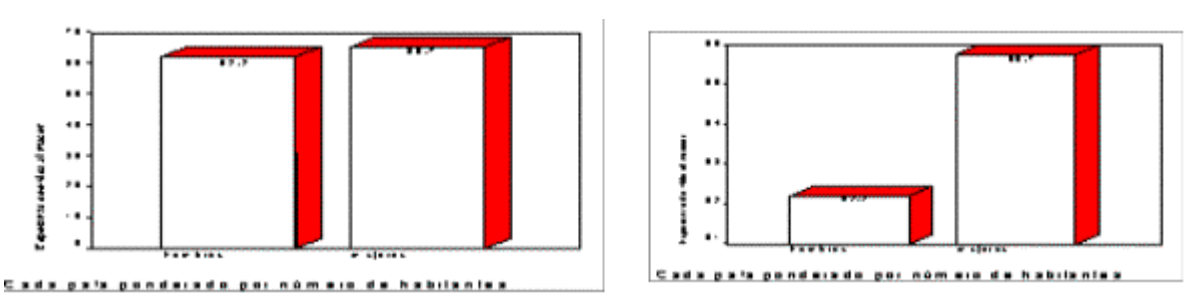
Actividades

3.8. Hay 10 personas en un ascensor, 4 mujeres y 6 hombres. El peso medio de las mujeres es de 60 kilos y el de los hombres de 80. ¿Cuál es el peso medio de las 10 personas del ascensor?

3.9. ¿Qué representa el valor obtenido al calcular la media aritmética simple de la esperanza media de vida al nacer en los 97 países del Proyecto 2? ¿Cómo habría que hacer para calcular la esperanza media de vida al nacer en hombres y mujeres, si no tenemos en cuenta el país de nacimiento?

3.10. En la figura 3.1 hemos representado la esperanza media de vida en hombres y mujeres con dos escalas diferentes. Comparar estos dos gráficos e indicar si te parecen o no adecuados para representar la diferencia entre la esperanza media de vida de mujeres y hombres. Uno de los dos gráficos ha sido obtenido directamente del ordenador, mientras que el otro ha sido manipulado. Averiguar cuál ha sido manipulado.

Figura 3.1. Esperanza de vida media en hombres y mujeres



Media aritmética ponderada

Un error muy frecuente en la actividad 3.8 es contestar que el peso medio es 70 kilos. Tenemos una tendencia a considerar que la media tiene la propiedad asociativa, es decir, que para calcular la media de un grupo de datos se puede calcular las medias parciales y luego promediar todas ellas para obtener el resultado final. Esto no es cierto, como podemos razonar con el siguiente ejemplo:

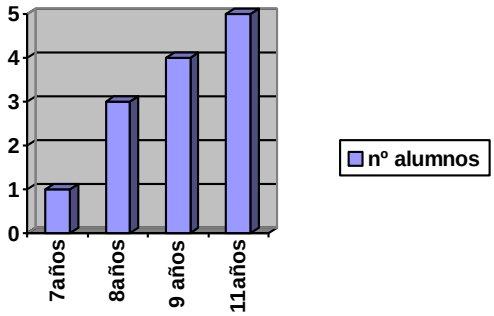
Ejemplo 3.1. Supongamos que una asignatura se divide en tres exámenes parciales, en los que han entrado respectivamente, 2, 3 y 5 temas. Si un alumno ha obtenido 3, 4 y 7 puntos, respectivamente en estos exámenes, ¿Estarías de acuerdo en darle como nota final un 4,3?

3.3. LA MODA

Cuando la variable es cualitativa no podemos calcular la media. Para describir un grupo podemos, entonces usar la moda M_O, que es el valor de la variable que tiene mayor frecuencia. En una distribución puede haber más de una moda. Si existe una sola moda se llama *unimodal,* si existen dos *bimodal,* si hay más de dos se llamará *multimodal.* Podemos también calcular la moda en variables numéricas y distinguiremos para su cálculo dos casos:

1. Variable cualitativa o numérica discreta: Su cálculo es sumamente sencillo, pues basta hallar en la tabla de frecuencias el valor de la variable que presenta frecuencia máxima.

Figura 3.2. Edad de un grupo de alumnos



Ejemplo 3.2. En la figura 3.2 mostramos la distribución de las edades de un grupo de alumnos. La moda es 11 años, pues es la edad más frecuente. Esta distribución es unimodal, pues tiene una sola moda.

2. Cuando la variable está agrupada en intervalos de clases (intervalos), la moda se encontrará en la clase de mayor frecuencia, pudiendo calcular su valor por medio de la expresión (3.2).

(3.2) $$Mo = E_i \frac{d_i}{d_i + d_{i+1}} a_i$$

Donde *Mo* representa la moda; *Ei* es el límite inferior real de la clase modal; d_i representa la diferencia entre la frecuencia absoluta de la clase modal y la clase anterior; $d_{i+1,}$ representa la diferencia entre la frecuencia absoluta de la clase modal y la siguiente $a_{i,}$ representa la amplitud del intervalo de la clase modal. De una forma aproximada podemos tomar como moda el centro del intervalo modal (intervalo de mayor frecuencia).

La moda presenta algunas limitaciones. Si las frecuencias se condensan fuertemente en algunos valores de la variable, la moda no es una medida eficaz de tendencia central.

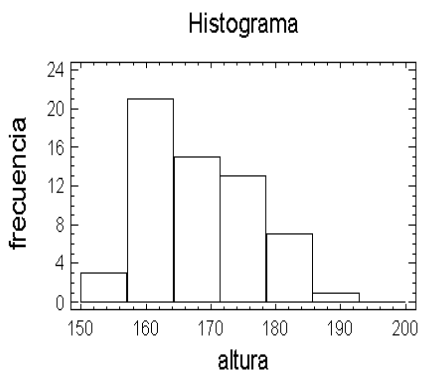
Ejemplo 3.3. Consideremos la siguiente distribución de las puntuaciones obtenidas por 40 alumnos en un examen:

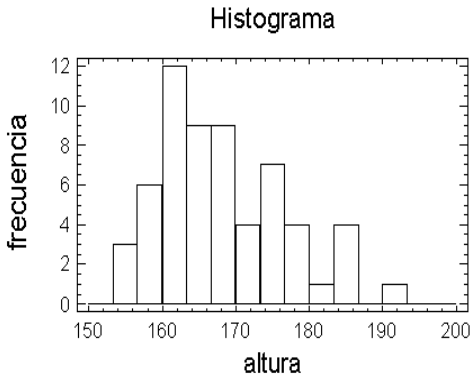
Puntuaciones:	0	1	2	3	4	5	6	7	8	9	10
Número alumnos	8	9	8	4	0	0	0	0	0	0	11

Decir que la moda es 10 (Sobresaliente), cuando el 72,5% no ha superado el examen, nos da idea de la limitación de la moda en este caso. Esto es debido a que en el cálculo de la moda no se tienen en cuenta todos los valores de la variable. Sin embargo, la media es 3,675, y en el cálculo de la media sí se tiene en cuenta todos los valores de la variable.

Una misma distribución con los valores agrupados en clases distintas, puede dar distintas modas, en el cálculo aproximado.

Ejemplo 3.4. Consideremos la altura de un grupo de alumnos cuyas frecuencias vienen representadas en los dos histogramas de frecuencia anteriores. Observamos que en el de la izquierda el intervalo modal es 160,0-165,0 y en el de la derecha160-162,5. Si calculamos la moda en la a) nos resulta 165 aproximadamente y en la b) 161,25.





3.4. MEDIANA Y ESTADISTICOS DE ORDEN

Son aquellos valores numéricos tales que nos indican su posición en el conjunto de datos ordenados, pues una fracción dada de los datos presenta un valor de la variable menor o igual que el estadístico. El más importante es la mediana, que también es una medida de posición central.

La mediana

Si suponemos ordenados de menor a mayor todos los valores de una variable estadística, se llama *mediana* al número tal que existen tantos valores de la variable superiores o iguales como inferiores o iguales a él. La representaremos por M_e. Para el cálculo de la mediana, distinguiremos entre datos no agrupados y agrupados en clases.

Datos presentados en forma de lista

Si el *número de valores es impar* la mediana es el valor del centro de la tabla, cuando los datos están ordenados

Ejemplo 3.5. Si tenemos las siguientes edades de un grupo de alumnos: Andrés: 8 años, María 8 años, Daniel 7 años, Pedro 9 años Luis 11 años. Al ordenar a los alumnos por edad obtenemos:

Daniel 7 años, Andrés: 8 años, María 8 años, Pedro 9 años Luis 11 años

Vemos que la edad del alumno que está en el centro (María) es 8 años. Este es el valor de la mediana.

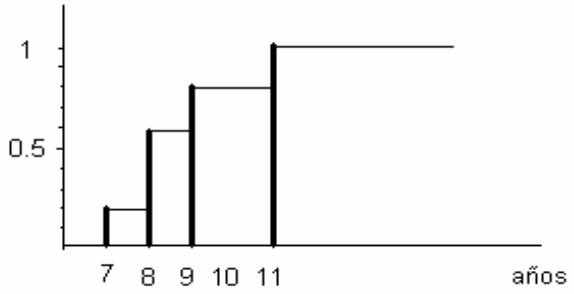
Si el *número de valores es par*, la mediana es la media aritmética de los dos valores que se encuentren en el centro de la tabla.

Ejemplo 3.6. En la actividad 3.11 el número de datos es par (54), Hay dos valores centrales, que corresponden a la oveja (75 pulsaciones) y el ganso (80), Por tanto la mediana es 77, 5 pulsaciones por minuto.

Datos presentados en una tabla de frecuencias

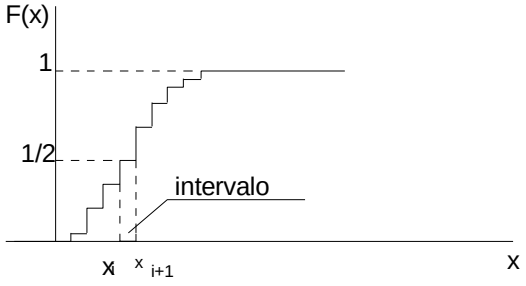
Si *los valores se presentan en una tabla de frecuencias*, es útil calcular las frecuencias acumuladas para hallar la mediana. El cálculo de la mediana se puede hacer en este caso gráficamente a partir del diagrama acumulativo de frecuencias. Una vez realizada la gráfica, procedemos al cálculo de la mediana. Para ello basta tener en cuenta que la frecuencia acumulada que corresponde a la mediana ha de ser igual a $n/2$, o bien, que la frecuencia relativa acumulada es igual a ½. Es posible que nos encontremos en uno de los dos casos siguientes:

Figura 3.3. Cálculo de la mediana con un número impar de datos



1. Si el número de datos es impar, el valor *n/2* corta a la gráfica precisamente en el salto que tiene el diagrama acumulativo para uno de los valores de la variable. Este valor es la mediana, ya que todos los valores de la variable comprendidos entre el lugar n_{i-1} y n_i son iguales a x_i y uno de ellos ocupa exactamente el lugar *n/2* (figura 3.3).
2. Si el número de datos es par, la mediana está indeterminada entre los valores x_i y x_{i+1} , ya que cualquiera de los valores de x incluidos en el intervalo (x_i, x_{i+1}) cumple la definición de mediana. El intervalo (x_i, x_{i+1}) se denomina mediano y suele tomarse como mediana la media aritmética de estos dos valores (figura 3.4).

Figura 3.4. Cálculo de la mediana con número par de datos



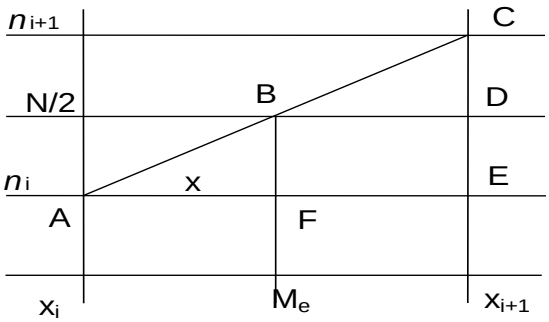
Determinación del intervalo

En la tabla estadística, la mediana se determina a partir de la columna que da las frecuencias (o las frecuencias absolutas) acumuladas, repitiendo el proceso que hemos descrito y finalizando, por tanto, en uno de los casos anteriores.

Datos agrupados en clases

Si los datos están agrupados en clases, se calculan las frecuencias acumuladas de las clases, comenzando el proceso obteniendo la clase mediana. Una vez calculadas estas frecuencias, se representa el polígono acumulativo de frecuencias y, mediante éste, se determina, gráfica o analíticamente, el valor de la variable cuya frecuencia acumulada es *n/2*.

Figura 3.5. Cálculo de la mediana en datos agrupados



Se determina la clase mediana a partir de las frecuencias absolutas acumuladas o de las frecuencias acumuladas y por interpolación lineal se obtiene la mediana. Observemos la figura 3.5 (detalle del diagrama acumulativo de frecuencias) donde:

$$\frac{x}{x_{i+1} - x_{i+1}} = \frac{\frac{N}{2} - f_i}{f_{i+1} - f_i}$$

$$\text{Me}=x_i+x, \quad \frac{x}{AE}=\frac{BF}{CE}$$

Puesto que $x_{i+1} - x_i$ es la amplitud del intervalo, CE la frecuencia en el intervalo mediano y BF la diferencia entre *N/2* y la frecuencia relativa acumulada en el intervalo mediano, obtenemos la cantidad que hay que sumar al extremo inferior del intervalo mediano para calcular la mediana.

Datos presentados en un diagrama de tronco

Si los datos se encuentran representados en un diagrama de tronco, se efectúa un recuento desde el tallo menor (arriba), anotando el número de hojas de cada tallo y acumulándolo a los anteriores, hasta que se supere el valor de *N/2*, siendo *N* el número total de datos; en ese momento se comienza el mismo recuento empezando por el tallo mayor (abajo) hasta llegar al tallo en que nos detuvimos antes.

La mediana se encontrará en el tallo cuyo recuento supera el valor de *N/2*, y sólo habrá que buscar el dato central de los valores de la distribución que se encuentra en este tallo. Si el número de datos el impar, la mediana será el valor que ocupa exactamente la posición central; mientras que si el número de datos es par, la mediana será la media aritmética de los dos valores que se encuentren exactamente en el centro de los datos. Por ejemplo, en el siguiente gráfico del tronco N/*2*=12,5 y la mediana es 27.

		recuento	
1	223455789	9	
2	5667**8**	14	Me =27
3	556779	11	
4	12446	5	

Este método tiene la ventaja de que se visualiza con bastante la claridad el significado de mediana como medida de posición de un conjunto de datos. Está basado en la búsqueda, entre todos los datos ordenados, de aquel que ocupa la posición central.

Actividades

3.11. A continuación reproducimos datos sobre número de pulsaciones por minuto en diversas especies animales[1]

a) ¿Te parece que la media sería un estadístico que representaría bien este conjunto de datos? ¿Y la moda?

[1] Ejemplo tomado de Friel. Mokros y Russell (1992). Statistics: Middles, means and in-betweens. Palo Alto, CA: Dayle Seymour.

b) ¿Encuentras que alguna de las especies es atípica, debido a que su número de pulsaciones está claramente alejada de la mayoría?

1	6	Ballena
2	5 9	Camello, Tiburón
3	0 5 5 7 8 8	Elefante, Caballo, Trucha, Merluza, Salmón, Dorada
4	0 0 2 4 7 8 8 8	Mula, Burro, León, Foca, Caimán, Cocodrilo, Bacalao, Rana
5	5 5 9 9	Vaca, Oso, Carpa, Perca
6	6	Jirafa
7	0 0 0 0 5	Hombre, Ciervo, Avestruz, Cerdo, Oveja
8	0	Ganso
9	0 2 5	Perdiguero, Mastín. Fox Terrier
10	0	Collie
11	0	Delfín
12	0 5	Canguro, Pekinés
13	0	Gato
14		
15	0	Conejo
16		
17	0	Paloma
18		
19		
20		
21	1	Pavo
22		
23		
24	0	Zorro
25		
26	8	Pavo
27		
28		
29		
30	0 1	Puercoespín, Aguila
31	2	Codorniz
32	0	Pollo
33		
34	2 7	Halcón, Buitre
35		
36		
37	8	Cuervo
38	0 8	Grajo Comadreja
39	0	Ardilla
40	1	Gaviota
.		
.		
58	8	Murciélago
59		
60	0	Ratón

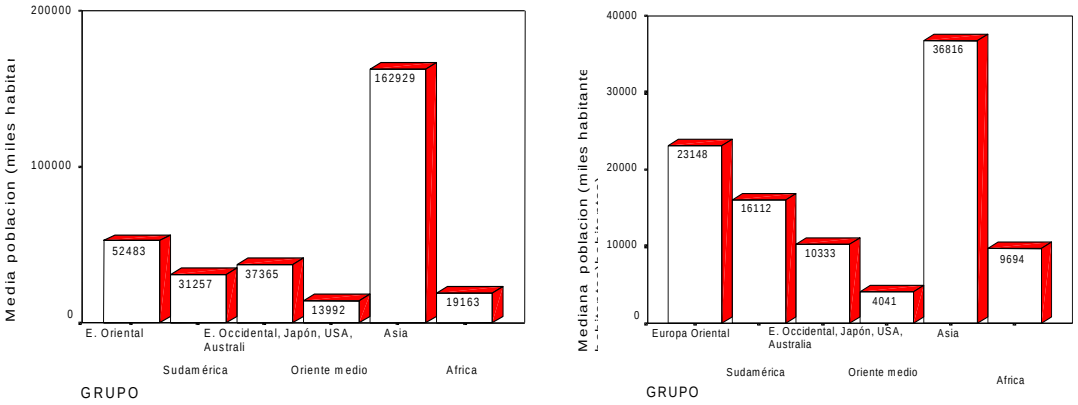
3.12. En las siguientes gráficas hemos representado el número promedio de habitantes en cada país según grupo, usando dos promedios diferentes: media y mediana.

a) Explica lo que representa cada uno de estos promedios

b) Elige el gráfico que mejor representa los datos argumentando la elección.

c) ¿Por qué los gráficos son tan diferentes? ¿Cuál de los dos promedios acentúa más las diferencias entre grupos de países?

Figura 3.6. Mediana y media del número de habitantes en los diferentes grupos de países



3.13. El ayuntamiento de un pueblo quiere estimar el número promedio de niños por familia. Dividen el número total de niños de la ciudad por 50 (que es el número total de familias) y obtienen 2.2. ¿Cuáles de las siguientes afirmaciones son ciertas?

a) La mitad de las familias de la ciudad tienen más de 2 niños

b) En la ciudad hay más familias con 3 niños que familias con 2 niños

c) Hay un total de 110 niños en la ciudad

d) Hay 2,2 niños por adulto en la ciudad

e) El número más común de niños por familia es 2

Propiedades características de la mediana

Al igual que la moda, la mediana también presenta limitaciones. Por un lado, al calcular la mediana no usamos todos los valores observados de la variable, lo que la limita como medida de tendencia central. Además, no puede ser aplicada a distribuciones de variables cualitativas.

Ejemplo 3.7. Supongamos que medimos la estatura de tres personas, de las cuales la primera mide 160 cm y la segunda 165 cm. Si la mediana es 165 cm, ¿Cuánto mide la tercera persona? Nadie podría dar un valor exacto como respuesta. Sin embargo, si la media aritmética es 165 cm, podemos afirmar que la tercera persona mide 170 cm.

Actividades

3.14. La mediana de las puntuaciones de un grupo de 8 alumnos es 6. Pon un ejemplo de posibles puntuaciones que podrían tener estos alumnos de forma que ningún alumno tenga una puntuación igual a 6 (las puntuaciones varían de 0 a 10). ¿Coincide la mediana con el centro del recorrido de los datos?

3.15. En la figura 3.7 presentamos las frecuencias acumuladas de altura de 1000 chicas.

a) Calcula aproximadamente la mediana. máximo y mínimo.

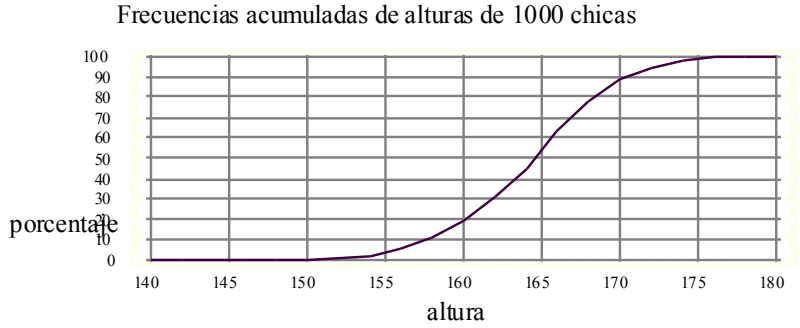
b) ¿Entre qué límites varía el 50 por ciento de los valores centrales?

c) ¿Cuál es el valor de la altura tal que el 70 % de las chicas tiene una altura igual o inferior (percentil del 70%)?

d) Si una chica mide 1.65, ¿En qué percentil está?

e) Compara tu altura con la de estas chicas. ¿Qué porcentaje de chicas son más altas/ bajas que tú?

f) ¿Qué valores de la estatura considerarías atípicos en esta distribución?



Como medida de tendencia central, la moda presenta ciertas ventajas:

a) No se ve afectada por valores extremos de las observaciones. *La mediana es invariante si se disminuye una observación inferior a ella o si se aumenta una superior*, puesto que sólo se tienen en cuenta los valores centrales de la variable. Por ello es adecuada para distribuciones asimétricas o cuando existen valores atípicos.

b) *Conserva los cambios de origen y de escala.* Si sumamos, restamos, multiplicamos o dividimos cada elemento del conjunto de datos por un mismo número esta operación se traslada a la mediana. Ello hace que ésta se exprese en la misma unidad de medida que los datos.

c) La mediana es *un estadístico resistente*: con pequeñas fluctuaciones de la muestra no cambia su valor. Se pueden cambiar uno o varios datos sin que por ello cambie el valor de la mediana, basta con no modificar las dos partes del mismo tamaño en que ésta divide a la distribución.

d) *Si los datos son ordinales la mediana existe*, mientras que la media no tiene sentido, puesto que su cálculo se basa en los valores (numéricos, necesariamente) de los datos.

e) *Para datos agrupados en intervalos con alguno de ellos abierto también es preferible la mediana a la media*. En estos casos, o bien se prescinde del intervalo abierto, o no es posible calcular la media ya que faltaría una de las marcas de clase, la correspondiente a este intervalo.

Actividades

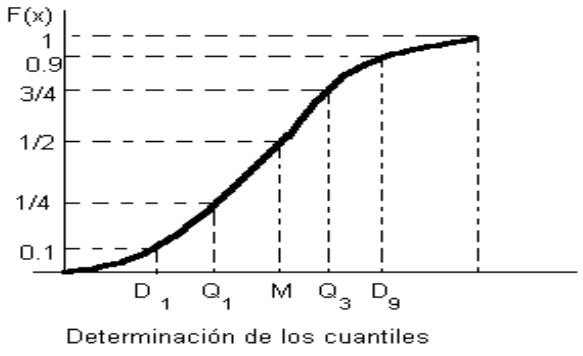
3.16. ¿Cuál de las medidas de posición central permanece constante si cambio un valor extremo de los datos?

3.17. La estatura mediana de un grupo de alumnos es de 156 cm. ¿Cuál será la nueva estatura si expresamos la estatura en metros?

Cuantiles

Además de la mediana, pueden definirse otros estadísticos de orden si, en lugar de considerar la mitad de los datos, tomamos otra fracción cualquiera de los mismos. Una vez ordenado el conjunto de datos, se llama *cuantil* de orden *r (0<r<1)* y se representa por x_r, al valor de la variable que por debajo de él la proporción r de los valores observados. Su cálculo es similar al de la mediana.

Figura 3.8. Cuantiles, cuartiles y mediana



Determinación de los cuantiles

Los cuantiles de uso más frecuente son los *cuartiles* Q_1 y Q_3*:* Q_1 es el cuantil de orden 1/4 y Q_3 el cuantil de orden 3/4. *La mediana es el percentil del 50%, el segundo cuartil y el decil 50.* La mediana y cuartiles dividen a la población en cuatro efectivos iguales. En la figura 3.8 mostramos gráficamente diferentes cuantiles de una distribución.

De la misma manera se definen los *deciles* (D_1 a D_9, cuantiles de orden entre 1/10 y 9/10, respectivamente), y los *percentiles* (cuantiles de orden entre 1/100 y 99/100). Una vez ordenado el conjunto de datos, se llama *percentil del k por ciento (0<k<100),* el valor de la variable que deja inferiores o iguales a él, el k por 100 de los valores observados. Lo representaremos por P_k. Su cálculo es similar al de la mediana.

Si una vez ordenado, el conjunto de datos lo dividimos en 10 partes iguales, se llama *decil k* el valor de la variable que deja inferiores o iguales a él las *k/10* partes del número de observaciones. Su cálculo es similar al de la mediana.

Ejemplo 3.8. Vamos a calcular P_{90} en el ejemplo 3.7. Como 90x5321/100 = 477,80, el percentil pertenece a la clase (110,5-130,5). De ello se deduce:

$$P_{90} = 110,5 + \frac{477,9 - 471}{4} \times 5 = 111,27$$

Como la mediana, los cuantiles se obtienen a partir de las frecuencias absolutas acumuladas o de las frecuencias acumuladas. Las observaciones que se han hecho a propósito de la mediana se pueden aplicar directamente al caso de los cuantiles.

Actividades

3.17. Con una puntuación de 100 María se situó en el percentil del 80 % respecto al total de alumnos de su clase. Supongamos que el profesor decide subir 5 puntos a todos los alumnos. ¿En qué percentil estaría María?

3.18. Supongamos que Pedro se sitúa en el percentil del 40% respecto a su clase y Carmen en el del 80% ¿Podemos decir que la puntuación obtenida por Carmen es doble que la de Juan?

3.5. CARACTERISTICAS DE DISPERSION

Las medidas de tendencia central nos indican los valores alrededor de los cuales se distribuyen los datos. Las características de *dispersión* son estadísticos que nos proporcionan una medida del mayor o menor agrupamiento de los datos respecto a los valores de tendencia central. Todas ellas son valores mayores o iguales a cero, indicando un valor 0 la ausencia de dispersión.

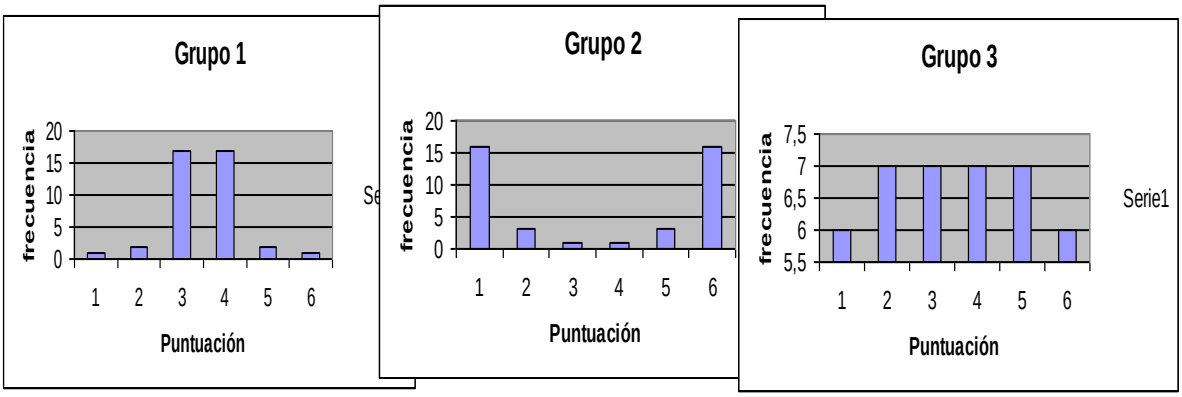
Ejemplo 3.9. Supongamos que hemos realizado una prueba con 5 ítems a 3 grupos de 40 alumnos obteniendo los resultados que se reflejan en la tabla 3.1, donde X_i es el número de ítems que un alumno ha resuelto correctos, f_i la frecuencia correspondiente.

Tabla 3.1

Grupo 1		Grupo 2		Grupo 3	
X_i	F_i	X_i	f_i	X_i	f_i
1	1	1	16	1	6
2	2	2	3	2	7
3	17	3	1	3	7
4	17	4	1	4	7
5	2	5	3	5	7
6	1	6	16	6	6

Las tres distribuciones tienen de media 2,5, ¿pero podemos afirmar que hay homogeneidad entre los tres grupos? Si los representamos gráficamente (figura 3.9) veremos que no. Para precisar mejor lo que denominamos como dispersión podemos calcular unos estadísticos que nos den esta información sin necesidad de representar los datos.

Figura 3.9. Puntuaciones en tres grupos de alumnos



Desviación media

Una primera medida de dispersión es la *desviación media*, que puede calcularse con respecto a cada uno de los valores centrales, media, mediana o moda. Se define como la media de las desviaciones respecto del valor central que se considere, tomadas en valor absoluto. Se calcula con la fórmula (3.4).

(3.4) $$D_c = \frac{\sum_{i=1}^{n} f_i \, x_i - c}{N}$$

donde *c* será, según los casos la media, mediana o moda.

Actividades

3.19. Una alumna tiene unas calificaciones de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Otra alumna tiene unas calificaciones de 1, 1, 1, 1, 1, 10, 10, 10, 10, 10. ¿Cuál de las dos tiene mayor dispersión en sus calificaciones?

Varianza

Es la media aritmética de los cuadrados de las desviaciones respecto a la media. Se representa por S^2 y se calcula mediante la fórmula (3.5).

(3.5) $$S^2 = \frac{\sum_{i=1}^{n} f_i \, (x_i - \bar{x})^2}{N}$$

Esta fórmula se puede simplificar, obteniéndose la (3.6).

(3.6) $$S^2 = \frac{\sum_{i=1}^{n} f_i \, x^2_i}{N} - \bar{x}^2$$

La varianza no varía cuando efectuamos una traslación, es decir, si sumamos o restamos la misma cantidad a todos los datos. En efecto, supongamos que la variable $x_i = z_i + a$, siendo a una constante real. Según vimos en las propiedades de la media $\bar{x} = \bar{z} + a$. Por tanto, sustituyendo en (3.5) los valores de x_i y x, tenemos:

$$S^2 = \frac{\sum_{i=1}^{n} f_i \, (x_i - \bar{x})^2}{N} = \frac{\sum_{i=1}^{n} f_i \, (z_i + a - (\bar{z} + a))^2}{N} = \frac{\sum_{i=1}^{n} f_i \, (z_i - \bar{z})^2}{N}$$

que es la varianza de la variable z_i. Un inconveniente de la varianza al ser utilizada como medida de dispersión respecto a la media, es que no viene expresada en la misma unidad de medida que ésta. Por ello suele utilizarse en su lugar el siguiente estadístico.

Desviación típica

Es la raíz cuadrada de la varianza. Se representa por y se calcula por una de las fórmulas (3.7) o (3.8).

(3.7) $$S = \sqrt{\frac{\sum_{i=1}^{n} f_i \, (x_i - \bar{x})^2}{N}}$$

(3.8) $$S = \sqrt{\frac{\sum_{i=1}^{n} f_i \, x^2_i}{N} - \bar{x}^2}$$

La desviación típica es invariante por traslaciones y viene expresada en la misma unidad de medida que la media y los datos.

Actividades

3.20. Supongamos que la desviación típica de la estatura de un grupo de estudiantes, medida en meros es igual a 2.3. ¿Qué valor tendrá la desviación típica de la estatura de los estudiantes si pasamos los datos a cm?

3.21. ¿Qué ocurre en un conjunto de datos si la varianza toma un valor cero?

3.22. Representa dos diagramas de barras sobre calificaciones de 10 alumnos de modo que la media sea igual en los dos conjuntos de datos pero la varianza sea diferente.

Recorrido, recorrido intercuartílico

A partir de los cuantiles se pueden definir algunos índices de dispersión. El más usado es la diferencia entre el tercer y el primer cuartiles, $Q_3 - Q_1$, llamado *recorrido intercuartílico*. Este contiene el 50%

de la población, dejando a la izquierda el 25% inferior de las observaciones y a la derecha el 25% superior. Otro índice de dispersión muy utilizado es el *recorrido*: es la diferencia entre el mayor y el menor valor posible de la variable. Es el intervalo intercuartílico extremo y es muy sensible a los valores erróneos y a las fluctuaciones del muestreo. Por el contrario su cálculo es extremadamente rápido, no necesitando la clasificación de todas las observaciones.

Coeficiente de variación

Los estadísticos anteriores han medido la dispersión en cifras absolutas. El coeficiente de variación CV es una medida de dispersión relativa y viene dado por (3.9).

(3.9)

$$CV = \frac{S}{\bar{X}}$$

Su utilidad radica en que es independiente de la unidad utilizada en los valores de la variable, por lo que se pueden comparar distribuciones cuyos datos estén medidos en distintas unidades, por ejemplo pesetas y dólares. Sin embargo es poco práctico cuando la media es próxima a cero, por el valor tan desmesurado que toma.

Actividades

3.23. ¿Cuál es la diferencia entre dispersión absoluta y dispersión relativa? Pon un ejemplo donde, en dos distribuciones una tenga mayor dispersión absoluta y otra tenga mayor dispersión relativa.

3.24. ¿Cuál de las medidas de posición central permanece constante si cambio un valor extremo de los datos? ¿Cuál de las medidas de dispersión permanece constante si cambio un valor central de los datos?

3.6. CARACTERISTICAS DE FORMA

Cuando conocemos las características de posición y las de dispersión es conveniente conocer la forma de la distribución, para ello estudiaremos la simetría, asimetría y curtosis.

Simetría y asimetría

Decimos que una distribución es *simétrica* cuando lo es su representación gráfica, es decir, los valores de la variable equidistantes a un valor central de la misma tienen frecuencias iguales. Este valor central coincide con la media y mediana. Si la distribución tiene una sola moda, ésta coincide también con las anteriores.

$$\bar{x} = M_e = M_o$$

Una distribución que no es simétrica se llama *asimétrica*. La asimetría se puede presentar a la derecha (positiva) o a la izquierda (negativa), según el lado a que se presente el descenso en la representación gráfica.

- En las distribuciones *asimétricas a la derecha* con una sola moda se cumple la relación (3.10).

(3.10) $$\bar{x} > M_e > M_o$$

- En las distribuciones *asimétricas a la izquierda* con una sola moda se cumple (3.11).

(3.11) $$\bar{x} < M_e < M_o$$

Coeficientes de asimetría

Para saber si una distribución con una sola moda es simétrica a la derecha o a la izquierda sin necesidad de representarla gráficamente, podemos utilizar el coeficiente de asimetría de Pearson, que se representa por A_p y se calcula por la fórmula (3.12).

(3.12) $$A_p = \frac{\bar{x} - M_o}{S}$$

- *En una distribución simétrica la mediana coincide con la media y la moda* (en distribuciones unimodales). En este tipo de distribuciones los datos se encuentran repartidos a lo largo del recorrido de forma que todas las medidas de tendencia central están justo en el centro del conjunto de datos. Si la distribución es simétrica $A_p = 0$, ya que $\bar{x} = M_o$
- *Si la distribución es asimétrica a la derecha el orden en que aparecen es moda-mediana-media*, puesto que es en el lado derecho dónde se concentran la mayor frecuencia de los datos y, por tanto la moda; *y si es asimétrica a la izquierda el orden es media-mediana-moda* (para

distribuciones unimodales). Si hay asimetría a la derecha $A_p > 0$, ya que $\bar{x} > M_O$.

- *Si la distribución es asimétrica es preferible la mediana a la media como medida de tendencia central*. En estos casos, tanto la media como la moda están desplazadas hacia uno de los extremos del conjunto de datos y son demasiado representativas de la distribución, a menos que se disponga de la información adicional aportada por las medidas de dispersión. Si hay asimetría a la izquierda $A_p < 0$, ya que $\bar{x} < M_O$.

Este coeficiente es, además, invariantes por traslaciones y cambios de escalas, debido a las propiedades de la media, moda y desviación típica.

Actividades

3.25. Buscar ejemplos de variables estadísticas en la vida real que tengan distribuciones asimétricas. ¿Qué signo tomaría el coeficiente de asimetría en cada caso?

3.26. El coeficiente de asimetría de la estatura de un grupo de alumnos medida en metros es 0.4. ¿Cuánto vale el coeficiente di pasamos la estatura a cm?

3.27. Dibujar el gráfico de la caja de una distribución que sea asimétrica a la derecha y el de otra distribución que sea asimétrica a la izquierda.

3.28. ¿Qué tipo de forma piensas tiene n las distribuciones de las siguientes variables?:

- Renta per cápita de las familias españolas
- Edad de los españoles
- Horas de duración de una bombilla que se funde
- Mes de nacimiento de un grupo de 100.000 personas
- Número de accidentes de tráfico diarios en una ciudad
- Peso en kg. de un recién nacido
- Calificaciones en las pruebas de selectividad
- Calificaciones de acceso a la Facultad de Psicología

Coeficiente de curtosis

Cuando una distribución es simétrica, a veces, es interesante saber si es más o menos apuntada que la curva normal. Esta es una distribución teórica que estudiaremos más adelante y tiene una forma característica, similar a una campana invertida. Si una distribución es más apuntada que la normal se llama *leptocúrtica.* Si es aproximadamente igual de apuntada que

la normal se llama *mesocúrtica.* Si es menos apuntada o más aplastada que la distribución normal se llama *platicúrtica.* Existe un coeficiente, ideado por Fisher, que mide el apuntamiento de una distribución y se llama coeficiente de curtosis. Se suele representar por *K* y se verifica:

- Si $K < 0$ la distribución es platicúrtica
- Si $K = 0$ es mesocúrtica
- Si $K > 0$ es leptocúrtica

3.7. GRÁFICO DE LA CAJA

El gráfico de la caja fue descrito por Tukey denominándolo *"box and whiskers".* Para su construcción se utilizan 5 estadísticos de la distribución de frecuencias: el mínimo, el primer cuartil Q_1, la mediana, el tercer cuartil Q_3, y el máximo. Explicaremos su construcción a partir del siguiente conjunto de datos (peso en kg. de un grupo de alumnos de bachillerato).

Peso en Kg.

Varones	Mujeres
55 64 70 74 75 70	60 45 46 50 47 55
64 93 60 62 70 80	49 52 50 46 50 52
61 60 62 68 65 65	52 48 52 63 53 54
66 68 70 72 72 71	54 54 53 55 57 44
	56 56 56 53 60 65
	67 61 68 55 64 60

1. Se traza una línea vertical u horizontal de longitud proporcional al recorrido de la variable, que llamaremos eje (véase la figura 3.10). Los extremos del eje serán el mínimo y el máximo de la distribución, que en nuestro caso son 44 y 93 kilos. En el interior del eje se señalarán las subdivisiones que creamos necesarias, para formar una escala.
2. Paralelamente al eje se construye una caja rectangular con altura arbitraria y cuya base abarca desde el primer cuartil al tercero. Como vemos esta "caja" indica gráficamente el intervalo de variación del cincuenta por ciento de valores centrales en una distribución que, para el peso de los estudiantes, abarca desde 53 a 66,5.
3. La caja se divide en dos partes, trazando una línea a la altura de la mediana (60 kg. en nuestro caso). Cada una de estas partes indica pues

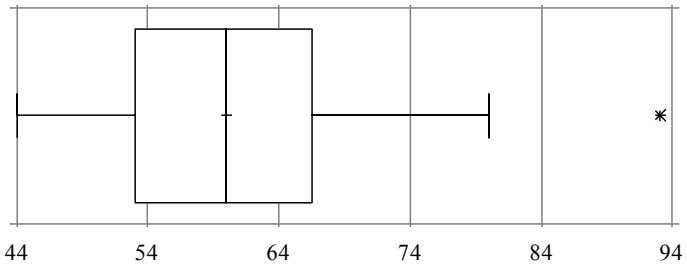
el intervalo de variabilidad de una cuarta parte de los datos. De este modo, en el ejemplo dado, una cuarta parte de los alumnos tiene un peso comprendido entre 44 y 53, estando incluidas las otras cuartas partes en los siguientes intervalos de peso: 53 a 60. 60 a 66,5 y 66,5 a 93.

4. A la caja así dibujada se añaden dos guías paralelas al eje, una a cada lado, de la forma siguiente: el primero de estos segmentos se prolonga desde el primer cuartil hasta el valor máximo entre el mínimo de la distribución y la diferencia entre el primer cuartil y una vez y media el recorrido intercuartílico. Como en nuestro caso el peso mínimo es 44 kilos, y el recorrido intercuartílico es 66,5 - 53 = 13,5, al restar al primer cuartil, Q_1= 53 una vez y media el recorrido intercuartílico obtenemos:

$$Q_1 - 1.5\ RI = 53 - 20{,}25 = 32{,}75$$

El máximo entre 44 y 32,75 es 44, por lo que el segmento inferior que debe dibujarse en el gráfico de la caja debe llegar hasta 44, como se muestra en la figura 3.10.

Figura 3.10. Gráfico de la caja para el peso de los alumnos



5. El segmento dibujado al otro lado de la caja abarca desde el tercer cuartil hasta el mínimo entre el mayor de los datos y la suma del tercer cuartil con una vez y media el recorrido intercuartílico. En el peso de los alumnos el máximo es 93 kilos y, al sumar una vez y media el recorrido intercuartílico al cuartil superior 66,5 obtenemos:

$$Q_3 + 1{,}5\ RI = 66{,}5 + 20{,}25 = 86{,}75$$

De este modo, el extremo superior del segmento debe prolongarse ahora sólo hasta 86.75

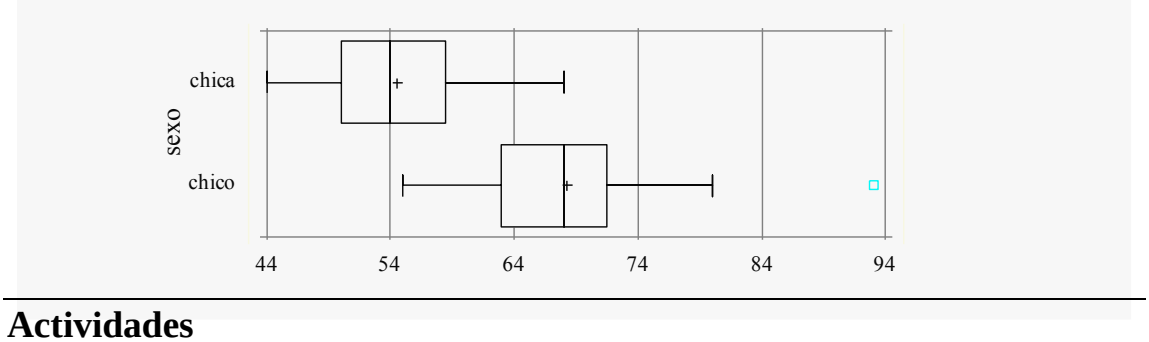
6. Si alguno de los datos queda fuera del intervalo cubierto por la caja y estos segmentos, como ocurre en el ejemplo con el alumno que pesa 93 Kg., se señala en el gráfico mediante un asterisco o cualquier otro símbolo, como puede verse en la figura 3.10.

Estos datos son los llamados *valores atípicos* (“outliers”), que son valores muy alejados de los valores centrales de la distribución. En la distribución normal, fuera del intervalo que resulta de extender los cuartiles en una vez y media el recorrido intercuartílico, sólo aparece un uno por ciento de los casos, por lo que estos valores, si no son debidos a errores, suelen ser casos excepcionales.

Utilidad del gráfico de la caja

Como vemos en el ejemplo dado, este gráfico nos proporciona, en primer lugar, la posición relativa de la mediana, cuartiles y extremos de la distribución. En segundo lugar, nos proporciona información sobre los valores atípicos, sugiriendo la necesidad o no de utilizar estadísticos robustos. En tercer lugar, nos informa de la simetría o asimetría de la distribución, y posible normalidad o no de la misma. El gráfico de la caja también se puede utilizar para comparar la misma variable en dos muestras distintas, como se muestra en la figura 3.11 al comparar los pesos de chicos y chicas

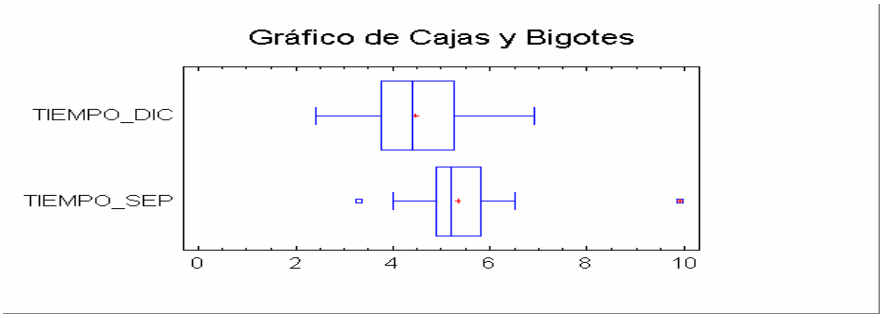
Figura 3.11. Gráfico de la caja para los pesos de chicos y chicas



Actividades

3.29. La figura adjunta representa los tiempos en segundos que tardan en recorrer 30 metros un grupo de deportistas en Septiembre y Diciembre. ¿Piensas que el entrenamiento durante los tres meses ha sido efectivo? ¿Qué puedes decir de la simetría de la distribución? ¿Hay valores atípicos?

Figura 3.12. Tiempos en recorrer 30 metros



3.8. CURVA EMPÍRICA DE DISTRIBUCIÓN

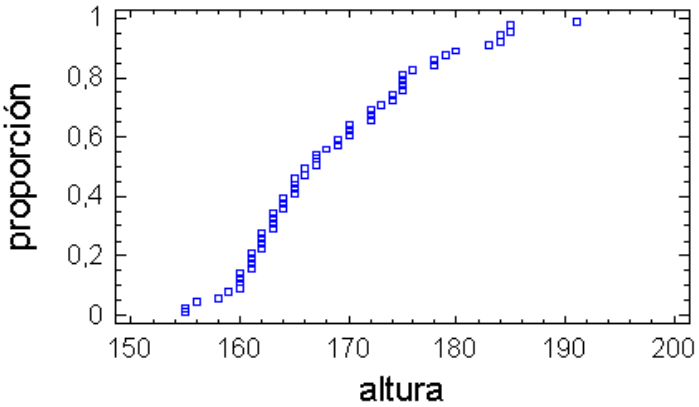
Al representar el polígono acumulativo de frecuencias, obtenemos un gráfico para las frecuencias acumuladas, que nos ha servido también para el cálculo de la mediana y otros estadísticos de orden. Sin embargo, hemos distorsionado el conjunto de datos original al agrupar las observaciones en intervalos, por lo que los valores obtenidos para dichos estadísticos tienen el carácter de aproximados.

Una representación gráfica para las frecuencias acumuladas, que conserva los valores originales de las observaciones, es la *curva empírica de distribución*. Para realizar dicha gráfica se ordenan los datos de menor a mayor. En un eje de coordenadas se dibuja un cuadrado cuya altura, colocada sobre el eje *Y* representa la frecuencia acumulada relativa, y la base los valores de la variable.

Para cada uno de los datos se representa un punto sobre dicho sistema de ejes. La coordenada *X* es el valor de la variable para esta observación y la coordenada *Y* la frecuencia relativa acumulada hasta este valor de la variable. En la figura 3.12 se muestra la curva empírica de distribución para la variable "altura" del conjunto de datos ALUMNOS. Puede observarse que el polígono acumulativo de frecuencias para esta misma variable es básicamente una versión simplificada de la curva empírica de distribución.

Mediante la curva empírica de distribución podemos obtener gráficamente la mediana y demás estadísticos de orden en una distribución de frecuencias.

Figura 3.12. Función de distribución empírica



Comparación de los estadísticos de orden en dos grupos

A veces, es necesario comparar una misma variable en dos grupos diferentes. Esta comparación puede hacerse de distintas formas. Podemos comprobar si la media de un grupo es superior a la del otro, o si una de ellos presenta más dispersión. Puede ser interesante, no obstante, comparar gráficamente las dos distribuciones para detectar cualquier tipo de diferencia. Esto puede hacerse de varios modos:

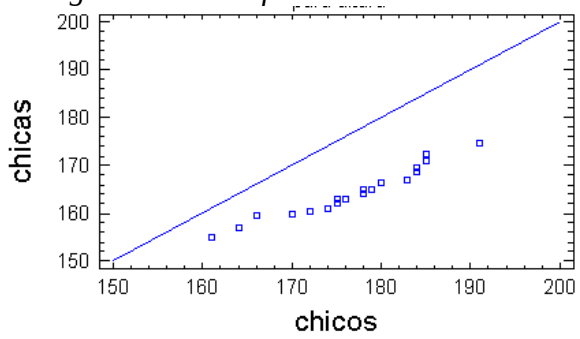
a) realizando separadamente el histograma o el gráfico de la caja en cada grupo, utilizando una misma escala para la variable.
b) realizando una representación superpuesta de las curvas empíricas de distribución
c) utilizando una representación *cuantil-cuantil*.

Para construir este nuevo gráfico se halla el mínimo y máximo del conjunto total de datos (reuniendo los dos grupos). Se construye un cuadrado sobre los ejes de coordenadas cuyo lado tiene por longitud la diferencia entre estos dos valores extremos. En la base del cuadrado, situada sobre el eje *X*, se representa el grupo con menor número de datos (Grupo A) el otro (Grupo B) se representa sobre el eje *Y*.

Por cada individuo *i* del grupo A se representa un punto, cuya coordenada *Xi* es el valor de la variable para este individuo. Sea *Pi* el percentil que en el grupo A corresponde al individuo *i*. Buscaremos en el grupo B el valor de la variable *X* que corresponde, en dicho grupo, al percentil *Pi;* este valor será utilizado como coordenada *Yi* del punto a representar.

Como ejemplo, en la figura 3.13 se muestra un gráfico cuantil-cuantil obtenido mediante Statgracphics para las alturas de chicos y chicas en una clase. Podemos ver como, para cualquier percentil, la altura de los chinos es mayor que la de las chicas, ya que la gráfica siempre se sitúa por debajo de la diagonal.

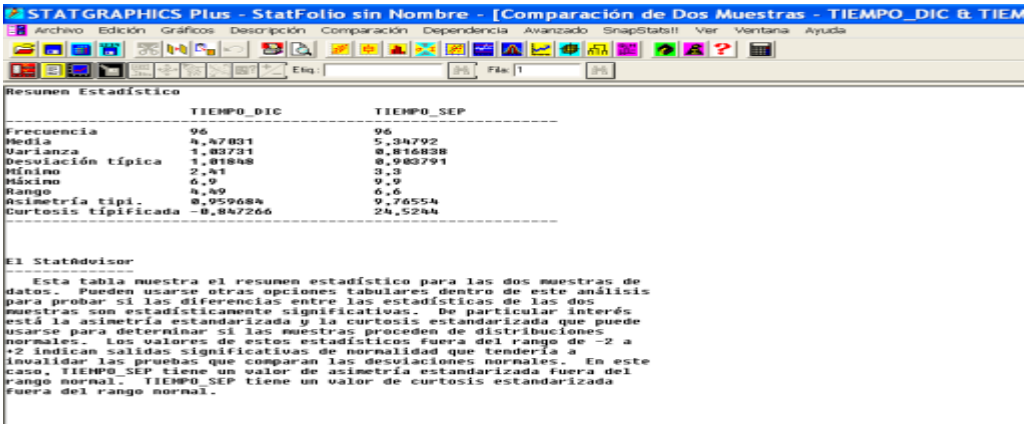
Figura 3.13. Gráfico Cuantil- Cuantil



3.9. CÁLCULO DE ESTADÍSTICOS CON STATGRAPHICS:

Los estadísticos se pueden obtener de la opción de Resumen numérico – Resumen estadístico, dentro de Descripción – Numeric Data – One Variable Analysis. En la figura 3.14 se muestran las medidas o parámetros que aparecen por defecto.

Figura 3.14. Estadísticos que proporciona RESUMEN ESTADÍSTICO por defecto



Pulsando en esta ventana con el botón derecho del ratón, se selecciona Opciones de Ventana donde se puede elegir los resúmenes estadísticos, que se quieren calcular y que se muestran en la figura 3.15. Por último la opción percentiles se muestra en la figura 3.16 y permite calcular percentiles.

Figura 3.15. Cálculo de percentiles

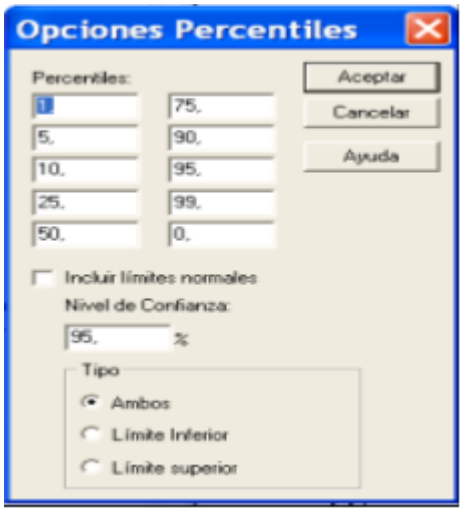
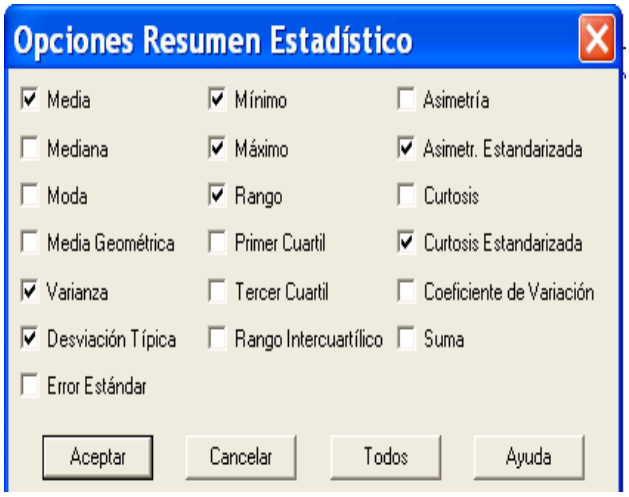


Figura 3.16. Resúmenes estadísticos



Análisis simultáneo de varias variables

Para realizar el cálculo de estadísticos de varias variables simultáneamente, ingresar al menú Descripción – Datos Numéricos – Análisis Multidimensional, se deberán seleccionar las variables que se desean usar como se muestra en la figura 3.17. En la ventana de análisis seleccionar Opciones Tabulares, se selecciona Resumen estadístico y luego se pueden seleccionar los estadísticos que se desean calcular de la misma forma que en el párrafo anterior.

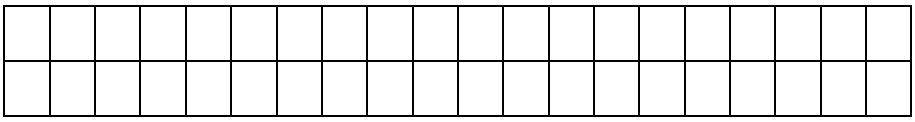
Figura 3.17. Selección de las variables

Cálculo de percentiles

Cuando se desea realizar el cálculo de percentiles se deben realizar los siguientes pasos:

1. Entrar al menú Descripción – Datos Numéricos – Análisis Unidimensional, allí aparecerá una ventana de análisis.
2. En la ventana, seleccionar el botón Opciones Tabulares, y allí seleccionar la opción Percentiles.
3. Una vez seleccionada la opción anterior se verá una ventana en la que aparecen algunos percentiles predefinidos. Si se desea modificar el valor de tales percentiles, seleccionar Opciones de Ventana del menú que aparece cuando se aprieta el botón derecho del ratón. En este caso aparecerá una ventana introduciendo en cada cuadro los valores que se necesiten pulsando Aceptar se obtienen los valores requeridos

Actividad 3.30. Supón que estás jugando a lanzar una moneda 40 veces y escribe los resultados que esperas obtener. Pon una C para indicar cara y + para indicar cruz:



- Calcula los siguientes valores de la sucesión producida: Número de caras, número de rachas (se produce una racha cuando aparecen varios resultados iguales; tomando la longitud igual a 1 si hay un cambio de resultado), longitud de la racha más larga, número de caras en la primera y segunda mitad.
- Repite la actividad pero lanzando realmente una moneda.
- Pide a 9 amigos más que realicen esta actividad. Puedes usar una hoja como la siguiente para recoger los datos

	Secuencia simulada			Secuencia real		
Amigo	N. Caras	N. Rachas	Longitud	N. Caras	N. Rachas	Longitud
1						
2						
3						
...						

Compara los resultados obtenidos al lanzar realmente la moneda y los simulados (cuando la moneda no se lanza realmente). ¿Piensas que tenemos, en general una buena intuición de las características de las secuencias aleatorias?

TEMA 4

INTRODUCCIÓN A LA PROBABILIDAD

4.1. EXPERIMENTO Y SUCESO ALEATORIO

Iniciaremos el estudio de las nociones probabilísticas analizando un ejemplo cotidiano -el pronóstico del tiempo-, en el que tenemos necesidad de realizar predicciones o tomar decisiones en situaciones de incertidumbre. Este ejemplo u otros sobre resultados de elecciones, esperanza de vida, accidentes, etc pueden servir de contextos sobre los cuales apreciar las características de los fenómenos para los que son pertinentes los modelos y nociones probabilísticas, es decir, los fenómenos aleatorios.

Actividades

4.1. Daniel y Ana son estudiantes cordobeses. Acuden a la misma escuela y su profesor les ha pedido que preparen una previsión del tiempo para el día 24 de Junio, fecha en que comenzarán sus vacaciones. Puesto que están aún en el mes de Mayo, Daniel y Ana no pueden predecir exactamente lo que ocurrirá. Por ello, han buscado una lista de expresiones para utilizar en la descripción del pronóstico. He aquí algunas de ellas:

cierto; posible; bastante probable; hay alguna posibilidad; seguro; es imposible; casi imposible; se espera que; incierto; hay igual probabilidad; puede ser; sin duda.

¿Podrías acabar de clasificar estas palabras según la mayor o menor confianza que expresan en que ocurra un suceso? Busca en el diccionario nuevas palabras o frases para referirte a hechos que pueden ocurrir y compáralas con las dadas anteriormente.

Busca en la prensa frases o previsiones sobre hechos futuros en que se usen las palabras anteriores. Clasifícalas según la confianza que tienes en que ocurran. Compara tu clasificación con la de otros compañeros.

El objetivo de la actividad 4.1 es reflexionar sobre el uso de palabras y expresiones del lenguaje ordinario en circunstancias en que se tienen distintos grados de confianza en que ocurrirá un suceso. Comparamos diferentes sucesos en función de la .confianza que se tenga en su ocurrencia. Se ordenarán los sucesos en base a las preferencias individuales; posteriormente se pueden emplear diversas expresiones lingüísticas para referirse a estas comparaciones: "más probable", "muy probable", etc.

La situación se refiere a fenómenos del mundo físico (previsión del tiempo) para los que habitualmente se aplican las técnicas de recogida de datos estadísticos y la modelización aleatoria. Utilizamos la expresión *"experimento aleatorio*" para describir este tipo de situaciones.

Llamaremos "*experimento*" tanto a los verdaderos experimentos que podamos provocar como a fenómenos observables en el mundo real; en éste último caso, la propia acción de observar el fenómeno se considera como un experimento. Por ejemplo, la comprobación del sexo de un recién nacido se puede considerar como la realización de un experimento. Diferenciamos entre *experimentos deterministas* y *aleatorios*. Los primeros son aquellos que, realizados en las mismas circunstancias sólo tienen un resultado posible. Por el contrario, un experimento aleatorio se caracteriza por la posibilidad de dar lugar, en idénticas condiciones, a diferentes efectos.

Suceso es cada uno de los posibles resultados de un experimento aleatorio. Distinguimos entre *sucesos elementales,* cuando no pueden descomponerse en otros más simples y *suceso compuestos* cuando se componen de dos o más sucesos elementales por medio de operaciones lógicas como la conjunción, disyunción o negación.

Actividades

4.2. Poner tres ejemplos de experimentos aleatorios y deterministas. Para cada uno de ellos describir un suceso simple y otro compuesto.

4.2. ESPACIO MUESTRAL Y OPERACIONES CON SUCESO

El conjunto de todos los resultados posibles de un experimento aleatorio se denomina *espacio muestral* o *suceso seguro*. Suele representarse mediante la letra *E*. Por ejemplo, el espacio muestral

obtenido al lanzar un dado sería *E={1,2,3,4,5,6}.* Este espacio muestral es finito, pero podemos considerar un espacio muestral con infinitos resultados posibles. Por ejemplo, la duración de una lámpara podría variar en un intervalo continuo [0, 1000], donde hay infinitos puntos. Otros casos serían el peso o la talla de una persona tomada al azar de una población.

Puesto que el suceso seguro consta de todos los resultados posibles, siempre se verifica. Teóricamente podríamos también pensar en un suceso que nunca pueda ocurrir, como obtener un 7 al lanzar un dado ordinario. Lo llamaremos *suceso imposible* y lo representamos por Ø.

Dentro de los posibles sucesos aleatorios asociados a este experimento podemos distinguir dos tipos: *sucesos elementales*, si no pueden ser descompuestos en otros más simples, y *sucesos compuestos*, si se componen de dos o más sucesos elementales.

Así, cuando realizamos el experimento consistente en lanzar un dado, un suceso simple sería: "obtener el número dos", y un suceso compuesto "obtener un número par".

Álgebra de Boole de sucesos

Un suceso compuesto está formado por varios sucesos elementales, por tanto, será cualquier subconjunto del espacio muestral. Así, al lanzar un dado, "obtener un 2" es un suceso simple, mientras que "obtener impar" es un suceso compuesto.

Inclusión de sucesos

Diremos que un suceso *A* está incluido en otro *B, A* $\subset$ *B,* si siempre que ocurre *A* ocurre *B*. Por ejemplo, obtener figura doble al lanzar dos dados está incluido en obtener suma par.

Sean *A* y *B* dos sucesos asociados a un mismo experimento. A partir de ellos podemos formar nuevos sucesos mediante las operaciones de unión e intersección.

Unión de sucesos

Llamaremos "unión" de los sucesos *A* y *B*, y representaremos por *A*$\cup$*B*, al suceso que se verifica cuando se produce al menos uno de los dos sucesos *A* ó *B*.

Intersección de sucesos

Llamaremos “intersección” de los sucesos A y B y representaremos por $A \cap B$ al suceso que ocurre cuando se verifican simultáneamente A y B.

Consideremos, por ejemplo, el experimento consistente en preguntar a un matrimonio de dos hijos el sexo de los mismos. Sea A el suceso "el mayor es hembra" y B el suceso "el menor es hembra". En dicho caso $A \cup B$ es el suceso "el matrimonio tiene al menos una hija", y $A \cap B$ "los dos son chicas".

Puede darse el caso de que al expresar la intersección de dos sucesos lleguemos a otro que no pueda realizarse, como es el caso de expresar $A \cap C$ en el ejemplo anterior, donde C fuese el suceso "el mayor de los hijos es varón". En este caso representaremos $A \cap C = \emptyset$ y llamaremos a dicho suceso “suceso imposible”. Los sucesos A y C se dicen que son incompatibles.

Puede observarse en las definiciones anteriores el paralelismo con las operaciones entre subconjuntos de un conjunto dado. Así, si adoptamos la convención de representar el espacio muestral asociado al experimento del ejemplo anterior como:

$E = \{vv,vh,hv,hh\}$, obtenemos

$A = \{hv,hh\}$, $B=\{vh,hh\}$, $C=\{vv,vh\}$

$A \cup B = \{hv,hh,vh\}$ $A \cap B = \{hh\}$ $A \cap C = \emptyset$

Suceso contrario

Por último, a cada suceso A posible en un experimento asociaremos otro suceso $\bar{A}$ que llamaremos “contrario” del dado, tal que $\bar{A}$ se verifica cuando no se verifica A. Así, en el ejemplo considerado,

$\bar{A}$ ="el primer hijo es varón" = C

$B = \{hv,vv\}$

Entre las propiedades del suceso contrario se hallan las siguientes:

$\bar{E} = \emptyset$ $\emptyset = E$ $A \cup \bar{A} = E$ $A \cap \bar{A} = \emptyset$

Vemos que el suceso contrario a uno dado representa el mismo papel que el conjunto complementario en Teoría de Conjuntos. Es fácil

demostrar que las operaciones definidas anteriormente cumplen las propiedades habituales del Álgebra de Conjuntos.

Actividades

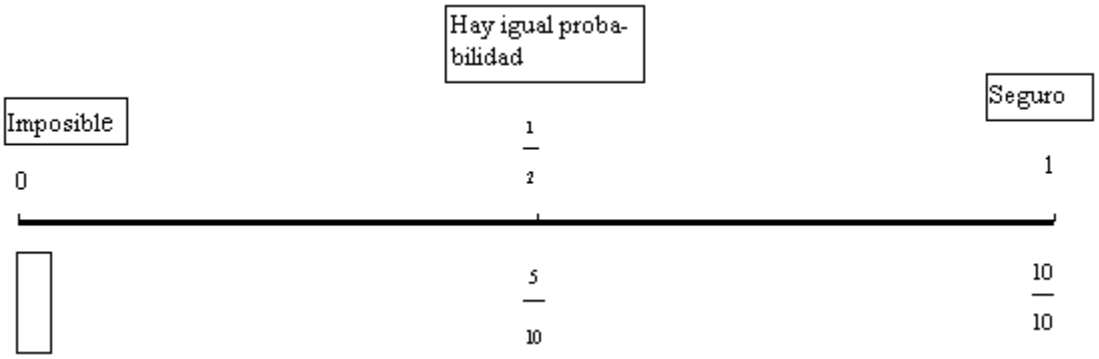
4.3. Describir el espacio muestral asociado a cada uno de los siguientes experimentos: a) lanzamiento simultáneo de tres monedas; b) suma de los puntos obtenidos al lanzar simultáneamente dos dados.

4.4. Describir un suceso imposible asociado a cada uno de los experimentos anteriores.

4.5. En una caja hay 4 bolas rojas, 3 verdes y 2 blancas. ¿Cuántas bolas se deben sacar sucesivamente para estar seguro de obtener una bola de cada color?

4.6. *La escala de la probabilidad.* Ana y Daniel han terminado su trabajo, pero no están satisfechos. Para completarlo van a asignar un número a cada una de las palabras utilizadas en la actividad 1. Esta es la escala que utilizan:

Figura 4.1



¿Podrías asignar un valor en la escala de probabilidad a las expresiones .de la actividad 4.1 a)?

4.3. ASIGNACIÓN DE PROBABILIDADES SUBJETIVAS

Para asignar probabilidades a sucesos, se hace corresponder un valor numérico entre 0 y 1 a los sucesos comparados anteriormente, o bien se situarán sobre un gráfico, mostrando la escala de la probabilidad. Una vez realizada la actividad individualmente o por parejas, pueden compararse los resultados de los diversos grupos. Se reflexionará sobre el carácter subjetivo de las probabilidades asignadas. Cuando existan diferencias notables en la asignación de probabilidades podría pedirse a los alumnos que las han hecho que expliquen la información en que se han basado o los

criterios seguidos en su asignación.

Finalmente, para obtener unas probabilidades en las que toda la clase se muestre de acuerdo, podría utilizarse el valor medio o la mediana de las probabilidades asignadas individualmente a los diversos sucesos por los diferentes alumnos. Precisamente este podría ser un contexto adecuado para dar sentido a las medidas de tendencia central, ya que se dispone de una serie de "medidas" del grado de ocurrencia de un suceso y deseamos obtener la mejor estimación.

Probabilidad, como grado de creencia

Para medir la mayor o menor posibilidad de que ocurra un suceso en un experimento, le asignamos un número entre 0 y 1 llamado su *probabilidad.*

La probabilidad varía entre 0 y 1

El suceso seguro siempre ocurre y el suceso imposible no puede ocurrir. Asignamos una probabilidad 0 a un suceso que nunca puede ocurrir, por ejemplo, que salga un 7 al lanzar el dado. Asignamos un 1 a un suceso que ocurre siempre que se realiza el experimento; por ejemplo, al lanzar una moneda es seguro que saldrá o "cara o cruz".

Entre estos dos casos se encuentran el resto de los sucesos asociados a cada experimento. A pesar de que no sabemos cuál de ellos ocurrirá en una prueba particular, algunos de ellos nos merecen más confianza que otros, en función de nuestros conocimientos sobre las condiciones de realización del experimento. Por medio de la probabilidad cuantificamos nuestro grado de creencia acerca de la ocurrencia de cada uno de los sucesos asociados a un experimento. *A cualquier otro suceso distinto del "imposible" y del "seguro" se le asigna un número entre 0 y 1.*

Este valor lo asignamos de acuerdo con nuestra información y la creencia que tengamos en la ocurrencia del suceso. Por ello, diferentes personas podrían asignar una probabilidad distinta al mismo suceso. Por ejemplo, si nos preguntan por la probabilidad de que una cierta persona llegue a cumplir 25 años, diremos que es muy alta. Pero, si su médico sabe que esta persona sufre una enfermedad incurable dará un valor bajo para esta misma probabilidad.

Actividades

4.7. *Esperanza de vida*: A partir de una tabla de vida, hacer predicciones sobre la probabilidad de vivir x años, o de vivir en el año 2000, según sea un chico o una chica, el profesor, etc.

4.8. *Investigación*. Discutir y ordenar la probabilidad de que se produzcan diversos inventos antes de 5, o 10 años (vacunas, viajes interplanetarios, energía,...)

4.9. *Accidentes*. Escribir una serie de frases sobre la reducción o aumento del número de accidentes, probabilidad de que se produzcan en una fecha dada y ordenarlas de mayor a menor probabilidad.

4.10. *Resultados de elecciones*. Con motivo de algunas elecciones escolares, locales, etc, plantear la mayor o menor probabilidad de que resulte elegido un candidato, o de que logre todos los votos, los 2/3, etc. Para ello utiliza los gráficos de alguna encuesta publicada en la prensa local (por ejemplo, un gráfico de barras o sectores).

4.11. Recoger de la prensa los datos de las temperaturas máxima y mínima durante una semana en las capitales de provincia. Confeccionar una tabla estadística con estos datos. ¿Cuál crees que será la temperatura máxima y mínima más probable la próxima semana?

4.12. Busca dos gráficos estadísticos diferentes que hayan aparecido en la prensa local recientemente. Para cada uno de ellos describe el experimento aleatorio al que se refieren; los sucesos asociados y cuál de ellos es más probable. .¿Podrías hacer un gráfico alternativo para representar la información en cada uno de los casos?

4.4. ESTIMACIÓN DE PROBABILIDADES A PARTIR DE LAS FRECUENCIA RELATIVAS

Cuando realizamos un experimento N veces, la frecuencia absoluta del suceso A es el número N_A de veces que ocurre A. El cociente $h(A)=N_A/N$ es la frecuencia relativa del suceso. Se pueden observar las tres propiedades siguientes en las frecuencias relativas:

1. La frecuencia relativa del suceso varía entre 0 y 1.
2. La frecuencia relativa del suceso seguro siempre es 1 en cualquier serie de ensayos.
3. Supongamos que un suceso A se forma uniendo sucesos que no tienen elementos comunes. En este caso, la frecuencia relativa del suceso A es la suma de las frecuencias relativas de los sucesos que lo componen. Por ejemplo, al lanzar un dado, $h(par)= h(2)+h(4)+h(6)$.

Actividades

4.13. *Juegos de dados.* Imagina que estás jugando a los dados con un amigo. Tu compañero indica que hay tres posibilidades diferentes al lanzar dos dados: a) que los dos números sean pares, b) que los dos sean impares y que c) haya un par y un impar. Afirma que los tres casos son igual de probables. ¿Tu qué opinas? Otro compañero sugiere que hagáis un experimento para resolver la discusión. Fíjate en la tabla que te presentamos.

Resultado	Recuento	Frecuencia absoluta	Frecuencia relativa	Nºesperado de veces
Dos números pares				
Dos impares				
Un par y un impar				
Total		20	1	20

a) Trata de adivinar cuantas veces, aproximadamente, saldrá el 3 y cuantas el 5 si lanzas un dado 20 veces. Escribe este número en la columna "número esperado de veces".

b) Lanza el dado 20 veces y anota los resultados en la tabla.

c) El profesor mostrará en la pizarra los resultados de toda la clase. Compara estos resultados con los vuestros y con la estimación que habéis hecho. ¿Cuál de los sucesos es más probable?

4.14. Con el fin de apreciar la ley de estabilidad de las frecuencias relativas y comparar los valores de la probabilidad asignados según la regla de Laplace con el correspondiente concepto frecuencial, se recomienda que los alumnos, por parejas, realicen algunos de los experimentos aleatorios, anotando los resultados de sus experimentos. A continuación, se recogerán todos los resultados de los distintos grupos en una hoja de registro como la siguiente:

Suceso observado:					
Pareja Nº	Nº de experimentos	Frecuencia absoluta	Nº de experimentos acumulados (N)	Frecuencia acumulada (A)	Frecuencia relativa (A/N)
1					
2					
.....					

En un diagrama cartesiano se representarán los puntos (N,A/N), número de experimentos acumulados, frecuencia relativa.

A pesar de que la ley de estabilidad de las frecuencias relativas es válida sólo cuando *n* crece indefinidamente, es posible que los alumnos aprecien una cierta regularidad o tendencia hacia el valor asignado "a

priori", aunque el número de experiencias de clase sea limitado.

El valor de la frecuencia relativa de un suceso no es fijo para *n*, puesto que se trata de un fenómeno aleatorio. Dos alumnos de la clase que realicen el mismo experimento 50 veces pueden obtener diferentes valores de las frecuencias absoluta y relativa del mismo suceso. Sin embargo, para una serie larga de ensayos, las fluctuaciones de la frecuencia relativa son cada vez más raras y de menor magnitud y oscila alrededor de un valor bien determinado. Este hecho .tiene una demostración matemática, en los teoremas conocidos como "*leyes de los grandes números*". También puede observarse experimentalmente; por ejemplo, en las estadísticas recogidas en grandes series de datos sobre natalidad, accidentes, fenómenos atmosféricos, etc…

La convergencia de las frecuencias relativas fue ya observada en el siglo XVIII; Buffon, en 4040 tiradas de una moneda obtuvo cara 2048 veces, siendo la frecuencia relativa de caras, por tanto, 0,5069. Pearson repitió este mismo experimento, obteniendo una frecuencia relativa de 0,5005 para 24000 tiradas. La estabilidad de frecuencias se presenta en fenómenos de tipo muy diverso: sexo, color de pelo o de ojos, accidentes o averías en maquinaria. Llamaremos *probabilidad* de un suceso aleatorio al valor alrededor del cual oscila la frecuencia relativa del mismo, al repetir la experiencia un número grande de veces.

Estimación frecuencial de la probabilidad

La estabilidad de la frecuencia relativa en largas series de ensayos, junto con el hecho de que haya fenómenos para los cuales los sucesos elementales no son equiprobables, hace que pueda estimarse el valor aproximado de la probabilidad de un suceso a partir de la frecuencia relativa obtenida en un número elevado de pruebas. Este es el único método de asignar probabilidades en experimentos tales como "lanzar una chincheta" o "tener un accidente de coche en una operación retorno". .Recuerda, no obstante, que el valor que obtenemos de esta forma es siempre aproximado, es decir, constituye una *estimación de la probabilidad.*

Sabemos, por ejemplo, que, debido a las leyes genéticas la probabilidad de nacer varón o mujer es aproximadamente la misma. Sin embargo, si en un hospital hacemos una estadística de nacimientos no sería raro que un día dado, de diez recién nacidos, 7 u 8 fuesen varones. Sería más raro que fuesen varones 70 o más entre cien recién nacidos, y todavía

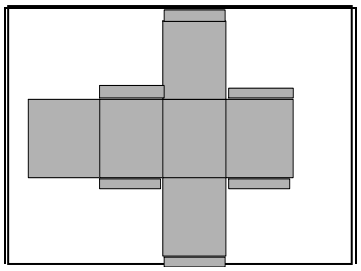
más difícil que más del 70% de entre 100000 recién nacidos lo fuesen.

Con este ejemplo, vemos también que es muy importante el tamaño de la muestra en la estimación de las probabilidades frecuenciales. A mayor tamaño de muestra mayor fiabilidad, porque hay más variabilidad en las muestras pequeñas que en las grandes.

Actividades

4.15. *Construcción de dados.* Un dado ordinario se puede construir recortando en cartulina el siguiente perfil

Figura 4.2



1. Construye un dado recortando en cartulina este perfil, pero numera dos caras con el número 5 y ninguna con el 1.
2. Comparar entre si las probabilidades de obtener un 5, un 3 y un 1. Compáralas también con 0, 1/2 y 1.
3. Construye un dado, recortando en cartulina el perfil dibujado. Pega un pequeño peso en la cara del 1, por ejemplo, un botón. De este modo hemos construido un dado SESGADO. ¿Qué consecuencias tiene el hecho de que una cara del dado pese más que las restantes? En este caso, obtener un 1 ¿es más, menos o igual de probable que antes? ¿Puedes construir un dado sesgado de tal manera que casi siempre salga el 5?

4.16. *Experimentos con chinchetas*. Por parejas, los alumnos lanzan una caja de chinchetas sobre una mesa, contando cuántas de ellas caen de punta o de cabeza. Con los resultados de toda la clase puede estimarse, aproximadamente, la probabilidad de estos dos sucesos y el profesor puede aprovechar para hacer observar a los chicos que existen ejemplos de experimentos en los que la aplicación de la regla de Laplace no es pertinente.

4.17. *Ruletas y tiro al blanco.* Construye una ruleta como la que representamos a continuación. Sólo necesitas un trozo de cartulina, un compás para trazar el contorno circular, un bolígrafo como eje de giro y un clip sujetapapeles parcialmente desenrollado.

a) Da un empujón al clip y observa en qué zona se para. Si se detiene en la zona rayada decimos que ha ocurrido el suceso simple R; si se para en la blanca ocurre el suceso simple B. Describe el espacio muestral

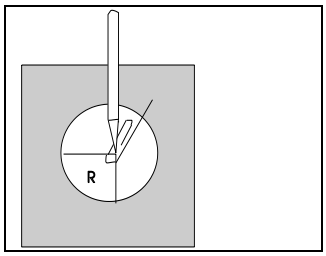


Figura 4.3

b) Asigna probabilidades a los sucesos R y B.

c) ¿Cuál sería la probabilidad de obtener dos veces consecutiva el rojo al girar la ruleta?

4.18. Se pidió a algunos niños lanzar una moneda 150 veces. Algunos lo hicieron correctamente. Otros hicieron trampas. Anotaron con la letra C la aparición de una cara y con X una cruz. Estos son los resultados de Daniel y Diana:

Daniel: c+c++cc++cc+c+c++c++c+ccc+++ccc++c++c+c+c++cc+ccc+

c+c+cc+++cc++c+c++cc+c++cc+c++cc+cc+c+++c++cc++c++

c+c+cc+c++cc+c+c++ccc+cc++c+c++cc+++c+++c+c++ccc++

Diana: +cc+++c++++c+cc+++cc+cc+++cc+ccc+++c++++++c+c+c+c+

+++cccccc+ccc+c+cc+ccccc+ccc++ccc+c+ccccccccc++c+

ccccccc+++++cccc++c+c+cc+cc+cc++++++c+cc++ccc++ccc

¿Hicieron trampas Daniel o Diana? ¿Por qué?.

Simulación de experimentos aleatorios

La realización de experimentos aleatorios usando dispositivos físicos, como dados, fichas, bolas, ruletas, etc. puede requerir bastante tiempo. A veces, incluso puede que no se dispongan de tales dispositivos en número suficiente para toda la clase. Una alternativa válida consiste en simular tales experimentos por medio de una tabla de números aleatorios. Este procedimiento incluso permite resolver problemas de probabilidad reales haciendo las simulaciones con un ordenador.

Llamamos *simulación* a la sustitución de un experimento aleatorio por otro equivalente con el cual se experimenta para obtener estimaciones de probabilidades de sucesos asociados al primer experimento. La estimación de la probabilidad que se obtiene con el experimento simulado es tan válida como si se tratase del experimento real. Este es el método que se emplea para obtener previsiones en las siguientes situaciones:

1. Experimentos complejos, como sería planificar el tráfico durante una

operación salida de vacaciones.

2. Experimentos peligrosos, como estimar la temperatura de control o la velocidad de reacción permitida en una central nuclear.
3. Situaciones futuras: estudios ecológicos o sobre contaminación ambiental.

Actividades

4.19. Explicar cómo usar la tabla de números aleatorios de la figura 4.4, o los números aleatorios generados por tu calculadora, para simular los siguientes experimentos:

Figura 4.4. Tabla de números aleatorios

2034	5600	2400	7583	1104	8422	9868	7768	2512	9575
8849	5451	8504	3811	0132	8635	1732	4345	9047	0199
8915	2894	5638	4436	9692	8061	4665	9252	6729	9605
6989	0682	0085	5906	8542	6884	5719	5081	8779	9071
5093	8880	3466	0212	9475	4957	8474	8550	9572	6770
7940	3305	1183	8918	4397	3167	7342	7780	6745	4688
9808	7499	9925	0695	4721	7597	0922	4715	6821	2259
5667	7590	8599	5032	3042	3666	1160	3413	2050	1796
0644	2848	7347	7161	6813	8276	8175	6534	6107	8350
4153	0293	0882	9755	5109	1484	4798	8039	3593	6369
4621	0121	0251	9783	7697	4079	8952	4884	8838	1587
8490	4941	5203	2932	1008	6544	1137	1018	5123	0347
3160	4107	2194	1314	1310	7060	3075	5273	6592	8875
0140	1600	8468	6585	5257	4874	9097	8684	7877	8881
0483	7097	5973	4235	7466	0821	3261	1359	3706	4676
2657	13867	6896	3132	2648	8947	9518	7472	9285	3067
4286	4327	3848	9128	5350	0407	6215	4059	4546	5170
8445	5087	0964	2800	9369	1980	8490	7760	7548	1060
4946	4327	0966	7861	8381	5865	4447	9063	2085	3635
9786	8853	0667	9100	2303	4455	0389	6145	2618	5401

a) Lanzar tres monedas. Calcular la probabilidad de obtener al menos dos caras.

b) Supongamos que el 10% de bombillas de una fábrica es defectuosa. Las bombillas se venden en cajas de 4 unidades. Simular el experimento consistente en abrir una caja y contar el número de defectos.

4.20. Si en una asignatura de 10 temas has estudiado 8 y el examen consta de dos preguntas, estimar la probabilidad de que no te toquen ninguna de las dos que no te sabes, usando la simulación.

4.21. Se toman 3 fichas de la misma forma y tamaño, de las cuales una es roja por ambas caras; otra es azul por una cara y roja por la otra, y la tercera es azul por las dos caras. El profesor coloca las tres fichas en una caja, que agita

convenientemente, antes de seleccionar una de las tres fichas, al azar. Muestra, a continuación, una de las caras de la ficha elegida, manteniendo la otra tapada, pidiendo a sus alumnos que adivinen el color de la cara oculta. Una vez hechas las apuestas, el profesor muestra la cara oculta. Cada alumno que haya acertado en la predicción efectuada, consigue un punto. Se trata de simular el juego y buscar la mejor estrategia en este juego.

4.5. ASIGNACIÓN DE PROBABILIDADES EN EL CASO DE SUCESOS ELEMENTALES EQUIPROBABLES. REGLA DE LAPLACE

Si un espacio muestral consta de un número finito *n* de sucesos elementales y no tenemos motivo para suponer que alguno de ellos pueda ocurrir con mayor frecuencia que los restantes, la probabilidad de cada uno de estos sucesos elementales es 1/n.

En estos casos, podemos aplicar la llamada *regla de Laplace* para calcular las probabilidades de los sucesos compuestos. Un suceso compuesto que se compone de *k* sucesos elementales tiene, en este caso, una probabilidad igual a *k/n* (regla de Laplace). En el caso de que tengamos motivos para pensar que algún suceso puede darse con mayor frecuencia que otros (por ejemplo, al usar un dado sesgado) o bien cuando el espacio muestral es infinito, no podemos aplicar esta regla.

Actividades

4.22. El experimento consiste en lanzar un dado con forma de dodecaedro, con los números del 1al 12 en sus caras. Encontrar la probabilidad de cada uno de los siguientes sucesos: a) Obtener un número par; b) Obtener un número primo; c) Obtener un divisor de 12.

4.23. Se lanza una moneda tres veces seguidas. a) ¿Cuál es la probabilidad de de obtener 2 caras? b) ¿Cuál es la probabilidad de obtener más caras que cruces?

4.24. Dos sucesos que no pueden ocurrir a la vez se llaman incompatibles. Por ejemplo, no pueden ocurrir a la vez los sucesos "obtener par" y "obtener impar" cuando lanzamos un dado. Tampoco podrían ocurrir a la vez "ser menor que 3" y "ser mayor .que 5". Describe otros ejemplos de otros sucesos incompatibles.

4.6. AXIOMAS DE LA PROBABILIDAD

En los apartados anteriores hemos visto tres modos diferentes de asignar probabilidades, según el tipo de experimento aleatorio:

1. En el caso de espacios muestrales con un número finito de sucesos elementales en los que pueda aplicarse el principio de indiferencia, calculamos las probabilidades usando la *regla de Laplace.*
2. Si no podemos usar la regla de Laplace, pero tenemos información estadística sobre las frecuencias relativas de aparición de distintos sucesos, podemos obtener una *estimación frecuencia*l de las probabilidades.
3. En los demás casos, el único modo de asignar las probabilidades a los sucesos es de modo *subjetivo.*

En todos los casos, las probabilidades cumplen unas mismas propiedades, que se recogen en la *definición axiomática de la probabilidad.* Toda teoría matemática se desarrolla a partir de una serie de axiomas. Generalmente estos axiomas se basan en la abstracción de ciertas propiedades de los fenómenos que se estudian, que para el caso de la probabilidad son las tres primeras propiedades que hemos citado sobre las frecuencias relativas.

Como consecuencia, se considera que la probabilidad es toda aplicación, definida en el conjunto de los sucesos asociados a un experimento aleatorio, que cumpla las tres siguientes propiedades:

- A todo suceso *A* le corresponde una probabilidad *P(A),* número comprendido entre 0 y 1.
- La probabilidad del suceso seguro es 1*, P(E)=1.*
- La probabilidad de un suceso que es unión de sucesos incompatibles es la suma de las probabilidades de los sucesos que lo componen.

Actividades

4.25. Carmen y Daniel han inventado un juego de dados con las siguientes reglas:

Lanzan dos dados sucesivamente y calculan la diferencia de puntos entre el mayor y el menor.

Si resulta una diferencia de 0, 1 o 2 entonces Carmen gana 1 ficha. - Si resulta 3, 4, o 5 es Daniel quien gana una ficha.

Comienzan con un total de 20 fichas y el juego termina cuando no quedan más. ¿Te parece que este juego es equitativo? Si tuvieras que jugar, ¿cuál jugador preferirías ser?

4.26. Se toma un número comprendido entre 0 y 999 ¿Cuál es la probabilidad de que la cifra central sea mayor que las otras dos? ¿Cuál es la probabilidad de que el número sea múltiplo de 5?

4.27. Se dispone de dos bolsas, cada una de las cuales contiene diez bolas numeradas del 0 al 9. Realizamos un experimento aleatorio consistente en extraer una bola de cada una de las bolsas. 1) Describir el espacio muestral asociado al experimento. 2) Hallar la probabilidad del suceso A "obtener dos bolas iguales".

4.28. A un congreso de científicos asisten 100 congresistas. De ellos, 80 hablan francés y 40 inglés. ¿Cual es la probabilidad de que dos congresistas elegidos al azar no puedan entenderse sin intérprete?

4.7. COMBINATORIA

Al aplicar la regla de Laplace, se presenta a menudo el problema de calcular el número de elementos de un cierto subconjunto del espacio muestral. Podemos utilizar, en este caso el cálculo combinatorio.

Regla del producto

Si disponemos de dos conjuntos de m y n elementos, $a_1,a_2,.....a_m$ y $b_1, b_2....b_n$, es posible formar $m \times n$ .parejas diferentes de la forma (a_i,b_j), en las que el primer elemento pertenece al primer conjunto y el segundo elemento al segundo conjunto.

Se pueden formar $n_1,n_2....n_r$ grupos diferentes de la forma $(a_{j1}, b_{j2}....., x_{jr})$, tomando a_{j1} del conjunto $a_1, a_2,.....a_{n1}$; b_{j2} del conjunto $b_1, b_2,.....b_{n2}$... y x_{jr} del conjunto $x_1,x_2....x_{nr}$.

Ejemplo 4.1. Para formar la matrícula de un coche con 2 letras y 4 cifras, podemos elegir entre 24*24*10*10*10*10 diferentes, supuesto que puede tomarse cualquier letra y cualquier dígito en cada posición.

Muestreo

En el análisis combinatorio, el problema es deducir el número de muestras diferentes que pueden formarse a partir de un conjunto dado. Podemos distinguir entre muestreo *con y sin reemplazamiento*, según que cada elemento de la población pueda formar o no más de una vez parte en la muestra.

Por otro lado, es preciso distinguir entre muestras ordenadas y no ordenadas. Una muestra se llama ordenada, cuando el orden en que han sido extraídos sus elementos es fundamental, y es por tanto, tenido en cuenta. En este caso, dos muestras formadas por los mismos elementos, pero difiriendo en el orden, son consideradas distintas. Cuando el orden no influye, de modo que dos muestras son diferentes sólo si varían en algún elemento, las muestras se llaman no ordenadas. De acuerdo con la clasificación anterior, a partir de un conjunto o población de m elementos, podemos formar cuatro tipos de muestras o grupos de tamaño *n*. En el estudio combinatorio clásico estos grupos reciben el nombre de variaciones con o sin repetición y combinaciones con o sin repetición.

Variaciones de m elementos tomados n a n (sin repetición):

Son las diferentes muestras ordenadas de tamaño n que pueden formarse de la población, sin reemplazamiento. Por tanto, no hay elementos repetidos, y dos grupos son considerados distintos si difieren, bien en un elemento, o si teniendo los mismos elementos, son cambiados de orden. Para aclarar la definición, formemos las variaciones de orden 3, de las letras A, B, C, y D

A	*AB*	*ABC ABD*
	AC	*ACB ACD*
	AD	*ADB ADC*
B	*BA*	*BAC BAD*
	BC	*BCA BCD*
	BD	*BDA BDC*
C	*CA*	*CAB CAD*
	CB	*CBA CBD*
	CD	*CDA CDB*
D	*DA*	*DAB DAC*
	DB	*DBA DBC*
	DC	*DCA DCB*

Para formar las variaciones de *m* elementos de orden 1 hay *m* posibilidades. Para formar las de orden 2, aplicando la regla del producto, y puesto que no podemos tomar el elemento ya elegido, hay *m(m-1)* posibilidades. En general, si denotamos por $V_{m,n}$ el número de variaciones sin repetición de m elementos tomados n a n se verifica:

(4.4) $$V_{m,n} = m(m-1)(m-2).....(m-n+1).$$

Ejemplo 4.2. ¿De cuantas formas posibles pueden colocarse en fila 5 personas? Hay 4*4*3*2*1=120 formas diferentes.

Un caso especial de las variaciones sin repetición es cuando *m=n*. En dicho caso, obtenemos todas las maneras posibles de colocar o "permutaciones" de *m* elementos. Se tiene que:

(4.5) $$V_{m,m} = m(m-1)(m-2).....3.2.1$$

Al número $V_{m,m}$ se le llama factorial de *m* y se representa por *m!* Por convenio se admite que *0!=1*.

Variaciones con repetición de m elementos tomados n a n.

Son las distintas muestras ordenadas de tamaño *n* que pueden formarse a partir de una población de m elementos, cuando el muestreo se efectúa con reemplazamiento. Existe, por tanto, la posibilidad de repetir elementos, y dos grupos son considerados distintos si tienen algún elemento diferente, o si han sido extraídos en distinto orden.

Veamos cuantas variaciones con repetición de tamaño 3 pueden formarse con las letras *A, B, C, y D.*

A *AA AAA AAB AAC AAD*
AB ABA ABB ABC ABD
AC ACA ACB ACC ACD
AD ADA ADB ADC ADD
B *BA BAA BAB BAC BAD*
BB BBA BBB BBC BBD
BC BCA BCB BCC BCD
BD BDA BDB BDC BDD
C *CA CAA CAB CAC CAD*
CB CBA CBB CBC CBD
CC CCA CCB CCC CCD
CD CDA CDB CDC CDD
D *DA DAA DAB DAC DAD*
DB DBA DBB DBC DBD
DC DCA DCB DCC DCD
DD DDA DDB DDC DDD

Si llamamos $VR_{m,n}$ al número de variaciones con repetición de m elementos tomados n a n, aplicando la regla del producto se deduce que:

$$VR_{m,\,n} = m^n$$

Combinaciones sin repetición de m elementos tomados n a n.

Son las distintas muestras no ordenadas sin reemplazamiento de n elementos, que pueden formarse a partir de m individuos dados. Por tanto, no se repiten elementos en la muestra, y dos grupos que constan de los mismos elementos se consideran idénticos, aunque estén colocados en orden diferente. En otras palabras, las combinaciones sin repetición de m elementos, tomados *n* a *n*, son los diferentes subconjuntos de tamaño n que pueden formarse a partir de un conjunto de m elementos. Formaremos las combinaciones sin repetición de orden 3 a partir de las letras A, *B, C y D.*

A *AB ABC ABD*
AC ACD
AD
B *BC BCD*
BD
C *CD*

Denotaremos como $C_{m,n}$, o bien como $\binom{m}{n}$ el número de combinaciones de m elementos tomados *n* a *n*. Para calcular su valor, basta tener en cuenta que, cada muestra no ordenada sin reemplazamineto de tamaño *n* puede dar lugar a *n!* muestras ordenadas con reemplazamiento distintas permutando sus elementos. Por tanto:

$$\binom{m}{n} n! = V_{m,n}$$

(4.6)
$$\binom{m}{n} = \frac{m(m-1)....(m-n+1)}{n!} = \frac{m!}{n!(m-n)!}$$

Los números $\binom{m}{n}$ reciben el nombre de números combinatorios y tienen, entre otras las siguientes propiedades:

(4.7)
$$\binom{n}{0} = 1 = \binom{n}{n}$$

(4.8) $$\binom{m}{n} = \binom{m}{m-n}$$

(4.9) $$\binom{m}{n} + \binom{m}{n+1} = \binom{m+1}{n+1}$$

Esta última propiedad se utiliza para calcular los números combinatorios recurrentemente, formando el llamado triángulo de Tartaglia, dos de cuyos lados están formados por unos, y el resto de los números se calcula como suma de los dos situados inmediatamente encima de él (figura 4.5).

Figura 4.5. Triangulo de Tartaglia

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1

Combinaciones con repetición

Por último, si suponemos que en una muestra no ordenada se permite el reemplazo de elementos, obtenemos las combinaciones con repetición. Formaremos las combinaciones con repetición de las letras A, B, C, y D de orden 3.

A *AA AAA AAB AAC AAD*
AB ABB ABC ABD
AC ACC ACD
AD ADD
B *BB BBB BBC BBD*
BC BCC BCD
BD BDD
C *CC CCC CCD*
CD CDD
D *DD DDD*

Puede mostrarse que el número de combinaciones con repetición de m elementos tomados *n* a *n* es:

(4.10) $$CR_{m,n} = C_{m+n-1,n}$$

Actividades

4.29. La señora Rodríguez tiene 6 sombreros, 4 camisas y 7 faldas diferentes. ¿De cuantas formas puede elegir un vestuario formado por falda, camisa y sombrero?

4.30. ¿De cuantas maneras pueden colocarse 8 torres en un tablero de ajedrez sin que ninguna pueda comer a las otras?

4.31. ¿Cuantos resultados distintos puede obtenerse al lanzar 6 veces una moneda?

4.32. ¿Cuantas palabras diferentes de 4 letras pueden formarse en código morse?

4.33. ¿De cuantas formas pueden colocarse 7 libros en un estante?

4.34. Diez jugadores de tenis compiten en un torneo. ¿De cuantas maneras pueden ordenarse para jugar el primer encuentro, si se dispone de una sola pista?

4.35. Una clase consta de 10 chicos y 10 chicas. ¿De cuantas formas se pueden dividir en 2 grupos de 10 estudiantes? ¿Cual es la probabilidad de que cada grupo está formado por 5 chicos y 5 chicas?

4.36. Un frutero vende plátanos, peras y manzanas a duro la pieza. Con 10 duros ¿Cuantas compras diferentes pueden hacerse?

4.37. En un polígono regular convexo de n lados se unen al azar dos vértices distintos. Hallar la probabilidad de obtener una diagonal.

4.38. Una persona ha colocado revueltos 10 pares diferentes de guantes en un cajón. ¿Cual es la probabilidad de que, al tomar dos guantes al azar, uno sea de la mano derecha y otro de la izquierda? ¿Cual será la probabilidad de que sean de un mismo par?

4.39. En una jaula hay 9 cobayas de los cuales 6 son machos y 3 hembras. ¿Cuál es la probabilidad de que al tomar 5 de ellos sólo se elija una hembra?

4.40. En un sorteo que consta de 500 números hay 10 premios. Una persona compra 5 papeletas. ¿Cuál es la probabilidad de recibir algún premio?.

4.8. PROBABILIDAD CONDICIONAL

En los ejemplos anteriores se ha considerado la probabilidad de cada suceso aisladamente. Así, hablamos de la probabilidad de nacer varón o hembra, o la probabilidad de obtener 14 aciertos en una quiniela. etc. Sin embargo, en algunas situaciones, podemos tener alguna información sobre la ocurrencia de sucesos que pensamos pueden estar relacionados con el que nos interesa. Así podemos preguntarnos, por ejemplo, por la probabilidad de conseguir los 14 aciertos de la quiniela, si sabemos que hemos acertado 12 resultados y aún no han terminado los dos últimos partidos; o bien un matrimonio que ya tiene tres hijos varones puede preguntarse por la probabilidad de que el próximo sea una hembra. En

estos casos, resulta conveniente introducir el concepto de probabilidad condicionada.

Sea E el espacio muestral asociado a un experimento aleatorio y *A* y *B* dos resultados posibles de dicho experimento. Si en *N* pruebas ha resultado N_A veces el suceso *A*, y entre estas ha resultado N_{AB} veces el *B*, tendremos:

$$h(A) = \frac{N_A}{N}; \qquad h(B/A) = \frac{N_{AB}}{N_A}; \qquad h(A \cap B) = \frac{N_{AB}}{N}$$

como

$$\frac{N_{AB}}{N} = \frac{N_A}{N} \frac{N_{AB}}{N_A}$$

resulta que para las frecuencias relativas se cumple:

$$h(A \cap B) = h(A) * h(B/A),$$

y en consecuencia,

$$h(B/A) = \frac{h(A \cap B)}{h(A)}$$

Al aumentar el número de pruebas *N*, las frecuencias relativas de *A* y $A \cap B$ oscilarán alrededor de las probabilidades *P(A)* y $P(A \cap B)$. Por ello, la frecuencia relativa *h(B/A)* oscilará alrededor del cociente

$$\frac{P(A \cap B)}{P(A)} \tag{4.11}$$

Por este motivo, se define la "probabilidad del suceso *B* condicionada por el suceso *A*" como el cociente dado en (4.11), y se representa por *P(B/A)*. Este número puede interpretarse como la probabilidad de que ocurra el suceso *B* en el caso de que se haya verificado el suceso *A*.

Ejemplo 4.3. Al lanzar un dado, sea *A* el suceso "obtener el número 2" y *B* el suceso "obtener número par". ¿Cual es la probabilidad de que el número obtenido sea 2, supuesto que ha resultado un número par?

En este caso tenemos:

$$E = \{1,2,3,4,5,6\}; \quad A = \{2\}; \qquad B = \{2,4,6\} \qquad A \cap B = \{2\}$$

Por tanto, se verifica: $P(A) = 1/6; \; P(B) = 3/6; P(A/B) = P(A \cap B)/P(B) = 1/3$

Ejemplo 4.4. Si un matrimonio tiene tres hijos varones, calcular la

probabilidad de que el cuarto sea mujer.

En este ejemplo haremos la simplificación de suponer igualmente probables el hecho de nacer varón o mujer. Si no se tuviese información sobre el género de los tres primeros hijos, las posibles combinaciones de sexos en un matrimonio de cuatro hijos serán:

E = {vvvv, mvvv, vmvv, vvmv, vvvm, vvmm, vmvm, vmmv,

mvvm, mvmv, mmvv, mmmv, mmvm, mvmm, vmmm, mmmm}

Sea A= "los tres primeros son varones" y *B*= "el cuarto es mujer", entonces tendremos,

A={vvvv,vvvm};
B={vvvm,vvmm,vmvm,mvvm,mmvm,mvmm,vmmm,mmmm}

A∩B = {vvvm}; P(B) = 8/16= 1/2; P(A) = 2/16 = 1/8. Por tanto,

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = 1/2$$

La probabilidad de que los tres primeros sean varones si el cuatro es mujer *P(A/B)* es 1/8, como puede comprobarse fácilmente.

Probabilidad de la intersección de sucesos

De la definición de la probabilidad condicional obtenemos las dos expresiones siguientes, que nos permiten calcular *P(A ∩ B):*

(4.12) $P(A \cap B) = P(A)*P(B/A)$

(4.13) $P(A \cap B) = P(B)*P(A/B)$

Así, en el ejemplo 4.4:

$$P(A \cap B) = P(vvvh) = \frac{1}{16} = \frac{1}{8} \cdot \frac{1}{2} = P(A/B) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{8} = P(B/A) \cdot P(A)$$

Esta regla, que se conoce como "*teorema de la probabilidad compuesta o del producto*", puede generalizarse sin dificultad a tres o más sucesos:

$$P(A \cap B \cap C) = P(A).P(B/A).P(C/A \cap B)$$

$$P(A_1 \cap A_2 \cap .... \cap A_n) = P(A_1)P(A_2/A_1)...P(A_n/A_1 \cap ... \cap A_{n-1})$$

Ejemplo 4.5. Supongamos que una urna contiene 3 bolas rojas y 2 blancas y nos preguntamos cual será la probabilidad de que, tomando 3 bolas de la urna, sin reemplazamiento, las 3 sean rojas.

Sea A_1 = "la 1ª bola es roja"

A_2 = " la 2ª bola es roja"

A_3 = " la 3ª bola es roja"

$$P(A_1 \cap A_2 \cap A_3) = P(A_1).P(A_2/A_1).P(A_3/A_1 \cap A_2) = (3/5)(2/4)(1/3)$$

Sucesos dependientes e independientes

En los ejemplos anteriores hemos observado que, a veces, la probabilidad de un suceso *B* condicionada por otro *A* es la misma que la probabilidad del suceso *B*, cuando no se impone ninguna condición. Esta propiedad se utiliza para la definición de dependencia e independencia entre sucesos aleatorios.

Sean *A* y *B* dos sucesos aleatorios asociados a un mismo experimento. Decimos que *A no depende de B* cuando $P(A/B) = P(A)$ y que A depende de B cuando P*(A/B) es diferente de P(A).* En el caso de que *A* no dependa de *B* se verifica: $.P(A \cap B) = P(A) \times P(B)$, y por tanto, $P(B/A) = P(B)$, por lo cual *B* tampoco depende *A*. Dos sucesos son, por tanto, dependientes o independientes mutuamente.

En algunas situaciones, como es el lanzamiento sucesivo de monedas, dados, etc, nos encontramos ante la situación de la repetición del mismo experimento un número dado de veces. Es fácil ver de la naturaleza de este tipo de pruebas que el resultado de cada una de ellas no influye en el resto. En dicho caso:

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1).\ P(A_2)\ldots P(A_n)$$

(*regla de multiplicación de probabilidades*), donde A_i es un suceso relacionado con la prueba *i.*

Ejemplo 4.6. Al lanzar dos dados, la probabilidad de obtener doble seis será:

$$P(6,6) = P(6).P(6) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$$

Ejemplo 4.7. La probabilidad de que un matrimonio de 4 hijos tenga al menos una niña:

$$P(\text{al menos una niña}) = 1 - Pr(vvvv) = 1 - \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = 15/16$$

Actividades

4.41. La probabilidad de que un hombre viva 65 años es 2/5 y la probabilidad de viva una mujer es 2/3. Se pide: a) Probabilidad de que ambos vivan 65 años; b) Probabilidad de que viva sólo el hombre; c) Probabilidad de que viva sólo la mujer; d) Probabilidad de viva uno de los dos al menos.

4.42. La tabla de longevidad en un cierto país indica que la probabilidad de llegar a los 25 años es 0,95, mientras que la de llegar a los 65 años es 0,64. Si una persona tiene 25 años, ¿cual es la probabilidad de que llegue a los 65 años?

4.43. Una caja contiene las 11 letras MIIIIPPSSSS. Las letras son extraídas una a una sin reemplazamiento y los resultados se registran en orden. Encontrar la probabilidad de que resulte MISSISSIPPI.

4.44. Se tienen dos cajas con las siguientes letras: SOS, SOS, SOS. Se debe elegir una de las dos cajas y a continuación extraer, al azar, tres letras, una a una sin reemplazamiento. Si el resultado es SOS entonces se gana un premio. ¿Qué caja elegirías?

4.45. Un temario de examen se compone de 40 temas de los que un estudiante conoce 30. El examen consta de 2 temas a los cuales se ha de contestar. ¿Cual es la probabilidad de aprobar el examen? ¿Y si el alumno puede elegir 2 temas entre 3?

4.46. En una facultad el 45% de los estudiantes dominan el inglés, el 25% tiene conocimientos de informática y un 10% las dos cosas. Si tomamos al azar un alumno de los que hablan inglés, ¿Cual será la probabilidad de que también tenga conocimientos de informática? Si tomamos un alumno al azar ¿Cual es la probabilidad de que no sepa inglés ni informática?

4.47. Un cierto análisis clínico da resultados positivos en 2 de cada 3 enfermos de hígado. Si a tres enfermos de hígado de les efectúa esta prueba ¿Cual es la probabilidad de obtener al menos un resultado positivo?

4.48. Cada vez que el señor García asiste a una reunión de 7 personas, apuesta 100 pesetas contra 1 a que dos de ellas al menos han nacido el mismo día de la semana. ¿Cuál es la probabilidad de que pierda su apuesta?

4.49. Dos chicos juegan al baloncesto. Pedro encesta 3 de cada 5 pelotas lanzadas, mientras que Juan logra 2 de cada 3 intentos. Si cada uno hace un lanzamiento. ¿Cual es la probabilidad de que ambos logren encestar?

4.50. ¿Cuál es la probabilidad de obtener 6 números diferentes al lanzar 6 veces un dado?

4.51. El 1 % de la población de un país es daltónica. Tomamos una muestra de n

personas. ¿Cual es el mínimo n, para que la probabilidad de obtener al menos un daltónico sea mayor de 0,95?

4.52. Cada uno de los motores de un avión puede averiarse durante un vuelo, con probabilidad 0,01. El avión puede continuar su vuelo si funcionan al menos la mitad de los motores. ¿Que es más seguro, un avión de 2 o de 4 motores?

4.53. Ruletas no transitivas: Supongamos que tenemos tres ruletas. Con la primera siempre obtenemos el número 3. Con la segunda obtenemos el número 1 con probabilidad 0,52 y el número 5 con probabilidad 0,48. Con la tercera obtenemos el número 0 con probabilidad 0,25 y el número 4 con probabilidad 0,74.

Jugamos a un juego en el que dos jugadores eligen una ruleta cada uno y gana aquél que consiga el número mayor al girar la ruleta. ¿Cuál jugador tiene ventaja, el que elige la ruleta en primer o segundo lugar?

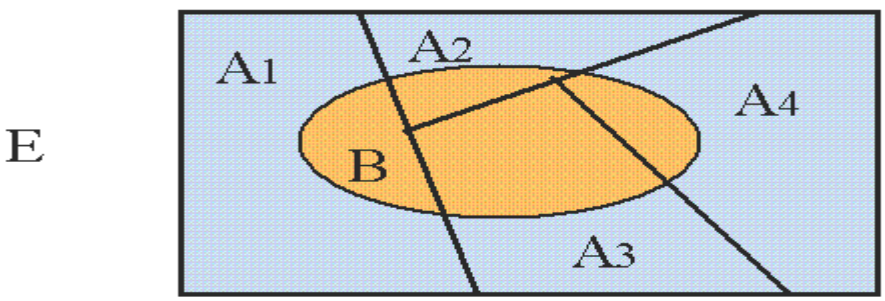
4.9. TEOREMAS DE LA PROBABILIDAD TOTAL Y DE BAYES

En el año 1763, dos años después de la muerte de *Thomas Bayes* (1702-1761), se publicó una memoria en la que aparece, por vez primera, la determinación de la probabilidad de las causas a partir de los efectos que han podido ser observados. El cálculo de dichas probabilidades recibe el nombre de teorema de Bayes.

El Teorema de Bayes es uno de los más importantes teoremas en estadística, porque nos permite aprender de la experiencia. Más concretamente, la regla de Bayes nos permite calcular cómo se modifican las probabilidades de determinados sucesos, cuando se conoce alguna información adicional.

Teorema de Bayes

Consideremos un experimento aleatorio y supongamos que su espacio muestral asociado es *E*. Sean los sucesos A_1, A_2, A_n una partición de *E*, cuyas probabilidades se conocen. Sea *B* un suceso cualquiera del espacio muestral, del que conocemos las probabilidades *P(B/Ai)*. El siguiente esquema representa esta situación.



El teorema de Bayes permite calcular las probabilidades $P(A_i/B)$, mediante la siguiente fórmula:

$$(4.14)\quad P(A_i / B) = \frac{P(A_i) \times P(B / A_i)}{P(A_1) \times P(B / A_1) + P(A_2) \times P(B / A_2) + \ldots + P(A_n) \times P(B / A_n)}$$

Las probabilidades $P(A_i)$ se denominan usualmente *probabilidades iniciales*, y $P(A_i/B)$ *probabilidades finales*, que se calcularán conocida la información $P(B/A_i)$, también llamadas *verosimilitudes*. El denominador de esta fórmula es el teorema de la probabilidad total $P(B)$. Es constante, por lo que si definimos *K* de la forma siguiente:

$$(4.15)\qquad K = \frac{1}{P(A_1) \times P(B / A_1) + P(A_2) \times P(B / A_2) + \ldots + P(A_n) \times P(B / A_n)}$$

La fórmula de Bayes quedaría así:

$$(4.16)\qquad \boldsymbol{P(A_i / B) = K \times P(A_i) \times P(B / A_i)}$$

El Teorema de Bayes nos indica que la probabilidad final de *Ai* dado *B* es proporcional al producto de la probabilidad inicial de *Ai* por la verosimilitud de los datos dado *Ai*. La constante de proporcionalidad *K* es la inversa de la probabilidad de *B* que se calcula en (4.15).

Organización de los cálculos

El teorema de Bayes puede también generalizarse a mayor número de sucesos, como en el ejemplo que mostramos a continuación:

Ejemplo 4.8. Tres máquinas denominadas A, B y C, producen un 43%, 26% y 31% de la producción total de una empresa respectivamente. Se ha detectado que un 8%, 2% y 1,6% del producto manufacturado por estas máquinas es defectuoso. Se selecciona un producto al azar y se encuentra que es defectuoso, ¿cuál es la probabilidad de que el producto haya sido fabricado en la máquina B? La probabilidad pedida se calcula usando el Teorema de Bayes:

$$P(B/D) = \frac{P(B \cap D)}{P(D)} = \frac{P(B) \times P(D/B)}{P(A) \times P(D/A) + P(B) \times P(D/B) + P(C) \times P(D/C)} =$$

$$P(B/D) = \frac{0,26 \times 0,02}{(0,43 \times 0,08) + (0,26 \times 0,02) + (0,31 \times 0,016)} = \frac{0,0052}{0,04456} = 0,116697$$

Una forma sencilla de organizar los cálculos para pasar de la probabilidades iniciales *P(A), P(B), P(C)* a las probabilidades finales *P(A/D), P(B/D), P(C/D),* es utilizando una tabla similar a la 4.1.

Tabla 4.1. Organización de cálculos de probabilidad final

(1) Sucesos de interés	(2) Probabilidad inicial	(3) Verosimilitud	(4) Producto	(5) Probabilidad final
A	0,43	0,08	0,0344	0,7720
B	0,26	0,02	0,0052	0,1167
C	0,31	0,016	0,00496	0,1113
Suma			0,04456	1

- En la columna (1) ponemos los sucesos de interés, en este caso las máquinas A, B y C.
- En la columna (2) ponemos las probabilidades inicial y en la (3) las verosimilitudes, que son los datos del problema.
- Calculamos ahora los numeradores de la fórmula de Bayes (producto (2)x(3) en la columna (4).
- Sumamos la columna (4) para obtener el denominador de la fórmula de Bayes.
- En la columna (5) obtenemos las probabilidades final, dividiendo cada celda en la columna (4) por la suma anterior.

Ejemplo 4.9. La narcolepsia es un trastorno primario del sueño, cuya sintomatología principal es la aparición recurrente e irresistible de ataques de sueño reparador. Las personas que tienen este trastorno no descansan bien; por tanto se vuelven irritables y no rinden lo que quisieran.

En una gran ciudad una de cada 1000 personas sufre narcolepsia.

Seleccionamos al azar una persona en una gran ciudad ¿Cuál es la probabilidad de que tenga narcolepsia?

- Probabilidad inicial: probabilidad inicial de sufrir narcolepsia en una persona tomada al azar de la población *P(N)= 1/1000*

Supón que el test es positivo en 99 de cada 100 personas enfermas y también en 2 de cada 100 personas sanas.

- *Verosimilitud:* probabilidad de que el test sea positivo si se tiene narcolepsia (sensibilidad del test) *P(+/N)= 99/100*
- Probabilidad *final:* probabilidad final de sufrir narcolepsia si el test ha dado positivo *P(N/+).* Para calcular esto, hay que aplicar el teorema de Bayes, que vemos a continuación

Vamos a completar en la tabla siguiente los datos que conocemos de la situación:

	Test +	Test -	Total
Sufren narcolepsia	99	1	100
No sufren narcolepsia	1998	97902	99.900
Total	2097	97903	100.000

Para calcular la probabilidad final, habrá que calcular:

$$P(N/+) = \frac{99}{99+1998} = 0{,}0472$$

El resultado no deja de sorprendernos a primera vista: ¡Puede parecer que las pruebas médicas son poco fiables! Lo que pasa es que el número de personas que tienen la enfermedad, en el total de la población es muy pequeño.

- La probabilidad de que el test de positivo si la persona está enferma es muy alta. Casi todas ellas son detectadas en el test (el test tiene mucha *sensibilidad).*
- Por otro lado, la probabilidad de un resultado positivo si se está sano (*falso positivo)* es muy pequeña.

- Pero un suceso con probabilidad pequeña no es un suceso *imposible* y puede ocurrir. Más aún, si el número de personas que pasan la prueba es muy grande, pueden aparecer más falsos positivos que positivos reales (como en el ejemplo).

Las pruebas médicas son bastante fiables, pero el resultado anterior se explica porque el número de personas sanas a las que la prueba da positiva es mucho mayor (aunque sólo una pequeña proporción de sanos tiene un test positivo) que el número de personas enfermas a las que la prueba da positiva (aunque casi todos los enfermos tienen el test positivo). ¡Pero es que el número de enfermos en la población es muy pequeño!

Ejemplo 4.9. ¿Qué ocurriría si, en lugar de narcolepsia, el test da positivo en insomnio?

*L*a prevalencia del insomnio en la población es bastante mayor que la de la narcolepsia (un 15%), es decir, la probabilidad inicial de insomnio es P(I) = 15/100.

La probabilidad de que el test de positivo en una persona con insomnio es de 0,99 y la probabilidad de que de positivo en una persona sana es de 0,02.

Calculemos ahora la probabilidad final de sufrir insomnio si la prueba es positiva

$$P(I/+)=\frac{P(I\cap +)}{P(+)}=0,8973$$

Vemos que el resultado de la prueba de insomnio es mucho más fiable, porque la prevalencia de insomnio en la población es mucho mayor que la narcolepsia.

Ejemplo 4.10. El 22% de los clientes de una compañía de seguros de accidentes de automóvil es menor de 30 años. Las estadísticas indican que el 11% de los conductores menores de 30 años tiene un accidente y que sólo el 5% de los mayores o iguales de 30 años tienen un accidente. ¿Qué porcentaje de accidentes pagados por la compañía son de clientes menores de 30 años?

- P(menor 30 años)=0,22; es una probabilidad inicial.

- P (mayor de 30 años)=0,78; es una probabilidad inicial.
- P(accidente /menor 30 años)=0,11; es una verosimilitud.
- P(accidente /mayor o igual de 30 años)=0,05; es una verosimilitud.

Para resolver el ejercicio colocamos en la tabla los datos conocidos (sucesos de interés, sus probabilidades iniciales y verosimilitudes. A continuación realizamos los productos de las columnas (2) y (3), obtenemos la suma de los productos (columna 4)

Finalmente dividimos cada uno de los productos (columna 4) por la suma de todos los productos (0,0632) para obtener las probabilidades finales. Si los cálculos son correctos la suma de las probabilidades finales es igual a 1

Tabla 4.2. Cálculos de probabilidad final

(1) Sucesos de interés	(2) Probabilidad inicial	(3) Verosimilitud de accidente	(4) Producto	(5) Probabilidad final
Menor 30 años	0,22	0,11	0,0242	0,3829
Mayor 30 años	0,88	0,05	0,044	0,6171
Suma			0,0682	1

El 38% de los accidentes que tiene que pagar la compañía son de conductores de menos de 30 años.

Ejemplo 4.11. Alrededor del 15% de las madres primerizas sienten ansiedad en los días posteriores al parto. En el 90% de los casos esta ansiedad disminuye y en pocos días desaparece, pero en los casos restantes puede aparecer una depresión que necesite tratamiento. Un 2% de madres primeriza que no sufrieron ansiedad en los primeros días desarrolla una depresión tras las primeras semanas del nacimiento. Si una madre primeriza tuvo una depresión post-parto, ¿Qué porcentaje de depresiones post- parto no fueron previstas los primeros días después del parto?

- P(ansiedad en las primeras semanas)=0,15; es una probabilidad inicial.
- P(no sufrir ansiedad en las primeras semanas)=0,85; es una probabilidad inicial.
- P(depresión /si hubo ansiedad inicial)=0,10; es una verosimilitud.

- P(depresión /si no hubo ansiedad inicial)=0,02; es una verosimilitud.

Puesto que tenemos que pasar de probabilidades iniciales a finales, hay que usar el teorema de Bayes. Usaremos la tabla.

(1) Sucesos de interés	(2) Probabilidad inicial	(3) Verosimilitud de depresión	(4) Producto	(5) Probabilidad final
Ansiedad inicial	0,15	0,1	0,015	0,4688
No ansiedad	0,85	0,02	0,017	0,5313
Suma			0,032	1

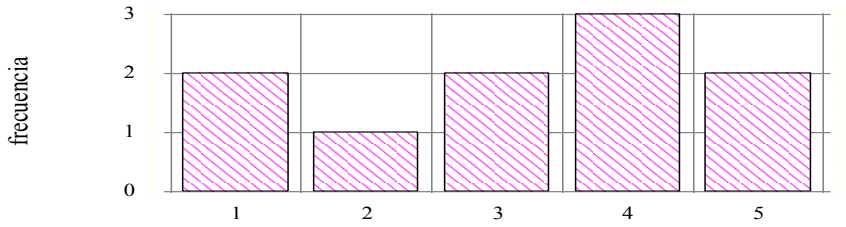
Casi la mitad de las depresiones post-parto no se previeron en las primeras semanas.

4.10. VARIABLE ALEATORIA DISCRETA

Si realizamos *n* pruebas o repeticiones de un experimento aleatorio, obtenemos un conjunto de *n* observaciones o resultados, que constituyen lo que se llama una muestra aleatoria de tamaño *n.* Este conjunto de resultados dará lugar a una tabla estadística en la cual a unos valores de la variable corresponden unas ciertas frecuencias. Así, si lanzamos un dado 10 veces, podríamos obtener la colección de resultados siguientes:

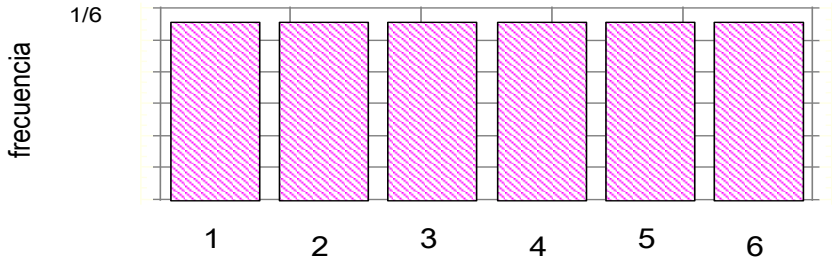
4, 3, 1, 5, 4, 4, 1, 2, 3, 5

La figura 4.4 indica la distribución de frecuencias correspondiente a estos datos, entre los cuales hemos supuesto que no ha aparecido el 6, ya que este hecho es plausible en una muestra de sólo 10 elementos. La variable *X* que representa únicamente los *n* resultados de *n* realizaciones de un experimento aleatorio recibe el nombre de *variable estadística.*



Si imaginamos que el experimento aleatorio se repite indefinidamente, la infinidad de resultados posibles da origen a la noción

de *variable aleatoria* asociada al experimento. En el ejemplo, si suponemos que se lanza el dado un número grande de veces, los resultados posibles serán 1,2,3,4,5,6 y, además, las frecuencias relativas de cada resultado tienden a la probabilidad, que es 1/6.La variable, que representamos por Γ, y que toma los valores 1,2,3,4,5,6, con probabilidad 1/6 para cada valor, recibe el nombre de *variable aleatoria* (figura 4.7).



Actividades

4.54. Consideramos el experimento de lanzar dos dados y anotar los resultados obtenidos. El espacio muestral será: E{(1,1), (1,2),...,(1,6), ..., (6,6)}. Podemos definir distintas variables aleatorias asociadas a este experimento. Una podría ser la correspondencia que asocia a cada elemento de E, la suma de puntos. Escribe en una tabla los valores posibles de esta variable y sus respectivas probabilidades.

4.55. Monedas dependientes. Una bolsa contiene 7 monedas de 100p, 50p, 50p, 50p, 10p, 10p, 10p. Sacamos dos monedas al azar. ¿Cuál es el valor esperado de su suma? ¿Depende este valor esperado de si la primera moneda es o no reemplazada? ¿Por qué?

4.56. Paradoja de Blythe. Tenemos tres ruletas: La primera siempre da como resultado el número 3. La segunda da como resultado 2 con probabilidad 0.51, 4 con probabilidad 0.29 y 6 con probabilidad 0.20. La tercera da como resultados 1 con probabilidad 0.52 y 5 con probabilidad 0.48. Si cada uno de dos jugadores tiene que elegir una ruleta, y gana el que obtenga el número mayor, ¿cual es la mejor elección para el primer jugador? ¿Cambie esta elección si son tres los jugadores?

De los ejemplos anteriores, podemos afirmar que una *variable aleatoria* es una variable cuyos valores dependen del resultado de un experimento aleatorio; Frecuentemente el resultado de un experimento se expresa en forma numérica y, en consecuencia, tal resultado es una variable aleatoria. Por ejemplo: "observar la temperatura diaria a las 8 h. en Jaén", "Observar la altura (o bien, el peso, pulsaciones por segundo, el C.I. etc), de un colectivo de individuos.

De modo similar a las variables estadísticas, clasificamos las variables aleatorias en discretas o continuas según que el conjunto de valores que puedan tomar sea o no numerable.

4.11. DISTRIBUCION DE PROBABILIDAD DE UNA VARIABLE ALEATORIA DISCRETA

Una variable aleatoria discreta queda especificada por su distribución de probabilidad, o relación en la que se exprese los posibles valores de la variable y, para cada uno de ellos, la probabilidad de que ocurra. Sea x_n uno de los valores posibles de una variable aleatoria. La probabilidad de que Γ tome el valor x_n, se suele representar por $P(\Gamma=x_n)$.

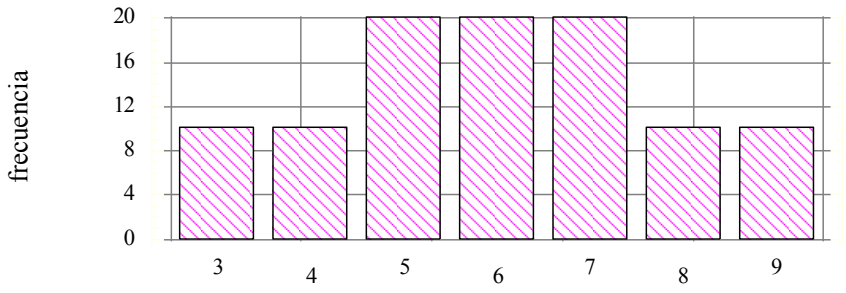
Ejemplo 4.12. Una urna contiene 5 fichas numeradas del 1 al 4. Sacamos sucesivamente dos fichas de la urna, sin reemplazamiento. Calculemos la distribución de probabilidad de la variable Γ = "suma de los números en las fichas".

Para ello consideramos el espacio muestral asociado al experimento y clasificamos sus puntos en subconjuntos distintos, de modo .que a todos los elementos de cada subconjunto les corresponde el mismo valor de la variable.

E muestral	valores de Γ	$P(x=x_i)$
12,21	3	1/10
13,31	4	1/10
14,23,32,41	5	2/10
15,24,42,51	6	2/10
25,34,43,52	7	2/10

35,53	8	1/10
45,54	9	1/10

Esta distribución de probabilidades puede representarse también gráficamente, utilizando el diagrama de barras de la figura 4.8.



Actividades

4.57. Se lanza una moneda 3 veces Representa gráficamente la distribución de probabilidad y la función de distribución de la variable aleatoria "número de caras obtenidas" ¿Cómo sería esta distribución si se considera que la moneda está sesgada y la probabilidad de obtener cara es p?

4.58. De un lote de 10 aparatos, en los que hay 3 defectuosos, se toman 2 al azar, si reemplazamiento Hallar la distribución de probabilidad de la variable aleatoria " número de defectos en la muestra" ¿Cual es la probabilidad de obtener a lo más un defecto?

4.59. Hallar la distribución de probabilidad de la variable aleatoria "número de veces que hay que lanzar un dado hasta obtener por primera vez un 6" ¿Cual es la probabilidad de que el número de lanzamientos sea par?

4.60. De una baraja española se extraen 6 cartas sin reemplazamiento Representar gráficamente la distribución de probabilidad y la función de distribución del número de ases obtenidos.

Esperanza matemática

Al estudiar las variables estadísticas, consideramos una serie de valores o características que sirven de resumen de la distribución de frecuencias. Igualmente es de interés definir las características de una variable aleatoria, como una serie de valores que resumen toda la distribución. Uno o varios de estos valores sirven, además, para especificar completamente la distribución de probabilidad y se suelen llamar

parámetros de la distribución. Uno de ellos es la media de la variable o *esperanza matemática.*

Sea Γ una variable aleatoria discreta, que toma los valores $x_1, x_2, \ldots, x_k$, con probabilidades $p_1, p_2, \ldots, p_k$. Se llama *media, esperanza matemática o valor esperado* de la variable a la suma:

$$\sum_{i=1}^{k} x_i p_i = \mu = E(\Gamma) \tag{4.17}$$

Si, en lugar de considerar Γ, estudiamos una función suya $g(\Gamma)$, obtenemos una nueva variable aleatoria. Para cualquier función $g(\Gamma)$ de la variable aleatoria se define como esperanza matemática de g la cantidad:

$$E\left[g(\Gamma)\right] = \sum_{i=1}^{k} g(x_i) p_i \tag{4.18}$$

Ejemplo 4.13. Pablo y María juegan a lanzar tres monedas. Si se obtienen 2 caras o 2 cruces, María paga a Pablo 10 euros. Si se obtienen 3 caras o tres cruces, Pablo paga 10 euros a María ¿Es un juego equitativo?

El concepto de juego justo o equitativo está estrechamente ligado con el de esperanza matemática. En un juego de este tipo, la esperanza matemática de la cantidad ganada por cada jugador ha de ser igual a cero.

Para contestar la pregunta, consideramos la variable aleatoria "dinero ganado por María", suponiendo dicha cantidad negativa si es ella la que ha de pagar la apuesta. En la tabla siguiente, hallamos la distribución de probabilidad y esperanza de esta variable.

E.Muestral	$\Gamma(x_i)$	$p(x_i)$	$x_i p(x_i)$
CCC XXX	10	2/8	20/8
CCX CXC XCC XXC XCX CXX	-10	6/8	-60/8
			-40/8=5

Del estudio de esta tabla deducimos que, si se jugase un gran número de veces el juego, en promedio, Pablo ganaría 5 euros cada jugada. No es por tanto un juego justo. Para lograr que el juego fuese equitativo, habría de pagarse a María 30 euros, cada vez que se obtuviesen 3 caras o tres

cruces. De esta forma:

$$E(\Gamma)=30x2/8-10x1/8=0$$

También podemos definir la *varianza de la variable:*

$$Var(\Gamma) = \sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 \, p_i$$

La varianza es una medida de dispersión, que toma valores positivos y es invariante por los cambios de origen. También se utiliza como medida de dispersión la desviación típica o raíz cuadrada de la varianza, que viene expresada en la misma unidad de medida de la variable.

Otras características de interés son la mediana y percentiles. Se define como percentil del *r%* aquel valor de la variable P_r que deja por debajo el r% de los posibles valores, es decir la probabilidad de obtener un valor menor que P_r es el *r %*. En particular, para *r*=50, 25 y 75 obtenemos la *mediana y los cuartiles,* que tienen la propiedad de dividir el recorrido de la variable en 4 intervalos de igual probabilidad.

Ejemplo 4.14. La tabla siguiente presenta la distribución de probabilidad de la variable aleatoria "mayor número consecutivo de caras en un lanzamiento de 4 monedas":

$\Gamma(x_i)$	$p(x_i)$	$x_i p(x_i)$	$x^2 p(x_i)$
0	1/16	0	0
1	7/16	7/16	7/16
2	5/17	10/16	20/16
3	2/16	6/16	18/16
4	1/16	4/16	16/16
			27/16

Para esta variable obtenemos las siguientes características:

$$\mu = 27/16 = 16875$$

$$Var = 61/16 - (27/16)^2$$

$$Me = 15$$

Otras características de cálculo sencillo son: recorrido = 4; Moda = 1 (valor más frecuente); Q_{25} = 1; Q_{75} =2; recorrido intercuartílico = RI = 1

Actividades

4.61. Para realizar un análisis de sangre a un grupo de r personas, con objeto de detectar una posible enfermedad, tenemos dos alternativas. La primera consiste en efectuar a cada uno una prueba. En la segunda, se mezcla la sangre de las r personas y se efectúa una prueba única. Si todos los individuos están sanos, el resultado del test es negativo, y se finaliza el análisis. Si uno al menos del grupo está enfermo, el test será positivo. En dicho caso, se hace un análisis individual a cada uno de los componentes del grupo para averiguar cual o cuales son los enfermos. Supuesto que la proporción de enfermos en la población es 0,1, describir la distribución del número de análisis necesarios para examinar a las r personas. Hallar la media de dicha variable. Usar distintos valores de r, y deducir cual es el agrupamiento que proporciona mayor economía.

4.62. Una moneda sesgada, tal que Pr(cara)=2/3, se lanza 4 veces. Hallar la media, mediana y moda del mayor número de caras consecutivas.

4.12. LA DISTRIBUCION BINOMIAL

Cuando se aplica la teoría de la probabilidad a situaciones reales, no es necesario encontrar una distribución distinta para cada modelo estudiado. A menudo nos encontramos con que muchas situaciones muestran una serie de aspectos comunes, aunque superficialmente parezcan diferentes. En este caso, podemos formular un modelo probabilístico aplicable a estas situaciones.

Desde los comienzos del Cálculo de Probabilidades hasta la fecha, se han desarrollado muchos de estos modelos, muy útiles a la hora de analizar problemas estadísticos. Generalmente son asignados a clases o familias de distribuciones, que se relacionan entre si mediante una función que incluye uno o varios parámetros, cuyos valores particulares definen la distribución de cada variable aleatoria concreta. En este capítulo estudiaremos la distribución binomial.

Consideremos un experimento aleatorio cualquiera, y en relación a él, estudiemos un suceso *A*, de probabilidad *p* y su contrario $\bar{A}$ de probabilidad *q=1-p*. Diremos que hemos tenido un éxito, si al realizar el experimento obtenemos el suceso *A*, y que hemos obtenido un fracaso en caso contrario.

Si, en lugar de realizar únicamente una vez el experimento, efectuamos una serie de repeticiones independientes del mismo, el número total de éxitos obtenido en las *n* realizaciones constituye una variable aleatoria Γ, que puede tomar los valores enteros comprendidos entre 0 y *n*.

Calcularemos la distribución y características de dicha variable aleatoria.

Ejemplo 4.15. Los tubos electrónicos producidos en una fábrica, pueden ser clasificados en correctos (suceso *A*) y defectuosos (suceso $\bar{A}$). Si estos tubos se venden en cajas de 3 elementos, el número de tubos defectuosos en cada caja puede ser 0, 1, 2 o 3. Si los tubos han sido colocados al azar en las cajas, esta variable aleatoria sigue la distribución binomial.

Supongamos que la proporción total de defectos es el 2 por ciento. El espacio muestral correspondiente al experimento que consiste en probar los tubos de una caja consecutivamente, para verificar su funcionamiento es:

E={$AAA\ \bar{A}AA\ A\bar{A}A\ AA\bar{A}\ \bar{A}\bar{A}A\ \bar{A}A\bar{A}\ A\bar{A}\bar{A}\ \bar{A}\bar{A}\bar{A}$}

Teniendo en cuenta la independencia de los ensayos, podemos calcular la distribución de la variable aleatoria

	$\Gamma=x$	$P(\Gamma=x)$
AAA	0	$0{,}98^3$
$\bar{A}$AA, A$\bar{A}$A, AA$\bar{A}$	1	$3*0{,}98^2*0{,}02$
$\bar{A}\bar{A}$A, $\bar{A}$A$\bar{A}$, A$\bar{A}\bar{A}$	2	$3*0{,}98*0{,}02^2$
$\bar{A}\bar{A}\bar{A}$	3	$0{,}02^3$

Supongamos ahora que en una repetición sucesiva de *n* ensayos independientes, hemos obtenido la sucesión $AAAAA\bar{A}\bar{A}\bar{A}$, que contiene *r* veces el suceso *A* y *n-r* veces el suceso $\bar{A}$. La probabilidad de ocurrencia de esta sucesión es $p^r q^{n-r}$. Ahora bien, todos los casos en que la variable toma el valor *r* vienen dados por las permutaciones de la anterior sucesión. Por tanto, al realizar *n* veces un experimento, la probabilidad de obtener *r* veces el suceso *A* viene dada por (4.19).

$$P(\Gamma = r) = \binom{n}{r} p^r q^{n-r} \qquad (4.19)$$

Esta es la distribución de probabilidades binomial, cuyo nombre proviene del hecho de que las probabilidades dadas en la expresión (44)

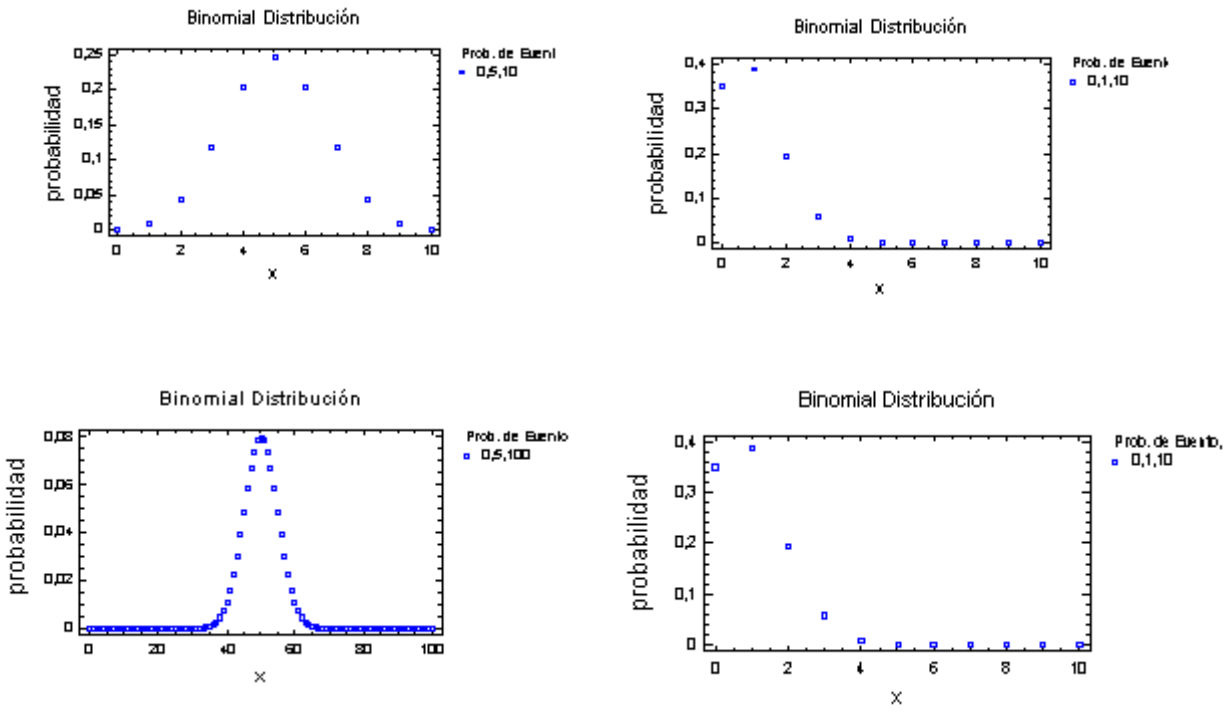
son los términos del desarrollo del binomio $(p+q)^n$. De esta expresión se deduce también que la distribución queda perfectamente determinada cuando se conocen los valores de p y n, que serán llamados parámetros de la distribución. En adelante representaremos la distribución binomial de parámetros n y p por $B(n,p)$.

Puede demostrarse que la media y varianza de dicha variable aleatoria se calculan mediante (4.20) y (4.21).

(4.20) $$\mu = np$$

(4.21) $$Var\ (\Gamma) = npq$$

Figura 4.9. Distribución binomial



En la figura 4.9 se muestra la distribución de probabilidades de la distribución binomial variando *p* y *n*. Obsérvese como cambia la forma de la distribución con el valor de los parámetros Para *p=05* o valores próximos se obtiene una distribución simétrica, que cambia a asimetría positiva o negativa, según *p* se aproxima a 0 o 1, respectivamente. Asimismo, para un mismo valor de *p*, la media y varianza de la distribución crecen con el valor de *n*.

Actividades

4.63. El 10 por ciento de una población tiene grupo sanguíneo 0 ¿Que probabilidad existe de que, al tomar 5 personas al azar, exactamente 3 sean de grupo 0?

4.64. Un autobús llega con retraso a su parada uno de cada diez dìas Si una persona toma una vez al día este autobús ¿Cual es la probabilidad de que en una semana no sufra retraso?

4.65. Un radar es capaz de detectar un blanco una de cada diez veces que efectúa un barrido de la zona Hallar la probabilidad de que el blanco no sea detectado en 4 barridas, en 10 barridas, en n barridas

4.66. Supóngase que el 85% de votantes de un distrito piensa acudir a realizar la

votación, en unas elecciones municipales Hallar la probabilidad de que en una familia compuesta por tres votantes, dos o más cumplan con esta obligación

4.67. Si el 6% de los niños en edad preescolar son disléxicos ¿Cual es la probabilidad de que entre 8 niños haya algún disléxico?

4.68. Dos jugadores A y B compiten en un torneo de ajedrez Se acuerda que el torneo conste de 6 partidas y que gane aquel que consiga mayor número de victorias Si A gana el 60% de las partidas que juega contra B ¿Cual es la probabilidad de que B sea el ganador?

4.69. Una cierta enfermedad tiene tasa de mortalidad del 10% .Al ensayar un nuevo tratamiento en un grupo de 10 pacientes, 4 de ellos fallecieron. ¿Hay evidencia suficiente para indicar que el tratamiento es inadecuado?

Ejemplo 4.16. En 10 lanzamientos de una moneda ¿Cual es la probabilidad de obtener menos de 3 caras? ¿Cual es la probabilidad de obtener 8 o más caras?

Sea Γ el número de caras al lanzar 10 monedas:

$$P(\Gamma<3)=P(\Gamma=0)+P(\Gamma=1)+P(\Gamma=2)=00547$$

Obsérvese que este resultado es la probabilidad acumulada para x=2, esto es, el valor de la función de distribución en dicho punto es,

$$P(\Gamma<3)=P(\Gamma\leq 2)=F(2)$$

$$P(\Gamma\geq 8)=1\text{-}P(\Gamma\leq 7)=1-0-9453=00547$$

Actividades

4.70. Supóngase que el 85% de votantes de un distrito piensa acudir a realizar la votación, en unas elecciones municipales. Hallar la probabilidad de que en una familia compuesta por tres votantes, dos o más cumplan con esta obligación.

4.71. Si el 6% de los niños en edad preescolar son disléxicos ¿Cuál es la probabilidad de que entre 8 niños haya algún disléxico?

4.72. Dos jugadores A y B compiten en un torneo de ajedrez Se acuerda que el torneo conste de 6 partidas y que gane aquel que consiga mayor n⌋mero de victorias Si A gana el 60% de las partidas que juega contra B ¿Cual es la probabilidad de que B sea el ganador?

4.73. Una cierta enfermedad tiene tasa de mortalidad del 10% Al ensayar un nuevo tratamiento en un grupo de 10 pacientes, 4 de ellos fallecieron ¿Hay evidencia suficiente para indicar que el tratamiento es inadecuado?

4.13. DISTRIBUCION DE POISSON

Si en la distribución binomial aumentamos indefinidamente el número de pruebas, manteniendo constante el producto $np = \lambda$, obtenemos una nueva distribución que recibe el nombre de distribución de Poisson.

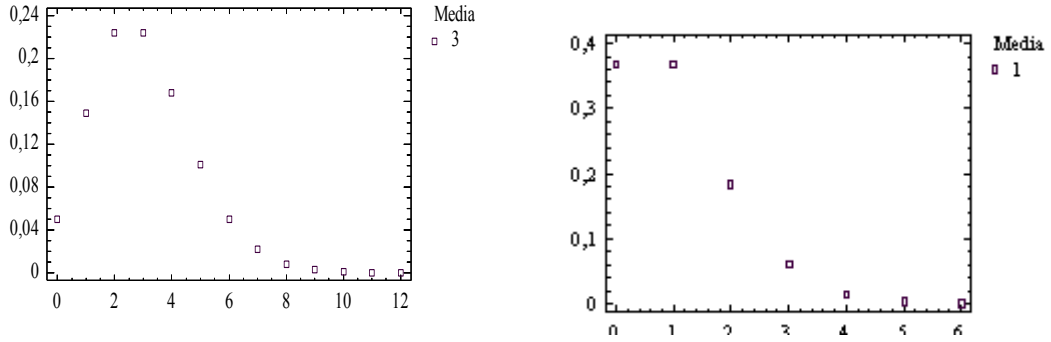
Diremos que una variable aleatoria discreta sigue la distribución de Poisson si toma los valores enteros 0, 1, 2 y su distribución de probabilidades es la dada por:

$$p(\Gamma=r)=e^{-\lambda}\,\lambda^r / r!$$

El valor λ es el único parámetro de esta distribución, que notaremos por $P(\lambda)$, y es igual a la media y varianza de la misma. En la figura 4.9 se muestra la distribución de probabilidades de la variable de Poisson *P(3.)* En la figura 4.10 aparecen la gráficas de las variables P(1) y P(3).

Nótese cómo aumentan la media y la varianza de la distribución, en función de λ

ıución de Poisson



La distribución de Poisson tiene muchas aplicaciones. Mostraremos en este texto las de mayor utilización La primera de ellas es la de aproximación de la distribución binomial, cuando *n* es grande y *p* pequeña Supongamos que manteniendo el producto $np=\lambda$ constante hacemos tender *n* a infinito:

$$\lim_{n\to\infty} P(\Gamma = r) = \lim_{n\to\infty} p^r q^{n-r} \binom{n}{r} = \lim_{n\to\infty} \left(\frac{\lambda}{n}\right)^r \left(1-\frac{\lambda}{n}\right)^{n-r} \binom{n}{r} = e^{-\lambda}\lambda^r / r!$$

Al igual que en la distribución binomial, pueden utilizarse tablas o bien programas para simplificar los cálculos.

Ejemplo 4.17. Una compañía de seguros sabe que la probabilidad de tener que indemnizar en caso de robo, cada año de la póliza es 00001. Si la compañía tiene 30000 asegurados ¿Cual es el número máximo de primas que habrá de pagar durante el año en curso, con probabilidad 099?

En este ejemplo, nos hallamos ante una distribución binomial B(30000, 00001), pues cada cliente tiene igual probabilidad de sufrir un robo, independientemente de los demás. Aproximaremos esta distribución por la de Poisson de parámetro =30000*00001=3.

Analizando la tabla de esta distribución, que aparece en la figura 4.9, podemos observar que la probabilidad de que aparezcan 8 o menos robos es 09962 y la de que aparezcan 7 o menos 09881. De estos datos se puede afirmar que 8 es la solución buscada.

Ejemplo 4.18. La tabla de mortalidad de un país indica que 1 de cada 1000 habitantes llega a centenario. Si una pequeña aldea tiene 1830 habitantes, ¿Cual es la probabilidad de que de ellos 0, 1, 2 lleguen a centenarios?

Resolveremos este segundo ejemplo con la ayuda de la tabla. En este caso, si *X* es el número de habitantes que llega a los cien años, X tiene distribución *B(1830, 0001)*. Aproximaremos esta distribución por la de Poisson de parámetro =1830*0001=183.

Este valor del parámetro no viene incluido en las tablas. Si no disponemos de un programa de cálculo, podemos utilizar un procedimiento análogo al seguido en el caso binomial. Por ello:

$$Pr(X\text{-}0)= 01496+(01653\text{-}01496)*7/10= 016059$$

El resto de las probabilidades se calcula de modo similar.

Otra aplicación de gran interés es la de recuento del numero de sucesos, cuando estos se producen a intervalos aleatorios de tiempo. Consideremos un cierto suceso aleatorio como mutación en un determinado gen, llegada de un cliente a una cola, etc que se produce a intervalos irregulares de tiempo. El número de estos sucesos ocurridos durante un intervalo de tiempo de longitud *t* es una variable aleatoria discreta. Estamos interesados en el cálculo de la probabilidad $p>k<(t)$, de

que este número sea exactamente *k*. Haremos las hipótesis siguientes:

1. Independencia El número de sucesos ocurridos en dos intervalos de tiempo que no se solapan son variables aleatorias independientes.
2. Homogeneidad en el tiempo Las probabilidades $p_k(t)$, solo dependen de *k* y de *t* , y no del instante que se toma como origen de tiempos.
3. Regularidad La probabilidad de que en un intervalo de longitud infinitesimal h se produzca un suceso es:

$$p_1(h) = \lambda h + o(h),$$

donde *o(h)* representa un infinitésimo respecto a *h*. La probabilidad de que en un intervalo *h* de longitud infinitesimal se produzcan dos o más sucesos es un infinitésimo respecto a *h*.

En estas condiciones, puede comprobarse que la variable aleatoria Γ sigue una distribución de Poisson de parámetro *t*. En realidad debe notarse que para cada valor de t obtenemos una variable aleatoria diferente. Una colección de variables aleatorias *{N(t)/t>0}* se conoce como proceso estocástico. Si, en particular, para cada *t la* variable aleatoria *N(t)* tiene distribución de Poisson, obtenemos el proceso de Poisson que tiene gran interés teórico y práctico.

Ejemplo 4.19. A una ventanilla de la oficina de Correos llega, en promedio, un cliente cada 5 minutos ¿Cual es la distribución del número de personas que acude a esta ventanilla cada hora? ¿Cual es la probabilidad de que en el próximo cuarto de hora lleguen más de 5 clientes a al ventanilla?

En este ejemplo, se cumplen, de forma aproximada, las condiciones del proceso de Poisson. Por ello, si cada 5 minutos llega una persona, es de suponer que en una hora lleguen 12. Por tanto, el número de clientes en una hora es una variable aleatoria con distribución de Poisson *P(12)*.

Análogamente, en un cuarto de hora, el número de personas sigue una distribución P(3). Utilizando de nuevo la tabla de la figura 4.9 obtenemos:

$$P(x>5) = 1 - F(5) = 1 - 09161 = 00839$$

Por último, la variable de Poisson aparece en el estudio de las

distribuciones espaciales. Cuando un cierto número de "partículas" (plantas, bacterias, glóbulos rojos, estrellas) se hallan repartidas al azar en un cierto medio (superficie de terreno, líquido, sangre, galaxia) y es λ el número medio de tales cuerpos por unidad de medio, la variable "número de partículas en u unidades de medio sigue una distribución de Poisson de parámetro λu.

Actividades

4.74. Estudiando la desintegración radioactiva, se ha comprobado que el número de partículas alfa que llegan a un cierto contador, por término medio, es de 10 partículas cada 30 segundos. Calcular la probabilidad de que en 30 segundos se obtengan menos de 4 partículas. Ídem de que en 15 segundos se obtenga alguna partícula.

4.75. Se supone que la demanda de una marca de relojes en un comercio sigue una distribución de Poisson, con media 10 unidades semanales ¿Cual es el stock que ha de tener el comerciante, a principios de semana, para tener una probabilidad de 095 de satisfacer la demanda?

4.76. Si la probabilidad de que un individuo sufra un accidente de tráfico un fin de semana es 00001, determinar la probabilidad de que se produzcan 2 o más accidentes entre un total de 5000 individuos.

4.77. Calcular la probabilidad de que entre 300 individuos tomados al azar, 4 al menos hayan nacido el día de Navidad.

4.78. En un libro de 400 páginas hay 40 erratas distribuidas al azar ¿Cual es el número de páginas libre de defectos? ¿Cual es la probabilidad de que en una página tenga más de 5 defectos?

4.79. Supongamos que un cable de acero tiene un promedio de un defecto cada 20. Si este cable se vende en rollos de 5 metros ¿Que porcentaje de rollos será defectuoso?

4.80 ¿Cuantas pasas hay que poner en un pastel de un kilo para que, al dividirlo en porciones de 50 gramos, la probabilidad de obtener una porción sin pasas sea como mucho 005?

4.14. REPRESENTACIÓN Y GENERACIÓN DE VALORES ALEATORIOS DE DISTRIBUCIONES TEÓRICAS

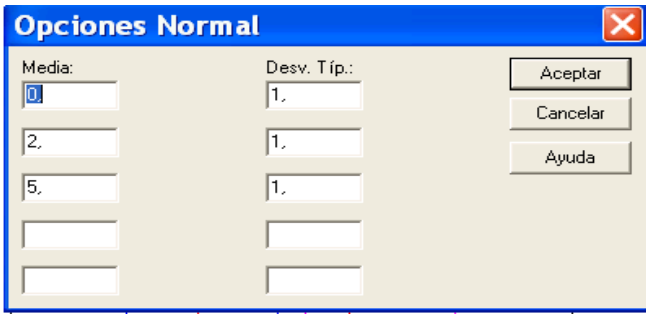
El programa Gráficos en Statgraphics representa las gráficas, realiza cálculos y genera valores aleatorios de diferentes distribuciones de probabilidad, por ejemplo, la distribución normal.

Cálculo de probabilidades teóricas

Se trata de calcular la probabilidad de que una cierta distribución teórica (por ejemplo, la normal) tome ciertos valores. Este programa hace el mismo papel que las tablas de distribuciones que aparecen en los libros de texto, con la ventaja que el programa nos da directamente los valores para una gran variedad de casos.

Al entrar al menú Gráficos – Distribuciones de probabilidad – aparece una ventana con diversos modelos de distribuciones. Si, por ejemplo, seleccionamos la distribución NORMAL, aparecerá una ventana de análisis. En ella pulsamos el botón derecho del ratón y seleccionamos Opciones de análisis Aparecerá un cuadro de diálogo como el de la figura 4.11, donde daremos los valores de la media y desviación típica correspondiente a la distribución que se está utilizando.

Figura 4.11. Diálogo para introducir la media y la desviación típica



Luego, aparecerá una ventana de análisis, en ella seleccionar el botón Opciones Tabulares, aparecerá una cuadro de diálogo, seleccionar la opción Distribución Acumulada. Al seleccionar la opción anterior aparecerá una ventana como la de la figura 4.12.

En dicha pantalla pueden observarse tres partes bien diferenciadas: Área de cola inferior (<), en la que se da la probabilidad de que se obtengan valores de la variable menores que el valor introducido Densidad de probabilidad, en la que se da la ordenada de la función de densidad, y Área de cola superior (>), en la que se da la probabilidad de obtener valores mayores que el valor dado

Figura 4.12. Ventana de resultados

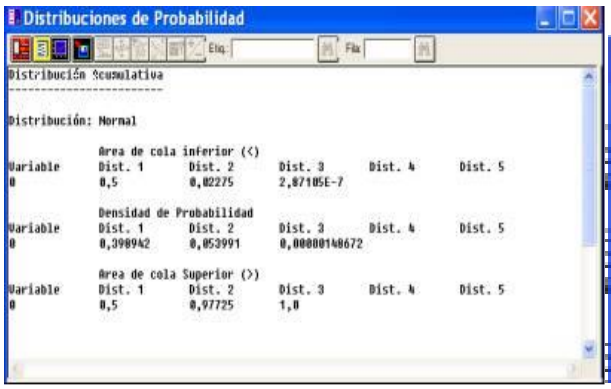
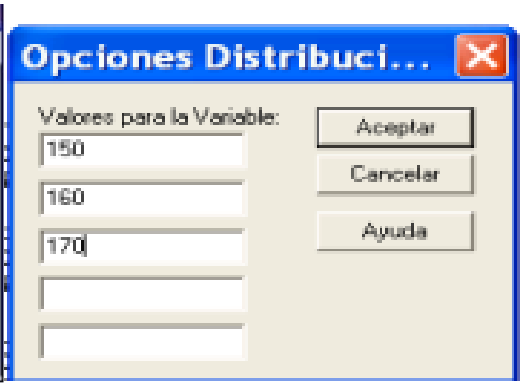


Figura 4.13. Distribución Acumulada



Sobre la ventana de la figura 4.12 haciendo clic con el botón derecho y seleccionando Opciones de ventana, aparecerá un cuadro de diálogo como el de la figura 4.13, en el que se pueden variar los valores de la variable para los cuales se desea calcular la probabilidad

Generación de números aleatorios

Para generar números aleatorios, ingresar al menú Gráficos – Distribuciones de probabilidad, allí aparecerá una ventana, en la que se deberá seleccionar el modelo de distribución con el que se desea trabajar, por ejemplo seleccionaremos la distribución discreta uniforme A continuación, aparecerá una ventana de análisis (figura 4.14), haciendo clic con el botón derecho sobre ella, se puede seleccionar Opciones de Análisis, allí aparecerá un cuadro de diálogo, en el que se ingresarán los límites inferior y superior de la variable que se desea generar

En la ventana de la figura 4.14, al seleccionar el botón Opciones tabulares, aparecerá un cuadro de diálogo, allí seleccionar Números aleatorios, aparecerá otra ventana de análisis, hacer clic con el botón derecho y entrar a Opciones de Ventana para definir el tamaño de la muestra que se desea generar; por defecto aparece el tamaño 100

Figura 4.14. Ventana de análisis

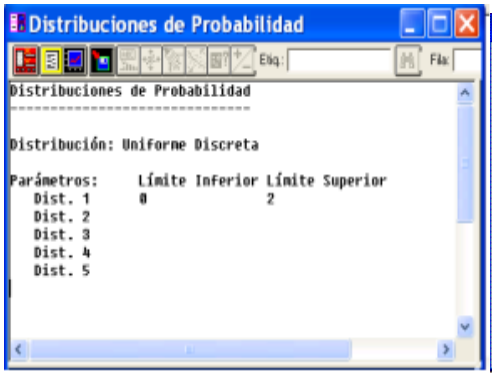
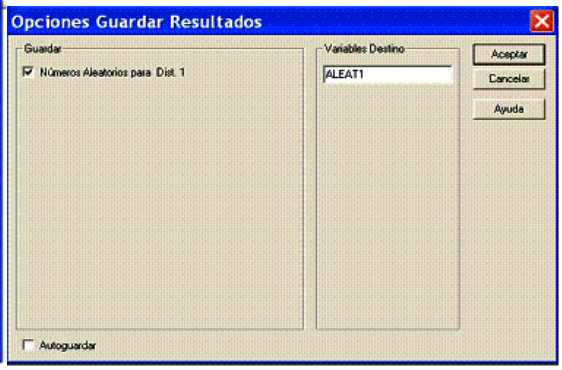


Figura 4.15. Grabación de la variable generada



Una vez que se generan los números aleatorios, se deben grabar, para ello entrar en el botón Guardar resultados (cuarto icono de la ventana de análisis), aparecerá una ventana como la de la figura 4.15, en la que se debe ingresar el nombre de la variable (en la figura aparece como ALEAT1) y seleccionar el campo Números aleatorios para Dist 1, luego hacer clic en el botón Aceptar. De esta manera se generará una nueva variable en la hoja de cálculo. Este procedimiento puede repetirse todas las veces que se desee para todas las variables que se quieran generar y también, pueden colocarse otros nombres distintos a los que aparecen por defecto en el cuadro Variables Destino.

TEMA 5
VARIABLE ALEATORIA CONTINUA

5.1. FUNCION DE DENSIDAD DE UNA VARIABLE ALEATORIA

En el tema 4 hemos estudiado el concepto de *variable aleatoria*, que se refiere a la variable estudiada en toda la población, mientras que la *variable estadística* se refiere a la misma variable estudiada sólo en la muestra.

La variable aleatoria se origina en un *experimento aleatorio*. Este experimento consiste en imaginar qué ocurriría si ampliáramos la muestra hasta tomar los datos de toda la población. Normalmente no es posible analizar toda la población, pero podemos pensar en un experimento teórico y preguntarnos por la *probabilidad* con que aparecen los diferentes valores en la población. Por ejemplo, en la actividad 5.1 nos podría interesar calcular la probabilidad de que una alumna, elegida al azar de la población, tenga una altura dada.

Si al realizar un experimento aleatorio y representar sus resultados mediante una variable, los valores que ésta puede tomar no son aislados, sino que pertenecen a un intervalo, diremos que dicha variable, es *continua*. Como ejemplos podemos citar cualquier experimento en el que se mida una magnitud continua un número ilimitado de veces, o en un colectivo de individuos muy grande, como el peso o talla de personas.

Actividad

5.1. La Tabla de frecuencias 5.1 ha sido obtenida con STATGRAPHICS a partir de los datos sobre altura de una muestra de 1000 chicas de edades comprendidas entre 15 y 20 años ¿Qué puedes deducir, sobre la forma del histograma y polígono de frecuencias de esta distribución? ¿En qué intervalo se encontrarían la moda y mediana? ¿Cuál sería su valor aproximado? ¿Podrías estimar la probabilidad de que una chica elegida al azar de la población de chicas de donde se ha tomado esta muestra tenga una altura entre 160 y 170? ¿Y que mida más de 174 cm?

Tabla 5.1. Tabla de frecuencias para altura

Clase	Lim, inferior	Lim. Superior	Puntomedio	Frecuencia absoluta	Frecuencia relativa	Acum..	Acum.. relativa
1	146,0	148,0	147,0	1	0,0010	1	0,0010
2	148,0	150,0	149,0	0	0,0000	1	0,0010
3	150,0	152,0	151,0	10	0,0100	11	0,0110
4	152,0	154,0	153,0	14	0,0140	25	0,0250
5	154,0	156,0	155,0	23	0,0230	48	0,0480
6	156,0	158,0	157,0	65	0,0650	113	0,1130
7	158,0	160,0	159,0	70	0,0700	183	0,1830
8	160,0	162,0	161,0	132	0,1320	315	0,3150
9	162,0	164,0	163,0	158	0,1580	473	0,4730
10	164,0	166,0	165,0	165	0,1650	638	0,6380
11	166,0	168,0	167,0	143	0,1430	781	0,7810
12	168,0	170,0	169,0	99	0,0990	880	0,8800
13	170,0	172,0	171,0	71	0,0710	951	0,9510
14	172,0	174,0	173,0	27	0,0270	978	0,9780
15	174,0	176,0	175,0	19	0,0190	997	0,9970
17	176,0	178,0	177,0	3	0,0030	1000	1,0000

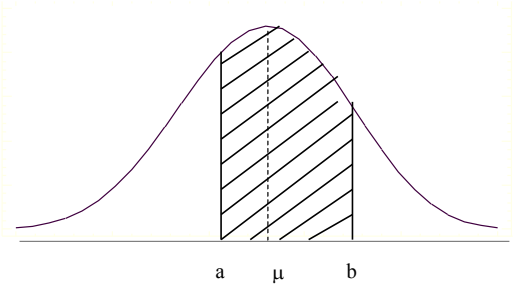
Media = 164,721 Desviación típica 4,92274 Varianza 24,2334
Asimetría= -0,165955 Curtosis = -0,0385743

En este tipo de variables, estamos interesados en calcular, no sólo la probabilidad de que tome un valor determinado $P(\xi=b)$, sino probabilidades como, por ejemplo, $P(a<\xi)$, $P(\xi>b)$, $P(a<\xi<b)$, etc. Por ejemplo podemos interesarnos por la probabilidad de que el peso de un recién nacido sea inferior a 2 kg, o que el índice de precios esta año supere lo marcado por el gobierno. Para resolver este tipo problema se asocia a cada variable aleatoria continua una función real, $f(x)$ definida en el conjunto de números reales llamada *función de densidad* de la variable aleatoria, tal que, para todo par de números a y b, se verifica:

$$P(a \le \xi \le b) = \int_a^b f(x)dx$$

Si observamos la figura 5.1, de la definición anterior se deduce que la probabilidad de que una variable aleatoria continua tome sus valores en un intervalo (a,b) viene dada por el área comprendida entre la función de densidad, el eje y los extremos a y b. Esta propiedad se corresponde con otra del histograma de frecuencias relativas en una variable estadística continua. En efecto, por construcción del mismo, la frecuencia relativa de valores de la variable en el intervalo (a,b) viene dada por el área comprendida entre el histograma, el eje X, y los extremos a y b.

Figu



Si subdividimos los intervalos de clase sucesivamente y construimos los polígonos de frecuencias correspondientes, obtenemos una colección de poligonales progresivamente más ajustadas entre sí, aproximándose a una curva que es la función de densidad. Podemos entonces definir la función de densidad como la curva límite a la que tiende el polígono de frecuencias relativas cuando aumentamos indefinidamente el número de pruebas, y la amplitud de los intervalos de clase tiende a cero.

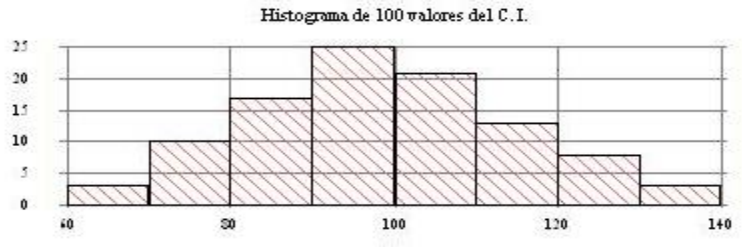
Ejemplo 5.1. El coeficiente intelectual de las personas (que denominaremos CI) es una variable teórica que mide la capacidad lógica y se obtiene a partir de ciertos cuestionarios que han sido validados y probados con un gran número de personas. En estos cuestionarios, una puntuación 100 corresponde al promedio, y una puntuación superior o inferior a 100 indica más o menos capacidad intelectual que el promedio de su edad. En la tabla 5.2 se muestra la puntuación obtenida por 100 personas seleccionadas aleatoriamente en el cuestionario que mide el C: I. y en la figura 5.2 el histograma de frecuencias correspondiente.

Tabla 5.2. Coeficientes intelectuales de 100 personas

Clase	Lim. inferior	Lim. Superior	Punto medio	Frecuencia absoluta	Frecuencia relativa	Acum..	Acum.. relativa
1	60,0	70,0	65,0	3	0,0300	3	0,0300
2	70,0	80,0	75,0	10	0,1000	13	0,1300
3	80,0	90,0	85,0	17	0,1700	30	0,3000
4	90,0	100,0	95,0	25	0,2500	55	0,5500
5	100,0	110,0	105,0	21	0,2100	76	0,7600
6	110,0	120,0	115,0	13	0,1300	89	0,8900
7	120,0	130,0	125,0	8	0,0800	97	0,9700
8	130,0	140,0	135,0	3	0,0300	100	1,0000

Media = 98,9919 Desviación Típica = 16,2713

Figura 5.2. Distribución del CI de 100 personas



El histograma es unimodal (una sola moda), y la moda se sitúa, aproximadamente, en el centro de la distribución. El mayor número de casos se concentra en el intervalo 90-100 y a ambos lados la distribución decrece rápidamente, aunque es todavía algo asimétrica. Al aumentar a la vez la muestra y el número de intervalos (Figuras 5.3 y 5.4) el histograma se aproximan a una curva continua que llamaremos *curva de densidad*. La función matemática correspondiente a dicha curva se llama **función de densidad.**

Figura 5.3. Distribución del CI de 1000 personas

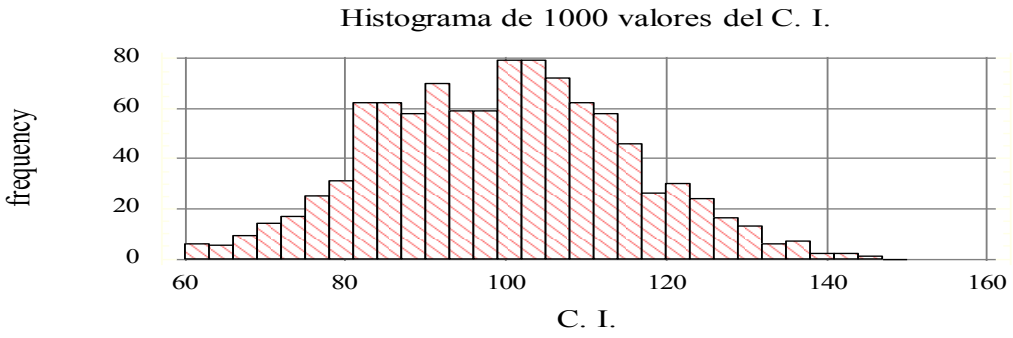
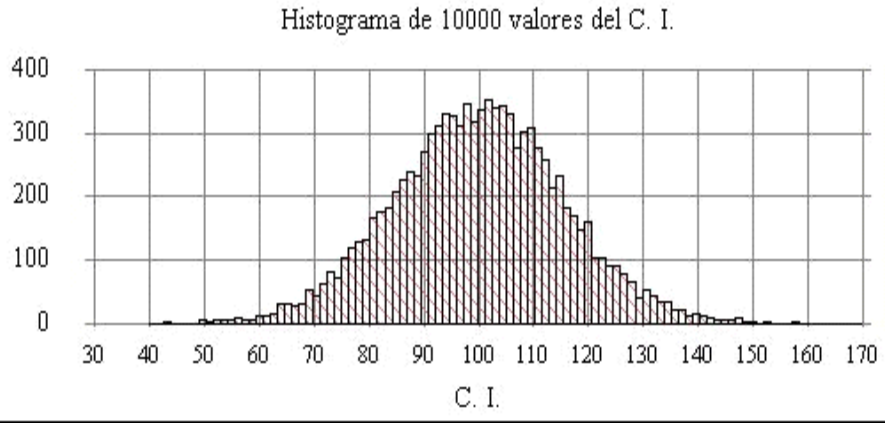


Figura 5.4. Distribución de C.I.



En la figura 5.3 sigue habiendo una sola moda, situada en el centro de la distribución, que empieza a tomar una forma característica, más cercana a una curva en forma de campana invertida. Esta forma se percibe más

claramente si continuamos el proceso de aumentar el tamaño de muestra y, a la vez el número de intervalos, como se puede apreciar en al figura 5.4 que corresponde a 10.000 puntuaciones del C. I.

Una función de densidad debe ser siempre positiva, lo cual implica que la gráfica de la función de densidad esté por encima del eje horizontal. Esto es debido a que la probabilidad es siembre igual o mayor que cero Mediante la función de densidad podemos calcular probabilidades de diverso tipo, como se muestra en los siguientes ejemplos.

- El área total bajo la curva y por encima del eje horizontal es igual a 1, al ser la suma de todas las áreas corresponde a la suma de todas las probabilidades, en consecuencia, dicha suma (integral) es 1 y lo expresamos en la forma siguiente:

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

- La probabilidad de que una variable continua tome un valor aislado es cero. Este tipo de suceso, que puede ocurrir y tiene probabilidad nula se llama "suceso casi imposible"

$$P(\xi = a) = P(a \le \xi \le a) = \int_{a}^{a} f(x)dx = 0$$

- Con un razonamiento análogo, deducimos que la probabilidad de que una variable aleatoria sea diferente de un valor dado es igual a uno. Este tipo de suceso, que no siempre se verifica, y sin embargo tiene probabilidad uno, se llama suceso "casi seguro".
- $P(a \le \xi \le b) = P(a < \xi < b)$, puesto que la probabilidad de un punto es cero

En los apartados anteriores, hemos supuesto que la variable toma valores en todo el eje real, por lo que los límites de integración son infinitos. En el caso en que la variable sea acotada, se utilizará los límites convenientes.

Actividad

5.2. En un hospital se comprobó que el peso de nacimiento de las niñas era una variable aleatoria que tomaba valores entre 2 y 4 kilos, siendo la función de densidad:

$$f(x)=\begin{cases} x/6 & \text{para } 2<x<4 \\ 0 & \text{fuera del intervalo} \end{cases}$$

¿Cual será la proporción de niñas con peso superior a 3 kilos?

5.9. FUNCION DE DISTRIBUCION

Al igual que en el caso de variables discretas, puede asociarse a cada variable aleatoria continua ξ una función real definida por la expresión (5.1).

(5.1) $$F(x) = P(\xi<x)$$

Para este tipo de variables la función de distribución verifica, para todo x la igualdad (5.2).

(5.2) $$F(x)=\int_{-\infty}^{x} f(x)dx$$

De esta igualdad, y debido a las propiedades de la integral definida, se verifican fácilmente las siguientes propiedades de la función de distribución:

a. $F(x)$ es una función monótona no decreciente:

b. $\lim_{x\to-\infty} F(x)=0$; $\lim_{x\to+\infty} F(x)=1$;

c. $P(a\leq b) = F(b) - F(a)$

También puede comprobarse sin dificultad que $F(x)$ es continua a la derecha de cada punto. En los puntos en que F es derivable, su derivada es igual a la función de densidad.

Ejemplo 5.2. Supongamos que una variable aleatoria toma valores en el intervalo (0,1) y tiene la siguiente función de densidad:

$$f(x) = \begin{cases} 1, \text{ si } 0\leq x\leq 1/2 \\ 2, \text{ si } 3/4\leq x\leq 1 \\ 0, \text{ para cualquier otro valor} \end{cases}$$

En este caso, la función $F(x)$ viene dada por,

$$F(x)= \begin{cases} 0, \text{ si } x\leq 0 \\ x, \text{ si } 0\leq x\leq 3/4 \\ 1/2, \text{ si } 1/2\leq x\leq 3/4 \\ 2x, \text{ si } 3/4\leq x\leq 1 \end{cases}$$

Puede observarse que *F* es derivable excepto en los puntos *x*=0, *1/2*, *3/4* y 1. En los casos en que es derivable su derivada es igual a *f(x).*

Actividades

5.3. Suponiendo que el tiempo de espera del metro es una variable aleatoria continua que tiene por función de distribución:

$$F(x)= \begin{cases} 0 & si\ x\leq 0 \\ x/2 & si\ 0\leq x\leq 1 \\ 1/2 & si\ 1(x(2 \\ x/4 & si\ 2(x(4 \\ 1 & si\ x>4 \end{cases}$$

Se pide: a) Dibujar la función de distribución; b) Hallar la función de densidad y dibujarla; c) Hallar la probabilidad de que el tiempo de espera sea menor de 3 minuto; d) Si una persona ha esperado ya 1 minuto, hallar la probabilidad de que el tiempo de espera sea menor que 3 minutos.

5.4. Sea $f(x) = 4x^3$, cuando $0<x<1$ la función de densidad de una variable aleatoria . Hallar un valor *a* tal que ξ tenga igual probabilidad de ser menor o de ser mayor que *a*. Calcular *b* tal $P(\xi>b) =0,05$

5.5. Sea ξ una variable aleatoria con función de densidad definida del siguiente modo

$$f(x) \begin{cases} = ae^{-ax}, \text{ para } x\geq 0 \\ =0 \text{ , para } x<0, \end{cases}$$

Determinar la función de distribución y calcular $P(\xi<100)$.

5.10. CARACTERISTICAS DE UNA VARIABLE ALEATORIA CONTINUA

Se define la esperanza matemática, valor esperado o media de una variable aleatoria continua mediante la expresión (5.3).

$$E[\xi] = \int_{-\infty}^{\infty} xf(x)dx \qquad (5.3)$$

Si la variable toma valores en un intervalo acotado *(a,b),* los límites de integración se extienden únicamente a este intervalo. Esta definición viene motivada por consideraciones de paso al límite en la definición de media de una variable estadística continua. En efecto, en este tipo de

variable $\bar{x} = \sum_{i=1}^{k} x_i h_i$, siendo x_i, las marcas de clase en el histograma de frecuencias y h_i la frecuencia relativa en el intervalo *i*. Ahora bien, en el histograma de frecuencias relativas cuando aumentamos el número de observaciones y hacemos tender la amplitud a cero esta expresión converge a la (5.3), pues el histograma de frecuencias converge a la función de densidad. Si en lugar de considerar la variable ξ tomamos una función de la misma, obtenemos (5.4).

(5.4) $$E[g(\xi)] = \int_{-\infty}^{\infty} g(x)\, f(x)dx$$

que, para distintas funciones g da lugar a la definición de los momentos de la variable aleatoria y se definen las características de la variable continua de la misma forma que se ha hecho para la discreta, con excepción de la moda. Esta se define como el valor de la variable en que la función de densidad presenta un máximo.

Actividades

5.6. Hallar el tiempo medio de espera en la actividad 5.3. Hallar la media y la varianza de la distribución de la actividad 5.5.

5.2. LA DISTRIBUCIÓN NORMAL

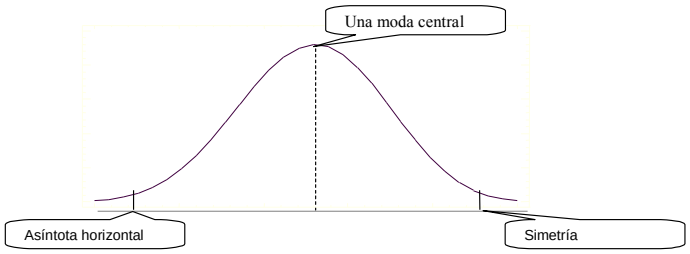
En el tema 4 hemos estudiado la *distribución binomial,* que es un modelo teórico que sirve para estudiar algunas variables discretas como número de votantes que votan a un cierto partido o número de averías en una caja de repuestos. Otro de los modelos estadísticos teóricos más importantes es la *distribución normal,* debido a su utilidad para describir variables que surgen en problemas reales en distintos campos, como, por ejemplo:

- Problemas biológicos: distribución de la tallas, pesos y otras medidas físicas de un conjunto numeroso de personas de una determinada edad;
- Datos psicológicos: coeficiente de inteligencia, tiempo de reacción, puntuaciones en un examen o test, amplitud de percepción;
- Problemas físicos: distribución de los errores de observación o medida que aparecen en los estudios acerca de fenómenos meteorológicos, físicos, astronómicos, etc. ;

- Datos económicos: distribución de las fluctuaciones de los índices de precio o de las cotizaciones en bolsa de un cierto valor alrededor de la línea de tendencia;

La distribución normal teórica tiene una forma muy característica. Su gráfica es simétrica respecto al centro de la distribución, que es donde se concentran la mayor parte de los valores y tiene la forma de una campana invertida (ver figura 5.5).

Figura 5.5. Función de densidad en la Distribución Normal



A partir del valor central la distribución de valores decrece suavemente hacia los extremos hasta que la gráfica se aproxima al eje horizontal. En estos casos, la curva normal puede ser un modelo adecuado.

Para que una variable aleatoria siga la distribución normal ha de ser cuantitativa y continua, por lo que teóricamente puede tomar todos los valores dentro de un intervalo dado (potencialmente infinito). En la práctica, podemos también considerar el caso de variables discretas con un número muy grande de valores. La función de densidad normal está definida por la siguientes expresión:

$$N(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)/\sigma]^2} \quad , \qquad -\infty < x < \infty \ ,$$

Actividades

5.7. Representa, aproximadamente, la función de densidad que correspondería a las alturas de las 1000 chicas, dadas en la Tabla 1, y compárala con la gráfica usual en las distribuciones normales. ¿Piensas que se obtendría una buena aproximación al representar los datos mediante una distribución normal? ¿Cuáles serían la media y desviación típica de dicha distribución normal teórica?

5.3. PROPIEDADES DE LA DISTRIBUCIÓN NORMAL

Simetría

La función de densidad normal es simétrica, respecto a su media, debido a que en su fórmula aparece una exponencial al cuadrado. Algunas propiedades derivadas de la simetría son las siguientes:

- Las dos áreas que se forman al dividir la gráfica por el eje de simetría (área superior e inferior), son iguales y cada una de ellas representa el 50 % de casos en el conjunto de datos.
- Puesto que la media, mediana y moda, en las distribuciones simétricas coinciden en un mismo punto, por lo tanto son iguales en las distribuciones normales.
- La moda, que es el punto sobre el eje horizontal donde la curva tiene su máximo, en la distribución normal coincide con la media. Por tanto los valores cercanos a la media son los que alcanzan la máxima probabilidad.

Actividades

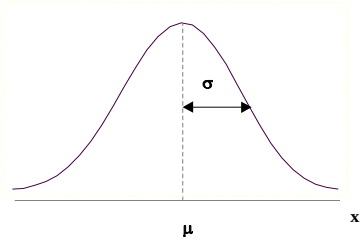
5.8. Supongamos que hacemos un estudio estadístico sobre los alumnos de la clase. Describir ejemplos de variables cuya distribución pudiera aproximarse bien mediante la distribución normal y otras para las que no sea adecuada dicha distribución.

Propiedades relacionadas con la media y la desviación típica

Si seguimos la curva desde el centro μ hacia ambos extremos, podremos observar que la curva cambia de sentido, de cóncava a convexa. El punto en donde se produce este cambio de sentido está localizado a una distancia σ a cada lado de la media.

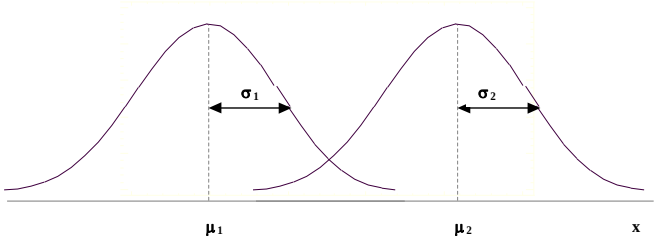
Figura 5.6. Significado de los parámetros en la distribución normal

μ es la media y σ es la desviación típica.

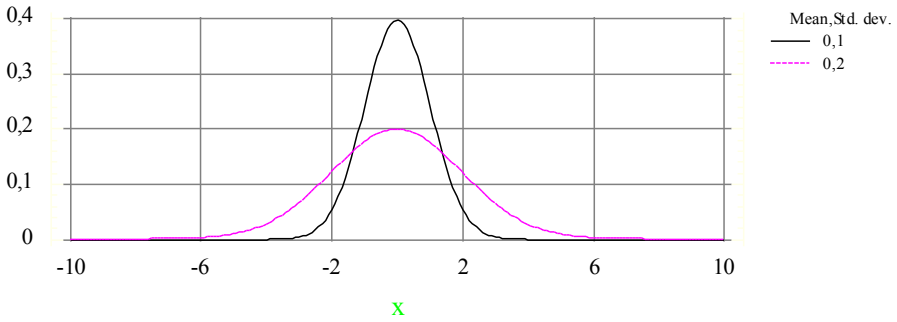


Dos curvas normales con la misma desviación típica pero diferentes medias, son idénticas pero se centran en diferentes posiciones a lo largo del eje horizontal (figura 5.7)

Figura 5.7. Distribuciones normales con igual desviación típica



Si las curvas tienen igual media y distintas desviaciones típicas, la curva con desviación típica mayor es más baja y más extendida, porque los datos están más dispersos, pero ambas curvas tienen su centro sobre el mismo valor en el eje horizontal (Figura 5.8).



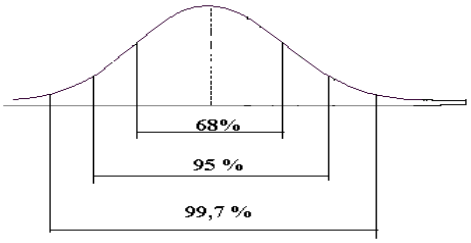
Distribución de casos en relación con la desviación típica

Una característica importante de las distribuciones normales es que la proporción de casos que se encuentran en el intervalo $(\mu - k\sigma, \mu + k\sigma)$ es siempre constante. Esta propiedad sirve para determinar entre qué valores podemos situar un porcentaje dado de casos centrales. También se utiliza para identificar la posición de un valor determinado con respecto a la media, o para saber si un determinado valor o intervalo es representativo o

no de la distribución (figura 5.9):

- El 68 % de las observaciones están a una distancia de la media μ. igual o menor que la desviación típica σ:
- El 95 % de los datos están a una distancia igual o menor que 2σ de la media μ.
- El 99,7 % de los datos están a una distancia igual 3σ de la media μ.

Figura 5.9. Distribución de casos en la curva normal



Actividades

5.9. Las puntuaciones en un test de inteligencia de un grupo de alumnos siguen una distribución normal con media 110 y desviación típica 25. ¿Qué proporción de alumnos puntúa por encima de 110? Obtener los valores de las puntuaciones tales que el 95% central de los casos esté comprendido entre dichos valores.

5.10. La temperatura media en Noviembre en Segovia sigue una distribución normal con 8 grados de media y 3 grados de desviación típica. ¿Cuál es la probabilidad de que la temperatura esté un día comprendida entre 5 y 11 grados? ¿Y entre 2 y 5 grados? ¿Cuál es la probabilidad de que la temperatura sea menor que 2 grados?

5.11. Dada una distribución de puntuaciones de un test que sigue la distribución normal de probabilidades, con media $\mu=12$ y $\sigma=4$, ¿Qué porcentaje de casos cae entre 8 y 16?

Ejemplo 5.3. El peso de los chicos de 18 a 24 años de edad es aproximadamente normal con media $\mu = 64{,}5$ kilos y una desviación típica $\sigma = 2{,}5$ kilos. Para calcular el intervalo que cubre el 95 % de los valores centrales debemos realizar las siguientes operaciones:

$$\mu - 2\sigma = 64{,}5 - 5 = 59{,}5 \text{ kilos}$$

$$\mu + 2\sigma = 64{,}5 + 5 = 69{,}5 \text{ kilos}$$

En conclusión, el 95 % central de los chicos está comprendido aproximadamente entre 59,5 y 69,5 kilos de peso. El otro 5 % de chicos tienen pesos que están fuera del intervalo (59,5 – 69,5). Pero como la distribución normal es simétrica, la mitad de este 5% de chicos se encontrará en cada una de las colas inferior y superior de la distribución. Por lo tanto el 2,5 % de los chicos tienen pesos menores que 59,5 kilos y el 2,5 % tiene pesos mayores que 69,5 kilos.

Ejemplo 5.4.

1) *¿Cuál es aproximadamente la proporción de personas que poseen una medida de CI menor que 100?* Puesto que la media es 100 y la distribución es simétrica, ***aproximadamente*** la mitad de las medidas de CI están a cada lado de la media 100, por lo tanto, la proporción de personas con un CI menor que 100 es igual al 50%.
2) *¿Cuál es el intervalo que contiene a ese 95 % central de valores para la distribución del CI?* Hemos visto que el 95% de casos centrales está a una distancia 2σ de la media μ. El intervalos es, por tanto (70, 130).
3) *Una persona con una medida de CI que excede los 130 puntos es considerada superdotada. ¿Cuál es la probabilidad de que una persona elegida en forma aleatoria esté dentro de esta categoría?* Puesto que fuera del intervalo anterior queda un 5% de casos repartido a ambos lados, la probabilidad pedida es 2,5 % .

5. 5. EVALUACIÓN DE LA NORMALIDAD DE UNA DISTRIBUCIÓN

Para decidir si podemos describir una distribución de datos dada por una curva normal, debemos comprobar si en nuestros datos se cumplen las propiedades de las distribuciones normales.

- En primer lugar, comprobaremos que nuestra variable es numérica, pues la distribución normal se refiere a variables numéricas y no a variables cualitativas. Será también necesario que la variable sea continua o que si es discreta, el número de valores distintos sea numeroso y la forma del histograma se aproxime a la distribución normal.
- Conviene representar los datos gráficamente y comparar con la función de densidad normal. Un histograma, un diagrama de tallos y hojas o un gráfico de caja, pueden revelar aspectos no normales de una distribución, tales como asimetría pronunciada, intervalos vacíos, o demasiados

valores atípicos.

- Se puede usar estos gráficos para evaluar si una distribución es o no normal, marcando los puntos $\bar{x}$, $\bar{x} \pm s$, y $\bar{x} \pm 2s$, sobre el eje x. Luego se compara la frecuencia de observaciones en cada intervalo con la regla 68 – 95 – 99,7 que hemos estudiado para las distribuciones normales.

Actividades

5.12. Dada una distribución de puntuaciones $N(16,4)$ ¿qué límites incluyen el 68 por ciento central de los casos? ¿Si queremos aprobar el 95 por ciento de los alumnos, a partir de qué nota debe considerarse aprobado?

5.13. Las puntuaciones obtenidas por 300 niños de un colegio de EGB al aplicarles un test de aritmética siguen una distribución normal de media 24 y desviación típica 4.¿Cuál es la probabilidad de obtener puntuación igual o inferior a 16? b) ¿Cuántos niños de dicho colegio tienen igual o mayor puntuación que 28?

5.14. Los errores aleatorios de una cierta medición obedecen a una ley normal con una desviación típica de un 1 mm y esperanza matemática 0. Hallar la probabilidad de que de dos observaciones independientes el error por lo menos en una de ellas no supere el valor absoluto de 1 mm.

Evaluación de la normalidad de una distribución por medio de STATGRAPHICS

Ejemplo 5.5. Usaremos como ejemplo un conjunto de datos sobre el CI de 1000 personas (figura 5.3). Se trata de una variable numérica que toma un número suficientemente grande de valores (de 40 a 150).

Simetría y unimodalidad. Tenemos varias formas para comprobarla; en primar lugar por la forma aproximada del histograma, además de la existencia de una sola moda y sin valores atípicos. Podemos comparar la posición relativa de media, mediana y moda, que en una distribución simétrica deben coincidir, como ocurre en este ejemplo. Podríamos estudiar el coeficiente de asimetría y asimetría tipificado. Para que la distribución sea simétrica, el coeficiente de asimetría debe ser próximo a cero y el coeficiente de asimetría tipificado debe estar comprendido en el intervalo (-2,2).

Curtosis. Para que la distribución sea normal, el coeficiente de curtosis debe ser próximo a cero y el de curtosis tipificado estar comprendido en el intervalo (-2, 2).

Porcentajes de casos alrededor de la media. Se puede estudiar el porcentaje de casos que se distribuye en los *($\bar{x}$–s; $\bar{x}$+s); ($\bar{x}$–s; $\bar{x}$+2 s); ($\bar{x}$–3s; $\bar{x}$+3s)*, siendo $\bar{x}$ la media de la muestra y s la desviación típica de la muestra y compararlos con los que esperamos en una distribución normal (68, 95 y 99,7). En el ejemplo 5.3, la proporción correspondiente al intervalo (84; 114) es 67,78%; al intervalo *(x–2s; $\bar{x}$+2s)* corresponde 96,6% y al *(x–3s; $\bar{x}$+3s)* = (55,17; 144,65) ≈ (55; 145) le corresponde el 99,3 %.; por tanto los datos son aproximadamente normales.

5.6. AJUSTE DE UNA DISTRIBUCIÓN NORMAL TEÓRICA A LOS DATOS OBTENIDOS PARA UNA VARIABLE DADA

Una vez que decidimos que la distribución normal podría ser un buen modelo para aproximar los datos *(distribución observada)*, el siguiente paso es elegir la distribución normal que mejor aproxima nuestros datos (*distribución normal teórica).* Tomaremos, por tanto la distribución normal que tiene como media y desviación típica las que hemos observado en la muestra. Cuando trabajamos con Statgraphics podemos hacer el ajuste de una distribución normal a una variable dada de dos formas:

- **Gráficamente,** utilizando: DESCRIPCIÓN – DISTRIBUCIONES – AJUSTE DE DISTRIBUCIONES – eligiendo en el botón de opciones gráficas (OPCIONES GRÁFICAS) – HISTOGRAMA DE FRECUENCIAS. Este programa dibuja una curva normal superpuesta al histograma de frecuencias, eligiendo la media y desviación típica adecuada.
- **Analíticamente** por medio de: DESCRIPCIÓN – DISTRIBUCIONES – AJUSTE DE DISTRIBUCINES – eligiendo el botón OPCIONES TABULARES – ÁREAS DE COLA. Esta opción permite calcular el área bajo la curva para los datos menores o iguales que un determinado valor.

Ejemplo 5.6. (Continuación). *¿Cuál es la probabilidad de que una persona escogida al azar tenga un coeficiente intelectual en el intervalo (70; 130)? ¿Cuál es la probabilidad de que se obtenga más de 140 puntos?*

Primeramente calculamos la media y desviación típica de la curva teórica que mejor se ajusta a los datos con la opción: DESCRIPCIÓN – DISTRIBUCIONES – AJUSTE DE DISTRIBUCIONES, de la que

obtenemos los siguientes datos:

Resumen del análisis

Datos: COEF_INT

1000 valores comprendidos desde 41,0 hasta 146, 0

Distribución normal ajustada

Media = 99,0551

Desviación Típica = 15,3527

Estos son precisamente la media y desviación típica en la muestra. Por medio del programa AREA de COLAS obtenemos los siguientes resultados:

Áreas de cola para la variable COEF_INT

Area por debajo 70,0 = 0,0224567

Area por debajo 130,0=0,978177

Por lo tanto el área comprendida entre 70 y 130 es: 0,978188 – 0,0224567 = 0,9557313 y la probabilidad de que una persona tenga un CI comprendido entre 70 y 130 es aproximadamente 95,56 %. Para resolver la segunda pregunta obtenemos con el programa:

Áreas de cola para COEF_INT

Área por debajo 140,0 = 0,996408

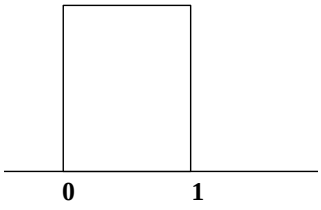
En consecuencia, el área que estamos buscando es 1 – 0,996 = 0,004, y la probabilidad de que una persona tenga un coeficiente mayor que 140 es: 0,4 %. En este ejemplo, el porcentaje de observaciones en los intervalos centrales y los valores de los coeficientes de simetría y de curtosis, nos indican que podemos aproximar bastante bien la distribución real de datos por medio de la distribución normal asociada a ella. Esto no ocurre con todas las variables

---

Actividades

5.15 La figura 5.14 muestra la curva de densidad de una distribución uniforme. La curva toma el valor constante 1 sobre el intervalo (0,1) y toma el valor 0 fuera de dicho intervalo. Esto significa que los datos descriptos por la distribución toman valores que se extienden uniformemente entre 0 y 1.

Figura 5.14



Utilice las áreas bajo esta curva de densidad para responder a las siguientes cuestiones:

a) ¿Qué porcentaje de las observaciones cae por encima de 0,8?

b) ¿Qué porcentaje de las observaciones cae por debajo de 0,6?

c) ¿Qué porcentaje de las observaciones cae entre 0,25 y 0,75?

5.16 La distribución de las alturas de hombres adultos es aproximadamente normal con una media de 69 pulgadas y una desviación típica de 2,5 pulgadas.

a. Traza una curva normal y sobre ella localiza la media y la desviación típica.

b. Usa la regla 68 – 95 – 99,7 para responder a las siguientes cuestiones: ¿Qué porcentaje de hombres tienen una altura mayor que 74 pulgadas?

c. ¿Entre qué alturas está comprendido el 95 % central de los hombres?

d. ¿Qué porcentaje de hombres tienen una altura menor a 66,5 pulgadas?

5.17 Las puntuaciones de un test es aproximadamente normal con $\mu = 110$ y $\sigma = 25$. Utilizando la regla 68 – 95 – 99,7 responde a las siguientes cuestiones:

a. ¿Qué porcentaje de personas tiene puntuaciones por encima de 110?

b. ¿Qué porcentaje de personas tiene puntuaciones por encima de 160?

c. ¿Cuál es el intervalo que abarca el 95 % central de los puntuaciones de CI?

5.18 Las medidas repetidas de la misma cantidad física generalmente tienen una distribución aproximadamente normal. A continuación se reproducen 29 medidas hechas por Cavendish de la densidad de la Tierra, realizadas en 1798 (Los datos dan la densidad de la Tierra como un múltiplo de la densidad del agua).

5,50	5,61	4,88	5,07	5,26	5,55	5,36	5,29	5,58	5,65
5,57	5,53	5,62	5,29	5,44	5,34	5,79	5,10	5,27	5,39
5,42	5,47	5,63	5,34	5,46	5,30	5,75	5,68	5,85	

Representa estos datos mediante un histograma y observa si una distribución normal puede ajustarse a estos datos. Luego comprueba tus conclusiones por medio de la regla 68 – 95 – 99,7. Para ello, calcula x y s, luego realiza el conteo del número de observaciones que caen dentro de los intervalos $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$. Compara los porcentajes de cada intervalo con los de la regla antes mencionada.

5.7. LA DISTRIBUCIÓN NORMAL TIPIFICADA

La regla 68 – 95 – 99,7 nos sugiere que todas las distribuciones normales son, en cierto modo, equivalentes, si usamos como unidades de

medida la desviación típica σ, y como origen de coordenadas la media μ. Esto puede ser útil en situaciones de comparación de variables diferentes, como en el ejemplo siguiente:

Ejemplo 5.7 *Ángel posee las siguientes calificaciones en un conjunto de asignaturas: 195 puntos en Inglés, 20 en Economía, 39 en Informática, 139 en Matemáticas y 41 en Física. ¿Es este estudiante mejor en Inglés que en Economía? ¿Será igualmente bueno en todas las asignaturas?*

Con la información proporcionada, no podemos responder a estas cuestiones, porque no sabemos cuál es el rango de calificaciones en cada asignatura, ni la distribución de las mismas en la clase, Las calificaciones de este estudiante y de otro compañero, así como los resultados de todos los alumnos de la clase se pueden observar en las columnas (2), (3) y (4) de la tabla 5.5.

Tabla 5.5. Calificaciones en 5 asignaturas

(1) Examen	(2) Media de la clase	(3) Desviación Típica de la clase	(4) Puntuaciones(X) Ángel	Carlos	(5) Desviaciones a la media (x) Ángel	Carlos	(6) Puntuaciones tipificadas (Z) Ángel	Carlos
Inglés	155,7	26,4	195	162	+39,3	+6,3	+1,49	+0,24
Economía	33,7	8,2	20	54	-13,7	+20,3	-1,67	+2,48
Informática	54,5	9,3	39	72	-15,5	+17,5	-1,67	+1,88
Matemáticas	87,1	25,8	139	84	+51,9	- 3,1	+2,01	-0,12
Física	24,8	6,8	41	25	+16,2	+ 0,2	+2,38	+0,03
Totales			434	397			+2,54	+4,51
Medias			86,8	79,4			+0,51	+0.90

Comparando las columnas (2) y (4) de la Tabla 5.5, podemos ver que Ángel está por encima de la media en Inglés, Matemáticas y Física, y está por debajo en Economía e Informática. Carlos, cuyas puntuaciones pueden verse en la columna 4, tiene puntuaciones mayores que el primero en dos asignaturas y puntuaciones menores para las otras tres. Sería injusto considerar sólo las puntuaciones absolutas para adjudicar la beca, debido a que cada asignatura puntúa en forma diferente. Necesitamos una escala común antes de realizar las comparaciones mencionadas anteriormente. Las puntuaciones típicas pueden proporcionarnos la escala común que estamos buscando.

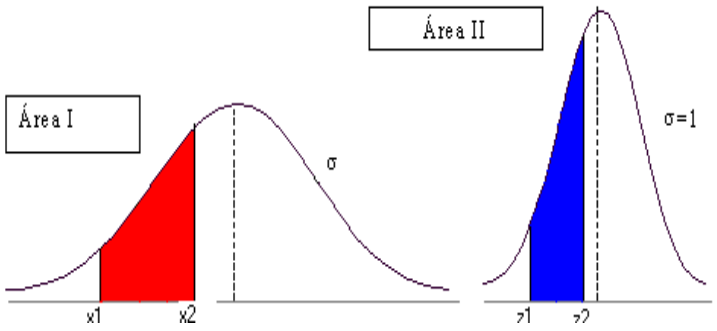
Como hemos comentado, la gráfica de todas las distribuciones normales podrían superponerse, si, en lugar de usar las puntuaciones

originales, las transformamos, usando como unidades de medida la desviación típica σ, y como origen de coordenadas la media μ. Este cambio de unidad de medida se llama *tipificación.* Si x es una observación de una distribución que tiene media μ y desviación típica σ, el valor tipificado de x es:

$$z = \frac{x - \mu}{\sigma}$$

Cuando hacemos una transformación en la variable, el área entre dos valores x_1 y x_2 en la distribución original (Área I) es igual al área entre los puntos transformados $z = z_1$ y $z = z_2$ (Área II) en la figura 18, puesto que la probabilidad de que la variable original toma valores comprendidos entre x_1 y x_2 es igual a la probabilidad que los valores transformados estén comprendidos entre $z = z_1$ y $z = z_2$.

Figura 5.10. Áreas equivalentes en la distribución original y tipificada



Cuando tipificamos varias variables con distribución normal, que, en principio tenían escalas diferentes, conseguimos pasar a un nuevo conjunto de variables que tienen una escala común.

Ejemplo 5.7. (continuación). Los dos estudiantes representados en la tabla 5.5 pueden ser comparados en términos de sus puntuaciones tipificadas z (columna 6). Ángel es superior a Carlos en inglés, matemáticas y física y deficiente en economía e informática.

La asignatura que mejor lleva Ángel es el inglés, pero en realidad, la mayor ventaja respecto a sus compañeros la lleva en física. Las puntuaciones originales de Carlos son más o menos las mismas en informática y matemáticas; sin embargo, tiene una ventaja bastante mayor en economía en términos de las puntuaciones tipificadas. Mientras que el

total de las puntuaciones originales da una ventaja a Ángel de 37 puntos, y en promedio una superioridad de cerca de 7 puntos, las puntuaciones tipificadas cambian el orden, dando a Carlos una ventaja de casi dos puntos y 0,39 en promedio. Por lo tanto, Carlos debería ganar la beca.

Actividades

5.19. Para comparar entre sí diferentes distribuciones normales, conviene tipificar la variable, restándole la media y dividiendo por su desviación típica, obteniendo de este modo las puntuaciones *Z* o puntuaciones tipificadas. Para la distribución de la actividad 1 (altura de chicas), tomando la $\mu = 165$ y $\sigma = 5$. a) ¿Cuáles serían las puntuaciones tipificadas para las alturas 164, 178, 150? b) ¿Qué alturas corresponden a las puntuaciones tipificadas *Z=0, Z=1, Z=-2?* Compara los resultados de ambos ítems.

5.20. ¿Cuál será la media y desviación típica de las puntuaciones tipificadas?

5.21. Dada una distribución de puntuaciones de un test que sigue la distribución normal de probabilidades, con media N(12,4), ¿Qué porcentaje de casos cae entre 8 y 16? ¿Qué proporción de casos se hayan por encima de la puntuación 18?

5.22. Dada una distribución N(29, 5), ¿qué tanto por ciento de la distribución caerá entre los valores 22 y 26?

5.23. Dada una distribución de puntuaciones N(16,4) ¿qué límites incluyen el 75 por ciento central de los casos? Si queremos aprobar el 75 por ciento de los alumnos, a partir de aquí nota debe considerarse aprobado?

5.24. Dada una distribución N(150, 25), ¿qué límites incluirán el 20 por ciento más alto de la distribución? ¿qué límites incluirán el 10 por ciento más bajo?

5.25. Las puntuaciones obtenidas por 300 niños de un colegio de EGB al aplicarles un test de aritmética siguen una distribución normal de media 24 y desviación típica 4. Calcula el cuartil inferior y el cuartil superior.

5.26. En una cierta población estudiantil el C.I. es una variable aleatoria N(100,18). De la experiencia se deduce que un estudiante de dicha población finalizará su carrera sin repetir ningún curso si su C.I. es al menos igual a 110. Calcular la proporción de estudiantes con coeficiente superior a 120 entre aquellos que finalizaron sus estudios sin repetir ningún curso.

5.27. Un camisero observa que el cuello de los jóvenes que concurren a su camisería es una variable aleatoria normal N(3.6, 7.5). De 3000 camisas que debe fabricar el próximo a˛o, ¿Cuántas han de estar comprendidas entre las siguientes medidas; 32-34; 34-35; 35-37?

5.31. El peso de los quesos fabricados en una cierta industria se distribuye normalmente. Se han fabricado 4000 piezas en un mes, de las cuales 800 pesaron menos de 1 kg y 1000 pesaron más de 2 kg. Determinar la media y desviación típica de dicha población normal.

5.28. Los errores aleatorios de una cierta medición obedecen a una ley normal con

una desviación típica de un 1 mm y esperanza matemática 0. Hallar la probabilidad de que de dos observaciones independientes el error por lo menos en una de ellas no supere el valor absoluto de 1.28 mm.

5.29. En una cierta población humana el índice cefálico se distribuye normalmente con media 74 y desviación típica 3. Hallar: a) La proporción de individuos que tiene un índice cefálico inferior a 75. b) Hallar los extremos entre los que varía el índice cefálico en 50 por ciento central de la población.

5.8. SUMA DE VARIABLES NORMALES INDEPENDIENTES

Sean $\xi_1, \xi_2 ... \xi_n$ n variables aleatorias normales independientes con distribuciones normales $N(\mu_i \sigma_i)$. Consideremos una nueva variable aleatoria ξ definida como la suma de las anteriores: $\xi = \xi_1 + \xi_2 + ... + \xi_n$. Debido a las propiedades lineales de la esperanza matemática y a la independencia de las variables, la media y varianza de ξ vendrán dadas por las expresiones siguientes:

(5.5) $$\mu = \mu_1 + \mu_2 + ... \mu_n$$

(5.6) $$\sigma^2 = \sigma^2_1 + \sigma^2_2 + ... \sigma^2_n$$

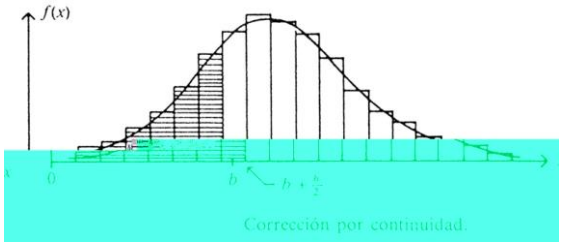
Puede demostrarse que ξ sigue una distribución normal con media y varianza dadas por las expresiones anteriores.

5.9. APROXIMACIÓN NORMAL A LA DISTRIBUCIÓN BINOMIAL.

Cuando *p* toma un valor moderado, las áreas bajo la curva normal pueden aproximar a la distribución normal, para valores grandes de *n*. Para entender esta aproximación, modifiquemos la representación gráfica de la distribución binomial, reemplazando el diagrama de barras, por un histograma. Puesto que esta variable toma valores enteros, será preciso utilizar estos valores enteros como centro de los intervalos en la construcción de dicho histograma. El área del mismo comprendido entre *x*-1/2 y *x*+1/2 será igual a la probabilidad de obtener el valor *x* (Figura 5.11).

Si sobre el histograma así construido representamos la gráfica de la curva normal de media *np* y varianza *npq*, podemos comprobar la equivalencia de las áreas, en especial cuando *n* toma valores suficientemente grandes. Como regla, puede utilizarse la aproximación normal a la distribución binomial cuando la menor de las cantidades *np* y *npq* es al menos igual a 5.

Figura 5.11



Supongamos que queremos calcular P(ξ <b) siendo *a* y *b* enteros, y ξ una variable aleatoria con distribución binomial *B(n,p).* En dicho caso podemos aproximar esta probabilidad de la siguiente manera:

$$P(\xi<b)=P(\xi<b+1/2)$$

siendo ξ una variable aleatoria con distribución normal *N(np,npq).* La cantidad 1/2 que se suma y resta a los extremos se conoce como corrección por continuidad.

Ejemplo 5.8. Sabiendo que el 60% de los glóbulos blancos pueden clasificarse como neutrófilos, hallar la probabilidad de que al realizar un análisis de 200 glóbulos blancos el número de neutrófilos encontrados esté comprendido entre 100 y 125. La variable aleatoria "número de neutrófilos en la muestra" sigue una distribución binomial *B*(200, 0.6). Por no disponer de valores tabulados para la distribución binomial con estos parámetros es preciso utilizar la aproximación normal, al tratarse de un valor p moderado.

Puesto que, en este caso $\mu=200*0.6=120$ y $\sigma^2=200*0.6*0.4=48$ tenemos:

$$P(100<\xi<125)=P(-3.54<Z<0.79)=0,7852-0,0003=0,7749$$

5.10. APROXIMACIÓN NORMAL A LA DISTRIBUCIÓN DE POISSON.

También la distribución de Poisson puede ser aproximada por la normal de igual media y varianza, para suficientemente grande.

Ejemplo 5.9. Un impresor produce inadvertidamente una errata por cada 5 páginas impresas. En un libro de 1000 páginas ¿Entre qué valores oscilará el número de erratas, con probabilidad 0,95? En este caso nos encontramos ante una distribución de Poisson de media $\mu=1000/5=200$. Por no disponer

de este valor tabulado, aproximaremos por la distribución normal de igual media y σ=14,14

P(200-a≤errores≤200+a)=P(-a/14,14≤Z≤a/14,14)=0,95, de donde: a/14,14=1.96; a=1.96*14,14=27,71.Por tanto, y tomando valores de a por exceso, el número de erratas del libro oscilará entre 172 y 228, con la probabilidad indicada.

Actividades

5.30. Se sabe que la probabilidad de que un matrimonio en el que ambos cónyuges son de genotipo A0 tenga un hijo de grupo 0 es 1/4. Consideremos una muestra de 400 hijos cuyos padres son A0. a) Calcula la probabilidad de que al menos 105 sean de grupo 0; b) ¿Entre qué límites oscilará el número con probabilidad 0.95?

5.31. En una fábrica hay 500 máquinas cada una de las cuales funciona sin problemas el 95 % de los días. Calcular la proporción de días en que más de 50 máquinas se habrán averiado.

5.32. Se ha comprobado que el 3 por ciento de las resistencias producidas en una fábrica son defectuosas. Si cada mes se fabrican 5000 resistencias, hallar: a) el número medio de resistencias defectuosas que resultan cada mes; b) la probabilidad de que un mes haya mas de 160 piezas defectuosas.

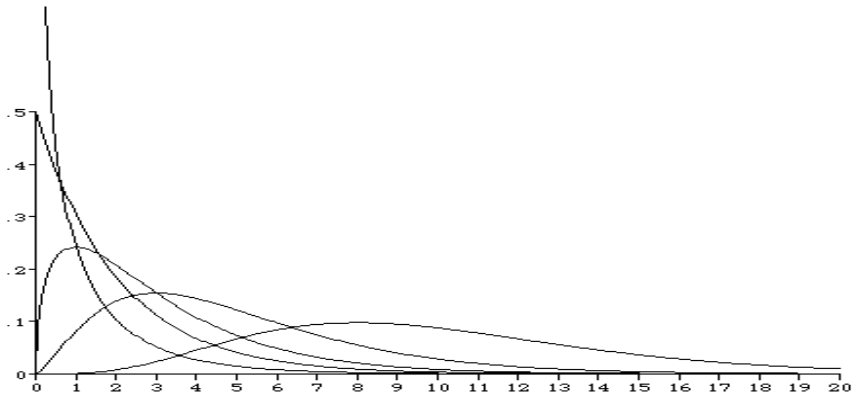
5.33. La probabilidad de sufrir reacción por una vacuna es 0.0001. En una ciudad, en la que se ha vacunado a 250.000 personas, ¿cuál es la probabilidad de obtener 30 0 más reacciones?

5.11. DISTRIBUCIONES RELACIONADAS CON LA NORMAL

La distribución Chi-cuadrado

Sean $\xi_1, \xi_2 ... \xi_n$ n variables aleatorias normales *N(0,1)* e independientes. Entonces la variable: $\chi^2 = \xi_1 + \xi_2 ... + \xi_n$ tiene una distribución que depende tan sólo del parámetro *n* y se conoce como distribución Chi-cuadrado. El parámetro *n* suele llamarse "grados de libertad". La media es igual a *n,* y la varianza a *2n*. Al ser la variable una suma de cuadrados, sólo toma valores positivos y su gráfica es asimétrica positiva, disminuyendo dicha asimetría conforme aumenta el valor de n, como puede observarse en los gráficos de la figura 5.12.

Figura 5.12. Gráficos Chi cuadrado para n creciente



Para *n* grande. la variable $\xi = \sqrt{2\chi^2} - \sqrt{2n-1}$ es aproximadamente una distribución normal N(0,1).

Ejemplo 5.10. En una distribución Chi-cuadrado con 100 grados de libertad, calcular el percentil del 90%. Queremos hallar a tal que P(χ^2 >a)=0.9 = $P(\sqrt{2\chi^2} - \sqrt{199} > \sqrt{2a} - \sqrt{199}) = P(Z > \sqrt{2a} - \sqrt{199})$. Puesto que, en la distribución normal N(0,1), el percentil del 90% es igual a 1,27 tenemos: $\sqrt{2a} - \sqrt{199}$ =1.28, de donde a=117,27

Actividades

5.34. Representar gráficamente, con la ayuda de un programa de ordenador, la distribución Chi-cuadrado, para diversos valores de *n* y comentar las diferencias.

5.35. Hallar las siguientes probabilidades en una distribución χ^2: P(χ^2 <28) para 15 grados de libertad; P(10< χ^2 <15) para 20 grados de libertad.

5.36. Hallar a tal que P(χ^2 <a)=0.1, siendo χ^2 una Chi-cuadrado con 10 g.l.

5.37. En una distribución Chi-cuadrado de 25 g.l. hallar la mediana y los cuartiles.

La distribución T

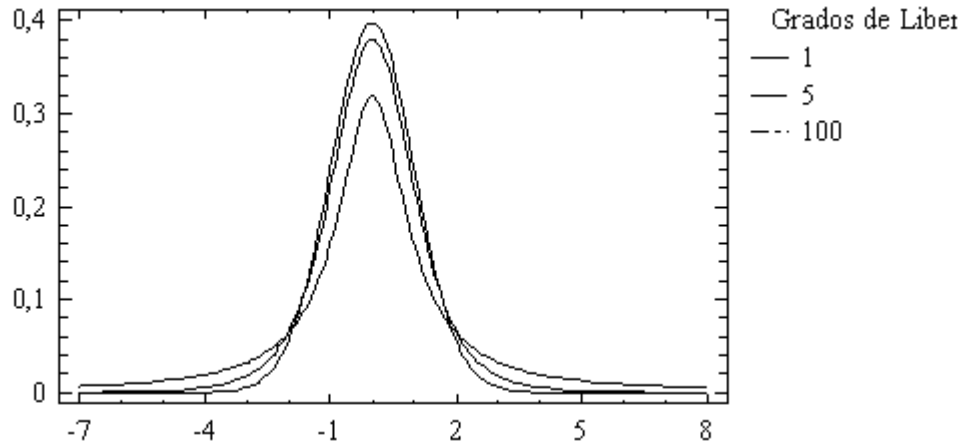
Otra distribución que tendrá gran utilización en inferencia es la T debida a "Student". Puede ser definida mediante la relación:

$$T = \frac{Z}{\sqrt{\chi^2 / n}}$$

Donde *Z* es una variable aleatoria normal *N(0,1)* y χ^2 una

distribución Chi-cuadrado con *n* grados de libertad. Esta distribución depende solamente del parámetro *n* o grados de libertad, y tiene media cero y varianza igual a *n/n-2*.

Figura 5.13. Distribución T



En la figura 5.13 se muestra la gráfica de la distribución *T*. Puede observarse que es simétrica respecto al origen de ordenadas, por lo que, al igual que en la normal standard, la media, moda y mediana de esta distribución es igual a cero. Su forma es parecida a la de la curva normal, mejorando la aproximación conforme aumenta el valor de *n*. Debido a este hecho la distribución normal puede servir como aproximación para la distribución *T*, para valores suficientemente grandes, Este hecho puede comprobarse comparando las tablas de las dos distribuciones.

Actividades

5.38. En una distribución *T* con 25 g.l. hallar a tal que $P(T > a)=0.9$. Hallar b tal que $P(T<b)=0.8$.

5.39. Con ayuda de una programa de ordenador, representar las gráficas de la distribución *T*, para diversos valores de *n* y comentar las diferencias.

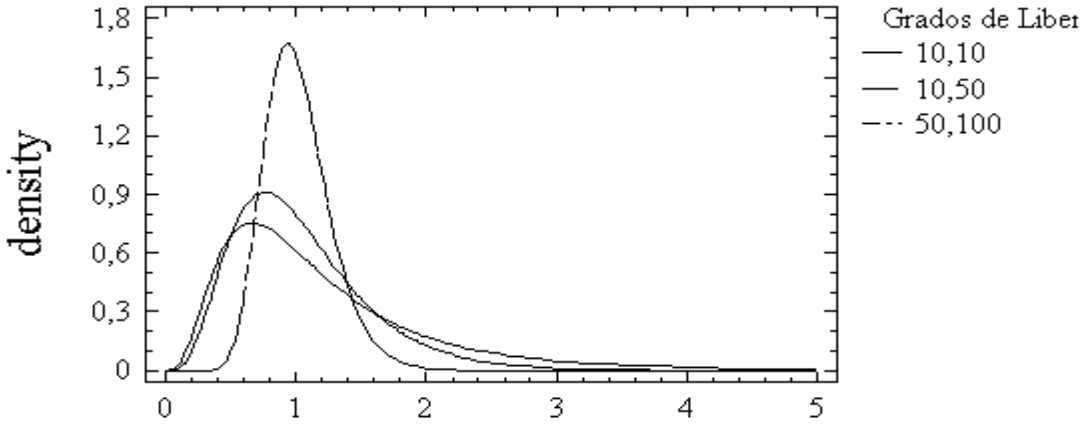
La distribución F

Dadas dos variables aleatorias *X* e *Y* que se distribuyen según una Chi-cuadrado de *m* y *n* grados de libertad respectivamente, el cociente:

$$F = \frac{X/m}{Y/n}$$

Es una variable aleatoria cuya distribución es conocida como distribución *F* con *(m,n)* grados de libertad. Depende de dos parámetros, y por su definición toma sólo valores positivos, como también puede apreciarse en las graficas de la figura (7.6). En dichas gráficas se puede observar la evolución de la distribución con la variación de los parámetros.

Figura 5.14.



Actividades

5.40. Con ayuda de un programa de ordenador, representar gráficamente la distribución *F* para diferentes valores de los parámetros.

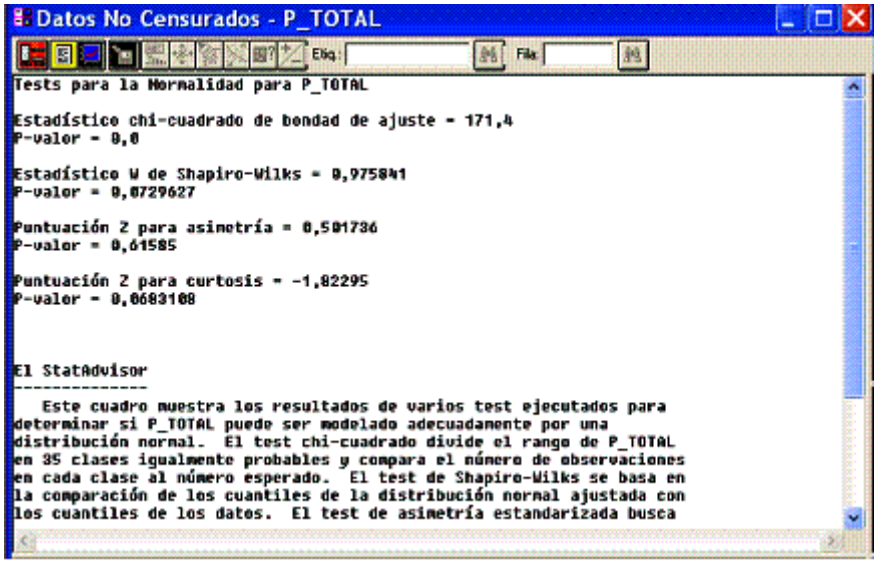
5.41. En una distribución *F*(7,10), hallar el percentil del 95 por ciento.

5.42. ¿Cual es la probabilidad de obtener un valor menor que 3.5 en una distribución *F*(8,12)?

5.15. AJUSTE DE DISTRIBUCIONES

En algunos casos queremos ver si algunos de los modelos teóricos de probabilidad podría ser una buena aproximación para nuestros datos. Un ejemplo es tratar de ver si la distribución normal teórica podría ser útil para describir un conjunto de datos (datos empíricos). Para ver si hay un *buen ajuste* entre el modelo teórico y los datos, usamos el programa Ajuste de Distribuciones.

Figura 5.15. Ventana con la opción Ajuste de Distribuciones



Entrando en el menú **Descripción**, seleccionamos la opción **Distribuciones** y luego **Ajuste de distribuciones**. Aparece un cuadro de diálogo, en el que se selecciona la variable que se quiere analizar, para ver si se ajusta bien a un modelo teórico. Por defecto obtenemos la pantalla de la figura 5.16, que se refiere al ajuste de la distribución normal. En ella se ve, por defecto, la media y la desviación típica de la variable que ha sido seleccionada y un comentario del Stat Advisor. Podríamos cambiar estos parámetros por defecto, mediante Opciones de análisis (pinchando con el botón derecho del ratón) y usar una distribución diferente de probabilidad. En general probamos diferentes modelos para ver cuál de ellos se ajuste mejor a los datos. El cuadro de selección de distribuciones es similar al de la figura 34.

Figura 5.16. Cuadro de diálogo para la selección de distribución

Cálculo de probabilidades a partir de la distribución ajustada

Una vez que hemos comprobado que el ajuste es bueno, estamos a veces interesados en calcular probabilidades de obtener ciertos valores, usando el modelo teórico. Estando en la ventana de la figura 35, se selecciona el icono Opciones Tabulares, y en el cuadro de diálogo que aparece en la figura 5.17 se selecciona Áreas de Cola.

Aparecerá una pantalla en la que están calculadas algunas probabilidades correspondientes a valores que aparecen por defecto. Estos valores pueden cambiarse, apretando el botón derecho sobre la pantalla y seleccionando Opciones de Ventana. Se podrán introducir hasta 5 valores de la variable. Los resultados obtenidos nos darán el valor del área bajo la curva o la probabilidad de que la variable tome un valor menor o igual que el valor dado por nosotros.

Figura 5.17. Cuadro de diálogo en el cálculo de áreas de cola

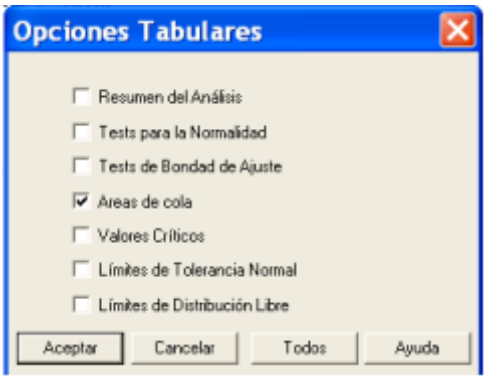
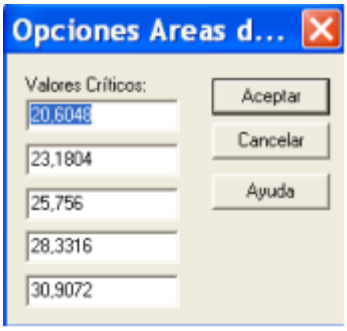


Figura 5.18. Introducción de valores para calcular la probabilidad



En el caso de que queramos calcular la probabilidad de que sea mayor que ese valor deberemos hacer un cálculo auxiliar, como por ejemplo tomar el resultado que nos da el programa y restárselo a 1. En la figura 5.18 aparece el cuadro de diálogo en el que pueden cambiarse los valores de la variable. Una vez que se hace clic sobre el botón Aceptar aparece una

ventana con los resultados solicitado, que se presentan en las figuras 5.19 y 5.20.

Figura 5.19. Cálculo de probabilidades

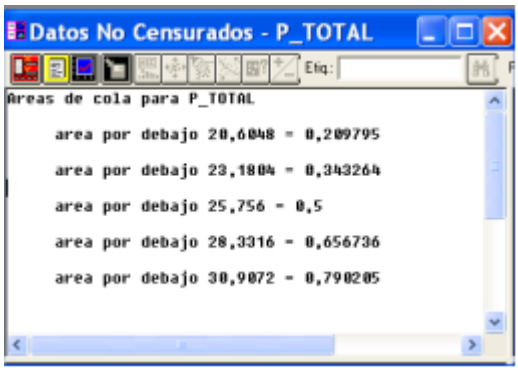
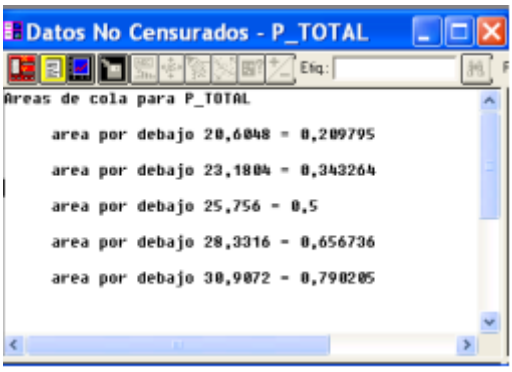


Figura 5.20. Valores críticos



En otras ocasiones, conocemos el valor de la probabilidad y queremos calcular a qué valores de la variable corresponde, por ejemplo, sabiendo que la probabilidad es del 90% (0,90) se desea conocer cuál es el extremo superior del área bajo la curva correspondiente a esa probabilidad. En estos casos, seleccionar el botón Opciones tabulares y aparecerá un cuadro de diálogo para seleccionar Valores críticos. También aquí, se pueden cambiar estos valores, pinchando con el botón derecho del ratón y seleccionando Opciones de ventana.

Representación de funciones de densidad y de distribución acumulada

Si queremos estudiar gráficamente la manera en que varía la función de densidad o la función de distribución acumulada de una distribución cualquiera, al variar el valor de sus parámetros, podemos graficar hasta cinco curvas en un mismo sistema de ejes, y de esta manera poder realizar el estudio gráfico.

Para realizar esto, debemos ingresar a la opción Distribuciones de probabilidad del menú Gráficos y se selecciona el modelo de distribución que deseamos (en el caso de la figura está seleccionada el modelo normal). Haciendo clic en el botón Aceptar de este cuadro, se obtiene una ventana de análisis, en la que aparecen por defecto los valores de los parámetros de la distribución normal típica (0,1). Se pueden ingresar hasta cinco pares de parámetros. En la figura 5.21 se han ingresado tres pares de parámetros con la misma desviación y distintas medias, que representan a tres distribuciones normales.

Figura 5.21. Tres análisis simultáneos

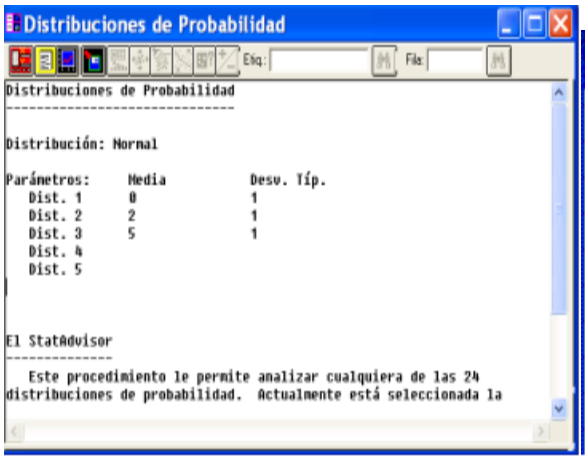
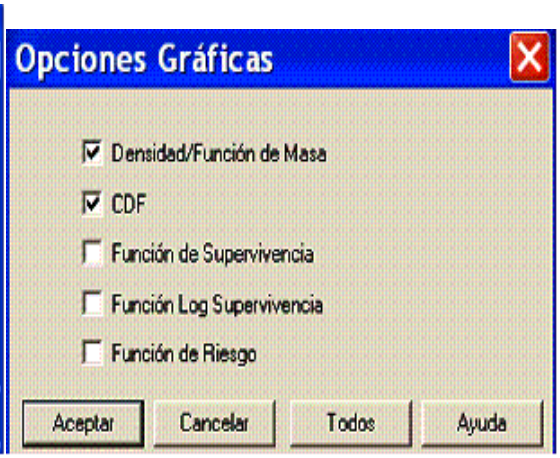
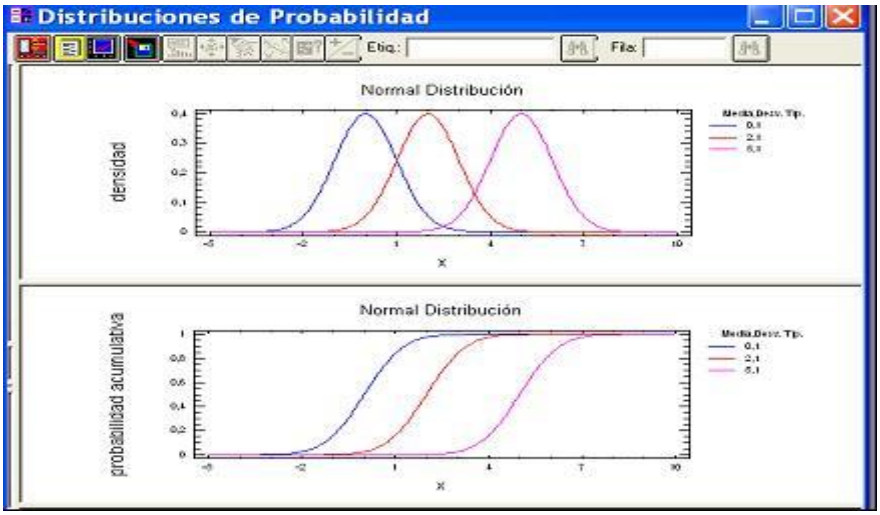


Figura 5.22. Opciones



Estando en la ventana de la figura 21, si se quiere obtener las gráficas de las funciones de densidad y de las funciones de distribución acumulada, se ingresa al botón Opciones Gráficas y en el cuadro de la figura 5.22, se selecciona Densidad/ Función de Masa (función de densidad) y CDF (función de distribución acumulada), o pueden seleccionarse de a una por vez, en cualquiera de estos casos aparecerán las ventanas de la figura 5.23.

Figura 5.23. Ventana de análisis gráfico



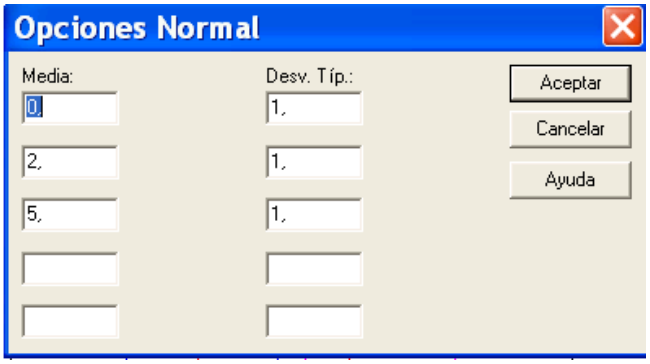
5.16. REPRESENTACIÓN Y GENERACIÓN DE VALORES ALEATORIOS DE DISTRIBUCIONES TEÓRICAS

El programa Gráficos representa las gráficas, realiza cálculos y genera valores aleatorios de diferentes distribuciones de probabilidad, por ejemplo, la distribución normal.

Cálculo de probabilidades teóricas

Se trata de calcular la probabilidad de que una cierta distribución teórica (por ejemplo, la normal) tome ciertos valores. Al entrar al menú Gráficos – Distribuciones de probabilidad – aparece una ventana con diversos modelos de distribuciones. Si, por ejemplo, seleccionamos la distribución NORMAL, aparecerá una ventana de análisis. En ella pulsamos el botón derecho del ratón y seleccionamos Opciones de análisis. Aparecerá un cuadro de diálogo como el de la figura 5.24, donde daremos los valores de la media y desviación típica correspondiente a la distribución que se está utilizando.

Figura 5.24. Diálogo para introducir la media y la desviación típica



Aparecerá una ventana de análisis, y al seleccionar Opciones Tabulares, aparecerá una cuadro de diálogo, seleccionar la opción Distribución Acumulada. Aparecerá una ventana con tres partes diferenciadas: Área de cola inferior (<), densidad de probabilidad, en la que se da la ordenada de la función de densidad, y Área de cola superior (>).

Figura 5.25. Ventana de resultados

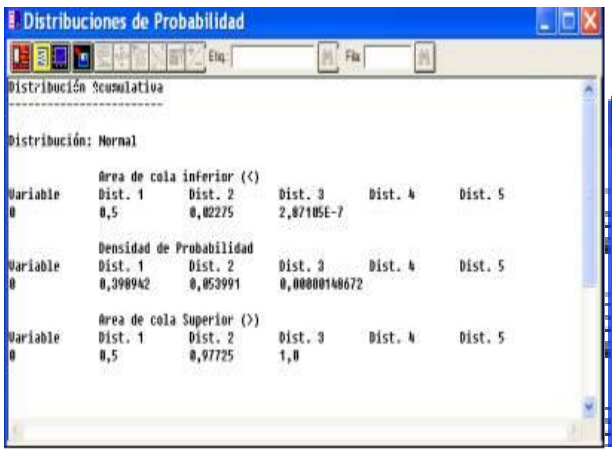
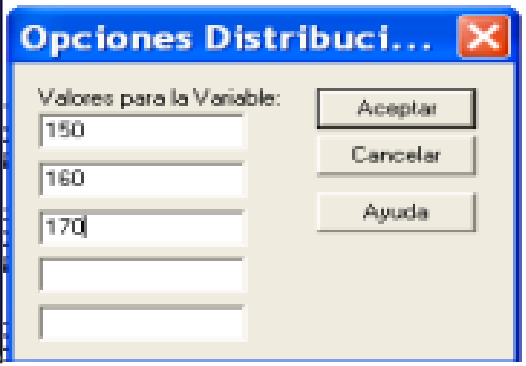


Figura 5.26. Distribución Acumulada



Sobre la ventana de la figura 5.25, haciendo clic con el botón derecho y seleccionando Opciones de ventana, aparecerá un cuadro de diálogo como el de la figura 5.26, en el que se pueden variar los valores de la variable para los cuales se desea calcular la probabilidad.

Generación de números aleatorios

Para generar números aleatorios, ingresar al menú Gráficos – Distribuciones de probabilidad, allí aparecerá una ventana como la de la figura 36, en la que se deberá seleccionar el modelo de distribución con el que se desea trabajar, por ejemplo seleccionaremos la distribución discreta uniforme. A continuación, aparecerá una ventana de análisis (Figura 5.27), donde se puede seleccionar Opciones de Análisis, allí aparecerá un cuadro de diálogo, en el que se ingresarán los límites inferior y superior de la variable que se desea generar. Al seleccionar el botón Opciones tabulares, aparecerá un cuadro de diálogo, allí seleccionar Números aleatorios, aparecerá otra ventana de análisis y entrar a Opciones de Ventana para definir el tamaño de la muestra que se desea generar; por defecto aparece el tamaño 100.

Figura 5.27. Ventana de análisis

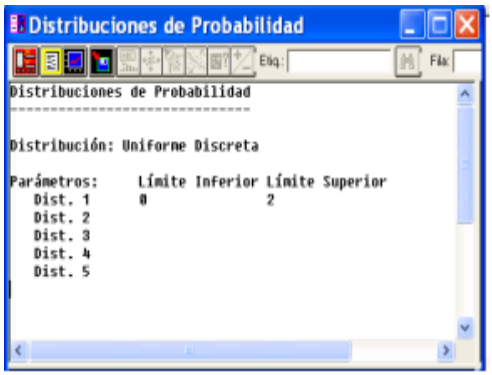
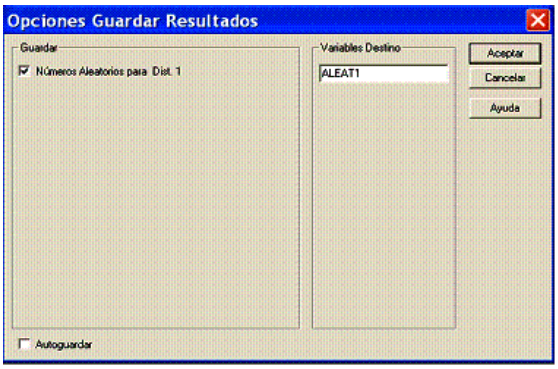


Figura 5.28. Grabación de la variable generada



Una vez que se generan los números aleatorios, se deben grabar, para ello entrar en el botón Guardar resultados, ingresar el nombre de la variable (en la figura aparece como ALEAT1) y seleccionar el campo Números aleatorios para Dist. 1, luego Aceptar. De esta manera se generará una nueva variable en la hoja de cálculo. Este procedimiento puede repetirse todas las veces que se desee.

TEMA 6

MUESTREO Y ESTIMACIÓN

6. 1. MUESTRAS Y POBLACIONES

En los temas anteriores hemos estudiado, por un lado, la Estadística Descriptiva, cuyo objeto es describir los datos obtenidos en observaciones u experimentos. Estos datos son usualmente representados por una o varias variables estadísticas, cuya distribución de frecuencias y demás características son obtenidas a partir de los datos, que en la mayor parte de los casos constituyen una muestra particular de la población. Por otro lado, mediante el Cálculo de Probabilidades, introducimos el concepto de variable aleatoria, al considerar que aumentamos indefinidamente las observaciones y representar todos los posibles valores que puede tomar un carácter en una población, o todos los posibles valores que pueden surgir como consecuencia de la realización de un cierto experimento.

Actividades

6. 1. Supongamos que se obtuvieron los siguientes resultados en las pasadas elecciones: El 40% del total de los votantes, votaron al PP, el 38% votó al PSOE y 9% votó a IU. Si en esta ciudad tomamos una muestra aleatoria de 100 votantes y les preguntamos a quien votaron (imaginamos que las personas a las que preguntamos son sinceras),

a) ¿Podemos decir que necesariamente de estos 100 votantes, 40 votaron al PP, 38 al PSOE y 9 a IU?

b) Supongamos que tomamos varias muestras aleatorias de 100 votantes. ¿Encontraremos siempre la misma proporción de votantes a cada partido en cada muestra? ¿Podrías adivinara, aproximadamente el porcentaje aproximado de personas que en cada muestra habrían votado al PP?

c) Supongamos ahora que tomamos una muestra de 100 votantes en el País Vasco. ¿Crees que variarían los resultados?

6.2. Supón que quieres comprar un coche nuevo y quieres decidir entre la marca A y B. En una revista de automóviles encuentras un estudio estadístico sobre reparaciones efectuadas el último año que muestra que la marca A tiene menos averías que la B. Sin embargo, te encuentras un amigo tuyo que te dice que compró el año pasado un coche B y no ha tenido más que problemas: primero se le estropeó la inyección de gasolina y gastó

120 €, luego tuvo que cambiar el eje trasero y al final, ha vendido el coche porque se le fue la transmisión. ¿Que decisión tomarías, comprar un coche A o B?

En muchos estudios estadísticos estamos interesados en obtener información acerca de una o varias variables en una población determinada. Aunque a veces es posible estudiar toda la población completa mediante un **censo**, otras veces es preciso contentarse con una *muestra* de la misma. La idea es obtener información de la población estudiando sólo una parte de la misma (la muestra). El proceso de generalizar los resultados obtenidos en la muestra a toda la población recibe el nombre de *inferencia* estadística. Hay dos características importantes en las muestras, que son:

- *Variabilidad muestral:* No todas las muestras son iguales. Los elementos de distintas muestras pueden ser diferentes, y, por tanto, los resultados de una muestra a otra pueden variar.
- *Representatividad:* Si elegimos una muestra adecuadamente, puede representar a la población, en el sentido de que los resultados en la muestra pueden servir para estimar los resultados en la población.

Los motivos que hacen necesario el uso de estas técnicas pueden ser económicos, ya que es más costoso y lleva más tiempo obtener información de toda la población. También puede darse el caso de que el experimento que debe realizarse tenga carácter destructivo, como ocurre en algunos ensayos de fiabilidad.

Otras veces la población está constituida por entes potenciales, como es el caso de los ensayos médicos en que se consideran los posibles enfermos con una dolencia; o bien se trata de una población infinita. Por último, la gran homogeneidad de algunas poblaciones hace innecesario el estudio de la totalidad de la misma, como ocurre al efectuar, por ejemplo, un análisis de sangre, con objeto de efectuar el recuento de hematíes.

Actividad

6.3. Discute en cuál de los siguientes estudios por muestreo habrá más variabilidad y en cuál habrá más representatividad

a) Tomar al azar muestras de 10 votantes para estimar la proporción de personas que votaron al PSOE ;
b) Tomar al azar muestras de 1000 votantes para estimar la proporción de personas que votaron al PSOE;
c) Tomar al azar muestras de 1000 votantes para estimar la proporción de personas que

votaron a IU;

d) Tomar al azar muestras de 10 votantes para estimar la proporción de personas que votaron a IU;
e) Tomar muestras de 1000 jubilados para estimar la proporción de personas que votaron al PSOE;
f) Tomar muestras de 10 personas al azar para estimar la proporción de mujeres .

6.2. TIPOS DE MUESTREO

Hay muchas formas diferentes de elegir las muestras. Por ejemplo, si queremos hacer un estudio de los alumnos de la Facultad de Ciencias de la Educación, podríamos formar una muestra con alumnos voluntarios. Sin embargo, si queremos que nuestros resultados sean generalizables, hay que planificar la elección de la muestra, siguiendo unos requisitos, que aseguren que la muestra ha sido elegida aleatoriamente de la población. Los métodos de inferencia estadística están basados en la utilización de unos métodos de muestreo probabilístico. El muestreo se dice probabilístico cuando puede calcularse de antemano la probabilidad de obtener cada una de las muestras que sea posible seleccionar. Para ello, es necesario que el proceso de selección pueda considerarse como un experimento aleatorio. Algunos tipos de muestreo probabilístico son:

- *Muestreo aleatorio simple:* Cuando los elementos de la muestra se eligen al azar de la población y cada elemento tiene la misma probabilidad de ser elegido. Puede realizarse con reemplazamiento (una vez elegido un elemento para formar parte de la muestra se puede volver a elegir de nuevo) o sin reemplazamiento.
- *Muestreo estratificado:* Primero dividimos la población en grupos de individuos homogéneos, llamados estratos. De cada estrato se toma una muestra aleatoria. El tamaño de la muestra global se divide proporcionalmente al tamaño de cada estrato.
- *Muestreo sistemático:* Se supone que los elementos de la población están ordenados. Si queremos tomar en la muestra uno de cada n elementos de la población, elegimos al azar un elemento entre los n primeros. A continuación sistemáticamente elegimos uno de cada n elementos.
- *Muestreo por conglomerados:* Se divide la población en unidades representativas de la misma (y por tanto heterogéneas) y se extrae aleatoriamente un grupo de éstas sobre las cuales se efectúa la medición. Por ejemplo, para realizar una encuesta sobre presupuestos familiares, la ciudad puede dividirse en manzanas de viviendas, y se toman al azar, varias de estas manzanas en las cuales se efectúa la encuesta a todos los vecinos de la misma.

- Puede realizarse un *muestreo en dos o más etapas*, cuando cada una de las unidades tomadas para el muestreo puede a su vez ser muestreada. En el ejemplo anterior, una vez elegida una manzana de viviendas para formar parte en la muestra, se sortea entre todas las viviendas que la componen para decidir cuales serán encuestadas.
- También puede realizarse un *muestreo opinático o intencional*. En este caso, la persona que selecciona la muestra es la que decide los elementos que la constituirán, procurando que ésta sea representativa de la población. Sin embargo, la representatividad real dependerá de las preferencias u opinión de esta persona y, por tanto, este tipo de muestreo carece de base te⌜rica suficiente.
- Por último, en el *muestreo sin norma,* se toma la muestra de cualquier manera y se obtiene así una parte de la población. Si esta es homogénea, la representatividad de la muestra puede ser satisfactoria. Este tipo de muestreo se emplea a menudo en la vida diaria (así, se prueba un trozo de queso o un sorbo de vino, etc, y se juzga el resto por el resultado).

Actividades

6.4. En una caja hay 3 bolas que pesan 1, 3 y 4 kg. respectivamente. ¿Cuáles son el peso medio y la varianza del peso en esa población? Si tomas muestras de 2 bolas con reemplazamiento: construye la distribución del peso medio muestral, su esperanza y su varianza. Repite el ejercicio pero sin reemplazamiento. Compara los resultados.

6.5. Se desea hacer una encuesta en la Facultad para averiguar el tiempo de desplazamiento de la Facultad a su domicilio de los estudiantes. Discute las diferentes formas de tomar una muestra de 1000 estudiantes y sus ventajas relativas.

Obtención de muestras aleatorias de una distribución teórica con STATGRAPHICS

Las *tablas de números aleatorios* contienen una secuencia de dígitos que han sido generados al azar, y han pasado una serie de pruebas para contrastar su aleatoriedad. En los microordenadores actuales se incluye usualmente una función que genera números aleatorios dentro de un rango fijado por el usuario.

Si, de una población de *N* elementos se desea tomar una muestra de tamaño n, se numeran de 1 a *N* todos los elementos de la población. A continuación, se seleccionan n números aleatorios comprendidos entre 1 y N, que serán los elementos a formar parte en la muestra. Dicha selección puede hacerse directamente, mediante un programa que genere los n números deseados, o a partir de una de las tablas de números aleatorios disponibles. Para utilizar una de ellas, basta tomar n números consecutivos comprendidos en el rango dado, a partir de uno cualquiera de los números de la tabla, leyendo en ella por filas o columnas.

Ejemplo 6.1. En el tema anterior vimos que la distribución de los coeficientes de inteligencia era aproximadamente normal, con media 100 y desviación típica 15. Es decir, μ=100, σ=15, cuando consideramos la variable aleatoria ξ: "Puntuación en la prueba del coeficiente de inteligencia de una persona extraída al azar". La población de referencia es la de todas las personas de una misma edad y la media μ ha sido calculada teóricamente, ajustando una distribución normal a los datos recogidos de cientos de miles de personas que han respondido al test.

Sin embargo y aunque la puntuación media teórica μ sea igual a 100, esto no quiere decir que cuando pasamos el test a una muestra de personas (por ejemplo en una clase) el valor medio $\bar{x}$ en la muestra sea igual exactamente a 100. Estudiaremos en este ejemplo el comportamiento de la media $\bar{x}$ en las muestras de valores del coeficiente de inteligencia, para distintos tamaños de muestras.

Para realizar este estudio, usaremos el programa Statgraphics, seleccionando la opción Gráficos, y dentro de ella Distribuciones de Probabilidad. Dentro de esta opción, tomaremos la Distribución Normal. En la pantalla Opciones Tabulares, seleccionamos la opción Números Aleatorios, que sirve para generar valores aleatorios de la distribución seleccionada.

Para ello basta seleccionar con el ratón el icono del disco y marcar la opción Guardar. Se generan 100 números aleatorios de la distribución normal N(0,1). Si queremos otro tamaño de muestra podemos cambiarlo mediante Opciones de Ventana. Si queremos cambiar los parámetros de la distribución normal, podemos hacerlo mediante Opciones de Análisis.

Nosotros hemos cambiado estos parámetros y hemos generado una muestra aleatoria de cuatro elementos de la distribución N(100, 15). Los valores obtenidos han sido: 118, 116, 78, 120.

De estos valores tres superan el valor medio y uno está por debajo. La media de los mismos es 108 que no coincide con el valor exacto 100, pero se aproxima. Tomemos una nueva muestra de cuatro valores al azar. Obtenemos: 88, 115, 89, 86. Ahora hay tres valores por debajo de 100 y uno por encima y el valor medio de los mismos es 94.5.

Estadísticos y parámetros

En el tema anterior hemos estudiado la distribución normal. Una distribución normal queda determinada por su media μ, y su desviación típica σ y la representamos por $N(\mu,\sigma)$. La media y desviación típica de la distribución normal determinan completamente la función de densidad. Por ello decimos que la media y la desviación típica son los *parámetros* de la distribución normal.

Si al realizar un estudio estadístico sospechamos que la variable de interés

podría ser aproximada adecuadamente mediante una distribución normal, nuestro interés se centrará en hallar el valor aproximado de estos parámetros (media y desviación típica), porque conocidos estos valores, habremos determinado la función de densidad de la variable y podremos calcular cualquier probabilidad relacionada con ella.

Recuerda:

- *Variable aleatoria* es la variable que surge de un *experimento aleatorio,* consistente en considerar todos los posibles valores de una variable en una población. La variable aleatoria se describe mediante su distribución de probabilidad. Si la variable aleatoria es cuantitativa y continua, viene descrita por su función de densidad.
- La *variable estadística* surge de un *experimento estadístico*, consistente en tomar datos de una variable aleatoria sólo en una muestra de la población. Describimos la variable estadística mediante la distribución de frecuencias y si es cuantitativa y continua la representamos gráficamente por medio del histograma.
- Llamamos *parámetros* a las medidas de posición central, dispersión y, en general cualquier resumen calculado en la variable aleatoria, es decir, en toda la población.
- Llamamos *estadísticos* a las mismas medidas cuando se refieren a la variable estadística, es decir, cuando se calculan sólo a partir de una muestra tomada de la población.

Actividad

6.6. En los siguientes enunciados identifica si los valores mencionados se refieren a un parámetro o a un estadístico y la población de interés a la que se refieren:

a) La proporción de todos los estudiantes de la facultad que han viajado al extranjero;
b) La proporción de estudiantes que han viajado al extranjero entre 100 estudiantes de la facultad elegidos al azar;
c) La proporción de los españoles que votaron al PSOE en las últimas elecciones;
d) La proporción de "caras" en 100 lanzamientos de una moneda;
e) El peso medio de 20 bolsas de patatas fritas de una cierta marca;
f) La proporción de personas que declararon votar al PSOE en una encuesta realizada después de las elecciones;
g) El peso medio de los chicos españoles de 18 años;
h) El peso medio de 10 chicos españoles.

6.7. ¿Por qué la proporción muestral es una variable aleatoria? Cita otras posibles variables aleatorias muestrales.

6.3. PROPIEDADES DE LOS ESTIMADORES

Al tratar extender los resultados de la muestra a la población podemos cometer dos tipos de errores:

- Errores sistemáticos o *sesgos.* Estos errores tienen siempre un mismo signo, se producen porque la muestra está sesgada y no es representativa de la población, o bien porque el instrumento que usamos para recoger los datos no es adecuado. Pueden ser controlados con una elección adecuada de la muestra y de los instrumentos (cuestionarios u otros) que empleamos para recoger los datos. La *validez* de un estudio es la ausencia de sesgos.
- *Errores aleatorios.* Estos errores pueden tener distintos signos, de modo que pueden compensarse entre sí al aumentar el tamaño de la muestra. La *precisión* de un estudio indica la magnitud del error aleatorio.

Un estimador $\bar{\theta}$ de un cierto parámetro θ se llama *insesgado* o centrado, cuando la esperanza matemática de su distribución de probabilidad coincide con el valor del parámetro que tratamos de estimar, esto es:

$$E[\ \bar{\theta}] = \theta$$

Así, por ejemplo, la media muestral $\bar{x}$ es un estimador insesgado de la media de la población, ya que:

$$E(\bar{x}) = \frac{E(x_1 + x_2 + ...x_n)}{n} = \frac{E(x_1) + E(x_2) + ...E(x_n)}{n} = \mu$$

Sin embargo, la varianza muestral no es un estimador insesgado de la varianza poblacional. Sea S_0^2 la varianza en la muestra. Puede demostrarse que:

$$E(S_0^2) = \frac{(n-1)\sigma^2}{n}$$

Decimos entonces que S_0^2 es un estimador *sesgado* de la varianza poblacional. A la cantidad $E[\ \bar{\theta}] - \theta$ se le llama *sesgo* del estimador. En el caso de la varianza el sesgo es igual a σ^2/n. Al ser la varianza muestral un estimador sesgado, se utiliza en su lugar la *cuasivarianza mues*tral, que viene definida por:

$$\tilde{S}_0^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad (6.1)$$

Además de la propiedad anterior, una característica importante del estimador es su precisión. Esta propiedad mide la dispersión del estadístico alrededor del barómetro que se trata de estimar. Como veremos mas adelante, en general, la precisión del estimador aumenta con el tamaño de la muestra.

De la propiedad opuesta a la precisión, que es la variabilidad, pueden darse diversas medidas. La mas utilizada es la varianza de la distribución del estadístico y su raíz cuadrada, conocida como *error de muestr*eo o *error estándar*.

Al utilizar el muestreo con reemplazamiento, el error de muestreo de la media es, por consiguiente igual a $\sigma/\sqrt{n}$, y se suele estimar por $S/\sqrt{n}$, cuando no se conoce σ. En las mismas condiciones, el error de muestreo de S^2 viene estimado por $S^2/\sqrt{2/(n-1)}$. Usualmente, los paquetes estadísticos incluyen el cálculo de los estimadores de diversos parámetros y de sus errores de muestreo.

Actividad

6.8. En una población, la varianza es igual a 200. Calcular el error de muestreo de la media muestral, si tomamos una muestra de n=10, 100, 1000 elementos, con reemplazamiento.

6.9. El sueldo medio de los trabajadores de un sector es de 1200 euros y la desviación típica 80 euros. Si se toman muestras de 100 trabajadores: ¿En qué porcentaje de muestras saldrá un sueldo medio menor que 1000 euros? ¿En qué porcentaje saldrá un sueldo medio mayor a 1300 euros?

6.4. DISTRIBUCIONES DE LOS ESTADÍSTICOS EN EL MUESTREO

Al comparar un *parámetro* (por ejemplo la media de la variable aleatoria en la población) con su correspondiente *estadístico* (la media de la variable en la muestra) vemos que:

- El *parámetro,* es un resumen de la distribución (por ejemplo la media, la varianza o el coeficiente de correlación). Se calcula en el total de la población, es un valor constante, pero desconocido.
- El *estadístico* es también un resumen, pero se refiere sólo a los datos de una muestra. Conocemos su valor una vez que tomamos la muestra, pero este valor puede variar en una muestra diferente. El estadístico es una variable aleatoria, porque tomar una muestra es un experimento aleatorio (no sabemos qué muestra saldrá) y el valor del estadístico cambia de una muestra a otra.

Ejemplo 6.2. Una cadena de televisión quiere estudiar los índices de audiencia de uno de sus programas, medido por la proporción de personas que ven el programa una determinada semana. Para ello diseñan un proceso de muestreo y eligen 1000 familias en forma que la muestra sea representativa de la población. En cada familia recogerán datos del número de personas de la familia que vio el programa esa semana y el total de personas que componen la familia:

- La proporción de personas que vio el programa esa semana en todo el país es un parámetro. Es un valor constante, pero no lo conocemos.
- La proporción de personas que vio el programa en la muestra es un estadístico. Supongamos que se obtuvo una proporción del 15% de audiencia en la muestra. En otra muestra de personas esta proporción podría variar, aunque si las muestras están bien elegidas esperamos que los valores se acerquen a la proporción (parámetro) en la población.

Actividades

6.10. Al experimento aleatorio consistente en lanzar un dado podemos asociarle la variable aleatoria "Número de puntos obtenidos". Representa, mediante un diagrama de barras la distribución de esta variable aleatoria. ¿Cuál es su valor medio μ? La población a que se refiere esta variable es la de todos los valores que podríamos obtener si imaginamos que lanzamos indefinidamente un dado y anotamos los valores obtenidos.

6.11. Supongamos que tomamos una muestra de dos valores al lanzar un dado. ¿Cuáles son las posibles muestras que podías obtener? ¿Cuál sería la media $\bar{x}$ de cada una de las muestras? Representa gráficamente la distribución de probabilidad de la variable aleatoria $\bar{x}$: "valor medio del número de puntos en una muestra de 2 lanzamientos de un dado" ¿Cuál es la media de esta variable aleatoria? Calcula la desviación típica.

6.12. Obtén 10 muestras de dos valores del lanzamiento de un dado y calcula la media de cada muestra. Representa los valores obtenidos, poniendo una cruz encima del valor obtenido en la siguiente gráfica (en rojo o con lápiz). Completa el gráfico representando los datos obtenidos por el resto de la clase (en un color diferente).

1 1.5 2 2.5 3 3.5 4 4.5 5 5.5 6

Hemos visto en la actividad anterior como la media de la muestra es una variable aleatoria. Sin embargo, si consideramos todas las muestras que podríamos obtener de la población, la media de todas las medias muestrales coincide con la media en la población. De la gráfica anterior, también podemos observar como los valores cercanos a la media de la población se obtienen con mayor frecuencia que los valores alejados.

Al realizar un proceso de muestreo, manejamos diferentes distribuciones de probabilidad o frecuencias. Consideraremos las siguientes:

- Distribución de probabilidades de la variable estudiada en la población.

- Distribución de frecuencias de los valores obtenidos de dicha variable al estudiar una muestra de n elementos.
- Distribución del estadístico utilizado en el muestreo.
- Distribución de probabilidades en la población.

Si consideramos un cierto carácter, (por ej. estatura, edad, etc.) susceptible de ser medido sobre los individuos de una población, podemos representar este carácter mediante una variable ξ que toma como valores los posibles resultados de dicha medida. ξ es una variable aleatoria, cuyas características deseamos estudiar. A veces conocemos, mediante razonamientos teóricos o por estudios anteriores, el tipo de distribución que sigue la variable bajo estudio (por ejemplo la estatura en un grupo humano se distribuye en forma aproximadamente normal), y deseamos conocer algunas de sus características, tales como la media μ o la desviación típica σ. Otras veces no conocemos el tipo de distribución, y el objeto del estudio es realizar una prueba -contraste de adherencia de ajuste- que permita decidir mediante el estudio de la muestra si la población sigue una distribución de tipo determinado.

Ejemplo 6.3. La superficie de la neurona de la especie Apodemus Sylváticus es una variable aleatoria ξ pues varía de una a otra neurona. Aquí la población bajo estudio es imposible de estudiar en su totalidad, ya que se trata de todas las posibles neuronas de todos los animales de esta especie. Supongamos que tenemos 92 neuronas disponibles para el análisis, y constituyen una muestra aleatoria de la población. El objeto del estudio es inferir las características de la variable en la población a partir del estudio de la muestra disponible.

Una primera distribución en este proceso de muestreo es la de la variable ξ (superficie neuronal en la población). De esta variable sabemos que es continua, y, al igual que otras medidas biológicas, presumiblemente normal. Si aceptamos la hipótesis de normalidad, sabemos que la distribución de ξ esta completamente especificada por su media μ y su varianza σ^2.

Distribución de frecuencias de la muestra particular

Una vez que hemos tomado una muestra de *n* elementos de la población, y medido sobre los mismos el carácter bajo estudio, tenemos una colección de n valores, que pueden ser representados mediante una variable estadística *X*, a la que corresponderá la distribución de frecuencias observada en la muestra particular. En virtud de la ley del azar, estas frecuencias, cuando el tamaño n de la muestra crece, tienden a estabilizarse y tomarán valores próximos a las probabilidades, de modo que el histograma de frecuencias de una muestra puede considerarse como una representación aproximada de la función de densidad de

la variable en la población. De igual modo, las características de la distribución de frecuencias de la muestra, como por ejemplo, la media muestral $\bar{x}$, serán valores aproximados a los correspondientes valores poblacionales o parámetros.

Ejemplo 6.4. La distribución de frecuencias de la superficie neuronal en una muestra de 92 neuronas, puede considerarse una representación aproximada de la función de densidad de la superficie neuronal en la población, por lo que sus características se aproximan a las correspondientes en la población.

Distribución del estadístico en el muestreo

Las características muestrales, cuando se utilizan para aproximar los valores de la población, reciben el nombre de *estadísticos*, y tienen el carácter de variables aleatorias. Por tanto, un estadístico es un valor deducido a partir de la muestra, que se obtiene con objeto de dar un valor para un parámetro desconocido en la población. Así, para estimar la media de una población, calculamos la correspondiente media muestral. Es importante observar que un estadístico es una variable aleatoria. Al haber obtenido la muestra mediante un experimento de azar, no podemos prever de antemano sus elementos, y por tanto, no podemos conocer antes de calcularlo el correspondiente valor del estadístico, que varia de unas muestras a otras. Como veremos más adelante, el conocimiento de la distribución del estadístico es fundamental para el proceso de inferencia.

Si aplicamos un test de inteligencia a una muestra de 500 universitarios obtenida al azar de toda la población universitaria española, podemos calcular la media resultante. Si obtenemos un número infinito de muestras de 500 universitarios, cada una de esas muestras tendrá una media. Entre esas infinitas medias algunas serán iguales, otras diferentes. Si hacemos una distribución de esas medias según su valor, resultará una distribución de medias de muestras, esto es, una distribución muestral de medias.

En general, diremos que la *distribución muestral* de un estadístico es la distribución de frecuencias de los valores que ese estadístico toma en un número infinito de muestras del mismo tipo y tamaño que la primera.

Ejemplo 6.5. Supongamos que tenemos una caja con tres fichas numeradas del 1 al 3. Tomamos al azar dos fichas, con reemplazamiento, y queremos deducir el valor de la media de las tres fichas, mediante la media obtenida en la muestra. En este caso, vemos que la media de la población toma un valor ½ y que la desviación típica es 1/3

Tomemos todas las muestras posibles, y calculemos la media de cada una de ellas:

Datos	Media	Datos	Media	Datos	Media
1,1	1	1,2	1.5	1,3	2
2,1	1.5	2,2	2	2,3	2.5
3,1	2	3,2	2.5	3,3	3

Como podemos observar, $\bar{x}$ es una variable aleatoria. Formemos ahora la distribución de probabilidades del estadístico $\bar{x}$

$\bar{x}$	P $\bar{x}$)	$\bar{x}$ P($\bar{x}$)	$\bar{x}^2$·P($\bar{x}$)
1	1/9	1/9	1/9
1.5	2/9	3/9	4.5/9
2	3/9	6/9	12/9
2.5	2/9	5/9	12.5/9
3	1/9	3/9	9/9
		18/9	39/9

Del examen de la distribución de probabilidad, observamos que el valor más frecuente, que coincide también con el valor medio del estadístico es 2, y corresponde con la media poblacional. Esta es una propiedad deseable, y a los estadísticos que la cumplen los llamaremos insesgados. Asimismo:

$$Var(x)=39/9-4=3/9=1/3$$

Por ello vemos que $Var(x)=\sigma^2/n$. La varianza de un estadístico disminuye, como era de esperar, con el tamaño de la muestra.

Actividades

6.13. Una población consta de los siguientes valores: 7, 14, 13, 10, 8, 4, 2, 10. Construir todas las muestras sin reemplazamiento de dos elementos que pueden hacerse en esta población, calcular la media de cada muestra y representar gráficamente la distribución obtenida.

6.14. En una bolsa hay una bola blanca y dos negras. Se hacen extracciones con reemplazamiento de muestras de tamaño 2. Escribir todas las muestras posibles y la probabilidad de obtener cada una. Hallar la distribución de la proporción muestral.

6.5. DISTRIBUCIÓN DE LA MEDIA EN EL MUESTREO

Como hemos indicado, el conocimiento de la distribución de los estadísticos en el muestreo es imprescindible para la aplicación de las distintas técnicas de inferencia. Esta distribución suele depender de las condiciones del problema de investigación. Vemos que la media de la muestra de valores de una misma población varía de una muestra a otra. Para tratar de estudiar los valores posibles de las medias de todas las muestras de cuatro valores del coeficiente de

inteligencia en el ejemplo 6.1 repetiremos el proceso 30 veces. Pinchando en el icono del Diskette y marcando la opción Guardar en la ventana de entrada de datos, hemos pedido que los resultados de la simulación se graben en las variables que hemos llamado Muestra1, Muestra2..., Muestra30. En la siguiente tabla presentamos los resultados.

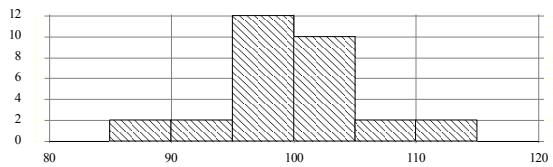
Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5	Muestra 6	Muestra 7	Muestra 8	Muestra 9	Muestra10
118	88	128	105	109	90	97	113	114	91
116	115	81	102	89	103	64	86	83	84
78	89	82	113	106	94	109	70	115	119
120	86	99	120	76	102	120	101	106	101
Muestra11	Muestra12	Muestra13	Muestra14	Muestra15	Muestra16	Muestra17	Muestra18	Muestra19	Muestra20
91	97	103	113	104	105	79	102	93	83
102	94	107	95	118	79	112	93	112	81
104	100	115	85	92	109	120	106	92	116
88	116	116	112	102	120	93	108	108	103
Muestra21	Muestra22	Muestra23	Muestra24	Muestra25	Muestra26	Muestra27	Muestra28	Muestra29	Muestra30
112	101	105	66	70	116	90	101	109	66
106	101	106	96	74	74	78	94	77	81
100	95	77	99	115	82	115	100	110	92
109	98	112	122	87	88	104	98	122	114

Aplicando a estas 30 variables la opción Descripción, Datos Numéricos, Análisis Multidimensional hemos calculado la media de cada una de estas muestras. Obtenemos los siguientes valores:

108, 94.5, 97.5, 110, 95, 97.25, 97.5, 92.5, 104.5, 98.75, 96.25, 101.75, 110.25, 101.25, 104, 103.25, 101, 102.25, 101.25, 95.75, 106.75, 98.75, 100, 95.75, 86.5, 90, 96.75, 98.25, 104.5, 88.25

Actividad

6.15. ¿Cuántos valores del estadístico (media de la muestra) del ejemplo anterior están por encima y por debajo del valor del parámetro (media de la población)? ¿Cuál es el valor máximo y mínimo de todas las medias de las muestras obtenidas? ¿Cuáles son los valores más frecuentes?



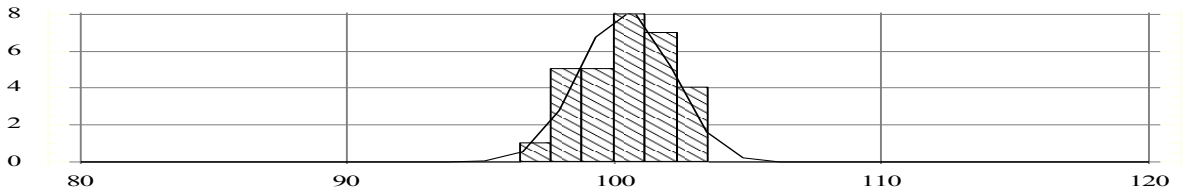
6.16. Hemos grabado los valores de todas estas medias en una nueva columna y hemos representado gráficamente su distribución. Observa los gráficos que hemos obtenido. ¿Piensas que podríamos usar la distribución normal para aproximar los valores de las medias de las muestras? ¿Cuál sería el valor medio de dicha distribución normal?

Hemos visto que dos características importantes de las muestras son su representatividad y variabilidad. Controlamos la representatividad procurando que no haya sesgos en la selección y eligiendo la muestra aleatoriamente, además de tomar un número suficiente de elementos en la muestra. Podemos controlar la variabilidad aumentando el tamaño de la muestra.

Veremos esto si tomamos ahora muestras aleatorias de 100 valores del coeficiente de inteligencia. Nosotros hemos usado el programa Statgraphics para simular 30 muestras, cada una con 100 valores del coeficiente de inteligencia y hemos calculado las medias de cada una de las 30 muestras. Estos son los valores obtenidos:

98.5, 98, 101.72, 98.29, 100.51, 99.75, 102.01, 100.05, 102.28, 98.53, 102.75, 99.52, 99.06, 100.75, 101.85, 101.28, 102, 98.57, 100.25, 103.12, 96.77, 98.78, 102.75, 101.01, 101.75, 101.23, 100.25, 101.84, 97.80, 100.65

En el diagrama hemos representado los resultados de ajustar una curva normal a la columna de datos formada por estas medias.



Actividad

6.17. Compara los gráficos de las medias de las muestras de cocientes intelectuales cuando el tamaño de la muestra es 4 y cuando es 100. ¿Podemos tomar la distribución normal como una buena aproximación para la distribución de las medias muestrales? ¿Cuál será en cada caso, aproximadamente la media de la distribución normal correspondiente? ¿En cuál de las dos distribuciones sería menor la desviación típica? ¿Es más fiable la muestra de cuatro elementos o la de 100 elementos? ¿Cómo podríamos disminuir el error al tratar de estimar la media de la población a partir de la media de una muestra?

6.18. Genera 50 muestras de tamaño 10 de las distribuciones a) uniforme [0,1]; b) Exponencial con $\lambda=3$. Calcula las medias de las muestras y prepara un histograma de las medias muestrales obtenidas. Compara los resultados.

En los ejemplos anteriores hemos visto que cuando la distribución de una variable es normal, y tomamos una gran cantidad de nuestras de valores de dicha variable, las medias de estas muestras también parecen que pueden ser descritas

apropiadamente por una distribución normal. Hemos visto también que, para estimar el valor de la media de una población, utilizamos la media muestral. Se verifica que $E[\bar{x}] = \mu$ y que $Var[\bar{x}] = \sigma^2/n$. Para obtener la distribución de $\bar{x}$ hemos de considerar varios casos.

Población de partida normal con desviación típica conocida

Supongamos que tenemos en una población una variable aleatoria que sigue la distribución normal $N(\mu,\sigma)$. Entonces, si tomamos una muestra aleatoria de n elementos de esta población la media de la muestra $\bar{x}$ sigue una distribución normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Este resultado explica lo que hemos observado en las simulaciones realizadas con Statgraphics, donde veíamos que, al aumentar el tamaño de la muestra, disminuye la dispersión.

Actividades

6.19. Si una variable aleatoria tiene distribución normal $N(100,5)$, ¿Cuál será la distribución de la media de la muestra de 25 elementos? ¿Cuál será la proporción de muestras cuya media estará comprendida entre 99 y 101? ¿Y entre 98,5 y 100,5?

6.20. En una población adulta, el nivel medio de inmunoglobulina medida en mg/100ml es una v. a. normal N(1100,350). Si tomamos 9 personas de esta población y calculamos el nivel medio de inmunoglobulina en esta muestra. ¿Entre que valores cabe esperar que se halle este nivel medio, con probabilidad 0.95? ¿Y si tomamos 100 personas?

6.21. La superficie de las hojas de la planta de berenjenas es de 800 cm^2 con desviación típica de 90 cm^2. Si tomamos una muestra de 100 hojas, ¿cuál es la probabilidad de que la media se sitúe entre 750 y 850 cm^2?

Población de partida normal, con desviación típica desconocida

En realidad es poco frecuente conocer la desviación típica de la población. Si no se conoce, hemos de aproximarla por *S*. Aunque en este caso no conocemos la distribución de $\bar{x}$, si efectuamos el cambio de variable (1), se obtiene una nueva variable aleatoria, que se distribuye con distribución *T* de *n-1* grados de libertad.

Esta distribución, para *n* grande, puede ser aproximada por la normal, como puede comprobarse al comparar los percentiles de las dos distribuciones en las tablas correspondientes.

(6.2) $$T = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

Población de partida no normal, cuando la muestra es grande. Teorema central del límite

Si la población de partida no es normal, en algunos supuestos, se ha deducido la distribución exacta de la media muestral. Sin embargo, en muchas ocasiones es preferible utilizar el siguiente teorema, que fija las condiciones bajo las cuales la media de una muestra tiene una distribución aproximadamente normal.

Supongamos que tenemos una variable aleatoria cuantitativa, con cualquier distribución siendo μ su media y σ su desviación típica valores finitos. Entonces, si tomamos una muestra aleatoria de *n* elementos de esta población la media de la muestra $\bar{x}$ sigue, cuando *n* es suficientemente grande, una distribución normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Este es uno de los teoremas más importantes de la estadística, porque nos permite estimar los valores de la media de una población a partir de una muestra suficientemente grande (en general *n=30* o más elementos).

Como consecuencia practica del teorema anterior, deducimos que, para valores de *n* suficientemente grandes la variable tipificada *Z* definida y la distribución normal *N*(0,1) serán aproximadamente iguales. Por ello podremos utilizar las tablas de la distribución normal para el cálculo de los percentiles de *Z*, siempre que *n* sea suficientemente grande.

Hay que tener en cuenta que, en la aproximación anterior, se supone conocida la desviación típica de la población. Si no es este el caso, es preciso efectuar una aproximación doble: en primer lugar por la distribución normal, y en segundo sustituir σ por *S*, lo que nos conduce a la distribución *T* de Student. Esta utilización de la distribución *T* se ha basado en la aproximación normal previa, y por tanto no puede ser usada con muestras pequeñas, a menos que la distribución de partida fuese normal.

Actividades

6.22. La altura en una población tiene una media de 170 cm y desviación típica de 7 cm. ¿Cuál será la distribución de la media muestral de muestras de tamaño 200?

6.23. En terreno arenoso se plantaron 50 arbolitos de cierto tipo y otros 50 en otra área con terreno arcilloso. Sea *X* = número de árboles plantados en terreno arenoso que sobreviven 1 año e *Y* = número de árboles plantados en terreno arcilloso que sobreviven 1 año. Si la probabilidad de que un árbol plantado en terreno arenoso sobreviva 1 año es 0,7 y la probabilidad de que sobreviva 1 año en terreno arcillos es 0,6, calcule una aproximación a $P(-5 \leq X-Y \leq 5)$.

6.24. Un sistema está formado por 100 componentes cada una de las cuales tiene una confiabilidad igual a 0,95. (Es decir, la probabilidad de que la componente funcione correctamente durante un tiempo específico es igual a 0,95). Si esas componentes

funcionan independientemente una de otra, y si el sistema completo funciona correctamente cuando al menos funcionan 80 componentes, ¿Cuál es la confiabilidad del sistema?

Muestreo en poblaciones finitas

En los casos anteriores hemos supuesto que el método de muestreo es con reemplazamiento, con lo que teóricamente se obtiene una población infinita. La varianza de los estimadores necesita ser corregida cuando efectuamos un muestreo sin reemplazamiento en una población finita de tamaño *N*. Si el tamaño de la muestra es *n* la varianza de la media muestral viene dada por:

(6.3) $$Var(\bar{x}) = \frac{N-n}{N-1} \times \frac{\sigma^2}{n}$$

Si el tamaño de la muestra es pequeño respecto al de la población, el factor $\frac{N-n}{N-1}$ es aproximadamente igual a uno, y el muestreo con y sin reemplazamiento coincide aproximadamente.

Actividades

6.25. De una población que consta de 200 elementos se toma una muestra de 50. Si la varianza de la población es 25, calcular el error de muestreo de la media muestral.

6.26. En la inspección por muestreo de cajas de tornillos, la longitud de los mismos es una variable aleatoria N(5,0.2) cm. Los tornillos se venden en cajas de 25 unidades. Si la longitud media de una caja es mayor de 5.1 cm se rechaza el lote. ¿Cual es la proporción de lotes rechazados?

6.27. Un investigador quiere estimar la media de una población normal, utilizando para ello la media de la muestra. ¿Cual será el tamaño de muestra que debe tomar para que con probabilidad 0.95 la diferencia entre las dos medias sea menor que la décima parte de la desviación típica de la población?

6.6. DISTRIBUCION DE LA CUASIVARIANZA MUESTRAL

Hemos visto que, para estimar la varianza de una población utilizamos la cuasivarianza. Consideraremos dos casos.

Población normal *N(0,1)*

En este caso la cuasivarianza se distribuye como una Chi-cuadrado con *n-1* grados de libertad, al ser suma de *n-1* variables aleatorias normales N(0,1) elevadas al cuadrado:

(6.4) $$\tilde{S}^2 = \frac{1}{n-1}\sum\left(X_i - \overline{X}\right)^2$$

Población normal $N(\mu,\sigma)$ con σ desconocida

En este caso, puede demostrarse que:

(6.5) $$\frac{\tilde{S}^2}{\sigma^2} = \sum\frac{\left(X_i - \overline{X}\right)^2}{\sigma^2}$$

se distribuye según una Chi-cuadrado con *n-1* grados de libertad.

Actividad

6.28. La estatura de un grupo de soldados es una variable aleatoria N(μ, 3). De 100 muestras de 25 soldados cada una ¿En cuantas cabe esperar que la cuasivarianza sea mayor que 15? ¿En cuantas menor que 7?

6.7. DISTRIBUCION DEL ESTIMADOR DE LA PROPORCION EN UNA POBLACION BINOMIAL

Si tratamos de estimar la proporción de valores φ con que en una población se presenta una cierta característica (como la proporción de personas que votan a un partido político) usamos para hacer la estimación la proporción p de valores en una muestra aleatoria. Para muestras suficientemente grandes puede demostrarse que la proporción p en la muestra sigue una distribución normal:

$$N\left(\varphi, \sqrt{\varphi(1-\frac{\varphi}{n})}\right)$$

Es decir, si en una población es φ la proporción de casos que tienen una cierta característica:

- La media de la distribución que sigue la proporción *p* de casos con esa característica en las muestras aleatorias de tamaño *n* es igual a φ.
- La desviación típica de la distribución que sigue la proporción p de casos con esa característica en las muestras aleatorias de tamaño n es igual a:

$$\sqrt{\varphi(1-\frac{\varphi}{n})}$$

- La distribución es aproximadamente normal para muestras de suficiente tamaño.

Actividad

6.29. Supongamos que tomamos muestras aleatorias de recién nacidos. Calcula la desviación típica de las distribuciones en el muestreo de la proporción de niñas, para cada uno de los siguientes valores del tamaño muestral:

n = 50, 100, 200, 400, 500, 800, 1000, 1600, 2000.

a) Construye un diagrama de dispersión de las desviaciones típicas calculadas frente a los tamaños muestrales n.

b) ¿En cuánto tiene que incrementarse el tamaño de muestra para reducir a la mitad las desviaciones típicas?

6.30. Sea p la proporción de votos recibidos por un candidato en unas elecciones. Supongamos que extraemos muestras de 100 votantes y calculamos las proporciones de votos del candidato.

a) Calcula las desviaciones típicas de las distribuciones muestrales de las proporciones calculadas para los siguientes valores de p: 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1

b) Representa mediante un diagrama de dispersión el par de variables: p, desviación típica calculada.

c) ¿Qué valores de p producen máxima variabilidad en las proporciones muestrales? ¿Y la mínima?

6.31. Un cierto tipo de artículo presenta un 25 por ciento de defectos. Si los artículos se venden en cajas de 4 unidades. ¿Cuál será la proporción de cajas con 1, 1, 2, 3, 4 defectos?

6.32. Si lanzamos 500 veces una moneda. ¿Cuál será la probabilidad de obtener una proporción de caras mayor de 0.52?

6.33. El 20 por ciento de una población está expuesta a los efectos de una cierta droga. Si se toman 200 personas de la población, ¿cuál es la probabilidad de encontrar entre el 18 y 22 % de personas sujetas a este efecto?

6.34. Durante cierta epidemia de gripe, enferma el 20 por ciento de la población. En un aula con 100 estudiantes ¿Cuál es la probabilidad de que al menos 30 padezcan la enfermedad?

6.8. DISTRIBUCION DEL ESTIMADOR DEL PARAMETRO EN LA DISTRIBUCION DE POISSON

Hemos visto que esta distribución tiene diversas aplicaciones. Por un lado, como aproximación de la distribución binomial. Por otro, se utiliza en el estudio de fenómenos aleatorios que se producen a lo largo del tiempo, o en el estudio de las distribuciones espaciales.

Consideremos un fenómeno que pueda ser descrito mediante una distribución de Poisson de parámetro λ. Como vimos al efectuar el estudio de la misma, dicho barómetro coincide con la media y varianza de la variable.

Un estimador posible para el parámetro vendrá dado por tanto por la media muestral:

(6.6) $$\lambda = \frac{x_1 + x_2 + \ldots x_n}{n}$$

La media de este estimador viene dada por:

(6.7) $$E\left[\lambda\right] = E\left[\frac{x_1 + x_2 + \ldots x_n}{n}\right] = \frac{n\lambda}{n} = \lambda$$

y es, por consiguiente, un estimador insesgado del parámetro . Puesto que $Var[\xi] = \lambda$, se verifica:

(6.8) $$Var\left[\lambda\right] = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$$

Puede demostrarse, utilizando la aproximación normal a la distribución de Poisson que, para valores de n suficientemente grandes, la distribución de es aproximadamente normal $N(\lambda , \lambda /n)$.

Actividad

6.35. El número de gotas de grasa en las células hepáticas de ratas hembra sometidas a un cierto tipo de dieta sigue una distribución de Poisson, con media 1,39 gotas. Si tomamos una muestra de 400 células de este tipo, ¿Cuales serán los límite entre los que oscilará el número medio de gotas de grasa en la muestra, con probabilidad 0,99?

6.36. El número de erratas en las páginas de un texto sigue la distribución de Poisson. Se examinan 40 páginas y se encuentra un total de 60 erratas. ¿Cuál sería el mejor estimador del número medio y varianza de erratas por páginas en el libro? Si el libro tiene 200 páginas, ¿cuántas erratas habría que esperar?

TEMA 7.
INTERVALOS DE CONFIANZA

7.1. INTRODUCCIÓN

En las lecciones anteriores hemos aprendido a predecir qué valor obtenemos para un estadístico (por ejemplo, para la media de una muestra) si conocemos el valor del parámetro (por ejemplo, si conocemos el valor de la media en la población). Sin embargo, lo que de verdad interesa en la práctica es lo contrario: Estimar el valor del parámetro en la población si conocemos el valor del estadístico en la muestra. Por ejemplo nos preguntamos:

- En un estudio médico sobre los efectos secundarios de un cierto analgésico, 23 pacientes de los 440 que siguieron un tratamiento con dicho analgésico tuvieron efectos secundarios. ¿Entre qué límites podemos estimar la proporción de enfermos que es propensa a tener efectos secundarios al tomar este analgésico?
- Si un fabricante vende paquetes de azúcar de 1 kilo y, al realizar un control de calidad y observar el peso medio de 100 paquetes observa que el peso medio es de 1050 grs. ¿Será que el proceso de llenado se ha descontrolado y está vendiendo más peso del exigido? (El fabricante sabe que la desviación típica debería ser de 80 grs).

Actividad

7.1. Supongamos que en una encuesta a 2500 personas el 36 % declara estar a favor de las medidas económicas del gobierno. ¿Cuál será el valor aproximado del % de personas a favor del gobierno en la población? ¿Cuál será aproximadamente la desviación típica de la distribución en el muestreo de la proporción de votantes en todas las muestras de 2500 personas?

Como vimos en el capitulo anterior, en la estimación por punto, cualquier parámetro desconocido se estima mediante un valor único. Una estimación de este tipo no es, en general, satisfactoria en los problemas prácticos. Es necesario obtener una medida de la precisión del estimador utilizado. Para ello, puede emplearse, en primer lugar, el error de muestreo que, al ser la desviación típica de la distribución muestral del estadístico, da una medida de su variabilidad. Otro enfoque posible es la construcción de un intervalo de confianza. Para ello, si θ es un estimador de un parámetro desconocido θ é intentamos hallar dos números positivos δ y ε tales que podamos asegurar se verifica la relación (7.1).

(7.1) $$P(\theta-\delta<\theta<\theta+\delta)=1-\varepsilon$$

Para una probabilidad dada 1- ε, una gran precisión del estimador estará asociada con valores pequeños de ε. En términos mas generales, la teoría de la estimación por intervalos intenta, para cada parámetro desconocido θ hallar dos funciones de los valores muestrales θ_1 y θ_2 tales que se cumpla la relación (7.2).

(7.2) $$P(\theta_1<\theta<\theta_2)=1-\varepsilon$$

Cuando exista tal intervalo, se denominar intervalo de confianza del parámetro é y la probabilidad 1- ε se llamará *coeficiente de confianz*a del intervalo.

Actividades

7.2. Indica cuál de las siguientes afirmaciones se cumple en un intervalo de confianza:

a. De una muestra a otra, el intervalo es constante
b. Se especifica un rango de valores dentro de los cuales supuestamente cae el parámetro con seguridad
c. Indica un intervalo de posibles valores para el parámetro, y un porcentaje de intervalos que cubrirán, aproximadamente dicho valor, para el mismo tamaño de muestra
d. Siempre contienen el parámetro poblacional

El propósito de un intervalo de confianza es estimar un parámetro

desconocido con indicación de la precisión de la estimación y del grado de confianza que tenemos en la estimación. Cuando calculamos un intervalo de confianza damos dos informaciones:

- Un *intervalo* de valores, calculado a partir de los datos
- Una *probabilidad o nivel de confianza.* En los ejemplos anteriores hemos usado el nivel de confianza del 95%, pero podríamos cambiarlo por otros valores. Es decir, para el caso de la proporción y el nivel de confianza del 99% podríamos asegurar que:

$$P(p- 3\sqrt{p(1-p)/ n} \leq \varphi \leq p+ 3\sqrt{p(1-p)/ n}) = 0,99$$

Actividad

7.3 ¿Entre qué valores podemos afirmar que se encontrará la proporción de personas favorables a la política económica del gobierno en el ejemplo anterior con una confianza del 99 %?

Vemos que la amplitud del nivel de confianza depende de los siguientes factores:

- El nivel de confianza
- El tamaño de la muestra
- La variabilidad en la población

Actividad

7.4. Discutir si el intervalo de confianza crece o decrece al aumentar cada uno de los factores anteriores.

7.2. INTERVALO DE CONFIANZA PARA LA MEDIA

En una distribución normal el 95% de los casos se encuentran a una distancia 2σ de la media. Si μ es la media de una población y σ su desviación típica, la media muestral $\bar{x}$ sigue una distribución aproximadamente normal $N(\mu,\sigma/\sqrt{n})$ siendo n el tamaño de la muestra para valores de n suficientemente elevados. Por ello, en el 95% de las muestras

la media muestral $\bar{x}$ estará a una distancia $2\sigma/\sqrt{n}$ de la verdadera media μ en la población. Recíprocamente, podemos deducir que el 95% de las muestras la media μ en la población estará dentro del intervalo $\bar{x} \pm 2\sigma/\sqrt{n}$. Este es el intervalo de confianza del 95%.

Por tanto, si $\bar{x}$ es el valor obtenido para la media en una muestra de tamaño *n*, y μ es el valor desconocido de la media en la población, y usando los intervalos en que se encuentran el 95% y 99% de casos en la distribución normal, podemos afirmar:

$P(\bar{x} - 2\sigma/\sqrt{n} \leq \mu \leq \bar{x} \pm 2\sigma/\sqrt{n}) = 0{,}95$ (Intervalo de confianza del 95%)

$P(\bar{x} - 3\sigma/\sqrt{n} \leq \mu \leq \bar{x} \pm 3\sigma/\sqrt{n}) = 0{,}99$ (Intervalo de confianza del 99%)

Ejemplo 7.2. En Stagraphics es posible calcular intervalos de confianza para las medias dentro de la opción DESCRIPCIÓN – DATOS NUMÉRICOS – ANÁLISIS UNIDIMENSIONAL (Opciones tabulares, Figura 7.1).

Figura 7.1.

Hemos usado esta opción para estimar el tiempo medio de respuesta a un estímulo en una población de adultos, dándoles los datos de un fichero que contiene una muestra de 96 adultos. El tiempo medio de reacción en esta muestra fue 2.5 segundos. A continuación incluimos los resultados obtenidos. El programa produce por defecto los intervalos de confianza del 95% y el 99%. Si queremos obtener otros intervalos de confianza, habrá que modificar el nivel de confianza mediante OPCIONES DE VENTANA.

```
Intervalo de confianza para tiempo

Intervalo de confianza del 95,0% para la media: 2,5 +/-0,22724 (2,27228,2,727772)

Intervalo de confianza para la desviación típica: 0.984303. 1.31001)
```

Es importante resaltar que estos programas calculan los intervalos de confianza para la media, incluso cuando no conocemos el verdadero valor de la desviación típica en la población. La desviación típica en la población es estimada a partir de los datos, mediante la fórmula: $\sigma \sim s\, n/(n-1)$, siendo s la desviación típica de la muestra.

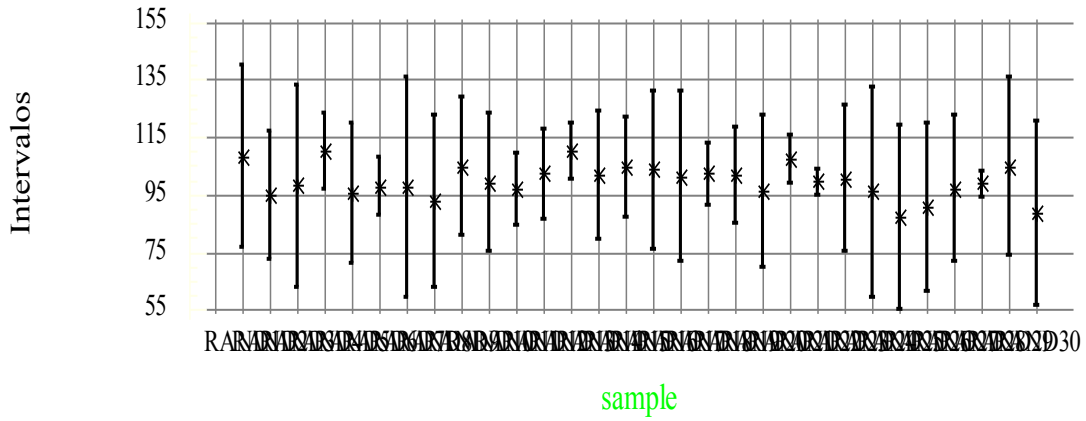
*Observación***.** Puesto que la media de la muestra varía de una muestra a otra, los intervalos de confianza variarán de una muestra a otra (lo mismo ocurre con la proporción). Lo que nos dice el coeficiente de confianza es que en un porcentaje dado de muestras, el verdadero valor del parámetro estará incluido en el intervalo.

Ejemplo 7.3. Mediante el programa DESCRIPCIÓN, DATOS NUMÉRICOS, ANÁLISIS MULTIDIMENSIONAL y tomando INTERVALOS DE CONFIANZA en las opciones tabulares hemos calculado los intervalos de confianza del 95% para las 30 muestras que se generaron en el ejemplo 2, obteniendo los siguientes resultados, donde se puede observar la variación de los intervalos de confianza (Recordemos que la media de esta población era igual a 100).

	Media	L. Inferior	L. Superior
MUESTRA1	108.428	76.707	140.148
MUESTRA2	94.9871	72.8822	117.092
MUESTRA3	98.112	63.0043	133.22
MUESTRA4	110.262	97.3364	123.188
MUESTRA5	95.7421	71.4131	120.071
MUESTRA6	97.795	87.729	107.861
MUESTRA7	97.9558	59.688	136.224
MUESTRA8	92.7246	62.7896	122.66
MUESTRA9	104.873	80.9692	128.777
MUESTRA10	99.1524	75.2227	123.082
MUESTRA11	96.8093	84.2132	109.405
MUESTRA12	102.487	86.824	118.15

MUESTRA13	110.352	100.431	120.272
MUESTRA14	101.743	79.6673	123.819
MUESTRA15	104.55	87.3908	121.71
MUESTRA16	103.807	76.3217	131.291
MUESTRA17	101.361	71.8713	130.85
MUESTRA18	102.516	91.7094	113.322
MUESTRA19	101.979	85.5198	118.438
MUESTRA20	96.3584	69.8985	122.818
MUESTRA21	107.135	98.8507	115.418
MUESTRA22	99.4508	95.117	103.785
MUESTRA23	100.608	75.2051	126.01
MUESTRA24	96.0852	59.4635	132.707
MUESTRA25	87.0429	55.0062	119.08
MUESTRA26	90.5363	61.2778	119.795
MUESTRA27	97.3316	71.6729	122.99
MUESTRA28	98.8195	94.2523	103.387
MUESTRA29	104.731	73.8401	135.622
MUESTRA30	88.6325	56.5146	120.75

La variación de los intervalos se ve mejor en el siguiente gráfico, aunque en este caso particular todos los intervalos cubren el verdadero valor del parámetro (100).



Actividades

7.5. Si un fabricante vende paquetes de azúcar de a kilo y, al realizar un control de calidad y observar el peso medio de 100 paquetes observa que el peso medio es de 1050 grs. Calcula el intervalo de confianza del peso medio real de los paquetes ¿Será que el proceso de llenado se ha descontrolado y está vendiendo más peso del exigido? (El fabricante sabe que la desviación típica debería ser de 80 grs).

7.6. Comparado a los intervalos de confianza calculados en muestras de tamaño n=4, el ancho de los intervalos de confianza de la media de la población calculado en muestras de tamaño $n = 50$:

a. Variará más que los anchos de los intervalos para muestras de tamaño n =4.
b. Variará un poco, pero no tanto como lo hicieron los anchos de los intervalos para muestras de tamaño n =4.
c. Tomarán valores parecidos.

7.7. ¿Cómo cambia el ancho del intervalo para la media si, manteniendo todos los datos fijos se reduce la varianza?

CASO A: Población normal, o muestras grandes, con desviación típica conocida

Si extraemos, al azar gran número de muestras de una población normal, de desviación tipica conocida σ y en cada una de ellas calculamos el estadístico $\bar{x}$, estas medias muestrales variarán entre sí, pero todas tenderán a agruparse alrededor de la verdadera media μ de la población. Si representamos gráficamente la distribución de dichos valores de $\bar{x}$, obtendremos una curva normal.

La desviación típica de esa distribución normal de las medias de las muestras será $\sigma_{\bar{x}} = \sigma / \sqrt{n}$. En consecuencia, cuanto mayor sea el tamaño de la muestra, menor será la desviación típica en la distribución de la media muestral.

El problema principal de la estadística inferencial o muestral es poder concretar lo más fielmente, a partir de los estadísticos calculados en un sola muestra, cual puede ser el valor más verosímil del parámetro de la población, o al menos entre qué límites se encuentra tal parámetro.

El problema se plantea partiendo del hecho de que sólo disponemos de los estadísticos de una sola muestra y no de infinitos estadísticos obtenidos de infinitas muestras. Es decir, si de toda una población elegimos al azar una muestra de n datos y calculamos la $\bar{x}$, ¿qué podemos decir acerca de la media verdadera de la población? El conocimiento de la distribución muestral de un estadístico nos permite dar respuestas útiles a esta pregunta.

Niveles de confianza y coeficientes de riesgo

Si tenemos en una bolsa 99 bolas blancas y una negra, podemos pronosticar que al sacar una bola al azar será blanca. No lo diremos con absoluta seguridad, sino con cierto riesgo de equivocarnos. Si hacemos muchos pronósticos de este tipo, nos equivocaremos a la larga, el 1 por

ciento de las veces. En global, tenemos una probabilidad del 99 por ciento de acertar y del 1por ciento de errar. Cuando hacemos esa clase de juicios, decimos que procedemos al *nivel de confianza* del 99 por ciento o con el coeficiente de riesgo del 1 por ciento. Si hubiera 95 bolas blancas y 5 negras, al afirmar que saldrá al azar una blanca emitiremos un juicio al nivel de confianza del 95 por ciento, o con el coeficiente de riesgo del 5 por ciento.

De igual manera, sabemos que en una distribución muestral normal, el 95 de cada 100 medias de las muestras elegidas al azar se encontrarán entre $1.96\,\sigma_{\bar{x}}$ por debajo y $1.96\,\sigma_{\bar{x}}$ por encima de la media μ

Ocurre, sin embargo, que no conocemos μ. Ahora bien, si decimos que la media $\bar{x}$ del 95 por ciento de las muestras no se apartará de la media verdadera de la población en más de $1.96\,\sigma_{\bar{x}}$, con 95% de probabilidades de confianza, recíprocamente también podemos decir, que la verdadera media de la población μ no se apartará de la media $\bar{x}$ de la muestra en más del $1.96\,\sigma_{\bar{x}}$ en el 95 por ciento de las muestras. Es decir:

$$P(\bar{x} - 1{,}96\sigma_{\bar{x}} < \mu < \bar{x} - 1{,}96\sigma_{\bar{x}}) = 0{,}95$$

Si tomamos el nivel de confianza del 99 por ciento, diremos entonces que la media de la población μ no se apartará de la x en más de $2.58\,\sigma_{\bar{x}}$ en el 99 por ciento de las muestras, ya que el error muestral máximo entre $\bar{x}$, y μ según la distribución muestral normal no puede ser mayor que $2.58\,\sigma_{\bar{x}}$. Es decir:

$$P(\bar{x} - 2{,}58\sigma_{\bar{x}} < \mu < \bar{x} - 2{,}58\sigma_{\bar{x}}) = 0{,}99$$

Para un coeficiente de confianza *1-ε* cualquiera, el correspondiente intervalo de confianza para la media de la población viene dado por la expresión (7.3), donde Z_ε es el percentil del (1- ε/2)100% de la curva normal. De dicha expresión pueden deducirse las siguientes propiedades del intervalo de confianza:

- La amplitud del intervalo disminuye al aumentar el tamaño de la muestra.
- Para un mayor nivel de confianza, se precisa un intervalo mayor.
- Para un mismo nivel de confianza y tamaño muestral, en poblaciones

mas homogéneas se obtiene un intervalo de confianza más preciso.

(7.3) $$(\bar{x} - Z_\varepsilon \sigma_{\bar{x}} < \mu < \bar{x} - Z_\varepsilon \sigma_{\bar{x}})$$

Cuando la población de partida no es normal, pero el tamaño de la muestra es lo suficientemente grande (n>30), puede utilizarse la expresión (7.3) para el cálculo del intervalo de confianza de la media, que en este caso tiene el carácter de aproximado.

EJEMPLO 7.3. La eliminación total por día de una cierta sustancia es una variable aleatoria normal, con desviación típica 1.9 mg. En una muestra de 100 personas se obtuvo una eliminación media de 12 mg. Calcularemos el intervalo de confianza de la media de esta población, para un coeficiente de confianza del 95%.

Puesto que, en este caso la media muestral tiene una distribución normal $N(\mu,0{,}19)$, y para $1\text{-}\varepsilon$=0.95 al percentil correspondiente de la distribución normal es $Z_\varepsilon = 1{,}96$ el intervalo pedido viene dado por:

$$12 \pm 1{,}96*0.19=(11{,}63\text{-}12{,}37)$$

Tamaño necesario de muestra para una precisión dada

Un problema de gran trascendencia en el diseño de un estudio estadístico es la determinación del tamaño de la muestra. Mientras que una muestra demasiado pequeña hace inútil el estudio, al no aportar la información deseada, una muestra excesiva es muy costosa, y quizás innecesaria, puesto que con menos elementos se consigue una precisión suficiente en las estimaciones. Afortunadamente, en muchos casos es posible determinar de antemano el tamaño óptimo de la muestra. En el caso que nos ocupa, al venir dada la precisión del estimador por:

$$\delta = Z_\varepsilon \, \sigma / \sqrt{n}$$

y siendo σ conocida, podemos disminuir δ y, por tanto, aumentar la precisión tanto como deseemos aumentando el tamaño muestral. Para conseguir un valor δ dado, basta tomar:

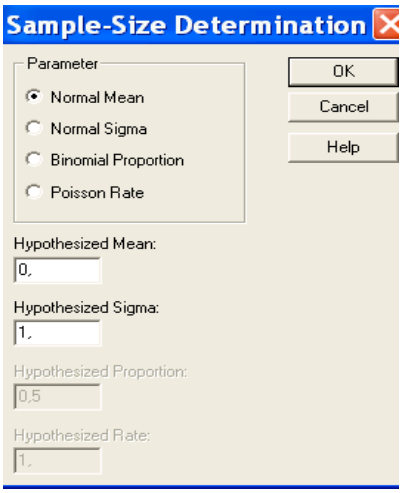
$$n > (Z_\varepsilon \sigma / \delta)^2$$

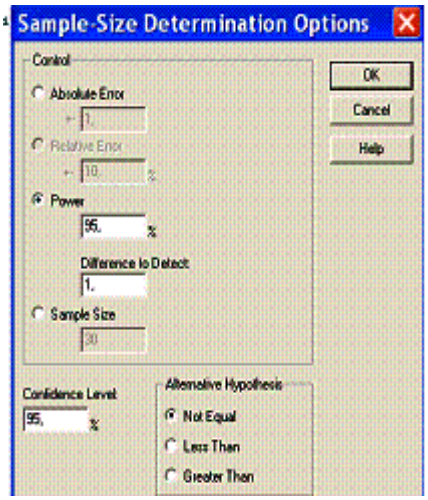
EJEMPLO 7.4. Si, en el ejemplo 7.3 queremos que el error δ=0,1 para el intervalo de confianza del 95%, basta tomar:

$$n>(1,96*1,9/0.1)^2=1386,81$$

es decir, para tamaños de muestra de 1387 o más elementos, podemos asegurar que δ será igual o menor que 0,1.

En Statgraphics es posible calcular los tamaños de muestra, dentro del programa opción DESCRIPCIÓN –DETERMINACIÓN DEL TAMAÑO DE MUESTRA, donde hay varias opciones (media o desviación típica de la distribución normal, parámetros de la distribución binomial o Poisson. Dependiendo del parámetro a estimar se piden los valores supuestos.





A continuación se abre otra ventana donde se puede controlar el error absoluto o la potencia. Si se elige controlar el valor absoluto se puede variar el coeficiente de confianza y elegir entre intervalo bilateral o unilateral (en este caso superior o inferior). Si se elige controlar la potencia, caso se puede variar el valor de la potencia y el tamaño del efecto, así como el valor del coeficiente de confianza. Un ejemplo de salida se obtiene a continuación

```
Sample-Size Determination
-------------------------

Parameter to be estimated: normal mean
Desired tolerance: +- 1,0
Confidence level: 95,0%
Assumed sigma: 1,0

The required sample size is n=7 observations.
```

Actividades

7.8. Hemos calculado un intervalo de confianza al 95% basado en el valor medio $\bar{x}$ obtenido de una muestra de 10 casos. Si incrementamos el tamaño de la muestra a 1000, y calculamos un segundo intervalo al 95 % de confianza, ¿debemos tener más o menos confianza en el resultado? ¿tendremos más o menos precisión?

7.9. Construya un intervalo de confianza al 95% para la media de una población normal de desviación típica σ desconocida si en una muestra de tamaño 10, la media de la muestra es $\bar{x}$ =25 y la estimación de la desviación típica en la muestra es $s = 6$.

7.10. Se sabe que el contenido de grasa de una magdalena sigue una distribución normal, cuya varianza es conocida, teniendo un valor de 0,25 gr. Se desea estimar el valor de la media poblacional con un error máximo de 0,2gr. y una confianza del 95%. ¿Cuál ha de ser el tamaño de la muestra?

7.11. La media de 100 estudiantes en una prueba fue de 6,5. Encuentre el intervalo de confianza al 95% para la media de la población asumiendo que $\sigma=0,7$.

7.12. El propietario de una tienda desea estimar el número promedio de envases vendidos por día. Una muestra aleatoria de 25 días dio un valor medio de 100 envases. La desviación estándar de la población es $\sigma=15$. Calcule el límite superior para un intervalo de confianza al 95%

7.13. Se ha tomado una muestra aleatoria de 100 individuos a los que se ha medido el nivel de glucosa en sangre, obteniéndose una media muestral de 110 mg/cc. Se sabe que la desviación típica de la población es de 20 mg/cc. Obtén un intervalo de confianza, al 90%, para el nivel de glucosa en sangre en la población ¿Qué error máximo se comete con la estimación anterior?

7.14. En una central telefónica se seleccionan 150 llamadas, observándose que el tiempo medio que tardan en descolgar el teléfono los receptores era de 2 segundos, con una desviación típica de 0,61 sg. Se pide, para un nivel de confianza de al menos el 99%, obtener un intervalo de confianza para el tiempo medio que tardan los usuarios en descolgar el teléfono, suponiendo que la desviación típica poblacional es 0,6.

CASO B: Población normal o muestras grandes, con desviación típica desconocida

En general, cuando se desea estimar la media de una población, no se conoce tampoco la desviación típica de la misma. Por ello, la aproximaremos mediante el valor *S* calculado en la muestra. La variable aleatoria:

$$T = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

sigue una distribución *T* con n-1 grados de libertad. El intervalo de confianza de coeficiente de confianza 1-ε para la media de la población vendrá dado por (7.4).

(7.4) $$\bar{x} - \frac{t_\varepsilon S}{\sqrt{n}} \leq \mu \leq \frac{t_\varepsilon S}{\sqrt{n}}$$

En esta expresión, t_ε es el percentil del (1- $\varepsilon/2$)100% de la distribución *T* con *n-1* grados de libertad. Puesto que, al crecer n, esta distribución se aproxima a la normal *N(0,1),* para tamaños suficientes de muestra puede utilizarse esta distribución, en lugar de la *T* de Student. Algunos autores recomiendan tomar valores de n al menos iguales a 30 para utilizar la distribución normal en este caso. Personalmente recomendamos siempre que sea posible utilizar la distribución *T*, y al menos obtener una muestra de 60 elementos antes de sustituirla por la distribución normal.

EJEMPLO 7.5. Se desea obtener una estimación del tiempo de latencia medio en un test MFF-20, con un tamaño de la muestra 58 (mayor de 30). El programa Statgraphics ha proporcionado la salida se muestra a continuación

---

CALCULO DE INTERVALOS DE CONFIANZA

MEDIA DE LA VARIABLE TIEMPO=12.84483

ERROR DE MUESTREO=1.412031

INTERVALO DE CONFIANZA DEL 95%=(10.01654 - 15.67311)

INTERVALO DE CONFIANZA DEL 99%=(9.079146 - 16.61051)

Como estimación puntual, hemos obtenido un tiempo medio de 12,85. Puesto que el error de muestreo $S/\sqrt{n}=1.412$, el intervalo de confianza se obtiene sumando y restando a la media muestral el producto de este error de muestreo por el correspondiente valor de la distribución *T* de 57 grados de libertad, esto es $T=2,00279$ y $T=2,66555$ respectivamente.

Los intervalos correspondientes de confianza, muestran la poca precisión de la estimación, debida a la variabilidad de los datos.

Nótese que, de haber utilizado la aproximación normal, se hubiera sustituido el valor *T=2,00279* por el *Z=1,96* y el *2,66555* por el *2,57* respectivamente.

EJEMPLO 7.6. Hemos usado el programa Statgraphics para estimar el peso medio de los alumnos de una Facultad. Los resultados se muestran a continuación. En este caso obtenemos mayor precisión. Puede tambien observarse que, en este caso, al aumentar el valor de *n* y, por tanto los grados de libertad, los valores de *T* utilizados son más aproximados a los correspondientes de la distribución normal.

CALCULO DE INTERVALOS DE CONFIANZA

MEDIA DE LA VARIABLE PESO=59.58333

ERROR DE MUESTREO=1.15159

INTERVALO DE CONFIANZA DEL 95%=(57.27841 - 61.88826)

INTERVALO DE CONFIANZA DEL 99%=(56.51581 - 62.65086)

Intervalos unilaterales de confianza

En algunos problemas reales estamos interesados en hallar tan sólo un límite superior o inferior para un parámetro de la población. Supongamos, por ejemplo, que un fabricante de vigas desea estimar la resistencia media de las mismas a la rotura. Puesto que, ordinariamente, no importará que el material sea muy resistente, puede estar interesado en hallar el límite inferior de la resistencia media, con un coeficiente de confianza del 95%, es decir un valor *a*, tal que $P(\mu>a)=0,95$. Este límite se obtendrá con la expresión siguiente:

$$a = \bar{x} - T_{0,95} S / \sqrt{n} < \mu$$

donde $T_{0,95}$ es el percentil del 95% en la distribución *T* con *n-1* grados de libertad. De forma análoga se halla un intervalo de confianza para el límite superior de la resistencia media.

Actividades

7.15. Un fabricante vende botellas que supuestamente tienen un litro de aceite. Al tomar una muestra de 16 botellas se determinó que en promedio contenían 0,94 litros, con desviación estándar 0,097. Construir un intervalo de confianza al 95 %, para el verdadero contenido promedio del envase. No se conoce la desviación típica de la población.

7.16. Supongamos que el cociente intelectual de un gran número de niños puede considerarse normalmente distribuido. Una muestra de 25 niños, dió un valor medio 114,5 y un valor *S*=12,1. Hallar un intervalo de confianza del 99% para el cociente intelectual medio de dicha población.

7.17. La superficie media en células hepáticas en la zona portal de 200 ratas fue 467,06 micras cuadradas, y el error de muestreo obtenido 25.,7. Hallar un intervalo de confianza del 99% para la superficie media de dichas células en la población.

7.18. Dos muestras diferentes se toman de una población donde la media poblacional y la desviación estándar poblacional son desconocidas. La primera muestra tiene 36 datos, y la segunda muestra 100 datos. Se construye un intervalo de confianza de 95% para cada muestra para estimar la media poblacional. Que intervalo de confianza esperaría que tenga mayor precisión?

7.19. Se han obtenido los siguientes datos de emisión diaria de óxidos de azufre, para una muestra de tamaño n=100, media: $\bar{x}$ =18 y cuasivarianza s^2=36. Elabore un intervalo de confianza de 95% para la verdadera emisión diaria promedio de óxidos de azufre.

7.20. Un estudiante de economía toma una muestra de 36 compañías a través de los Estados Unidos. Imagine que el salario medio ofrecido por esas 36 compañías es de 30000 dólares con una desviación estándar de 20000 dólares. Obtener un intervalo de confianza al 95% para el verdadero salario medio.

7.21. La media de edad de los alumnos de una clase es 18,1 años, y la desviación típica 0,6 años. ¿Qué tamaño debe tener una muestra de dicha población para que su media esté comprendida entre 17,9 y 18,3 años, con una confianza del 99,5%?

7.5. INTERVALO DE CONFIANZA PARA LA VARIANZA DE UNA POBLACION NORMAL

Anteriormente vimos que si una variable aleatoria tiene distribución normal N(μσ), la magnitud aleatoria:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

se distribuye según una Chi-cuadrado con *n-1* grados de libertad. Esta distribución no es simétrica y toma sólo valores positivos. Por este motivo,

la construcción del intervalo de confianza para la varianza, es algo diferente de las mostradas en la sección anterior. Supongamos que queremos determinar las cantidades positivas a y b tales que, para un nivel dado de confianza (1- ε) se verifique:

$$P(a \le \sigma^2 \le b) = 1 - \varepsilon$$

Esto es equivalente a que se cumpla:

$$1 - \varepsilon = P(\frac{1}{b} \le \frac{1}{\sigma^2} \le \frac{1}{a}) = P(\frac{(n-1)S^2}{b} \le \frac{(n-1)S^2}{\sigma^2} \le \frac{(n-1)S^2}{a}$$

Pero $\frac{(n-1)S^2}{b}$ se distribuye como una Chi-cuadrado con n-1 grados de libertad. Para determinar los valores a y b de los extremos del intervalo, basta despejar a y b en las igualdades (7.5) y (7.6).

(7.5) $$\chi^2_{\varepsilon/2} = \frac{(n-1)S^2}{b}$$

(7.6) $$\chi^2_{1-\varepsilon/2} = \frac{(n-1)S^2}{a}$$

En dichas expresiones $\chi^2_{\varepsilon/2}$ y $\chi^2_{1-\varepsilon/2}$ son los percentiles del *ε*100/2 %* y del *(1- ε/2)*100%* de la distribución Chi-cuadrado con *n-1* grados de libertad. Aunque S^2 es un estimador insesgado de la varianza, no ocurre lo mismo con *S* respecto a σ Se obtiene un ligero sesgo que decrece con el valor de *n.* Sin embargo, si se desea calcular el correspondiente intervalo de confianza para la desviación típica de la población, basta obtener la raiz cuadrada de los extremos del intervalo de confianza para la varianza.

EJEMPLO 7.7. En la figura (9.5) se muestra la salida de un programa que efectua estimacion y contraste de hipótesis para la varianza de una población. Se ha utilizado como variable el peso de los alumnos.

Puesto que *n=60* y *S=79.57*, para el intervalo de confianza del 95% los extremos requeridos son:

b=59*79,57/39,65=118,4015

a=59*79,57/82,1274=57,16277

Extrayendo las raices cuadradas de a y b se obtiene:

$$\sqrt{b}=10{,}88125 \text{ y } \sqrt{a}=7{,}560606,$$

que constituyen los extremos del intervalo de confianza del 95% para la desviación típica de la población.

CALCULO DE INTERVALOS DE CONFIANZA

CUASIVARIANZA DE LA VARIABLE PESO=79,56956

GRADOS DE LIBERTAD 59

INTERVALO DE CONFIANZA DEL 95% VARIANZA= (57,16-118,40)

INTERVALO DE CONFIANZA DEL 99% VARIANZA= (57,71-135,19)

INTERVALO DE CONFIANZA DEL 95% (D. TIPICA)=(7,56-10,88)

INTERVALO DE CONFIANZA DEL 99% (D. TIPICA)= (7,19-11,63)

Actividades

7.22. Al examinar 1.200 células hepáticas, se obtuvo un error de muestreo igual a 3.9 al estimar la media de la población. Calcular un intervalo de confianza del 99% para la varianza y desviación típica de dicha variable.

7.23. En una muestra de 26 elementos se obtuvo un valor S=5. Hallar un intervalo de confianza del 95% para la varianza de dicha población.

7.24. Sea σ^2 la varianza de la distribución de la tensión en un dispositivo. El valor calculado de la varianza muestral es s^2=13700, n=16. Calcular el intervalo de confianza de 95% para σ.

7.25. La cantidad de dióxido de carbono (CO_2) liquido presente en un proceso inclusión geológico en cinco días distintos en una roca cristalizada tuvo una varianza muestral igual a 80Haga una estimación de la precisión de la técnica LRM estableciendo un intervalo de confianza de 99% para la variación en las mediciones de concentración de CO_2

7.6. INTERVALO DE CONFIANZA PARA LA PROPORCION EN UNA POBLACION BINOMIAL

En una distribución normal el 95% de los casos se encuentran a una distancia 2σ de la media. Sabemos que la proporción muestral p sigue una distribución aproximadamente normal $N(\varphi, \sqrt{\varphi(1-\varphi/n)})$, siendo φ la proporción en la población. Por ello, en el 95% de las muestras la proporción muestral p estará a una distancia $2\sqrt{\varphi(1-\varphi)}$/n de la verdadera

proporción φ en la población.

Recíprocamente, podemos deducir que el 95% de las muestras la proporción φ en la población estará dentro del intervalo $p \pm 2\sqrt{p(1-p)/n}$. Este es el intervalo de confianza del 95%.

Por tanto, si p es el valor obtenido para la proporción en una muestra de tamaño n, y φ es el valor desconocido del parámetro en la población, y usando los intervalos en que se encuentran el 95% y 99% de casos en la distribución normal, podemos afirmar que:

$$P(p - 2\sqrt{p(1-p)/n} \leq \varphi \leq p + 2\sqrt{p(1-p)/n}) = 0.95$$

Ejemplo 7.1. Si en una caja de 100 bombillas el 10% son defectuosas. ¿Entre qué límites varía la proporción de defectos en la población con una confianza del 95%?

Puesto que p=0,1, y n=100, $\sqrt{p(1-p)/n}$=√0,1*0,9/100=0,03. El intervalo de confianza del 95% será (0.1-0,03, 0,1+0,03)=(0,07, 0,13).

En resumen, cuando se dispone de un tamaño de muestra suficiente (n>30), la distribución de la proporción muestral es aproximadamente normal $N(p, \sqrt{pq/n})$. Podemos pues, en estas condiciones, construir un intervalo aproximado de confianza basado en la distribución normal, que viene dado, para un coeficiente de confianza $1-\varepsilon$ por la expresión (7.7).

(7.7) $$p - Z_\varepsilon\sqrt{pq/n} \leq p \leq p + Z_\varepsilon\sqrt{pq/n}$$

donde Z_ε es el percentil del $(1-\varepsilon/2)100\%$ de la distribución normal $N(0,1)$.

Una dificultad que se plantea en la construcción de este intervalo, es que el valor p que aparece en la expresión del error de muestreo del estimador, es precisamente el valor que tratamos de estimar y por tanto es desconocido. Como cabe pensar que dicho valor será próximo al obtenido en la muestra, se toma, para $n>80$, como intervalo de confianza el (7.8).

(7.8) $$p - Z_\varepsilon\sqrt{pq/n} \leq p \leq p + Z_\varepsilon\sqrt{pq/n}$$

EJEMPLO 7.8. A continuación se muestra la salida de un programa que realiza el cálculo de intervalos de confianza. En este caso, se desea estimar la proporción de alumnos que fuman, considerando los datos disponibles como una muestra de los alumnos de su misma especialidad.

CALCULO DE INTERVALOS DE CONFIANZA

PROPORCION DE PACIENTES OBSERVADOS EN LA VARIABLE FUMA= ,6

INTERVALO DE CONFIANZA DEL 95%= (,47601- ,723989)

INTERVALO DE CONFIANZA DEL 99%= (.43697- .76302)

Tamaño de muestra:

Si, al efectuar una estimación del parámetro *p*, queremos conseguir una precisión dada (7.8), nos encontraremos con que no podemos despejar de dicha expresión *n,* por intervenir en ella el valor p que tratamos de estimar.

$$\delta = Z_{\varepsilon}\sqrt{pq/n}$$

Sin embargo, aunque no conocemos *p*, sabemos que $\sqrt{pq} < 1/2$. Por ello basta tomar:

$$n = (Z_{\varepsilon} / 2\delta)^2$$

Actividades

7.26. Supongamos que queremos estimar la proporción del grupo RH negativo en una población. Si de una muestra de 400 personas, 35 tuvieron dicho grupo, hallar un intervalo de confianza del 95% de la proporción en la población.

7.27. Una muestra de 100 votantes, elegida al azar de entre los de un distrito, indicó que el 35 por ciento estaban a favor de un cierto candidato. Calcular los límites del intervalo de confianza para proporción real de votantes favorables a dicho candidato, con un coeficiente de confianza del 95%. Calcular el tamaño de muestra que permite estimar con una confianza del 95% la proporción de votantes, con un error menor de 0.03 en la estimación.

7.27. Se dispone de 140 animales de la misma especie, a los que se inocula una suspensión virulenta. El número observado de muertes fue 78. Calcular un intervalo de confianza del 99% para la proporción de muertes en la población.

7.28. En un hospital nacen cada día aproximadamente 16 niños. En otro hospital nacen aproximadamente 100 niños. Calcula los límites en el que variará la proporción de niñas en cada hospital el 95% de los días. ¿En cuál de los dos

hospitales es más variable la proporción de niñas?

7.29. En una muestra aleatoria de 100 rodamientos, 10 tienen un acabado de especificaciones defectuoso. Calcular el intervalo de confianza de 95% para la proporción verdadera de rodamientos defectuosos.

7.30. En un estudio con 240 jóvenes estadounidenses cuyas edades van de 16 a 19 años, seleccionados al azar, 36 presentaron problemas graves de sobrepeso. Obtenga un intervalo de confianza de 99% para la verdadera proporción p de jóvenes de esta población con problemas graves de sobrepeso.

7.31. Un estudio de mercado con 100 personas encontró 37 que habían consumido alguna vez un determinado producto. Calcule el intervalo de confianza del 95% para la proporción de personas que ha consumido dicho producto.

7.6. INTERVALO DE CONFIANZA PARA EL PARAMETRO DE UNA DISTRIBUCION DE POISSON

En este caso, y para muestras grandes, vimos que tambien puede utilizarse la aproximación normal. Para un coeficiente de confianza dado 1-ε el intervalo de confianza será el indicado por (7.10).

(7.10) $$\lambda - Z_{\varepsilon}\sqrt{\lambda} \leq \lambda \leq \lambda + Z_{\varepsilon}\sqrt{\lambda}$$

Al igual que en el caso anterior, en esta expresión hemos sustituido el valor desconocido λ por el valor λ obtenido en la muestra.

Actividad

7.32. En un estudio realizado en 1979 para la provincia de Jaén se alcanzó un total de 85 casos de lepra, entre 100.000 habitantes. Hallar un intervalo de confianza del 99% para la proporción real de leprosos en dicha provincia.

7.33. El número de casos ocurridos durante un mes de una enfermedad rara fue de 15. Calcule un intervalo de confianza del 95% para el número esperado de casos mensuales.

TEMA 8

CONTRASTE DE HIPOTESIS

8.1. INTRODUCCIÓN

Cuando en una rama de las Matemáticas como es el Análisis o la Geometría quiere probarse una cierta conjetura, se realiza un procedimiento de demostración de la misma, que, una vez comprobado que es correcta, establece la certeza de la hipótesis.

En las ciencias experimentales, se plantean en ocasiones ciertas hipótesis que no pueden ser comprobadas de la forma anterior. Así podemos tener motivo para suponer que las personas de una cierta comarca tendrán, en promedio mayor estatura que las de otra, o que una vacuna es efectiva en más del 90% de las personas para la prevención de la gripe. La única forma de comprobar una hipótesis de este tipo, sería efectuar un censo o estudio de toda la población para, en vista de los resultados, aceptarla o no.

Este procedimiento es inviable en la mayoría de los casos. Basándose en la Inferencia estadística podemos, sin embargo, a partir de la información suministrada por una muestra, comprobar con ciertos márgenes de error si dichas hipótesis deben ser admitidas o rechazadas. Llamaremos procedimiento estadístico de contraste de hipótesis al conjunto de operaciones necesarias para llegar a la aceptación o rechazo de una hipótesis estadística. Consta de los pasos siguientes:

- Fijación de las probabilidades de error admisibles
- Determinación del tamaño de la muestra
- Tratamiento de los datos, y formación a partir de ellos de una función de decisión
- Decisión de aceptar o rechazar la hipótesis

Estudiaremos en esta lección el contraste de hipótesis relativas a valores de parámetros en las poblaciones o contrastes paramétricos, desde un punto de vista práctico. Otros tipos de contrastes relativos a la forma de la distribución, dependencia entre variables etc. serán estudiados en los temas posteriores.

8.2. CONCEPTOS FUNDAMENTALES PARA LA REALIZACION DE UN CONTRASTE

El esquema de la realización de un contraste estadístico es siempre el mismo: En primer lugar se propone una hipótesis a comprobar, que llamaremos *hipótesis nula*, y denotaremos por H_0. También ha de proponerse una *hipótesis alternativa* de la anterior H_1 que será la que ha de aceptarse en el caso de que se decida rechazar la hipótesis nula.

Así, si queremos comprobar como hipótesis nula que un somnífero es efectivo en el 90% de los pacientes de insomnio, la hipótesis alternativa puede ser que esta proporción es diferente del 90% (tanto en más como en menos). Otro contraste distinto sería el realizado para comprobar la hipótesis de que la proporción de personas en las cuales el medicamento surte efecto es el 90% o más, contra la alternativa de que esta cantidad es menor del 90%.

Las hipótesis se llaman simples, si se contrasta un único valor del parámetro (como en la hipótesis nula del primer caso), y compuesta si se contrasta un conjunto de valores del parámetro, como en el resto de las hipótesis del ejemplo.

Tipos de errores posibles en la realización de un contraste

Para considerar los errores posibles, estudiemos la situación que se produce según la hipótesis verdadera, y la decisión final tomada. En la tabla 8.1, se presenta un esquema de estas situaciones.

Tabla 8.1. Situaciones en la toma de decisión del contraste de hipótesis

		Decisión tomada	
		Aceptar H_0	Rechazar H_0
Hipótesis cierta	H_0		Error 1
	H_1	Error 2	

Del esquema mostrado, se observa que existen dos tipos de errores:

- El error tipo 1 es el que se comete cuando se rechaza la hipótesis nula, siendo cierta. La probabilidad de cometer este error, que representaremos por α se fija al inicio del contraste, y se conoce como nivel de significación.
- El error de tipo 2 es el que se comete cuando se acepta la hipótesis siendo falsa. En dicho caso, aceptamos como valor del parámetro desconocido uno que no es el verdadero. La probabilidad de cometer este error es función de este valor verdadero θ desconocido, y por tanto la representaremos por $\beta(\theta)$, de forma que, para cada posible valor de θ se obtiene un valor $\beta(\theta)$, que es función del parámetro.

La gráfica de la función de potencia suele estar tabulada para los parámetros más habituales, en función de θ y del tamaño de la muestra, y se conoce como curva característica operativa.

Estadístico utilizado como criterio de verificación de la hipótesis

Para decidir entre las dos hipótesis planteadas, se toma una muestra de la población. Sobre ella se calcula un cierto estadístico, cuya distribución muestral es conocida, relacionado con el parámetro que se desea contrastar. Este estadístico se utiliza como función de decisión. Llamaremos valor observado del estadístico al obtenido en la muestra.

Tras elegir el estadístico utilizado, se divide el conjunto de sus posibles valores en dos conjuntos mutuamente excluyentes:

- Uno de ellos contiene todos los valores del estadístico para los cuales se acepta la hipótesis nula y se llama región de aceptación.
- El otro, los valores para los cuales se rechaza la hipótesis nula y se acepta la alternativa, y se conoce como región crítica.

Llamaremos puntos críticos los que separan las dos regiones. Si obtenemos un punto crítico como valor del estadístico, al no poder tomar una decisión, deberemos aumentar el tamaño de la muestra y repetir el contraste.

Puesto que el nivel de significación α es la probabilidad de rechazar la hipótesis H_0 siendo cierta, y esta hipótesis es rechazada cuando el estadístico usado como función de decisión toma un valor de la región

crítica, esta se determina de forma que la probabilidad de obtener un valor del estadístico en la región crítica, cuando H_0 es cierta sea igual a α.

Potencia del criterio

Se llama potencia del criterio a la probabilidad de que el estadístico tome un valor de la región crítica, cuando es cierta la hipótesis alternativa. Esta probabilidad es la contraria de que el estadístico caiga en la región de aceptación, siendo cierta H_1 y por tanto es igual a $1-\beta(\theta)$. El valor de la potencia, depende del verdadero valor de θ y es por tanto una función de θ Si $\theta= \theta_0$, $\beta(\theta_0)=1- \alpha$ que es la probabilidad de aceptar la hipótesis nula, siendo cierta.

Es claro que cuando menores sean los valores de α y β tanto mejor será el contraste. Sin embargo estos errores están ligados entre si, de modo que, al disminuir uno aumenta el otro. Para comprender esto basta considerar que si hacemos $\alpha =0$, esto es si aceptamos siempre la hipótesis H_0, $\beta(\theta) =1$ para cualquier valor posible de θ distinto del supuesto, ya que como siempre rechazamos la hipótesis alternativa, lo haremos así, aún en el caso de que sea cierta. La única forma posible de disminuir a la vez α y β es aumentar el tamaño de la muestra.

Para un nivel de significación α fijo, es posible sin embargo, determinar la región crítica, de forma que se obtenga la potencia máxima, y por tanto disminuya lo más posible $\beta(\theta)$ para los valores de θ distintos del supuesto. En los ejemplos de contrastes que estudiamos a continuación, se utiliza este método de construcción de la región crítica.

Actividades

8.1. Analiza las definiciones del término "hipótesis" que puedes obtener en un texto de estadística y en un libro de metodología de investigación. Compara con el significado de las hipótesis en otras ramas de las matemáticas (por ejemplo la geometría). Indica la semejanzas y diferencias.

8.2. ¿Equivale la obtención de un resultado estadísticamente significativo a la refutación lógica de la hipótesis nula? ¿Por qué?

8.3. ¿Por qué no son lógicamente equivalentes el rechazo y la aceptación de una hipótesis en un contraste estadístico? ¿Cuáles son las conclusiones cuando se obtiene un resultado que no es estadísticamente significativo?

8.4. Un nivel de significación del 5% significa que, en promedio 5 de cada 100 veces que rechacemos la hipótesis nula estaremos equivocados (verdadero /falso). Justifica tu respuesta.

8.5. Un nivel de significación del 5% significa que, en promedio, 5 de cada 100 veces que la hipótesis nula es cierta la rechazaremos (verdadero / falso). Justifica tu respuesta.

8.6. Un contraste estadístico de hipótesis correctamente realizado establece la verdad de una de las dos hipótesis nula o alternativa. Analiza este enunciado y razona si es verdadero o falso.

8.6. ¿Qué ocurre cuando pasamos de un nivel de significación de 0.01 a otro de 0.05?
a. Hay menos riesgo de error Tipo I
b. Hay más riesgo de error Tipo I
c: Hay menos riesgo de error Tipo II

8.7. ¿Cuál de las siguientes no es una hipótesis nula legítima?
a. $\therefore_x = 10$; b. $\Phi_x = 3$; c. $\therefore_1 = \therefore_2$; d. $\square x_1 = 35$

8.3. COMPARACION DE LA MEDIA MUESTRAL CON UN VALOR HIPOTETICO PARA LA MEDIA DE LA POBLACION

Supongamos que deseamos efectuar un contraste para decidir si el valor medio μ de una población es igual al valor supuesto μ_0. Distinguiremos varios casos:

Población normal, o muestras grandes, con desviación típica conocida: Test bilateral

Se trata de decidir, con un nivel de significación α entre las dos hipótesis siguientes:

$H_0 \equiv \mu = \mu_0$

$H_1 \equiv \mu \neq \mu_0$

Para verificarlas, se toma como estadístico de contraste la media muestral $\bar{x}$, cuyo valor se calcula en la muestra. Puesto que, de ser cierta H_0, $\bar{x}$ sigue una distribución normal $N(\mu_0, \sigma/\sqrt{n})$, se cumple que el valor Z dado en (8.1) sigue una distribución normal $N(0,1)$. Es de esperar, por tanto, que en el caso de verificarse la hipótesis nula el valor Z obtenido en la muestra sea próximo a cero, siendo muy improbables los valores alejados del origen.

(8.1) $$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Al ser α la probabilidad de que el estadístico caiga en la región crítica, cuando $\mu=\mu_{0,}$ tomaremos como regla de decisión la siguiente:

- Si $-Z_{\alpha/2} < Z < Z_{1-\alpha/2}$ decidimos aceptar H_0
- En caso contrario decidimos aceptar H_1

La región crítica de este contraste, está formada pues por dos intervalos: los valores de Z mayores que $Z_{\alpha/2}$ y los menores que $-Z_{\alpha/2}$. Se dice que el contraste es bilateral.

Este mismo contraste puede efectuarse aunque la población de partida no sea normal, utilizando el teorema central del límite. Hay que tener en cuenta que en dicho caso, el método tiene el carácter de aproximado, y será tanto más preciso cuando mayor sea el tamaño de la muestra.

Población de partida normal o muestras grandes con desviación típica conocida: test unilateral

En algunas ocasiones, para realizar el contraste sobre un determinado valor de la media, se desea tomar una alternativa unilateral. Supongamos, por ejemplo que un cierto fabricante produce lámparas cuya duración es una variable aleatoria normal *N(500,50)* y desea cambiar la maquinaria. Sin embargo, antes de realizar el cambio, quiere estar seguro que la vida media resultante con el nuevo procedimiento será mayor que la anterior. Si tiene motivos para suponer que esto no es cierto, sobre la base de una cierta muestra puede realizar un contraste para decidir entre las dos hipótesis siguientes:

$H_0 \equiv \mu \leq 500$

$H_1 \equiv \mu > 500$

Para la realización del contraste ha de tomarse una región de aceptación:

(8.2) $$\bar{x} \leq \mu_0 + \frac{50 Z_{1-\alpha}}{\sqrt{n}}$$

Siendo $z_{1-\alpha}$ el percentil del (1- α)100% en la distribución normal. La región crítica recibe el nombre de unilateral. Si al calcular la media de la muestra, esta cae en la región crítica, tendremos motivos suficientemente fundados para rechazar el nuevo procedimiento de fabricación. La región crítica correspondiente sería:

(8.3) $$\bar{x} \geq 500 - \frac{50 \cdot Z_{1-\alpha}}{\sqrt{n}}$$

En general, en algunos problemas de investigación se desea realizar un contraste para decidir entre las dos hipótesis:

$H_0 \equiv \mu \leq \mu_0$

$H_1 \equiv \mu > \mu_0$

o entre las dos siguientes:

$H_0 \equiv \mu \geq \mu_0$

$H_1 \equiv \mu < \mu_0$

Nótese que en estos casos nos hallamos ante una hipótesis nula compuesta, por lo cual esta no especifica completamente la distribución del estadístico de contraste. Efectivamente, puesto que la distribución de $\bar{x}$, es $N(\mu_0, \sigma/\sqrt{n})$, para los distintos valores posibles de μ menores que μ_0 se obtienen distintas distribuciones de $\bar{x}$.

Sin embargo, como lo que se desea es construir una región crítica, de modo que la probabilidad de error tipo 1 no sobrepase la cantidad α, basta calcular esta región crítica para el caso en que esta probabilidad sea la mayor entre todos los valores del parámetro que componen la hipótesis nula. Para ello, hay que tomar como regiones críticas y de aceptación las obtenidas en el caso anterior. Para dichas regiones puede asegurarse que *max P(error tipo 1)= α*

Población de partida normal o muestras grandes con desviación típica desconocida: test bilateral

Cuando no se conoce la desviación típica de la población, que es el caso más habitual, para decidir, con un nivel de significación α entre las hipótesis:

$H_0 \equiv \mu = \mu_0$

$H_1 \equiv \mu \neq \mu_0$

Se calcula en la muestra el valor del estadístico de contraste que es el dado en la expresión (8.4).

(8.4) $$T = \frac{\bar{x} - \mu_0}{S / \sqrt{n}}$$

Este estadístico tiene una distribución T con $n-1$ grados de libertad. Al ser esta distribución simétrica respecto al origen de coordenadas, es de esperar, por tanto, que en el caso de verificarse la hipótesis nula, el valor T obtenido en la muestra sea próximo a cero, siendo muy improbables los valores alejados del origen. Por ser α la probabilidad de que el estadístico caiga en la región crítica, cuando $\mu=\mu_0$, tomaremos como regla de decisión la siguiente:

- Si $-T_{\alpha/2} < T < T_{\alpha/2}$ decidimos aceptar H_0.
- En caso contrario decidimos aceptar H_1, siendo $-T_{\alpha/2}$ el percentil del $\alpha 100/2$ % de la distribución T de $n-1$ grados de libertad. Obsérvese que la probabilidad de rechazar la hipótesis en el caso de ser cierta es precisamente igual a α

Ejemplo 8.1. Supongamos que tenemos motivo para suponer que la distribución del tiempo de latencia tiene una media de 10 minutos. Deseamos pues, decidir, con un nivel de significación del 5% entre las hipótesis:

$H_0 \equiv \mu = 10$

$H_1 \equiv \mu \neq 10$

En la figura 8.1 se muestra los resultados del cálculo en la hipótesis supuesta. El programa suministra el valor T experimental, y los puntos críticos, que determinan las regiones de aceptación y rechazo, para el test unilateral y bilateral al nivel de significación de 1% y 5%. El estadístico de contraste en este caso, o valor T experimental se calcula de acuerdo con la expresión (8.4) y hemos obtenido un valor 2.0147.

Figura 8.1. Contraste de hipótesis sobre tiempo de latencia

```
Hypothesis Tests for tiempolat
Sample mean = 12,8448
t-test
------
Null hypothesis: mean = 10,0
Alternative: not equal
Computed t statistic = 2,01471
P-Value = 0,0486618
Reject the null hypothesis for alpha = 0,05.
```

Al tratarse de un contraste bilateral, la región de aceptación de la hipótesis nula es simétrica alrededor del origen, y estará formada por los valores *T* incluidos en el intervalo (-2 ,00279, 2 ,00279), para un nivel de significación del 5%. La región crítica está formada por los valores experimentales de *T* que no pertenecen a dicho intervalo. En nuestro caso, hemos obtenido un valor que pertenece a la región crítica, pero muy cerca de los puntos críticos. Resultaría conveniente aumentar el tamaño de la muestra. El posible error a cometer sería el Error Tipo I pues estamos rechazando la hipótesis. Su probabilidad sería el valor Alfa, esto es 0,05.

Actividades

8.8. Un test de una cola es apropiado si:
a. Se desea estar seguro que los resultados serán significativos.
b. Si hay una buena razón para especificar una hipótesis alternativa direccional.
c. Si se quiere ser conservador respecto al nivel de significación.
d. Si se conoce el resultado del experimento antes de hacer el test.

8.9. ¿Es posible que una hipótesis nula pueda ser rechazada en un test de dos colas y la misma hipótesis con el mismo valor del estadístico muestral pueda ser aceptado en un test de una cola?

8.10. Se sabe que la desviación típica de las notas de cierto examen es 2,4. Para una muestra de 36 estudiantes se obtuvo una nota media de 5,6. ¿Sirven estos datos para confirmar la hipótesis de que la nota media del examen fue de 6, a un nivel de significación de 0,05?

8.11. Se cree que la altura media de los habitantes de cierta población es como mucho 170 cm, con una desviación típica de 8 cm. En una muestra de 100 personas se observa una altura media de 172 cm. ¿Podemos aceptar la hipótesis con un nivel de significación del 5%?

8.12. Si tomamos como hipótesis nula H_o μ= 100 y en una muestra de 30 elementos se obtuvo una media x_1= 100, ¿qué decisión debemos tomar?
a. Aceptar la hipótesis nula, sabiendo que hemos tomado la decisión correcta.
b. Aceptar la hipótesis nula, aunque no sabemos si hemos tomado la decisión correcta.
c. Rechazar la hipótesis nula, ya que hemos obtenido un suceso improbable.
d. Necesitamos más información, ya que no conocemos la desviación típica de la población.

Población de partida normal o muestras grandes, con desviación típica desconocida: contraste unilateral

Al igual que en el caso de la desviación típica conocida, podemos estar interesados en decidir entre las dos hipótesis:

$H_0 \equiv \mu \leq \mu_0$

$H_1 \equiv \mu > \mu_0$

Para ello, se calcula en la muestra el valor del estadístico de contraste dado por (8.4) y adoptamos el criterio siguiente: Si $T<T_{1-\alpha}$ decidimos aceptar H_0; en caso contrario decidimos aceptar H_1. Para decidir entre las dos hipótesis:

$H_0 \equiv \mu \geq \mu_0$

$H_1 \equiv \mu < \mu_0$

Se calcula en la muestra el valor del estadístico de contraste (8.4) y la regla de decisión *es: Si* $T_\alpha<T$ *decidimos aceptar* H_0*; en caso contrario decidimos aceptar* H_1

Ejemplo 8.2. Supongamos que el peso medio de los habitantes de Granada es de 65 Kg. Sin embargo, es plausible que el peso medio del grupo de alumnos de Psicología sea algo menor, debido a que entre dichos alumnos abunda la gente joven. Para comprobar esta conjetura tomamos una muestra de 60 alumnos y encontramos un peso medio de 62,38 con desviación típica muestral de 8,56. Se trata de realizar un contraste para decidir entre las hipótesis:

$H_0 \equiv \mu = 63$

$H_1 \equiv \mu < 63$

Puesto que es poco probable que los alumnos tengan un peso superior a la media, no me preocupo de esa posibilidad.

Figura 8.2. Contraste unilateral sobre peso de alumnos

```
Hypothesis Tests for peso
Sample mean = 62,3833
t-test
------
Null hypothesis: mean = 65,0
Alternative: less than
Computed t statistic = -2,36536
P-Value = 0,0106595
Reject the null hypothesis for alpha = 0,05.
```

El valor experimental obtenido es –2,36536 que indica un peso medio inferior al de la población. El valor *p* obtenido es aproximadamente del 1% Por tanto se decide rechazar la hipótesis nula y, en consecuencia decidir que el peso medio de los estudiantes es inferior a 63 kg. La probabilidad de error tipo I es 0,05.

Actividades

8.13. Un estudiante de zootecnia analizó el fósforo en el suero sanguíneo de 9 animales de una cierta especie, obteniendo x= 2,944 mg/l y S= 0,6527. La teoría y la práctica indican que el promedio de fósforo en el suero sanguíneo de una población no deficiente ha de ser 5 mg/l. ¿Son deficientes en fósforo los animales de la muestra? Hallar el nivel de significación mínimo para el contraste unilateral.

8.14. En el total de las células hepáticas, la superficie media es de 291 micras cuadradas. ¿Puede admitirse que la zona portal tiene una superficie mayor, a la vista de los datos del ejercicio 9,4? Hallar el nivel de significación mínimo para el contraste.

8.15. Al examinar una muestra de 60 niños de un colegio español, se obtuvo una puntuación media de 24,3 en un test de intuición probabilística y un valor S=6,12. Este mismo test fue efectuado en una amplia población de escolares ingleses de la

misma edad, alcanzándose un valor medio 20,3. ¿Puede admitirse que el nivel de intuición probabilística de los niños encuestados no difiere significativamente del de sus compañeros ingleses?

8.16. Se obtuvieron los siguientes datos en la medición de la intensidad de una corriente: 3,823 3,844 3,762 3,871 3,762

Realizar un contraste para decidir si la intensidad real es significativamente menor que 3,90.

8.4. COMPARACION DE LA VARIANZA DE UNA POBLACION NORMAL CON UN VALOR SUPUESTO

En ocasiones, el problema que se plantea al investigador es el de realizar un contraste sobre los posibles valores de la varianza poblacional. Puesto que, en general no se conocerá la media de la población, nos basaremos en la distribución del estadístico para esta hipótesis.

Contraste bilateral

Para decidir entre las hipótesis:

$H_0 \equiv \sigma^2 = \sigma^2_0$

$H_1 \equiv \sigma^2 \neq \sigma^2_0$

A partir de una muestra de n valores de una población normal, se calcula en la misma la cuasivarianza muestral S y se acepta la hipótesis nula cuando se obtiene:

(8.5) $$\chi^2_{\alpha/2} < \frac{S^2(n-1)}{\sigma_0^2} < \chi^2_{1-\alpha/2}$$

Y se rechaza cuando se obtiene un valor muestral que no cumple la desigualdad (8.5)

Contrate unilateral

A veces estamos interesados en decidir entre las hipótesis:

$H_0 \equiv \sigma^2 \leq \sigma^2_0$

$H_1 \equiv \sigma^2 > \sigma^2_0$

Para ello, tomaremos una muestra de n valores de una población normal, y calcularemos en la misma la cuasivarianza muestral S. Si esta verifica la relación:

(8.6) $$\chi^2_\alpha < \frac{S^2(n-1)}{\sigma_0^2}$$

Rechazaremos la hipótesis nula, y la aceptaremos cuando se obtiene un valor muestral que no cumple la desigualdad (8.6). Para decidir entre las hipótesis:

$H_0 \equiv \sigma^2 \geq \sigma^2_0$

$H_1 \equiv \sigma^2 < \sigma^2_0$

Se tomará como región crítica: $\frac{S^2(n-1)}{\sigma_0^2} < \chi^2_{1-\alpha}$ y como región de aceptación su complementaria.

Actividades

8.17. Una muestra de 30 personas muestra una desviación típica en el tiempo de ejecución de una tarea de reconocimiento de palabras en una matriz de letras de 120 segundos. ¿Es compatible este resultado, a un nivel de confianza de 99%, con el supuesto de que la variabilidad en la población, con una distribución normal, es de 150 segundos?

8.18. Para evaluar los conocimientos matemáticos de los alumnos de un colegio, un profesor utiliza una prueba que él construye. Con esta prueba viene obteniendo una media de 20 y una varianza de 6. El profesor aplica la prueba a 30 de sus alumnos seleccionados al azar y obtiene una media de 19,5 y una varianza insesgada de 10. ¿Es razonable pensar que la variabilidad de sus alumnos es mayor que lo esperado?

8.5. COMPARACION DE LA PROPORCION MUESTRAL CON EL VALOR SUPUESTO DE LA PROPORCION EN UNA POBLACION BINOMIAL

Contraste bilateral

Supongamos que para un número *n* suficientemente grande de experimentos, en cada uno de los cuales hay una probabilidad *p* desconocida de que se produzca un cierto suceso, se halló una proporción muestral *p*. Se desea decidir entre las dos hipótesis siguientes:

$H_0 \equiv p = p_0$

$H_1 \equiv p \neq p_1$

Tomaremos como criterio de verificación de la hipótesis nula la magnitud aleatoria (8.7) que, en el caso de ser $p=p_0$, y el tamaño de la muestra n lo suficientemente grande, se distribuye aproximadamente como una normal $N(0,1)$. Para un nivel de significación α la regla de decisión será pues, la siguiente:

(8.7) $$z = \frac{p - p_o}{\sqrt{p_o q_o}}$$

Si $-Z_{\alpha/2} < Z < Z_{\alpha/2}$ decidimos aceptar H_0 y en caso contrario decidimos aceptar H_1, donde $-Z_{\alpha/2}$ es el percentil del $\alpha/2100/2$ % de la distribución normal.

Contraste unilateral

Para realizar un contraste unilateral del tipo:

$H_0 \equiv p \leq p_0$

$H_1 \equiv p > p_1$

Se utiliza el estadístico (8.7), aceptando la primera de las hipótesis cuando se verifique: $Z < Z_{\alpha}$. Para decidir entre las hipótesis:

$H_0 \equiv p \geq p_0$

$H_1 \equiv p < p_1$

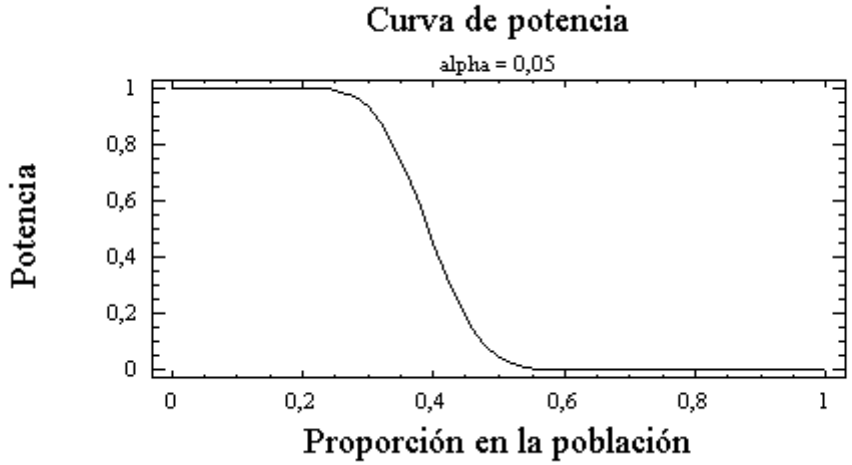
Calcularemos la expresión (8.7), y aceptaremos la primera de las hipótesis en el caso de que: $Z_{1-\alpha} < Z$.

Ejemplo 8.5. En la figura (8.3) se muestra la realización de un contraste para decidir si la proporción de alumnos que practica deporte habitualmente es mayor del 50%. En una muestra de 60 alumnos se obtuvo una proporción del 53%. El programa proporciona el nivel de significación mínimo del contraste, es decir la probabilidad de obtener una diferencia igual o mayor que la obtenida entre el valor p hipotético y el experimental.

Puesto que, en nuestro caso, esta probabilidad es 0,74, no tenemos suficiente motivos para rechazar la hipótesis nula.

Figura 8.3. Contraste unilateral para una proporción y curva de potencia

```
Hypothesis Tests
----------------
Sample proportion = 0,53
Sample size = 60

Approximate 95,0% upper confidence bound for p: [0,641254]

Null Hypothesis: proportion = 0,5
Alternative: less than
P-Value = 0,740521
Do not reject the null hypothesis for alpha = 0,05.
```



El error que podríamos cometer en este caso es el Error Tipo II pues no rechazamos la hipótesis nula. La curva de potencia representada en la parte inferior de la Figura 8.3 me representa la probabilidad de rechazar la hipótesis nula para diferentes posibles valores de *p*. Observamos que esta probabilidad sería muy pequeña si el valor de la proporción en la población fuese realmente mayor que 0,5 y cada vez menor cuanto mayor fuese dicha proporción. Todo ello nos tranquiliza respecto a la decisión tomada.

Actividades

8.19. En una encuesta electoral, se desea averiguar si hay más candidatos a favor de la política económica del presidente que en contra de la misma. Si *p* representa la probabilidad asociada a los habitantes que están de acuerdo con dicha política

económica y $q=1-p$. ¿Cuál de las siguientes hipótesis elegirías como hipótesis nula: a: $p>q$; b: $p=q=1/2$; c: $q>p$

8.20. De entre 340 enfermos que acudieron un cierto día a consulta de atención primaria, 167 eran pensionistas. ¿Están de acuerdo estos datos con la hipótesis de que al menos el 50 por ciento de los enfermos que acuden a consulta son pensionistas?

8.21. En un test de intuición probabilística, de 251 escolares españoles encuestados, 42 obtuvieron el nivel máximo. Entre los escolares ingleses de su edad el 14,8% alcanzaron dicho nivel. ¿Pueden considerarse similares ambas proporciones?

8.22. Se realizan 200 lanzamientos de una moneda y salen 120 caras, ¿podemos aceptar que la moneda no está trucada con un nivel de significación del 5%?

8.23. Una máquina fabrica piezas de precisión y se garantiza que la proporción de piezas correctas producidas es al menos del 97%. Un cliente recibe un lote de 200 piezas y aparecen 8 piezas defectuosas; a un nivel de confianza del 95% ¿rechazará el lote por no cumplir las condiciones de la garantía?

8.6. COMPARACION DE DOS MUESTRAS

Al comparar dos muestras, una de las preguntas que con mayor frecuencia se plantea el investigador, es si dichas muestras provienen o no de la misma población. Aunque la respuesta pueda parecer obvia, sin embargo en la práctica ocurre que, incluso aunque las muestras hayan sido tomadas realmente de una misma población, rara vez producen exactamente la misma información.

La pregunta a la que se ha de responder, es si las diferencias observadas pueden ser atribuidas al azar o al efecto del muestreo, o si realmente son evidencia suficiente de una diferencia entre las poblaciones. Estadísticamente, el problema que se plantea es tomar una decisión acerca de si puede considerarse que los dos conjuntos de datos son observaciones de una misma variable aleatoria, o bien si se trata de dos variables aleatorias con diferente distribución.

A continuación estudiaremos los principales métodos paramétricos de comparación de dos distribuciones. Esto es, se trata de decidir si los valores de ciertos parámetros de las poblaciones (medias varianzas y proporciones) son o no significativamente diferentes. Puesto que en las principales distribuciones de probabilidad, el conocimiento del valor dado del parámetro especifica la distribución, y puesto que en este capítulo haremos de ordinario hipótesis sobre la forma de la misma, estos métodos se conocen también como métodos dependientes de la distribución.

Muestras independientes y muestras relacionadas

Una cuestión primordial al elegir el método estadístico a emplear en la comparación de dos muestras es la hipótesis de independencia.

Diremos que dos muestras son *independientes*, si cada una de las observaciones tomadas en la primera de ellas es, por su naturaleza, independiente de todas las observaciones tomadas en la segunda y recíprocamente. No existe relación entre los individuos de una u otra muestra, ni en el orden en que han sido tomados los datos, puesto que cada conjunto de valores ha sido tomado separadamente por un procedimiento aleatorio.

Ejemplo 8.6. Los valores de la superficie neuronal de la zona dorsal de Apodemus Sylváticus (DATOSA) y los correspondientes a la zona ventral son dos muestras independientes. Se trata de células diferentes elegidas al azar entre las de cada zona.

Por el contrario, diremos que dos muestras están *relacionadas*, o constituyen muestras apareadas, si cada observación de la primera puede ponerse en correspondencia con la que ocupa el mismo lugar en la segunda, de forma que los valores de ambas observaciones constituyen variables aleatorias dependientes.

Ejemplo 8.7. Si durante una semana tomamos los valores de las temperaturas máxima y mínima en la ciudad de Granada, los valores de las temperaturas máximas y los de las temperaturas mínimas constituyen dos conjuntos de datos relacionados. La temperatura máxima cada día, en general dependerá de la mínima, efecto que se manifiesta aún más si consideramos las temperaturas en ciudades diferentes o en diferentes meses.

En los ejemplos anteriores puede observarse que las muestras relacionadas han de tener obligatoriamente el mismo número de elementos. Las muestras independientes pueden o no tener el mismo número de datos.

8.7. COMPARACION DE MEDIAS EN MUESTRAS RELACIONADAS

Contraste bilateral.

Supongamos que disponemos de dos conjuntos relacionados de valores. El primero de ellos x_1, x_2, x_3..... x_n son observaciones de una variable aleatoria ζ_1, cuya media denotaremos por μ_1, y el segundo y_1, y_2,

y_3..... y_n, de la variable aleatoria ζ_2, cuya media llamaremos μ_2. Si queremos efectuar el contraste:

$H_0 \equiv \mu_1 = \mu_2$

$H_1 \equiv \mu_1 \neq \mu_2$

Podemos sustituirlo por otro que se expresará en la forma siguiente:

$H_0 \equiv \mu_1 - \mu_2 = 0$

$H_1 \equiv \mu_1 - \mu_2 \neq 0$

Para realizarlo, sustituimos las observaciones originales por las diferencias: $d_1=x_1-y_1$, $d_2=x_2-y_2$,.....$d_n=x_n-y_n$. Estas diferencias constituyen una muestra de la variable aleatoria $D=\zeta_1-\zeta_2$ cuya media $\bar{d}$, viene dada por la expresión (8.8).

(8.8) $$d = \mu_1 - \mu_2$$

El problema de estimación y contraste de diferencia de medias en muestras relacionadas se reduce, por tanto, al de estimación y contraste de una media en una población, por lo que aplicaremos en este caso los resultados de los apartados anteriores. Nos limitaremos a considerar poblaciones normales, -o muestras grandes- en las que no se conoce la desviación típica de la población de diferencias, que es el caso más habitual. Como vimos, para decidir, con un nivel de significación α entre las hipótesis dadas, se calcula en la muestra el valor del estadístico de contraste que es:

(8.9) $$T = \frac{\bar{d}}{S_{\bar{d}} / \sqrt{n}}$$

Siendo $S_{\bar{d}}$ la raíz cuadrada de la cuasivarianza de las diferencias d_i. Este estadístico tiene una distribución T con $n-1$ grados de libertad. Al ser esta distribución simétrica respecto al origen de coordenadas, es de esperar que, en el caso de verificarse la hipótesis nula, el valor T obtenido en la muestra sea próximo a cero, siendo muy improbables los valores alejados del origen. Al ser α la probabilidad de que el estadístico caiga en la región crítica, tomaremos como regla de decisión la siguiente: Si $-T_{\alpha/2} < T < T_{\alpha/2}$ decidimos aceptar H_0 y en caso contrario decidimos aceptar H_1, siendo $-T_{\alpha/2}$

el percentil del α.100/2 % de la distribución T de $n-1$ grados de libertad. Obsérvese que la probabilidad de rechazar la hipótesis, en el caso de ser cierta, es precisamente igual a α.

Ejemplo 8.8. Al medir la presión sistólica antes y después de haber efectuado un cierto tratamiento médico a un grupo de 10 mujeres se obtuvo los valores siguientes:

Antes	115 112 107 119 115 138 126 105 104 115
Después	128 115 106 128 122 145 132 109 102 117

Se desea decidir si la presión es la misma antes y después del tratamiento. Puesto que se han tomado dos medidas en cada una de las 10 mujeres, las muestras están relacionadas, ya que el valor de la presión dependerá, en general, del sujeto en el que se toma. Para este caso obtenemos los valores siguientes:

d=4.8 $\quad S_{\bar{d}}^{\,2}$=20.84 $\quad S_{\bar{d}}/\sqrt{n}$ =1.444 $\quad T_{exp}$=3.31

Puesto que, para nueve grados de libertad y una significación del 5% se obtiene un valor crítico T=2.26 y para una significación de 0,01 T=3.24, tomamos la decisión de rechazar la hipótesis de igualdad de medias. El nivel de significación del contraste es próximo a 0,01.

Contraste unilateral

En algunas ocasiones podemos tener motivos para suponer que una de las medias es mayor que la otra. Estaremos interesados en decidir entre las dos hipótesis:

$H_0 \equiv d \leq 0$

$H_1 \equiv d > 0$

En este caso, se calcula en la muestra el valor del estadístico de contraste dado por (8.9) y adoptamos el criterio siguiente: Si $T<T_{1-\alpha}$ decidimos aceptar H_0 y en caso contrario decidimos aceptar H_1.

Para decidir entre las dos hipótesis:

$H_0 \equiv d \geq 0$

$H_1 \equiv d < 0$

Se calcula en la muestra el valor del estadístico de contraste (8.9) y la regla de decisión es: Si $T>T_\alpha$ decidimos aceptar H_0 y en caso contrario decidimos aceptar H_1

Algunos autores previenen contra el uso inadecuado de los contrastes unilaterales, pues el valor crítico utilizado es, en general, menor que en el contraste bilateral, para una misma significación. Por ello, dichos autores recomiendan usar el contraste unilateral, sólo después que ha resultado un contraste bilateral significativo.

Ejemplo 8.9. Si en el ejemplo anterior utilizamos el contraste unilateral, para decidir si la segunda media es mayor que la primera, obtendremos una significación de 0,005. En este caso está justificado el uso del contraste unilateral, puesto que en el contraste bilateral habíamos obtenido una diferencia significativa.

Intervalo de confianza para la diferencia de medias.

En general, una vez rechazada la hipótesis de igualdad entre las dos medias, es deseable cuantificar la magnitud de las diferencias. Esto se consigue mediante el cálculo del intervalo de confianza de la diferencia entre las medias, que viene dado por la expresión (8.10).

(8.10) $$\bar{x}-\bar{y}-\frac{t_\varepsilon S_d}{\sqrt{n}} \le \mu_1-\mu_2 \le \bar{x}-\bar{y}+\frac{t_\varepsilon S_d}{\sqrt{n}}$$

En dicha expresión, t_ε es el percentil del *(1-α/2)*100% de la distribución *T* con n-1 grados de libertad.

Ejemplo 8.10. El intervalo de confianza para la diferencia de presiones sistólicas para un coeficiente de confianza del 95% viene dado por:

$$4.8 \pm 2.26*1.444 = 4.8 \pm 3.26 = (1.537, 8.06)$$

Actividades

8.24. A un grupo de pacientes portadores de lente intraocular, se les midió la tonometría previa a la operación de implante y en la fecha en que se obtuvo el alta. Los datos son los siguientes:

P: 14 16 14 15 16 28 14 10 17 16 12 20 20 12 20 12 27 20 14 17

A: 20 12 10 13 16 16 10 16 16 5 12 14 20 12 18 20 23 18 20 12

A la vista de los datos y supuestas las poblaciones normales ¿Puede deducirse que la operación efectuada aumenta la tonometría del paciente?

8.25. Un investigador sospecha que los hombres y las mujeres difieren en sus actitudes hacia el aborto. Para confirmar sus sospechas selecciona aleatoriamente 30 varones y 30 mujeres y les pasa una escala para medir la mencionada actitud. Los resultados obtenidos son los siguientes:

- Hombres: media 38; desviación típica 6
- Mujeres: media 31; desviación típica 5

Sabiendo que cuanto mayores son las puntuaciones en la escala más favorable es la actitud hacia el aborto, ¿qué concluirá el investigador con un nivel de confianza de 0,95?

8.26. Supongamos que, sobre una misma muestra de estudiantes estudiamos las calificaciones en 10 asignaturas diferentes y nos interesa analizar cuáles de estas calificaciones difieren significativamente. Si usamos el test T de diferencias de medias relacionadas, a un nivel de significación del 0.01. ¿Cuántas diferencias habría que esperar resultasen significativas, simplemente por las fluctuaciones del muestreo, en el caso de que no existiese ninguna diferencia real entre las calificaciones? ¿Cómo podríamos solucionar este problema?

8.8. COMPARACION DE MEDIAS EN MUESTRAS INDEPENDIENTES DE POBLACIONES DE VARIANZA CONOCIDA.

Contraste bilateral

Supongamos que disponemos ahora de dos conjuntos independientes de datos. El primero de ellos $x_1, x_2, x_3 \ldots\ldots x_n$ son observaciones de una variable aleatoria ζ_1 cuya media y varianza denotaremos por μ_1 y σ^2_1, respectivamente, y el segundo $y_1, y_2, y_3 \ldots\ldots y_m$, de la variable aleatoria ζ_2, cuya media y varianza llamaremos μ_2 y σ^2_2. En general, los valores m y n serán diferentes. Aunque, en los casos reales no suelen conocerse los valores de σ^2_1 y σ^2_2, resulta conveniente proceder en principio al estudio de este caso hipotético pues, como veremos, puede ser utilizado, con carácter aproximado para muestras lo suficientemente grandes. En este caso, procedemos a calcular el valor S del error de muestreo de la diferencia de medias dado por (8.11).

(8.11) $$S = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

Si queremos realizar el contraste:

$H_0 \equiv \mu_1 - \mu_2 = 0$

$H_1 \equiv \mu_1 - \mu_2 \neq 0$

El estadístico a utilizar es:

(8.12) $$Z = \frac{\bar{x} - \bar{y}}{S}$$

Siendo S el valor dado en la expresión (8.11). Este estadístico tiene una distribución N (0,1). Tomaremos como regla de decisión la siguiente: Si $-Z_{\alpha/2} < Z < Z_{\alpha/2}$ decidimos aceptar H_0 y en caso contrario, decidimos aceptar H_1, siendo $-Z_{\alpha/2}$ el percentil del $\alpha 100/2$ % de la distribución normal.

Contraste unilateral

Si estamos interesados en decidir entre las dos hipótesis:

$H_0 \equiv \mu_1 - \mu_2 \leq 0$

$H_1 \equiv \mu_1 - \mu_2 > 0$

Se calcula en la muestra el valor del estadístico de contraste dado por (8.12). La región crítica del contraste está constituida por los valores Z tales que $Z < Z_\alpha$. El contraste unilateral de sentido inverso, se realiza en forma similar.

Intervalo de confianza para la diferencia de medias

El intervalo de confianza de la diferencia entre las medias, en este caso, viene dado por la expresión (8.13).

(8.13) $$\bar{x} - \bar{y} - Z_\varepsilon S \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} - Z_\varepsilon S$$

En la expresión (8.13), Z_ε es el percentil del $(1 - \varepsilon/2)100\%$ de la distribución normal, y S viene dado por la expresión (8.11).

Actividades

8.27 Elegimos aleatoriamente 50 alumnos de Psicología de la Universidad Autónoma de Madrid y 120 de la Universidad Complutense. Supongamos que

cada universidad sigue un método distinto de enseñanza de la asignatura de "Análisis de datos". Sea $\overline{X_1}$ (la media de los alumnos de la Autónoma) igual a 74 y $\overline{X_2}$ (la media de los alumnos de la Complutense) igual a 79. Sabiendo que las desviaciones típicas de la población son 12 y 18 respectivamente, deseamos contrastar la hipótesis de si la enseñanza tiene efecto

---

8.9. COMPARACION DE MEDIAS EN MUESTRAS INDEPENDIENTES EN POBLACIONES DE IGUAL VARIANZA.

Contraste bilateral.

En general, los valores σ^2_1 y σ^2_2 de las varianzas poblacionales no serán conocidos. En el estudio de la diferencia de medias en muestras independientes de varianza desconocida, el método utilizado será diferente según que las varianzas de las poblaciones puedan considerarse o no idénticas. Por ello, el primer paso a realizar en un contraste de este tipo es una prueba de homogeneidad de varianzas. Supondremos que, por el resultado de dicha prueba, podemos suponer que las dos variables aleatorias poseen una desviación típica común σ. Esta viene estimada por la expresión (8.14).

(8.14)
$$S = \frac{\sqrt{(nS_x^2 + mS_y^2)(\frac{1}{n} + \frac{1}{m})}}{\sqrt{n+m-2}}$$

Para decidir entre las hipótesis:

$H_0 \equiv \mu_1 - \mu_2 = 0$

$H_1 \equiv \mu_1 - \mu_2 \neq 0$

El estadístico a calcular es:

(8.15)
$$T = \frac{\bar{x} - \bar{y}}{S}$$

Siendo S el valor dado en la expresión (8.14). En este caso, el estadístico tiene una distribución T con $n+m-2$ grados de libertad, por lo

que aceptamos H_0 si $-T_{\alpha/2}<T<T_{\alpha/2}$ y aceptamos H_1 en caso contrario. $-T_{\alpha/2}$es el percentil del α 100/2 % de la distribución T de $n-1$ grados de libertad.

Ejemplo 8.11. Al medir los niveles de inmunoglobulina IgD en escolares de ambos sexos se obtuvo los datos siguientes:

VARONES: 12.0 0.0 9.3 8.1 5.8 6.8 3.6 9.5 8.6 9.3

HEMBRAS: 5.8 0.0 7.0 0.0 7.5 2.6 5.5 7.2 7.3 3.3

Efectuaremos un contraste para decidir si el nivel de inmunoglobulina es similar en ambos sexos. Aunque tenemos 10 escolares en cada grupo, las muestras son independientes, pues se trata de sujetos diferentes, sin relación entre ellos (podría ser distinto el caso, si se tratase de hermanos, etc.). En este caso, puede considerarse que las varianzas de las poblaciones, aunque desconocidas, son idénticas. Obtenemos los siguientes valores en los cálculos:

x=7,1 y=4,02 S_x^2 =10,37 S_y^2 =8,71 T_{exp}=1,75

Puesto que, para 18 grados de libertad el valor crítico de T es, para una significación del 5%, 2,101 aceptamos que las medias son iguales.

Contraste unilateral

Si estamos interesados en decidir entre las dos hipótesis:

$H_0 \equiv \mu_1-\mu_2 \leq 0$

$H_1 \equiv \mu_1-\mu_2 > 0$

Se calcula en la muestra el valor del estadístico de contraste dado por (8.15) y adoptamos el criterio siguiente: Si $T<T_{1-\alpha}$ decidimos aceptar H_0 y en otro caso aceptamos H_1. De forma semejante se realiza el contraste unilateral de sentido inverso.

Ejemplo 8.12. Al realizar el contraste unilateral en el ejemplo anterior, el valor T crítico para un nivel de significación del 5% es 1,734. Este valor también es superior al experimental, por lo que no puede tampoco en este caso aceptarse la hipótesis alternativa. Nótese, sin embargo que el valor experimental es muy próximo al crítico en este ejemplo. Ello obliga a

considerar el peligro de aceptar un contraste unilateral, sin haber estudiado previamente el caso bilateral.

Intervalo de confianza para la diferencia de medias.

Para evaluar la magnitud de las diferencias entre las medias, calcularemos el intervalo de confianza, que en este caso viene dado por la expresión (8.16).

(8.16) $$\bar{x}-\bar{y}-T_{\varepsilon}\,S \leq \mu_1-\mu_2 \leq \bar{x}-\bar{y}-T_{\varepsilon}\,S$$

En donde T_{ε} es el percentil del (1-ε/2)100% de la distribución T con n-1 grados de libertad, y S viene dado por la expresión (8.14).

8.10. COMPARACION DE MEDIAS EN MUESTRAS INDEPENDIENTES EN POBLACIONES DE VARIANZA DIFERENTE.

Contraste bilateral.

Supongamos que al realizar la prueba de homogeneidad de varianzas a los conjuntos independientes de datos $x_1, x_2, x_3 \ldots\ldots x_n$ e $y_1, y_2, y_3 \ldots\ldots y_m$ llegamos a la conclusión de que las variables ζ_1 y ζ_2 poseen varianzas diferentes. En dicho caso, Welch ha sugerido un procedimiento que tiene carácter aproximado. Calcularemos el valor S dado por la expresión (8.17), en la que S_1 y S_2 son, respectivamente las cuasivarianzas de la primera y segunda muestra.

(8.17) $$S = \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}$$

Para realizar el contraste:

$H_0 \equiv \mu_1-\mu_2=0$

$H_1 \equiv \mu_1-\mu_2 \neq 0$

El estadístico a utilizar es:

(8.18) $$T = \frac{\bar{x}-\bar{y}}{S}$$

Siendo S el valor dado en la expresión (8.17). Este estadístico tiene una distribución T con f grados de libertad, donde f viene dado por la expresión (8.19).

(8.19)
$$f = \frac{\left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right)^2}{\frac{\left(S_1^2/n\right)^2}{n+1} + \frac{(m)^2}{m+1}} - 2$$

Tomaremos como regla de decisión la siguiente: si $-T_{\alpha/2}<T<T_{\alpha/2}$ aceptamos H_o en caso contrario aceptamos H_1. $T_{\alpha/2}$ es el percentil del α 100/2 % de la distribución T de n-1 grados de libertad. Otra alternativa es utilizar los percentiles correspondientes a la distribución normal, cuando los grados de libertad son lo suficientemente elevados.

Contraste unilateral

Si estamos interesados en decidir entre las dos hipótesis:

$H_0 \equiv \mu_1-\mu_2 \leq 0$

$H_1 \equiv \mu_1-\mu_2 > 0$

Se calcula en la muestra el valor del estadístico de contraste dado por (8.18) y adoptamos el criterio siguiente: Si $T<T_{1-\alpha}$ decidimos aceptar H_0 y en otro caso aceptamos H_1. De forma análoga se realiza el contraste unilateral de sentido inverso.

Intervalo de confianza para la diferencia de medias.

El intervalo de confianza de la diferencia entre las medias, en este caso viene dado por la expresión (8.20).

(8.20)
$$\bar{x} - \bar{y} - T_\varepsilon S \leq \mu_1-\mu_2 \leq \bar{x} - \bar{y} - T_\varepsilon S$$

En donde T_ε es el percentil del (1-ε/2)100% de la distribución T con n-1 grados de libertad, y S viene dado por la expresión (8.17).

8.11. COMPARACION DE VARIANZAS EN POBLACIONES NORMALES

En los apartados anteriores, hemos visto la necesidad de estudiar la homogeneidad de las varianzas de dos poblaciones, con objeto de elegir adecuadamente el método estadístico a aplicar en cada caso. Por otro lado, en ciertos problemas de investigación es más importante contrastar la igualdad de varianzas que la de medias. Una modificación en la variabilidad de potencia de un medicamento, por ejemplo, aunque tal vez sea menos importante que una modificación en la potencia media, podría tener como resultado la producción de un porcentaje demasiado elevado de lotes ineficaces por su baja potencia o peligrosos por su elevado efecto.

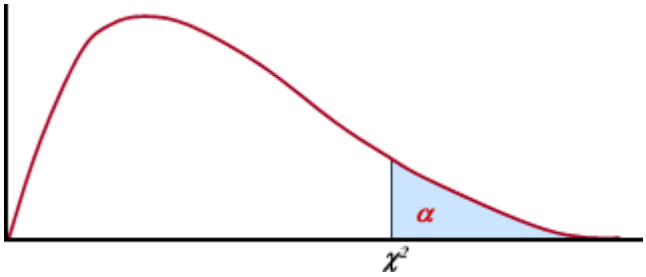
Contraste unilateral

Para el cálculo de intervalos de confianza y la realización de contrastes sobre el cociente da varianzas σ_1^2/σ_2^2 de dos poblaciones se utiliza el estadístico F dado en (8.21)

(8.21) $$F= S_1^2/S_2^2$$

Este estadístico tiene una distribución F con n-1 y m-1 grados de libertad, siendo n y m el número de elementos de las muestras utilizadas para el cálculo de las cuasivarianzas S_1^2 y S_2^2 respectivamente. La distribución F, al igual que ocurre con la distribución Chi-cuadrado es no simétrica. En la figura 8.8 se muestra una de las posibles gráficas de dicha distribución.

Figura (8.8). Ejemplo de distribución Chi-cuadrado



Comenzaremos por la realización del contraste unilateral, para decidir si dos varianzas son o no diferentes. Para decidir entre las hipótesis:

$H_0 \equiv \sigma_1^2/\sigma_2^2=1$

$H_1 \equiv \sigma_1^2/\sigma_2^2>1$

Calcularemos el valor F dado en (8.21) y adoptaremos la siguiente regla de decisión: Si $F>F_{1-\alpha}$ rechazamos la hipótesis nula y en caso contrario la aceptamos. En este caso, $F_{1-\alpha}$ es el percentil del 100(1-α)% de la distribución F con n-1 y m-1 grados de libertad. Las tablas habitualmente disponibles sólo presentan los percentiles superiores de la distribución **F**. Si quisiéramos invertir el sentido del anterior contraste unilateral, podríamos utilizar los percentiles inferiores de la distribución F, que se obtienen mediante la relación (8.22).

(8.22) $$F_\alpha(n,m)=1/F_{1-\alpha}(m,n)$$

Ejemplo 8.13. Con los datos del ejemplo 8.11, efectuaremos un contraste de igualdad de varianzas: En este caso:

$$S_1^2/S_2^2 = 10.37/8.71 = 1.304$$

Puesto que el valor crítico F para $n=m=9$ grados de libertad es, para una significación de 5% igual a 3.18, deducimos la igualdad de las varianzas.

Contraste bilateral

Para realizar el contraste bilateral de homogeneidad de dos varianzas, basta proceder como en el caso anterior y tomar como regla de decisión la siguiente: Si $F>F_{1-\alpha/2}$ rechazamos la hipótesis nula y en caso contrario la aceptamos. En este caso, $F_{1-\alpha/2}$ es el percentil del 100(1-$\alpha/2$)% de la distribución F con n-1 y m-1 grados de libertad. El nivel de significación del contraste es, sin embargo α al considerar como hipótesis alternativa tanto $\sigma_1^2>\sigma^2_2$ como el caso contrario.

Intervalo de confianza.

Para un coeficiente de confianza dado, 1-α el correspondiente intervalo de confianza para el cociente de las varianzas viene dado por la expresión (8.23).

(8.23) $$\frac{S_1^2/S_1^2}{F_{1-\varepsilon/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2/S_1^2}{F_{\varepsilon/2}}$$

Ejemplo 8.14. Calcularemos el intervalo de confianza del 95% en el ejemplo. En este caso:

$$F_{97.5}=4.43 \text{ y } F_{2.5}=1/4.43=.2257$$

$$a=1.304/4.43=0.2943 \text{ y } b=1.304/0.2257=5.778$$

Actividades

8.28 En 200 células de la zona portal del hígado en ratas hembras, el porcentaje medio de grasa citoplasmática fue 42,13, con un error de muestreo de 1.5. En el mismo número de células de ratas macho el porcentaje medio obtenido fue 22,49 con un error de muestreo de 0.95. ¿Puede considerarse igual las varianzas de ambas poblaciones? Calcular un intervalo de confianza del 95% para el cociente de varianzas.

8.29. En el ejercicio 10,4, ¿pueden considerarse iguales las medias? Calcular un intervalo de confianza para la diferencia de medias.

8.30. Al realizar una encuesta de lecturas infantiles, entre 143 niños el número medio de autores citados fue 7,37 con un valor S_1=9.52. Entre 209 niñas, el número medio de autores citados fue 9,7 con un valor S_2=1,.95. ¿Son significativas las diferencias entre medias?

8.12. COMPARACION DE DOS PROPORCIONES EN MUESTRAS INDEPENDIENTES

Contraste bilateral

Así como al comparar dos distribuciones continuas nos hemos preocupado de la diferencia de medias o varianzas, al tratarse de variables dicotómicas suele ser deseable comparar los parámetros de dos distribuciones binomiales. Supongamos pues que tenemos una muestra de n_1 observaciones de una variable ζ_1, con distribución binomial $B(n_1,p_1)$ en la que se ha obtenido x_1 veces la característica considerada, y otra muestra de n_2 observaciones de la variable ζ_2 que tiene distribución $B(n_2,p_2)$. De esta segunda muestra se obtuvo un total de x_2 apariciones de la característica. Para realizar el contraste:

$H_0 \equiv p_1=p_2$

$H_1 \equiv p_1 \neq p_2$

Se calcula de las muestras dadas el estimador para la proporción, supuestamente común de las poblaciones, que viene dado por (8.24)

(8.24) $$p = \frac{x_1 + x_1}{n_1 + n_1}$$

El estadístico de contraste viene dado por (8.25) y la regla de decisión es análoga a las utilizadas en otros contrastes anteriores. De igual forma se realizan los contrastes unilaterales.

(8.25) $$\frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}$$

Ejemplo 8.15. A dos grupos de personas se les hizo una prueba de destreza manual. Del grupo A, 44 superaron la prueba y 10 no. Del grupo B, 81 superó la prueba y 35 fallaron ¿Son igualmente diestros ambos grupos?

En este ejemplo, obtenemos los siguientes valores:

p_1=0.185 p_2=0.432 p_2 - p_1=0.247

p =0.33 Z=0.247/0.0828=2.98

El valor Z es significativo al 1%, por lo que se rechaza la hipótesis de igualdad entre los grupos.

Intervalo de confianza

Una vez decidida la diferencia entre dos proporciones, conviene cuantificarla. El intervalo de confianza para un coeficiente dado de confianza 1-ε viene dado por (8.27), en donde S viene dado por (8.26)

(8.26) $$S = \sqrt{\frac{x_1}{n_1 - x_1} + \frac{x_2}{n_2 - x_2}}$$

(8.27) $$p_1 - p_2 - S.Z_{\varepsilon/2} < p_1 - p_2 < p_1 - p_2 + S.Z_{\varepsilon/2}$$

Ejemplo 8.16. El intervalo de confianza de la diferencia de proporciones en el ejemplo anterior para un coeficiente del 95% viene dado por:

$$0.247 \pm 1.96*09762 = (0.097,0396)$$

siendo S=0.0762

Actividades

8.31. En el estudio de pacientes portadores de lente intraocular, de un total de 101 varones, 9 presentaron patología oftálmica previa y 8 de 63 mujeres. ¿Son similares ambas proporciones?

8.32. Al compara dos técnicas de radioterapia A y B se obtuvo resultados positivos con la técnica A en un 40% de 215 casos, y con la B en un 30% de 150 casos. Hallar un intervalo de confianza del 95% para la diferencia entre ambas proporciones.

8.33. En el estudio sobre la lepra realizado en la provincia de Jaén, de entre 184 enfermos que presentaban la forma clínica lepromatosa 117 seguían el tratamiento con regularidad. De entre 105 pacientes que presentaban otra de las posibles formas clínicas, 49 seguían con regularidad el tratamiento. ¿Puede deducirse de los datos, que los enfermos con forma clínica lepromatosa son más regulares en el tratamiento?

TEMA 9.

ANALISIS DE LA VARIANZA

9.1. INTRODUCCION

En el capítulo anterior se estudiaron diversos métodos de comparación de dos muestras. En muchos casos, sin embargo, nos vemos obligados a comparar tres o más muestras. En este capítulo estudiaremos el Análisis de la Varianza, procedimiento estadístico que permite comprobar si *r* muestras provienen o no de poblaciones con la misma media, cuando se dan ciertas condiciones establecidas.

Ejemplo 9.1. En una investigación se aplicó un test sobre intuición probabilísticas elaborado por Green, a una muestra de 248 escolares de la ciudad de Jaén. Este test consta de un total de 50 ítems de opciones múltiples. La puntuación total se desglosa en tres componentes: combinatoria, verbal y probabilística, que intenta estudiar estos tres aspectos dentro de la intuición probabilística. También se evalúa el nivel probabilístico de los niños (que varía de 0 a 3), en base a haberse alcanzado unos objetivos mínimos.

Las variables que consideraremos en la investigación son las siguientes: COLEGIO, SEXO, CURSO (Curso de EGB, de 6º a 8º), AM (Aptitud matemática, asignada por el profesor del niño), PC(Puntuación combinatoria), PV(Puntuación verbal), PP(Puntuación probabilística), PT(Puntuación total) y NP(Nivel probabilístico).

El objeto de esta investigación es descubrir si existen diferencias entre sexo, curso o colegio en alguna de las variables, y comparar los resultados obtenidos en la muestra con los de los escolares ingleses.

En principio, podría pensarse que el problema de comparación de varias muestras podría solucionarse estudiando las muestras dos a dos, para intentar detectar las diferencias significativas. Aparece aquí el problema de las "comparaciones múltiples". Cuando efectuamos *k* comparaciones

teniendo cada una de ellas un nivel de significación individual α, el nivel de significación global toma el valor *kα*, por lo que crece la probabilidad de encontrar diferencias significativas aún en el caso de que las muestras provengan en realidad de la misma población.

El Análisis de la varianza intenta paliar este problema, a la vez que resulta un procedimiento más eficaz de análisis. Del mismo modo que al realizar un contraste entre dos medias, podemos encontrarnos en el caso de muestras independientes o relacionadas. Por medio del análisis de la varianza de un factor se estudia el caso de muestras independientes. El caso de muestras relacionadas, corresponde al análisis de la varianza de 2 o más factores.

9.2. ANALISIS DE LA VARIANZA CON UN FACTOR: MODELO DE EFECTOS FIJOS

Supongamos que disponemos de *k* grupos, de modo que en el grupo *i*-ésimo hay n_i observaciones. Sea x_{ij} la *j*-ésima observación en el grupo *i*. Todos los elementos del mismo grupo *i* se dice que están "sujetos al tratamiento *i*". Este nombre proviene de las primitivas aplicaciones del análisis de varianza en el trabajo experimental, sobre todo, en agricultura. Haremos las siguientes hipótesis:

- *Independencia*: Cada una de las observaciones es independiente de las demás.
- *Linealidad:* Cada uno de los valores observados x_{ij} puede descomponerse en la forma (9.1).

(9.1) $$x_{ij}=\mu+\alpha_i+\varepsilon_{ij}$$

con la condición de que $\Sigma\alpha_i=0$. Es decir la suma de las diferencias de cada grupo a la media global es igual a cero. En la expresión (9.1) μ es una constante, que representa la media global del conjunto de las *k* muestras; α_i es una constante dentro del grupo *i* y representa la diferencia de la media de este grupo con la media del grupo. El valor α_i se suele conocer como "efecto debido al tratamiento *i*". Por último, ε_{ij} es una variable aleatoria con distribución normal $N(0, \sigma)$. Así, si en el ejemplo 1 tomamos la puntuación x_{ij} de un niño de 6º en el test de Green, μ sería la puntuación media de todos los niños, α_i sería la

diferencia entre la puntuación media de los niños de 6º respecto a la de todos los niños y ε_{ij} la diferencia entre la puntuación de este niño concreto y todos los de 6º.

- ***Homocedasticidad.*** Se supone una varianza común σ para todos los grupos. Por tanto, podemos decir que la variable aleatoria x_{ij} sigue una distribución normal $N(\mu+\alpha_i, \sigma)$.

El Análisis de la Varianza consiste en la realización de un contraste estadístico para decidir entre las dos hipótesis siguientes:

$H_0 \equiv \alpha_1=\alpha_2=.....=\alpha_k=0$

$H_1 \equiv$ al menos un α_i es diferente de cero.

Podemos expresar el contraste anterior de otra forma, diciendo que, bajo la hipótesis nula, todos los grupos provienen de poblaciones con igual media, y, en consecuencia, todos los tratamientos producen un efecto nulo. Bajo la hipótesis alternativa, al menos un tratamiento produce un efecto no nulo, por lo que algunas de las poblaciones tendrán diferentes medias.

En el modelo de efectos fijos se supone que los k grupos de que disponemos son los únicos existentes en la población. Por ejemplo, el factor "género" es fijo, porque en la población sólo hay dos grupos, chicos y chicas, que son los que se usan en el análisis. Las conclusiones que obtendremos se referirán, por tanto, a esos k valores α_i y no a otros distintos y el modelo se llama "de efectos fijos".

Cuando los valores α_i no son los únicos posibles, sino que representan una muestra posible de una población de valores α que es, a su vez una variable aleatoria, nos hallaremos ante el modelo de efectos aleatorios, que se estudiará en la sección 9.3. Por ejemplo, en el ejemplo 9.1, los colegios de la muestra no son todos los colegios de Jaén. Por tanto, el factor colegios es aleatorio.

Planteamiento de las hipótesis

Para realizar el contraste, representaremos la desviación de la observación x_{ij} respecto a la media total en la forma (9.2).

(9.2) $$x_{ij}-x=(x_{ij}-x_i)+(x_i-x)$$

En la expresión (9.2) $(x_{ij}-x_i)$ es la desviación de la observación respecto a la media de la muestra *i*, y representa la variabilidad dentro de la población *i*. Por otro lado, (x_i-x) es la desviación entre la media de la muestra *i* y la media global, y mide la variabilidad entre los grupos.

Si la variabilidad entre grupos fuese grande respecto a la que hay dentro de cada grupo, pensaríamos que nos hallamos ante poblaciones con medias diferentes, y rechazaríamos la hipótesis nula. En la figura 9.1, se muestran dos ejemplos. En el primero, la variabilidad entre grupos es mayor que la que hay en cada grupo. Por el contrario, en la segunda, al haber una gran dispersión dentro de cada grupo, no permite apreciar si hay una diferencia real de medias en las poblaciones.

Figura 9.1

a) b)

Disposición de los cálculos.

Para efectuar el análisis es preciso calcular los elementos de la tabla del análisis de la varianza, dada en la tabla 9.1.

Tabla 9.2. Tabla del análisis de varianza

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	*F*
Entre grupos	*SCE*	*k-1*	$CME = \frac{SCE}{k-1}$	$F_{\exp} = \frac{CME}{CMD}$
Dentro de los grupos (residual)	*SCD*	*n-k*	$CMD = \frac{SCD}{n-k}$	
Total	*SCT*	*n-1*		

En la tabla 9.1, los distintos componentes se calculan mediante las igualdades (9.3) a (9.5), donde *SCT* representa la suma total de cuadrados,

SCD la suma de cuadrados dentro de los grupos y *SCE* la suma de cuadrados entre grupos, y se verifica que *SCT=SCD+SCE*.

(9.3) $$SCT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

(9.4) $$SCD = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

(9.5) $$SCE = \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2$$

Hay que hacer notar que *SCD* es en realidad la varianza de los valores x_{ij} respecto a la media global de todas las muestras, multiplicada por *n.* Por tanto, el cuadrado medio *CMD* estima la varianza de los valores x_{ij} respecto a la media global, esto es, σ^2. Por su parte *CME* estima la varianza de las medias de cada muestra respecto a la media global. Si la hipótesis nula fuese cierta, estas dos varianzas serían aproximadamente iguales, siendo las diferencias observadas pequeñas y debidas únicamente al error del muestreo.

En el caso de ser mayor las diferencias entre grupos a las diferencias dentro de los grupos, el valor F_{exp} será mayor que la unidad. Puede observarse que el razonamiento seguido para efectuar este contraste es parecido al utilizado en el estudio de homogeneidad de varianzas, en el capítulo anterior. El estadístico F_{exp} sigue la distribución *F* con *k-1* y *n-k* grados de libertad.

Adoptaremos, en consecuencia, la siguiente regla de decisión: Si el valor $F_{exp} \leq F_{k-1,n-k}$ se acepta H_0 En caso contrario aceptamos H_1

Ejemplo 9.2. En la tabla 9.2 presentamos los resultados de haber aplicado un análisis de la varianza a la puntuación total clasificada por curso, en el conjunto de datos TESTPR. El problema que nos planteamos es decidir entre las dos hipótesis siguientes:

$H_0 \equiv$ La puntuación total media en los cursos 6º a 8º es la misma

$H_1 \equiv$ Al menos dos de estos cursos tienen diferente puntuación total media.

Tabla 9.2. Tabla del análisis de varianza

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Entre los cursos	650,63	2	325,31	8,42
Dentro de los cursos	9465,60	245	38,63	
Total	10116,23	247		

Puesto que, con los grados de libertad del ejemplo, el valor F obtenido corresponde a un nivel de significación menor de 0,0003, decidimos rechazar la hipótesis nula, y concluimos que la puntuación total varía entre los cursos.

Actividades

9.1. Se desea analizar si existen diferencias en el gasto medio en medicamentos efectuado por las familias de renta alta, media y baja. Para poder utilizar el contraste F de análisis de la varianza es necesario suponer que:

a. La varianza poblacional del gasto en medicamentos es la misma para los tres niveles de renta
b. El gasto medio poblacional en medicamentos es el mismo para los tres niveles de renta
c. La varianza muestral del gasto en medicamentos es la misma para los tres niveles de renta
d. El gasto medio en medicamento en la muestra es el mismo en los tres grupos

9.2. En el análisis de varianza llamamos factor:

a. A las variables extrañas
b. A las variables dependientes
c. A las variables independientes

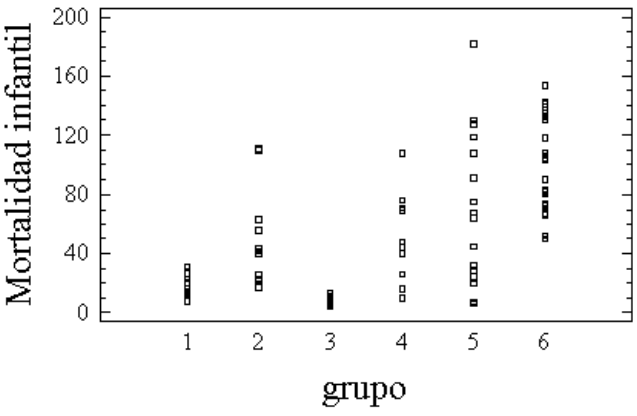
9.3. La varianza entre grupos en el análisis de varianza es:

a. La varianza muestral
b. La atribuible al error
c. La atribuible a las diferencias entre grupos

9.4. Si en un ANOVA de un factor y cuatro niveles del factor se rechaza la hipótesis nula, esto implica que:

a. Debemos concluir la igualdad de medias poblacionales en todos los grupos
b. Debemos concluir que las medias poblacionales de todos los grupos serán diferentes unas de otras.
c. Algunas medias poblacionales de los grupos serán diferentes entre si.

9.5. En la figura adjunta se representa gráficamente la tasa de mortalidad infantil en una serie de paises clasificados por zona geográfica en la forma siguiente: 1=Europa Oriental: 2= Ibero América; 3=Europa Occidental, Norte América, Japón, Australia, Nueva Zelanda; 4 = Oriente Medio; 5= Asia; 6 = África. ¿Es mayor la variación de la mortalidad entre grupos o dentro del os grupos? Si se aplicase el análisis de varianza, ¿cuál sería la variable dependiente y cuál el factor? ¿Cúantos niveles habría? ¿Qué resultado cabe esperar?



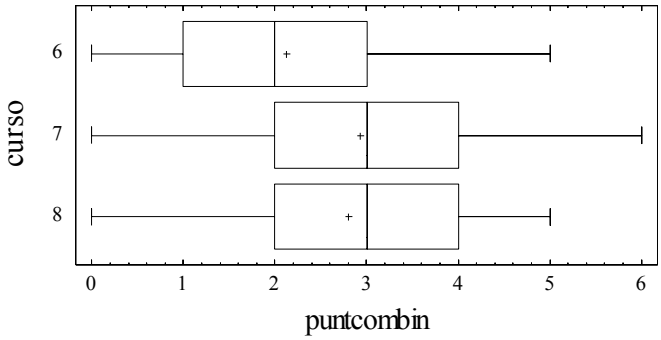
9.6. La siguiente tabla de análisis de varianza se obtuvo al estudiar la puntuación matemática de los estudiantes de un a muestra por curso. ¿Cuántos cursos hay en la muestra? Establece las hipótesis adecuadas. Completa la tabla ¿Puede deducirse la existencia de diferencias por curso?

Fuente	Suna cuadrados	G.L.	Cuadrado medio	F	p
Entre grupos	0,703472	2			
Dentro	1182,45	247			
Total	1183,16	249			

9.7. Completa la siguiente tabla de análisis de varianza de puntuación combinatoria de alumnos de una muestra por curso ¿Varía la puntuación verbal en los diferentes cursos?

Fuente	Suna cuadrados	G.L.	Cuadrado medio	F	p
Entre grupos	33,4657	2			
Dentro	536,534	247			
Total	570,0	249			

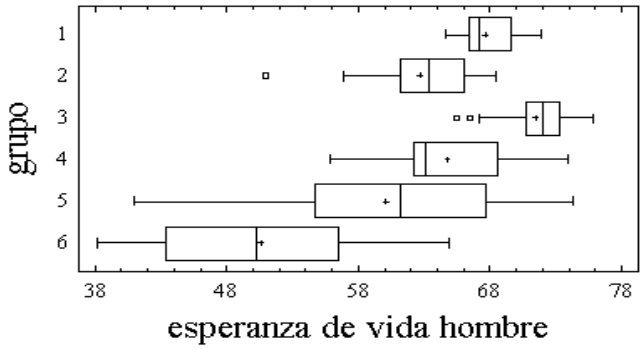
9.8. Considerando el gráfico siguiente y el resultado del análisis de varianza del ejercicio 9.7, ¿En cuáles cursos podría decirse que existe diferencia media en



9.9. En un estudio sobre los efectos de un determinado gen en la protección del organismo se analiza una muestra de 5 ratones, a 3 de los cuales se elimina el gen y 2 no. Posteriormente se miden sus niveles de células cancerıgenas. Se obtiene que la media para los que no tienen el gen es 126 y para los que lo tienen es 109. ¿Cuál sería la variable dependiente y cuál el factor si se considera para hacer un anova de una vía? ¿Cuáles serían los y a que niveles? Si sabemos que el estadıstico del test de la *F* que se obtiene es 11.82, ¿Qué conclusión sobre el experimento se obtiene con $\alpha = 0.05$?

9.10. La tabla siguiente presenta el resultado del análisis de varianza de la esperanza de vida del hombre por zona geográfica (ver ejercicio 9.5)

Fuente	Suna cuadrados	G.L.	Cuadrado medio	F	p
Entre grupos	5675,93	5	1135,19	32,14	0,0000
Dentro	3178,43	90	35,3159		
Tal	8854,36	95			



Teniendo en cuenta los resultados del Anova y las gráficas de cajas en la figura adjunta, ¿Qué conclusión se puede obtener sobre la esperanza de vida en diferentes zonas geográficas?

Escribe formalmente las hipótesis en este contraste.

Estimación de los efectos del factor

Una vez rechazada la hipótesis nula, pueden calcularse intervalos de confianza para los valores α_i. Puesto que, según hemos supuesto, las *k* poblaciones tienen una varianza común σ, y ésta viene estimada por *CMD*, aplicando el cálculo del intervalo de confianza de la media de una población, cuando no se conoce la desviación típica, obtenemos para α_i el intervalo de confianza (9.6), con coeficiente de confianza (1-ε). En dicha expresión t_ε es el percentil del *(1-ε)100%* de la distribución *T* con *n-k* grados de libertad.

(9.6) $$\bar{x}_i - \frac{t_\varepsilon \sqrt{CMD}}{\sqrt{n_i}} \le \mu_i \le \bar{x}_i + \frac{t_\varepsilon \sqrt{CMD}}{\sqrt{n_i}}$$

Ejemplo 9.3. Estimaremos la puntuación total media de 6º curso, en el ejemplo 9.1. Puesto que en este caso:

x_i=26.644 n_i=90 $\sqrt{cmd}$=6,2157

y los grados de libertad correspondientes para la distribución *T* son 245, podemos aproximar ésta por la normal. Para un intervalo del 95% de confianza *T*=1,96, por lo que se obtiene un intervalo:

26,644±6,2157*1,96/√90=26,644±1,2842= (25,3598, 27,9882)

También puede hallarse un intervalo de confianza para la diferencia de efectos entre dos tratamientos, que coincide con la diferencia de medias entre las dos subpoblaciones, mediante la expresión (9.7).

$$(9.7) \qquad \bar{x}_i - \bar{x}_j - t_\varepsilon \sqrt{CMD\left(\frac{1}{n_i}+\frac{1}{n_i}\right)} \le \mu_i \le \bar{x}_i - \bar{x}_j + t_\varepsilon \sqrt{CMD\left(\frac{1}{n_i}+\frac{1}{n_i}\right)}$$

Ejemplo 9.4. Estimaremos la diferencia de puntuaciones medias en los curso 6º y 7º, con un coeficiente de confianza del 95%. En este caso el número de grados de libertad es 90+87-2=175, por lo que de nuevo aproximaremos *T*=1,96 y x_i-x_j=3,517, obteniendo un intervalo 3,517±1,832

Actividades

9.11. La tabla siguiente presenta los intervalos LSD para las medias de la esperanza de vida del hombre en diferentes zonas geográficas (ver ejercicio 9.5). ¿Qué se puede concluir sobre la existencia de diferencias estadísticamente significativas?

				Intervalos LSD del 95%	
grupo	N	Media	D. Típica	L. Inferior	L. Superior
1	11	67,6909	1,7918	65,1738	70,208
2	12	62,7083	1,71552	60,2984	65,1183
3	19	71,5	1,36335	69,5848	73,4152
4	11	64,8182	1,7918	62,3011	67,3353
5	16	60,1312	1,48568	58,0442	62,2183
6	27	50,637	1,14368	49,0304	52,2437

9.12. En la tabla adjunta se presentan los intervalos LSD de confianza del 95% para la diferencia del tiempo en responder un test para tres grupos de niños (reflexivos, impulsivos y normales). Si la

Contraste	Limites de las diferencias	
1-2	*-14,5731	5,15826
1-3	-3,58974	6,05258
2-3	*10,9833	6,3328

media del grupo 1 es 7,0793 hallar las medias de los otros grupos e indicar qué diferencias son estadísticamente significativas.

Comparaciones específicas de medias de tratamientos

Una vez rechazada la hipótesis de igualdad de las medias en los diferentes tratamientos, estaremos interesados, en general, en decidir si algunos tratamientos prefijados son o no significativamente diferentes. Para contrastar las hipótesis:

$H_0 \equiv \alpha_i = \alpha_j$

$H_1 \equiv \alpha_i \neq \alpha_j$

adoptaremos la siguiente regla de decisión: Si la diferencia de medias entre los grupos *i*,*j* es menor que $t_{\alpha/2}\sqrt{CMD\left(\frac{1}{n_i}+\frac{1}{n_i}\right)}$) aceptamos la hipótesis nula, y en caso contrario la rechazamos. Nótese que hemos realizado una prueba *T* de diferencia de medias, estimando la varianza común mediante *CMD*. Puesto que en el cálculo de *CMD* hay *n-k* grados de libertad, estos serán los utilizados.

Ejemplo 9.5. En el análisis de varianza del ejemplo 9.1, para contrastar con un nivel de significación del 5 por ciento las hipótesis:

$H_0 \equiv$La puntuación media es igual en 6º que en 7º

$H_1 \equiv$Estos cursos tienen diferente puntuación media,

adoptaremos la siguiente regla de decisión: Si la diferencia entre las medias es menor que $t_{\alpha/2}\sqrt{CMD\left(\frac{1}{n_i}+\frac{1}{n_i}\right)}$=1,83169, aceptamos la hipótesis nula, y la rechazaremos en caso contrario.

Puesto que hemos obtenido una diferencia igual a 3,517, rechazaremos la hipótesis nula, a un nivel de significación del 5%. Para hallar el nivel mínimo de significación de este contraste, realizamos los cálculos siguientes:

$$t_{\alpha/2}\sqrt{CMD\left(\frac{1}{n_i}+\frac{1}{n_i}\right)}=1,83169/1,96=0,9345$$

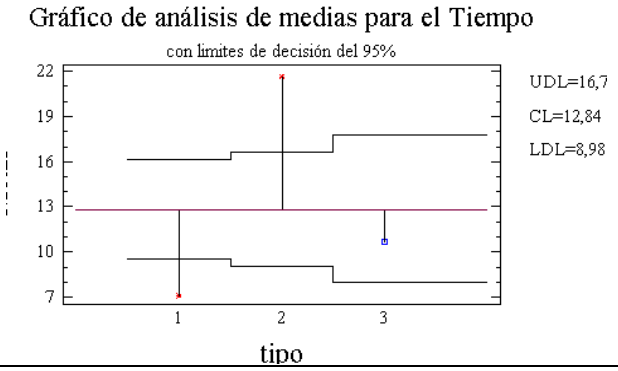
t=3,517/0,9345=3,7635 que corresponde a un nivel de significación de 0,0001.

Actividades

9.13. La tabla siguiente presenta los contrastes LSD para las medias de la esperanza de vida del hombre en diferentes zonas geográficas (ver ejercicio 9.5). Compara los resultados con las conclusiones obtenidas en el ejercicio 9.10.

Contraste	Limites de las diferencias		Contraste	Limites de las diferencias	
1-2	*4,98258	4,92822	3-4	*6,68182	4,47301
1-3	-3,80909	4,47301	3-5	*11,3688	4,00599
1-4	2,87273	5,03421	3-6	*20,863	3,53535
1-5	*7,55966	4,62422	4-5	*4,68693	4,62422
1-6	*17,0539	4,22305	4-6	*14,1811	4,22305
2-3	*-8,79167	4,35338	5-6	*9,49421	3,72482
2-4	-2,10985	4,92822			
2-5	2,57708	4,50859			
2-6	*12,0713	4,09612			

9.14. El gráfico adjunto analiza las medias del tiempo en el cuestionario para los tres grupos de alumnos del ejercicio 9.12. Interprete el gráfico y compare los resultados con los del ejercicio 9.12.



Comparaciones múltiples

El contraste anterior puede aplicarse para probar todas las posibles diferencias entre medias, cuando el número de éstas es pequeño. Sin embargo, con un número grande de comparaciones, crece la probabilidad de detectar como diferencias significativas algunas que no lo son realmente. Si, por ejemplo, tenemos 10 pares de medias, es de esperar 0,05*45=2 diferencias significativas simplemente por azar.

El problema de las "comparaciones múltiples" ha recibido mucha atención por parte de diversos investigadores en Estadística. Consiste en el hallazgo de un tipo de test que permitan efectuar una serie de comparaciones, manteniendo un nivel de significación global α prefijado.

Entre estos diversos métodos, expondremos el de Scheffé, que permite efectuar contrastes de tipo muy general.

Definición. Un contraste entre los parámetros $\alpha_1,....,\alpha_k$ es una función lineal Φ de los α_i tal que :

$$\Phi=\Sigma\alpha_i\, c_i \quad y\ \Sigma c_i=0$$

Un contraste como el definido, admite como estimador insesgado el siguiente: $\Phi = \sum \overline{x_i} c_i$ y su varianza se estima por:

(9.8) $$\sigma_\Phi^2 = CMD\sum \frac{c^2{}_i}{n_i}$$

Además, se verifica el siguiente teorema

Teorema: Existe una probabilidad 1-α de que todos los contrastes satisfagan a la vez las desigualdades (9.9).

(9.9) $$\Phi - \sigma_\Phi\sqrt{(k-1)F} \le \Phi \le \Phi - \sigma_\Phi\sqrt{(k-1)F}$$

siendo F el percentil del (1-α/2)100% de la distribución F con k-1, n-k grados de libertad.

Ejemplo 9.6. Compararemos en el ejemplo 9.1 todas las posibles diferencias entre medias, utilizando el método de Scheffé.

Cada una de las comparaciones μ_i-$\mu_j=0$ es un contraste, pues en este caso los coeficientes *c* son iguales a 1 y -1 y, por tanto su suma es nula. La varianza de dicho contraste viene dada por:

$$\sigma_\Phi^2 = CMD\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

Es decir, efectuando operaciones, obtenemos para cada uno de los contrastes:

$H_0 \equiv \mu_1-\mu_2=0$

$H_1 \equiv \mu_1-\mu_2\neq 0 \quad \sigma_\Phi=0{,}9345376$

$H_0 \equiv \mu_1 - \mu_3 = 0$

$H_0 \equiv \mu_1 - \mu_3 \neq 0 \quad \sigma_\Phi = 0{,}9866282$

$H_0 \equiv \mu_2 - \mu_3 = 0$

$H_0 \equiv \mu_2 - \mu_3 \neq 0 \quad \sigma_\Phi = 0{,}9941016$

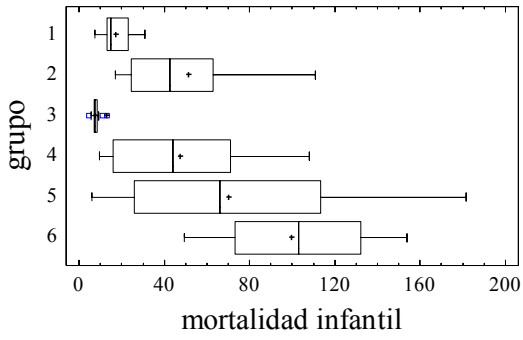
Teniendo en cuenta que para 2 y 245 g.l. los valores críticos de la distribución *F* son, *F*=3 al nivel 5% y *F*=3.6 al nivel 2.5%, y aplicando (9.9), obtenemos la tabla siguiente:

		Mínima diferencia significativa	
Cursos comparados	Diferencia de medias	nivel 5%	nivel 2.5%
6º y 7º	3,517	2,8036	3,4484
6º y 8º	3,159	2,9588	3,6406
7º y 8º	0,358	2,9823	3,6682

A la vista de esta tabla, deducimos que existe una diferencia significativa al nivel 2,5% entre los cursos 6º y 7º y al nivel 5% entre los cursos 6º y 8º, no habiendo diferencia entre 7º y 8º. Comparando con el ejemplo anterior, vemos que se obtiene un valor mayor para el nivel de significación del contraste de cada diferencia particular con este método. Ello es debido a que de esta forma logramos mantener a un nivel fijo la probabilidad global de error.

Actividades

9.15. En la figura adjunta se pres
la distribución de la tasa
mortalidad infantil en diferentes z
geográficas y a continuación la
de análisis de varianza.
conclusiones puede obtenerse sob
diferencia de tasa de morta
infantil en distintas zonas geográf



Fuente	Suna cuadrados	G.L.	Cuadrado medio	F	p
Entre grupos	116526,0	5	23305,2	24,49	0,0000
Dentro	85641,3	90	951,57		
Tal	202167,0	95			

9.3.MODELO DE EFECTOS ALEATORIOS

En este caso, suponemos que hemos seleccionado aleatoriamente k niveles del factor de una población infinita de posibles valores para el mismo. Esto ocurre cuando, por ejemplo, tomamos al azar k animales de una especie, para estudiar la variabilidad de algún parámetro dentro de la misma. El problema que se plantea ahora es que no estamos interesados únicamente en los n animales estudiados, sino en todos los elementos de la población que representan. Teóricamente, el modelo planteado es el siguiente:

$$x_{ij}=\mu+\alpha_i+\varepsilon_{ij}$$

en donde α_i son v. a. independientes y $N(0,\sigma_\alpha)$, ε_{ij} son v.a. independientes y $N(0,\sigma)$, ε_{ij} y α_i son independientes, μ es una constante, que representa la media global.

Mientras que en el modelo de efectos fijos estamos interesados en contrastar la hipótesis de que todos los niveles del factor producen el mismo efecto y, en caso contrario, en estimar los valores α_i, en el modelo de efectos aleatorios estamos interesados en comprobar si σ_α es igual a cero, y en caso contrario, en estimar su valor.

Procedimiento de cálculo

Para realizar el análisis de la varianza con efectos aleatorios, se calcula, en primer lugar la tabla 9.2. Necesitamos, además, estimar la esperanza matemática de los cuadrados medios o cuadrados medios esperados, cuyos valores se muestran en la tabla 9.3.

Tabla 9.3. Estimación de componentes de la varianza

Fuente de variación	Suma de cuadrados
Entre grupos	$\sigma^2+C\sigma^2_\alpha$
Dentro de los grupos (residual)	σ^2

donde C viene dado por la expresión (9.10)

(9.10)
$$C=\frac{1}{k-1}\sum_{i=1}^{k}n_i-\frac{\sum_{i=1}^{k}n^2{}_i}{\sum_{i=1}^{k}n_i}$$

Para decidir entre las dos hipótesis:

$H_0 \equiv \sigma^2_\alpha = 0$

$H_0 \equiv \sigma^2_\alpha \neq 0$

calcularemos el valor F_{exp} dado en la tabla 9.2 y adoptaremos la siguiente regla de decisión: Si el valor es menor que el percentil $(1-\alpha)100\%$ de la distribución F con $k-1$ y $n-k$ grados de libertad, aceptamos la hipótesis nula, y en caso contrario la rechazamos. Puede observarse que el procedimiento de contraste es idéntico al seguido en el caso de efectos fijos, aunque es diferente la hipótesis contrastada.

Una vez rechazada la hipótesis nula, para estimar el valor σ^2_α utilizando los datos de la tabla 9.3, obtenemos:

(9.11) $$\sigma^2_\alpha = \frac{CME - CMD}{C}$$

donde C viene dado por la expresión (9.10).

Ejemplo 9.7. En la siguiente tabla presentamos los resultados de haber aplicado un análisis de varianza a la puntuación total alcanzada en el test de Green clasificando los niños por colegios.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Entre colegios	817,26	2	408,63	10,77
Dentro de colegios	9298,98	245	37,95	
Total	10116,23	247		

Puesto que los tres colegios disponibles son solamente una muestra de los posibles colegios de la provincia, el problema que se plantea ahora es decidir entre las siguientes hipótesis:

H_0 ≡Todos los colegios posibles tienen una puntuación homogénea, por lo que $\sigma^2_\alpha = 0$,

H_1 ≡Existe variabilidad entre colegios.

Del resultado del análisis se deduce la necesidad de rechazar la hipótesis nula, por lo que concluímos que σ_α=0. A continuación, procederemos a estimar su valor. Al ser el número de alumnos en los diferentes centros 80, 80 y 88, aplicando (9.10):

$$C = \frac{248}{2} - \frac{80^2 + 80^2 + 88^2}{248} = 41,66$$

Por último, deducimos de (9.11) el siguiente valor para σ^2_α:

$$\sigma^2_\alpha = (408,63\text{-}37,95)/41,16 = 9,005$$

Actividades

9.16. Considere el modelo de bloques aleatorizados, $y_{ij} = \mu + \alpha_i + \beta_j + u_{ij}$. Las dos hipótesis nulas habituales para los contrastes de análisis de la varianza implican que

a. Las medias poblacionales son iguales para todas las categorías del factor y todas las categorías del bloque
b. Las medias muestrales son iguales para todas las categorías del factor
c. Las medias muestrales son iguales para todas las categorías del bloque
d. Las medias muestrales son distintas para todas las categorías del factor

9.17. La siguiente tabla presenta el análisis de varianza de la variable "número de alumnos por colegio" en una muestra de colegios de la provincia de Jaén. ¿Cuántos colegios había en la muestra? ¿Por qué el modelo debe ser de efectos aleatorios? Estimar la varianza del número de alumnos por colegio en la población de colegios.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F	P
Entre colegios	1,85862E7	320	58081,9	516,28	0,0346
Dentro de colegios	112,5	1	112,5		
Total	1,85863E7	321			

9.4.- ANALISIS DE VARIANZA CON DOS FACTORES: MODELO DE EFECTOS FIJOS

Consideremos ahora los grupos clasificados por dos variables diferentes. Mediante el análisis de varianza de dos factores queremos estimar el efecto de cada una de estas variables, cuando mantenemos la otra bajo control.

Ejemplo 9.8. En el ejemplo 9.1 hemos llegado a la conclusión de que la puntuación total varía en los diferentes cursos escolares. Si sospechamos que en alguno de los cursos es diferente el número de varones al de hembras, podría pensarse que la diferencia obtenida viene motivada por una diferencia de intuición probabilística entre ambos sexos. Para descartar esta hipótesis conviene efectuar a los datos un análisis de la varianza con dos factores: sexo y curso escolar.

Definición: Llamamos interacción entre dos variables al efecto de una de ellas cuando éste depende del nivel de la otra. Así en el ejemplo anterior, existiría interacción si la mejora de puntuación de un curso al superior sólo se verificara en uno de los dos sexos. Consideraremos el modelo siguiente:

$$x_{ijk}=\mu+\alpha_i+ß_j+\delta_{ij}+\varepsilon_{ijk}$$

donde μ es una constante que representa la media global, α_i representa el efecto del nivel i del factor A, $ß_j$ representa el efecto de nivel j del factor B, δ_{ij} representa la interaccion del nivel i del factor A con el j del factor B, ε_{ijk} es el efecto del azar y se considera $N(0, \sigma)$.Suponemos, además, $\Sigma\alpha_i=\Sigma ß_j=\Sigma\delta_{ij}=0$.

En este modelo, podemos descomponer las desviaciones entre los valores observados y la media global de la muestra en la forma siguiente:

$$x_{ijk}-x=(x_{ijk}-x_{ij})+(x_{i.}-x)+(x_{.j}-x)+(x_{ij}-x_{i.}-x_{.j}+x)$$

En dicha descomposición, cada uno de los sumandos tiene la siguiente significación: $(x_{ijk}-x_{ij})$ representa la desviación de una observación respecto a la media de su grupo, y se conoce como término de error, (x_i-x) es la desviación de la media del grupo con nivel i en el factor A respecto a la media global, y representa el efecto de dicho nivel, $(x_{.j}-x)$ es el efecto del nivel j del factor B. El último sumando es el efecto de la interacción.

En el análisis de la varianza con dos factores estamos interesados en contrastar varios grupos de hipótesis. Para ello, procederemos en primer lugar al cálculo de la tabla del análisis que viene representada en la tabla 9.5.

Tabla 9.5. Análisis de varianza de dos vías

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Factor A	SCA	a-1	$CMA = \frac{SCA}{a-1}$	$F_A = \frac{SCA}{SCE}$
Factor B	SCB	b-1	$CMA = \frac{SCB}{b-1}$	$F_B = \frac{SCB}{SCE}$
Interacción	SCI	(a-1)(b-1)	$CMA = \frac{SCI}{(a-1)(b-1)}$	$F_I = \frac{SCI}{SCE}$
Error	SCE	n-ab	$CMA = \frac{SCE}{n-ab}$	
Total	SCT	n-1		

Contraste de hipótesis referente al factor A

Para probar las hipótesis:

$H_0 \equiv$ todas las α_i son iguales a cero,

$H_1 \equiv$ algún α_i es diferente de cero,

calculamos el estadístico F_A y aceptamos la hipótesis nula en el caso de que el valor obtenido sea menor que el percentil *(1-α/2)100%* de la distribución *F* con *a-1* y n-ab grados de libertad, rechazándola en caso contrario.

Contraste de hipótesis referente al factor B

Para probar las hipótesis:

$H_0 \equiv$ todas las $ß_j$ son iguales a cero

$H_1 \equiv$ algún $ß_j$ es diferente de cero

calculamos el estadístico F_B y aceptamos la hipótesis nula, en el caso de que el valor obtenido sea menor que el percentil *(1-α/2)100%* de la distribución *F* con *b-1* y *n-ab* grados de libertad, rechazándola en caso contrario.

Contraste de hipótesis sobre la interacción

Para probar las hipótesis:

$H_0 \equiv$ todas las δ_{ij} son iguales a cero

$H_0 \equiv$ algun δ_{ij} es diferente de cero

calculamos el estadístico F_I y aceptamos la hipótesis nula, en el caso de que el valor obtenido sea menor que el percentil *(1-α/2)100%* de la distribución *F* con *(a-1)(b-1)* y *n-ab* grados de libertad, rechazándola en caso contrario.

Ejemplo 9.9. En el siguiente cuadro se presentan los resultados de efectuar un análisis de varianza por sexo y curso, sobre la puntuación total en el test de Green.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F	Valor p
SEXO	22,2978	1	22,2978	,58	,4486
CURSO	692,4876	2	346,2438	8,94	,0002
INTER.	40,5151	2	20,2576	,52	,5932
ERROR	9406,5885	243	38,7102		

Analizando los resultados obtenidos, deducimos que el único efecto significativo sobre la puntuación total es el debido al curso escolar, no habiendo diferencias en cuanto a sexo, ni interacción entre los factores.

Una vez efectuado los contrastes anteriores, si resulta significativo el efecto de alguno de los factores, puede continuarse el estudio, aplicando los métodos de estimación y contraste que hemos estudiado en el caso de un solo factor, corrigiendo debidamente el número de g. l. y demás parámetros de las distribuciones utilizadas.

Actividades

9.18. En una empresa disponemos del salario medio de los hombre y del salario medio de las mujeres para cada grupo de edad. Queremos detectar si existe discriminación salarial y si varía de unos grupos de edad a otros. Para ello con estos datos podemos realizar un análisis estadístico utilizando:

a. Análisis de la Varianza de un factor, efectos fijos
b. Análisis de la Varianza de dos factores
c. Análisis de la Varianza de un factor, efectos aleatorio

Se ha experimentado la pérdida de peso de cuatro materiales M1, M2, M3 y M4, sujetos a tres condiciones C1, C2 y C3. El resultado del experimento viene recogido en la siguiente tabla:

	M1	M2	M3	M4
C1	11	-35	5	4
C2	40	11	43	6
C3	44	-12	0	-3

a. Construir tabla ANOVA de dos vias
b. Contrastar que los materiales son idénticos
c. Contrastar que las condiciones no influyen.

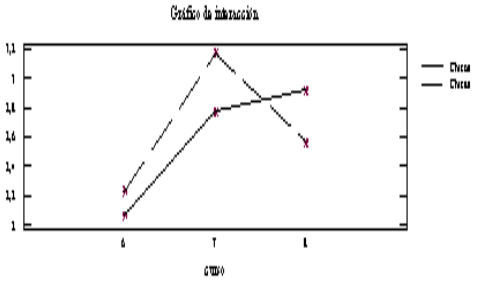
9.19. Indicar el máximo numero de factores variando en dos niveles que pueden analizarse con 8 observaciones:
a. 8
b. 7
c. 3

9.20.Un anova bifactorial es equilibrado si:
a. Los grupos que definen los dos factores tienen igual varianza
b. Los grupos tienen el mismo número de sujetos
c. Los grupos tienen la misma media
d. Los grupos fueron tomados aleatoriamente

9.21.En la tabla siguiente se presenta la tabla del análisis de varianza del número de alumnos por colegio en una muestra de colegios de la provincia de Jaén, clasificado por zona (rural/urbana) y tipo (privado/público). ¿Por qué en este caso se usa el modelo de efectos fijos? ¿Cuáles son los factores y sus niveles? ¿Hay efecto de alguno de los factores? ¿Y de la interacción?

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F	P
Zona	1173,85	1	1173,85	14,02	0,0002
Tipo	601,553	1	601,553	7,19	0,0077
Interacción	80,5555	1	80,5555	0,96	0,3273
Residual	26616,7	318	83,7002		
Total	31549,6	321			

9.22. El gráfico adjunto describe la interacción en un análisis de dos vias de la puntuación en combinatoria de un grupo de alumnos en función del género y curso escolar. ¿Cuáles son las variables dependiente y factores en el estudio? ¿Sugiere el gráfico la presencia de interacción y efecto de los factores?



9.5. COMPROBACION DE LAS HIPOTESIS DEL MODELO Y TRANSFORMACIONES A LOS DATOS.

Aunque el modelo del análisis de la varianza supone que las muestras disponibles provienen de poblaciones normales de igual varianza, puede ocurrir que en un caso concreto no se verifiquen estos supuestos, ni siquiera aproximadamente. Por tanto, es de interés conocer la forma de averiguar si estamos en este caso, y cual es la manera de proceder si uno de los supuestos no se cumple.

En muchas ocasiones se violan simultáneamente las hipótesis de normalidad y homocedasticidad. La explicación de ello es que, mientras en una distribución normal la varianza no depende del valor medio, hay otros tipos de distribuciones, como la binomial, Poisson, exponencial. en que estos dos parámetros son dependientes. Por ello, cuando queremos estudiar la diferencia de medias en varias poblaciones que son, por ejemplo de Poisson, nos encontraremos con que, además de violarse el supuesto de normalidad, las poblaciones tienen varianzas diferentes.

Una primera forma de saber si se cumplen o no las hipótesis requeridas para poder aplicar correctamente el análisis de la varianza, consiste en inspeccionar los datos. Así, la representación gráfica dela distribución de frecuencias nos permite observar visualmente la forma no simétrica de la misma, que hace sospechar que la variable proviene de una población no normal. Igualmente, en algunos casos es evidente la desigualdad de las varianzas. En caso de duda, puede aplicarse uno de los contrastes que existen para este fin.

Cuando en el estudio de los datos se revela una desviación acusada de la normalidad, o una gran diferencia entre las varianzas, se puede tratar de solucionar el problema efectuando a los datos una transformación adecuada. Todos los cálculos del análisis se harán con las variables transformadas. Para expresar las conclusiones obtenidas (por ejemplo los intervalos de confianza) en la escala primitiva, basta con efectuar a los resultados la transformación inversa. Las principales transformaciones utilizadas son tres:

Transformación logarítmica

Si las desviaciones típicas de los diferentes grupos son proporcionales a las medias de los mismos, se aplica el siguiente cambio de variable:

$$Y = log\ X$$

Transformación raiz cuadrada

Se emplea cuando las varianzas de los distintos grupos son función de los valores medios.

$$Y=\sqrt{x}$$

Transformación arco seno

Se usa cuando no producen efecto las anteriores, y los datos provienen de distribuciones binomiales

$$Y=arcsen(X)$$

Si, tras aplicar el cambio de variable, siguen incumpliéndose las hipótesis podemos seguir dos caminos:

- Si son pocos los grupos a comparar, pueden estudiarse dos a dos, utilizando el test *T*, teniendo en cuenta las recomendaciones que se han hecho en el estudio de las comparaciones múltiples.
- Puede aplicarse un método no paramétrico de comparación de varias muestras.

Actividades

9.23. El supuesto de independencia se aplica:

a. Sólo a las observaciones
b. Sólo a las muestras
c. A las observaciones y las muestras

9.24. El test de Bartlet se aplica para comprobar:

a. La independencia de las observaciones
b. La igualdad de las varianzas
c. La normalidad

9.25. Al aplicar el test de Bartlet se obtuvo un valor p=0,0001. ¿Se debe concluir que los grupos analizados tienen la misma varianza?

9.26. El gráfico adjunto presenta los residuos de la esperanza de vida de la mujer respecto a la media en diferentes zonas geográficas. ¿Es plausible la hipótesis de igualdad de varianzas a la vista del gráfico de residuos? ¿Qué implicaciones tendría si lo único que se

desea es contrastar la igualdad de las medias en las poblaciones?

9.27. El gráfico adjunto presenta los residuos de la puntuación combinatoria de un grupo de estudiantes en función del curso. ¿Es plausible la hipótesis de igualdad de varianzas a la vista del gráfico de residuos?

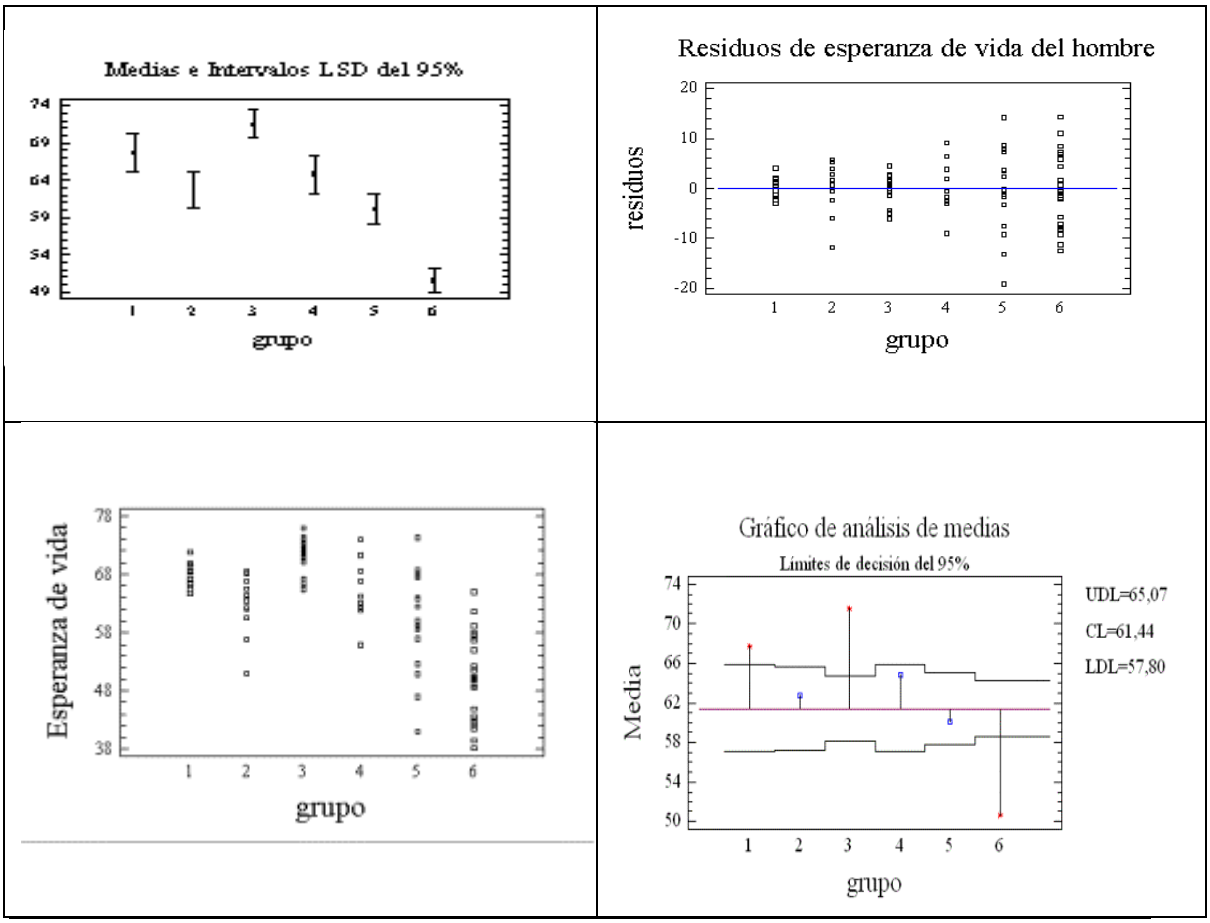
9.6. ANÁLISIS DE VARIANZA CON STATGRAPHICS

El paquete Statgraphics presenta una completa colección de programas de análisis de varianza, que incluye, anova de una y varias vias y estudio de los componentes de la varianza. Se encuentran dentro del programa COMPARAR, donde hay un submenú específico. Dentro del mismo aparecen tres subprogramas:

- *Anova de una vía:* Incluye en sus opciones tabulares el cálculo de estadísticos resumen, la tabla de análisis de varianza, tabla de medias con cálculo de intervalos (donde se puede elegir el coeficiente de confianza y tipo de intervalo, como LSD, Tukey, Bonferroni o intervalos de confianza con varianzas iguales o diferentes. Se proporciona también las compraciones múltiples LSD, Tukey, Bonferroni y otras, varias pruebas de igualdad de varianzas y análisis de varianza no paramétrico de Kruskal-Wallis.
- *Anova multifactorial:* Admite varios factores y se puede diseñar el modelo, eligiendo las interacciones y efectos que se desea probar. Además de la tabla anova, añade las tablas de medias con cálculo de intervalo y comparaciones múltiples, con las mismas posibilidades que el caso anterior.
- *Componentes de la varianza:* Calcula el porcentaje de varianza explicado por cada factor

También se presenta una completa colección de gráficos que pueden ayudar al análisis, incluyendo gráficos de puntos y cajas de la variable dependiente en función de los diferentes niveles del factor, intervalos LSD para las medias, gráficos de análisis de medias y varios gráficos de residuos. Algunos de estos gráficos se presentan en la Figura 9.2.

Figura 9.2. Algunas salidas gráficas en los programas de Anova de Statgraphics



TEMA 10
VARIABLES ESTADISTICAS BIDIMENSIONALES

10.1. DEPENDENCIA FUNCIONAL Y DEPENDENCIA ALEATORIA

Cuando se realiza un estudio estadístico, generalmente, se está interesado en más de un carácter de los individuos de la población. Una de las preguntas a las cuales se trata de dar respuesta es si existe alguna relación entre dos variables *X* e *Y*. Para algunos fenómenos, es posible encontrar una fórmula que exprese exactamente los valores de una variable en función de la otra: son los fenómenos llamados deterministas.

Ejemplo 10.1. Al estudiar la caída libre de un cuerpo, la Física ha encontrado que el espacio recorrido, *Y*, está relacionado con el tiempo desde su lanzamiento, *X*, por la expresión (10.1), siendo g la constante 9,8 m/s^2.

(10.1) $$Y = \tfrac{1}{2}\, g\, X^2$$

Este es un caso de dependencia funcional entre dos variables. Para este fenómeno, si realizamos el experimento de medir el espacio recorrido para valores del tiempo *X=5* seg, 10seg.,... hasta 30 seg., por ejemplo, obtendremos una tabla como la indicada en la Tabla 1. Si representamos en un sistema de ejes cartesianos estos pares de valores, se obtendrá una colección de 10 puntos (Figura 10.1), por los cuales es posible trazar la parábola cuya ecuación es dada por la fórmula (10.1).

Tabla 10.1. Datos sobre caída libre de un cuerpo

X (seg)	*Y* (mts.)
5	122,5
10	490,0
15	1102,5
20	1960,0
25	3062,5
30	4410,0

Figura 10.1: Curva ajustada a los datos

En este tipo de relación, los valores que toma la variable *Y* quedan determinados, de un modo preciso, por los valores que toma la otra variable, que se considera como independiente.

Dependencia aleatoria

Existen muchos fenómenos en los que, al observar pares de valores correspondientes a variables estadísticas, no es posible encontrar una fórmula que relacione, de un modo funcional, esas variables. Si dichos pares de valores los representamos en un sistema cartesiano, los puntos, en general, no se ajustan de un modo preciso a una curva plana, sino que se obtiene un conjunto de puntos más o menos dispersos. Una representación de ese tipo recibe el nombre de *nube de puntos o diagrama de dispersión*.

Ejemplo 10.2. En las figuras 10.2 y 10.3 hemos representado la esperanza de vida del hombre en función de la tasa de mortalidad y el PNB para cada un conjunto de países.

Figura 10. 2

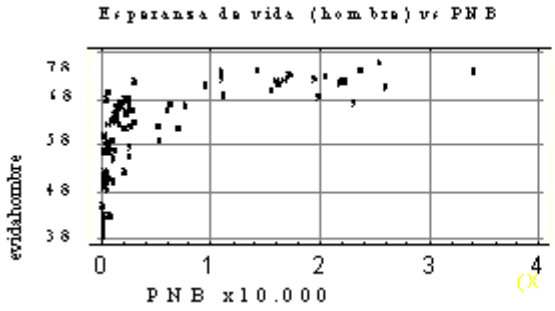
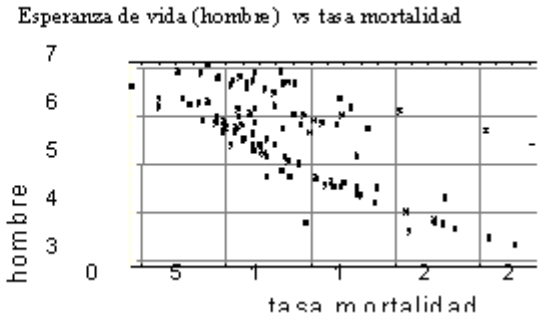


Figura 10. 3



Aunque puede apreciarse que en ninguno de los dos casos es posible encontrar una relación funcional entre las dos variables, sin embargo, observamos una variación conjunta de las variables. En el primer caso la relación es directa, puesto que al crecer el PNB crece la esperanza de vida y en el segundo inversa (la esperanza de vida decrece al aumentar la tasa de mortalidad). Mientras que en el segundo caso podríamos aproximar la relación entre las variables mediante una recta (dependencia lineal) en el primero habría que usar otro tipo de función, posiblemente una parábola o una función exponencial.

Figura 10.4

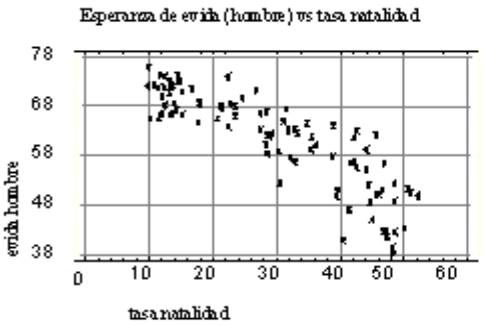
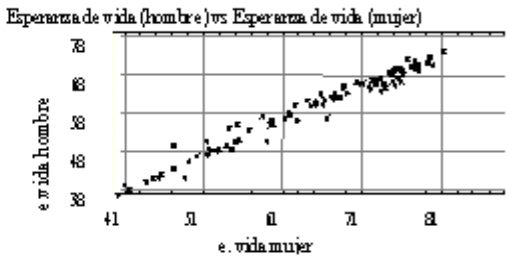


Figura 10.5



Actividades

10.1. En las Figuras 10.4 y 10.5 hemos representado la esperanza de vida del hombre en una serie de países en función de otras dos variables. Discute en cada caso si la relación es directa o inversa, lineal o no. ¿Respecto a cuál variable la relación es más intensa? ¿Cuál serviría mejor para predecir la esperanza de vida del hombre? ¿Qué significa para ti una causa y un efecto? ¿En qué casos de los mostrados en las figuras 10.2 a 10.5 consideraría la relación entre la esperanza de vida del hombre y otra variable de tipo causal ?

10.2. EL CONCEPTO DE ASOCIACIÓN

El estudio de la posible relación entre dos variables cuantitativas suele iniciarse mediante la observación del correspondiente diagrama de dispersión o "nube de puntos". La presencia de una relación entre las variables se pondrá de manifiesto en el diagrama por una cierta tendencia de los puntos a acumularse en las proximidades de una línea, como hemos visto en los ejemplos anteriores.

En otros casos nos interesa analizar si dos variables cualitativas están relacionadas entre sí, como en la actividad 10.2, o si una variable cuantitativa está relacionada con otra cualitativa como en la actividad 10.3.

Actividades

10.2. Se quiere estudiar si un cierto medicamento produce trastornos digestivos en los ancianos. Para ello se han observado durante un período suficiente de tiempo a 25 ancianos obteniendo los siguientes resultados de la Tabla 10.2. Utilizando los datos de la tabla, razona si en estos ancianos, el padecer trastornos digestivos está relacionado con haber tomado o no el medicamento, indicando cómo has usado los datos.

Tabla 10.2. Sintomatología digestiva según se toma o no una medicina

	Molestias digestivas	No tiene molestias	Total
Toma la medicina	9	8	17
No la toma	7	1	8
Total	16	9	25

10.3. Al medir la presión sanguínea antes y después de haber efectuado un cierto tratamiento médico a un grupo de 10 mujeres, se obtuvieron los valores de la Tabla 10.3. Utilizando los datos de la tabla estudia si la presión está relacionada con el momento en que se toma (antes o después del tratamiento).

Tabla 10.3. Presión sanguínea antes y después de un tratamiento

	Presión sanguínea en cada mujer									
Mujer	A	B	C	D	E	F	G	H	I	J
Antes del tratamiento	115	112	107	119	115	138	126	105	104	115
Después del tratamiento	128	115	106	128	122	145	132	109	102	117

Al tratar de estudiar si existe o no una relación entre dos variables estadísticas, tratamos de contestar a las preguntas siguientes:

- ¿Hay algún tipo de relación entre las variables?
- ¿Podría medir la intensidad de esta relación mediante un coeficiente (coeficiente de asociación)?
- ¿Sirve este coeficiente para poder comparar la intensidad de la relación de diferentes variables? ¿Cómo puedo interpretarlo?

En los ejemplos anteriores hemos visto que se nos pueden presentar tres tipos de estudio de la relación entre variables según la naturaleza de las mismas:

- Dos variables cualitativas, como en la actividad 10.2. Estudiaremos la asociación entre las variables cualitativas mediante el análisis de las *tablas de contingencia*.
- Una variable cuantitativa y otra cualitativa, como en la actividad 10.3. Observa que lo estudiado en los temas anteriores nos podría servir para analizar de forma intuitiva la asociación entre estos tipos de variables. Por ejemplo, analizando la diferencia entre las dos medias y comparando los intervalos de confianza de las dos medias podríamos deducir si existe asociación en estos casos. Hay otros procedimientos estadísticos específicos, como el test *T* de diferencias de medias.
- Dos variables cuantitativas como en los ejemplos 10.1 y 10.2. En este caso específico, si las variables están relacionadas, hablamos de *correlación* entre las variables.

Mediante la observación de los diagramas de puntos podemos obtener alguna información sobre la correlación entre las variables numéricas *X, Y.*

Las figuras 10.1, 10.2 y 10.4 sugieren que los valores de *Y* crecen en promedio, a medida que los de *X* aumentan, y que, por tanto, la regresión de *Y* sobre *X* es directa. La diferencia entre estos casos es que en la figura 1 los puntos están sobre la curva, porque la relación es de tipo funcional, mientras

que en los demás casos la relación es de tipo aleatorio. En la figura 10.4 los puntos están mucho más cerca de la línea de regresión, porque la correlación entre las variables es más intensa. En las figuras 10.3 y 10.5 la relación sería inversa. Podría darse el caso de que no se observara ninguna relación entre las variables y hablaríamos de independencia.

10.3. TABLAS DE CONTINGENCIA

En algunos estudios estadísticos tomamos para cada individuo valores de dos variables estadísticas: *X* que toma los *valores* x_1, x_2,.. x_r, e *Y*, que toman valores y_1, y_2, ..., y_c. Podemos escribir los datos recogidos en forma de *listado*, como se indica en la Figura 10.5 o bien, cuando todos o algunos pares se repiten, pueden escribirse como una *tabla de doble entrada* (o tabla de contingencia) (Figura 10.6), donde f_{ij} indica la frecuencia absoluta con que aparece el par (x_i, y_j).

Si representamos mediante h_{ij} la frecuencia relativa del par de valores (x_i, y_j), se verifica la relación siguiente. Llamamos a esta frecuencia relativa de cada celda respecto al total de datos *frecuencia relativa doble.*

$$h_{ij}=f_{ij}/n$$

Figura 10.6

X	Y
x_1	y_1
x_2	y_2
.	.
x_i	y_i
x_n	y_n

Figura 10.7

	y_1	y_j	y_c	
x_1				$f_{1.}$
x_2				$f_{2.}$
.				
x_i		f_{ij}		$f_{i.}$
x_r				$f_{r.}$
	$f_{.1}$	$f_{.j}$	$f_{.c}$	n

Distribuciones marginales y condicionadas

A partir de la tabla de frecuencias bidimensional (figura 10.6), pueden obtenerse diferentes distribuciones unidimensionales. Si en la tabla de frecuencias se suman las frecuencias por columnas, obtengo en cada columna j, el número de individuos $f_{\cdot j}$ con un valor de la variable $Y=y_j$, independientemente del valor X. A la distribución así obtenida se le conoce como *distribución marginal* de la variable Y. De forma análoga podemos definir la distribución marginal de la variable X.

Ejemplo 10.3. Al clasificar una serie de modelos de automóviles por el número de cilindros y su origen se obtuvo la tabla 10.4. De ella podemos obtener dos distribuciones condicionales.

Sumando por filas obtenemos la distribución de coches según su origen y sumando por columnas la distribución de coches según su número de cilindros. Nótese que, al ser estas distribuciones de una sola variable, podemos realizar con ellas gráficas y tablas, como mostramos en las Figuras 10.8 y 10.9.

Tabla 10.4. Distribución del número de cilindros de automóviles según origen

	N. cilindros			
Origen	4	6	8	Total
Europa	140	57	51	248
E.U.	40	12	20	72
Japón	27	15	36	78
Total	207	84	107	398

Figura 10.8

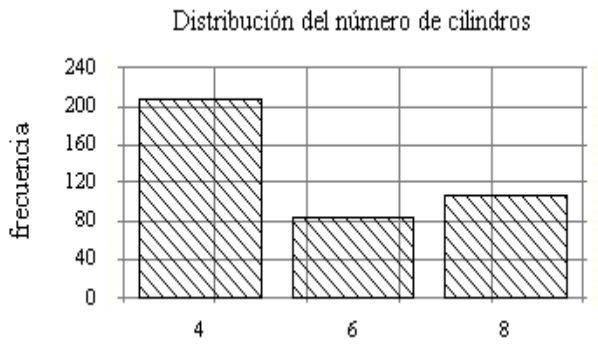
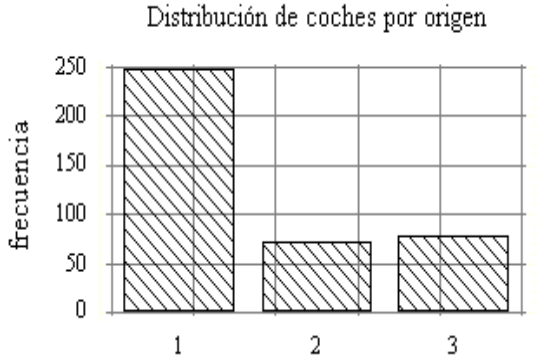


Figura 10.9

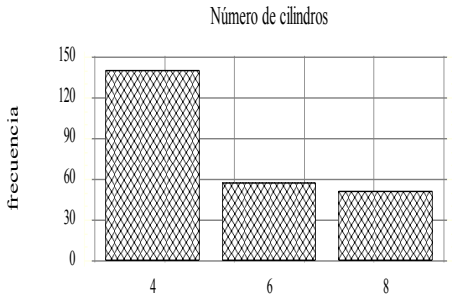


En particular, si *X* o *Y* son cuantitativas, podríamos calcular su media y varianza. Así, por ejemplo, el número medio de cilindros de todos los coches en el ejemplo 10.3 es 5,49 y su varianza 2,91. Otro tipo de distribución para la variable *X* es la que puede obtenerse fijando un valor $Y=y_j$, que se conoce como *distribución de X condicionada* para $Y=y_j$. Así en la Tabla 10.4 podríamos analizar la distribución de coches europeos según el número de cilindros, y obtener la tabla 10.5 y Figura 10.10.

Tabla 10.5. Número de cilindros en coches europeos

	N. cilindros			
	4	6	8	Total
Frecuencia	140	57	51	248
Porcentaje	56.45	22.98	20.56	

Figura 10.10. Número de



Igualmente podríamos haber obtenido la distribución del número de cilindros para los coches americanos o japoneses. Es decir, existen tantas distribuciones condicionadas diferentes para la variable *Y*, como valores distintos toma *X*.

Observamos que la frecuencia absoluta de la distribución de *X* condicionada por un valor de $Y=y_j$ coincide con f_{ij}, es decir, con la de la variable bidimensional. Si representamos por $h(x_i|y_j)$ la frecuencia relativa condicional del valor x_i entre los individuos que presentan el carácter y_j, obtenemos la igualdad (10.2).

$$h(x_i|y_j) = \frac{f_{i,j}}{f_{\cdot j}} = \frac{h_{i,j}}{h_{\cdot j}} \qquad (10.2)$$

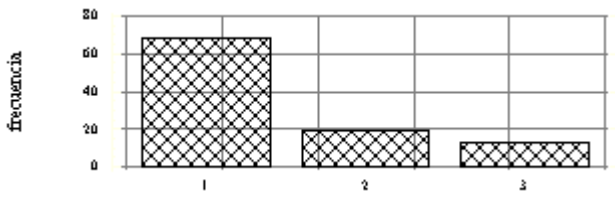
En el ejemplo anterior hemos hallado la distribución condicional de la variable *Y* en función de uno de los valores de *X*, es decir, la distribución condicional de filas en la tabla de contingencia, cuando sólo tomamos los datos de una de las columnas.

Podríamos intercambiar los papeles de filas y columnas y obtener la distribución condicional de *X* en función de alguno de los valores de *Y*. En la tabla 10.4, podríamos obtener la distribución del origen de los coches de 4 cilindros, obteniendo la tabla 10.6 y figura 10.11.

Tabla 10.6. Origen de coches de 4 cilindros

Origen	Frecuencia	Porcentaje
Europa	140	67.63
E.U.	40	19.32
Japón	27	16.04
Total	207	

Figura 10.11. Coches de 4 cilindros



Podemos ahora obtener la frecuencia relativa de y_j condicionada por $x=x_i$ mediante la expresión (10.3):

$$h(y_j|x_i) = \frac{f_{i,j}}{fi.} = \frac{h_{i,j}}{hi.} \qquad (10.3)$$

Como consecuencia se verifica la igualdad (10.4) que nos permite obtener la frecuencia relativa doble a partir de las condicionales y marginales.

$$h_{i,j} = h(x_i|y_j)\, h_{.j} = h(y_j|x_i)\, h_{i.} \qquad (10.4)$$

Actividades

10.4. Se quiere estudiar si un cierto medicamento produce trastornos digestivos en los ancianos. Para ello se han observado durante un periodo suficiente de tiempo a 25 ancianos obteniendo los siguientes resultados:

	Molestias digestivas	No tiene molestias	Total
Toma la medicina	9	8	17
No la toma	7	1	8
Total	16	9	25

Utilizando los datos de la tabla, razona si en estos ancianos, el padecer trastornos digestivos depende o no del medicamento.

10.5. En la siguiente tabla se muestra la edad actual de un grupo de pacientes clasificados por sexos. Calcular la edad media de las distribuciones condicionadas de varones y hembras.

EDAD	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Hombre	6	9	38	49	38	14	13	17
Mujer	6	12	25	23	29	23	7	1

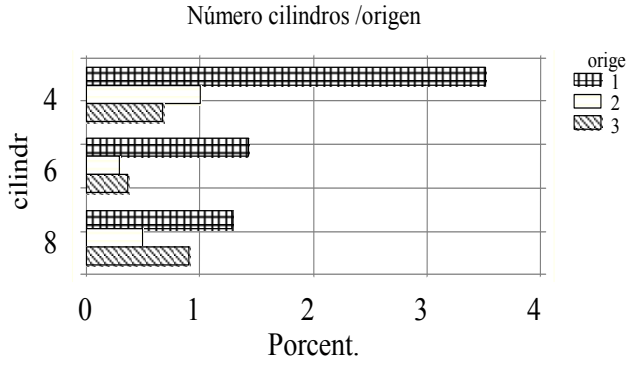
10. 4. TABLAS DE CONTINGENCIA Y REPRESENTACIONES ASOCIADAS EN STATGRAPHICS

Con Statgraphics podemos realizar una tabla bidimensional para dos variables cualitativas. Para ello elegiremos el menú DESCRIPCIÓN, en la opción DATOS CUALITATIVOS – TABULACIÓN CRUZADA. Aparecerá un cuadro de diálogo en el que se debe elegir la variable que irá en las filas y la variable que se tomará en las columnas. Por ejemplo, si elegimos como filas el número de cilindros y columna el origen de los coches de la muestra analizada en el ejemplo 10.3, obtendremos la tabla 10.7.

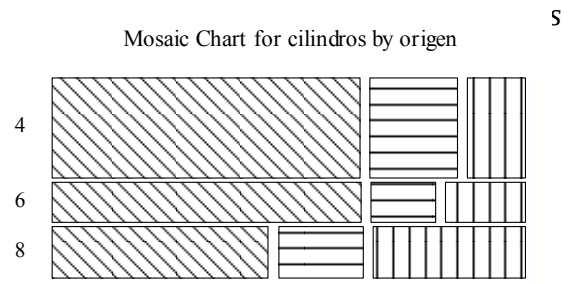
Tabla 10.7. Tabla de frecuencias

Row	1	2	3	Total
4	140	40	27	207
	35.18	10.05	6.78	52.01
6	57	12	15	84
	14.32	3.02	3.77	21.11
8	51	20	36	107
	12.81	5.03	9.05	26.88
Column	248	72	78	398
Total	62.31	18.09	19.60	100.00

Figura 10.12. Diagrama de barras adosado



Es importante fijarse que, además de las frecuencias absolutas dobles y marginales, la tabla proporciona las frecuencias relativas dobles, esto es, respecto al total de datos o h_{ij}. Podemos comprobarlo al ver que sumando todas estas frecuencias relativas obtendremos 100. Para cada fila, aparece el total de la fila y la frecuencia relativa de la fila respecto al total de datos (*frecuencia relativa marginal* $f_{i.}$ de la fila i). Para cada columna obtenemos el total de la columna y la frecuencia relativa de la columna respecto al total (*frecuencia relativa marginal* $f_{.j}$ de la columna *j*).

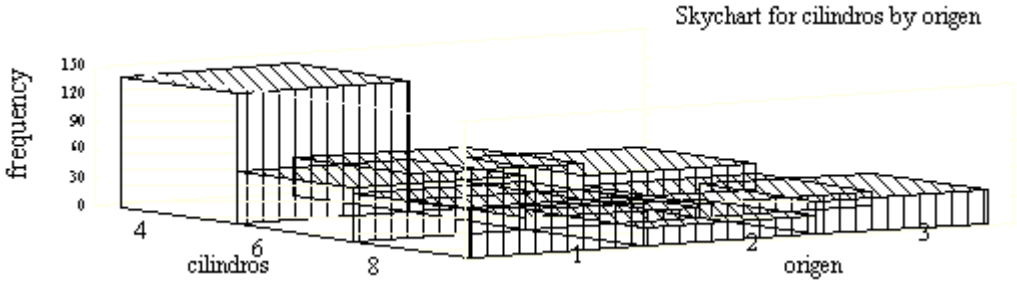


Como opciones gráficas tenemos el diagrama de barras adosado (Figura 10.12), que clasifica los datos primeramente por filas en la tabla y dentro de cada fila por columnas. Este diagrama puede cambiarse a formato apilado y en lugar de frecuencias absolutas a porcentajes, pero éstos siempre se refieren al total de los datos. Otros dos gráficos disponibles son el gráfico de mosaicos e histograma tridimensional.

En el gráfico de mosaicos (Figura 10.13) se divide primero el eje y segmentos proporcionales a la frecuencia relativa de cada categoría en filas, es decir, en el eje y tenemos representadas las frecuencias relativas marginales de la variable en filas. En cada uno de los rectángulos resultantes se divide el eje x en partes proporcionales a la frecuencia condicional de las columnas de la tabla respecto a la fila dada. Es decir, el gráfico de mosaico visualiza las frecuencias marginales de las filas y las frecuencias condicionales de columnas respecto a cada una de las filas de la tabla.

En el diagrama tridimensional de barras, se representan en dos ejes las categorías de filas y columnas y en el eje *Z* las frecuencias dobles (Figura 10.14).

Figura 10.14. Diagrama de barras tridimensional



Actividades

10.6. Compara las tres representaciones gráficas obtenidas del programa TABULACION CRUZADA en las figuras 10.10, 10.11 y 10.12. ¿Cuál de ellas representa las frecuencias relativas dobles? ¿Cuál de ellas representa las frecuencias relativas condicionales?

Distribuciones condicionadas por filas y columnas

Para obtener en la tabla de contingencia las distribuciones condicionadas podemos usar la opción OPCIONES DE VENTANA. Si pedimos que los porcentajes de la tabla se calculen respecto al total de las filas, obtendremos la distribución condicionada de columnas respecto a cada una de las filas (como se muestra en la tabla 10.8, donde se presentan las distribuciones condicionadas del origen de los coches de 4, 6 y 8 cilindros). En este caso la suma de los porcentajes dentro de las diferentes celdas de una misma fila suma 100.

Tabla 10.8. Distribuciones condicionadas por filas y columnas

	1	2	3	Total
4	140	40	27	207
	67.63	19.32	13.04	52.01
6	57	12	15	84
	67.86	14.29	17.86	21.11
8	51	20	36	107
	47.66	18.69	33.64	26.88
	248	72	78	398
	62.31	18.09	19.60	100.00

	1	2	3	Total
4	140	40	27	207
	56.45	55.56	34.62	52.01
6	57	12	15	84
	22.98	16.67	19.23	21.11
8	51	20	36	107
	20.56	27.78	46.15	26.88
	248	72	78	398
	62.31	18.09	19.60	100.00

Si pedimos que las frecuencias relativas se calculen respecto al total de las columnas obtenemos las distribuciones condicionales de las filas respecto a cada una de las columnas (en la tabla 10.8 se presentan las distribuciones condicionadas del número de cilindros en los coches según su origen). En este caso al sumar las frecuencias relativas de una misma columna obtenemos la suma 100.

Intercambio de filas y columnas

Si intercambiamos filas y columnas en la ventana de entrada de variables, observaremos un cambio en la tabla y gráficos. La primera variable de clasificación es siempre la variable que situamos en las filas y la variable en columnas se usa como segunda variable de clasificación.

Actividades

10.7. En una Facultad se preguntó a los alumnos si fumaban y también si fumaban sus padres, obteniéndose los siguientes datos:

	El alumno fuma	El alumno no fuma
Los dos padres fuman	400	1380
Sólo fuma uno de los padres	416	1823
Ninguno de los dos padres fuma	188	1168

Compara la distribución de alumnos fumadores y no fumadores, según fumen los dos padres, uno sólo o ninguno. ¿Piensas que hay alguna relación entre si los padres fuman o no y si fuman los hijos?

10.5. DEPENDENCIA E INDEPENDENCIA

El mayor interés del estudio de las distribuciones condicionadas, es que a partir de ellas estamos en condiciones de definir el concepto de dependencia aleatoria.

Diremos que la variable X es independiente de Y si todas las distribuciones de frecuencias relativas que se obtienen al condicionar X por diferentes valores de $Y= y_j$ son iguales entre si, e iguales a la distribución marginal de la variable X, es decir, cuando se verifica (10.5) para todo par de valores i, j.

(10.5) $$h(x_i|y_j)= h_i.$$

Esta propiedad significa que todas las distribuciones condicionales por columna coinciden con la distribución marginal de la variable *X* o lo que es lo mismo, la distribución de *X* no cambia cuando se condiciona por un valor de *Y*.

En el caso de independencia, se cumplen, además, las propiedades (10.6) a (10.8). La propiedad (10.7) quiere decir que la frecuencia relativa respecto al total en cada celda es igual al producto de las frecuencias relativas de su fila y su columna.

(10.6) $$h_{i,j}= h_i. h._j,$$ para todo *i, j*

(10.7) $$h(y_j|x_i)= h._j,$$ es decir *Y* no depende de *X*

La propiedad (10.7) indica que las distribuciones condicionales por filas son todas iguales y coinciden con la distribución marginal de la variable *Y*, es decir que la distribución de *Y* no cambia cuando condiciono por un valor de X. Finalmente la propiedad (10.8) nos da un método de cálculo de las frecuencias teóricas en caso de independencia

(10.8) $$f_{i,j} = \frac{f_i . f._j}{n}$$

Actividades

10.8. Demostrar las relaciones (10.6), (10.7) y (10.8)

10.9. En la siguiente tabla hemos clasificado un grupo de estudiantes por sexo y si va o no al cine asiduamente.

	Va al cine con frecuencia	Va al cine raramente
Chicos	90	60
Chicas	60	40

Comprueba si se cumplen las propiedades (6) a (9) en esta tabla. ¿Piensas que la afición al cine en esta muestra de estudiantes depende del sexo?

10.6. ANALISIS DE LAS TABLAS DE CONTINGENCIA

La utilización de tablas de doble entrada para las distribuciones de frecuencias de variables bidimensionales es obligada cuando las variables son discretas, con pocos valores distintos - lo que no justificar。a hacer un agrupamiento en intervalos- y cuando alguna de las variables es cualitativa, y por tanto, ha sido medida con escala nominal.

En este último caso, en la primera fila y la primera columna se indicaran las distintas modalidades de los caracteres que se estudian. La tabla 10.9 muestra la distribución de frecuencias para el par de variables cualitativas "Sexo" y "Fuma/ no fuma" para un grupo de estudiantes.

Tabla 10.9. Clasificacion por Sexo y Fuma/ N

	Fuma / No fuma		
	Fuma	No fuma	Total
Hombre	26	23	49
	53.1	46.9	
	50.0	65.1	52.7
Mujer	26	18	44
	59.1	40.9	
	50.0	43.9	47.3
Total	52	42	93
	51.9	410.1	100.0

La confección de tablas de doble entrada - o tabulaciones cruzadas - es una operación que se debe realizar siempre que se trate de estudiar la posible relación o asociación entre dos variables o factores cualitativos, cada uno de los cuales puede tomar distintos "valores" o modalidades. Estas tablas reciben el nombre de *tablas de contingencia,* y sobre ellas suele plantearse dos hipótesis principales.

Contraste de homogeneidad

Un primer caso que podemos encontrar al estudiar una tabla de contingencia es aquél en que se dispone de una población X clasificada en r subpoblaciones $x_1, x_2,...,x_r$. En cada una de estas poblaciones se toma una muestra, y los individuos de la misma se clasifican según una variable Y que puede tomar c valores posibles $y_1, y_2.....y_m$. Sea p_{ij} la proporción de individuos que, en la población x_i tiene como valor de $Y=y_j$.

Un contraste de homogeneidad entre las muestras es aquel que consiste en decidir entre una de las hipótesis siguientes:

$H_0 \equiv p_{1j} = p_{2j} = ... = p_{mj}$ para todo j

$H_1 \equiv$ algunas de estas proporciones son diferentes.

Esto es, se desea decidir si las muestras provienen de poblaciones con igual o diferente distribución de probabilidad. Una de las aplicaciones prácticas de este test es la posibilidad de combinar los resultados experimentales obtenidos por diferentes investigadores sobre un mismo trabajo. Generalmente, en estos casos, es preciso asegurarse de que los datos de las diferentes muestras que se pretende agrupar son realmente homogéneos.

Contraste de independencia

En este supuesto, la utilización de la tabla viene motivada por el interés de estudiar la asociación entre las variables cualitativas observadas sobre una misma población. Un ejemplo podría ser el siguiente: ¿existe relación entre el número de niños de una familia y el nivel de estudios de la madre? En principio podemos sospechar que si, pero de lo que se trata es de definir, en términos estadísticos, una medida de la mayor o menor asociación entre ambos factores. Este concepto, que en las variables cuantitativas aparece claro, necesita ser interpretado para las variables cualitativas. Diremos que no existe asociación, si la probabilidad de observar una cierta categoría de una variable no está afectada por la observación de ninguna otra categoría del otro factor en el mismo individuo. En este caso, las variables cualitativas correspondientes se dice que son independientes.

En las tablas de contingencia de las variables independientes, las frecuencias relativas de cada valor del carácter X es la misma para cada valor

distinto del carácter *Y*, como se muestra en la Figura 10.15a) Diremos que la asociación entre dos variables es "perfecta" si cada categoría de una de ellas se produce con una categoría de la otra, tal como se muestra en la Figura 10.15.b) Si unos valores de la variable *X* se presentan con mas frecuencia que otros para alguna de las categorías de la variable *Y* diremos que existe una "asociación parcial" (Figura 10.15.c)

Figura 10.15. Diferentes tipos de asociación

a) Independencia total

	Y_1	Y_2	Y_3	Total
X_1	10	20	70	100
X_2	20	40	140	200
Total	30	60	210	300

b) Asociación perfecta

	Y_1	Y_2	Y_3	Total
X_1	100			100
X_3			200	200
X_3		100		100

c) Asociación parcial

	Y_1	Y_2	Y_3	Total
X_1	10	80	10	100
X_2	80	20	100	200
Total	90	100	110	300

Al contrario que para las variables cuantitativas no existe una medida de asociación que tome el valor 1 siempre que la asociación sea perfecta y 0 en caso de independencia, para cualquier tipo de tabla de contingencia. Por este motivo, se han definido diversas medidas de este tipo, cada una de las cuales tiene sus ventajas e inconvenientes. En las siguientes secciones estudiaremos algunas de las más utilizadas.

10.7. EL TEST CHI-CUADRADO

Tanto para realizar una prueba de homogeneidad como de independencia seguiremos un procedimiento análogo, que pasamos a describir. Una primera forma de medir la diferencia entre las subpoblaciones, en el caso de un estudio de homogeneidad, o la asociación entre las variables, es mediante el estudio de las llamadas "frecuencias esperadas en el caso de ser cierta H_0", que se obtienen mediante la expresión (10.9).

(10.9) $$e_{i,j} = \frac{f_i \times f_j}{n}$$

Una medida de la discrepancia entre las frecuencias esperadas y las observadas, viene dada por el estadístico "Chi-cuadrado", que se define por (10.10).

(10.10) $$\chi^2 = \sum_i \sum_j \frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}}$$

Este estadístico toma siempre un valor mayor o igual que cero, correspondiendo el valor cero al caso de ser cierta Ho y sigue una distribución Ji-cuadrado con *(n-1)*(m-1)* grados de libertad. Procederemos entonces de la forma siguiente:

- Si el valor experimental $\chi^2 \leq \chi^2_{1-\alpha}$ (percentil del (1-α)*100% de la distribución Chi-cuadrado con (*n-1)*(m-1*) g.l.) aceptamos la hipótesis nula, para un nivel de significación α.
- Si el valor experimental $\chi^2 \geq \chi^2_{1-\alpha}$, rechazamos *Ho* y aceptamos H_1, puesto que en este caso se detecta una diferencia entre las frecuencias observadas y las esperadas que tiene una probabilidad menor de α.

Actividades

10.10. Calcular las frecuencias esperadas en la tabla de contingencia obtenida en las tablas de la figura 10.13. A la vista de los resultados ¿Crees que existe algún tipo de asociación entre las variables? Calcular el estadístico Chi-cuadrado y realizar un contraste de independencia.

10.11. Queremos saber si hay relación entre dos variables cualitativas. El valor del estadístico Chi-cuadrado en la tabla de contingencia (asumiendo que no hay relación entre variables para el cálculo de las frecuencias teóricas) fue 8,2 (el número de grados de libertad es 3). ¿Podemos indicar que hay relación entre ambas variables?

Caso especial de la tabla 2x2

En este caso podemos utilizar la forma alternativa (10.11) para el cálculo de Chi-cuadrado:

$$\chi^2 = \frac{(f_{11} f_{22} - f_{12} f_{21})^2 n}{f_1 \times f_2 \times f_1 \times f_2} \tag{10.11}$$

Frecuencias esperadas pequeñas

Puesto que la distribución del estadístico utilizado para el contraste se basa en la aproximación normal, no debe ser utilizado cuando las frecuencias de la tabla son demasiado pequeñas. Diversos autores dan reglas más o menos restrictivas sobre cuando es válido la aplicación del método. Una regla a seguir es que la mínima frecuencia esperada debe ser al menos igual a 1, y no más del 20% de las frecuencias esperadas serán menores de 5. En caso de tener que estudiar tablas que no cumplan estas condiciones caben varias alternativas, que pasamos a enumerar

Corrección de continuidad

Es una corrección para las tablas 2x2 parecida a la que se usó al aproximar la distribución binomial por la normal. Consiste en usar la expresión (10.12) modificada de Chi-cuadrado:

$$\chi^2 = \frac{\left[(f_{11} f_{22} - f_{12} f_{21}) - n/2\right]^2 n}{f_1 \times f_2 \times f_1 \times f_2} \tag{10.12}$$

Ejemplo 10.4. Si calculamos el estadístico Chi-cuadrado en la tabla de la Tabla 10.9 mediante la expresión (10.11) obtenemos:

$$\chi^2 = \frac{(468-598)^2 \times 93}{52 \times 41 x 49 x 44} = 0,3419$$

En cambio, aplicando la fórmula (10.12) obtenemos:

$$\chi^2 = \frac{\left[(468-598)-46\,5\right]^2 \times 93}{52 \times 41x49x44} = 0,141$$

Prueba "exacta" de Fisher

Si todas las frecuencias esperadas son pequeñas, incluso la utilización de la fórmula anterior es desaconsejable. Para el caso de la tabla 2x2 es, sin embargo, relativamente sencillo calcular la probabilidad de obtener la frecuencias observadas u otras que se aparten aún más de las frecuencias esperadas, utilizando la distribución hipergeométrica. Explicaremos el método con un ejemplo.

Ejemplo 10.5. Supongamos clasificados 13 individuos por la presencia o ausencia de dos factores *A* y *B*

	B	$\overline{B}$	Total
A	5	2	7
$\bar{A}$	3	3	6
Total 8	5	13	

Calculemos la probabilidad de obtener esta tabla en el caso de que los caracteres *A* y *B* fuesen independientes. Esta probabilidad viene dada por:

$$P = \frac{f_1.!\, f_2.!\, f_{\cdot 1}!\, f_{\cdot 2}!}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!\, n!}$$

esto es;

$$P = \frac{7!6!8!5!}{5!2!3!3!13!} = 0,3263$$

Otras tablas con valores más extremos que la dada, con la misma distribución marginal son:

B	$\overline{B}$	Total		B	$\overline{B}$	Total

A	6	1	7	A	7	0	7
$\bar{A}$	2	4	6	Ā	1	5	6
Total	8	5	13	Total	8	5	13

a las que corresponde unas probabilidades:

$$P = \frac{7!6!8!5!}{6!1!2!4!13!} = 0,816 \qquad P = \frac{7!6!8!5!}{7!0!1!5!13!} = 0,0047$$

Por tanto, la probabilidad de obtener los valores observados o más extremos es:

$$P = 0,3263 + 0,0816 + 0, 0047 = 0,426$$

Vemos que es bastante probable obtener la tabla dada en el caso de que las variables fuesen independientes. En consecuencia, como no tenemos suficiente motivo para establecer una asociación entre las variables, aceptamos la hipótesis nula.

10.8. MEDIDAS DE ASOCIACION EN DATOS NOMINALES (TABLAS 2x2)

Una vez rechazada la hipótesis de independencia, interesa dar una medida de la intensidad de la asociación entre las variables. Consideremos de nuevo una muestra de individuos de una población, clasificada según la presencia o ausencia de dos factores *A* y *B*.

	B	$\overline{B}$	Total
A	f_{11}	f_{12}	$f_{1.}$
$\overline{A}$	f_{21}	f_{22}	$f_{2.}$
Total	$f_{.1}$	$f_{.2}$	n

Una primera medida de la asociación entre *A* y *B* es el valor Ji-cuadrado obtenido de la tabla. Sin embargo, este valor depende del tamaño n de la muestra, como se aprecia en la fórmula (10.12). Para resolver este problema Pearson definió como medida de asociación la dada en (10.13).

(10.13) $$\Phi = \sqrt{\chi^2 / n}$$

El coeficiente Phi de Pearson toma valores entre 0 y 1. En caso de independencia total de las variables toma el valor 0, correspondiendo el 1 a la asociación perfecta. Puede demostrarse que es equivalente al coeficiente de correlación cuando se codifican los valores *A* y *B* por 0 y $\bar{A}$ y $\overline{B}$ por 1.

Riesgo relativo

Viene definido por (10.14).

(10.14) $$RR = \frac{P(A/B)}{P(A/\overline{B})} = \frac{f_{11} f_{\cdot 2}}{f_{\cdot 1} f_{12}}$$

El cociente *RR* indica, por tanto, cuanto más probable es la presencia de *A* entre los individuos con *factor B* que entre aquellos que no lo poseen.

Razón de productos cruzados

Se define esta medida por (10.15).

(10.15) $$RC = \frac{f_{11} f_{22}}{f_{21} f_{12}} = \frac{f_{11} / f_{21}}{f_{12} f_{22}} = \frac{C_1}{C_2}$$

Vemos que la razón de productos cruzados es una razón de cocientes. El cociente C_1 indica la razón de casos en que se presenta *A* y los que no se

presenta *A* cuando está presente *B*. El cociente C_2 indica la razón de casos *A* y no *A* cuando no está presente el factor *B*. Conviene observar que *RR* es una medida no simétrica. Es decir, A hace el papel de variable dependiente y *B* de independiente.

Ejemplo 10.6. Al clasificar 216 pacientes según incidencia de infarto de miocardio y hábito de fumar se obtuvo la siguiente tabla:

	Infarto	No tuvo	Total
Fuma	45	83	123
No Fuma	14	74	88
Total	59	157	216

$\chi^2 = 9.73$ $p \leq 0.001$

En esta tabla las medidas de asociación son las siguientes:

$$RR=(45 \times 88) / (128 \times 14) = 2{,}209$$

$$RC=(45 \times 74) / (14 \times 83) = 2{,}86$$

$$\Phi = \sqrt{9\text{-}73/216} = 0.2122$$

Al comparar estas medidas vemos que se ha obtenido un valor de Chi-cuadrado significativo. El valor *RR* indica que la probabilidad de infarto en los fumadores es 2,21 veces la de los no fumadores. El cociente *RC* indica que la razón infarto/ no infarto es doble en fumadores. El valor Phi no es muy grande. Si embargo el valor de este coeficiente suele reducirse cuando hay diferencia acusada en los totales de los márgenes de la variable en filas. Por otro lado, Phi es una medida simétrica, y tiene en cuenta la interdependencia conjunta de las variables, al contrario de las otras dos en que solo se tiene en cuenta la dependencia de *A* sobre *B.*

Actividades

10.12. En un estudio sobre la lepra se halló que de cada 100 personas sanas 49 son varones y de cada 100 enfermos 58 son varones. Indican estos datos la existencia de una asociación entre las variables sexo y padecer/ no padecer la lepra?

10.13. La siguiente tabla muestra datos sobre una enfermedad. Calcule la razón de riesgos relativos y productos cruzados para mujeres vs. hombres.

	Enfermos	Sanos	
Mujeres	46	1438	1484
Hombres	18	1401	1419

10.9. MEDIDAS DE ASOCIACION PARA TABLAS rxc.

Medidas basadas en el estadístico Chi-cuadrado

La interpretación y análisis de las tablas de *r* filas y *c* columnas es aún más compleja. Estudiaremos en esta sección algunas medidas de asociación. Para más detalles sobre medidas de asociación, estimación y análisis de las tablas de este tipo, puede consultarse la bibliografía referenciada en la sección anterior.

Un coeficiente parecido al Φ, utilizado en las tablas *rxc* es el coeficiente de contingencia de Pearson (10.16).

(10.16) $$C = \sqrt{\chi^2 /(\chi^2 + n)}$$

Un valor *C* igual a cero indica independencia absoluta. Sin embargo, esta medida no siempre alcanza el valor 1, siendo su máximo:

$$C\max = \frac{\min(r-1, c-1)}{1+\min(r-1, c-1)}$$

Por ello, algunos investigadores, ajustan el valor C calculado dividiéndolo por el máximo posible en una tabla de sus dimensiones.

Otro coeficiente basado en Ji-cuadrado es el *V* definido por Cramer:

(10.17) $$V = \sqrt{\chi^2 / n(p-1)}$$

donde p es el mínimo del número de filas y columnas. Este coeficiente varía entre 0 y 1 aún en tablas no simétricas.

Medidas basadas en la reducción proporcional del error:

Puesto que los coeficientes anteriores a veces no tienen una interpretación sencilla, algunos autores consideran medidas de asociación basadas en la cuantificación de la reducción del error que se comete al predecir el valor de una variable, cuando se conoce el valor de la otra.

La construcción de estas medidas está basada en un razonamiento del tipo siguiente: Supongamos que quiero predecir el valor de la característica X (variable en filas) en un individuo tomado al azar en la población. Si no tuviera ninguna información sobre el mismo, y lo asignara a la clase x_i la probabilidad de cometer un error en la clasificación sería:

$$P(\text{Error regla 1}) = (n - f_i)/n$$

Cuando dispongo de información del valor de la variable $Y=y_j$, en general no asignaré el valor X al azar, sino siguiendo una cierta regla (regla 2) que tenga en cuenta cual es el valor más probable de X para $Y=y_j$. Llamaremos medida de la reducción proporcional del error (PRE) al cociente:

$$\text{Medida PRE} = \frac{\text{P(error regla 1)-P(error regla 2)}}{\text{P(error regla 1)}}$$

Es decir, una medida PRE indica cual es el porcentaje de error que se ve reducido al predecir el valor de la variable dependiente *(X)*, conocido el valor de la variable independiente *(Y)*, en lugar de asignar al azar el valor de X. Una de estas medidas es la lambda de Goodman y Kruskal, dada en (10.18).

(10.18) $$\lambda_x = \frac{(\sum f_{mj}) - f_{m+}}{N - f_{m+}}$$

En la expresión (10.18) f_{m+} es la mayor frecuencia marginal en filas y f_{mj} es la mayor frecuencia en la columna j-ésima.

Ejemplo 10.7. La población de enfermos de Lepra en la provincia de Ja O se clasificó según forma clínica y situación médica de la forma siguiente:

	SITUACION		
F. CLINICA	CONTROL ALTA	CONDICIONAL	TOTAL
LEPROMATOSA	219	20	239
TUBERCULOIDE	63	128	191
OTRAS	26	18	44
TOTAL	308	166	474

Si calculamos el coeficiente Lambda para la Forma clínica como variable dependiente obtenemos:

$$\lambda_x = (219+128-239) / (474-239) = 0{,}4595$$

que indica que se reducen en un 50% los errores al predecir la forma clínica en función de la situación. Si calculamos Lambda para la situación como dependiente obtenemos:

$$\lambda_x = (219+128+26-308) / (474-308) = 0{,}3915$$

que indica una reducción del 40 %.

Actividades

10.14. En un estudio sobre la lepra se obtuvo la siguiente tabla de la prueba de mitsuda y forma clínica:

	FORMA CLINICA			
	I	L	T	B

MITSUDA+	12	130	17	7
MITSUDA-	10	38	46	6

¿Indican los datos un valor predictivo de la prueba mitsuda sobre la forma clínica?

10.15. Analizar las siguientes medidas de asociación entre variables cualitativas, desde el punto de vista de su equivalencia por la información que proporcionan sobre la existencia de una relación causal entre las variables intervinientes

- Dos variables *A* y *B* se correlacionan positivamente si y solamente si la probabilidad de que ocurran simultáneamente *A* y *B* es mayor que el producto de las probabilidades de *A* y *B* (Kendall, Lazarsfeld y Nagel): $P(A \bullet B) - P(A)P(B) > 0$.
- Dos variables están positivamente correlacionadas si y solamente si la probabilidad de *B* condicionada a *A* menos la probabilidad de B es mayor que cero (Reinchebach y Suppes). Esto es, A y B están correlacionadas positivamente si y solamente si $P(B/A) - P(B) > 0$.
- Dos variables están positivamente correlacionadas si y solamente si la probabilidad de *B* condicionada a *A* menos la probabilidad de *B* condicionado a no *A* es mayor que cero (Salmon y Suppes). Esto es, *A* y B están correlacionadas positivamente si y solamente si: $P(B/A) - P(B/A) > 0$.

10.10. MEDIDAS DE ASOCIACION PARA VARIABLES ORDINALES

Algunas veces las diferentes categorías de las variables están ordenadas, aun cuando los datos no puedan ser medidos en escala de intervalo o razón. Para estos casos puede ampliarse el análisis de la tabla. En particular, se definen medidas de asociación que miden cuando la dirección de las ordenaciones coincide o no en sentido. Para mayor claridad explicaremos algunas de ellas utilizando sólo un pequeño conjunto de datos, que no necesitan ser clasificados en tabla de contingencia.

Correlación por rangos de Spearman

Esta medida de asociación, también llamada coeficiente de Spearman o de ordenaciones, se utiliza especialmente en Psicología y Educación para estudiar la relación entre los órdenes de los valores de dos variables, esto es, con variables medidas con escalas ordinales. Mostraremos su cálculo con un ejemplo.

Ejemplo 10.8. La tabla siguiente da las calificaciones de 6 alumnos en dos asignaturas A y B.

A	B	X	Y	d	d^2
8.3	7.5	1	2	-1	1
8.1	7.6	2	1	1	1
6.2	7.2	3	3	0	0
12.0	4.1	5	4	1	1
12.1	4.0	4	5	-1	1
3.5	3.8	6	6	0	0

Las columnas X_i, Y_i indican el orden que cada alumno tiene en cada asignatura y la $d_i = X_i - Y_i$. El coeficiente de Spearman da una medida de la asociación entre dichos órdenes. Se calcula por la fórmula:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} \tag{10.19}$$

n representa el número de pares de observaciones. La interpretación de r_s es similar a la de r, tomando valores en el intervalo [-1,1]. En el ejemplo propuesto, aplicando (12.22) se obtiene r_s =0.89, siendo, por tanto una correlación alta.

Puede ocurrir que dos o más puntuaciones de las variables coincidan, en cuyo caso no quedara determinado el orden que debe atribuirse a cada una. Por ejemplo, si en la tabla anterior se incluyen los pares de puntuaciones (9.5,8) y (9.5,8.3) uno de los valores de A debería ser el 7º y el otro el 8º. El convenio que se toma para calcular r_s es asignar a esas puntuaciones el valor 7.5, media entre 7 y 8. En consecuencia, añadiremos los órdenes (7.5,7) y (7.5,8) para las variables ordenadas (X,Y).

Tau de Kendall

Para calcular este coeficiente, calculamos en primer lugar los valores P, Q y S definidos en la forma siguiente:

P= nº de pares que tienen el mismo orden en las clasificaciones X e Y

Q= nº de pares para los cuales los órdenes no concuerdan.

$$S=P-Q$$

$$\tau = \frac{2S}{n^2(n-1)}$$

Gamma de Goodman y Kruskal

$$\Gamma = \frac{S}{P+Q}$$

El coeficiente Gamma da la diferencia de las probabilidades de tener o no el mismo rango en las dos ordenaciones.

Ejemplo 10.9. Para calcular Tau y Gamma en el ejemplo, colocamos los valores *X* ordenados de menor a mayor:

X	1	2	3	4	5	6
Y	2	1	3	5	4	6

El valor *Y* que corresponde a *X* = 1 es 2, que tiene a su derecha 4 valores mayores que *M,* y, por tanto, bien clasificados respecto a 2 y 1 menor que *M*, mal clasificado. Por consiguiente, anotamos 4 puntos para *P* y 1 para *Q*. De igual manera hacemos con los otros valores de *Y*.

$$P = 4+4+3+1+1=13$$

$$Q = 1+0+0+1+0=2$$

$$\tau = 2+11/30=0.73$$

$$\Gamma = 11/13=0,846$$

En ambos casos obtenemos una correlación alta.

Ejemplos 10.10. Al realizar una clasificación, mediante un paquete estadístico, de 164 enfermos operados de implante intraocular según sexo y patología previa se obtuvo la siguiente tabla de contingencia y medidas de asociación con el programa Statgraphics

	Patología Previa		
	No	Si	Total
Varón	73	28	101
	72,28	27,72	61,59
	65,18	53,85	
Hembra	39	24	63
	61,90	38,10	38,41
	34,82	46,15	
Total	112	52	164
	68,29	31,71	

Ji-cuadrado(correc. de continuidad)	=1.4785 ($p \leq 0.2240$)
Phi	=0.1084
Razón de productos cruzados	=1.5996
Coeficiente de contingencia C	=0.1078
Tau de Kendall	=0.1084
Lambda de Goodman con X dep.	=0.0000
Lambda de Goodman con Y dep.	=0.6790

Como podemos apreciar, este paquete realiza el cálculo de todas las medidas de asociación (en el ejemplo hemos suprimido del listado otras medidas calculadas), y es el investigador el que debe tomar la adecuada a cada caso.

Vemos que no se ha obtenido un valor Ji-cuadrado significativo, por lo que se rechaza la hipótesis de independencia. Del mismo modo, los valores obtenidos para las medidas de asociación son poco significativos.

Para calcular el riesgo relativo, no suministrado por el programa procedemos en la forma siguiente:

$$RR = 73 \times 52 / (112 \times 28) = 1{,}21$$

que indica que la probabilidad de que al tomar al azar un enfermo con patología previa sea varón es sólo 1,21 veces la de que al tomar un enfermo sin patología previa sea varón.

Actividades

10.16. Al clasificar 96 enfermos portadores de lente intraocular por sexo y resultado de prueba fluorescencia se obtuvo:

	Fluorescencia Retina		
	Hiperfluorescencia	Normal	Total
Varón	13	48	61
Hembra	15	20	35
Total	28	68	96

¿Es diferente la proporción de hiperfluorescencia en varones y hembras? Calcule e interprete las medidas de asociación

10.17. Con la finalidad de averiguar si las bajas notas finales obtenidas en el curso de Estadística General es producto de las pocas horas dedicadas al estudio del curso durante el ciclo, se obtuvo la siguiente información. Pruebe la hipótesis de relación.

Horas de estudio	0-5	5-10	10-15	15-20	Total
0-3	25	15	8	1	49
3-6	20	10	11	3	44
6-9	15	8	15	10	48
Total	60	33	34	14	141

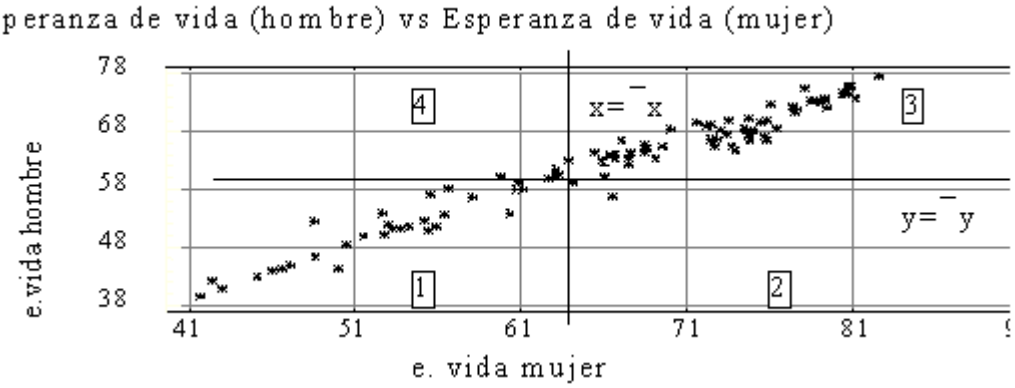
10.11. COVARIANZA Y CORRELACIÓN EN VARIABLES NUMÉRICAS

En el caso de variables numéricas podemos emplear algunos coeficientes cuyo valor nos indica el tipo de relación entre las variables. El primero de ellos es la covarianza S_{xy} cuya fórmula de cálculo viene dada en la expresión (10.20).

(10.20) $$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Es decir, para calcular la covarianza, para cada uno de los puntos (x_i, y_i) restamos a cada valor x_i su media $\bar{x}$ y el resultado lo multiplicamos por la diferencia entre y_i y su media $\bar{y}$. La covarianza tiene la propiedad de ser igual a cero si las variables son independientes, positiva si las variables tienen dependencia directa, y negativa en el caso de dependencia inversa. Podemos ver esto de forma intuitiva si razonamos del siguiente modo (Ver figura 10.13).

Figura 10.16. División del plano en cuatro cuadrantes al trazar las rectas $X= \bar{x}$ *e* $Y= \bar{y}$



En la figura 10.16 trazamos las dos rectas $X= \bar{x}$ e $Y= \bar{y}$. El diagrama queda dividido en cuatro regiones que en la figura hemos numerado de 1 a 4. Pueden darse tres casos, según el tipo de dependencia:

1. Si la dependencia entre las variables es directa como en la figura 10.16, la mayor parte de los puntos del diagrama se sitúan en los cuadrantes (1) y (3). Ahora bien, si un punto está en el cuadrante (1) su valor x_i es inferior al de la media $\bar{x}$ y su valor y_i es inferior al de la media $\bar{y}$. El producto $(x_i - \bar{x})(y_i - \bar{y})$ tiene signo positivo. Igualmente, para los puntos situados en el cuadrante (3) el producto $(x_i - \bar{x})(y_i - \bar{y})$ tiene signo positivo. Por tanto, en el caso de dependencia directa el signo de la covarianza será positivo, puesto que la mayoría de los sumandos son positivos.
2. Si la dependencia entre las variables es inversa podemos mostrar de forma análoga que el signo de la covarianza es negativo.

3. El caso restante, de independencia, corresponde a la covarianza nula.

Actividades

10.18. Razona por qué en caso de dependencia inversa entre variables numéricas el signo de la covarianza es negativo.

10.19.¿Cuál de los siguientes enunciados es cierto si dos variables están correlacionadas positivamente:

1. Cuando una aumenta la otra también aumenta
2. Cuando una disminuye la otra también aumenta
3. Cuando una disminuye la otra también disminuye
4. La relación entre las variables es de tipo lineal

10.20. Comprobar que la covarianza es invariante por traslaciones, pero no por cambio de escala.

10.21. Las calificaciones en dos exámenes han sido:

Primer examen 7 9 5 6 4 4 5 1 6 4 7 2 8 5 4 2 4 5 7 2

Segundo examen 6 7 7 5 5 3 4 1 6 5 6 3 6 5 6 5 3 4 5 3

Calcular la covarianza, y a la vista de su valor indicar el tipo de dependencia entre las dos calificaciones.

10.22. Las estadísticas muestran que casi todos los accidentes de circulación se producen entre vehículos que ruedan a velocidad moderada. Muy pocos ocurren a más de 150 Km. por hora. ¿Significa esto que resulta más seguro conducir a gran velocidad?

Coeficiente de correlación

Un problema con la covarianza es que no hay un máximo para el valor que puede tomar, por lo cual no nos sirve para comparar la mayor o menor intensidad de la relación entre las variables. Un coeficiente que permite estudiar no sólo la dirección de la relación sino también su intensidad es el *coeficiente de correlacion lineal* o coeficiente de Pearson, que se define por la relación (10.21), siendo s_X , s_Y las desviaciones típicas de las variables X e Y en la muestra analizada.

(10.21)
$$r = \frac{S_{xy}}{S_x S_y}$$

Puesto que las desviaciones típicas son siempre positivas, r tiene el mismo signo que la covarianza y por tanto:

- Si $r>0$ la relación entre las variables es directa;
- Si $r<0$ la relación entre las variables es inversa;
- Si $r=0$ las variables son independientes.

Además, el coeficiente de correlación r es siempre un número real comprendido entre -1 y 1.

- Cuando existe una relación lineal funcional, esto es todos los puntos se encuentran sobre una recta - que es el caso de máxima asociación - el valor de r será 1 si la recta es creciente (relación directa) o -1 si la recta es decreciente (relación inversa);
- Cuando las variables son independientes, $r=0$ porque la covarianza es igual a cero;
- Los casos intermedios son aquellos en que existe dependencia aleatoria entre las variables. Esta dependencia será más intensa cuanto más se aproxime a 1 o -1 el coeficiente de correlación.

Actividades

10.23. Ordena los siguientes coeficientes de correlación según indiquen mayor o menor intensidad en la relación de las variables X e Y. Indica cuáles corresponden a cada una de las gráficas 2, 3, 4 y 5.

r= 0.982; r=0.637; r=-0.7346; r= -0.8665; r=0.

10.24. Indica cuáles de los siguientes enunciados sobre la covarianza son ciertos:
Cuando la covarianza entre X e Y es mayor que cero, entonces:

1. La correlación entre X e Y es positiva
2. X e Y pueden tener una relación no lineal
3. La nube de puntos es decreciente

10.25. Juan calcula la correlación entre pesos y alturas de los chicos de la clase. Mide el peso en kilos y la altura en metros. Angela mide la altura en cm. y el peso en grs. y

calcula también la correlación ¿Cuál de los dos obtiene un coeficiente mayor?

10.26. ¿Cuál de las siguientes afirmaciones sobre el coeficiente de correlación r es cierta?

1. Si r=0 las variables son independientes
2. Si las variables son independientes, r=0
3. r puede interpretarse como un porcentaje de la varianza
4. Si la relación es funcional r=1 0 r=-1

10.27. Analiza qué tipos de relaciones entre variables pueden originar la existencia de correlación y cuáles de ellos son de naturaleza causal. Pon ejemplos de relaciones causales que den origen a un bajo coeficiente de correlación.

10.12. AJUSTE DE UNA LÍNEA DE REGRESIÓN A LOS DATOS

En el caso de que exista una correlación suficiente entre dos variables numéricas podemos plantearnos un nuevo problema que consiste en tratar de determinar la ecuación de una función matemática que nos permita predecir una de las variables *(Y)* cuando conocemos la otra variable *(X).* Esta función será la *línea de regresión de Y en función de X.*

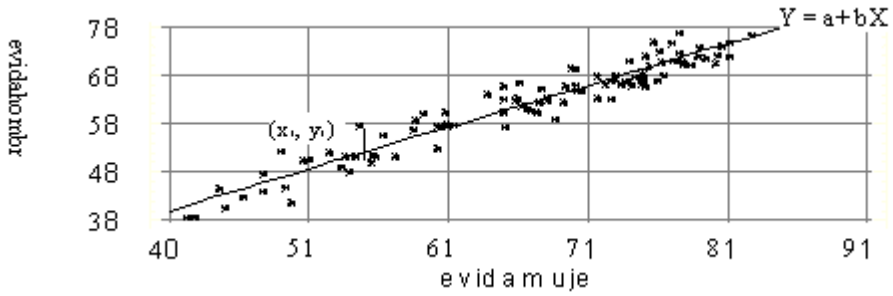
Esto puede ser útil cuando la variable *Y* se refiere a un acontecimiento futuro, mientras que X se refiere al presente o pasado, por ejemplo, si queremos predecir la nota media del expediente académico de un alumno que ingresa en la Facultad a partir de su nota en el examen de selectividad. En otros casos la variable *X* es más fácil de medir que la *Y*.

Para cualquier tipo de función de regresión que sea necesario ajustar a una cierta nube de puntos, el problema que se plantea es determinar los parámetros de la curva particular - perteneciente a una familia de funciones posibles - que mejor se adapte a la muestra de datos de que se disponga. Por ejemplo, en las gráficas 10.3, 10.4 y 10.5 la función que mejor se ajustaría a la nube de puntos sería una recta (creciente en la gráfica 10.5 y decreciente en las gráficas 10.3 y 10.4). Para la gráfica 10.2 habría que buscar otro tipo diferente de función, como una exponencial.

Cuando la forma de la nube de puntos sugiere que una recta de ecuación $Y = a+bX$ puede ser apropiada como línea de regresión, será necesario calcular las constantes *a* y *b.* Si, por el contrario, es más apropiada una parábola de ecuación $Y=a+bX+cX^2$, se precisa determinar tres constantes: *a, b, c.*

El principio general que se utiliza para calcular dichas constantes se conoce con el nombre de *criterio de los mínimos cuadrados.* Está basado en la idea de que a medida que una curva se ajusta mejor a una nube de puntos, la suma de los cuadrados de las desviaciones d_i (Figura 10.17), sumadas para todos los puntos, es más pequeña. La desviación o residuo del punto (x_i,y_i) respecto de la curva es la diferencia entre la ordenada y_i del punto y la ordenada de un punto de la curva que tiene la misma abscisa x_i. Es decir $d_i = y_i - (a+b\,x_i)$.

Figura 10.17. Desviaciones de los puntos a la recta de regresión



El procedimiento de obtención de las constantes será hacer mínima la cantidad *D,* dada por (10.22), siendo $f(x_i)=a+bx_i$, si lo que se trata de ajustar es una línea recta.

(10.22) $$D= \Sigma [y_i-f(x_i)]^2$$

Como consecuencia se obtienen las cantidades a y b que vienen dadas por (10.23):

(10.23) $$b = \frac{S_{xy}}{S_{x^2}}; \quad a = \bar{y} - b\,\bar{x}$$

siendo S^2_x la varianza de la variable *X*, $\bar{x}$, $\bar{y}$ las medias de *X* e *Y* y S_{xy} la covarianza.

Ejemplo 10.11. Estudiando la población total en 1986 en función de la población total en 1970 en los diferentes municipios de la provincia de Jaén se obtuvo la siguiente ecuación de la línea de regresión:

$$Y = -0{,}0688 + 0{,}97658\,X$$

Como se ve, la población en un municipio es aproximadamente el 98 % de la que había en 1970. Aunque el conjunto de población ha aumentado en la provincia, sin embargo el valor ligeramente inferior a uno de la pendiente de la recta se explica debido a las emigraciones de los pueblos pequeños (la mayoría de los datos) a las cabeceras de comarca. Aunque los puntos no se encuentran colocados exactamente sobre la recta, esta nos muestra los valores de los datos en forma aproximada. Como consecuencia la ecuación de la recta de regresión de Y sobre X puede escribirse según la relación (10.24).

$$y_i - \bar{y} = (x_i - \bar{x})\frac{\bar{S}_{xy}}{S_{x^2}} \qquad (10.24)$$

Como puede apreciarse, el par $(\bar{x}, \bar{y})$, satisface la ecuación (10.24); esto es, la recta pasa por el "centro de gravedad" de la nube de puntos, formado por las dos medias. La constante *b*, pendiente de la recta de regresión, recibe el nombre de *coeficiente de regresión* de *Y* sobre *X* .

Nótese que la ecuación de la recta de regresión de *Y* sobre *X* expresa los valores medios de la variable *Y* para cada valor fijo de *X*. También puede plantearse el problema de hallar la recta que determine los valores medios de *X* en función de cada valor de *Y,* es decir el cálculo de la recta de regresión de *X* sobre *Y*. En este caso, intercambiando en la expresión (26) los papeles de las variables, obtenemos (10.25).

$$x_i - \bar{x} = (y_i - \bar{y})\frac{\bar{S}_{xy}}{S_{y^2}} \qquad (10.25)$$

Esta recta es, en general, diferente de la dada en la ecuación (25), aunque también pasa por el punto *(x,y).*

La cantidad *D/n* o *varianza residual* representa la fracción de la varianza

de *Y* que es debida al azar, o sea, a las desviaciones de las observaciones y_i respecto de la recta de regresión *y* puede demostrarse que es igual a $1- r^2$, siendo *r* el coeficiente de correlación. El cuadrado del coeficiente de correlación r^2 -llamado *coeficiente de determinación*- representa la fracción de la varianza de *Y* debida o explicada por la regresión.

En el caso de que se desee ajustar a la nube de puntos una parábola de grado n, será preciso minimizar la expresión (10.26), obteniendo un sistema de ecuaciones que nos permite determinar los parámetros $a, b_1, b_2.....b_n$.

(10.26) $$S^2_r = - \Sigma [y_i - (a+b_1x+b_2x_2+....b_nx^n)]^2$$

Actividades

10.28. Ajustar una recta de regresión de Y sobre X a los datos del ejercicio 13.4. ¿Cual será la nota esperada en el segundo parcial para un alumno que ha obtenido un 6.5 en el primero?

10.29. Al observar la densidad por hectárea de ciertas comunidades de aves mediante dos métodos diferentes de muestreo se obtuvieron los datos siguientes. Calcúlense las dos rectas de regresión

Parcela	0	1,1	1,66	1,1	3,32	5,54	2,77	18,84	6,65
Taxiado	,28	0	,7	,55	2,37	3,79	1,95	14,17	2,94

10.30. Un comercio estudia la relación entre el número de cajeras y el tiempo de espera en cola, obteniendo los siguientes datos

Nº cajeras	10	12	14	12	18	20
Tiempo espera	59	51	42	32	22	18

Calcule la ecuación de la recta que da el tiempo de espera en función del número de cajeras. ¿Cuál sería el tiempo con 16 cajeras? ¿Cuántas cajeras habría que instalar para que el tiempo fuese menor a 15?

10.13. REGRESIÓN Y CORRELACIÓN CON STATGRAPHICS

En Statgraphics hay varios programas relacionados con la correlación y regresión. Uno de ellos se obtiene a partir de las opciones DEPENDENCIA –

REGRESIÓN SIMPLE, cuya ventana de entrada de variables nos pide las variables que tomamos como *Y* (variable dependiente o explicada) y *X* (variable independiente o explicativa).

Es importante darse cuenta cuál variable tomamos como *Y* y como *X*, porque el programa encontrará una ecuación de *Y* en función de *X* (que no siempre coincide con la ecuación que da *X* en función de *Y*). En la Figura 10.18 presentamos el resultado que se obtiene en RESUMEN ESTADÍSTICO cuando elegimos como variable *Y* (dependiente) la esperanza de vida del hombre y como variable *X* (independiente) la esperanza de vida de la mujer en el fichero DEMOGRAFÍA.

Este programa presenta una gran cantidad de información, pero nosotros sólo tendremos en cuenta la siguiente:

- Modelo ajustado: Se indica en la primera línea (Modelo Lineal: Y = a + b*X);
- Variables dependiente e independiente: (líneas segunda y tercera);
- Parámetros del modelo (*a* es la intersección con el origen o "ordenada"; en nuestro caso, a = 4,69411; *b* es la pendiente o "pendiente"; en nuestro caso b= 0,858511); Por tanto, en nuestro caso Y= 4,69 + 0,86X es la ecuación de la recta de regresión que da la esperanza de vida del hombre en función de la de la mujer. En promedio el hombre vive el 86% de lo que vive la mujer más unos cinco años;
- Coeficiente de correlación; en nuestro caso Coeficiente de Correlación = 0,982558, tiene signo positivo (dependencia directa) e intensidad muy fuerte porque es muy próximo a 1, que es el máximo valor del coeficiente de correlación;
- Coeficiente de determinación o correlación al cuadrado. En nuestro caso R-cuadrado = 96,54; indica que el 96,54 por ciento de la variabilidad de la esperanza de vida del hombre queda explicada por la esperanza de vida de la mujer.

Figura 10.18. Resultados del Programa de Regresión Simple

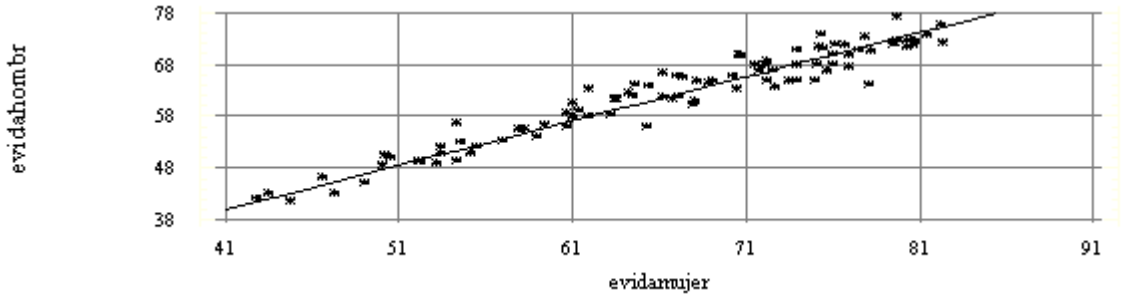
```
Análisis de Regresión - Modelo Lineal Y = a + b*X
------------------------------------------------------------------------------
Variable dependiente: evidahombr
Variable independiente: evidamujer
------------------------------------------------------------------------------
                              Error     Estadístico
Parámetro     Estimación      estándar         T             P-Valor
------------------------------------------------------------------------------
Ordenada         4,69411       1,11775         4,1996         0,0001
Pendiente       0,858511     0,0166701        51,4999         0,0000
------------------------------------------------------------------------------


                          Análisis de la Varianza
------------------------------------------------------------------------------
Fuente          Suma de cuadrados     GL  Cuadrado medio Cociente-F   P-Valor
------------------------------------------------------------------------------
Modelo                   8569,86      1      8569,86    2652,24       0,0000
Residuo                  306,962     95      3,23117
------------------------------------------------------------------------------
Total (Corr.)            8876,82     96

Coeficiente de Correlación = 0,982558
R-cuadrado = 96,542 porcentaje
```

Nota importante. Que una variable quede explicada por otra no quiere decir que haya una relación de causa y efecto. En el ejemplo analizado tanto la esperanza de vida del hombre como la de la mujer tienen su causa en una serie de factores que afectan a las dos variables simultáneamente y se refieren al desarrollo económico de un país y sus condiciones de vida, salud, etc. "Quedar explicado" en regresión significa que una variable sirve para predecir la otra, como hemos visto en el ejemplo.

En el ejemplo, hemos utilizado la regresión lineal porque en la gráfica se puede observar con claridad que la función que mejor aproxima los datos es una línea recta. En otros casos será preferible usar una función diferente. En el programa Statgraphics mediante OPCIONES DE ANÁLISIS es posible realizar ajuste con una variedad de curvas, aunque la interpretación es muy similar a la que hemos hecho para el caso de la recta.

El programa tiene diversas representaciones gráficas. La más útil es la de GRÁFICO DEL MODELO AJUSTADO que dibuja la curva ajustada sobre la nube de puntos. Cambiando el tipo de función en OPCIONES DE ANÁLISIS podemos ver también visualmente cuál de los modelos es más ajustado a los datos. El coeficiente de correlación calculado para cada modelo y su cuadrado (proporción de varianza explicada) nos permite elegir entre varios modelos aquél que proporciona la mayor proporción de varianza explicada para el conjunto de datos.

Figura 10.19. Dibujo del modelo ajustado a la nube de puntos



Actividades

10.31. ¿Cuál de los siguientes enunciados es cierto?

Cuando la intensidad de la relación entre dos variables decrece:

1. La pendiente de la recta de regresión de Y sobre X crece
2. La pendiente de la recta de regresión de X sobre Y crece
3. Hay mayor dispersión en la nube de puntos
4. La covarianza aumenta en valor absoluto

10.32. ¿Cuál de los siguientes enunciados es cierto si el coeficiente de correlación entre dos variables es nulo?:

1. Las rectas de regresión Y sobre X y X sobre Y son paralelas
2. Las rectas de regresión Y sobre X y X sobre Y son perpendiculares
3. Las rectas de regresión Y sobre X y X sobre Y coinciden
4. La covarianza es nula

10.33. ¿Cuál es el valor del coeficiente de correlación, si las dos rectas de regresión tienen la misma pendiente?

a)0; b) 1; c) -1

10.34. Si X e Y tienen una correlación perfecta, ¿Cuál es el ángulo que forman las dos rectas de regresión?

a) 120 b) 90; c) 45; d) 0

10.35. Una recta de regresión tiene una pendiente igual a 16 y corta al eje de ordenadas en el punto Y= 4. Si la media de la variable independiente es 8, ¿cuál es la media de la variable dependiente?

10.14. INFERENCIAS SOBRE LOS PARAMETROS DE LA RECTA DE REGRESION

Los cálculos realizados en el apartado anterior se refieren a los valores muestrales. En la mayor parte de los casos, sin embargo, estamos interesados en hallar la fórmula que expresa la relación entre las variables en la población. En dichos casos, se supone que los valores $y_1,...y_n$ son valores observados de una variable aleatoria *Y*. Haremos las siguientes hipótesis:

- *Linealidad:* Existe una relación lineal entre los valores x_i e y_i que puede expresarse en la forma siguiente:

$y_i=ß_0+ß_1x_i+\in_i$, para cada i,

donde $ß_0$ y $ß_1$ son parĀmetros desconocidos - ordenada en el origen y pendiente de la recta - referidos a la población, x_i es un valor fijo e y_i es una observación de la variable aleatoria Y.

- *Homocedasticidad:* El valor $\in_i$, denominado "residuo", representa la diferencia del valor y_i con el punto de la recta que tiene como coordenada $X=x_i$. Para cada x_i, $\in_i$ tiene una misma varianza, σ_r^2, que llamaremos "varianza residual".
- *Normalidad:* Las variables $\in_i$ tienen distribución $N(0, \sigma_r)$.

Puede demostrarse que las mejores estimaciones de $ß_0$ y $ß_1$ son precisamente los valores a y b obtenidos al calcular en la muestra la ecuación de la recta de regresión. Por tanto, la ecuación de la recta de regresión estimada es (10.27)

(10.27) $$Y=a+bX$$

y la estimación del valor medio de la variable *Y*, para un x_i dado es (10.28).

(10.28) $$y_i = a+bx_i$$

Para realizar inferencias sobre $ß_0$, $ß_1$ e y_i, necesitamos estimar el valor σ_r^2 de la varianza común que tiene como estimador insesgado la expresión (10.29).

(10.29) $$S_r^2 = \frac{\sum (Y_i - a - bx_i)^2}{n-2}$$

Este valor viene calculado ordinariamente en los diferentes paquetes estadísticos en la tabla del análisis de la varianza de la regresión, que aparece en la tabla del análisis de varianza que reproducimos a continuación y que utilizaremos también para diferentes contrastes.

TABLA DEL ANALISIS DE LA VARIANZA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Debida a regresión	SCA	1	CMA	CMA/CMB
Residual	SCB	n-2	CMB	
Total	SCT	n-1		

En la tabla del análisis de varianza se verifican las relaciones siguientes:

$$SCT=SCA+SCB= \sigma^2_y n$$

$$SCB= \Sigma (y_i - a - bx_i)^2 = S^2_r (n-2)$$

$$SCA = \Sigma (a+bx_i-y)^2 = S_{xy}^2 n / S_x^2 = b^2 S_x^2 n$$

$$CMA=SCA$$

$$CMB=SCB/(n-2)=S_r^2$$

Nótese que CMA puede expresarse en función del coeficiente de

regresión. Por ello, recibe el nombre de "varianza debida a regresión". Su valor será mayor cuanto más grande sea el valor de *b*.

Por otro lado *CMB* estima la varianza de los residuos. La suma de estas dos varianzas es igual a la varianza de la variable *Y*, que puede ser descompuesta en dos sumandos. El primero de ellos explica la variación de *Y*, que es explicada por la regresión sobre *X*. El otro es la variación de los "residuos" o diferencia entre la nube de puntos y la recta de regresión.

Inferencias sobre el coeficiente de regresión ß1.

En caso de cumplirse las hipótesis del modelo, el estadístico b tiene una distribución normal con media igual a ß1 y desviación típica (r/((x(n). Por ello, el estadístico (10.30) sigue una distribución T con n-2 grados de libertad.

(10.30) $$T = \frac{b - \beta_1}{S_r / S_x \sqrt{n}}$$

Por tanto, para decidir entre las hipótesis:

$H_0 \equiv \beta_1 = c$

$H_1 \equiv ß_1 = c$

se calcula el estadístico *T* dado en (10.30), sustituyendo $ß_1$ por *c,* y se decide aceptar la hipótesis nula, cuando el valor experimental obtenido es menor en valor absoluto que el percentil del $(1-\alpha/2)100\%$ de la distribución *T*. De forma similar se realizan contrastes unilaterales. El denominador de la expresión (10.30) suele aparecer en los estudios de regresión realizados con paquetes estadísticos como error de muestreo del coeficiente de regresión o de la "pendiente".

Si observamos la tabla del análisis de la varianza, vemos que el valor *F=CMA/ CMB* es igual a al cuadrado del valor *T*, en el caso de que *c=0*. Por ello, puede utilizarse este valor *F* para probar la hipótesis de no regresión.

Un intervalo de confianza para el parámetro $ß_1$ con coeficiente de confianza $(1-\in)100\%$ vendrá dado por (10.31).

(10.31) $$b \pm T_{1-\epsilon/2}\, S_r / (S_x \sqrt{n})$$

Ejemplo 10.12. La tabla del análisis de varianza de la regresión del ejemplo 10.11 es la siguiente:

TABLA DEL ANALISIS DE LA VARIANZA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Debida a regresión	417939,38	1	417939,38	39153,21
Residual	10110,07	95	10,67	
Total	417950	96		

De ella se deduce que S_r^2=10,67, S_d^2 =417939,38 S_y^2=417950, S_x^2=4449,81. Para decidir la hipótesis de que la pendiente de la recta de regresión es igual a 1 se calcula (10.30).

$$T = (1-0{,}97658)\ \sqrt{96}/\sqrt{10{,}67/4449{,}81} = 10{,}686$$

Como el valor de *T* obtenido es mayor que el percentil del 99% de la distribución *T*, rechazamos la hipótesis con un nivel de significación del 1%. Un intervalo de confianza para $ß_1$ con un coeficiente de confianza del 95% viene dado por:

$$0{,}9758 \pm 1{,}96 * 0{,}0010{,}998 = (0{,}9708\text{-}0{,}98157)$$

Inferencias sobre la ordenada en el origen $ß_0$.

El error de muestreo del parámetro $ß_0$ viene dado por EMA, donde su cuadrado viene dado en (10.32).

(10.32) $$EMA^2 = \frac{S_y^2 \sum x^2}{n^2 S_x^2}$$

Este error de muestreo puede ser utilizado para la realización de contrastes sobre $ß_0$. Para decidir entre las hipótesis:

$$H_0 \equiv ß_0 = c$$
$$H_1 \equiv ß_0 \neq c$$

se calcula el valor *T* que sigue una distribución T con n-2 g. l. y se procede en la forma habitual.

$$T = \frac{a - c}{EMA}$$

Los intervalos de confianza para α vienen dados por:

$$a \pm EMA*T$$

Ejemplo 10.13. El error de muestreo de la ordenada en el origen a para la recta de regresión empírica del ejemplo es EMA = 0,33856. Un intervalo de confianza del 95% para el parámetro $ß_1$ vendrá dado por:

$$-0.0678 \pm 1.96 * 0.33856 = (-0.7322, 0.5947)$$

Intervalos para el valor de $Y=y_i$ para un valor dado de $X=x_i$

El valor de $Y=y_i$ para un $X=x_i$ dado puede tener dos interpretaciones. En primer lugar podemos desear calcular un intervalo para el valor medio de la distribución de *Y* cuando *X* toma un valor fijo. Este intervalo viene dado por (10.33).

(10.33)
$$y_i \pm S_y \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{nS_x^2} \right)^{1/2} T$$

donde *T* es el limite correspondiente de la distribución *T* con *n-2 g. l.* Debe observarse que la amplitud del intervalo depende de la diferencia x_c-$\bar{x}$, por lo que la estimación se hace menos precisa conforme el valor x_i se aparta de la media.

Cuando interpretamos y_i como un simple valor de la variable Y para un x_i

dado, el intervalo de confianza viene dado por (10.34).

$$y_i \pm S_y \left(1 + \frac{1}{n} + \frac{(x_c - \bar{x})^2}{nS_x^2} \right)^{1/2} T \qquad (10.34)$$

En el caso de que el tamaño n de la muestra sea grande, este intervalo es aproximadamente igual a $y_i \pm S_r T$, por lo que la raíz cuadrada de la varianza residual es, aproximadamente, el error de muestreo de la estimación de un valor y_i para un x_i dado.

Ejemplo 10.110. Para una población *X = 100000* habitantes en 1970 podemos estimar el número de habitantes Y en 1986, con un coeficiente de confianza del 95%.

$$Y = -0.0678 + 97.658 \pm 1.96 * 3.267 = (97651, 97665)$$

Actividades

10.36. En el ejercicio 13.9 contrastar la hipótesis de no regresión.

10.37. En el ejercicio 13.5, calcular un intervalo de confianza para la pendiente de la recta de regresión.

10.38. Al estudiar la relación entre las concentraciones urbanas de los años 1970 (X) y 1986 (Y), se obtuvo la siguiente tabla del análisis de la varianza:

Fuente de variación	Grados de libertad	Suma de cuadrados
Regresión	1	19506.3
residual	95	8743.5

Completa la tabla y contrastar la hipótesis de no regresión.

10.15. INFERENCIAS SOBRE EL COEFICIENTE DE CORRELACION.

Mientras que en el modelo de regresión suponemos que los valores de X son determinados de antemano, y sólo hay una variable aleatoria, en el modelo de correlación se supone que la muestra disponible es un conjunto de observaciones de una variable aleatoria bidimensional. Conviene, pues,

distinguir entre el valor *r* muestral y el coeficiente de correlación ρ de la población, que se define como:

$$\rho = Cov(X,Y)/\sigma_x\sigma_y$$

Como $\rho=ß_1\sigma_x/\sigma_y$, $\rho=0$ si y sólo si $ß_1=0$, para probar la hipótesis :

$H_0 \equiv \rho = 0$

$H_1 \equiv \rho \neq 0$

se procede en forma análoga que para probar que el coeficiente de regresión $ß_1=0$. Para probar la hipótesis:

$H_0 \equiv \rho = \rho_0$

$H_1 \equiv \rho \neq \rho_0$

hacemos uso del valor *r* observado en la muestra y de la transformación de Fisher:

$$v = 1/2 \log (1+r)/(1-r)$$

que sigue una distribución aproximadamente normal de media $\mu_v=1/2log(1+ \rho_0)/(1- \rho_0)$ y desviación típica $\sigma_v= \sqrt{1/(n-3)}$. Calcularemos a partir de la muestra el estadístico *Z* dado por (10.35) que tiene distribución *N(0,1),* y procederemos en la forma acostumbrada.

(10.35) $$Z= (v- \mu_v)/\sigma_v$$

Un intervalo de confianza para μ_v viene dado por $v\pm\sigma_v Z$, en donde *Z* es el valor correspondiente de la distribución normal. Para hallar un intervalo de confianza para el coeficiente de correlación basta usar la transformación inversa de Fisher, que viene dada por (10.36).

(10.36) $$r = \frac{e^{2v}-1}{e^{2v}+1}$$

Ejemplo 10.6: El coeficiente de correlación muestral entre las variables del Ejemplo 13,2 es 0,9987, y el número de municipios muestreados 96. Consideremos estos municipios como muestra de una población de municipios andaluces. Para hallar un intervalo de confianza para calculamos en primer lugar la transformada de Fisher:

$$v=1/2 \ln 1,9787/0,0013=3,6689$$

$$\sigma_v=1/\sqrt{93}=0,103695$$

Un intervalo para μ_v viene dado por: 3,6689±1,96x 0,103695=(3,4657-3,8722). A continuación aplicamos a los extremos del intervalo la transformación inversa de Fisher, obteniendo para el coeficiente de correlación el intervalo: (0,998, 0,9991).

Actividades

10.39. Si la ecuación obtenida de la recta de regresión fue: Y = 21.2+0.82X, calcular un intervalo aproximado de confianza del 95% para la X=20 si el coeficiente de correlación es 0,7.

10.40. El coeficiente de correlación entre la altitud sobre el nivel del mar y la concentración urbana en una provincia con 96 municipios fue -0.3116. ¿Puede admitirse que el coeficiente es significativamente distinto de 0? Hallar un intervalo de confianza del 99%.

10.41. Hallar un intervalo de confianza del 95% para el coeficiente de correlación de las variables altitud y distancia a la capital, sabiendo que en la muestra de 96 municipios se obtuvo un valor r=0.5131.

10.16. EXAMEN DE LOS RESIDUOS

Aunque en la mayor parte de los paquetes se dispone de diversos test para la comprobación de las hipótesis del modelo, esta comprobación también puede hacerse gráficamente mediante la observación de los residuos $y-a-bx_i$. Para ello se prepara una gráfica de puntos en los que el eje *X* represente los valores de *X* y el *Y* los residuos, disponible en Statgraphics, donde también se pueden representar los valores observados en función de los predichos y otros

gráficos. En caso de cumplirse las hipótesis, al menos de forma aproximada, la gráfica de los residuos formará una banda aproximadamente horizontal, alrededor del valor medio cero, como se muestra en la figura 10.20. La presencia de heterocedasticidad, se pondría de manifiesto cuando la dispersión de los residuos dependa del valor de *X*. Una banda de forma lineal, como en la figura 10.21 presupondría la existencia de otra variable que depende de *X* y es la causa de la variación.

Figura 10.20. Residuos en una correlación lineal

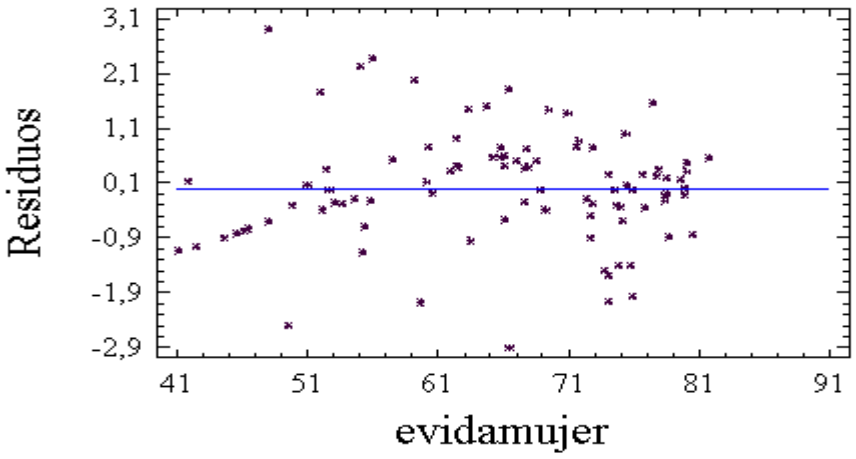
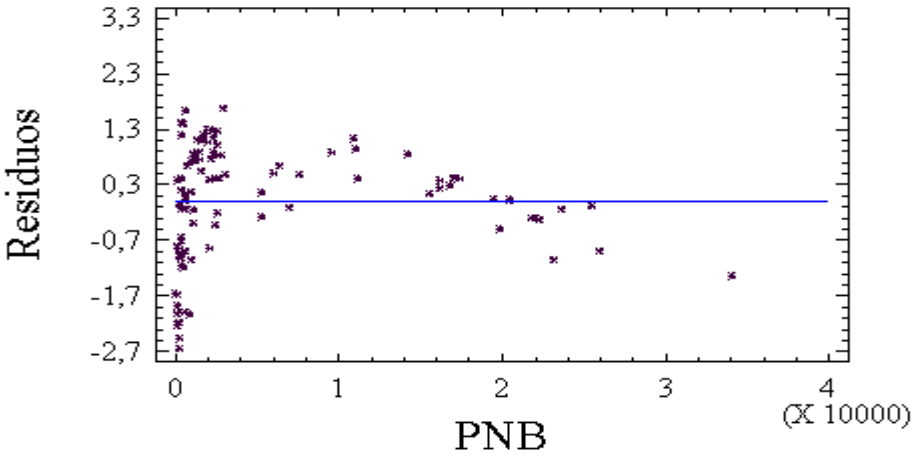


Figura 10.21. Residuos en correlación no lineal



REFERENCIAS

Abelson, R.P. (1998). *La estadística razonada: Reglas y principios*. Barcelona: Paidós.

Afifi, A, A. y Clark, V. (1990). *Computer-aided multivariate analysis.* New York: Van Nostrand.

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Amón, J. (1984). *Estadística para psicólogos I*. Madrid: Pirámide.

Amón, J. (1987). *Estadística para psicólogos II*. Estadística inferencial. Madrid: Pirámide.

Arteaga, P., Batanero, C., Cañadas, G., & Contreras, J. M. (2011). Las tablas y gráficos estadísticos como objetos culturales. *Números*, *76*, 55-67.

Arteaga, P., Batanero, C., Díaz, C., & Contreras, J. M. (2009). El lenguaje de los gráficos estadísticos. *Revista Iberoamericana de Educación Matemática*, *18*, 93-104.

Azorín, F. y Sánchez-Crespo, J. L. (1986). *Métodos y Aplicaciones del Muestreo.* Madrid: Alianza Editorial.

Batanero, C., & Díaz, C. (2011). Estadística con proyectos. *Recuperado el*, *13*.

Batanero, C. y Godino, J. (1991). Análisis de datos y su didáctica. Granada: Los autores.

Botella, B. y Barriopedro, M.I. (1991). *Problemas y ejercicios de psicoestadística*. Madrid: Pirámide.

Botella, J., León, O. G. y San Martín, R. (1993). *Análisis de datos en psicología I.* Madrid: Pirámide.

Calot, G. (1974). *Curso de estadística descriptiva*. Madrid: Paraninfo, Madrid.

Canavos, G. (1992). *Probabilidad y estadística*. México: McGraw Hill.

Castillo, J. (1990). *Estadística inferencial básica*. México: UNAM.

Cuadras, C. (1999). *Problemas de probabilidades y estadística*. Barcelona: EUB.

Cuadras, C. M., Echevarría B., Mateo, J. y Sánchez, P. (1984). *Fundamentos de estadística. Aplicación a las ciencias humanas.* Madrid: Promociones Publicaciones Universitarias.

De Groot, M. (1988). *Probabilidad y estadística*. Wilmington: Addison-Wesley.

Fernandes, J. A., Batanero, C., Contreras, J. M., & Díaz, C. (2009). A simulação em Probabilidades e Estatística: potencialidades e limitações. *Quadrante, XVIII*, *1*, 161-183.

Freund, J. E., Miller, I. y Miller, M. (2000). *Estadística matemática con aplicaciones*. Prentice Hall..

García Ferrando, M. (1989). *Socioestadística. Introducción a la estadística en Sociología*. Madrid, Alianza Universidad.

Glass y Stanley (1974). *Métodos estadísticos aplicados a las ciencias sociales*. México: Prentice Hill.

Hopkins, K.D., Hopkins, B.R. y Glass, G.V. (1997, 3ª ed). Estadística básica para las ciencias sociales y del somportamiento. México: Prentice-Hall Hispanoamericana.

Johnson, R. y Kuby, P. (2004). *Estadística elemental*. México: Thompson.

Kalbfleisch, J. (1984). *Probabilidad e inferencia estadística*. Madrid: AC.

Macía, A., Lubin, P. y Rubio, P. (2000). *Psicología matemática II*. Madrid: UNED.

MacRae, S. (1995). *Modelos y métodos para las ciencias del comportamiento*. Barcelona: Ariel.

Martín Andrés, A. y Luna del Castillo, J.D. (2005). *Bioestadística para las Ciencias de la Salud*. Madrid: Norma.

Martínez G., A. (2000). *Diseños experimentales. Métodos y elementos de teoría.* México: Trillas.

Martínez, R., Maciá, M. y Pérez, J. (1998). *Psicología Matemática II.* Madrid: U.N.E.D.

Mendenhall, W., Wackerly, D. y Scheaffer R. (1994). *Estadística matemática con aplicaciones*. México: Grupo Editorial Iberoamericana.

Merino, J.M., Moreno, E., Padilla, M., Rodríguez Miñon, P. y Villarino, A. (2002). *Análisis de datos en psicología 1*. Madrid: UNED.

Meyer, P. (1992). *Probabilidad y aplicaciones estadística*. México: Addison-Wesley.

Moore, D. S. (1998). *Estadística aplicada básica*. Barcelona: Antoni Bosch, editor.

Mullor R. y Fajardo M.D. (2000), *Manual práctico de Estadística aplicada a las Ciencias Sociales.* Barcelona, Ariel.

Nortes Checa, A. (1993). *Estadística teórica y aplicada*. Barcelona: PPU.

Pardo, A. (2002). *Análisis de datos categóricos.* Madrid: UNED.

Pardo, A. y San Martín, R. (1994). *Análisis de datos en psicología II*. Madrid. Pirámide.

Peña, D. (1995). *Estadística. Modelos y métodos*. Madrid: Alianza Universitaria.

Peña, D. (2002). *Regresión y diseño de experimentos*. Madrid: Alianza Editorial.

Peña. D. y Romo, J. (1997). *Introducción a la estadística para las ciencias sociales.* Madrid: McGraw-Hill.

Padilla, M., Merino, J.M. y Pardo, A. (1986). *Psicología matemática I: Ejercicios resueltos*. Madrid: UNED.

Rios, S. (1967). *Métodos estadísticos.* Madrid: Ediciones del Castillo.

Silva, L.C. (1993). *Muestreo para la investigación en ciencias de la salud.* Madrid: Diaz de Santos.

San Martín, R. y cols. (1987). *Psicoestadística: Estimación y contraste*. Madrid: Pirámide

San Martín, R., Espinosa, L. y Fernández, L. (1987). *Psicoestadística descriptiva*. Madrid: Pirámide.

San Martín, R. y Pardo, A. (1989). *Psicoestadística. Contrastes paramétricos y no paramétricos*. Madrid: Pirámide.

Spiegel, M. R (1991). *Estadística*. Madrid: Mc Graw Hill.

Tanur, J. M., Mosteller, F., Kruskal, W. y otros. (1972). *Statistics: a guide to the unknown.* Holden Day. California.

Tukey, J. W. (1977). *Exploratory data analysis.* Nueva York: Addisson Wesley.

Wonnacot, T.H. y Wonnacot, R.J. (1991). *Estadística básica práctica*. México: Limusa.

Yáñez, L. (1989). *Fundamentos de psicología matemática*. Madrid: Pirámide.