Análisis de Datos y su Didáctica

Carmen Batanero y Juan D. Godino

GRANADA, 2001

Departamento de Didáctica de la Matemática
Universidad de Granada

Documentos de trabajo para la asignatura de libre configuración,

ANÁLISIS DE DATOS Y SU DIDÁCTICA

Profesores:

Carmen Batanero y Juan D. Godino

ISBN: 84-699-4296-6

Publica:
Grupo de Investigación en Educación Estadística
Departamento de Didáctica de la Matemática
Universidad de Granada

Imprime:
Servicio de Reprografía de la Facultad de Ciencias
Universidad de Granada
Avda. Fuentenueva s/n 18071 Granada

Financiación:
Proyecto BSO2000-1507, DGES, Ministerio de Educación y Ciencia.
Grupo de Investigación FQM-126. Consejería de Educación. Junta de Andalucía.

INDICE

ORIENTACION GENERAL DEL CURSO

Este material se ha preparado para que sirva de complemento al trabajo desarrollado en la asignatura de libre configuración "Análisis de datos y su didáctica", ofrecida por el Departamento de Didáctica de la matemática en la Facultad de Ciencias de la Educación de la Universidad de Granada.

El enfoque que damos al curso se basa en tres ideas fundamentales:

- Las aplicaciones de la estadística, que es hoy día un instrumento metodológico básico tanto en la investigación experimental, como en el mundo empresarial, político y profesional;
- El uso de ordenadores, que en la actualidad son un recurso imprescindible en el análisis de datos;
- El análisis didáctico de los conceptos y técnicas estadísticas.

En primer lugar, se trata de un curso con carácter aplicado, dirigido a estudiantes que usarán la estadística como herramienta en su trabajo futuro. Deseamos huir, por tanto, del desarrollo teórico excesivo para el cual existen en la actualidad cursos y libros de texto excelentes. Pensamos en el estudiante universitario **no matemático**, por lo que pretendemos que, desde el primer momento quede patente la aplicabilidad de cada uno de los temas expuestos. En éste sentido incluimos dos tipos de aplicaciones.

- Discusión de una serie de casos prácticos (proyectos) de análisis de datos. Cada uno de estos casos se refiere a un ejemplo específico de investigación (generalmente simplificado), y utiliza uno de los conjuntos de datos que se usará como material complementario. En cada ejemplo se analizan diferentes variables, ya que los fenómenos investigados en la realidad son, generalmente complejos, y no pueden reducirse a una única dimensión, sino que por el contrario, abarcan diversas variables, que se trata de relacionar. Estos proyectos serán utilizados para mostrar el uso de los programas de ordenador (Statgraphics) para la realización de diversos análisis. Cuando es posible se comparan los diversos métodos de análisis disponibles, así como las posibilidades y limitaciones de los mismos.
- Inclusión de ejemplos y actividades complementarias que, aunque resueltos con menor detalle, sirven para presentar ejemplos de la aplicabilidad de los métodos en áreas diversas: pedagogía, psicología, ciencias sociales y económicas, y para reflexionar sobre propiedades específicas de los conceptos y técnicas estadísticas.

La segunda idea a desarrollar en el curso será el uso de los ordenadores en una doble vertiente: como instrumento de cálculo y como recurso didáctico. Como hemos comentado, hoy día los cálculos y gráficos estadísticos se realizan con paquetes de programas y algoritmos. Los ordenadores actuales y la difusión del software estadístico han permitido el uso generalizado de la estadística, porque ha liberado al investigador de los cálculos (generalmente demasiado complejos para que puedan hacerse con calculadoras). Pero un buen instrumento de cálculo es ineficiente si el investigador no dedica el tiempo suficiente a la mejora de sus datos - mediante un adecuado diseño experimental - y a la interpretación de los resultados de los mismos.

Tenemos también que resaltar que la estadística, hoy día se usa con frecuencia incorrectamente, debido al desconocimiento de los usuarios de los principios básicos subyacentes. El hecho de aprender a manejar un programa de cálculo no evita que se aplique en situaciones inadecuadas o que se haga una interpretación incorrecta de sus resultados. Por este motivo, el énfasis del curso se hace sobre los aspectos de comprensión de los conceptos e

hipótesis supuestas, incidiendo menos en el detalle de los algoritmos. También queremos concienciar a los estudiantes de la importancia que tiene la consulta a un estadístico profesional en la fase de diseño de su trabajo. Mientras que nuestra intención es preparar a los alumnos para un uso autónomo de los procedimientos elementales, aconsejamos el recurso a un consultor estadístico para trabajos más complejos.

Respecto al uso de los ordenadores como recurso didáctico en la enseñanza de la estadística, observamos que la presencia de los ordenadores en los centros educativos va siendo un hecho cada vez más habitual, por lo que pensamos que pueden utilizarse sus posibilidades de cálculo y representación gráfica para cambiar la metodología de enseñanza de ciertos contenidos, simulando algunos de los fenómenos probabilísticos más característicos.

Consideramos que entre las dos alternativas extremas de presentar los resultados deductivamente, en su formulación final, o darles la correspondiente regla de aplicación práctica, cabe una tercera solución consistente en una aproximación gradual, intuitiva y experimental, orientada hacia la comprensión de los conceptos y de los teoremas, y a la captación de su necesidad y utilidad.

Es preciso intentar incorporar el enfoque heurístico y constructivo- frente al enfoque deductivista - en la presentación de los principales conceptos matemáticos a los alumnos de los primeros ciclos universitarios. Con frecuencia, en la metodología deductivista, se presentan los distintos teoremas ya "acabados" y "perfectos". Sin embargo, para llegar a cualquiera de estos resultados ha sido necesario, en primer lugar, su intuición, el comprobar para valores particulares cómo se cumple, y por último, tras varios intentos la demostración definitiva. Pensamos que, a través del uso del ordenador, se puede mostrar este proceso al estudiante, lo cual es más formativo y, además, no elimina, sino que refuerza, la "intuición matemática" y el razonamiento plausible.

El tercer objetivo del curso consiste en proporcionar a los estudiantes universitarios, interesados por la estadística aplicada, oportunidad de entrar en contacto con un componente importante de la didáctica de la matemática, como es la educación estadística. El conocimiento de algunos aspectos epistemológicos, psicológicos, curriculares y didácticos de los principales conceptos y métodos estadísticos ayudará al estudiante a comprender de una manera más completa la estadística y a mejorar su aplicación práctica.

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 que se encuentra dentro de una vaina de guisantes, o al observar la duración de un televisor, o el tiempo transcurrido entre dos llamadas a una central telefónica, etc. Igualmente, el resultado de cualquiera de los denominados juegos de azar, como lotería, dados, monedas, etc., 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 fué 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 espectacular 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.

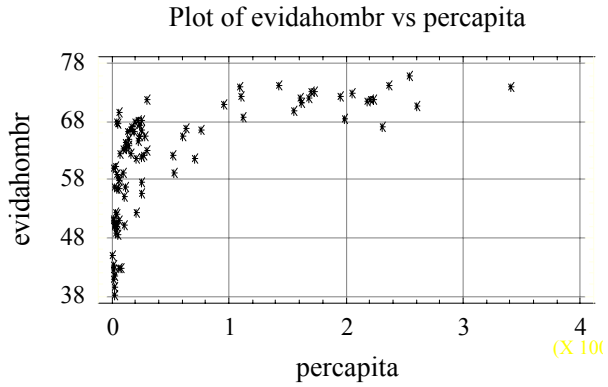
Con todo ello suge una nueva filosofía en los estudios estadísticos: el *análisis exploratorio de datos*, introducido por Tukey, quien compara la labor del estadístico con la de un detective.

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 presenten un valor estadísticamente significativo (en el ejemplo, el coeficiente de correlación sea significativamente distinto de cero), la relación entre las variables no se ajusta bien a una línea recta. En este caso, si no se representasen gráficamente los datos al investigador le faltaría descubrir algo importante: el modelo que mejor se ajusta a los datos no es una línea recta.



Características educativas del análisis exploratorio de datos

Al considerar la conveniencia o no de incluir un tema como objeto de enseñanza hemos de tener en cuenta su utilidad y que este tema se halle al alcance de los alumnos. Además de la utilidad, ya razonada, el análisis exploratorio de datos tiene las siguientes características que lo hacen un tema apropiado de estudio:

Posibilidad de generar situaciones de aprendizaje referidas a temas de interés del alumno: Lo usual es trabajar sobre un proyecto en el que se recogen datos, tomados de internet o un anuario estadístico u obtenidos mediante experimentos o la realización de una encuesta. Esto puede motivar bastante a los estudiantes, quienes pueden ver la utilidad dela estadística en la investigación.

Fuerte apoyo en representaciones gráficas: Una idea fundamental del análisis exploratorio de datos es que el uso de representaciones múltiples de los datos se convierte en un medio de desarrollar nuevos conocimientos y perspectivas y esto coincide con la importancia que se da a la representación gráfica en los nuevos diseños curriculares.

No necesita una teoría matemática compleja: Como el análisis de datos no supone que estos se distribuyen según una ley de probabilidad clásica (frecuentemente la normal), no

utiliza sino nociones matemáticas muy elementales y procedimientos gráficos fáciles de realizar y así cualquier alumno puede hacer análisis de datos.

Actividades

1.10. El análisis de datos se basa en el método de elaboración de proyectos por parte de los estudiantes. Describe algunos proyectos sencillos en los que los alumnos de secundaria puedan 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 tiene 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 psicofisiologia, 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.

Aplicaciones en el campo empresarial

Particularmente en este campo citaremos las siguientes aplicaciones:

- Volumen de ventas: Evolución en la empresa, valores medios, distribución territorial, oscilaciones estacionarias, tendencia. Estructura del volumen de venta. Relación de los colectivos parciales entre sí, y con el total. Concentración. Índices de ventas.
- Zonas de ventas: Estructuras de las zonas. Datos de mercado relacionados: población, empresas competidoras, renta, actividad económica, superficie bruta de viviendas..
- Servicio exterior: Personal, proveedores..
- Costes de comercialización y producción; gestión financiera.
- Stock de fabricación...

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 destreza 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 INTRODUCTORIOS

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 (http://www2.ncsu.edu/ncsu/pams/stat/info/jse/homepage.html) de la revista Journal of Statistical Education. 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;

PNB. 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 1990 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.

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 Saudita	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

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.

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 siguiente escala de actitudes:

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 algio de estadistica para ser un consumidor inteligente.
1 2 3 4 5
3. Ya que es fácil menir con la estadísticao, 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 aprendear 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

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 las componentes de las actitudes, así como la actitud global hacia la estadística y comparar según sexos, especialidades y estudios previos del tema.

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:

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, grupo sanguíneo, provincia de origen, etc.

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. Características de este tipo son: grado de mejoría de un paciente, las puntuaciones en test de aptitud, etc.

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 volumen de ventas, coste de producción, edad, cotización de un cierto tipo de acciones, etc.

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, si tiene o no carnet de conducir, 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:
- Tiempo para completar una tarea
- Número de años de escolaridad
- Número de sillas en una habitación

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

1.6. CODIFICACION DE LOS DATOS

Una de las fases del análisis de los datos es la preparación de los mismos para ser introducidos en el ordenador. Si el volumen de datos es pequeño, podemos realizar los

cálculos manualmente o con ayuda de una calculadora, en cuyo caso esta fase no será precisa.

Sin embargo, hoy día es frecuente realizar más de un análisis diferente sobre un mismo conjunto de datos. Podemos, por ejemplo, desear obtener ciertas representaciones gráficas y estadísticas simples de cada una de las variables en estudio y también efectuar un análisis de asociación o correlación entre dos o más variables.

Por otro lado, puede que nos interese informatizar nuestros datos por otros motivos. Un fichero en disco es fácilmente duplicable y ocupa menos espacio que otro convencional. Quizás precisemos conservar la información que poseemos para completarla posteriormente o bien intercambiarla con la de otro colega.

En todos estos casos, es necesario realizar una serie de operaciones preliminares con nuestros datos. La primera de ellas es la de *codificación.* Con este nombre se entiende el proceso de asignar un número u otro símbolo, de forma unívoca, a cada uno de los valores posibles de las variables que estamos utilizando, y transcribir nuestras observaciones, con un formato especialmente diseñado, a unos impresos preparados con tal fin que se conocen como *hojas de codificación*. Este proceso es necesario para la organización informática de nuestros datos. En la sección siguiente se comenta cómo tiene lugar esta organización.

En la Figura 2.2 se muestra un ejemplo de hoja de codificación utilizada en un estudio de seguimiento médico. Como puede observarse, está dividida en filas y columnas. Cada fila constituye la información disponible sobre un mismo enfermo y se conoce con el nombre de *registro.*

Figura 2.2. Hoja de codificación de un fichero de seguimiento médico

Paciente			V1	V2	V3				V4	V5	V6		. . .				
			Edad	Sexo	Fecha de la operación												
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0	0	1															
0	0	2															
.																	

Tradicionalmente, cada línea de la hoja de codificación está dividida en 80 columnas. Esto es debido a que el medio de almacenamiento primitivamente empleado - la ficha perforada - tenía capacidad para 80 caracteres (letras o números). Sin embargo, los modernos disquetes o CDs no poseen esta limitación, por lo que, si es preciso puede diseñarse una hoja de codificación de más de 80 columnas.

Cada una de las columnas de esta hoja se usará, por tanto, para escribir en ella un número, letra o signo (por ejemplo, el punto decimal), que represente los valores observados en las variables estadísticas estudiadas. Según el tipo de variable podemos necesitar para codificarlas una o varias columnas. El conjunto de columnas que se destinan

a contener la información sobre una variable se conoce como *campo*. Así, para codificar el sexo de un individuo podemos dedicar a ello una sóla columna y adoptar la siguiente regla de codificación:

1 : indica varón; 2: indica hembra

Para codificar el número de miembros de una unidad familiar podemos destinar dos columnas y escribir en ellas dicho número, pues es posible tener familias de 10 o más miembros y no es probable que encontremos ninguna de mas de 92. Para codificar el peso de un individuo habrá que especificar el número de cifras decimales deseadas. Supongamos que decidimos emplear dos decimales. Entonces, precisaremos 6 columnas para la variable peso: las tres primeras para la parte entera, la cuarta para el punto decimal y las dos últimas para las cifras decimales. Podemos representar simbólicamente esta codificación en la forma ###.##, donde el símbolo # (se suele leer 'almohadilla') representa un dígito.

Una vez decidido el sistema de codificación, se prepara la hoja de codificación apropiada, en la que se ha indicado, en la cabecera cada uno de los campos, separando convenientemente el grupo de columnas que lo forman.

Actividades

1.18. En una encuesta codifico la provincia de nacimiento con un número de 1 a 50. ¿Qué tipo 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?

REFERENCIAS

Cabriá, S. (1994). *Filosofía de la estadística.* Servicio de Publicaciones de la Universidad de Valencia.

Tanur, J. M., Mosteller, F., Kruskal, W. y otros. (1972). *Statistics: a guide to the unknown.* Holden Day. California.

Rouncenfield (1995). The statistics of poverty and inequality. *Journal of Statistics Education,* 3(2).

TEMA 2

DISTRIBUCIONES DE FRECUENCIAS Y GRAFICOS

2.1. DISTRIBUCIÓN DE FRECUENCIAS DE VARIABLES ESTADÍSTICAS CUALITATIVAS

Cuando se comienza a analizar una nueva variable estamos interesados es saber los valores que puede tomar, el número total de datos y cuantas 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 dicho anteriormente 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.

Porcentajes

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

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

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 practican 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.

Figura 2.1. Población activa en Jaén (1980) según relación laboral

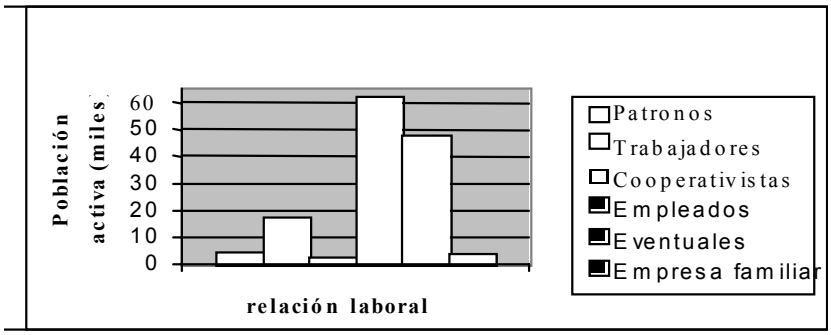
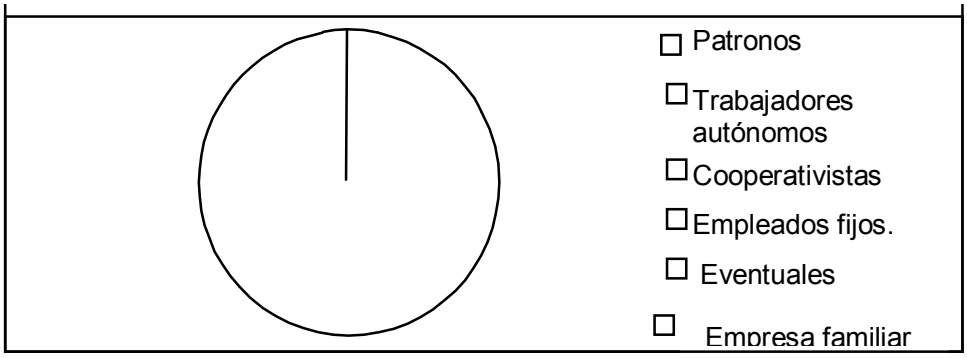


Gráfico de sectores

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



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 DESCRIBE, y dentro de él CATEGORICAL DATA. Allí se selecciona TABULATION, en la ventana TABULAR OPTIONS y en la ventana de diálogo que aparece, se selecciona la variable que se desea. Por ejemplo, en la Tabla 2.3. presentamos la tabla de frecuencias de la variable A2 en el fichero ACTITUDES, del 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?

El significado de las columnas es:

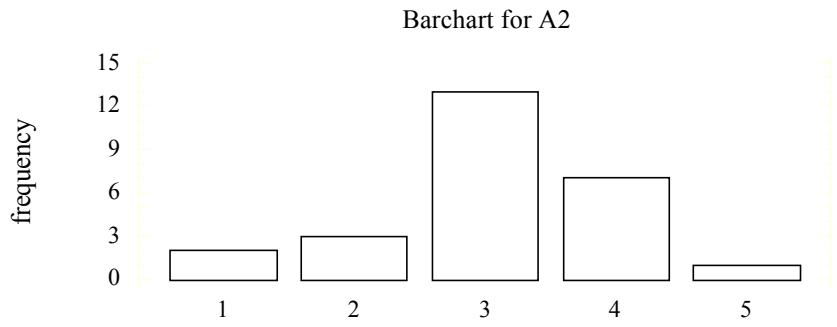
Class: clase
Value: Valor de la variable
Frequency: Frecuencia absoluta
Relative Frequency: Frecuencia relativa
Cumulative Frequency: Frecuencia acumulada
Cum. Rel. Frequency: Frecuencia acumulada relativa

Tabla 2.3. Tabla de frecuencias para la puntuación en A2

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
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 CATEGORICAL DATA, en GRAPHICAL OPTIONS se pueden realizar el diagrama de barras (BARCHART) y el gráfico de sectores (PIE CHART). Cada uno de ellos tiene diversas opciones que pueden explorarse pulsando la tecla PANE OPTIONS. Como ejemplo, en la figura 2.3. representamos el diagrama de barras correspondiente a la tabla 2.3.

Figura 2.3. Diagrama de barras de puntuación en utilidad de la estadística



2.3. VARIABLES CUANTITATIVAS: FRECUENCIAS ACUMULADAS

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

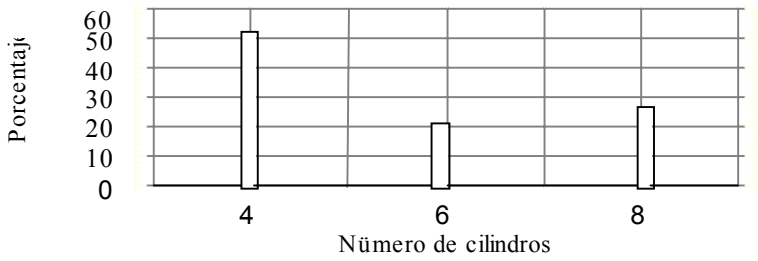
Ejemplo 2.2. En la Tabla 2.4 podemos observar la distribución de frecuencias de la variable "número de cilindros" de un conjunto de 398 tipos de automóviles de diferentes marcas y modelos, fabricados en Europa, Japón y Estados Unidos y en la Figura 2.3 representamos el diagrama de barras correspondiente.

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		

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.

Figura 2.4. Distribución del número de cilindros en automóviles



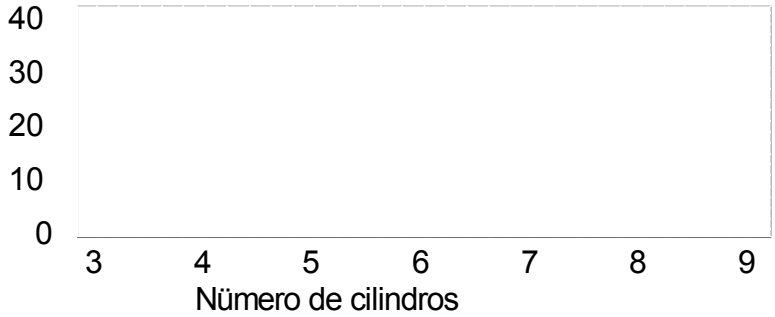
Frecuencias acumuladas

Algunas veces, en una variable estadística, es interesante conocer el número de valores que son menores que uno 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 3 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 3.3, 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 2.4. En esta gráfica podemos ver cómo las frecuencias acumuladas experimentan un aumento en cada valor de la variable.

Figura 2.5. Distribución del número de cilindros en automóviles



Actividades

2.4. Sabiendo que la frecuencia absoluta de alumnos que tienen 3 hermanos es 30 y que la frecuencia acumulada de alumnos que tienen 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?

2.4. VARIABLES AGRUPADAS: INTERVALOS DE CLASE

Algunas variables cuantitativas toman valores aislados -variable discreta- (nº 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 altura 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

Tabla 2.6. Frequency Tabulation for alturas

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		140,0		0	0,0000	0	0,0000
1	140,0	150,0	145,0	2	0,0339	2	0,0339
2	150,0	160,0	155,0	15	0,2542	17	0,2881
3	160,0	170,0	165,0	22	0,3729	39	0,6610
4	170,0	180,0	175,0	18	0,3051	57	0,9661
5	180,0	190,0	185,0	2	0,0339	59	1,0000
above	190,0			0	0,0000	59	1,0000

Mean = 166,542 Standard deviation = 8,15438

Obtención de tablas de frecuencias agrupadas con STATGRAPHICS

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. Para ello se siguen los pasos siguientes:

- Ir al menú DESCRIBE. Elegir la opción NUMERIC DATA y luego ONE VARIABLE ANALYSIS. Seleccionar la variable con la que se desea trabajar;

- Pulsar el botón TABULAR OPTIONS, seleccionando FREQUENCY TABULATION en los resúmenes numéricos;
- Para cambiar el ancho de intervalo o la cantidad de intervalos: estando en la tabla de frecuencias, con el botón derecho del ratón seleccionar la opción PANE OPTIONS, aparecerá un cuadro de diálogo en el que se puede definir la cantidad de intervalos (NUMBER OF CLASSES), el límite inferior (LOWER LIMIT) del recorrido de la variable y el límite superior (UPPER LIMIT).

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.

Actividades

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

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

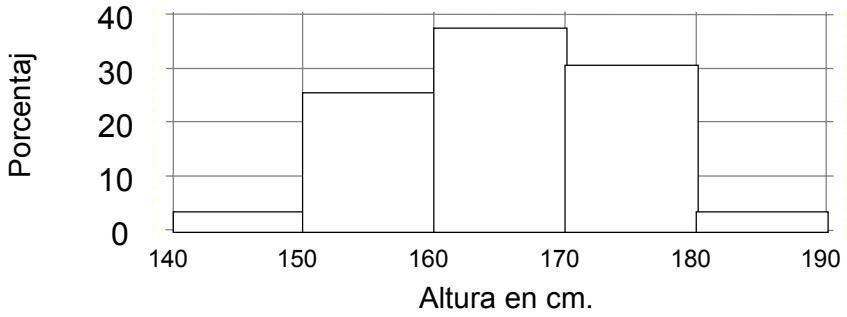
2.5. HISTOGRAMAS Y POLÍGONOS DE FRECUENCIAS

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 2.6 se presenta el histograma de frecuencias de la variable altura del conjunto de datos ALUMNOS, correspondientes a la Tabla 2.6.

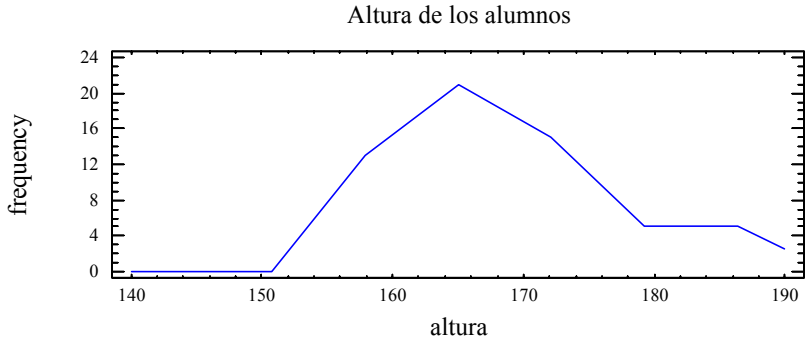
Figura 2.6. Distribución de la altura de un grupo de



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 2.7 se representa un polígono de frecuencias de la variable altura del conjunto de los datos de la tabla 2.6.

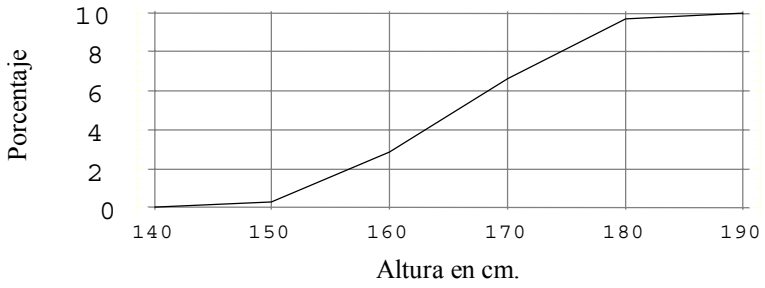
Figura 2.7. Polígono de frecuencias de la distribución de las alturas de los alumnos



Polígono acumulativo de frecuencias

La Figura 2.7. representa el polígonos acumulativos 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 2.8. Polígono acumulativo de alturas de un grupo de estudiantes



Para realizar el histograma y polígonos con STATGRAPHICS:

- Ir al menú DESCRIBE. Elegir la opción NUMERIC DATA y luego ONE VARIABLE ANALYSIS. Seleccionar la variable con la que se desea trabajar;
- Ubicados en la ventana de opción de gráficos (GRAPHICAL OPTIONS), se selecciona el histograma de frecuencias (FREQUENCY HISTOGRAM):
- Para realizar el polígono de frecuencias: Sobre la ventana en la que aparece el histograma, hacer clic con el botón derecho del ratón (PANE OPTIONS), aparecerá un cuadro de diálogo en el que aparece seleccionado por defecto el histograma, seleccionar el polígono (POLYGON) y si lo que se desea es trabajar con las frecuencias relativas, seleccionar la opción RELATIVE .

Actividad 2.8. En las figuras 2.9 y 2.10 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 2.9. Histogramas. Distribución de la esperanza de vida en hombres y mujeres

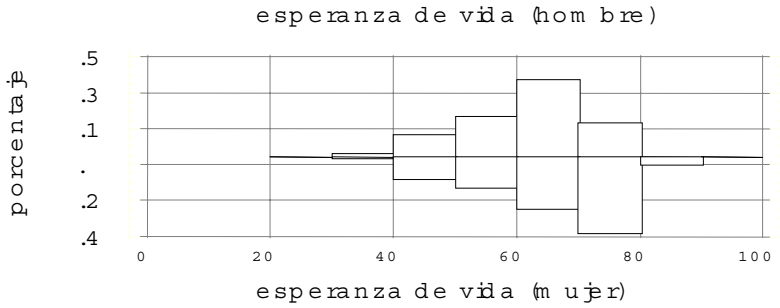
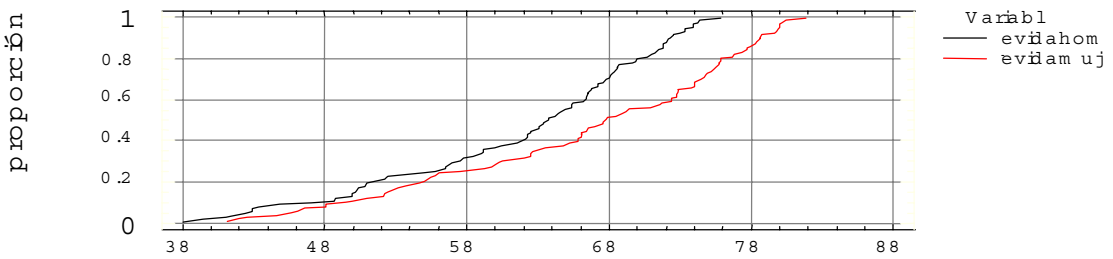


Figura 2.10. Distribución acumulativa de la esperanza de vida en hombres y mujeres



2.6. GRÁFICO DEL TRONCO

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 "el tronco" (o también, gráfico del tallo y las 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 "tronco". Los dígitos que nos quedaban los vamos escribiendo en la fila correspondiente de forma ordenada.

Por ejemplo, tomaremos el conjunto de datos que representan la talla de los alumnos de la Tabla 2.4

1. Los ordenamos de menor a mayor en la Tabla 2.7:

Tabla 2.7

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

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 2.11

Figura 2.11. Gráfico del tronco

```
14 | 9
15 | 0134669
16 | 00000000011122244466778899 99
17 |  00001222233344556688 99
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 tronco 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. El gráfico de la Figura 2.11. 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 2.12.

Figura 2.12. Gráfico del tronco extendido

```
------------------------------
14* |
14. | 9
15* | 0134
15. | 669
16* | 000000000111222444
16. | 6677889999
17* | 00001222233344
17. | 55668899
18* | 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 del tronco, 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 a comparaciones inapropiadas. Dos distribuciones con distinta escala vertical son difíciles de comparar.

Actividades

2.9. En cada uno de los siguientes gráficos (Figura 2.13) 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 2.13 Gráficos de 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

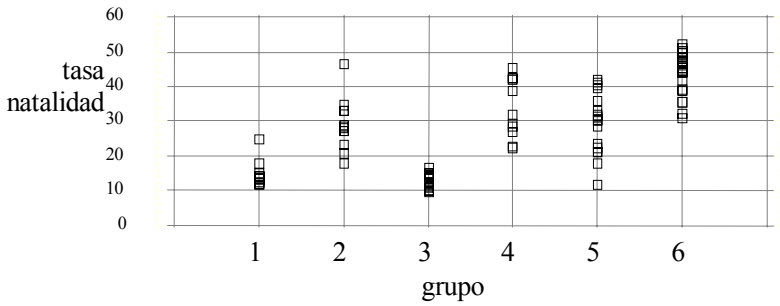
2.10. En la figura 2.14 representamos la distribución de las tasas de natalidad del conjunto de datos relativo al Proyecto 1.¿Por qué en este caso conviene agrupar en intervalos? ¿Cuántos intervalos conviene usar en la tabla de frecuencias? Construye un histograma de frecuencias. Compara con tus compañeros cómo cambia la forma al variar el número de intervalos.

En la figura 2.15 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 grupo.

Figura 2.14: Gráfico de tallo y hojas: Tasa de natalidad

0|99
1|00111122222333333444444
1|556778
2|011222334
2|677888899
3|00111122234
3|5568899
4|011122224444
4|55566677788888
5|0012

Figura 2.15: Diagrama de puntos



2.7. NIVELES Y DIFICULTADES EN LA COMPRENSIÓN DE GRÁFICOS

La destreza en la lectura crítica de datos es un componente de la alfabetización cuantitativa y una necesidad en nuestra sociedad tecnológica. Curcio (1989) describe tres niveles distintos de comprensión de los gráficos:

(a) "Leer los datos": este nivel de comprensión requiere una lectura literal del gráfico; no se realiza interpretación de la información contenida en el mismo.

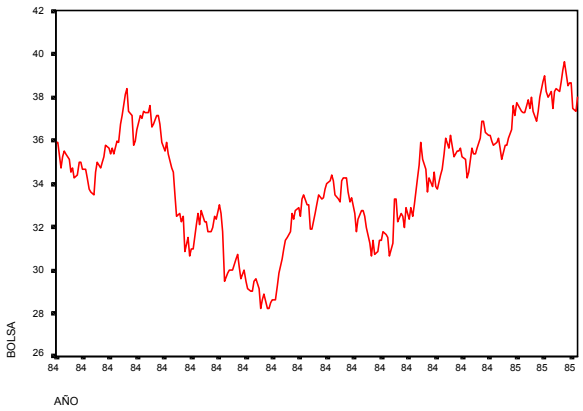
(b) "Leer dentro de los datos": incluye la interpretación e integración de los datos en el gráfico; requiere la habilidad para comparar cantidades y el uso de otros conceptos y destrezas matemáticas.

(c) "Leer más allá de los datos": requiere que el lector realice predicciones e inferencias a partir de los datos sobre informaciones que no se reflejan directamente en el gráfico.

(d) "Leer detrás de los datos": supone valorar la fiabilidad y completitud de los datos.

Por ejemplo, si analizamos las tareas que se requieren en la interpretación de la figura 2.16, "leer los datos" se refiere a cuestiones sobre la lectura de las escalas o encontrar el valor de una de las coordenadas de uno de los puntos, dado el valor de la otra coordenada. "Leer dentro de los datos" se refiere, por ejemplo, a cuestiones sobre la tendencia, sobre si podría ser representada o no mediante una función lineal o sobre si se observan ciclos. La predicción del comportamiento de la serie para los próximos meses, requeriría el trabajo en el nivel de "leer más allá de los datos". "Leer detrás de los datos" supondría valorar si los datos son completos, analizar la forma en que fueron recogidos y detectar posibles sesgos.

Figura 2.16. Evolución de cotizaciones



Curcio estudió el efecto que sobre la comprensión de las relaciones matemáticas expresadas en los gráficos tienen los siguientes factores:

- conocimiento previo del tema al que se refiere el gráfico; si el alumno está o no familiarizado con el contexto.
- conocimiento previo del contenido matemático del gráfico, esto es, los conceptos numéricos, relaciones y operaciones contenidas en el mismo;
- conocimiento previo del tipo de gráfico empleado (gráfico de barras, pictograma, etc.).

Todos estos puntos intervienen en la dificultad que tienen los alumnos para interpretar los gráficos. Encontró que las principales dificultades aparecen en los niveles superiores ("leer dentro de los datos", "leer más allá de los datos", "leer detrás de los datos"), pero que las dificultades disminuyen con la edad.

Li y Shen (1992) muestran ejemplos de elección incorrecta del tipo de gráfico en los proyectos estadísticos realizados por los estudiantes de secundaria. Algunos alumnos utilizaron un polígono de frecuencias con variables cualitativas, o un diagrama de barras horizontal para representar la evolución del índice de producción industrial a lo largo de una serie de años. Con frecuencia la elección de las escalas de representación son poco adecuadas para el objetivo pretendido. Los autores incluyen, además, una lista de errores de carácter técnico entre los cuales destacamos los siguientes:

- omitir las escalas en alguno de los ejes horizontal o vertical, o en ambos;

- no especificar el origen de coordenadas;

- no proporcionar suficientes divisiones en las escalas de los ejes.

Este problema se agrava por la disponibilidad de "software" para la representación gráfica y el desconocimiento del modo correcto en que debe ser empleado por parte de los alumnos. Otras veces, el empleo inadecuado del "software" gráfico se debe a las concepciones incorrectas del estudiante, como al obtener un diagrama de sectores en los que éstos no son proporcionales a las frecuencias de las categorías, o comparar cantidades heterogéneas en un mismo gráfico.

Actividades

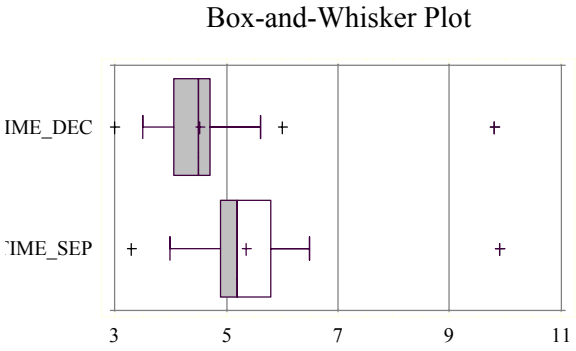
2.11. Buscar ejemplos en la prensa de tablas estadísticas o gráficos que presenten errores de construcción o que induzcan a obtener conclusiones equivocadas. Elaborar una lista de los principales tipos de errores detectados.

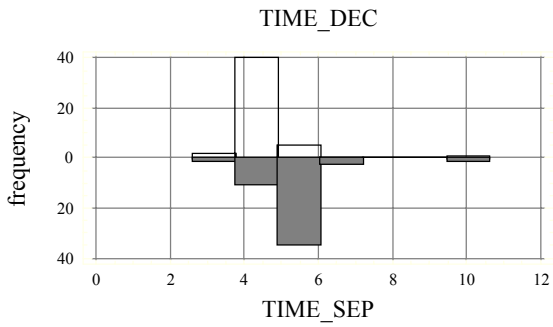
2.12. Analizar los errores posibles en la interpretación y elaboración de los siguientes gráficos estadísticos: diagrama de sectores, histograma, polígono acumulativo de frecuencias, gráfico del tronco.

2.13. Buscar ejemplos de gráficos incorrectos o instrucciones inapropiadas para la realización de gráficos estadísticos en los libros de texto de enseñanza primaria o secundaria. ¿Qué obstáculos didácticos se deducirían para los estudiantes?

2.14. Sobre un mismo conjunto de datos produce un gráfico que los represente adecuadamente y otro en que los datos queden distorsionados.

2.15. Supongamos que queremos estudiar si existe o no diferencia en el tiempo que tardan unos alumnos en correr 30 metros en Septiembre y en Diciembre (después de 3 meses de entrenamiento). 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

MEDIDAS DE TENDENCIA CENTRAL, DISPERSIÓN Y FORMA 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 este 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.

Actividad 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).

Actividad 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	Gia	Hilda	Ines	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

Actividad 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)

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

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.

Actividad 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?

Actividad 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.

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

Actividad 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 3.2 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 3.2 a 3.7. Pon otros ejemplos de situaciones en que se apliquen las propiedades que no correspondan a ninguna de las actividades anteriores.

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.

$$\Sigma(x_i - \bar{x})f_i = 0$$

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

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

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:

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

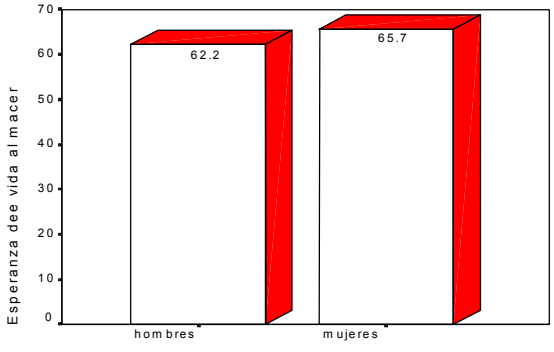
Esta propiedad implica que, al efectuar un cambio en el origen desde el que se han medido los datos, la media queda afectada por dicho cambio de origen.

Actividad 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?

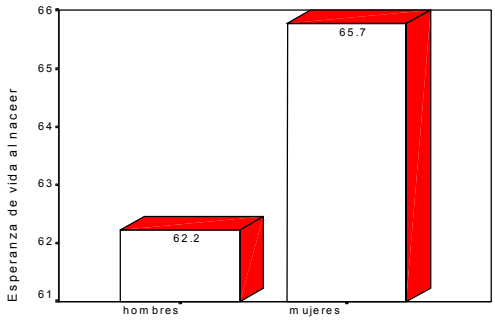
Actividad 3.9. ¿Qué representa el valor obtenido al calcular la media aritmética simple de la esperanza media de vida al nacer en los 97 países del Proyecto 2? ¿Cómo habría que hacer para calcular la esperanza media de vida al nacer en hombres y mujeres, si no tenemos en cuenta el país de nacimiento?

En la Figura 3 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. Esperanza de vida media en hombres y mujeres



Cada país ponderado por número de habitantes



Cada país ponderado por número de habitantes

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;

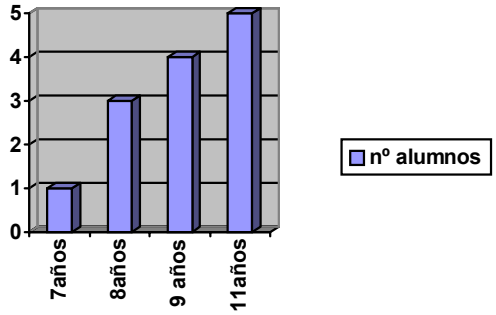
> **Ejemplo 3.1**. Supongamos que una asignatura se divide en tres exámenes parciales, en los que han entrado respectivamente, 2, 3 y 5 temas. Si un alumno ha obtenido 3, 4 y 7 puntos, respectivamente en estos exámenes, ¿Estarías de acuerdo en darle como nota final un 4,3?

3.3. LA MODA

Cuando la variable es cualitativa no podemos calcular la media,. Para describir un grupo podemos, entonces usar la moda $\mathbf{M_O}$, 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:

a) 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 3.1. Edad de un grupo de alumnos



Ejemplo 3.2. En la figura 3.1 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.

b) 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).

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

donde M_O, representa la moda, $\mathbf{E_i}$, 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 como medida de tendencia central. Veamos dos de ellas.

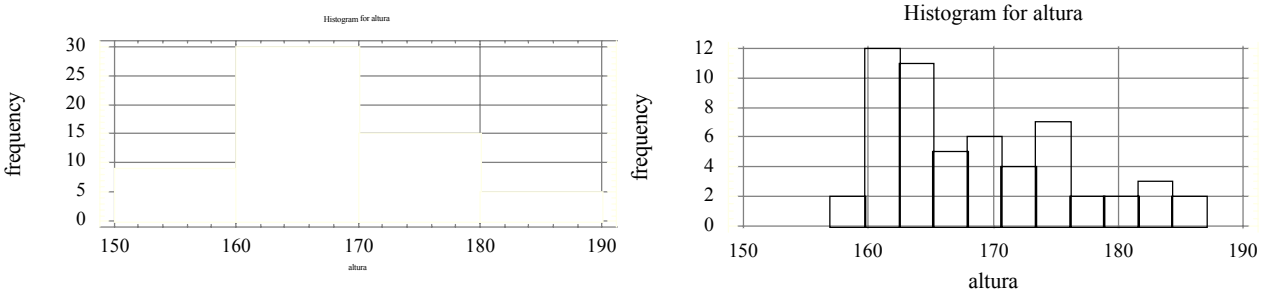
a) Si las frecuencias se condensan fuertemente en algunos valores de la variable, la moda no es una medida eficaz de tendencia central.

   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
Nº	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.

b) Una misma distribución con los valores agrupados en clases distintas, puede dar distinta moda, en el cálculo aproximado.



Ejemplo 3.4. Consideremos en el conjunto de datos ALUMNOS la variable altura, cuyas frecuencias vienen representadas en los dos histogramas de frecuencia siguientes. 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.

Algunos problemas en el aprendizaje de los promedios

Además de ser uno de los principales conceptos estadísticos, la media tiene muchas aplicaciones en cuestiones prácticas de la vida diaria. Sin embargo no siempre se usa adecuadamente. Un error muy frecuente es no ponderar adecuadamente en el cálculo de promedios, como en el problema presentado en la actividad 3.8. Las situaciones en las cuales se debe calcular una media ponderada y la selección de los correspondientes pesos no son fácilmente identificados por los estudiantes. Por ejemplo, cuando los datos se agrupan en intervalos, los estudiantes olvidan con frecuencia que cada uno de estos grupos debería ponderarse de modo distinto al calcular la media.

Respecto a la comprensión de los aspectos conceptuales, Strauss y Bichler (1988) investigaron el desarrollo evolutivo de la comprensión de esta noción en alumnos de 8 a 12 años, distinguiendo las siguientes propiedades:

a) La media es un valor comprendido entre los extremos de la distribución.
b) La suma de las desviaciones de los datos respecto de la media es cero.
c) El valor medio es influenciado por los valores de cada uno de los datos.
d) La media no tiene por qué ser igual a uno de los valores de los datos.
e) El valor obtenido de la media puede ser una fracción (ello puede no tener sentido para la variable considerada).
f) Hay que tener en cuenta los valores nulos en el cálculo de la media.
g) La media es un "representante" de los datos a partir de los que ha sido calculada.

Como se sabe la media es un valor “típico ‘‘ o “representativo” de los datos. Campbell (1974) observa que, debido a ello, se tiende a situar la media en el centro del recorrido de la distribución, propiedad que es cierta para distribuciones simétricas. Pero cuando la distribución es muy asimétrica la media se desplaza hacia uno de los extremos y la moda o la mediana serían un valor más representativo del conjunto de datos. La comprensión de la idea de “valor típico” implica, según Russel y Mokros (1991), tres tipos diferentes de capacidades:

- Dado un conjunto de datos, comprender la necesidad de emplear un valor central, y elegir el más adecuado.
- Construir un conjunto de datos que tenga un promedio dado.
- Comprender el efecto que, sobre los promedios (media, mediana o moda), tiene un cambio en todos los datos o parte de ellos.

Watson y Moritz (2000), analizan el significado intuitivo dado por los niños al término "promedio" y hallan un gran número de niños para los cuales el promedio es simplemente un valor en el centro de la distribución (es una idea próxima al concepto de mediana). Pocas veces se relaciona la palabra "promedio" con la moda y menos aún con la media aritmética. Las siguientes definiciones de "promedio" fueron obtenidas en entrevistas a niños realizadas: "*Significa igual", "que es normal", "no eres realmente bueno, pero tampoco malo".*

Al preguntar que quiere decir que el número medio de niños por familia es 2'3, obtienen respuestas correctas y otras como las siguientes: *"Que tienen dos niños grandes y otro que no ha crecido todavía", "que en las familias australianas el número más frecuente de niños es 2'3", "el '3 es un niño que tiene que crecer para hacerse mayor. Por ejemplo, tiene 3 años ahora y cuando cumpla 10, contará como 1 y entonces el número promedio de niños será 3".* Para el profesor los enunciados sobre los promedios pueden parecer muy claros, pero estas respuestas indican la necesidad de poner atención al significado que las palabras y valores numéricos tienen para los estudiantes en relación a contextos específicos.

Podría parecer que este error solo se presenta en los niños. Sin embargo, Eisenbach (1994) plantea a estudiantes universitarios en un curso introductorio de estadística el significado de la frase: "*¿Qué quiere decir que el salario medio de un empleado es 3.600 dólares?"* obteniendo respuestas como "*que la mayoría de los empleados gana alrededor de 3.600 dólares",* o que *"es el salario central; los otros trabajadores ganan más o menos de 3600 dólares",* que muestran la confusión terminológica entre las palabras "media", "mediana" y "moda".

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.

Actividad 3.10. A continuación reproducimos datos sobre número de pulsaciones por minuto en diversas especies animales[1]

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

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

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

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

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

o iguales como inferiores o iguales a él. La representaremos por $\mathbf{M_e}$. Para el cálculo de la mediana, distinguiremos entre datos no agrupados y agrupados en clases.

1) 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 3.4. 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 3.5. En la actividad 3.9 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.

2)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:

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 3.2)

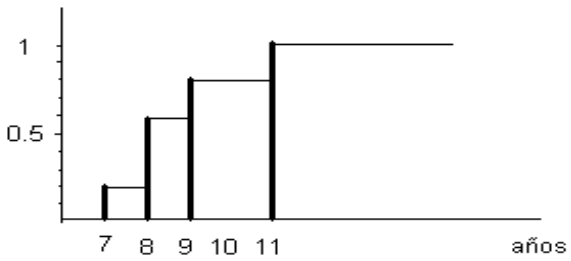
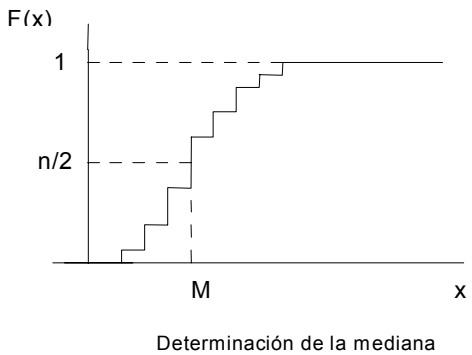


Figura 3.2. Cálculo de la mediana con un número impar de datos



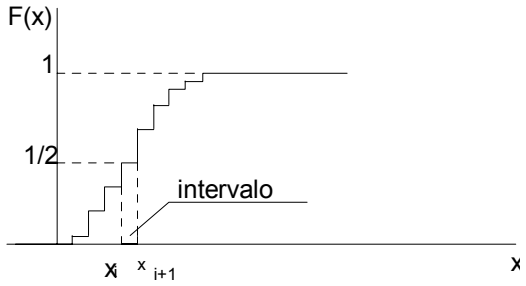
Determinación de la mediana

Figura 3.3. Distribución acumulada de edades

Ejemplo 3.6. En la figura 3.3 presentamos el diagrama de frecuencias acumuladas correspondiente a la distribución de edades de la figura 3.1. El valor de la variable correspondiente a la ordenada 0.5 es 8 años, luego este es el valor de la mediana.

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

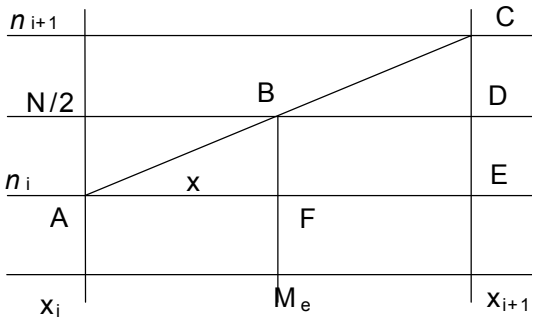
En el ejemplo 3.6 si añadimos un alumno de 11 años, tanto los valores 8 como 9 cumplen la definición de mediana. Tomamos, entonces como edad mediana 8.5

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:

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

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

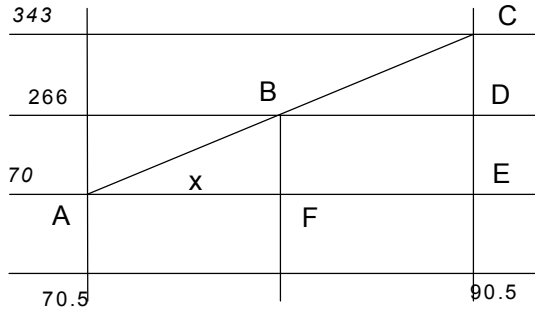
Puesto que $x_{i+1} - x_i$ es la amplitud del intervalo, CE la frecuencia en el intervalo mediano y BF la diferencia entre N/2 y la frecuencia relativa acumulada en el intervalo mediano, obtenemos la cantidad que hay que sumar al extremo inferior del intervalo mediano para calcular la mediana.

Ejemplo 3.7. La siguiente tabla se refiere al sueldo mensual en miles de pesetas, de los 531 trabajadores de una fábrica.

Sueldo	f_i	F_i
30.5-50.5	10	10
50.5-70.5	60	70
70.5-90.5	273	343
90.5-110.5	128	471
110.5-130.5	45	516
130.5-150.5	13	529
150.5-170.5	2	531

En la Figura 3.6 se representa el polígono acumulativo de frecuencias de la distribución anterior.

Figura 3.6. Detalle del polígono de frecuencias acumuladas en el ejemplo 3.7.



Para más claridad, solamente hemos representado la clase mediana, la anterior y posterior. Los triángulos rectángulos ABF y ACE son semejantes lo que nos da la relación (3.3)

$$(3.3).....\frac{AB}{AE}=\frac{CD}{CE}$$

La mediana será el punto de abscisa 70.5 + x, siendo x la longitud del segmento AF. La longitud de los segmentos que nos interesa es:

AF = x, AE = 90.5 - 70.5 = 20, CE = 343 - 70 = 273, DE = 266 - 70 = 196

Aplicando (3.3) tenemos,

$$\frac{x}{20}=\frac{196}{273};\quad x = 14.36,$$

y la mediana será $M_e = 70.5 + x = 84.86$

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	56678	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.

Actividad 3.11. En las siguientes gráficas hemos representado el número promedio de habitantes en cada país, de los datos del Proyecto 2, según grupo, usando dos promedios diferentes: media y mediana.

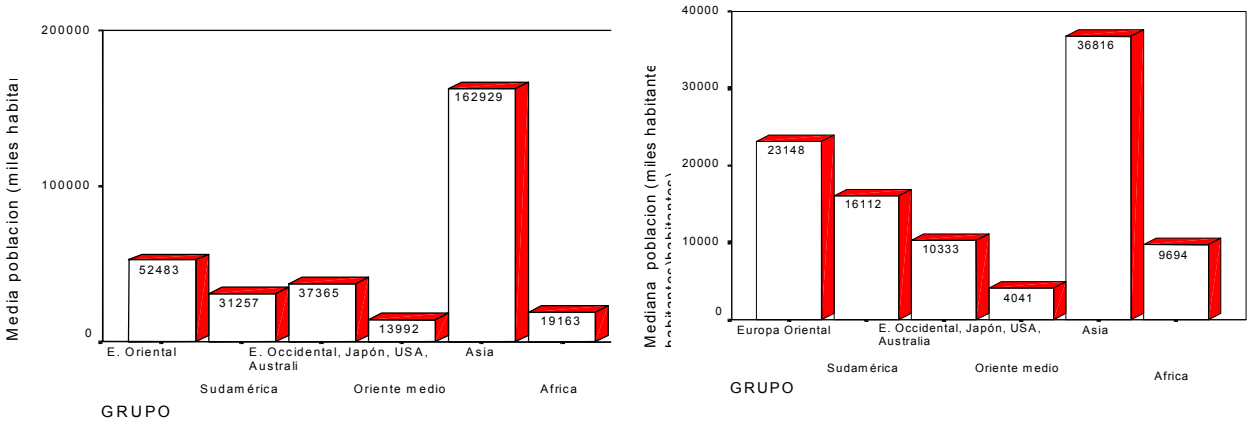


Figura 2: Mediana y media del número de habitantes en los diferentes grupos de países

a) Explicar 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?

Actividad 3.12 . 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.

a) Al calcular la mediana no usamos todos los valores observados de la variable, lo que la limita como medida de tendencia central.

Ejemplo 3.8. 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.

b) No puede ser aplicada a distribuciones de variables cualitativas.

Actividad 3.13. 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?

Actividad 3.14. 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?

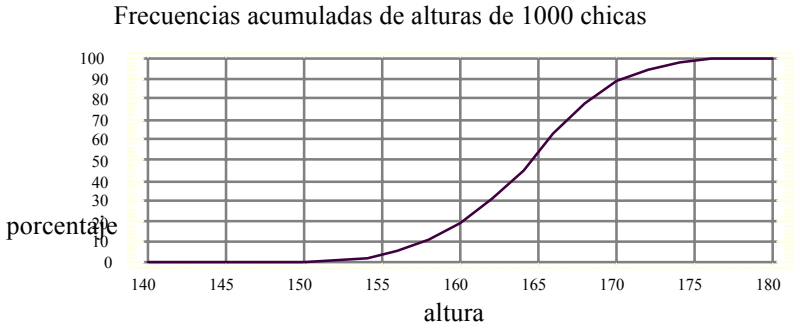
c) ¿Cuál es el valor de la altura tal que el 70 % de las chicas tiene una altura igual o inferior (percentil del 70%)?

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?

Figura 3.7.



---

c) Como medida de tendencia central, presenta ciertas ventajas frente a la media en algunas distribuciones ya que 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.

d) *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.

Actividad 3.15. ¿Cuál de las medidas de posición central permanece constante si cambio un valor extremo de los datos?

Actividad 3.16. La estatura mediana de un grupo de alumnos es de 156 cm. ¿Cuál será la nueva estatura si expresamos la estatura en metros?

---

c) La mediana e*s 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.

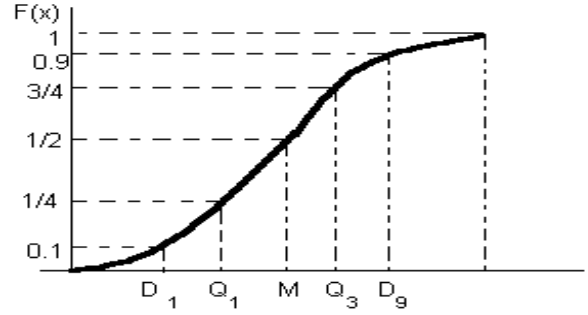
Cuantiles

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.

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.

Figura 3.8. Cuantiles, cuartiles y mediana



Determinación de los cuantiles

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 90*5321/100 = 477.80, el percentil pertenece a la clase (110.5-130.5). De ello se deduce:

$$P_{90} = 110.5 + \frac{477.90 - 471}{45} \times 5 = 111.27$$

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.

Actividad 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?

Actividad 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?

Dificultades en el estudio de los estadísticos de orden

El estudio de los estadísticos de orden presenta dificultades. En primer lugar, el cálculo de la mediana, percentiles y rango de percentiles se enseña empleando un algoritmo diferente para el caso de variables estadísticas agrupadas en intervalos o no agrupadas. Como sabemos, la opción de agrupar o no en intervalos se toma a juicio del que analiza los datos. Por ello, incluso los alumnos universitarios encuentran difícil aceptar que se pueda emplear dos algoritmos diferentes de cálculo para el mismo promedio y que puedan obtenerse valores distintos para el mismo parámetro, al variar la amplitud de los intervalos de clase.

Otros problemas se presentan al interpretar la gráfica de frecuencias acumuladas de variables discretas, debido a que presenta discontinuidades de salto y su inversa no es una aplicación: en esta correspondencia un punto puede tener más de una imagen, o vanos puntos pueden tener la misma imagen. Este es un tipo de función al que los alumnos no están acostumbrados.

Hay también bastante diferencia entre la definición de la mediana y el método de cálculo que se emplea para obtener su valor. Desde la definición de la mediana como "valor de la variable estadística que divide en dos efectivos iguales a los individuos de la población supuestos ordenados por el valor creciente del carácter", hasta su cálculo basado en la gráfica de frecuencias acumuladas intervienen una serie de pasos no siempre suficientemente comprendidos.

Barr (1980) llama, la atención sobre la falta de comprensión de los estudiantes sobre la mediana en un estudio llevado a cabo con estudiantes de edades entre 17 y 21 años. El 49% dio una respuesta incorrecta a la cuestión siguiente:

La mediana del siguiente conjunto de números.

1, 5, 1, 6. 1, 6, 8 es

a) 1; b) 4; c) *5; d) 6: e) (otro valor); f) no sé:*

La mayoría de los alumnos entiende la idea de mediana como valor central, pero no tienen claro a que secuencia numérica se refiere ese valor central. Los estudiante pueden interpretar la mediana como el valor central de los valores de la variable, de las frecuencias o incluso de la serie de datos antes de ser ordenada.

3.5. CARACTERISTICAS DE DISPERSION

Las medidas de tendencia central nos indican los valores alrededor de los cuales se distribuyen los datos.

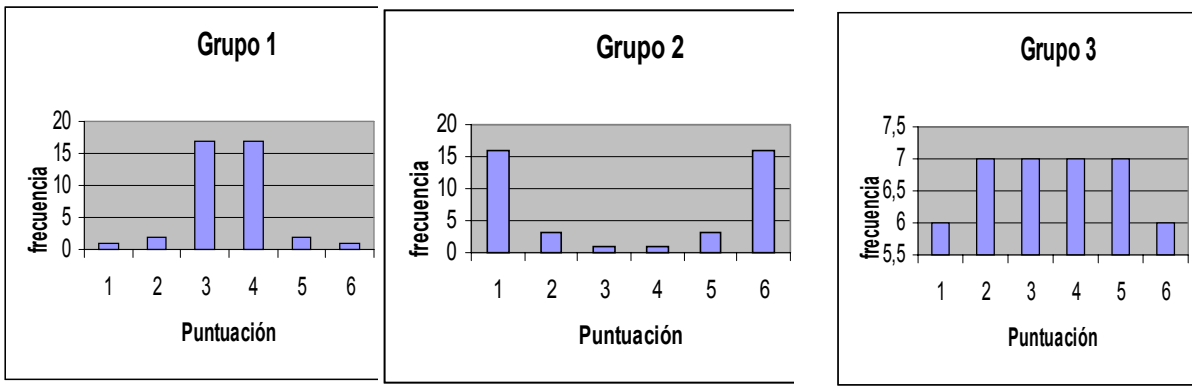
Las características de *dispersión* son estadísticos que nos proporcionan una medida del mayor o menor agrupamiento de los datos respecto a los valores de tendencia central. Todas ellas son valores mayores o iguales a cero, indicando un valor 0 la ausencia de dispersión.

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 el número de items que un alumno ha resuelto correctos, f_i la frecuencia correspondiente.

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).

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

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

Actividad 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 ^{2} y se calcula mediante la fórmula (3.5).

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

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

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

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).

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

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

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

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

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

Actividad 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 intercuantílico 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.

Actividad 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.

Actividad 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?

Algunas dificultades en la comprensión de la idea de dispersión

El estudio de una distribución de frecuencias no puede reducirse al de sus promedios, ya que distribuciones con medias o medianas iguales pueden tener distintos grados de variabilidad. Para Campbell (1974) un error frecuente es ignorar la dispersión de los datos cuando se efectúan comparaciones entre dos o más muestras o poblaciones.

La desviación típica mide la intensidad con que los datos se desvían respecto de la media. Loosen y cols. (1985) hicieron notar que muchos libros de texto ponen mayor énfasis en la heterogeneidad entre las observaciones que en su desviación respecto de la posición central. Como señalan Loosen y cols., las palabras empleadas: variación, dispersión, diversidad, fluctuación, etc. están abierta a diferentes interpretaciones. Es claro para el profesor, pero no para el estudiante, cuándo estas palabras se refieren a una diversidad relativa a la media o en términos absolutos.

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 utilzar el coeficiente de asimetría de Pearson, que se representa por A_p y se calcula por la fórmula (3.12).

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

- *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 no 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, invariante por traslaciones y cambios de escalas, debido a las propiedades de la media, moda y desviación típica.

Actividad 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?

Actividad 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?

Actividad 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.

Actividad 3.28. ¿Qué tipo de forma piensas tienen 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 Medicina

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 *leptocurtica.* Si es aproximadamente igual de apuntada que la

normal se llama *mesocurtica.* Si es menos apuntada o más aplastada que la distribución normal se llama *platicurtica.*

Existe un coeficiente, ideado por Fisher, que mide el apuntamiento de una distribución y se llama coeficiente de curtosis. Se suele representar por g_2 y se verifica:

- Si $g_2 < 0$ la distribución es platicútica
- Si $g_2 = 0$ es mesocurtica
- Si $g_2 > 0$ 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	Hembras
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 raso). 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 intercuartilico. Como en nuestro caso el peso mínimo es 44 kilos, y el recorrido intercuartilico 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.

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 intercuartilico 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 6.

   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 intercuartilico, sólo aparecen un uno por ciento de los casos, por lo que estos valores, si no son debidos a errores, suelen ser casos excepcionales.

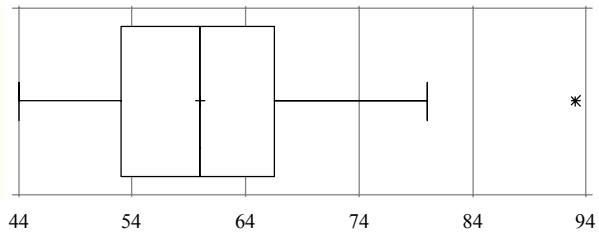


Figura 3.10. Gráfico de la caja para el peso de los alumnos

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

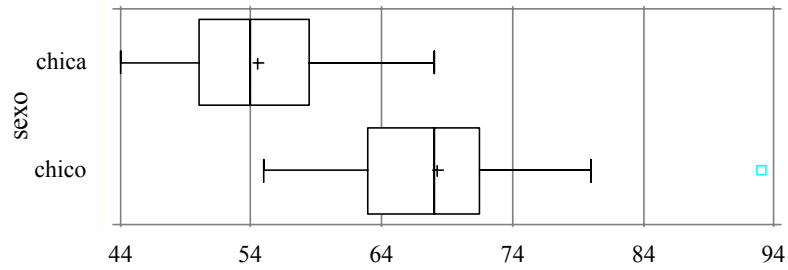


Figura 3.11Gráfico de la caja para los pesos de chicos y chicas

---

Actividad 3.29. La figura 3.12 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

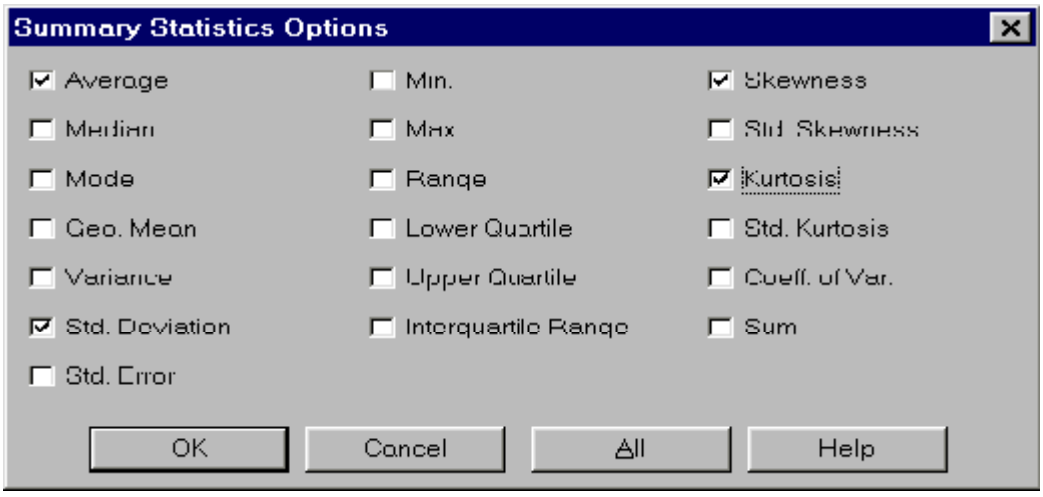
Cálculo de estadísticos con STATGRAPHICS:

Los estadísticos se pueden obtener de la opción de resumen numérico (SUMMARY STATISTICS), dentro de DESCRIBE, NUMERIC DATA- ONE VARIABLE ANALYSIS. En la figura 10 se muestran las medidas o parámetros que aparecen por defecto.

Para cambiar estas opciones, pulsar el botón derecho del ratón y seleccionar PANE OPTIONS, aparecerá un cuadro de diálogo como el de la figura 3.13. Sobre él seleccionar los parámetros o medidas que desean calcularse,

Figura 3.13. Estadísticos que proporciona SUMMARY STATISTICS por defecto

---



La traducción de estos estadísticos es la siguiente:

Average: Media	Min: Mïnimo	Kewness: Asimetría
Median: Mediana	Máx: Máximo	Std. Kewness: Asimetría estandarizado
Mode: Moda	Range: Recorrido	Kurtosis: Curtosis
Geo Mean: Media geométrica	Lower Quartile: Primer Cuartil	Std. Kurtosis: Curtoris estandarizado
Variance: Varianza	Upper Quartile: Tercer Cuartil	Coeff. of Var: Coeficiente de variación

Std. Deviation: Desviación típica	Interquartile Tange: Recorrido intercuartílico	Sum: Suma de los datos
Std. Error: Error de muestreo de la media. Es la desviación típica de todas las medias que pueden obtenerse en las diferentes muestras de la población		

Proyecto 3.1. 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 recojer 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

VARIABLES ESTADISTICAS BIDIMENSIONALES

4.1. DEPENDENCIA FUNCIONAL Y DEPENDENCIA ALEATORIA ENTRE VARIABLES

Generalmente, cuando se realiza un estudio estadístico, 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 4.1. Al estudiar la caída libre de un cuerpo, para el cual la Física ha encontrado que el espacio recorrido, Y, está relacionado con el tiempo desde su lanzamiento, X, por la expresión (4.1), siendo g la constante 9.8 m/s^2.

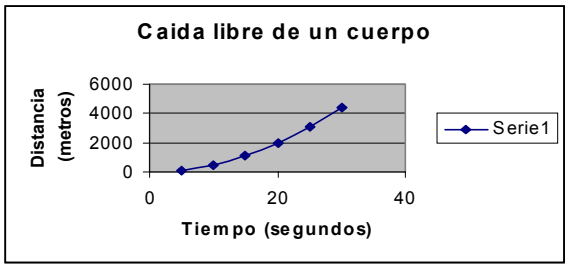
(4.1) $$Y = \tfrac{1}{2}\, g\, X^2$$

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., 10 seg.,... 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 4.1), por los cuales es posible trazar la parábola cuya ecuación es dada por la fórmula (4.1).

Tabla 4.1. 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 4.1: 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 4.2. En las figuras 4.2 y 4.3 hemos representado, a partir del fichero DEMOGRAFÍA, la esperanza de vida del hombre en función de la tasa de mortalidad y el PNB para cada país de la muestra.

Figura 4. 2

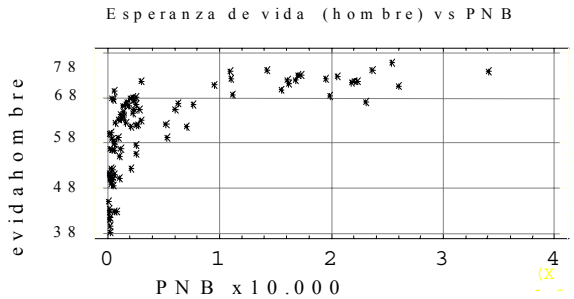
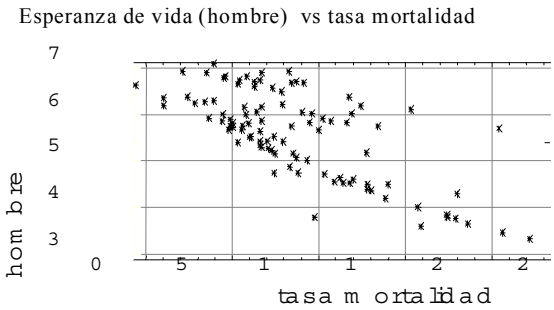


Figura 4. 3



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.

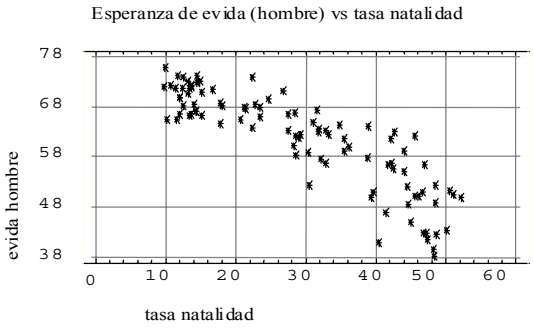


Figura 4.4

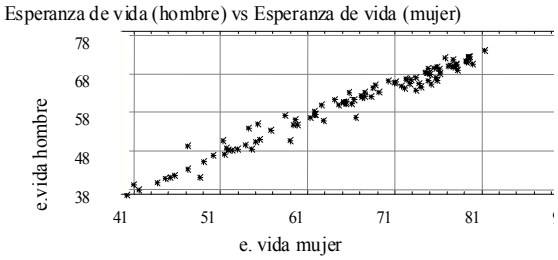


Figura 4.5

Actividad 4.1. En las Figuras 4.4 y 4.5 hemos representado la esperanza de vida del hombre en los países del fichero DEMOGRAFÍA 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 4.2 a 4.5 consideraría la relación entre la esperanza de vida del hombre y otra variable de tipo causal ?

---

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 4.2, o si una variable cuantitativa está relacionada con otra cualitativa como en la actividad 4.3.

Actividad 4.2. 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 4.2.

Tabla 4.2. 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

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.

Actividad 4.3. 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 4.3.

Tabla 4.3. 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

Utilizando los datos de la tabla estudia si la presión está relacionada con el momento en que se toma (antes o después del tratamiento).

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 4.2. 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 4.3. 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, que no estudiaremos por falta de tiempo.
- Dos variables cuantitativas como en los ejemplos 4.1 y 4.2. 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 4.1, 4.2 y 4.4 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 4.4 los puntos están mucho más cerca de la línea de regresión, porque la correlación entre las variables es más intensa. En las figuras 4.3 y 4.5 la relación sería inversa. Podría darse el caso de que no se observara ninguna relación entre las variables y hablaríamos de independencia.

3. DISTRIBUCIÓN CONJUNTA DE DOS VARIABLES ESTADÍSTICAS. 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 4.5 o bien, cuando todos o algunos pares se repiten, pueden escribirse como una *tabla de doble entrada* (o tabla de contingencia) (Figura 4.6), 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 (2). Llamamos a esta frecuencia relativa de cada celda respecto al total de datos *frecuencia relativa doble.*

(2) $$h_{ij}=f_{ij}/n$$

Figura 4.5

X	Y
x_1	y_1
x_2	y_2
.	.
x_i	y_i
x_n	y_n

Figura 4.6

	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 4.6), 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 $f_{.j}$ 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 4.3. Al clasificar una serie de modelos de automóviles por el número de cilindros y su origen se obtuvo la tabla 4.4.

Tabla 4.4. Distribución del número de cilindros en una muestra de automóviles según su 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

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 4.7 y 4.8.

Figura 4.7

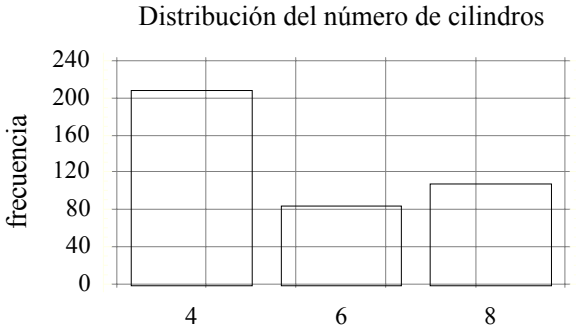
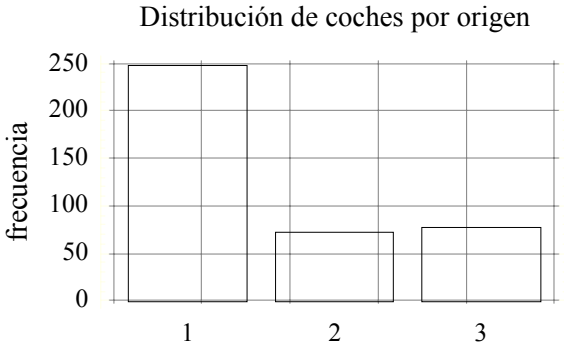


Figura 4.8



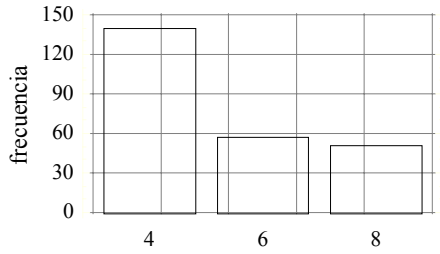
En particular, si X o Y son cuantitativas, podríamos calcular su media y varianza. Así, por ejemplo, el número medio de cilindros de todos los coches en el ejemplo 4.3 es 5.49 y su varianza 2.91.

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 4.4 podríamos analizar la distribución de coches europeos según el número de cilindros, y obtener la tabla 4.5 y Figura 4.9.

Tabla 4.5. Número de cilindros en coches europeos europeos

	N. cilindros			
	4	6	8	Total
Frecuencia	140	57	51	248
Porcentaje	56.45	22.98	20.56	

Figura 4.9. Número de cilindros en coches



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 (3).

(3) $$h(x_i \mid y_j) = \frac{f_{i,j}}{f_{.j}} = \frac{h_{i,j}}{h_{.j}}$$

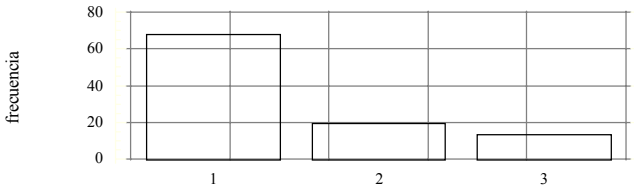
En el ejemplo anterior hemos hallado la distribución condicional de la variable Y en función de uno de los valores de X, es decir, la distribución condicional de filas en la tabla de contingencia, cuando sólo tomamos los datos de una de las columnas.

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 4.4, podríamos obtener la distribución del origen de los coches de 4 cilindros, obteniendo la tabla 4.6 y figura 4.10.

Tabla 4.6. Origen de los coches de 4 cilindros

Origen	Frecuencia	Porcentaje
Europa	140	67.63
E.U.	40	19.32
Japón	27	16.04
Total	207	

Figura 4.10. Origen de los coches de 4 cilindros



Podemos ahora obtener la frecuencia relativa de y_j condicionada por $x=x_i$ mediante la expresión (4):

(4) $$h(y_j \mid x_i) = \frac{f_{i,j}}{fi.} = \frac{h_{i,j}}{hi.}$$

Como consecuencia se verifica la igualdad (5) que nos permite obtener la frecuencia relativa doble a partir de las condicionales y marginales.

(5) $$h_{i,j} = h(x_i|y_j)\, h_{.j} = h(y_j|x_i)\, h_{i.}$$

Actividad 4.4. 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
VARON	6	9	38	49	38	17	14	13
HEMBRA	6	12	23	29	23	25	7	1

4. TABLAS DE CONTINGENCIA Y REPRESENTACIONES ASOCIADAS EN STATGRAPHICS

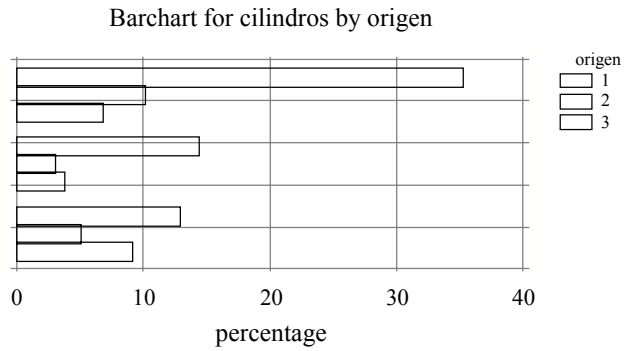
Con Statgraphics podemos realizar una tabla bidimensional para dos variables cualitativas. Para ello elegiremos el menú DESCRIBE, en la opción CATEGORICAL DATA - CROSSTABULATION. Aparecerá un cuadro de diálogo en el que se debe elegir la variable que irá en las filas (campo Row variable) y la variable que se tomará en las columnas (campo Column Variable). Por ejemplo, si elegimos como filas el número de

cilindros y columna el origen de los coches de la muestra analizada en el ejemplo 4.3, obtendremos la tabla 4.7.

Tabla 4.7. Frequency Table for cilindros by origen

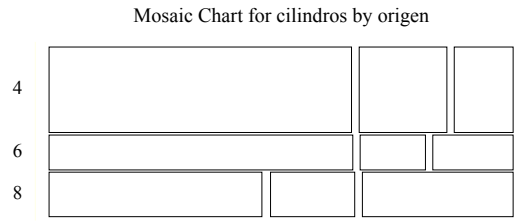
Row	1	2	3	Total
4	140	40	27	207
	35.18	10.05	6.78	52.01
6	57	12	15	84
	14.32	3.02	3.77	21.11
8	51	20	36	107
	12.81	5.03	9.05	26.88
Column	248	72	78	398
Total	62.31	18.09	19.60	100.00

Figura 4.11. 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).

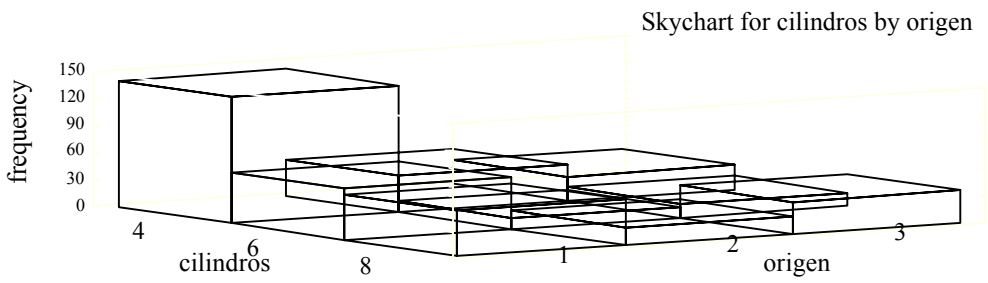
Figura 4. 12. Gráfico de mosaicos.



Como opciones gráficas tenemos el diagrama de barras adosado (Figura 4.11), 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 4.12) 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 *skychart* o diagrama tridimensional de barras, se representan en dos ejes las categorías de filas y columnas y en el eje Z las frecuencias dobles (Figura 4.13).

Figura 4.13. Diagrama de barras tridimensional



Actividad 4.5. Compara las tres representaciones gráficas obtenidas del programa CROSTABULATION en las figuras 4.10, 4.11 y 4.12. ¿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 PANE OPTIONS. 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 4.8, 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.

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 4.9 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.

Tabla 4.8. Distribuciones condicionadas por filas

	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
Column	248	72	78	398
Total	62.31	18.09	19.60	100.00

Tabla 4.9. Distribuciones condicionadas por columnas

	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
Column	248	72	78	398
Total	62.31	18.09	19.60	100.00

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.

Actividad 6. 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?

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 (6) para todo par de valores i, j.

(6) $$h(x_i|y_j)= h_{i.}$$

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 condiciono por un valor de Y.

En el caso de independencia, se cumplen, además, las propiedades (7) a (9). La propiedad (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.

(7) $$h_{i,j}= h_{i.}\, h_{.j}, \text{ para todo i, j}$$

(8) $$h(y_j|x_i)= h_{.j}, \text{ es decir Y no depende de X}$$

La propiedad (8) 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 (9) nos da un método de cálculo de las frecuencias teóricas en caso de

(9) $$f_{i,j} = \frac{f_{i.}\, f_{.j}}{n}$$

independencia

Actividad 4.7. 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?

---

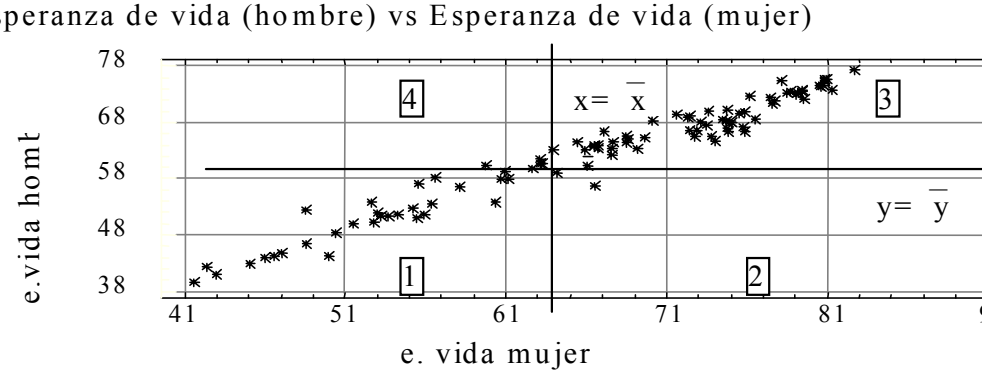
6. COVARIANZA Y CORRELACIÓN EN VARIABLES NUMÉRICAS

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).

(10) $$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

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 4.14).

Figura 4.14. División del plano en cuatro cuadrantes al trazar las rectas X= $\bar{x}$ e Y= $\bar{y}$



Supongamos que en un diagrama de dispersión, como el mostrado en la figura 4.13 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 4.1, 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.

Actividad 4.8. Razona por qué en caso de dependencia inversa entre variables numéricas el signo de la covarianza es negativo.

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 correlación lineal** o coeficiente de Pearson, que se define por la relación (11), siendo s_x , s_y las desviaciones típicas de las variables X e Y en la muestra analizada.

$$r = \frac{s_{xy}}{s_x s_y} \qquad (11)$$

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.

Actividad 4.9. 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.

Actividad 4.10 . Cuando la covarianza entre X e Y es mayor que cero, entonces (V/F):

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

Actividad 4.11. 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?

7. AJUSTE DE UNA LÍNEA DE REGRESIÓN A LOS DATOS

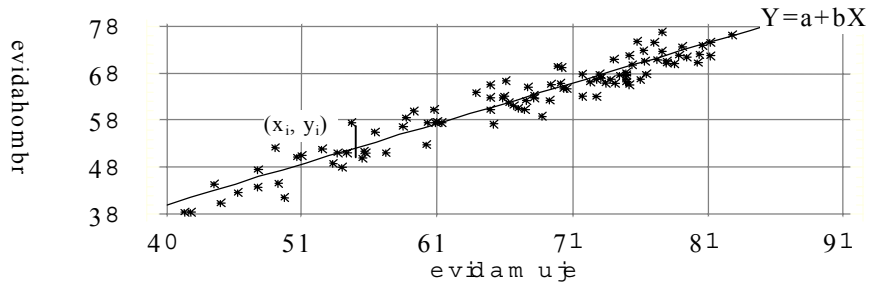
En el caso de que exista una correlación suficiente entre dos variables numéricas podemos plantearnos un nuevo problema que consiste en tratar de determinar la ecuación de una función matemática que nos permita predecir una de las variables (Y) cuando conocemos la otra variable (X). Esta función será 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 4.3, 4.4 y 4.5 la función que mejor se ajustaría a la nube de puntos sería una recta (creciente en la gráfica 4.5 y decreciente en las gráficas 4.3 y 4.4). Para la gráfica 4.2 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 4.15), 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 4.15. Desviaciones de los puntos a la recta de regresión



El procedimiento de obtención de las constantes será hacer mínima la cantidad D, dada por (12), siendo $f(x_i)=a+bx_i$, si lo que se trata de ajustar es una línea recta.

(12) $$D= \Sigma\ [y_i-f(x_i)]^2$$

(14) $$b=\frac{S_{xy}}{S_x^2};.....a=\bar{y}-b\bar{x}$$

Como consecuencia se obtienen las cantidades a y b que vienen dadas por (14), 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 4. 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 de (14), la ecuación de la recta de regresión de Y sobre X puede escribirse según la relación (15).

(15) $$y_i - \bar{y} = (x_i - \bar{x})\frac{S_{xy}}{S_x^2}$$

Como puede apreciarse, el par ($\bar{x}$, $\bar{y}$), satisface la ecuación (15); 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 (16) los papeles de las variables, obtenemos (16).

(16) $$x_i - \bar{x} = (y_i - \bar{y})\frac{S_{xy}}{S_x^2}$$

Esta recta es, en general, diferente de la dada en la ecuación (15), aunque también pasa por el punto ($\bar{x}$, $\bar{y}$).

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.

8. 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 RELATE- SIMPLE REGRESSION, 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 4.16 presentamos el resultado que se obtiene en SUMMARY STATISTICS 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

Figura 4.16. Resultados del Programa de Regresión Simple

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: evidahombr
Independent variable: evidamujer
------------------------------------------------------------------------------
                                 Standard          T
Parameter       Estimate          Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept        4.69411         1.11775         4.1996         0.0001
Slope           0.858511       0.0166701        51.4999         0.0000
------------------------------------------------------------------------------

                            Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-
Value
------------------------------------------------------------------------------
Model                       8569.86      1      8569.86    2652.24
0.0000
Residual                    306.962     95      3.23117
------------------------------------------------------------------------------
Total (Corr.)               8876.82     96

Correlation Coefficient = 0.982558
```

Este programa presenta una gran cantidad de información, pero nosotros sólo tendremos en cuenta la siguiente:

- Modelo ajustado: Se indica en la primera línea (Linear model: Y = a + b*X);
- Variables dependiente e independiente: (líneas segunda y tercera; Dependent variable: evidahombr;Independent variable: evidamujer);
- Parámetros del modelo (*a* es la intersección con el origen o "intercept"; en nuestro caso, a = 4.69411; *b* es la pendiente o "slope"; 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 Correlation Coefficient = 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-squared = 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.

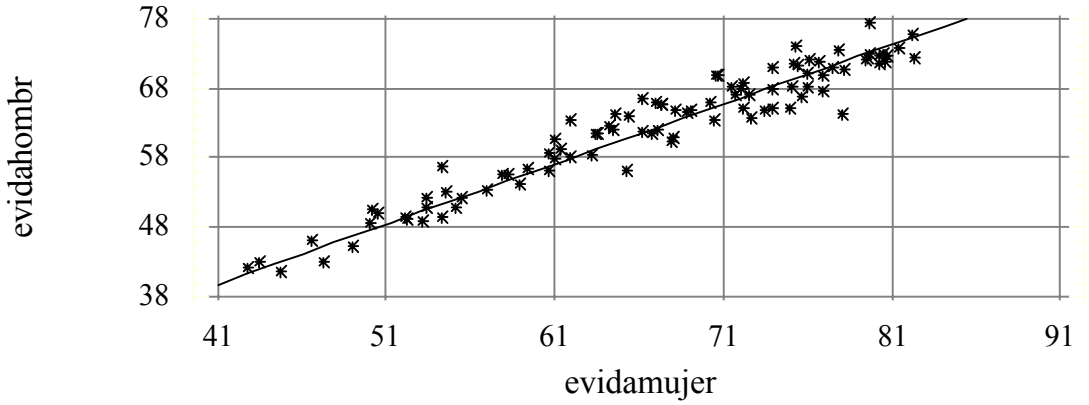
Nota importante. Que una variable quede explicada por otra no quiere decir que haya una relación de causa y efecto. En el ejemplo analizado tanto la esperanza de vida del hombre como la de la mujer tienen su causa en una serie de factores que afectan a las dos variables simultáneamente y se refieren 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 ANALYSIS OPTIONS 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 PLOT OF FITTED MODEL que dibuja la curva ajustada sobre la nube de puntos. Cambiando el tipo de función en ANALYSIS OPTIONS 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 4.17. Dibujo del modelo ajustado a la nube de puntos



Razonamiento espontáneo sobre las tablas de contingencia

La idea de asociación estadística extiende la de dependencia funcional, y es fundamental en muchos métodos estadísticos que permiten modelizar numerosos fenómenos en las diversas ciencias.

Como hemos visto, una tabla de contingencia o clasificación cruzada de dos variables sirve para presentar en forma resumida la distribución de frecuencias de una población o muestra, clasificada respecto a dos variables estadísticas.

En su forma más simple, cuando las variables poseen sólo dos categorías, toma la forma de la Tabla 4.16

Tabla 4.16: Formato típico de la tabla de contingencia 2x2

	A	no A	Total
B	a	b	a+b
no B	c	d	c+d
Total	a+c	b+d	a+b+c+d

Incluso la interpretación de las frecuencias reviste dificultad, ya que, a partir de la frecuencia absoluta de una celda, por ejemplo, la celda *a*, podemos obtener tres frecuencias relativas diferentes: la frecuencia relativa doble [a/(a+b+c+d)], la frecuencia relativa condicional respecto a su fila [a/(a+b)] y la frecuencia relativa condicional respecto a su columna [a/(a+c)].

La investigación sobre el razonamiento respecto asociación ha sido objeto de gran interés en psicología y ha estado ligada a los estudios sobre toma de decisiones en ambiente de incertidumbre, ya que la toma de decisiones precisa, generalmente, un juicio previo sobre la asociación entre variables. La mayor parte de estas investigaciones han empleado tablas 2x2, como la mostrada en el ejemplo siguiente:

Se quiere estudiar si el fumar produce trastornos respiratorios a un grupo de personas. Para ello se han observado durante un periodo suficiente de tiempo a 250 personas obteniendo los siguientes resultados:

	Molestias respiratorias	No tiene molestias	Total
Fuma	60	40	100
No fuma	90	60	150
Total	150	90	250

Utilizando los datos de la tabla, razona si en esta muestra, el padecer trastornos respiratorios depende o no de fumar.

Si analizamos con detalle la tarea presentada podemos observar que, a pesar de su aparente simplicidad, es para el alumno un problema complejo y su dificultad depende de ciertos datos del enunciado.

En el ejemplo dado, no aparece una asociación puesto que la proporción de personas con trastornos es la misma entre fumadores y no fumadores. No obstante, según los valores dados a las cuatro casillas de la tabla, puede aparecer asociación directa, inversa o independencia.

Otro hecho que complica esta tarea es que el número de ancianos en ambos grupos no es el mismo, esto es, que la distribución marginal de la variable (fumar o no) no tiene la misma frecuencia para sus diferentes valores.

El estudio del razonamiento sobre la asociación estadística fue iniciado por Piaget e Inhelder (1951), quienes consideraron que la comprensión de la idea de asociación implica las de proporción y probabilidad. Por esta razón, sólo estudiaron

este problema con chicos a partir de los 13-14 años. El contexto empleado fue estudiar la asociación entre el color de los ojos y el de los cabellos. Para ello emplean cartas con dibujos de rostros en los que los ojos y el cabello están coloreados, preguntando al sujeto si existe o no una relación entre el color de los ojos y el del cabello, no en forma general, sino cuando se consideran los únicos datos presentados.

Inhelder y Piaget encuentran que algunos sujetos analizan solamente la relación entre los casos favorables positivos (casilla a en la tabla 4.16) en relación a los casos totales (valor a+b+c+d en la tabla 4.16). En nuestro ejemplo, estos sujetos deducirían incorrectamente la existencia de una asociación directa entre las variables ya que el número de personas con trastornos que fuman es superior a cualquiera de las otras tres categorías.

En un nivel más avanzado de razonamiento los adolescentes comparan las casillas dos a dos. En la tabla 4.16, una vez admitido que también los casos (d) (ausencia-ausencia) son favorables a la existencia de asociación, no calculan la relación entre los casos que confirman la asociación (a+d) y el resto de los casos (b+c), lo que se produce sólo a partir de los 15 años según Piaget e Inhelder.

Estas mismas conclusiones son obtenidas en trabajos con estudiantes adultos. La mayor parte de los estudiantes adultos basan su juicio, bien en la casilla (a) o comparando (a) con (b), esto es, empleando sólo la distribución condicional de tener o no trastornos, en los que fuman. Con los datos del ejemplo, esta estrategia llevaría a concluir incorrectamente la existencia de una relación directa entre las variables, puesto que si nos restringimos a las personas que fuman, hay más con trastornos que sin ellos.

Como dificultad añadida al tema, Chapman y Chapman (1967) mostraron que hay expectativas y creencias sobre las relaciones entre variables que producen la impresión de contingencias empíricas. Este fenómeno ha sido llamado "correlación ilusoria", porque los sujetos mantienen sus creencias y sobreestiman la asociación cuando piensan que existe causación entre dos variables. Por ejemplo, en la figura 4.5 algunos sujetos defenderían que no puede haber correlación, a pesar de la forma de la gráfica, porque no hay relación causal directa entre lo que vive un hombre y lo que vive una mujer.

Todas estas dificultades influyen en la investigación e incluso en la toma de decisiones. Un ejemplo notable fue el caso del aceite de Colza, en el que se detectó un síndrome que, en principio fue denominado "neumonía atípica". La enfermedad, con síntomas parecidos a la neumonía se producía solo en familias completas y no parecía ser transmitida en la forma habitual. Se tardó bastante tiempo en descubrir que en realidad era una intoxicación producida por un aceite tóxico. Ello fue debido a que solo se tuvieron en cuenta los casos de personas que consumían el aceite y tenían la enfermedad y los que consumiéndola no la tenían (es decir, las casillas a y b en la Tabla 4.16). Pero no se comparó hasta bastante tarde la proporción de enfermos entre los consumidores del aceite con la de enfermos entre no consumidores, y esto hizo que en un principio no se detectase la asociación y se siguiesen más bien las creencias previas de que era un virus lo que causaba la enfermedad.

TEMA 5:

INTRODUCCIÓN A LA PROBABILIDAD

5.1. 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.

Actividad 5.1. *Fenómenos atmosféricos*:

a) 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.

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

El objetivo de la actividad 2.1. 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.

Actividad 5.2. Poner tres ejemplos de experimentos aleatorios y deterministas. Para cada uno de ellos describir un suceso simple y otro compuesto.

Suceso seguro e imposible.

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 ∅

Actividades

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

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

5.5. 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?

5.6. *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:

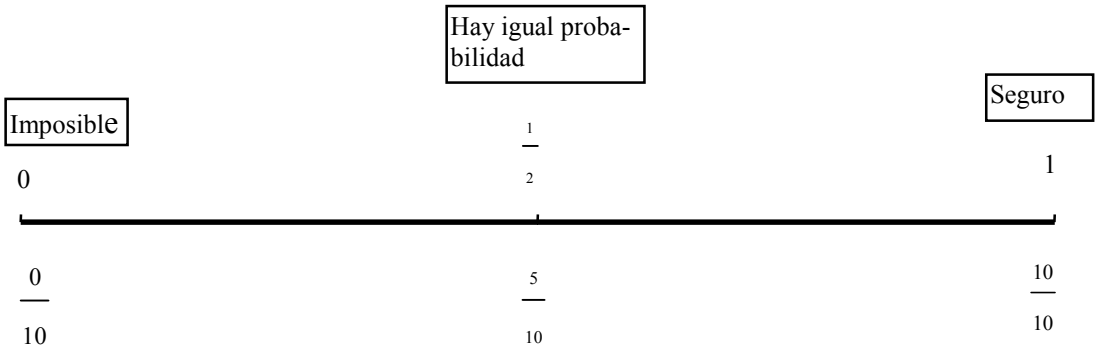


Figura 5.1

¿Podrías asignar un valor en la escala de probabilidad a las expresiones de la actividad 5.1. a)?

5.2. 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

5.7. *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.

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

5.9. *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.

5.10. *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).

5.11. 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?

5.12. 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?

5.3. ESTIMACIÓN DE PROBABILIDADES A PARTIR DE LAS FRECUENCIA RELATIVAS

Actividad 5.13. *Juegos de dados*

a) 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. Trata de adivinar cuantas veces, aproximadamente, saldrán dos números impares y cuantas uno par y otro impar si lanzas los dados 20 veces. Escribe este número en la columna "número esperado de veces".

¡Error! Marcador no definido. Resultado	Recuento	Frecuencia absoluta	Frecuencia relativa	N. esperado de veces
Dos números pares				
Dos números impares				
Un par y un impar				
Total		20	1	20

b) Lanza los dados 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?

Frecuencia absoluta y relativa. Estabilidad de las frecuencias 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).

Actividad 3.2: Ley del azar

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 24.000 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 100.000 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.

Actividad

5.14. *Construcción de dados*

Un dado ordinario se puede construir recortando en cartulina el siguiente perfil

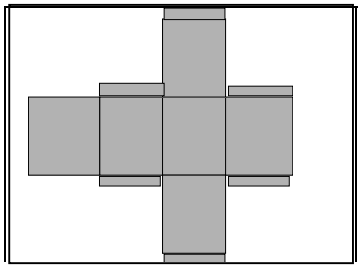


Figura 5.2

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?

5.15. *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.

5.16. *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. El conjunto de todos los sucesos elementales es, por tanto: $E = \{ R, B\}$
b) Si tiramos 40 veces, ¿alrededor de cuantas veces ocurrirá R?; ¿y B?
Haz este experimento con un compañero y completa la tabla siguiente:

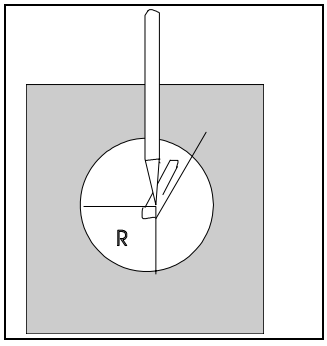


Figura 5.3

¡Error! Marcador no definido. Suceso	Recuento	Nº de veces obtenidos	Frecuencia relativa	Nº de veces esperado
R				
B				
Total		40	1	40

d) Asigna probabilidades a los sucesos R y B: P(R) = ; P(B) =
e) ¿Cuál sería la probabilidad de obtener dos veces consecutiva el rojo al girar la ruleta?

Simulación de experimentos aleatorios

La realización de experimentos aleatorios usando dispositivos físicos, como dados, fichas, bolas, ruletas, etc. puede requirir 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:

a) Experimentos complejos, como sería planificar el tráfico durante una operación salida de vacaciones.
b) Experimentos peligrosos, como estimar la temperatura de control o la velocidad de reacción permitida en una central nuclear.
c) Situaciones futuras: estudios ecológicos o sobre contaminación ambiental.

Actividades

5.16. Explicar cómo usar la tabla de números aleatorios de la figura 5.5, o los números aleatorios generados por tu calculadora, para simular los siguientes experimentos:

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.
c) Girar una ruleta como la de la figura 5.4.

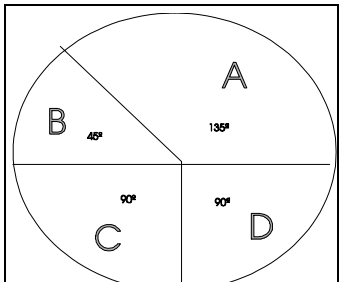


Figura 5.4

Figura 5.5: 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

5.17. 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.

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

Actividades

5.18. 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.

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

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.

Actividad 5.20. 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.

Axiomas de la probabilidad

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

a) 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.*
b) 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.
c) 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:

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

Actividades

5.21. 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?

5.23. 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?

5.24. 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".

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

5.5. VARIABLE ALEATORIA DISCRETA

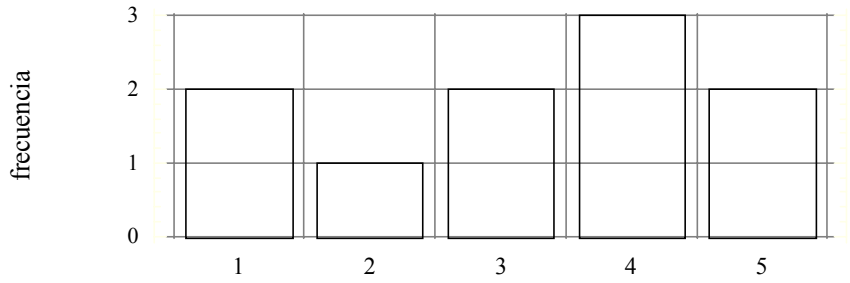
Si realizamos n pruebas o repeticiones de un experimento aleatorio, obtenemos un conjunto de n observaciones o resultados, que constituyen lo que se llama una muestra

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 5.5. 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**.

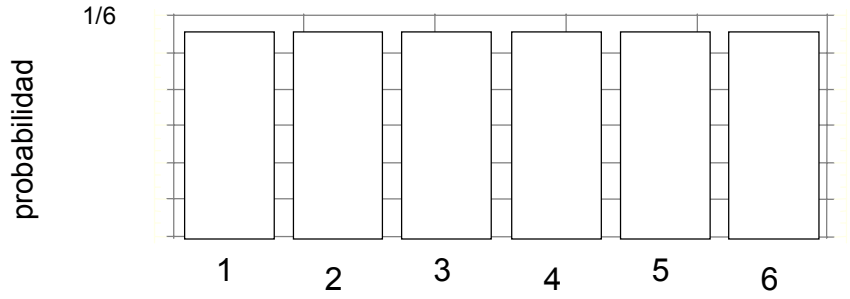
Figura 5.5. Variable estadística. Número de puntos obtenidos en 10 lanzamientos de un dado



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 que estamos considerando, 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 ξ, 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 5.6).

Figura 5.6. Variable aleatoria "número posible de puntos al lanzar un dado"



Actividad 5.26. 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, esto es:

Escribe en una tabla los valores posibles de esta variable y sus respectivas probabilidades

---

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 de 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.

5.6. DISTRIBUCION DE PROBABILIDAD DE UNA VARIABLE ALEATORIA DISCRETA

Una variable aleatoria discreta queda especificada por los valores que toma 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 tome el valor x_n, se suele representar por $P(\xi=x_n)$, donde ξ representa la variable aleatoria.

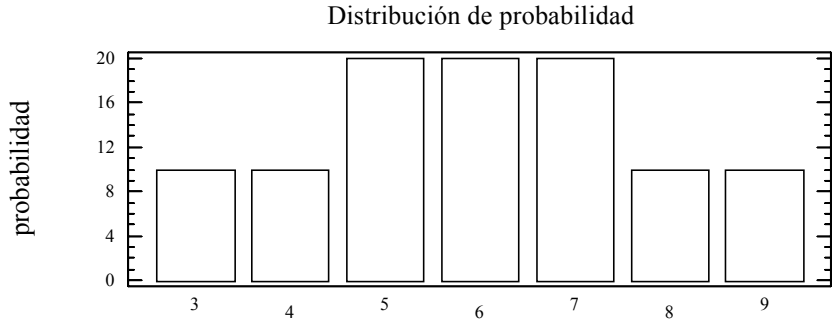
Ejemplo 5.1. Una urna contiene 5 fichas numeradas del 1 al 5. Sacamos sucesivamente dos fichas de la urna, sin reemplazamiento. Calculemos la distribución de probabilidad de la variable = "suma de los números en las fichas"

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	$P(\xi=x_i)$
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 5.7.

Figura 5.7.



Actividades

5.27. 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?

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

5.29. 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". ¿Cuál es la probabilidad de que el número de lanzamientos sea par?

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

$$\sum_{i=1}^{k} x_i p_i = \mu = E(\xi) \qquad (5.2)$$

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

$$E[g(\xi)] = \sum_{i=1}^{k} g(x_i)p_i \qquad (5.3)$$

Ejemplo 5.2. Pablo y María juegan a lanzar tres monedas. Si se obtienen 2 caras o 2 cruces, María paga a Pablo 100 pesetas. Si se obtienen 3 caras o tres cruces, Pablo paga 100 pesetas 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	(x_i)	$p(x_i)$	$x_i p(x_i)$
CCC XXX	100	2/8	200/8
CCX CXC XCC	100	6/8	600/8
XXC XCX CX			400/8=50

Del estudio de esta tabla deducimos que, si se jugase un gran número de veces el juego, en promedio, Pablo ganaría 50 Pts. cada jugada. No es por tanto un juego justo. Para lograr que el juego fuese equitativo, habría de pagarse a María 300 Pts., cada vez que se obtuviesen 3 caras o tres cruces. De esta forma:

$$E(\xi)=300*2/8-100*6/8=0$$

También podemos definir la *varianza de la variable:*

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

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 .

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

(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

Para esta variable obtenemos las siguientes características

$\mu = 27/16 = 1.6875$

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

$Me = 1.5$

Otras características de cálculo sencillo son: recorrido = 4;

Moda = 1 (valor más frecuente); $Q_{25} = 1$; $Q_{75} = 2$; recorrido intercuartílico = RI = 1.

Actividades

5.31. 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.

5.32. 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.

5.7. 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 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 , que puede tomar los valores enteros comprendidos entre 0 y n. Calcularemos la distribución y características de dicha variable aleatoria.

Ejemplo 5.4. Los tubos electrónicos producidos en una fábrica, pueden ser clasificados en correctos (suceso A) y defectuosos (suceso **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 AA**A** A**A**A **A**AA A**AA** **AA**A **AAA** **AAA**}

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

	ξ =x	P(ξ=x)
AAA	0	0.98^3
AAA, AAA, AAA	1	$3*0.98^2*0.02$
AAA, AAA, AAA	2	$3*0.98*0.02^2$
AAA	3	0.02^3

Supongamos ahora que en una repetición sucesiva de n ensayos independientes, hemos obtenido la sucesión AA**A**A**A**.....AA**A**A**A**, que contiene r veces el suceso A y n-r veces el suceso **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 (5.4).

(5.4) $P(\xi=r)= C_{n,r}\ p^r\ q^{n-r}$

Esta es la distribución de probabilidades binomial, cuyo nombre proviene del hecho de que las probabilidades dadas en la expresión (5.4) 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).

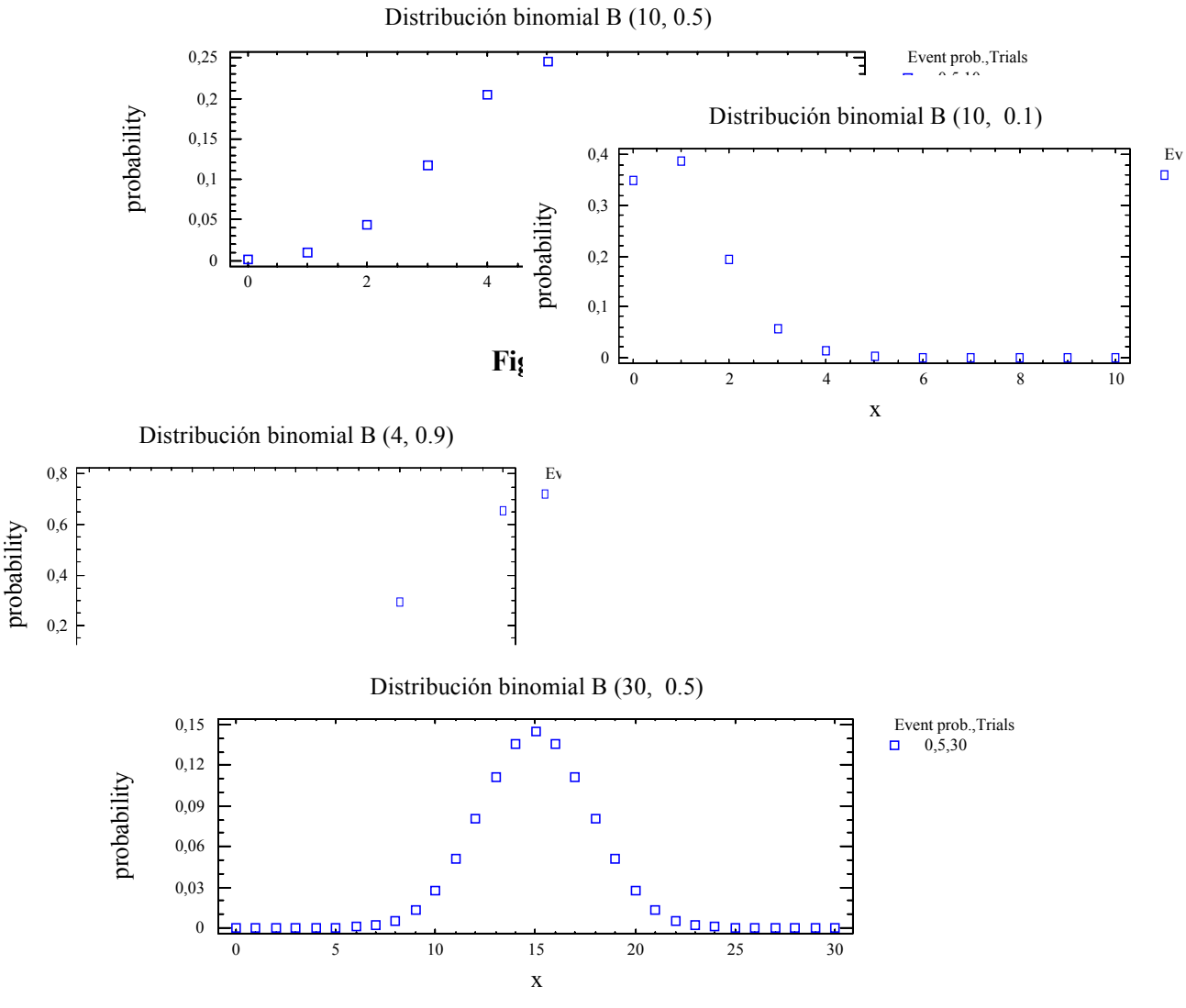
Puede demostrarse que la media y varianza de dicha variable aleatoria se calculan mediante (5.5) y (5.6).

(5.5) $\mu=np$

(5.6) $Var\ (\xi)=npq$

En la figura 5.4 se muestra la distribución de probabilidades y función de distribución binomial B(10,0.5). En la figura 5.5 se muestran gráficamente las distribuciones de las variable binomiales B(10,0.5), B(10,0.1), B(20,0.5) y B(4,0.9). Obsérvese como cambia la forma de la distribución con el valor de los parámetros. Para p=0.5 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.

Figura 5.4.



Actividades

5.33. 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?.

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

5.35. 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.

5.36. 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.

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

5.38. 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?

5.39. 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?

5.8. PROBABILIDAD Y ESTADÍSTICA EN LOS CURRÍCULOS DE PRIMARIA

Las razones por las que un tema cualquiera debe ser incluido en el currículo de la educación obligatoria pueden sintetizarse en las siguientes:

- Ser una parte de la educación general deseable para los futuros ciudadanos.
- Ser útil para la vida posterior, bien para el trabajo o para el tiempo libre.
- Ayudar al desarrollo personal.
- Ayudar a comprender los restantes temas del currículo, tanto de la educación obligatoria como posterior.
- Constituir una base para una especialización posterior en el mismo tema u otros relacionados.

Estas cinco razones son ampliamente cubiertas por las nociones estadísticas y probabilísticas, a las cuales nos referiremos conjuntamente con el nombre de nociones estocásticas. Además, la *probabilidad y la estadística pueden ser aplicadas a la realidad tan directamente como la aritmética elemental,* no siendo preciso el conocimiento de teorías físicas ni de técnicas matemáticas complicadas. Por sus muchas aplicaciones, la probabilidad y la estadística proporcionan una excelente oportunidad para mostrar a los estudiantes cómo matematizar, cómo aplicar la matemática para resolver problemas reales. En consecuencia, la enseñanza de las nociones estocásticas puede ser llevada a cabo mediante una *metodología heurística y activa,* a través del planteamiento de problemas concretos y la realización de experimentos reales o simulados.

Otro aspecto señalado por Fischbein es *el carácter exclusivamente determinista de los currículos actuales*, y la necesidad de mostrar al alumno una imagen más equilibrada de la realidad: "En el mundo contemporáneo, la educación científica no puede reducirse a una interpretación unívoca y determinista de los sucesos. Una cultura científica eficiente reclama

una educación en el pensamiento estadístico y probabilístico". Esta tendencia determinista de la enseñanza no es motivada por razones científicas. A pesar del carácter aproximado de las leyes del azar, desde el momento en que se conoce su grado de aproximación, es posible hacer predicciones, como ocurre con las restantes leyes experimentales, ya que ninguna magnitud se puede medir con una precisión absoluta.

A los argumentos que acabamos de exponer, podemos añadir algunos matices. En primer lugar, ¿qué niño de estas edades no practica juegos de azar en casa o con otros compañeros?. Los juegos como el parchís, la oca, etc., están fuertemente enraizados en la vida del niño. En consecuencia, nos parece conveniente utilizarlos con fines educativos. Por ejemplo, incluso un alumno de preescolar, lanzando una simple moneda al aire (una ficha, etc.) puede contar el número de veces que resulta cara o cruz, y esta actividad puede ser útil en el aprendizaje de los primeros conceptos numéricos, al mismo tiempo que el alumno toma contacto con un experimento aleatorio.

Por último, aportamos una nueva razón de tipo social a favor de tratar de educar la intuición probabilística de todo ciudadano en el período de enseñanza obligatoria. Se trata de hacerles conscientes de la naturaleza probabilística de los distintos juegos de azar (loterías, máquinas "tragaperras", bingos, etc. Con frecuencia estos juegos constituyen magníficos negocios para sus promotores, pero usados desproporcionadamente por el ciudadano puede no ser una mera actividad lúdica, sino un riesgo desproporcionado de perder su dinero.

Creemos que las razones expuestas son suficientes para concluir que es preciso incorporar en los currículos de la enseñanza un objetivo referente al razonamiento estocástico a partir de los niveles educativos en que esto sea posible.

La Estadística y Probabilidad en los currículos

Decreto de Educación Primaria (Junta de Andalucía, BOJA, 20-6-92)

El Decreto de Educación Primaria de la Junta de Andalucía hace una tímida mención al tratamiento de las situaciones aleatorias dentro del bloque de contenidos denominado "Operaciones". Concretamente dice: "En casos sencillos se pondrá a los alumnos en situaciones de exploración de la noción de casualidad, pretendiendo el descubrimiento del carácter aleatorio de algunas experiencias y la representación sencilla del grado de probabilidad de un suceso experimentado" (Pág. 110)

Respecto al tratamiento de información estadística, el objetivo general número 6 hace una mención explícita de la misma en los siguientes términos: "6. Utilizar técnicas elementales de recogida de datos para obtener información sobre fenómenos y situaciones de su entorno; representarla de forma gráfica y numérica y formarse un juicio sobre la misma.

La recogida, organización y presentación de datos así como la interpretación y las posibles predicciones basadas en los mismos, son conocimientos que tienen cada vez más importancia en nuestro medio lo que hace deseable su aprendizaje y utilización.

Ha de considerarse que las sencillas actividades estadísticas pueden representar para los alumnos de estas edades aplicaciones de las matemáticas al medio real, prestando significado al mismo, haciéndolo más inteligible. Al mismo tiempo representan ocasiones para la exploración matemática ya que implican la formulación de preguntas, conjeturas, la búsqueda de relaciones, la toma de decisiones sobre qué información hace falta y cómo obtenerla, etc." (Pág.. 104)

Decreto de enseñanzas mínimas para la Educación Primaria (M.E.C., BOE, 26-6-91)

El objetivo general 6 para el área de Matemáticas formulado por el M.E.C. dice: "Utilizar técnicas elementales de recogida de datos para obtener información sobre fenómenos y situaciones de su entorno; representarla de forma gráfica y numérica y formarse un juicio sobre la misma". Este objetivo es desarrollado en el bloque de contenidos referido a organización de la información en los siguientes términos:

Conceptos:

1. La representación gráfica
2. Las tablas de datos.
3. Tipos de gráficos estadísticos: bloques de barras, diagramas lineales, etc.
4. Carácter aleatorio de algunas experiencias.

Procedimientos:

1. Exploración sistemática, descripción verbal e interpretación de los elementos significativos de gráficos sencillos relativos a fenómenos familiares.
2. Recogida y registro de datos sobre objetos, fenómenos y situaciones familiares utilizando técnicas elementales de encuesta, observación y medición.
3. Elaboración de gráficos estadísticos con datos poco numerosos relativos a situaciones familiares.
4. Expresión sencilla del grado de probabilidad de un suceso experimentado por el alumno.

Actitudes:

1. Actitud crítica ante las informaciones y mensajes transmitidos de forma gráfica y tendencia a explorar todos los elementos significativos.
2. Valoración de la expresividad del lenguaje gráfico como forma de representar muchos datos.
3. Sensibilidad y gusto por las cualidades estéticas de los gráficos observados o elaborados.

Respecto a criterios de evaluación sobre los contenidos estocásticos el M.E.C. especifica:

10. Realizar, leer e interpretar representaciones gráficas de un conjunto de datos relativos al entorno inmediato.
11. Hacer estimaciones basadas en la experiencia sobre el resultado de juegos de azar sencillos, y comprobar dicho resultado.

Estándares curriculares y de evaluación para la educación matemática (N.C.T.M.; USA)

Un documento curricular de interés para los profesores es el elaborado por la prestigiosa asociación de profesores de matemáticas de EE.UU., National Council of Teachers of Mathematics, conocido como "Estándares curriculares y de evaluación para la educación matemática", del cual existe traducción al castellano realizada por la Sociedad de Profesores de Matemáticas de Andalucía "Thales".

Para el nivel P-4 (Preescolar a 9 años) propone que el currículo incluya experiencias con análisis de datos y probabilidades para que los alumnos sean capaces de -

- recoger, organizar y describir datos;
- construir, leer e interpretar datos presentados de manera organizada;
- formular y resolver problemas que impliquen la recogida y análisis de datos;
- explorar el concepto de casualidad.

En los niveles 5-8 (que corresponden a los dos últimos cursos de enseñanza primaria en España y los dos primeros de enseñanza secundaria obligatoria) el currículo de matemáticas debe incluir la exploración de la estadística en situaciones del mundo real para que el estudiante sea capaz de:

- recoger, organizar y analizar datos de forma sistemática;
- elaborar, leer e interpretar tablas y diversas representaciones gráficas;
- formular inferencias y argumentos convincentes que se basen en el análisis de datos;
- evaluar argumentos que estén basados en el análisis de datos;
- llegar a apreciar los métodos estadísticos como medios potentes en la toma de decisiones.

En el "Estándar" sobre 'probabilidad' de estos niveles indica que el currículo de matemática debe incluir la exploración de la probabilidad en el mundo real para que los estudiantes sean capaces de:

- elaborar modelos de situaciones diseñando y llevando a cabo experimentos o simulaciones para estimar probabilidades;

- elaborar modelos de situaciones construyendo un espacio muestral para determinar probabilidades;
- apreciar las posibilidades de usar un modelo de probabilidad comparando los resultados experimentales con soluciones matemáticas esperadas;
- realizar predicciones que se basen en probabilidades experimentales o teóricas;
- llegar a reconocer el uso constante que se hace de la probabilidad en el mundo real.

1.3. Evaluación inicial de nociones estocásticas

Un punto esencial en el proceso educativo es el conocimiento, por parte del profesor, de las ideas previas de los alumnos sobre las nociones que trata de enseñarles. Las siguientes preguntas han sido tomadas de diferentes investigaciones didácticas que han estudiado las intuiciones y dificultades iniciales de los alumnos sobre la probabilidad y estadística. Para cada una de estas preguntas, indica las posibles respuestas de los alumnos y cuáles de ellas serían correctas o incorrectas, analizando las posibles causas de sus errores.

Prueba de evaluación de nociones estocásticas

1. Se introducen las siguientes cantidades de canicas rojas y azules en dos cajas:

Caja	Rojas	Azules
A	6	4
B	60	40

Cada caja se agita fuertemente. Después de elegir una caja, metes la mano y, sin mirar, sacas una canica. Si la canica es azul, ganas 500 Ptas. ¿Qué caja te da mejores posibilidades de extraer una bola azul?

2. ¿Cuál de las siguientes sucesiones es más probable que resulte al lanzar una moneda equilibrada 5 veces? ¿Cuál es la menos probable? ¿Por qué?

_____a. CCC++
_____b. +CC+C
_____c. +C+++
_____d. C+C+C
_____e. Las cuatro sucesiones son igualmente probables

3. Cinco caras de un dado equilibrado se pintan de negro y una se pinta de blanco. Se lanza el dado seis veces. ¿Cuál de los siguientes resultados es más probable?

____a. Cara negra en cinco lanzamientos y cara blanca en el otro.
____b. Cara negra en los seis lanzamientos.
____c: Igual probabilidad.

4 Cuando lanzamos simultáneamente dos dados es posible que ocurra uno de los dos resultados siguientes:
Resultado 1: Se obtiene un 5 y un 6
Resultado 2: Se obtiene el cinco dos veces.
¿Son estos resultados igualmente probables?

5. La mitad de todos los recién nacidos son niñas y la otra mitad niños. El hospital A registra un promedio de 50 nacimientos al día y el hospital B un promedio de 10 nacimientos al día. En un día particular, ¿cuál de los dos hospitales es más probable que registre un 80% o más nacimientos de niñas?

____a. El hospital A (50 nacimientos al día)
____b. El hospital B (10 nacimientos al día)
____c. Los dos hospitales tienen igual posibilidad de registrar este suceso.

6. En una lotería semanal se colocan en un recipiente los números 1 a 36. Se toman seis de estos números al azar sin reemplazamiento. Para ganar, un jugador debe predecir correctamente estos seis números.¿Son tus posibilidades de ganar la lotería mayores si juegas a los mismos números una semana tras otra, o si cambias los números cada semana? ¿Qué piensas de ello?

7. Un centro meteorológico quiso determinar la precisión de su meteorólogo. Buscaron los datos de aquellos días en los que el meteorólogo había informado que había un 70% de posibilidades de lluvia. Compararon estas predicciones con los registros que indicaban si llovió o no en esos días en particular.
La predicción del 70% de posibilidades de lluvia puede considerarse muy precisa, si llovió:
____a. Entre el 95% y el 100% de esos días
____b. Entre el 85% y el 94% de esos días
____c. Entre el 75% y el 84% de esos días
____d. Entre el 65% y el 74% de esos días
____e. Entre el 55% y el 64% de esos días

8. Una profesora quiere cambiar la colocación de los niños en su clase, con la intención de aumentar su participación. Primero, estudió el número de preguntas que hicieron los niños durante una semana con la colocación actual, obteniendo la tabla siguiente:

	Iniciales de los alumnos							
	A.A	R.F.	A.G.	J.G.	C.P.	J.L.	A.V.	C.B.
Nº de preguntas	0	5	3	22	3	2	1	2

La profesora quiere resumir estos datos, calculando el número típico de preguntas hechas esa semana. ¿Cuál de los siguientes métodos le recomendarías que usara?
_____a. Usar el número más frecuente, que es el 2.
_____b. Sumar los 8 números y dividir por 8.
_____c. Descartar el 22, sumar los otros 7 números y dividir por 7.
_____d. Descatar el 0, sumar los otros 7 números y dividir por 7

9. Se está experimentando una nueva medicina para determinar su efectividad en el tratamiento del eczema, una enfermedad inflamatoria de la piel. Se seleccionan treinta pacientes con eczema para participar en el estudio. Los pacientes se dividen al azar en dos grupos. Veinte pacientes en un grupo experimental recibieron la medicina, mientras que diez pacientes en un grupo de control no recibieron medicación. Abajo mostramos los resultados transcurridos dos meses.

Grupo experimental (medicación)
.....Mejoraron 8
..... No mejoraron 12

Grupo control (No medicación)
..... Mejoraron 2
..... No mejoraron 8

Basándote en estos datos, crees que el medicamento era:
_____1. algo efectivo _____2. básicamente no efectivo

Si escogiste la opción 1, selecciona la explicación que mejor describa tu razonamiento.

_____a. El 40% de las personas del grupo experimental (8/20) mejoró.
_____b. 8 personas mejoraron en el grupo experimental, mientras que sólo 2 mejoraron en el grupo control.
_____c. En el grupo experimental, el número de personas que mejoró es sólo 4 menos que el número de las que no mejoraron (12-8), mientras que en el grupo de control la diferencia es 6 (8-2).
_____d. El 40% de los pacientes en el grupo experimental mejoró (8/20), mientras que sólo el 20% mejoró en el grupo control (2/10).

Si escogiste la opción 2, selecciona la explicación que mejor describa tu razonamiento.

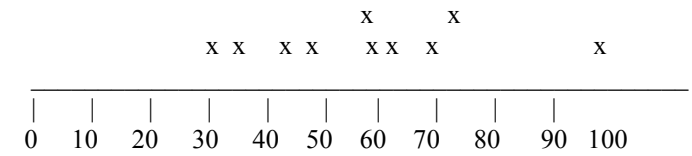
_____a. En el grupo de control, 2 personas mejoraron, incluso sin medicación.
_____b. En el grupo experimental hubo más gente que no mejoró que la que sí lo hizo (12 frente a 8).
_____c. Las diferencias entre los números de los que mejoraron y los que no mejoraron es casi la misma en los dos grupos (4 frente a 6).
_____d. En el grupo experimental sólo mejoró el 40% de los pacientes (8/20).

10. Abajo listamos varias posibles razones que podrían cuestionar los resultados del experimento que hemos descrito antes. Coloca una marca en cada una de las razones con las que estés de acuerdo.

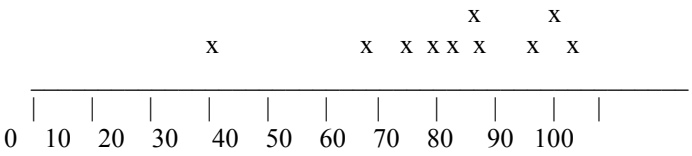
_____a. No es correcto comparar los dos grupos, porque hay diferente número de pacientes en cada grupo.

_____b. La muestra de 30 es demasiado pequeña para permitir obtener conclusiones.
_____c. Los pacientes no deberían haber sido asignados aleatoriamente a los grupos, porque los casos más graves podrían por azar haber coincidido en uno de los grupos.
_____d. No tengo bastante información acerca de cómo decidieron los médicos si un paciente mejoraba o no. Los médicos podrían haber estado sesgados en su juicio.

11. Veinte estudiantes universitarios participaron en un estudio sobre el efecto del sueño en las calificaciones en los exámenes. Diez de los estudiantes voluntariamente estuvieron despiertos estudiando toda la noche anterior al examen (grupo que no durmió). Los otros 10 estudiantes (el grupo control) se acostaron a las 11 la noche anterior al examen. Las puntuaciones en el examen se muestran en los gráficos siguientes. Cada X representa la puntuación de un estudiante particular. Por ejemplo, las dos Xs encima del número 80 en el gráfico inferior indican que dos estudiantes en el grupo control tuvieron una puntuación de 80 en el examen.



Puntuaciones en el examen del grupo que no durmió



Puntuaciones en el examen del grupo de control

Examina los dos gráficos con cuidado. Luego escoge entre las 6 posibles conclusiones que se listan a continuación aquella con la que estés más de acuerdo.
_____a. El grupo que no durmió lo hizo mejor porque ninguno de estos estudiantes puntuó por debajo de 40 y la máxima puntuación fue obtenida por un estudiante de ese grupo.
_____b. El grupo que no durmió lo hizo mejor porque su promedio parece ser un poco más alto que el promedio del grupo control.
_____c. No hay diferencia entre los dos grupos, porque hay un solapamiento considerable en las puntuaciones de los dos grupos.
_____d. No hay diferencia entre los dos grupos, porque la diferencia entre sus promedios es pequeña, comparada con la cantidad de variación de sus puntuaciones.
_____e. El grupo control lo hizo mejor porque hubo en ese grupo más estudiantes que puntuaron 80 o por encima.
_____f. El grupo control lo hizo mejor, porque su promedio parece ser un poco mayor que el promedio del grupo que no durmió.

12. El comité escolar de una pequeña ciudad quiere determinar el número promedio de niños por familia en su ciudad. Dividen el número total de niños de la ciudad por 50, que es el número total de familias. ¿Cuál de las siguientes frases debe ser cierta si el número promedio de niños por familia es 2.2?
____a. La mitad de las familias de la ciudad tienen más de 2 niños.
____b Más familias tienen 3 niños que 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 en una familia es 2.

13. Una empresa consultora informó que el 58% de una muestra aleatoria de adultos aprueban la actuación del Presidente del Gobierno. El informe dice que el margen de error de la encuesta es el 3%. ¿Qué significa este margen de error?

5.9. CONCEPTOS DE PROBABILIDAD

Como cualquier otro vocablo importante, la probabilidad tiene muchos matices de significación y admite variedad de usos. Un estudio de los términos utilizados en el lenguaje ordinario, a través de los "diccionarios de uso" revela que el azar y la incertidumbre se aprecia como cualidades graduables. Entre lo cierto o lo seguro (lo que ocurrirá necesariamente o lo que es verdadero sin ninguna duda) y lo imposible (lo que no puede ocurrir nunca) está lo probable, término que define (M. Moliner (1983):

"se dice de lo que, en opinión del que habla, es más fácil que ocurra que que deje de ocurrir"

Para expresar estas tres circunstancias (imposible, probable, seguro) existen una gran variedad de términos. Así, por ejemplo, un suceso que es probable ("es probable que llueva") se puede expresar con los adjetivos:

POSIBLE: "es posible que llueva"
PREVISIBLE: "es previsible que mañana haga frio"
PRESUMIBLE: "es presumible que apruebe el examen"
FACTIBLE: "es factible que termine a tiempo"
VIABLE: "es viable que ocurra"

Estos términos funcionan en el lenguaje ordinario como operadores modales, esto es, podemos afirmar un cierto enunciado rotundamente, comprometiéndonos categóricamente con su verdad, o podemos afirmarlo gradualmente. Los enunciados:

"lloverá mañana"

"Probablemente lloverá mañana"

describen la misma realidad. La diferencia estriba en el modo de afirmación: el primero es categórico, incondicional y el segundo es gradual y cauteloso.

El término probabilidad lo define M. Moliner como "Cualidad de probable o circunstancia de ser probable una cosa:

“La probabilidad de su hallazgo es cada vez menor”

“Hay probabilidad (probabilidades) de encontrarla”

La mayor o menor probabilidad de ocurrencia de un suceso puede graduarse mediante adverbios de cantidad o número:

'Hay algo de probabilidad de que se marche”; "algo" puede ser sustituido por: alguna, muchas, pocas, grandes, ..)

Otras veces, especialmente en el contexto de apuestas, se estima la probabilidad mediante la comparación de posibilidades a favor y en contra de un resultado:

”El caballo X tiene tres posibilidades contra una de resultar ganador”;

“Las apuestas son cinco a uno a favor del equipo X”.

También puede la probabilidad ser interpretada como "propiedad" de la persona o cosa a que afecta:

"Tiene algunas probabilidades de colocarse"

"La bala tiene muchas probabilidades de dar en el blanco"

La incertidumbre no sólo afecta a la ocurrencia de sucesos, sino que también puede afectar a la veracidad de las proposiciones o leyes. En castellano la palabra **verosímil**, "que tiene apariencia de verdadero", se utiliza especialmente con dicha finalidad, aunque también se usa **probable** en dicho contexto lógico:

"Es probable que lo que dice sea verdad"

"Lo que dice es verosímil"

La variedad y riqueza de términos que puede encontrarse en el diccionario para expresar lo incierto o verosímil es exponente de la amplitud de contextos, situaciones y matices en que estas características se presentan, y al mismo tiempo de la necesidad de proceder a un análisis filosófico y matemático del problema. Los usos formales del término probabilidad en el campo de la ciencia y la filosofía - construcción de modelos para los fenómenos aleatorios, el diseño de procedimientos de inferencia y toma de decisiones, etc - han llevado a definir, de un modo cuantitativo y preciso, la noción de probabilidad.

Estos esfuerzos no han cristalizado, sin embargo, en una única teoría filosófica, sino que han conducido a la formulación de distintas puntos de vista sobre la naturaleza de la probabilidad, que se describen a continuación.

Teoria clasica: Laplace

El primer intento de definir con rigor matemático la noción de probabilidad es debido a Laplace. En su obra "Teoríe analytique des probabilités" (1812), Laplace dio la definición que se conoce como clásica de probabilidad de un suceso que puede ocurrir sólamente en un número finito de modalidades como *la proporción del número de casos favorables al número de casos posibles, siempre que todos los resultados sean igualmente "probables".*

De acuerdo con esta definición, el cálculo de la probabilidad de los sucesos se reducía a problemas de análisis combinatorio. Pero incluso en su misma época, esta definición se encontró inadecuada. Además de ser circular y restrictiva, no dío respuesta a la pregunta de lo que realmente es la probabilidad; sólo proporcionó un método práctico de cálculo de probabilidades de algunos sucesos sencillos. Laplace, siguiendo a Bernouilli (1713) usó el principio de razón insuficiente, que considera las alternativas como equiprobables en la ausencia de razón conocida para esperar lo contrario. Más recientemente, para justificar la asignación de probabilidades por la regla de Laplace ha sido formulado el principio de indiferencia, que considera las alternativas como equiprobables cuando hay un balance de evidencia a favor de cada alternativa.

La definición de Laplace supone que siempre es posible seleccionar, como espacio muestral asociado a un experimento aleatorio, un conjunto de sucesos elementales que satisfacen las condiciones de simetría que garanticen la equiprobabilidad. Pero la aplicación del principio de indiferencia no es satisfactoria en general, ya que la evidencia nunca es perfectamente simétrica con respecto a un número de alternativas. Es inútil en los casos numerosos en que las posibilidades a analizar no pueden inventariarse en un conjunto de alternativas simétricas. En todo caso no es apropiado cuando realmente se carece de razones a favor de cada resultado ("paridad de ignoracia") o cuando la variable cuyo valor tiene que determinarse es continua.

La aproximación escolar tradicional hacia la probabilidad es teórica y "a priori", basada sobre la noción de sucesos equiprobables. Se dice a los niños que la probabilidad de obtener "un uno" o "un cinco" en una tirada de un dado es 1/6. De hecho esto entra en conflicto con la experiencia que puedan tener jugando, por ejemplo, al parchís, cuando a veces han debido de esperar bastante tiempo para poder comenzar a mover fichas porque no les salía el "cinco" requerido. Como su experiencia es limitada, pueden tener la impresión de que obtener "un cinco" es más difícil que obtener otros números.

Probabilidad frecuencial o empirica

En este punto de vista se considera que la probabilidad se calcula a partir de las frecuencias relativas observadas de cada uno de los diferentes resultados en pruebas repetidas. El principal elemento en este enfoque es que el concepto de probabilidad debe ser "objetivo", separado de cualquier consideración de factores personales y sujeto a demostración práctica a través de la experimentación.

La teoría frecuencial ha sido defendida en los tiempos modernos especialmente por Richard von Mises ("Probability, Statistics and Truth; 1919), aunque ya en 1888 John Venn en "The logic of chance" defendió explícitamente el cálculo de la probabilidad a partir de las frecuencias relativas de ocurrencias de sucesos. También son partidarios de este enfoque Hans Reichenbach y Kolmogorov.

El enfoque frecuencial descansa en dos características observables del comportamiento de los resultados de las realizaciones repetidas. En primer lugar, es un hecho que los resultados varían de una repetición a otra de una manera imprevisible. Esto es lo que significa el término

"variación aleatoria". En segundo lugar, se observa cómo un hecho empírico a corto plazo puede ser desordenado, pero a la larga surge una cierta regularidad. Esta pauta se demuestra de la siguiente forma. Supongamos un suceso particular A que nos interesa; tomamos observaciones repetidas anotando las ocasiones en que ocurre A; entonces la razón entre el número de veces que sucede A, n_A, y el número total de repeticiones n (razón frecuencial o frecuencia relativa de que A ocurra n_A/n) parece tender a un límite cuando n tiende a infinito. En esta aproximación, la idea de la probabilidad surge como el valor hacia el cual tiende la frecuencia relativa de una secuencia de resultados.

Aunque el planteamiento frecuencial atrae a los estadísticos profesionales y, en general, es útil siempre que se manejan grandes cantidades de datos (mecánica estadística, seguros, etc.), tiene inconvenientes desde los puntos de vista filosófico, conceptual y práctico relacionados con la noción de número infinito de experimentos. No se puede evaluar una probabilidad con precisión, porque el número de ensayos es siempre limitado. Además, existen situaciones donde no es posible conducir ensayos repetidos bajo condiciones experimentales fijas. Incluso con el lanzamiento de un dado es difícil estar razonablemente seguro de que no está sesgado examinando y realizando, por ejemplo, 1000 ensayos. Sin embargo, para la enseñanza elemental el enfoque frecuencial es muy adecuado en la asignación de probabilidades de fenómenos, como la distribución de sexos, que tienen una fuerte evidencia experimental. Además, la introducción del ordenador en el aula nos permite abordar en la escuela la probabilidad desde este punto de vista, ya que no resulta costoso, en esfuerzo ni en tiempo, simular un gran número de lanzamientos de un dado, por ejemplo, u otros experimentos similares.

Probabilidad subjetiva

En esta aproximación, la probabilidad es una expresión de la creencia o percepción personal. Este punto de vista mantiene que la probabilidad mide la confianza que un individuo particular tiene sobre la verdad de una proposición particular y, por tanto, no está unívocamente determinada. Este concepto no descansa en la repetibilidad de ningún proceso, por lo que se puede evaluar la probabilidad de un suceso que puede ocurrir una sola una vez, como por ejemplo, la probabilidad de que se descubra un medicamento que cure el cancer en el próximo año. Los subjetivistas consideran que se trata de un grado de creencia "personal" que un individuo sostiene sobre la base de su propia experiencia. Diferentes personas pueden asignar probabilidades distintas para un mismo suceso.

Aunque esta interpretación fue presentada por primera vez en 1926 por F.P. Ramsey y defendida en 1937 por B. de Finetti, ha sido L.J. Savage (1954) quien en los primeros años de la década de los 50 le ha dado un ímpetu considerable.

En esta concepción, a cualquier entidad aleatoria se puede atribuir una probabilidad. Esta puede ser asignada de cualquier modo, pero con la condición de que uno esté preparado para aceptar apuestas basadas en dicha asignación. Por ejemplo, si una persona cree que para un cierto dado la probabilidad de obtener un "uno" es 0.5, debe estar preparada para pagar 100 Pts. si el resultado de una tirada no es un "uno" y para ganar 100 Pts. si resulta un "uno", es decir para esta persona hay tantas posibilidades a favor como en contra del número uno.

Otro criterio que De Finetti (1974) postula es la condición de coherencia. En el ejemplo anterior no sería inteligente hacer apuestas con otra persona con la regla de "pagar 100 Pts. si sale un número distinto de dos y ganar 100 Pts. si sale dos". A menos que uno esté seguro de que los números mayores que dos no pueden salir nunca, se está abocado a perder sistemáticamente con ese tipo de apuesta. La condición de coherencia es la siguiente: "Se supone que no deseas hacer apuestas que con seguridad conducirán a una pérdida". A partir de este criterio se pueden derivar las leyes básicas de probabilidad. Por tanto, el criterio de coherencia es notablemente potente, proporcionando un fundamento intuitivo pero suficiente para la teoría. La probabilidad subjetiva puede ser un precursor fundamental para la formal enseñada en la universidad. La concepción clásica requiere cierta destreza con las fracciones,

mientras que la subjetiva puede depender sólamente de comparaciones de verosimilitudes percibidas.

Probabilidad formal

Hablamos de probabilidad formal cuando ésta se calcula con precisión usando las leyes matemáticas de la teoría axiomática correspondiente. Se conoce también como probabilidad objetiva o normativa. La base matemática puede reflejar hipótesis hechas en las concepciones clásica, frecuencial o subjetiva.

La teoría matemática de la probabilidad, tal y como hoy se conoce, es de un origen comparativamente reciente. Fué Kolmogorof quien la axiomatizó en su trabajo fundamental publicado en 1933 y traducido posteriormente al inglés con el título "Foundations of the theory of probability". Según este autor los sucesos se representan por conjuntos y la probabilidad es una medida normada definida sobre estos conjuntos. Los axiomas propuestos están basados en la abstracción de las propiedades de las frecuencias relativas: Si E es el espacio muestral asociado a un experimento aleatorio y *A* un s-álgebra de sucesos de E, una función P definida sobre *A* es una medida de probabilidad si:

1) A todo suceso S Ì *A* corresponde un número P(S), tal que 0 £ P(S) £ 1.

2) La probabilidad del suceso seguro es uno (P(E)= 1).

3) Si $(S_i)_{i \in I}$ son sucesos incompatibles dos a dos, siendo el conjunto de índices I finito o numerable, se verifica:

$$P(\bigcup S_i) = \sum P(S_i)$$

Este desarrollo, basado en la teoría de la medida, no sólo proporcionó un fundamento lógico consistente para el Calculo de Probabilidades, sino que también la conectó con la corriente principal de la matemática moderna.

La teoría axiomática de Kolmogorov surgió como consecuencia de las restricciones que el concepto clásico laplaciano imponía sobre la equiprobabilidad de los sucesos y la finitud del espacio muestral correspondiente. Una primera extensión de la definición de Laplace fué usada para calcular las probabilidades de sucesos con resultados infinitos. La noción de igual verosimilitud de ciertos sucesos jugó un papel clave en esta extensión. Según este desarrollo si E es alguna región con una medida conocida (longitud, área, volumen) la probabilidad de que un punto elegido al azar pertenezca a un subconjunto A de E es el cociente entre la medida de A y la medida de E.

Las dificultades conceptuales y de índole matemática que ésta aproximación a la probabilidad comporta desaconseja su tratamiento en el período de enseñanza obligatoria, de modo que cuando se habla de probabilidad en primaria o secundaria obligatoria, sin duda no se habla de la probabilidad bajo un punto de vista formal - axiomático.

5.10. DESARROLLO PSICOLOGICO DE LA INTUICION PROBABILÍSTICA EN EL NIÑO

Los textos más significativos sobre el desarrollo de la cognición probabilística son los clásicos de Piaget e Inhelder (1951) y Fischbein (1975). Mientras que Piaget tiende a definir el nivel de desarrollo en que se encuentra el niño, proporcionando razones para el retraso en la acción del profesor, algunos de los trabajos citados por Fischbein se preocupan de analizar el efecto de la instrucción en el proceso de aprendizaje.

Fischbein concede, además, una gran importancia a la intuición como parte integrante de la conducta inteligente. Las intuiciones son, según Fischbein, adquisiciones cognitivas que intervienen directamente en las acciones prácticas o mentales, en virtud de sus características de inmediatez, globalidad, capacidad extrapolatoria, estructurabilidad y auto-evidencia. Establece

varias clasificaciones de las intuiciones, distinguiendo, en primer lugar, entre intuiciones primarias y secundarias.

Las intuiciones primarias son adquisiciones cognitivas que se derivan directamente de la experiencia, sin necesidad de ninguna instrucción sistemática. Ejemplo de ellas son las intuiciones espaciales elementales, como el cálculo de distancia y localización de objetos, o la apreciación de que al lanzar un dado todas las caras tienen la misma probabilidad de salir.

Por el contrario, las *intuiciones secundarias* consisten en adquisiciones que tienen todas las características de las intuiciones, pero que son formadas por la educación científica, principalmente en la escuela. Como ejemplo puede servir la idea de que un móvil conserva su estado de movimiento o de reposo mientras no intervenga una fuerza exterior.

En el campo de la probabilidad, una intuición secundaria, aunque mal concebida, podría ser la llamada "falacia del jugador", por la cual, después de lanzar una moneda tres veces y haber obtenido tres caras, el sujeto tiende a predecir que la próxima vez es más probable que salga cruz. Esto se debe a una mala interpretación de la ley de los grandes números. Una intuición secundaria no se reduce a una simple fórmula aceptada o utilizada automáticamente; lo más interesante es que la adquisición se transforma en convicción, en creencia, en un sentimiento de evidencia. Pero para convertir una información en una intuición no es suficiente una simple explicación teórica, sino que el alumno ha de utilizarla en sus propias acciones y predicciones a lo largo de gran parte de su desarrollo intelectual.

Siguiendo a Fischbein, resumimos, a continuación, los principales resultados hallados en la bibliografía acerca de la génesis de la idea de azar y probabilidad desde la infancia a la adolescencia. Para cada estadio se estudian las siguientes facetas:

- la intuición del azar;
- la intuición de la frecuencia relativa;
- la estimación de probabilidades;
- operaciones combinatorias;
- el efecto de la instrucción sobre cada una de estas facetas.

El niño de preescolar

La intuicion del azar

Piaget e Inhelder concluyen de sus experimentos que no hay una intuición del azar innata en el niño, como no existía tampoco en el hombre primitivo, que atribuía los sucesos aleatorios a causas ocultas o a la "voluntad de los dioses". Para Piaget la comprensión del azar presupone la apreciación del carácter irreversible de una mezcla, y, por tanto, la posesión de un esquema combinatorio. Un experimento piagetiano clásico utiliza una bandeja con dos compartimentos. En los dos compartimentos de ésta se colocan ocho bolas blancas y ocho rojas. Al bascular la bandeja se produce la mezcla progresiva de las dos clases de bolas. En la primera etapa del desarrollo del concepto de azar (preoperacional), los niños afirman que las bolas vuelven nuevamente a su lugar original, o bien que el conjunto completo de blancas acabarán en el lugar ocupado originalmente por las rojas, y viceversa. Piaget interpreta esta reacción en el sentido de que el niño, antes de los 7 años, no comprende la naturaleza irreversible de la mezcla aleatoria y esto le impide una apreciación del azar.

Sin embargo, la opinión de Piaget es rechazada por Fischbein para quien la **intuición primaria** del azar, esto es, la distinción entre fenómeno aleatorio y determinista sin instrucción previa, está presente en la conducta diaria de cada niño, incluso antes de la edad de 7 años. El azar es equivalente a impredecibilidad y cuando el número de posibilidades, y consiguientemente el número de combinaciones posibles, es pequeño, el niño de preescolar razona correctamente y a veces, como se ha puesto de manifiesto en algunas investigaciones, más correctamente que el niño que ha alcanzado la etapa de las operaciones formales.

La intuicion de la frecuencia relativa

Diferentes investigadores han llevado a cabo experimentos de aprendizaje probabilístico, en los cuales se trata de estudiar las predicciones de los sujetos ante situaciones en que un suceso se repite con una determinada frecuencia relativa. Un ejemplo de esta clase de experiencias consiste en presentar al alumno dos luces de color diferente (pueden ser rojo y verde) que se irán encendiendo intermitente y aleatoriamente con una determinada frecuencia, por ejemplo, el 70 y el 30%, respectivamente. El alumno debe predecir cuál de las dos luces se encenderá la próxima vez. Los resultados obtenidos en este tipo de experimentos apoyan fuertemente la conclusión de que el niño de preescolar adapta sus predicciones a las probabilidades de los sucesos que se le presentan como estímulo, aunque sus respuestas no llegan a coincidir totalmente con la frecuencia de los mismos.

La estimacion de posibilidades y la nocion de probabilidad

Distintos autores han afirmado que el niño de preescolar es incapaz de estimar correctamente las posibilidades a favor y en contra de los sucesos aleatorios, basándose en que el niño de esta edad no posee los recursos necesarios:

- la habilidad de distinguir entre el azar y lo deducible;
- el concepto de proporción;
- los procedimientos combinatorios;

Sin embargo, para Fischbein, estas carencias no impiden al niño hacer juicios probabilísticos. Un experimento que pone de manifiesto esta capacidad consiste en presentar a los alumnos cinco conjuntos de canales por los que una bola puede rodar recorriendo distintas trayectorias. A los niños se le plantean preguntas como: "Si lanzo una bola por cada canal, ¿en cuál de ellos es más probable que salga la bola por el orificio 1?", o bien, "Si lanzo la bola muchas veces seguidas, ¿crees que saldrá el mismo número de veces por cada orificio en todos los canales, o saldrá por uno con más frecuencia que por otros?.

Otro experimento consiste en la elección, por parte del alumno, entre dos urnas o cajas con diferente contenido, aquella que ofrezca más posibilidades de obtener una bola de un color determinado. Si se realiza un adecuado control experimental (posición de los objetos en el espacio, preferencia de color, etc) y las operaciones auxiliares de comparación y cálculo requeridas son simples, el niño de preescolar es capaz de hacer apuestas basadas en una estimación probabilística.

Operaciones combinatorias

Piaget e Inhelder han probado que el niño de preescolar sólo puede hacer algunas combinaciones, permutaciones y variaciones de una manera empírica, y no intentan encontrar un método de realizar un inventario exhaustivo.

El efecto de la instruccion

Usando un procedimiento de instrucción elemental, Fischbein y sus colaboradores han intentado mejorar las respuestas de los niños a cuestiones que implican la comparación de posibilidades en situaciones donde las razones no tenían iguales términos. Este intento no tuvo éxito. Es posible que, a esta edad, los niños no puedan asimilar un esquema que implique una comparación doble.

El periodo de las operaciones concretas

La intuicion del azar

A través de la adquisición de esquemas operacionales espacio-temporales y lógico-matemáticos, el niño adquiere la capacidad de distinguir entre el azar y lo deducible, incluso al nivel conceptual. Es consciente de que, por ejemplo, al lanzar 15 monedas es muy difícil obtener 15 cruces. Claramente, este proceso no se completa durante este período, puesto

que el pensamiento está todavía muy ligado al nivel concreto. No obstante, la representación del azar, que no es sino una intuición primaria en el niño de preescolar, se convierte en una estructura conceptual distinta y organizada después de la edad de los 7 años. El niño comienza a comprender la interacción de cadenas causales que conducen a sucesos impredecibles, y la irreversibilidad de los fenómenos aleatorios.

La intuicion de la frecuencia relativa

La mayoría de los investigadores han encontrado que la intuición de la frecuencia relativa de sucesos, puesta de manifiesto a través de experimentos de aprendizaje probabilístico, mejora con la edad. Si la intuición se ve como el resultado cognitivamente fijado de experiencias acumuladas, parece razonable que la intuición de la frecuencia relativa se desarrolle de un modo natural como resultado de las experiencias del niño con situaciones que implican sucesos aleatorios, en los cuales las respuestas deben expresar una estimación correcta de las frecuencias relativas de los fenómenos.

La estimacion de posibilidades y la nocion de probabilidad

Los niños de 9-10 años pueden resolver problemas que impliquen comparación de probabilidades de un mismo suceso A en dos experimentos diferentes sólo en situaciones donde, bien el número de casos favorables o el número de casos no favorables a A son iguales en ambos experimentos (sus estimaciones se basan en comparaciones binarias).

En problemas donde las posibilidades son referidas a proporciones de elementos discretos (bolas en un recipiente), las respuesta de los niños de 9-10 años no son mejores que las que se obtendría por una respuesta al azar, y no son significativamente mejores que las respuestas de los niños de preescolar, excepto en el caso citado. En problemas donde las posibilidades tienen que ser determinadas a partir de una configuración geométrica (canales bifurcados por donde unas bolas pueden circular de un modo aleatorio) el porcentaje de respuestas correctas decrece incluso con la edad.

Las operaciones combinatorias

Durante el período de las operaciones concretas, los niños buscan modos de realizar inventarios de todas las permutaciones, variaciones y combinaciones posibles en un conjunto dado con un número pequeño de elementos, y llegan a procedimientos rudimentarios de cálculo mediante ensayo y error.

Los experimentos de Fischbein han demostrado que, al final de este periodo (10-11 años) los niños pueden, con la ayuda de instrucción, asimilar los procedimientos enumerativos usados en la construcción de diagramas en árbol.

El efecto de la instruccion

Con la instrucción, las respuestas de los niños de 9-10 años pueden mejorar significativamente en problemas que no pueden ser reducidos a comparaciones binarias. Fischbein ha demostrado que, por medio del uso de procedimientos figurativos, pueden ser construidos, al nivel de las operaciones concretas, esquemas considerados por Piaget e Inhelder como accesibles sólo al nivel de las operaciones formales. La ausencia de proporcionalidad no es un obstáculo para aprender el concepto de probabilidad. Incluso antes de la edad de 10 años, el niño es capaz de asimilar este esquema con la ayuda de instrucción elemental.

El periodo de las operaciones formales

La intuicion del azar

Según Piaget e Inhelder, el adolescente agrupa las relaciones no determinadas de fenómenos aleatorios según esquemas operacionales. Una vez que se presenta una situación aleatoria, por medio del uso de estos esquemas se hace inteligible, y la síntesis entre el azar y lo operacional conduce al adolescente al concepto de probabilidad.

Fischbein sostiene que la síntesis entre el azar y lo deducible no se realiza espontánea y completamente al nivel de las operaciones formales. En experimentos donde se requiere al sujeto reconocer probabilidades iguales en diferentes condiciones experimentales, es el adolescente quien evita lo impredecible, y busca dependencias causales que reduzcan lo incierto, incluso en situaciones donde no existen tales dependencias.

La estructura operacional del pensamiento formal por sí sola no puede hacer inteligible al azar, incluso aunque pueda proporcionar los esquemas que son necesarios para esto, o sea, capacidad combinatoria, proporcionalidad, e implicación. La explicación para esta deficiencia es que las tradiciones culturales y educativas de la sociedad moderna orientan el pensamiento hacia explicaciones deterministas unívocas, según las cuales los sucesos aleatorios caen fuera de los límites de lo racional y científico.

La intuicion de la frecuencia relativa

Las investigaciones que se han realizado con diferentes niveles de edad han demostrado que el adolescente ha hecho progresos en comparación a los niños más pequeños en lo que se refiere a la intuición de la frecuencia relativa, particularmente en casos donde las predicciones tienen algún resultado práctico. La estrategia óptima ante decisiones en condiciones aleatorias muestra los efectos favorables del desarrollo de la inteligencia sobre las predicciones en ciertas condiciones experimentales.

La estimacion de posibilidades y la nocion de probabilidad

El logro de los adolescentes estimando posibilidades a favor y en contra de un resultado es superior al de los niños pequeños. Cuando el material experimental consiste en un recipiente con bolas, los niños de 12 años dan respuestas correctas desde el principio, incluso en casos en que tienen que comparar razones con términos desiguales. Tal descubrimiento es previsto por la teoría de Piaget. Lo que Fischbein añade a esto es el hecho de que incluso niños de 9-10 años pueden responder correctamente a tales situaciones, si tienen la instrucción adecuada.

Las operaciones combinatorias

Piaget e Inhelder afirman que, durante la etapa de las operaciones formales, el niño adquiere la capacidad de usar procedimientos sistemáticos para realizar inventarios de todas las permutaciones posibles, variaciones y combinaciones de un conjunto dado de elementos.

La investigación de Fischbein ha demostrado, sin embargo, que esto es sólo una potencialidad para la mayoría de los sujetos. Bajo su punto de vista, sería más preciso afirmar que estos alumnos son capaces de asimilar procedimientos combinatorios con la ayuda de la instrucción, y que esto es también verdad para los niños de 10 años. Aunque hay diferencias en la realización entre estos dos niveles de edad, estas diferencias son bastante pequeñas.

El efecto de la instruccion

Las lecciones experimentales realizadas por Fischbein y sus colaboradores se hicieron con niños de 12-14 años. Dichas lecciones trataron los siguientes conceptos y procedimientos: suceso, espacio muestral, suceso elemental y compuesto, probabilidad como una medida del azar, frecuencia relativa, y análisis combinatorio.

Los resultados de la instrucción revelaron un mayor interés y receptividad de los adolescentes en lo que se refiere a las ideas de probabilidad y estadística. Estos sujetos son capaces de comprender y aplicar correctamente los conceptos enseñados. Para este autor, los modelos generativos (por ejemplo, diagramas en arbol, en el caso de las operaciones

combinatorias) son los mejores dispositivos de enseñanza para la construcción de intuiciones secundarias.

TEMA 6

LA DISTRIBUCIÓN NORMAL

6.1. INTRODUCCIÓN

En los temas 1 a 4 hemos estudiado las variables estadísticas, y las características de su distribución de frecuencias, es decir, las medidas de posición central, dispersión y forma. En dichos temas solo nos hemos interesado por el estudio exclusivo de las muestras, sin tener la pretensión de generalizar los resultados a las poblaciones de donde se habían tomado estas muestras. Por ejemplo, en el fichero ALUMNOS, usado en las clases de prácticas hemos analizado la altura y el peso de 60 alumnos que cursaron la asignatura los años anteriores, sin tratar de extender las conclusiones a otros alumnos.

Sin embargo, la mayor utilidad de la estadística se encuentra, precisamente al tratar de predecir el comportamiento de una o varias variables en una población, a partir de los datos de estas variables en una muestra aleatoria de la población. Por ejemplo, podríamos estar interesados en predecir cómo se distribuirían las alturas de los alumnos de la Facultad de Educación o, en general, de los alumnos universitarios. En este tema iniciaremos el estudio de la forma en que estas *inferencias* pueden llevarse a cabo.

Para ello necesitaremos *modelos teóricos* de la distribución de los datos en la población. En el tema 5 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;
- Problemas técnicos: distribución de las medidas de piezas manufacturadas, etc.

Antes de comenzar este tema, te contaremos brevemente cómo surge esta distribución. Al tratar de resolver un problema planteado por Jacob Bernoulli sobre la forma de estimar un valor medio de la población, a partir de una muestra de valores, Abraham DeMoivre encuentra en 1733 que la ecuación de la distribución normal describe la distribución de los valores de las medias muestrales alrededor de la media de la población. Con ello proporcionó una base sobre la cual se fundamenta gran parte de la teoría estadística inductiva. A la distribución normal, frecuentemente, se la llama **distribución gaussiana**, en honor a Karl Friedrich Gauss (1777 – 1855), quien también obtuvo su ecuación al estudiar la distribución de los errores en mediciones repetidas de la misma cantidad.

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. Cuando queremos decidir si la distribución normal sería una buena aproximación a un conjunto de datos, comenzamos por estudiar la forma de la distribución de estos datos, y el porcentaje de casos que se distribuye alrededor de la media, como haremos en la Actividad 5.1.

Actividades

6.1. La Tabla de frecuencias 6.1 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. La salida de ordenador proporciona también algunos resúmenes estadísticos.

Tabla 6.1. Frequency Tabulation for altura

Class	Lower Limit	Upper Limit	Midpoint	Abolute Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel Frequency
at or below		140,0		0	0,0000	0	0,0000
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
above	180,0			0	0,0000	1000	1,0000

Mean = 164,721 Standard deviation = 4,92274 Variance = 24,2334
Skewness = -0,165955 Kurtosis = -0,0385743

a) ¿Qué características puedes deducir, sobre la forma de las representaciones gráficas del histograma y polígono de frecuencias de esta distribución? ¿Es la distribución aproximadamente simétrica respecto a su centro? ¿Qué nos indica el coeficiente de apuntamiento?.
b) ¿En qué intervalo se encontrarían la moda y mediana? ¿Cuál sería su valor aproximado? ¿Recuerdas el significado de estas medidas ?
c) Calcula, a partir de la tabla y de un modo aproximado, el porcentaje de chicas en este grupo cuya altura está comprendida en el intervalo ($\bar{x}$ -2s, $\bar{x}$ + 2s), donde con $\bar{x}$ indicamos la media y con s la desviación típica de esta muestra. En una distribución normal teórica el porcentaje de casos que está situado a menos de dos desviaciones típicas de la media es el 95%.

Variables aleatorias continuas

En el tema 5 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 6.1 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.

Al considerar tal tipo de variables, estamos interesados en calcular, no sólo la probabilidad de que tome un valor determinado $P(\xi=b)$, sino probabilidades como las siguientes: $P(a \leq \xi \leq b)$, $P(\xi \leq a)$, $P(\xi \geq b.)$, etc.

Actividad 6.1 (continuación)

d) Supongamos que escribimos el nombre de cada chica que tomó parte en la muestra anterior en un papel y elegimos uno de ellos al azar, ¿Cuál será la probabilidad de que la chica en cuestión tenga una altura comprendida en el intervalo $(\bar{x} - 2s, \bar{x} + 2s)$? ¿Y que tenga una altura que caiga fuera del intervalo?

e) En un histograma, las áreas de cada rectángulo representan las frecuencias en el intervalo. Recíprocamente, a partir de las frecuencias podemos calcular el área que, en el histograma corresponde a un intervalo dado. En el histograma de frecuencias relativas:
   1. ¿Cuál sería el área correspondiente al intervalo (160-170)?
   2. ¿Cuál sería el área aproximada en este intervalo en el polígono de frecuencias?
   3. ¿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?
   4. ¿Y que mida más de 174 cm?

6.2. LA DISTRIBUCIÓN NORMAL

La distribución normal es un modelo teórico que puede servir para representar, en forma aproximada, algunas distribuciones de datos continuos. Al considerar en una población una variable continua o con un número grande de valores, ocurre a veces que la distribución es simétrica respecto a su valor medio, la mayor parte de valores se concentra alrededor de la media, y los valores sean menos probables cuanto más se alejen de la media 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.

Ejemplo 6.1. Distribución del CI.

El coeficiente intelectual de las personas (que denominaremos **CI**) es una variable teórica que supuestamente 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 se supone que se alcanza cuando el desarrollo intelectual de una persona es el promedio del correspondiente a su edad. Una puntuación superior o inferior a 100 indica más o menos capacidad intelectual que el promedio de su edad. En la tabla 6.2 se muestra la puntuación obtenida por 100 personas seleccionadas aleatoriamente en el cuestionario que mide el C: I. y en la figura 6.1 se representa el histograma de frecuencias correspondiente.

Tabla 6.2. Coeficientes intelectuales de 100 personas

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		60,0		0	0,0000	0	0,0000
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
above	140,0			0	0,0000	100	1,0000

Mean = 98,9919 Standard deviation = 16,2713

Figura 6.1. Distribución del CI de 100 personas

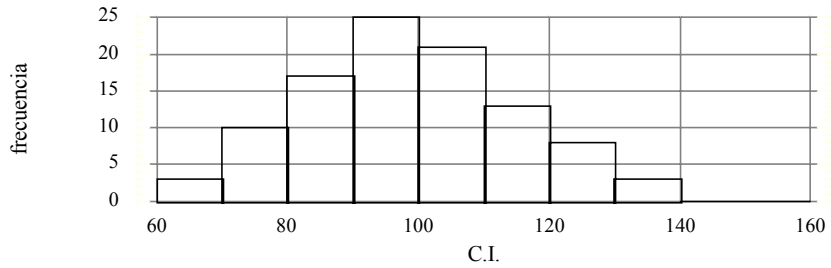


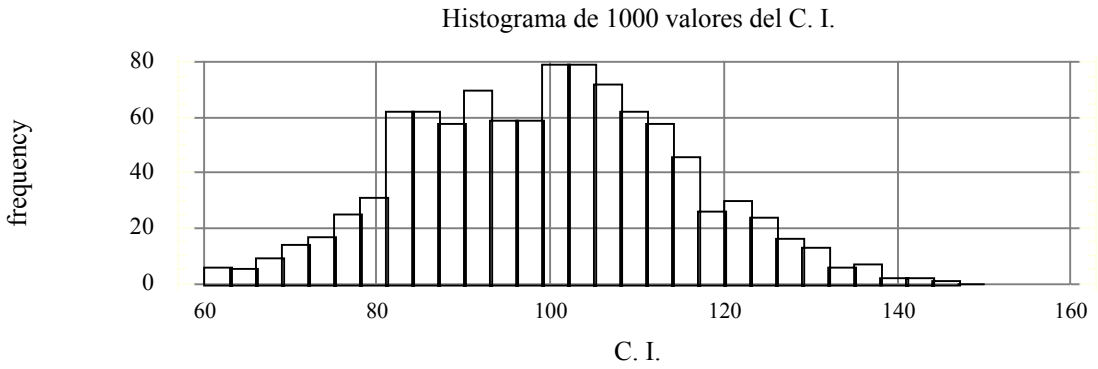
Tabla 6.3. Coeficientes intelectuales de 1000 personas

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		60,0		0	0,0000	0	0,0000
1	60,0	63,0	61,5	6	0,0060	6	0,0060
2	63,0	66,0	64,5	5	0,0050	11	0,0110
3	66,0	69,0	67,5	9	0,0090	20	0,0200
4	69,0	72,0	70,5	14	0,0140	34	0,0340
5	72,0	75,0	73,5	17	0,0170	51	0,0510
6	75,0	78,0	76,5	25	0,0250	76	0,0760
7	78,0	81,0	79,5	31	0,0310	107	0,1070
8	81,0	84,0	82,5	62	0,0620	169	0,1690
9	84,0	87,0	85,5	62	0,0620	231	0,2310
10	87,0	90,0	88,5	58	0,0580	289	0,2890
11	90,0	93,0	91,5	70	0,0700	359	0,3590
12	93,0	96,0	94,5	59	0,0590	418	0,4180
13	96,0	99,0	97,5	59	0,0590	477	0,4770
14	99,0	102,0	100,5	79	0,0790	556	0,5560
15	102,0	105,0	103,5	79	0,0790	635	0,6350
16	105,0	108,0	106,5	72	0,0720	707	0,7070
17	108,0	111,0	109,5	62	0,0620	769	0,7690
18	111,0	114,0	112,5	58	0,0580	827	0,8270
19	114,0	117,0	115,5	46	0,0460	873	0,8730
20	117,0	120,0	118,5	26	0,0260	899	0,8990
21	120,0	123,0	121,5	30	0,0300	929	0,9290
22	123,0	126,0	124,5	24	0,0240	953	0,9530
23	126,0	129,0	127,5	16	0,0160	969	0,9690
24	129,0	132,0	130,5	13	0,0130	982	0,9820
25	132,0	135,0	133,5	6	0,0060	988	0,9880
26	135,0	138,0	136,5	7	0,0070	995	0,9950
27	138,0	141,0	139,5	2	0,0020	997	0,9970
28	141,0	144,0	142,5	2	0,0020	999	0,9990
29	144,0	147,0	145,5	1	0,0010	1000	1,0000
30	147,0	150,0	148,5	0	0,0000	1000	1,0000
above	150,0			0	0,0000	1000	1,0000

Mean = 99,4992 Standard deviation = 15,4664

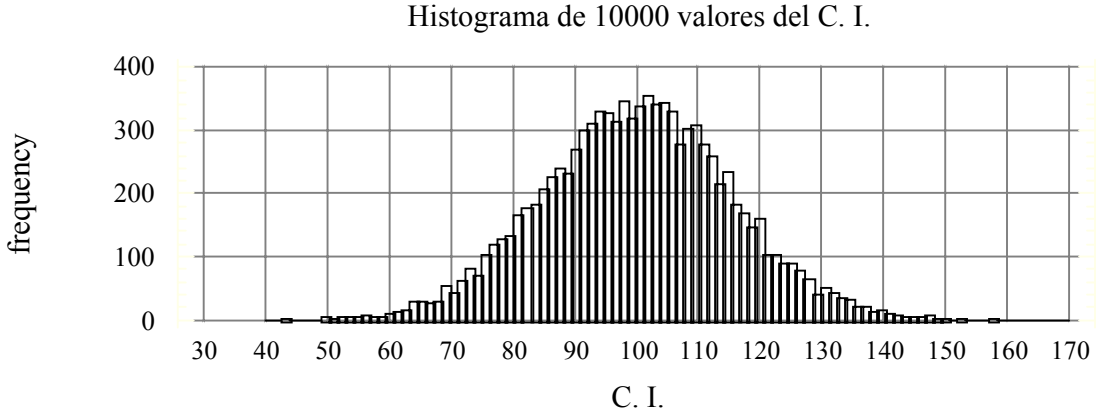
Vemos que el histograma es unimodal (una sola moda), y la moda se sitúa, aproximadamente, en el centro de la distribución. El mayor número de casos se concentra en el intervalo 90-100 y a ambos lados la distribución decrece rápidamente, aunque es todavía algo asimétrica. Veamos qué sucede si aumentamos la muestra a 1000 personas (tabla 6.3) y Figura 6.2) y a la vez aumentamos el número de intervalos.

Figura 6.2. Distribución del CI de 1000 personas



En la figura 5.2, 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 6.3. que corresponde a 10.000 puntuaciones del C. I.

Figura 6. 3. Distribución del CI de 1000 personas



Vemos que a medida que, simultáneamente aumentamos el número de valores recogidos de una variable estadística y reducimos el ancho del intervalo, el histograma (y también el polígono de frecuencias) se aproxima a una curva continua y podemos sustituirlo por dicha curva, que llamaremos ***curva de densidad***. La función matemática correspondiente a dicha curva se llama ***función de densidad.***

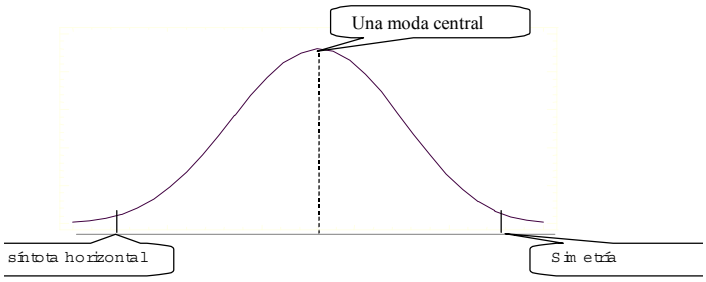
Podemos usar la función de densidad para resolver problemas tales como el cálculo de la probabilidad de que el C. I. sea mayor que 120. Otros ejemplos serían calcular la probabilidad de que el peso de un recién nacido sea inferior a 2 Kg., o que la cotización del dólar la próxima semana supere las 180 Pts.

Para resolver este tipo problema usaremos el hecho de 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 X y los extremos a y b.

Algunas distribuciones de datos, como el C. I. pueden aproximarse bien por una curva de densidad en forma de campana invertida, con la forma característica que mostramos en la figura 6.4. Esta curva corresponde a una función matemática de ecuación conocida que se denomina ***función de densidad normal*** La variable aleatoria cuya función de densidad es la ***función de densidad normal*** se conoce como ***distribución normal.***

Esta distribución es un modelo teórico, que es una buena aproximación de algunos modelos reales de datos y su función de densidad posee algunos elementos característicos que se muestran en la figura 6. 4: Una moda central y dos colas simétricas a cada lado de la media, así como una asíntota horizontal.

Figura 6. 4. Forma característica de la función de densidad en la Distribución Normal



Si volvemos al ejemplo del test de CI, podemos ver que podríamos sustituir el histograma de la figura 6.3. por una función de densidad normal, es decir, vemos que el C. I. sigue una distribución aproximadamente normal. Es más fácil trabajar con la función de densidad normal que con el histograma, ya que la forma del histograma depende de nuestra elección de los intervalos de clases, mientras la curva de densidad normal no depende de dicha elección. A continuación, veremos de qué forma podemos trabajar con la curva de densidad normal:

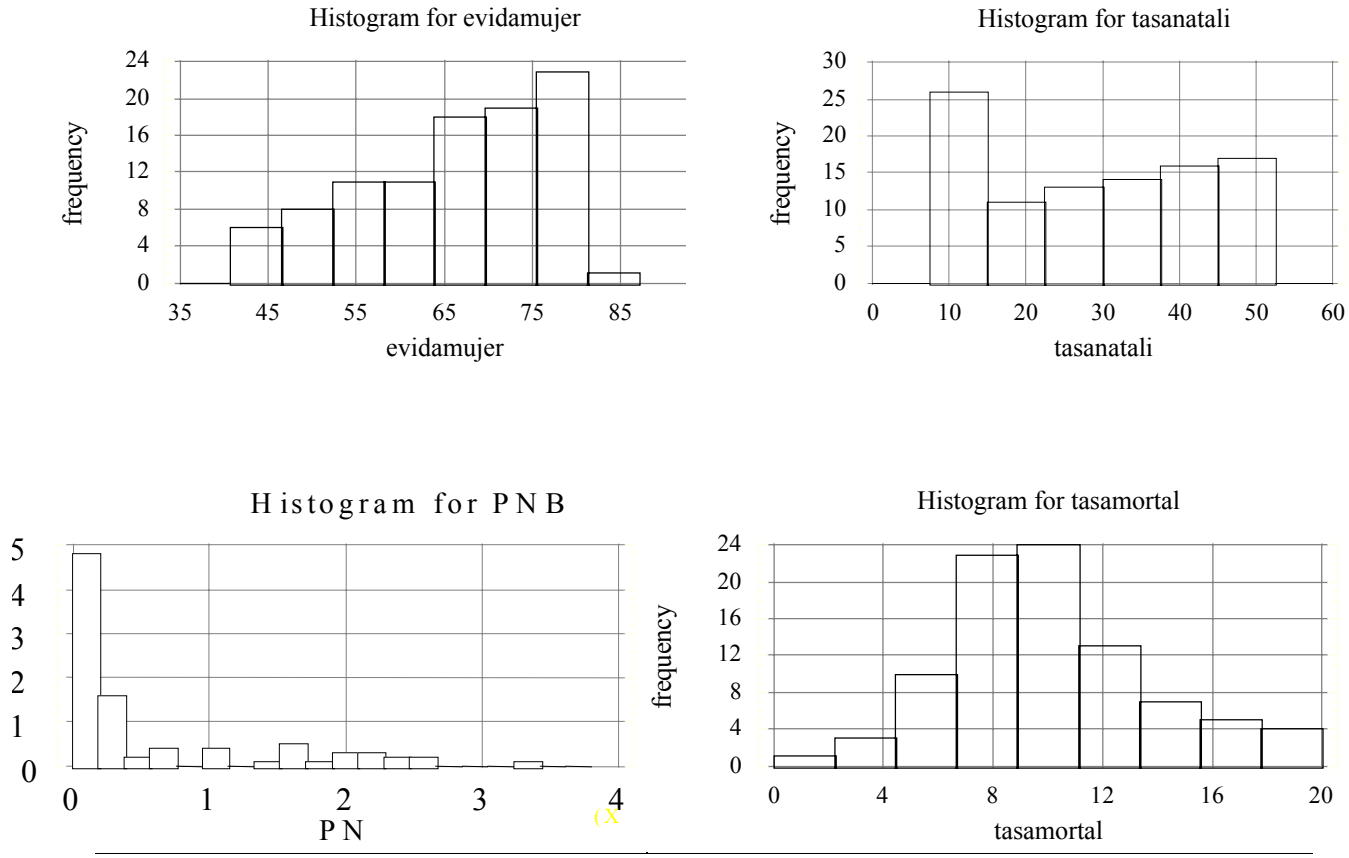
- Podemos usarla para *estimar la probabilidad* de que una observación futura caiga dentro de cada rango de valores. Las áreas de los rectángulos del histograma nos dan la frecuencia relativa de datos que caen en cada clase en la muestra tomada. Puesto que la curva de densidad se aproxima al histograma de frecuencias relativas al aumentar el número de casos, si recogíesemos datos de toda la población llegarían a coincidir. En consecuencia *las áreas bajo la curva de densidad dentro de un rango de valores nos da la probabilidad de que una nueva observación esté incluida en dicho rango.*

  Por ejemplo, en la Figura 6.3 vemos que para el C. I. la mitad de las 10.000 observaciones aproximadamente caen por encima de 100 y la otra mitad por debajo. Podemos, por tanto estimar que, si damos a una nueva persona el cuestionario que mide el C. I. la probabilidad de que su puntuación sea mayor que 100 es aproximadamente igual a 1/2, sin tener que medir el coeficiente intelectual de todas las personas de la población.

- *El área bajo la curva de densidad es exactamente igual a 1.* Esta propiedad es cierta para cualquier otra función de densidad, aunque no siga la distribución normal. El área bajo la curva de densidad en cualquier rango de valores sobre el eje horizontal nos da la probabilidad de que una observación caiga en este rango. Si considero todo el conjunto de valores, el área debe ser igual a 1, que es la probabilidad de que la variable tome cualquier valor posible.

Nota: Aunque muchas variables continuas siguen la distribución normal, también hay otras cuya distribución es claramente no normal. En la figura 6.5. mostramos los histogramas de diversas variables del proyecto 2, donde podemos ver algunas que claramente presentan una forma diferente de la típica en la distribución normal y otras que podrían aproximarse por esta distribución.

Figura 6.5.



6. 3. DEFINICIÓN DE LA DISTRIBUCIÓN NORMAL

Como hemos visto en la discusión anterior, la idea de variable aleatoria surge como generalización de los conceptos de histograma, polígono de frecuencias y variable estadística. La variable estadística se refiere al conjunto de datos particular, en el que estaremos interesados en los diferentes valores que toma la variable, y las frecuencias y proporciones con que los diferentes valores aparecen en las observaciones. Generalmente los datos se han obtenido de una **muestra** que suponemos representativa y aleatoriamente extraída de una **población**.

Para que una variable aleatoria siga la distribución normal, la primera condición es que sea **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, que haga necesaria su agrupación en intervalos.

La función de densidad normal está definida por una fórmula matemática que depende de dos **parámetros** μ y σ, que son su media y su desviación típica y que la determinan por completo. La función de densidad de la variable aleatoria normal X, con media μ y varianza σ^2, es:

$$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)/\sigma]^2}, \qquad -\infty < x < \infty,$$

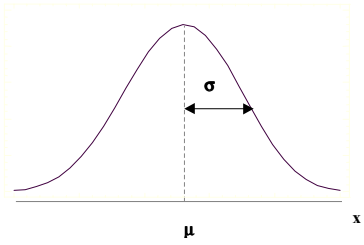
donde: $\pi = 3.14159\ldots$ y $e = 2.71828\ldots$.

Veamos ahora el significado de estos parámetros:

Sabemos que la **media** μ es una característica de posición central que se obtiene sumando el valor de todos los datos y dividiendo por el número de datos. Por ejemplo, en el CI, la media 100 indica que si promediamos las puntuaciones de todos los sujetos de la población, obtenemos 100 como valor medio. En el caso de una distribución normal la media μ coincide con la moda, es decir el valor de la variable aleatoria X que aparece con mayor probabilidad, lo que se observa claramente en el ejemplo del C. I.

La **desviación típica** σ es una medida de la dispersión de los datos. En una distribución normal, la desviación típica puede determinarse en forma aproximada gráficamente. 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 6.5. Significado de la media y desviación típica (parámetros) en la distribución normal



μ es la media y σ es la desviación típica.

Actividades

6. 2. La función de densidad es una función matemática que usamos como modelo de una distribución de datos. Con ella podemos representar aproximadamente y hacer cálculos aproximados sobre la distribución de datos. 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?

6.4. PROPIEDADES DE LA DISTRIBUCIÓN NORMAL

Aunque en secciones anteriores nos hemos referido a algunas de las propiedades de la distribución normal, a continuación, enumeraremos las propiedades más importantes:

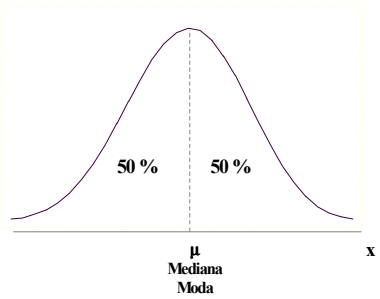
Simetría:

La función de densidad normal es simétrica, respecto a un eje que pasa por μ, debido a que en su fórmula aparece una exponencial al cuadrado. La simetría es importante en la distribución normal, pues permite que el cálculo de áreas resulte más sencillo. A continuación veremos algunas propiedades derivadas de la simetría:

- La distribución normal es simétrica respecto del eje vertical que pasa por su valor medio. 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. Es decir: existe la misma proporción de casos por encima y por debajo de la media.
- Sabemos que la mediana en un conjunto de datos es el valor tal que la mitad de los datos son menores o iguales y el resto mayores o iguales que la mediana. En consecuencia, la mediana de una curva de densidad es el punto que la divide en áreas iguales. y la

probabilidad de ser menor a ella es igual a ½. En la distribución normal, por tanto, la mediana coincide con la media.

Figura 6. Propiedad de simetría de la normal



- 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

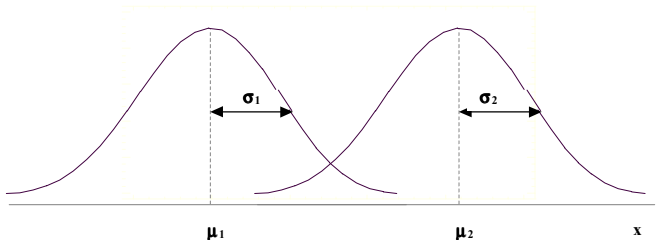
6.3. Trazar, a mano alzada, una curva de densidad normal. Trazar una curva de densidad que sea simétrica, pero cuya forma sea diferente a la de la distribución normal.

6.4. 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

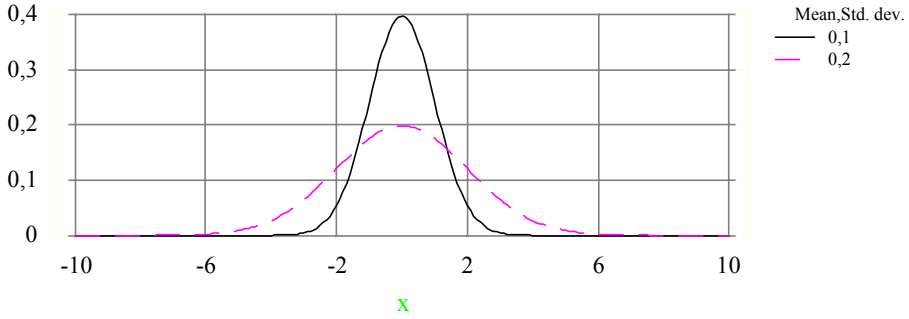
Dos curvas normales con la misma desviación típica pero diferentes medias, son idénticas pero se centran en diferentes posiciones a lo largo del eje horizontal. Si observamos la figura 6.7, podemos ver que $\sigma_1 = \sigma_2$ pero las medias son distintas, en consecuencia, las curvas están desfasadas sobre el eje horizontal.

Figura 6.7. 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 6.8).

Figura 6. 8. Distribuciones normales con igual media

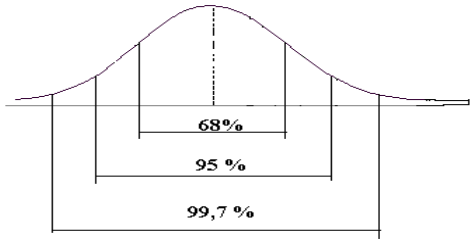


Distribución de casos en relación con la desviación típica

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. En cualquier distribución normal con media μ y desviación típica σ, se verifica (Figura 6.9):

- 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 2 σ de la media μ.
- El 99,7 % de los datos están a una distancia igual 3σ de la media μ.

Figura 6.9. Distribución de casos en la curva normal



Actividades

6.5. Las puntuaciones obtenidas en un test de inteligencia por un grupo de alumnos siguen una distribución normal con media 110 y desviación típica 25. A) ¿Qué proporción de alumnos puntúa por encima de 110? B) Obtener los valores de las puntuaciones tales que el 95% central de los casos esté comprendido entre dichos valores.

6.6. La temperatura media en Noviembre en Nueva York 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?

6.7. Dada una distribución de puntuaciones de un test que sigue la distribución normal de probabilidades, con media μ=12 y =4, [N(12,4)], a) ¿qué porcentaje de casos cae entre 8 y 16?

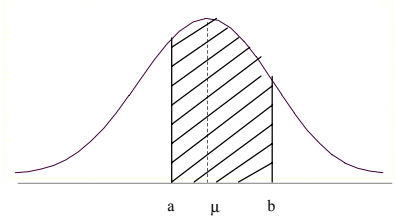
6.5. CÁLCULO DE PROBABILIDADES UTILIZANDO LA DISTRIBUCIÓN NORMAL

Las propiedades anteriores son importantes para el cálculo de probabilidades. Debes recordar los siguientes puntos:

- La probabilidad de que un valor esté entre los puntos a y b es igual al área bajo la curva normal comprendido entre los puntos a y b (figura 6.10).
-

Figura 6.10. Area bajo la curva normal

- 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 asociada a cualquier valor x es siempre mayor o igual a cero.



$p(x_i) \geq 0$ para todo x_i

- El área total bajo la curva y por encima del eje horizontal es igual a 1. Como cada área limitada

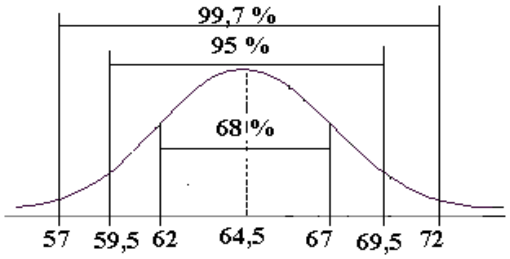
por dos valores cualesquiera representa la probabilidad entre esos dos valores, la suma de todas las áreas corresponde a la suma de todas las probabilidades, en consecuencia, dicha suma (integral) es 1 y lo expresamos en la forma siguiente:

$$\int_{-\infty}^{+\infty} f\ (\ x\ )\ dx \quad = \quad 1$$

A continuación mostraremos de qué forma se puede usar la curva normal para calcular probabilidades.

Ejemplo 6.2.El peso de los chicos de 18 a 24 años de edad es aproximadamente normal con media μ = 64,5 kilos y una desviación típica σ = 2,5 kilos. La figura 6.11 indica cómo se utiliza la regla del 68 – 95 – 99,7 % en este ejemplo**.**

Figura 6.11. Porcentajes centrales de pesos de chicos



Como la desviación típica es de 2,5 kilos dos desviaciones típicas serán 5 kilos para esta distribución. Si deseamos saber cuál es 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 entre 59,5 y 69,5 kilos de peso. Este resultado es exacto si la distribución fuera exactamente normal, pero como la distribución con la que estamos trabajando es aproximadamente normal, este cálculo nos da un resultado más o menos aproximado a la realidad.

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.

Si deseamos saber cuál es el intervalo que cubre el 99,7 % de los valores centrales debemos realizar las siguientes operaciones:

$$\mu - 3\sigma = 64{,}5 - 7{,}5 = 57 \text{ kilos}$$
$$\mu + 3\sigma = 64{,}5 + 7{,}5 = 72 \text{ kilos}$$

En conclusión, el 99,7 % central de los chicos está comprendido entre 57 y 72 kilos de peso.

Ejemplo 6. 1 (continuación)

Para mostrar como podemos usar estas propiedades para resolver problemas prácticos, responderemos a las siguientes preguntas, sobre el ejemplo 1, sobre el CI, que sigue aproximadamente una distribución normal N (100, 15).

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 2 σ de la media μ. El intervalo 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 % .

6. 6. EVALUACIÓN DE LA NORMALIDAD DE UNA DISTRIBUCIÓN

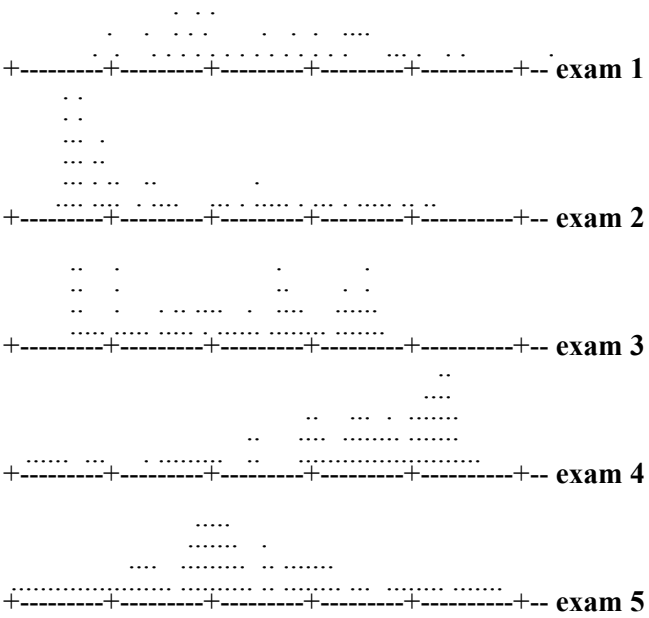
Para decidir si podemos describir una distribución de datos dada por una curva normal, debemos comprobar si en nuestros datos se cumplen las propiedades de las distribuciones normales que hemos descrito. Podemos analizar estas propiedades usando algunos gráficos y cálculos estadísticos, para poder juzgar acertadamente si nuestros datos son aproximadamente 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

Antes de continuar, recordaremos que debemos hacer énfasis en la distinción de la **media de la muestra** y **la media de la población.** A la primera la designamos con el símbolo $\bar{x}$ y a la segunda con el símbolo μ, asimismo s es la desviación típica de la muestra y σ la desviación típica de la población.

Actividades

6.8. Los siguientes gráficos muestran las distribuciones de las puntuaciones de 5 exámenes. Dos de estas muestras han sido extraídas de poblaciones normalmente distribuidas. Identifica las tres muestras que no proceden de una distribución normal, indicando en qué te basas.



6.9. 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?

6.10. 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. Calcular: a) ¿cual es la probabilidad de obtener puntuación igual o inferior a 16? b) ¿cuantos niños de dicho colegio tienen igual o mayor puntuación que 28?

6.11. 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

En lo que sigue analizaremos los pasos que debes seguir para decidir si la distribución normal proporciona o no una adecuada aproximación para una variable dada. Estos pasos se pueden resumir en el esquema siguiente:

1. Analizar de qué tipo es la variable
2. Analizar la forma de la distribución (unimodalidad, simetría y curtosis)
3. Analizar si se cumple la regla 68 – 95 – 99,7 para los intervalos de casos alrededor de la media

Es importante tener en cuenta que hay que comprobar todos los puntos. No basta con que se cumpla uno de ellos, sino que deben cumplirse todos los requisitos, si queremos

obtener una buena aproximación. A continuación te enseñamos como usar STATGRAPHICS para analizar estos puntos

Ejemplo 6.3: Usaremos como ejemplo un conjunto de datos sobre el CI de 1000 personas. La tabla 5, muestra la distribución de frecuencias de esta variable, obtenida con el paquete estadístico Statgraphics (opción DESCRIBE- NUMERICAL DATA- ONE VARIABLE ANALYSIS). La variable correspondiente se denomina COEF_INT y se presenta en la tabla de frecuencias 6.4.

Tabla 6.4. Frequency Tabulation for Coef-Int

Class	Lower Limit	Upper Limit	Midpoint	Abolute Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel Frequency
at or below	40.0			0	0.0000	0	0.0000
5	40.0	50.0	45.0	6	0.0060	6	0.0060
6	50.0	60.0	55.0	5	0.0050	11	0.0110
7	60.0	70.0	65.0	18	0.0180	29	0.0290
8	70.0	80.0	75.0	85	0.0850	114	0.1140
9	80.0	90.0	85.0	142	0.1420	256	0.2560
10	90.0	100.0	95.0	272	0.2720	528	0.5280
11	100.0	110.0	105.0	259	0.2590	787	0.7870
12	110.0	120.0	115.0	137	0.1370	924	0.9240
13	120.0	130.0	125.0	64	0.0640	988	0.9880
14	130.0	140.0	135.0	10	0.0100	998	0.9980
15	140.0	150.0	145.0	2	0.0020	1000	1.0000
above	150.0			0	0.0000	1000	1.0000

Mean = 99.0551 Standar deviation = 15.3527

1. Tipo de variable. En primer lugar comprobamos que se trata de una variable numérica que toma un número suficientemente grande de valores (de 40 a 150).

*2. Forma de la distribución***:** *debemos comprobar su simetría, unimodalidad y curtosis.*

2. a) Simetría y unimodalidad. Tenemos varias formas para comprobarla*:*

2.a1) Podemos analizar la forma aproximada del histograma y del polígono de frecuencias asociado, observando si son aproximadamente simétricos o no y si tienen una o varias modas, así como si aparecen valores atípicos.

Recordemos que para que la distribución normal proporcione un buen ajuste a los datos las gráficas deben ser aproximadamente simétricas, con una sola moda en el centro de la gráfica y sin valores atípicos. Podemos ver si hay valores atípicos a partir del gráfico de la caja o del gráfico del tronco, en donde aparecen marcados. La existencia de muchos valores atípicos o una asimetría muy pronunciada es suficiente para rechazar la normalidad de la distribución. Ten en cuenta, no obstante que la forma del histograma puede cambiar al variar el número de intervalos. Conviene probar con varias longitudes diferentes de intervalos, antes de aceptar si la forma de una distribución se asemeja a la normal.

2. a2) Podemos también deducir la simetría y unimodalidad observando la columna de frecuencias absolutas en la Tabla 6.4, donde vemos que en los intervalos inferiores y superiores las frecuencias son bastante pequeñas, mientras que en los intervalos centrales se acumula la mayor cantidad de datos. De la observación de las frecuencias absolutas podemos

inferir que el polígono de frecuencias será aproximadamente simétrico, con un solo pico (o moda).

2.a3) Podemos comparar la posición relativa de media, mediana y moda, que en una distribución simétrica deben coincidir.

Sabemos que la moda es el valor más frecuente. En el ejemplo, como estamos trabajando con intervalos de clase, buscaremos el intervalo modal que es (90; 100]. La moda es el valor central 95.

La mediana es el valor que divide a la distribución por la mitad. Para encontrar la mediana buscamos el valor de la variable que corresponda a una frecuencia acumulada igual a 500, que es la mitad del total de datos, que también corresponde al intervalo (90; 100]. Podemos suponer un valor aproximado de 95.

En este ejemplo, la media es 99,0551, vemos que media, mediana y moda toman valores próximos, lo que es una condición necesaria para que la distribución sea simétrica (aunque no es suficiente). Si en un caso dado media, mediana y moda son muy diferentes, entonces la distribución no sería simétrica y por tanto, no sería normal

2.a4) 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). Además sabemos que:

- Si el coeficiente de asimetría es menor que 0, significa que la distribución es asimétrica a izquierda, o presenta asimetría negativa.
- Si el coeficiente de asimetría es igual a 0, significa que la distribución es simétrica.
- Si el coeficiente de asimetría es mayor que 0, significa que la distribución es asimétrica a derecha, o presenta asimetría positiva.

2 b) *Curtosis.* Además de ser simétrica, debemos comprobar el apuntamiento o curtosis de la distribución, mediante el coeficiente de curtosis y curtosis tipificado. 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). Además sabemos que:*:*

- Si el coeficiente de curtosis es menor que 0, entonces la curva es un poco más aplanada que la normal.
- Si el coeficiente de curtosis es igual a 0, entonces la curva coincide con la normal.
- Si el coeficiente de curtosis es mayor que 0, entonces la curva es un poco menos aplanada que la normal.

3. Porcentajes de casos alrededor de la media. El tercer punto a comprobar es el porcentaje de casos que se distribuye en los ($\bar{x} - s$; $\bar{x} + s$); ($\bar{x} - 2s$; $\bar{x} + 2s$); ($\bar{x} - 3s$; $\bar{x} + 3s$), 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 3, $\bar{x} = 99{,}0551$ y $s = 15{,}3527$ (ver tabla 5). El intervalo: ($\bar{x} - s$; $\bar{x} + s$), será (99,0551 – 15,3527 ; 99,0551 + 15,3527) = (83,7024 ; 114,4078). Como los intervalos formados en la tabla de frecuencias son números enteros, realizaremos un redondeo y trabajaremos con el intervalo (84; 114).

Utilizando la opción PANE OPTIONS podemos cambiar los extremos y número de intervalos para encontrar las frecuencias relativas correspondientes a los intervalos comprendidos entre 84 y 114, cuyo resultado es 0,6778. Por lo tanto, la proporción

correspondiente al intervalo (84; 114) es 67,78% de personas que poseen un CI comprendido entre 84 y 114.

De la misma forma trabajamos para obtener los porcentajes relativos a los otros dos intervalos: $(\bar{x} - 2s;\ \bar{x} + 2s) = (70{,}0815\ ;\ 129{,}7363) \approx (70;\ 130)$ al que corresponde el porcentaje es 96,6 %.

$(\bar{x} - 3s;\ \bar{x} + 3s) = (55{,}1678\ ;\ 144{,}65) \approx (55\ ;\ 145)$ al que corresponde el 99,3 %.

Resumiendo, las proporciones obtenidas son: 67,7 – 96,6 – 99,3. Si comparamos con la regla 68 – 95 – 99,7, podemos ver que los datos son aproximadamente normales.

6.7. 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 una de las variables que hemos obtenido en una cierta investigación o medición, el siguiente paso es saber cual es la distribución normal que mejor aproxima nuestros datos.

Recordemos que la curva normal viene definida por su media y desviación típica. Llamaremos *distribución observada* a los datos obtenidos en una muestra particular para la variable que estamos estudiando y *distribución normal teórica* la curva normal que usaremos para poder hacer inferencias sobre la variable en la población. Ajustar una curva normal a los datos es calcular los parámetros de la distribución normal teórica que mejor aproxima los datos, es decir, su media y desviación típica. Tomaremos, por tanto la distribución normal que tiene como media y desviación típica las que hemos observado en la muestra.

Cuando trabajamos con el programa Statgraphics podemos hacer el ajuste de una distribución normal a una variable dada de dos formas:

- **Gráficamente,** utilizando: DESCRIBE – NUMERIC DATA – DISTRIBUTION FITTING – eligiendo en el botón de opciones gráficas (GRAPHICS OPTIONS) – FREQUENCY HISTOGRAM. Este programa dibuja una curva normal superpuesta al histograma de frecuencias, eligiendo la media y desviación típica adecuada.

- **Analíticamente** por medio de: DESCRIBE – NUMERIC DATA – DISTRIBUTION FITTING –eligiendo el botón TABULAR OPTIONS – TAIL AREAS. Esta opción permite calcular el área bajo la curva para los datos menores o iguales que un determinado valor.

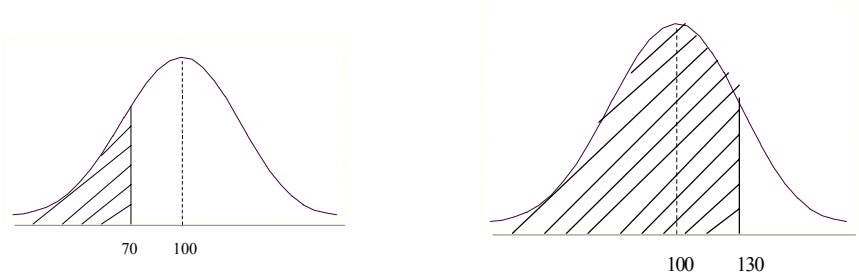
EJEMPLO 3 (continuación)

Utilicemos la opción TAIL AREAS para comparar si realmente la distribución normal teórica se aproxima a nuestros datos. Comprobemos, por ejemplo, el porcentaje de casos alrededor de la media en la distribución normal teórica y comparemos con el porcentaje de casos obtenido en los datos reales.

a) ¿Cuál es el área que en la distribución normal teórica ajustada al ejemplo 3 corresponde al intervalo (70; 130)? Puesto que la opción TAIL AREAS sólo calcula áreas bajo la curva hasta un cierto valor, para contestar esta pregunta debemos calcular el área inferior a 70 y el área

inferior a 130 y, luego realizar la diferencia de estas dos. Los resultados que nos entregará el programa, se corresponden con las situaciones gráficas presentadas en la figura 6.12.

Figura 6. 12. Áreas bajo la curva



La gráfica izquierda representa el área menor que 70, y la derecha representa el área menor que 130. Por medio del programa obtenemos los siguientes resultados:

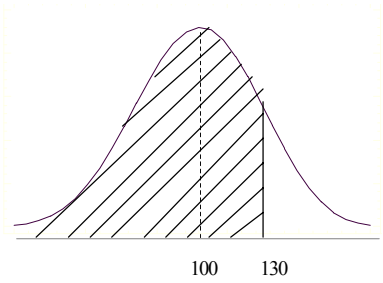
Tail Areas for COEF_INT
Area below 70,0 = 0,0224567
Area below 130,0=0,978177

Por lo tanto el área comprendida entre 70 y 130 es: 0,978188 – 0,0224567 = 0,9557313

b) *¿Cuál es la probabilidad de que una persona escogida al azar tenga un coeficiente intelectual en el intervalo (70; 130)?*La probabilidad de que una persona tenga un CI comprendido entre 70 y 130 viene dada por el área bajo la curva en el intervalo y es, por tanto: 95,55713 %, que se obtiene de multiplicar el área obtenida en el ítem a) por 100.

c)¿Cuál es la probabilidad de que se obtenga más de 140 puntos? En este caso, la figura 6.13. nos muestra el área bajo la curva para valores mayores que 140. El resultado que obtendríamos con el programa sería:

Figura 6.13. Area bajo la curva



Tail areas for COEF_INT
Area below 140,0 = 0,996408

Por lo tanto, el área que estamos buscando es 1 – 0,996408 = 0,003592. Y la probabilidad de que una persona tenga un coeficiente mayor que 140 es: 0,3592 %.

d)¿ Cuál es la probabilidad de que en esta curva se obtenga menos de 50 puntos? ¿Y más de 50 puntos? En el primer caso, podemos obtener el resultado directamente con el programa:

Tail areas for COEF_INT
Area below 50,0 = 0,0040935

La probabilidad buscada es**:** 0,40935 %. Como el área total bajo la curva es igual a 1, y queremos calcular el área para los valores mayores que 50, podemos hacer: 1 – 0,004092 = 0,995908. Por lo que obtenemos una proporción de 99,5908 %

NOTA: No debemos olvidar que los cálculos que hemos realizado se refieren al modelo teórico, y, por tanto, son valores aproximados respecto a los datos reales.

g) *Utilizando los datos de la tabla de frecuencias 5, calcula la proporción de personas de la muestra con un coeficiente intelectual menor de 50 puntos y con un CI mayor a 50 puntos ¿Crees que la curva normal da un ajuste aceptable a los datos?*

Debemos recordar que, ahora vamos a trabajar con la tabla de frecuencias de los datos reales, no con el modelo teórico de la distribución normal. La proporción de personas de la muestra con un CI menor que 50 es de 0,6 %. Este valor resulta de multiplicar el valor de la columna de frecuencias relativas acumuladas (Cumulative relative frequency) por 100.

Si observamos la columna de frecuencias acumuladas relativas (Cumulative Relative Frequency), vemos que el total es igual a 1. La proporción de personas con un CI mayor que 50 es, por lo tanto: 1 -0,006 = 0,994, es decir 99,4 %.

Observando los resultados obtenidos en el ejercicio anterior, podemos concluir que la curva normal realiza un ajuste bastante aproximado a los datos reales, ya que:

- En el modelo teórico de la distribución normal, la proporción de personas con CI mayor que 50 es: 99,59 %.
- En la distribución de datos obtenidos en una muestra real de personas, la proporción de personas con CI mayor que 50 es: 99,4 %.

h) Calcula la media μ y la desviación típica σ de la distribución normal que se ajusta a los datos de la muestra. Podemos calcular la media y desviación típica de la curva teórica que mejor se ajusta a los datos con la opción: DESCRIBE – NUMERIC DATA – DISTRIBUTION FITTING, de la que obtenemos

Analysis Summary
Data variable COEF_INT
1000 values ranging from 41,0 to 146, 0
Fitted normal distribution
Mean = 99.0551
Standard deviation = 15.3527

Si comparamos estos resultados, con la media y desviación típica en nuestra muestra, obtenidas en la tabla de frecuencias 5, mediante DESCRIBE- NUMERIC DATA- ONE VARIABLE ANALYSIS, vemos que los parámetros de la curva normal ajustada a los datos coinciden con los de nuestra muestra de datos. Esto es lógico, porque la distribución normal que mejor ajusta un conjunto de datos es la que tiene su misma media y desviación típica.

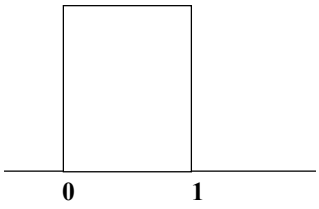
En nuestro ejemplo particular, el cálculo del porcentaje de observaciones en los intervalos centrales y los valores de los coeficientes de simetría y de curtosis, nos indica que podemos aproximar bastante bien la distribución real de datos por medio de la distribución normal asociada a ella. Esto no ocurre con todas las variables, ya que el modelo teórico sólo

se ajusta en algunos casos y por ello debemos hacer comprobaciones similares a las que hemos visto en este capítulo, para poder asegurar que se puede realizar tal aproximación

Actividades

6.12. La figura 6.14. 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 6.14



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?

6.13. 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?

6.14. 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?

6.15. 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 $\bar{x}$ 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.

6.8. 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 4: *Angel 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 6.5.

Tabla 6.5. Calificaciones en 5 asignaturas

(1)	*(2)*	*(3)*	*(4)*		*(5)*		*(6)*	
Examen	*Media de la clase*	*Desviación Típica de la clase*	*Puntuaciones (X)* Angel	Carlos	*Desviaciones a la media (x)* Angel	Carlos	*Puntuaciones tipificadas (Z)* Angel	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 6.5, podemos ver que Angel 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.

Algunas cuestiones a analizar

1. *Carlos está más cerca de la media en Inglés que Angel. ¿Y en las otras asignaturas?*
2. *Angel está 39,3 puntos por encima de la media en inglés y 16,2 puntos por encima de la media en informática (ver columna 5). ¿Es este estudiante mejor en Inglés que en informática?.*
3. *Carlos está 20,3 puntos por encima de la media en economía y 17,5 en informática. ¿Es este estudiante igualmente bueno en las dos materias?*
4. *¿Cómo podríamos comparar a los dos estudiantes?. Angel parece superior en tres materias y Carlos en las otras dos. Pero, suponiendo que los dos están compitiendo para una beca en la universidad;¿ cuál la merece más?.*
5. *La suma de todas las puntuaciones son 434 y 397, y favorece a Angel. ¿Sería justo dar la beca a Angel, ya que tiene mayor suma de puntuaciones?.*

Las preguntas 2, 3 y 5 tienen una respuesta negativa. 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. Para poder hacer comparaciones acertadas entre las diferentes calificaciones, se deben cumplir dos condiciones:

1. La calificación de los estudiantes de donde se extrajo estos datos debe tener igual media y dispersión en todas las asignaturas;

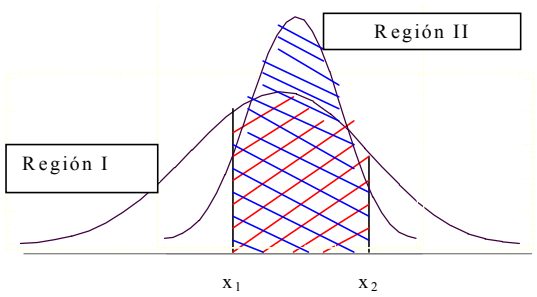
2. La forma de la distribución (curtosis y simetría), debe ser muy similar en la distribución de las calificaciones.

Las puntuaciones típicas pueden proporcionarnos la escala común que estamos buscando. Como hemos comentado al comienzo de esta sección, 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.* Para tipificar un valor, se resta a éste la media de la distribución y se divide por la desviación típica.

En resumen: *si x es una observación de una distribución que tiene media μ y desviación típica σ, el valor tipificado de x es:*

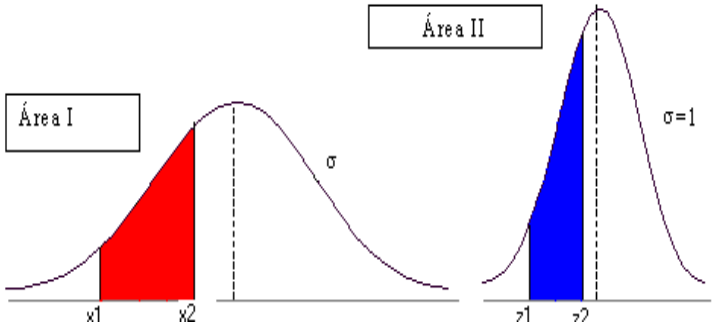
$$z = \frac{x - \mu}{\sigma}$$

Figura 18. Probabilidad en un intervalo



Cuando tipificamos una variable que tiene una distribución normal, obtenemos una nueva variable que tiene *distribución normal estándar o tipificada*. Una distribución normal típica tiene media 0 y desviación típica igual a 1.

Figura 19: Areas equivalentes en dos distribuciones normales



Sabemos que la gráfica de función de densidad de cualquier variable aleatoria continua es tal que el área bajo la curva limitada por los dos puntos $x = x_1$ y $x = x_2$ es igual a la probabilidad de que la variable aleatoria ξ asuma un valor entre $x = x_1$ y $x=x_2$. (Figura 18).

Puesto que la curva normal depende de la media y de la desviación típica, el área bajo la curva entre cualesquiera dos puntos debe, entonces, depender también de los valores μ y σ. Esto es evidente en la figura 18, donde se han sombreado las regiones correspondientes a $P(x_1 < \xi < x_2)$ para dos curvas con medias y desviaciones típicas diferentes.

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 19, 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 $z = z_2$

En síntesis, 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. Para comparar dos valores originales de variables diferentes, podemos pasar a los valores transformados, ya que se conservan las probabilidades, con la ventaja de poder ahora trabajar en una escala única. La principal razón para tipificar valores de una variable aleatoria es tener escalas comparables.

Tanto en educación como en psicología, es muy importante trabajar con medidas tipificadas, ya que no existen medidas absolutas sobre el comportamiento humano. Por lo tanto, debe tomarse un punto de referencia para poder dar una interpretación significativa a los resultados de los tests y exámenes; un punto de referencia muy importante es la media de una población de estudiantes. Además, si queremos tener en cuenta la variabilidad de la población, también usamos la desviación típica de las puntuaciones de los estudiantes.

Ejemplo 4 (continuación)

Observando la tabla 6, podemos extraer las conclusiones que necesitamos para decidir cuál estudiante seleccionar para la beca.

Los dos estudiantes representados en la tabla anterior pueden ser comparados en términos de sus puntuaciones tipificadas z (columna 6). Angel es superior a Carlos en inglés, matemáticas y física. En términos de las puntuaciones tipificadas encontramos que Angel es deficiente en economía e informática.

Comparando los dos estudiantes en términos de las puntuaciones originales, podríamos concluir que la asignatura que mejor lleva Angel es el inglés; en términos de las desviaciones, pensaríamos que es mejor en matemáticas, 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; respecto a las desviaciones, parece tener una calificación similar en economía e informática. 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 Angel 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, el estudiante Carlos debería ganar la beca.

Actividades

6.16. 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.

6.17. La puntuación total en probabilidad en el fichero TEST_P (que hemos usado en las clases prácticas), toma un valor medio de 25 puntos con desviación típica 6 ¿Qué puntuación en probabilidad corresponde a un alumno que tenga una puntuación tipificada de 1,2 y -1,5?

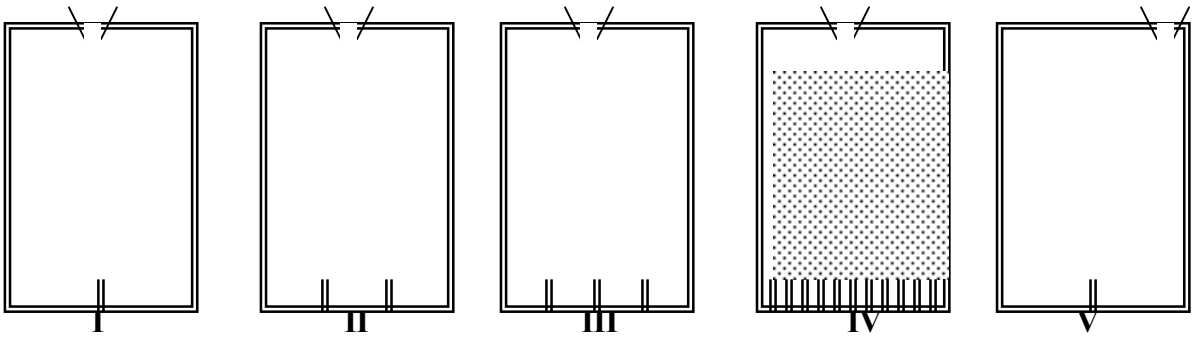
6.18.. ¿Cuál será la media y desviación típica de las puntuaciones tipificadas?

6.9. COMPRENSIÓN DE LA IDEA DE DISTRIBUCIÓN POR LOS NIÑOS

Uno de los problemas que estudian estos autores es la comprensión de la idea de distribución normal que se produce, por ejemplo cuando los granos de arena caen a través de un pequeño orificio (en un aparato de Galton o en un reloj de arena). Para Piaget, para comprender el mecanismo de esta distribución es preciso captar la simetría de las trayectorias posibles de los granos al caer por el orificio, porque hay las mismas posibilidades para cada grano de orientarse a derecha o izquierda.

Piaget indica que la representación matemática de dicha distribución corresponde a la curva de Gauss. Se emplea a menudo para hacer comprender su significado el aparato construido por Galton, el Quincux, que tiene forma de plano inclinado de madera donde se han colocado unos clavos regularmente espaciados. Al dejar caer algunas bolas por el orificio superior, se dispersan al azar, recogiéndose en unas casillas colocadas al final del plano inclinado. Las casillas centrales reciben más bolas que las extremas y la disposición es simétrica. Piaget utiliza un dispositivo semejante, pero simplificado en sus experimentos. Utiliza 5 aparatos, de los cuales cuatro tienen la abertura superior en la parte central y uno la tiene en el extremo superior derecho. Los aparatos se diferencian por el número de bifurcaciones (con 2, 3, 4 y un gran número de casillas finales en los aparatos 1 a 4 y 2 casillas finales en el aparato 5). (Ver figuras 3.1)

Figura 3.1. Esquema de los aparatos de Galton utilizados en la experiencia



En cada una de las cajas se comienza introduciendo una bola, después una segunda y tercera, preguntando al niño donde cree que va a caer y por qué. Una vez comprendida la experiencia se pide al niño que explique la forma que tomaría la distribución cuando se dejasen caer un gran número de bolas. Finalmente se dejan caer las bolas y se pide al niño que interprete la distribución obtenida.

El primer estadio (hasta 7 años) se caracteriza por la ausencia de la idea de distribución. En la caja I generalmente los niños hacen apuestas simples a favor de uno de los dos casilleros, pero sin la idea de igualación al aumentar el número de bolas. En la caja 2 el sujeto prevé bien una distribución igual en los tres casilleros o bien que todas las bolas irán a parar a uno de los casilleros. Con la caja III el niño apuesta por un de los casilleros centrales o por una distribución irregular. En la caja IV se espera en general una distribución irregular.

El segundo estadio (7 a 11 años) se caracteriza por un principio de distribución de conjunto generalizable, reconocible. En particular el sujeto prevé la desigualdad entre las frecuencias laterales y centrales. Pero esta distribución permanece insuficientemente cuantificada, falta de comprensión de la ley de los grandes números. Por ello, aunque hay una simetría global, no hay aún equivalencia entre los sectores correspondientes. En la caja I el sujeto prevé una igualdad aproximada entre los dos casilleros, pero sin que esta igualdad se consiga progresivamente con el número de bolas. En la caja II se prevé un máximo central, pero sin

igualdad en los casilleros laterales. La caja III da lugar a la previsión correcta de la ventaja de las casillas centrales, pero sin equivalencia entre ellas ni entre las laterales. La caja IV no provoca la previsión de una distribución simétrica regular, pero comienza a haber una generalización de las experiencias anteriores sobre la configuración de conjunto.

El tercer estadio (a partir de 12 años) está marcado por la cuantificación de la distribución de conjunto, es decir, por la previsión de una equivalencia entre las partes simétricas correspondientes de la dispersión. Esto es claro para las cajas 1 a II; la caja IV da lugar a ensayos de graduación hasta descubrir la distribución en forma de campana, el progreso más notable es la comprensión del papel de los grandes números en la regularidad de la distribución.

6.10. COMPRENSIÓN DE LA DISTRIBUCIÓN NORMAL Y TEOREMA CENTRAL DEL LÍMITE POR ESTUDIANTES UNIVERSITARIOS

El teorema central del límite es uno de los teoremas estadísticos más importantes que explica el uso generalizado de la distribución normal. Méndez (1991) estudia la comprensión de este teorema en alumnos universitarios, indicando cuatro propiedades básicas en la comprensión del teorema:

1. La media de la distribución muestral es igual a la media de la población, e igual a la media de una muestra cuando el tamaño de la muestra tiende al infinito.
2. La varianza de la distribución muestral es menor que la de la población (cuando N>1).
3. La forma de la distribución muestral tiende a ser acampanada a medida que se incrementa el tamaño muestral, y aproximadamente normal, independiente de la forma de la distribución en la población.
4. La forma de la distribución muestral crece en altura y decrece en dispersión a medida que el tamaño muestral crece.

Definiendo dos niveles de comprensión: El primer nivel esta formado por las habilidades y conocimientos que permiten resolver los ejercicios tipo presentados en los libros de texto. El segundo nivel representa la capacidad de aplicación a problemas reales. Observa que la comprensión del teorema se limita al nivel 1 y es excesivamente formal. Los estudiantes no son capaces de dar una explicación intuitiva de su significado o aplicarlo a casos reales.

- **Factores y objetivos aprendizaje**

Otras investigaciones estudian los factores que determinan el aprendizaje. Wilensky (1993, 1995, 1997), por ejemplo, muestra el influjo de los sentimientos y actitudes de los individuos frente a conceptos que conocen teóricamente y que manejan formalmente, pero cuyo significado no comprenden. Define la *ansiedad epistemológica*, como el sentimiento de descontento, confusión e indecisión que sienten la mayoría de los estudiantes frente a las distintas posibilidades o vías de resolución de un problema al no conocer los propósitos, orígenes o legitimidad de los objetos matemáticos que manipula y utiliza.

Esta *ansiedad* está reforzada por las prácticas de enseñanza empleadas en la clase de matemáticas y por la cultura matemática "protectora" que no promueve el diálogo entre docente y alumno y entre alumnos entre sí para ir construyendo el concepto.

Pensamos que hay muchas razones para que la *ansiedad epistemológica* sea particularmente pronunciada en el campo de la Estocástica. Todavía en la actualidad las nociones básicas de la teoría de probabilidad, como por ejemplo: aleatoriedad, distribución y probabilidad son bastante controvertidas. No existe un acuerdo generalizado sobre la interpretación de los conceptos de probabilidad o su aplicabilidad a casos particulares. Todo esto lleva a una confusión acerca de la aplicabilidad de la teoría de probabilidad a situaciones prácticas en nuestras vidas, lo que suscita dudas sobre su utilidad.

Wilensky concluye que la posibilidad de utilizar un medio informático que permita construir modelos de fenómenos probabilísticos que el alumno pueda revisar progresivamente ayuda a tener más seguridad y a reducir la ansiedad epistemológica. El autor enfatiza la importancia de la resolución de problemas, y principalmente sobre la proposición de problemas (o problem posing), ya que esta última permite proponer un problema de interés del propio alumno

Garfield (1981) indica los siguientes objetivos generales en el aprendizaje de la distribución normal eran los siguientes:

- Conocer las características de la distribución normal y de la distribución normal típica y usarla en la resolución de problemas
- Usar la aproximación de la normal a la binomial
- Comprender el teorema central del límite y construir distribuciones muestrales de poblaciones finitas.
- Responder a cuestiones de probabilidad transformando medias de variables normales a valores tipificados.

Además plantea objetivos específicos de razonamiento estadístico, clasificados de la siguiente forma:

Objetivos relacionados con la comprensión de los problemas

- Identificar términos estadísticos, tales como distribución normal y normal típica, factor de corrección de continuidad, variables aleatorias continuas y discretas, área bajo la curva, distribución muestral, error típico de la media.
- Identificar la información conocida y desconocida en el problema
- Identificar el tipo de problema. Ser capaz de determinar si los datos de un problema están relacionados con una variable normal o con los valores de medias muestrales, identificar si un problema puede ser resuelto utilizando la distribución normal o la corrección por continuidad para aproximar a la distribución binomial.

Objetivos relacionados con la planificación y la ejecución

- Seleccionar el procedimiento estadístico que debe utilizarse. Ser capaz de seleccionar un procedimiento para usar en la resolución de un problema.
- Aplicar un procedimiento estadístico a los datos. Ser capaz de realizar los cálculos que requiera el procedimiento seleccionado, entregando un conjunto de datos o un resumen estadístico.

Objetivos relacionados con la evaluación e interpretación

- Verificar una solución, comprobar que la solución tenga sentido. Ser capaz de comprobar si los valores z calculados son razonables en tamaño y signo, comprobar que las probabilidades no sean negativas, comprobar si las respuestas son razonables en función del contexto.
- Interpretar y responder estadísticamente. Ser capaz de dar conclusiones sobre la aproximación de probabilidades empíricas y teóricas, interpretar enunciados de probabilidad para una variable aleatoria y para una media muestral.

- Interpretar una respuesta en términos del contexto del problema. Ser capaz de determinar si una variable está distribuida normalmente, interpretar probabilidades o percentiles en relación con el contexto del problema.

Dificultad de comprensión de las puntuaciones tipificadas

Uno de los usos más comunes de la media y desviación típica es el cálculo de puntuaciones Z (o puntuaciones tipificadas). La mayoría de los estudiantes no tienen dificultad en comprender este concepto ni en calcular las puntuaciones Z para un conjunto de datos particular. Sin embargo hay dos concepciones erróneas ampliamente extendidas entre los estudiantes, referentes al rango de variación de las puntuaciones Z, cuando se calculan a partir de una muestra finita o una distribución uniforme.

Por un lado, algunos alumnos creen que todas las puntuaciones Z han de tomar un valor comprendido entre -3 y +3. Otros estudiantes piensan que no hay límite para los valores máximo y mínimo de las puntuaciones Z. Cada una de estas creencias está ligada a una concepción errónea sobre la distribución normal. Los alumnos que piensan que las puntuaciones Z siempre varían de -3 a +3, han usado frecuentemente una tabla o gráfico de la curva normal N(0,1) con este rango de variación. De igual modo, los estudiantes que creen que las puntuaciones Z no tienen límite superior ni inferior, han aprendido que las colas de la curva normal son asintóticas a la abscisa y hacen una generalización incorrecta.

Por ejemplo, si consideramos el número de niñas entre diez recién nacidos, obtenemos una variable aleatoria X que sigue la distribución binomial con n=10 y p=0.5. La media de esta variable es np=5 y la varianza npq=2.5. Por ello, la puntuación Z máxima que puede obtenerse en esta distribución es Z =3.16 que es un límite finito pero mayor que 3.

TEMA 7

MUESTREO Y ESTIMACIÓN

7. 1. MUESTRAS Y POBLACIONES

Actividades

7. 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 sinceros):

a) ¿Podemos decir que necesariamente de estos 100 votantes, 40 votaron al PP, 38 al PSOE y 9 a IU?

b) Supongamos que tomamos una 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?

7.2. Supón que quieres comprar un coche nuevo y quieres decidir entre la marca A y B. En una revista de automóviles encuentras una estudio estadístico sobre reparaciones efectuadas el último años 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 25.000 Ptas., 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.

Actividad 7.3. 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 .

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. 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.

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 7.4. 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.

Tipos de errores en un estudio por muestreo

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.

7.2. 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 7.1. 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.

Actividad 7.5. 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.

Actividad 7.6. 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.

Actividad 7.7. 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.

Obtención de muestras aleatorias de una distribución teórica con STATGRAPHICS

Ejemplo 7.2. 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 PLOT, y dentro de ella PROBABILITY DISTRIBUTIONS. Dentro de esta opción, tomaremos la Distribución Normal. En la pantalla TABULAR OPTIONS (Opciones tabulares), seleccionamos la opción Random Numbers, 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 SAVE (salvar). Se generan 100 números aleatorios de la distribución normal N(0,1). Si queremos otro tamaño de muestra podemos cambiarlo mediante PANE OPTIONS. Si queremos cambiar los parámetros de la distribución normal, podemos hacerlo mediante ANALYSIS OPTIONS.

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.

Distribución de la media en el muestreo

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 anterior, repetiremos el proceso 30 veces. Pinchando en el icono del DISKETTE y marcando la opción SAVE 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.... Muestra 30.

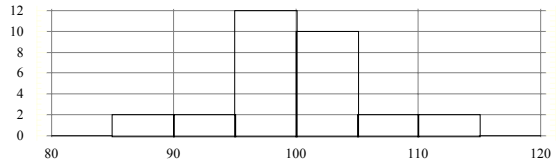
En la siguiente tabla presentamos los resultados.

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 DESCRIBE, NUMERIC DATA, MULTIPLE VARIABLE ANALYSIS hemos calculado la media de cada una de estas muestras. Obtenemos los siguientes valores:

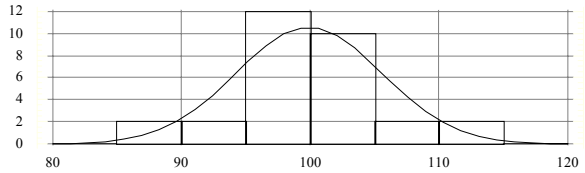
108, 94.5, 97.5, 110, 95, 97.25, 97.5, 92.5, 104.5, 98.75, 96.25, 101.75, 110.25, 101.25, 104, 103.25, 101, 102.25, 101.25, 95.75, 106.75, 98.75, 100, 95.75, 86.5, 90, 96.75 , 98.25, 104.5, 88.25

Actividad 7.8. ¿Cuántos valores del estadístico (media de la muestra) 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?.



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.



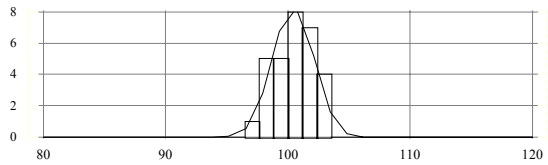
Veremos esto si tomamos ahora muestras aleatorias de 100 valores del coeficiente de inteligencia. Nosotros hemos usado el programa Statgraphics para simular 30 muestras, cada una con 100 valores del coeficiente de inteligencia y hemos calculado las medias de cada una de las 30 muestras. Estos son los 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 7.9. 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?

7.3. EL TEOREMA CENTRAL DEL LIMITE

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. La media de dicha distribución normal coincidiría con la media de la variable en la población y la desviación típica disminuye con el tamaño de la muestra. Aunque no lo demostraremos, matemáticamente puede comprobarse el siguiente resultado.

Distribución en el muestreo de la media de una variable aleatoria N (μ,σ).

Supongamos que tenemos en una población una variable aleatoria que sigue la distribución normal N(μ,σ). 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.

Por otro lado, también cuando realizamos la experiencia de obtener muestras del lanzamiento de dos dados, los valores más probables de las medias de estas muestras eran los cercanos a la media de la población (3.5), y la distribución de todas las medias de las muestras era simétrica, a pesar de que la distribución de partida no era normal. ¿Quiere decir esto que, si aumentásemos el tamaño de la muestra, llegaríamos a obtener una distribución normal de las medias de las muestras, aunque la variable de partida no fuese normal? Este es precisamente el resultado que establece el siguiente teorema.

Teorema central del límite

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,

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

cuando n es suficientemente grande, una distribución normal.

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).

Distribución muestral de la proporción

En particular 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:

$$N\left(\varphi, \sqrt{\frac{\varphi(1-\varphi)}{n}}\right)$$

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

$$\sqrt{\frac{\varphi(1-\varphi)}{n}}$$

La distribución es aproximadamente normal para muestras de suficiente tamaño.

Actividad 7.10. 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?

Actividad 7.11. 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?

7.4. INTERVALOS DE CONFIANZA

En las secciones 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 7.12. 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?

Intervalo de confianza para la proporción

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:

$$P(p - 2\sqrt{p(1-p)/n} \leq \varphi \leq p + 2\sqrt{p(1-p)/n}) = 0.95$$

Ejemplo 7.3. 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, $\sqrt{p(1-p)/n} = \sqrt{0.1*0.9/100}=0.03$. El intervalo de confianza del 95% será (0.1-0.03, 0.1+0.03)=(0.07, 0.13).

Actividad 7.13 ¿Entre qué valores podemos afirmar que se encontrará la proporción de personas favorables a la política económica del gobierno en la actividad 11 con una confianza del 95%?

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 7.14 ¿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 7.15. Discutir si el intervalo de confianza crece o decrece al aumentar cada uno de los factores anteriores.

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:

$$P(\bar{x} - 2\sigma/\sqrt{n} \le \mu \le \bar{x} \pm 2\sigma/\sqrt{n}) = 0.95 \text{ (Intervalo de confianza del 95\%)}$$

$$P(\bar{x} - 3\sigma/\sqrt{n} \le \mu \le \bar{x} \pm 3\sigma/\sqrt{n}) = 0.99 \text{ (Intervalo de confianza del 99\%)}$$

Ejemplo 7.4. En Statgraphics es posible calcular intervalos de confianza para las medias dentro de la opción DESCRIBE NUMERIC DATA ONE VARIABLE ANALYSIS (Opciones tabulares). 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 PANE OPTIONS.

```
Confidence Intervals for time

95.0% confidence interval for mean: 2.5 +/- 0.227724    [2.27228,2.72772]

95.0% confidence interval for standard deviation: [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

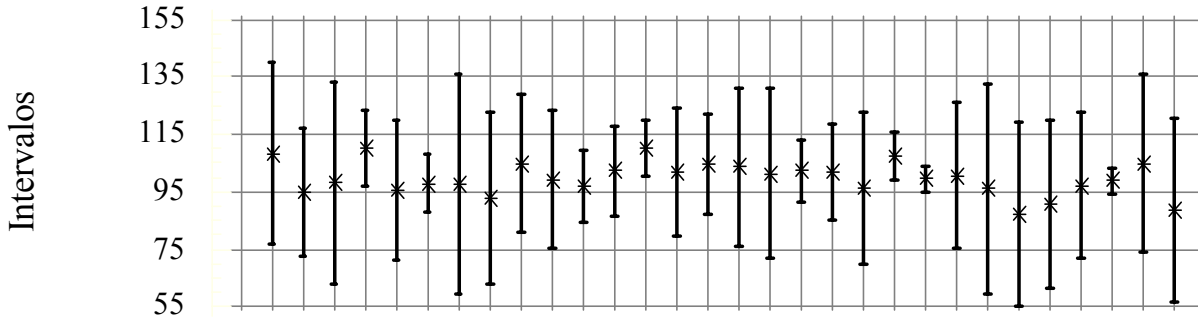
dice el coeficiente de confianza es que en un porcentaje dado de muestras, el verdadero valor del parámetro estará incluido en el intervalo.

Ejemplo 7.4.

Mediante el programa DESCRIBE, NUMERIC DATA, MULTIPLE VARIABLE ANALYSIS y tomando CONFIDENCE INTERVALS 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).

	Mean	Lower limit	Upper limit
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).



Problemas complementarios

7.16. 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).

7.17. 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?

7.5. Dificultades en la comprensión del Muestreo y la Inferencia

La idea central de la inferencia es que una muestra proporciona "alguna" información sobre la población y de este modo aumenta nuestro conocimiento sobre la misma. Se puede pensar en la inferencia estadística como una colección de métodos para aprender de la experiencia y, en la práctica, esto implica la posibilidad obtener intervalos de confianza para los parámetros de las poblaciones.

La comprensión de esta idea básica implica el equilibrio adecuado entre dos ideas aparentemente antagónicas: la representatividad muestral y la variabilidad muestral. La primera de estas ideas nos sugiere que la muestra tendrá a menudo características similares a las de la población, si ha sido elegida con las precauciones adecuadas. La segunda, el hecho de que no todas las muestras son iguales entre sí. El punto adecuado de equilibrio entre los extremos de información total e información nula respecto a la población es complejo, puesto que depende de tres factores: variabilidad de la población, tamaño de la muestra y coeficiente de confianza.

Los estudios sobre errores referidos al muestreo han tomado una gran importancia en el campo de la psicología, en el contexto de toma de decisiones. Un resumen de estos trabajos se presenta en Kahneman, Slovic, y Tversky (1982), quienes atribuyen estos errores al empleo de heurísticas en la resolución de problemas de decisión.

La heurística de la representatividad

La heurística de representatividad consiste en calcular la probabilidad de un suceso sobre la base de la representatividad del mismo respecto a la población de la que proviene. En esta heurística, se prescinde del tamaño de la muestra, y de la variabilidad del muestreo,

produciéndose una confianza indebida en las pequeñas muestras. Se supone que cada repetición del experimento, aunque sea limitada, ha de reproducir todas las características de la población. Por ejemplo, se espera que la frecuencia de una característica en la muestra coincida con la proporción de la misma en la población; por ello, tras una racha larga de aparición de un suceso se espera intuitivamente la aparición del suceso contrario, olvidando la independencia de los ensayos repetidos.

Actividades

7.18- La probabilidad de que nazca un varón es 1/2. A lo largo de un año completo, habrá más días en los cuales al menos el 60% de los nacimientos corresponden a varones:

a) en un hospital grande (100 nacimientos al día)

b) en un hospital pequeño (10 nacimientos al día)

c) no hay ninguna diferencia

7.19. Un taxi se ve implicado en un accidente de tráfico que ocurre de noche, en una ciudad con dos tipos de taxis de color verde y azul. Sabes los datos siguientes: a) 85% de los taxis son verdes y el resto azules; b) Un testigo que vio el accidente dice que el taxi era azul; c) El 80 % de los testigos hacen una identificación correcta de los datos del accidente. ¿Cuál es la probabilidad de que el taxi del accidente fuese azul?

7.20. Dos recipientes etiquetados como A y B se llenan con canicas rojas y azules en las siguientes proporciones:

Recipiente	Rojas	Azules
A	6	4
B	60	40

¿Cuál recipiente da más probabilidad de obtener una bola azul? Analiza el razonamiento que seguiría un alumno que eligiese el recipiente B. ¿Por qué la solución de este problema es diferente al del ítem de la actividad 1?

7.21. Se aplica una prueba de tuberculina en una población, conociéndose que el test da resultado positivo en 1 de cada 10.000 personas sanas en 99 de cada 100 personas enfermas. La incidencia de enfermos en la población es 3 de cada 100.000 personas. ¿Qué pensarías si te haces la prueba y resulta positiva? Analiza en este ejemplo la importancia del razonamiento estocástico correcto en la toma de decisiones.

La representatividad se suele usar para predecir sucesos, ya que, normalmente, los acontecimientos más probables son más representativos que los menos probables, o bien se sobreestima la correlación entre una causa y su efecto. Pero su uso inapropiado da lugar a diferentes sesgos en los juicios probabilísticos. Estos sesgos no son debidos a la no comprensión de las normas probabilistas o estadísticas, sino que incluso expertos en el tema llegan a cometerlos. Los sesgos más comunes que surgen de la utilización de esta heurística son:

i - Insensibilidad al tamaño de la muestra.

ii- Concepción errónea de las secuencias aleatorias.

iii - Intuiciones erróneas sobre las probabilidades de experimentos compuestos.

- Insensibilidad al tamaño de la muestra.

Según indican Tversky y Kahneman (1971) se hace una extensión indebida de la ley de los grandes números, creyendo en la existencia de una "Ley de los pequeños números", por la que se espera que incluso una muestra pequeña representante en todas sus características estadísticas de las poblaciones de donde procede.

- Concepciones erróneas de las secuencias aleatorias.

En un proceso aleatorio, se espera una serie corta de resultados represente fielmente el proceso. Por ello, secuencias relativamente ordenadas no parecen el resultado de un proceso aleatorio. Un evidente ejemplo se da entre los jugadores de la Loto, que mayoritariamente creen que los números han de salir sin un orden, error que también ha sido descrito en investigaciones con escolares. En este sentido, un error típico es la llamada falacia del jugador.

-Intuiciones erróneas sobre las probabilidades de experimentos compuestos.

Otro error típico de esta heurística se suele dar en el cálculo de la probabilidad de un suceso en un espacio muestral producto. Se puede llegar a creer que es más probable la intersección de dos sucesos que su unión. Así, suele pensarse que es más probable encontrar un loco asesino a encontrar un hombre que solamente sea loco o que solamente sea asesino.

Actividades

7.22. Analizar el enunciado del ítem siguiente y estudio las posibles interpretaciones erróneas del enunciado que pudieran llevar a un niño a dar como correcta la respuesta A

> **Item:** La probabilidad de que un niño nazca varón es aproximadamente 1/2. ¿Cuál de las siguientes secuencias de sexos es más probable que ocurra en seis nacimientos?
>
> a) VHHVHV; b) VHHHHV; c) las dos son igual de probables.

7.23. Si observamos los siguientes 10 nacimientos, ¿Qué te parece más probable?

a) La fracción de varones será mayor o igual a 7/10.

b) La fracción de varones será menor o igual a 3/10.

c) La fracción de varones estará comprendida entre 4/10 y 6/10.

d) Las tres opciones anteriores a), b), c) son igual de probables.

7.24. Cinco caras de un dado equilibrado se pintan de negro y una se pinta de blanco. Se lanza el dado seis veces, ¿Cuál de los siguientes resultados es más probable?

a) Cara negra en cinco lanzamientos y blanca en el otro

b) Cara negra en los seis lanzamientos

La heurística de la disponibilidad

Otra de las heurísticas usadas en el razonamiento probabilístico es la disponibilidad, por la que se juzgan más probables los sucesos más fáciles de recordar. Igual que con la representatividad, la presencia de errores derivados del uso de esta heurística se evalúa después de su utilización y no hay intento de predecir a priori bajo qué condiciones van a aparecer. Por ejemplo se suponen más peligroso viajar en avión que en coche, porque los accidentes de avión tienen más cobertura en la prensa.

Actividades

Una persona debe seleccionar comités a partir de un grupo de 10 personas (Cada persona puede formar parte de más de un comité).

a) Hay más comités distintos formados por 8 personas

b) Hay más comités distintos formados por 2 personas

c) Hay el mismo número de comités de 8 que de 2

Razona la respuesta y analiza el razonamiento que seguiría un alumno según la opción elegida.

Otro problema relacionado con el muestreo son los diferentes niveles de concreción de un mismo concepto en estadística descriptiva e inferencia. En la estadística descriptiva la unidad de análisis es una observación (una persona, un objeto) y calculamos la media x de una muestra de tales objetos. En inferencia, estamos interesados por obtener información de la media teórica o esperanza matemática $E(\xi)$ de la población de la que ha sido tomada la muestra dada.

Consideramos tal muestra como una observación de otra población diferente, la población de todas las posibles muestras de tamaño similar al dado, que podrían extraerse de la población de referencia. Hemos cambiado, en consecuencia, la unidad de análisis, que es ahora la muestra, y hablamos de que la media de la muestra es una variable aleatoria. Estudiamos la distribución de la media X en el muestreo y la media E(X) de esta variable aleatoria. Es preciso distinguir, por tanto, entre la media teórica en la población (que es una constante desconocida), la media particular obtenida en muestra; los posibles valores de las diferentes medias que se obtendrían en las diferentes muestras aleatorias de tamaño n (que es una variable aleatoria) y la media teórica de esta variable aleatoria, que coincide con la media de la población en el muestreo aleatorio. Esto supone una gran dificultad conceptual.

ANEXO A:

DESCRIPCIÓN DE FICHEROS DE DATOS

Fichero: MFF2O

DESCRIPCION:
Este fichero contiene información correspondiente a una investigación para estudiar la relación entre el carácter impulsivo y reflexivo de una muestra de niños y los errores en exactitud lectora. La información almacenada es la siguiente: número del alumno, grupo a que pertenece, sexo, tipo (impulsivo, reflexivo, normal), tiempo de latencia (obtenido en la aplicación del testo MFF-20), errores en sílabas, palabras y comprensión lectora.

VARIABLES:
grupo (1= grupo A; 2= grupo B); sexo (1= varón; 2= hembra); tipo (1= impulsivo; 2= reflexivo; 3= normal); tiempolat (tiempo de latencia, ##);errorsilab (errores en sílabas, ##1); errorpalab (errores en palabras, ##);
errorlectu (errores en comprensión lectora, ##)

fila	numero	grupo	*sexo*	tipo	tíempolat	errorsílab	errorpa1abra	errorlectu
1	1	1	1	1	6	10	4	4
2	2	1	2	1	10	16	8	14
3	3	1	1	2	15	0	0	2
4	4	1	1	2	15	4	2	0
5	5	1	2	1	4	15	3	6
6	6	1	2	2	18	0	0	0
7	7	1	2	2	14	1	2	1
8	8	1	1	2	31	2	1	6
9	9	1	2	1	12	0	0	0
10	10	1	2.	1	6	6	3	3
11	11	1	2	2	17	0	0	0
12	12	1	1	2	16	4	1	1
13	13	1	1	3	9	0	0	0
14	14	1	2	2	22	0	0	1
15	15	1	1	3	10	3	0	0
16	16	1	2	3	14	3.3	2	2
17	17	1	2	3	10	5	5	1
18	18	1	1	1	6	4	3	5
19	19	2	2	3	9	0	0	5
20	20	2	2	1	8	5	3	5
21	21	2	2	2	19	0	0	0
22	22	2	1	2	21	0	0	2
23	23	2	1	2.	9	7	4	0
24	24	2	2	2	78	0	0	0
25	25	2	2	2	14	0	0	.0

(58 registros)

Fichero: ALUMNOS

DESCRIPCIÓN:
Este fichero contiene información correspondiente a una encuesta realizada a nuestros alumnos para la realización de un trabajo práctico. Se les ha pedido que cumplimenten un cuestionario con los siguientes datos: sexo, si practican o no deporte (nada, poco, mucho), peso, altura, longitud de los brazos extendidos, número de calzado y 'cantidad de pesetas que llevaban en ese momento). Se han incluido variables cualitativas (sexo y deporte, cuantitativa discreta (número de calzado, pesetas) y cuantitativas continuas (peso, altura y longitud de brazos extendidos)
La finalidad de esta aplicación es exclusivamente didáctica, como ejemplo de recogida de datos estadísticos fácilmente disponibles en una clase de cualquier nivel, mediante los cuales se pueden mostrar los principales conceptos del análisis de datos estadísticos.

VARIABLES:
sexo (1 = varón; 2= hembra);
deporte (si practica o no deporte, 1= nada; 2 poco; 3 = mucho)
peso (en kg.); altura (en cm); longitud (longitud de brazos extendidos en cm)
calzado (número de calzado); pesetas (cantidad de pts en el bolsillo)

fila	sexo	deporte	peso	altura	longitud	calzado	pesetas
1	2	2	59	161	160	37	770
2	1	1	62	178	181	41	385
3	2	2	50	159	153	36	500
4	1	2	69	176	179	42	325
5	1	2	74	175	179	43	740
6	2	3	62	169	165	37	2600
7	2	2	56	162	158	36	250
8	2	2	58	162	163	37	225
9	2	1	52	170	171	38	501
10	1	2.	68	170	172	42	5450
11	1	3	72	184	185	43	7500
12	1	2	74	180	182	42	1785
13	1	2	66	175	177	41	0
14	2	2	60	170	168	38	200
15	2	1	60	165	161	38	4400
16	2	3	55	163	160	36	700
17	2	2	60	167	165	37	120
18	2	2	50	167	165	37	700
19	2	2	52	160	157	35	2016
20	2	1	53	164	160	37	875
21	2	2	58	163	166	38	285
22	2	2	74	175	178	40	560
23	2	2	63	173	180	39	3010
24	2	2	60	161	164	38	500
25	2	2	53	162	162	37	1000

(60 registros)

Fichero: NEURONAS

DESCRIPCIÓN:

Este conjunto de datos es una muestra de *135* neuronas de una especie de roedores d< las 3000 estudiadas en una investigación. La finalidad de dicho trabajo es obtener estimaciones del valor medio de las dimensiones de las distintas partes de las neuronas y determinar si existe alguna diferencia significativa entre los valores medios según la localización de las mismas. Se ha simplificado la estructura de la aplicación, suprimiendo algunas variables controladas. Los datos que presentamos representan medidas de la superficie, núcleo, nucleolo, y heterocromatina de las citadas neuronas. Se han considerado diversas localizaciones de las células en el cerebro: dorsal y ventral.

VARIABLES:

zona (1= dorsal; 2= ventral); supcelular (superficie celular, *###.##)*
supnucleo (superficie del núcleo, *##.##);*
supnucleol (superficie del nucleolo, *#.##)*
supheteroc (superficie de la heterocromatina, *#.#)*

fila	numero	zona	supcelular	supnucleo	supnucleol	supheteroc
1	1	1	66.99	47.39	2.10	2.10
2	2	1	99.10	51.69	4.10	1.93
3	3	1	77.60	52.75	3.12	2.93
4	4	1	71.78	46.01	5.14	0.55
5	5	1	67.18	38.07	3.74	0.78
6	6	1	54.27	34.28	3.02	3.03
7	7	1	73.71	34.56	1.79	0.79
8	8	1	59.27	37.24	2.83	1.70
9	9	1	48.56	33.67	2.72	1.43
10	10	1	39.52	23.45	1.16	1.01
11	11	1	47.55	31.04	2.14	1.39
12	12	1	37.40	25.76	2.57	1.82
13	13	1	64.00	37.80	2.89	1.57
14	14	1	50.20	33.41	1.29	0.78
15	15	1	59.91	32.17	2.65	2.68
16	16	1	74.77	40.07	1.57	0.71
17	17	1	83.35	49.97	0.73	0.15
18	18	1	53.28	31.19	1.40	1.22
19	19	1	48.72	28.71	2.52	1.79
20	20	1	45.59	26.41	3.13	0.94
21	21	1	85.70	55.08	2.68	2.28
22	22	1	48.38	28.18	1.25	1.86
23	23	1	79.89	46.02	2.72	1.70
24	24	1	47.64	29.72	3.10	2.00
25	25	1	50.01	31.78	2.82	1.99

(135 registros)

Fichero: TONO

DESCRIPCIÓN:
La operación de implante de lente intraocular, a pesar de sus grandes ventajas, se encuentra en la actualidad controvertida. Uno de los problemas que suele presentar consiste en el aumento de la tensión ocular del enfermo en los primeros días posteriores a la operación. Con
objeto de comprobar que la tensión ocular en estos enfermos se mantiene prácticamente constante, o incluso inferior, a la previa a la operación se realizó una prueba sobre 137 enfermos operados de cataratas cuyos datos se incluyen en esta aplicación.

VARIABLES:
tensionpre (tensión previa); tensionalt (tensión al alta);
tensionmes (tensión al mes de la operación)

fila	numero	edad	tensionpre	tensionalt	tensionmes
1	1	61	14	20	16
2	2	59	16	12	10
3	3	60	14	10	14
4	4	68	13	16	12
5	5	55	15	13	13
6	6	54	16	16	16
7	7	79	24	14	14
8	8	48	20	28	16
9	9	76	20	14	10
lO	10	50	15	10	16
11	11	58	17	17	16
12	12	64	18	16	16
13	13	83	19	16	5
14	14	58	15	12	12
15	15	66	18	18	32
16	16	70	16	20	14
17	17	58	16	20	20
18	18	64	12	12	12
19	19	69	16	12	12
20	20	60	19	20	18
21	21	64	17	12	20
22	22	56	15	27	23
23	23	72	12	20	18
24	24	56	16	14	20
25	25	61	16	10	12

(137 registros)

Fichero: TESTP

DESCRIPCIÓN:
Contiene datos relativos a los resultados de aplicar un test de evaluación de intuiciones probabilísticas primarias a niños del Ciclo Superior de EGB aplicado a tres colegios de Jaén. La finalidad el trabajo es doble:
a) Por una parte, conocer los errores y sesgos sobre intuición del azar y estimación de las probabilidades de sucesos sencillos, existentes en los niños de nuestro entorno sociocultural, así como relacionar dichos errores con las variables sexo, nivel y aptitud matemática. Se obtienen del test varias puntuaciones: puntuación combinatoria, verbal, probabilística y nivel probabilístico
b) Comparar los resultados de nuestra muestra piloto con las puntuaciones alcanzadas por niños ingleses en una investigación semejante.

VARIABLES:
colegio (1, 2, 3); sexo (1= varón; 2= hembra); curso (6º, 7º, 8º); aptitudmat (aptitud para las matemáticas, puntuación 0 a 10); puntcombin (puntuación en las preguntas sobre combinatoria, 1 a 6) puntverbal (puntuación en las preguntas sobre cuestiones de lenguaje probabilístico) puntprobab (puntuación en las preguntas de cálculo probabilístico)

fila	numero	colegio	sexo	curso	aptitudmat	puntcombin	puntverbal	puntprobab
1	1	1	1	7	8	4	12	21
2	2	1	1	7	9	4	11	19
3	3	1	1	7	5	O	10	14
4	4	1	1	7	8	4	10	19
5	5	1	1	7	6	5	10	17
6	6	1	1	7	8	4	11	16
7	7	1	1	7	4	2	6	20
8	8	1	1	7	5	4	8	13
9	9	1	1	7	6	2	11	17
10	10	1	1	7	8	3	12	16
11	11	1	1	7	9	4	12	22
12	12	1	1	7	4	3	11	14
13	13	1	1	7	2	3	7	10
14	14	1	1	7	8	4	7	17
15	15	1	1	7	4	3	12	15
16	16	1	1	7	4	3	5	8
17	17	1	1	7	6	3	13	11
18	18	1	1	7	10	5	13	25
19	19	1	1	7	6	4	12	19
20	20	1	1	7	8	5	11	20
21	21	1	1	7	3	2	10	17
22	22	1	1	7	8	4	11	21
23	23	1	1	7	4	3	11	12
24	24	1	1	7	4	5	11	14

(250 registros)

Fichero COLEGIOS

DESCRIPCIÓN

Este fichero contiene datos relativos a los colegios de enseñanza primaria y educación preescolar en la Provincia de Jaén en el curso escolar 1985-86. Estos datos fueron utilizados para la selección de una muestra de centros escolares, utilizando un muestreo en dos estapas con estratificación y definición de conglomerados artificiales, en una investigación sobre salud ocular de la población escolar. La técnica de muestreo empleada permitió controlar la variabilidad de las estimaciones y asegurar la aleatoriedad de la muestra de alumnos. El fichero que presentamos es una versión simplificada del original y contiene los siguientes datos:

Variable	Col.	Códigos
Código del centro	1 a 3	1 a 322
Tipo de centro	4	1: público; 2: privado
Zona	5	1: rural; 2: urbana
N. Unidades escolares	6-7	
N. Alumnos EGB	8-11	
N. Unidades escolares EGB	12-13	
N. Alumnos preescolar	14-16	
N. Unidades escolares preescolar	17-18	
N. profesores	19-20	

Número total de registros: 322; formato ascii

```
001 120100170101150001
002120100140101100001
003120100081000800001
004120100150101200001
005120100260100900001
006120100140101400001
007120100140101000001
008120100070101000001
009120100120100400001
010120100110100600001
011120100130100500001
012120100060100500001
013120100190100500001
014120100060100100001
015120100210100200001
```

Prácticas de Análisis de Datos

Práctica Nº 1:

TABLAS DE FRECUENCIAS Y GRÁFICOS DE VARIABLES CUALITATIVAS. REDACCIÓN DE INFORMES

- Una vez arrancado el programa Statgraphics, lee el fichero ACTITUD grabado en la última práctica, donde recogíamos los resultados del test de actitudes. Como recuerdas, en este cuestionario los alumnos puntuaban de 1 a 5 su grado de acuerdo con una serie de afirmaciones. El cuestionario se describe en la página 1-11 de los apuntes.

1. En la página 2-3 de los apuntes se explica como obtener tablas de frecuencia para datos cualitativos y variables discretas con Statgraphics. A partir de los datos contenidos en el fichero ACTITUD prepara tablas de frecuencias, diagramas de barras y diagramas de sectores de las siguientes variables estadísticas:
   a) Carrera que sigue el alumno. ¿Cuál es la carrera predominante?
   b) Curso en que se encuentra este año ¿Cuál es el curso predominante?

2. Completa los datos de la tabla 1, elaborando para ello una tabla de frecuencias de las puntuaciones dadas por los alumnos a los diferentes ítems del cuestionario de actitudes.

Tabla 1. Porcentaje de puntuación dada a cada ítem

Puntuación	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10

3. Si consultas en la copia del cuestionario incluida en los apuntes, verás que algunos de los ítems sobre actitudes puntúan en forma inversa al resto, es decir, cuanto mayor es la puntuación la actitud del alumno es más negativa. ¿Cuáles ítems puntúan en forma negativa? ¿Por qué?

4. En la página C-9 del anexo de los apuntes (guía para las clases prácticas) se explica el uso de la recodificación de datos. Lee las explicaciones y piensa cómo habría que recodificar los ítems que puntúan en sentido inverso para lograr que la puntuación tenga el mismo sentido que el resto de los ítems. Con la opción RECODE transforma los datos de estos ítems. Graba el fichero de datos una vez que estén recodificados, pulsando el icono del disquete.

5. Ahora que todos los ítems puntúan en el mismo sentido, podríamos definir para cada alumno un total en el test de actitudes. Define una nueva columna para la variable TOTAL en la que guardaremos la puntuación total del test para cada alumno. Consulta la página C-8 como se calculan nuevas variables y por medio de la opción generar datos de EDIT calcula la puntuación total en la prueba de cada alumno, almacenándola en esta variable. Graba el fichero de datos una vez obtenido.

Prácticas de Análisis de Datos

Práctica Nº 2:

DISTRIBUCIONES DE FRECUENCIAS Y GRÁFICOS DE VARIABLES NOMINALES Y DISCRETAS

Resolver las siguientes cuestiones referidas a datos del fichero MFF20 explicando los procedimientos usados y justificando su pertinencia así como la validez de las respuestas.

1. En la variable "errores en comprensión lectora" (errolectu),
a) ¿Qué porcentaje de niños no presenta errores?
b) ¿Qué porcentaje de niños presenta 10 o más errores?
c) ¿Cuál es el número más frecuente de errores?

2. ¿Son más impulsivos los niños, o las niñas?

3. ¿Son más impulsivos los niños (chicos y chicas) del grupo A o los del B?

4. Estudiar la variable "errores en palabras" (errorpalab). ¿Es mayor la proporción de alumnos sin errores entre los chicos o entre las chicas? ¿Es mayor la proporción de niños con errores entre los impulsivos que en el resto del grupo?

5. ¿Cuál es el número máximo de errores (palabras, sílabas y comprensión) para el 90 por ciento de los alumnos?

6. Enuncia y resuelve otras cuestiones que consideres de interés sobre el conjunto de datos MFF20.

Prácticas de Análisis de Datos

Práctica Nº 3:

DISTRIBUCIONES DE FRECUENCIAS Y GRÁFICOS PARA VARIABLES AGRUPADAS EN INTERVALOS

Una vez arrancado el programa Statgraphics, lee el fichero DEMOGRAFIA, (disco U, carpeta Datos) donde recogemos datos demográficos, PNB y población de 97 países. Repasa el significado de las variables del fichero en la página 1-9 de los documentos de trabajo.

En las páginas C11-13 del Anexo C se explica el funcionamiento del programa DESCRIBE- NUMERICAL DATA- ONE VARIABLE ANALYSIS . Recuerda que este programa prepara tablas de frecuencias y gráficos para variables agrupadas en intervalos, mientras que la opción DESCRIBE -CATEGORICAL DATA- TABULATION prepara tablas de frecuencias y gráficos para variables cualitativas y numéricas sin agrupar valores.

1. Prepara una tabla de frecuencias para la tasa de esperanza de vida en el hombre con intervalos de 10 años. Cambia las opciones por defecto si es necesario con PANE OPTIONS.

2. Copia la tabla de frecuencia en un fichero WORD y comenta las principales características que ves en la tabla: a) rango de variación de la esperanza de vida, b) valores más frecuentes e intervalo modal. Graba el fichero en tu carpeta incluyendo tu nombre al principio del texto.

3. Representa el polígono de frecuencias e inclúyelo en el fichero WORD (El polígono se obtiene en la opción histogramas, cambiando las opciones por defecto con PANE OPTIONS). Representa el gráfico del tronco *(* opciones numéricas, "stem and leaf"). ¿Te parece que la tasa de natalidad tiene una distribución simétrica? ¿Hay algún valor que te parezca atípico? ¿Por qué ? ¿Qué porcentaje de países tiene una esperanza de vida en el hombre menor que 60 años?

4. Prepara un polígono de frecuencias y gráfico del tronco para la esperanza de vida de la mujer, usando los mismos intervalos que hayas empleado en el punto 3. ¿Qué semejanzas y diferencias encuentras entre la distribución de esperanza de vida de mujeres y hombres en los distintos países?

5. Repite la tabla de frecuencia que has preparado para el punto 4, pero seleccionando solo los países africanos. Para ello, en la ventana de entrada de variables, debes indicar GRUPO=6 (podrías seleccionar otro grupo cambiando el número del código. ¿Cuáles son las principales diferencias en la esperanza de vida de la mujer de los países africanos respecto a la de todo el conjunto de países?

Prácticas de Análisis de Datos

Práctica Nº 4:

GRÁFICO DEL TRONCO Y PERCENTILES

---

1. Abre el programa Statgraphics y lee el fichero ACTITUD que se refiere al test de actitudes pasado en clase. Prepara un gráfico del tronco de la puntuación total de los alumnos en su actitud hacia la estadística, usando la opción STEM-AND LEAF DISPLAY en el programa DESCRIBE - NUMERIC DATA-ONE VARIABLE ANALYSIS. Copia el gráfico en un fichero WORD correspondiente a la práctica 4. ¿Hay algunos valores atípicos? ¿Te parece que la distribución de frecuencias de la variable 'puntuación total" es aproximadamente simétrica?

2. Calcula, a partir del gráfico del tronco la mediana y los cuartiles de la puntuación total. Comprueba los cálculos con la opción PERCENTILES dentro del mismo programa (pg. C-14 de los apuntes).

3. Con ayuda de la opción PERCENTILES haya la puntuación tal que el 80% de los alumnos tiene esa puntuación o menos? Si un alumno tiene una puntuación de 35 ¿Qué porcentaje de alumnos lo supera? ¿A qué percentil corresponde? Tienes para ello que modificar las opciones por defecto en PANE OPTIONS

4. Usando la opción SELECT en la ventana de entrada de variables selecciona ahora sólo los alumnos de pedagogía. Copia en el informe el gráfico del tronco para esta parte del fichero y comenta las diferencias que presenta respecto al global de los alumnos.

Prácticas de Análisis de Datos

Práctica Nº 5:

INTERPRETACIÓN DE ESTADÍSTICOS (CENTRALIZACIÓN, DISPERSIÓN Y FORMA). GRAFICO DE LA CAJA.

---

1. Abre el programa Statgraphics y lee el fichero ALUMNOS de la unidad U (carpeta datos). Este fichero contiene datos sobre altura, peso y otras características de un grupo de alumnos y viene descrito en los apuntes. Lee el significado de las variables de este fichero.

2) Calcula los resúmenes estadísticos de las variables *peso, altura* y *calzado.* Puedes obtenerlo de dos formas diferentes: a) usando tres veces sucesivas la opción SUMMARY STATISTICS en DESCRIBE- NUMERIC DATA- ONE VARIABLE ANALYSIS, dando en cada caso una de las variables; b) usanso una sóla vez la opción SUMMARY STATISTICS en DESCRIBE- NUMERIC DATA- MIJLTIPLE VARIABLE ANALYSIS. Este programa permite analizar varias variables de una vez. En la página C-14 de los apuntes viene la descripción de los dos métodos y el significado de los estadísticos calculados. Copia el resultado de tu análisis en un fichero WORD y guárdalo con el nombre de PRACTICA5 (escribe al comienzo del archivo tu nombre y el de tu compañero/a)..

3. ¿Cuáles son los valores de las medida de posición central (media, mediana y moda) para el peso, la altura y el calzado ? ¿Cuál sería, en tu opinión el peso, la talla y el número de calzado típico o representativo de estos alumnos?

4) ¿Crees que sería mejor separar varones y hembras en el cálculo de estos estadísticos, si queremos obtener valores representativos de los datos. ¿Por qué? Vuelve a repetir el cálculo de estadísticos separadamente, para chicos y para chicas. ¿Cuál es el peso, talla y número de calzado de la chica típica? ¿Y del chico típico?

5) Compara los valores de la varianza, desviación típica y coeficiente de variación de la variable *altura* para chicos y chicas. ¿Qué grupo presenta mayor dispersión? ¿Cuál presenta mayor dispersión respecto a la media?

6) Compara los coeficientes de asimetría para las variables *longitud* y *peso* ¿Qué distribución presenta mayor asimetría? ¿Es una asimetría positiva o negativa? Comprobar este resultado comparando histogramas de frecuencias y gráfico de la caja de ambas distribuciones.

7) Calcula la media, moda y la mediana del número de *pesetas* en el bolsillo. ¿Qué medida de tendencia central te parece, en este caso, más adecuada para representar el dinero que un alumno típico lleva en el bolsillo?.

8) Calcula los estadísticos necesarios para hacer un gráfico de la caja de la variable *longitud* y cópialos en el informe. Comprueba que has hecho bien los cálculos representando el gráfico de la caja de esta variable.

9) ¿Es cierto que en la muestra de alumnos del fichero ALUMNOS las mujeres tienen los brazos más cortos que los hombres y que también son más variables que los hombres en ese rasgos? ¿Hay algún alumno/a atípico/a en este rasgo? ¿Podrías identificarlos?

10) Construye el gráfico de la caja de la variable *pesetas,* y razona si hay valores atípicos.

Prácticas de Análisis de Datos

Práctica Nº 6:

DISTRIBUCIONES DE PROBABILIDAD. LA CURVA NORMAL (1)

- Abre el programa STATGRAPHICS y carga de la unidad U (carpeta DATOS) el fichero TESTP. Lee en los apuntes la descripción del fichero y el significado de las variables.
- En esta práctica y la que sigue vamos a analizar en este fichero algunas variables, para decidir si es o no adecuado ajustar una distribución normal a las mismas. En la página 6.14 y siguientes de los apuntes sobre la distribución normal se incluye un ejemplo, que puede servirte en la realización de esta práctica. Analizaremos sólo variables numéricas, ya que la distribución normal no se aplica a variables cualitativas.
- En esta práctica vamos a analizar la puntuación total en el test de probabilidad (fichero TESTP), para ver si la distribución normal sería una aproximación aceptable para esta variable. Estudiaremos, primeramente la forma de la distribución, para comparar con la esperada en una curva normal.

1. Con ayuda de la opción DESCRIBE; ONE VARIABLE NUMERICAL, prepara una tabla de frecuencias y un polígono de frecuencias relativas de la puntuación total en el test de probabilidad (ptotal). Estudia la forma del polígono de frecuencias relativas y la función de densidad (que puede obtenerse con la opción DENSITY TRACE). a) ¿Son el polígono y la función de densidad aproximadamente simétricos? b) ¿Tiene una o varias modas?

2. Mediante SUMMARY STATISTICS calcula los valores del coeficiente de asimetría y de curtosis y valores tipificados (si hace falta, usa PANE OPTIONS). ¿Son los valores obtenidos aceptables para una distribución normal?

3. Has estudiado en clase teórica la regla de los intervalos $\mu \pm \sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$, en una distribución normal. Utilizando la tabla de frecuencias, calcula el porcentaje de niños cuya puntuación total está incluida en el intervalo $x \pm s$, $x + 2s$ $x \pm 3s$. El valor de la media y desviación típica viene dado en la tabla de frecuencias que has construido anteriormente. Si es necesario, cambia la amplitud de los intervalos para que los extremos coincidan con los valores que te interesan. ¿Se cumple la regla 68- 95- 99?

4. Teniendo en cuenta los puntos 1, 2, 3 ¿Crees que la distribución de esta variable es aproximadamente normal?

Prácticas de Análisis de Datos

Práctica Nº 7:

DISTRIBUCIONES DE PROBABILIDAD. LA CURVA NORMAL (2)

- Ejecuta el programa STATGRAPHICS y carga de la unidad U (carpeta DATOS) el fichero TESTP.
- En esta práctica vamos a analizar la variable nivel probabilístico (nivel-proba), para ver si la distribución normal sería una aproximación aceptable para esta variable. Estudiaremos, primeramente la forma de la distribución, para comparar con la esperada en una curva normal.

1. Con ayuda de la opción DESCRIBE; ONE VARIABLE NUMERICAL, prepara una tabla de frecuencias y un polígono de frecuencias relativas del nivel probabilístico. a) ¿Son el polígono y la función de densidad aproximadamente simétricos? b) ¿Tiene una o varias modas?

2. Mediante SUMMARY STATISTICS calcula los valores del coeficiente de asimetría y de curtosis y valores tipificados (si hace falta, usa PANE OPTIONS). ¿Son los valores obtenidos aceptables para una distribución normal?

3. Has estudiado en clase teórica la regla de los intervalos $\mu \pm \sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$, en una distribución normal. Utilizando la tabla de frecuencias, calcula el porcentaje de niños cuyo nivel probabilístico está incluido en el intervalo x +s x ±2s x ±3s. ¿Se cumple la regla 68- 95- 99? Teniendo en cuenta los puntos 1, 2, 3 ¿Crees que la distribución de esta variable es aproximadamente normal?

4. Usa ahora la opción FITTING DISTRIBUTIONS, dentro de DESCRIBE NUMERIC VARIABLE ONE VARIABLE ANALYSIS ajusta una curva normal a los datos. ¿Cuáles son los parámetros de la curva normal ajustada?

5. Utilizando la option TAIL AREAS en este programa, que proporciona probabilidades para la curva normal ajustada ¿Cuál es la probabilidad de que, en esta curva teórica ajustada a los datos se obtenga un nivel igual o menor a 2?

6. Usando de nuevo la opción ONE VARIABLE NUMERICAL, y la tabla de frecuencias, calcula la proporción de niños en la muestra con nivel igual o menor a 2. ¿Crees que la curva normal da un ajuste aceptable de los datos?

Prácticas de Análisis de Datos

Práctica Nº 8:

LA CURVA NORMAL (3)

- Ejecuta el programa STATGRAPHICS y carga el fichero altu 1000 del disco U (carpeta datos). Este fichero contiene datos sobre la altura de 1000 chicas de entre 18 y 20 años que se pueden aproximar mediante una distribución normal.

1. Usando la opción DESCRIBE, NUMERIC DATA, DISTRIBUTION FITTING calcula la media y desviación típica de la distribución normal que aproximaría las alturas de las 1000 chicas

2. Usando la opción TAIL AREAS del programa, y, a partir de la curva normal ajustada, calcula la propobabilidad de chicas que, en la población donde se han tomado los datos tendrían más de 150, 160 y 170 cm de altura. ¿Qué proporción de chicas se encontrarían en los siguientes intervalos (160-170), (150-170)?

3. Mediante la opción CRITICAL VALUES, halla los valores de las alturas que corresponden a los cuartiles y a los percentiles del 10 y 90 por ciento en la curva normal ajustada.

- La opción PLOT sirve para representar gráficamente y calcular probabilidades de diversas distribuciones. Consulta en los apuntes la descripción de esta opción del programa. Ejecútala y selecciona la distribución normal. Cambia ANALYSIS OPTIONS para que el programa realice cálculos con una distribución normal ajustada a los datos de la muestra, es decir, con la media y desviación típica halladas en el punto
  1.

4. Representa gráficamente la función de densidad y función de distribución de la normal ajustada a la altura de las chicas de la muestra (DENSITY/MASS FUNCTION e INVERSE CDF). ¿Cómo piensas que cambiarían estas gráficas si, conservando la misma desviación típica aumentamos la media en 10 cm? ¿Y si disminuimos la media en 20 cm? Comprueba si has acertado en tu pronóstico, representando gráficamente las nuevas distribuciones y explica las diferencias encontradas.

5. ¿Cómo piensas que cambia la distribución normal si, conservando la misma media aumentamos o disminuimos la desviación típica? Comprueba tus pronósticos representando gráficamente 3 distribuciones normales de igual media y diferente desviación típica y explica las diferencias encontradas.

Prácticas de Análisis de Datos

Práctica Nº 9:

DISTRIBUCIONES MUESTRALES

- El propósito de esta práctica es estudiar la opción PLOT, y dentro de ella la opción RANDOM NUMBERS de PROBABILITY DISTRIBUTIONS para seleccionar muestras aleatorias de valores de una distribución de probabilidad teórica. En primer lugar vamos a simular la obtención de muestras de 100 lanzamientos de un dado, para comparar los resultados con los que obtuvimos en la clase de teoría al obtener muestras del lanzamiento de 2 dados.

1. Selecciona dentro de esta opción la Distribución discreta uniforme. Esta distribución produce al azar números equiprobables en un intervalo de valores. En la pantalla TABULAR OPTIONS (Opciones tabulares), selecciona la opción Random Numbers. Para simular 100 resultados aleatorios de un dado (en general de la distribución seleccionada), tienes primero que ajustar los parámetros, pinchando con el ratón en ANALYSIS OPTIONS. Luego basta pinchar con el ratón el icono del disco y marcar la opción SAVE (salvar). Se generan 100 números aleatorios de la distribución y se guardan en la variable RAND1, señalando para ello la opción salvar en la ventana de entrada de variables. Podrías si quieres tomar otro tamaño de muestra cambiando el número n en PANE OPTIONS.

2. Repite la simulación otras nueve veces, hasta obtener 10 muestras aleatorias, cada una con 100 valores obtenidos al lanzar un dado. En cada ejecución graba los datos en una nueva columna (RAND2, RAND3,... RANDIO) cambiando el nombre en la ventana de entrada de variables. Estas variables se van incorporando de manera automática en el fichero UNTITLED. Cuando termines con la opción SAVE DATA FILE graba este fichero de datos con el nombre MUESTRAS.

3. Vamos a calcular ahora la media de estas 10 muestras, para estudiar su variabilidad y comparar con la media en la población que es *3.5*. Con la opción DESCRIBFJ NUMERIC DATA-MULTIPLE VARIABLES ANALYSIS calcula la media de cada una de las 10 muestras obtenidas. Compara estos valores con la media teórica de la población y con las obtenidas en clase con muestras de tamaño 2. ¿Se acercan más a la media teórica las medias de la muestra de tamaño 2 o de tamaño 100? ¿Por qué?

4. En el Teorema central del límite ha estudiado que la media de una muestra sigue la distribución normal N (μ,σ/n), donde μ es la media de la población, σ la desviación típica de la población y n el tamaño de la muestra, si n es suficientemente grande. La media del número de puntos al lanzar un dado es 3.5 y la desviación típica 1.87. ¿Cuál será la media y desviación típica de la distribución normal que describe la variación de las medias en las muestras que has obtenido?

5. Usando la regla de los intervalos $\mu \pm K\sigma$ en una distribución normal, ¿Entre qué límites varía la media en el 95% las muestras de 100 lanzamientos de un dado? ¿Entre qué limites han variado las medias en las muestra obtenidas en tu simulación?

Prácticas de Análisis de Datos

Práctica Nº 10:

TABLAS DE CONTINGENCIA

- Lee el fichero ALUMNOS de la carpeta Datos en la Unidad U. Este fichero contiene datos de una muestra de alumnos y el significado de las variable se describe en los apuntes.

1. Con los datos del fichero ALUMNOS, prepara una tabla de contingencia para relacionar la práctica de deporte *(deporte)* con el sexo *(sexo),* usando la opción DESCRIBE, CATEGORICAL DAlA, CROSTABULATION. Identifica en la tabla las distribuciones marginales de las dos variables.

2. Construye ahora las distribuciones condicionales por filas y por columnas.
¿Son iguales las proporciones de los que practican deporte entre chicos y chicas? ¿Se puede decir que la práctica de deporte depende del sexo del alumno? Puedes basarte en la interpretación de las frecuencias por filas o por columnas y en el gráfico de mosaicos de la práctica de deporte respecto al sexo.

- Lee ahora el fichero TESTP de la carpeta datos. Este fichero contiene resultados sobre un test de intuiciones probabilísticas con varias puntuaciones parciales, pasado a una muestra de alumnos de primaria.

3 . En el conjunto de datos TESTP preparar una tabla de contingencia para relacionar el nivel probabilístico *(nivelproba)* con el sexo. ¿se observa una variación del nivel probabilístico en niños y niñas?

4. Prepara ahora una tabla de contingencia que relacione el nivel probabilístico con el curso. ¿Se puede decir que el nivel probabilístico varía significativamente con el curso?

5 ¿Se puede afirmar, apoyándose en los datos del fichero TESTP, que el nivel probabilístico *(nivelproba)* depende de la puntuación en combinatoria? (puntcombi).

Prácticas de Análisis de Datos

Práctica Nº 11:

REGRESIÓN Y CORRELACIÓN BIVARIANTE

- Abre el fichero DEMOGRAFÍA
- Ejecuta el programa RELATE; SIMPLE REGRESSION de Statgraphics

El objetivo de esta práctica es estudiar qué tipo de relación hay entre la variable lASA DE NATALIDAD y otras variables del fichero.

1. Representa gráficamente la tasa de natalidad en función de la tasa de mortalidad. ¿Es la relación directa o inversa? ¿Es lineal o no?

2. ¿Cuál es la ecuación de la recta de regresión que permite calcular la tasa de natalidad en función de la tasa de mortalidad?

3. ¿Cuál es el valor del coeficiente de correlación? ¿Cuál es la proporción de la varianza de la tasa de natalidad explicada por la tasa de mortalidad?

4. Representa ahora la tasa de natalidad en función de otras variables (esperanza de vida del hombre; esperanza de vida de la mujer, PNB, mortalidad infantil).
¿En cuál variable la relación es más intensa? (No copies los gráficos, sino sólo justifica por qué).

5. ¿Hay alguna variable respecto a la cuál la relación sea inversa? ¿Por qué?

6. ¿Hay alguna variable respecto a la cual la relación es no lineal?

PRÁCTICA 12. SIMULACIÓN DE EXPERIMENTOS ALEATORIOS.

- El propósito de esta práctica es usar el programa PLOT (Probability distributions) para realizar simulaciones de experimentos aleatorios. Este programa permite obtener datos al azar de diferentes modelos teóricos de distribuciones de probabilidad.

1. *La distribución uniforme discreta.* Abre PLOT, Probability Distributions y elige la opción Discrete Uniform. Esta opción permite obtener números al azar entre dos valores dados, de tal modo que todos tienen igual probabilidad de salir. Por defecto elige valores entre 0 y 2. Puedes cambiar estos valores con PANE OPTIONS.

Usa PANE OPTIONS para añadir una segunda distribución que tome valores entre 1 y 6 (sería como obtener los valores de lanzar un dado). Modifica la definición de las variables para que sean enteros (sin decimales)

Representa gráficamente la distribución de probabilidad y la función de distribución acumulada de estas dos variables uniforme discretas (con valores entre 0 y 2; con valores entre 1 y 6) y comenta las diferencias que observas.

2. *Simulación de lanzamientos de dados.* Vamos a resolver el siguiente problema mediante simulación:

Problema 1. 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?

Simula el experimento de lanzar 2 dados con la opción RANDOM NUMBERS, del icono de opciones numéricas. Para ello, cambia en PANE OPTIONS el valor "size" o tamaño de la muestra de números aleatorios que quieres generar a 2. Pulsando el icono del diskette (en la barra de menus del Statgraphics) aparece la ventana SAVE RESULTS OPTIONS. Marca Random numbers for Dist 1 y se grabarán los resultados de la simulación en la variable RAND1. Comprueba si los dos números que has obtenidos son pares, impares o un par /impar.

Repite la simulación 10 veces (grabando los resultados en variables diferentes y completa la tabla siguiente:

Frecuencias de resultados al simular el lanzamiento de dos dados

Dos números pares		Dos números impares		Par /impar	
Frec Ab.	Frec. Rel	Frec Ab.	Frec. Rel	Frec Ab.	Frec. Rel

Compara con los compañeros de tu fila o los que se sientan delante/ detrás.

a) ¿Puedes ahora resolver el problema 1?
b) ¿Concuerdan estos resultados experimentales con tus previsiones sobre este experimento aleatorio?
c) ¿A qué crees que se deben las diferencias entre las frecuencias observadas y las esperadas calculadas estas bajo la hipótesis probabilística de equiprobabilidad de los sucesos elementales al lanzar un dado equilibrado?

3. *Simulación de nacimientos en un hospital*. Piensa cómo puedes usar la distribución discreta uniforme para simular el nacimiento de niños y niñas y resolver el siguiente problema.

Problema 2. Supongamos que la mitad de todos los recién nacidos son niñas y la otra mitad niños en la población en general. El hospital A registra un promedio de 100 nacimientos al día y el hospital B un promedio de 10 nacimientos al día. En un día particular, ¿cuál de los dos hospitales es más probable que registre un 70% o más nacimientos de niñas?
____a. El hospital A (100 nacimientos al día)
____b. El hospital B (10 nacimientos al día)
____c. Los dos hospitales tienen igual posibilidad de registrar este suceso.

Simula los nacimientos del hospital A durante 10 días y graba los resultados en 10 variables. Usa la opción DESCRIBE para producir la tabla de frecuencias de niños y niñas en cada simulación y completa la tabla siguiente:

Frecuencias de resultados al simular nacimientos durante 10 días en el hospital A

	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7	Dia 8	Dia 9	Dia 10
Niños										
Niñas										

¿Cuál es el porcentaje de dias en los que el número de nacimientos de chicas es de 70 o más? ____ %

Simula los nacimientos del hospital B durante 10 días y completa la tabla:

Frecuencias de resultados al simular nacimientos durante 10 días en el hospital B

	Dia 1	Dia 2	Dia 3	Dia 4	Dia 5	Dia 6	Dia 7	Dia 8	Dia 9	Dia 10
Niños										
Niñas										

¿Cuál es el porcentaje de dias en los que el número de nacimientos de chicas es de 7 o más? ____ %

Compara con los compañeros de tu fila o los que se sientan delante/ detrás. ¿Puedes ahora resolver el problema 1?

4. Define qué es la "distribución binomial de probabilidad" (consulta en los apuntes del curso, tema 5). ¿Cuántos parámetros intervienen en su definición? ¿Qué significan? ¿Qué expresión matemática permite calcular las expresiones binomiales?

5. *Resolución del problema 2 (hospitales) mediante un modelo teórico: Distribución Binomial.* Usa la distribución binomial en PLOT (Probability distributions) para resolver el problema 2, comparando los gráficos de probabilidad y distribución de la distribución Binomial con (p= 0.5 y n= 10; hospital B) y (p= 0.5 y n= 100; hospital A).

Prácticas de Análisis de Datos

Práctica Nº 13:

ESTUDIO DESCRIPTIVO DE VARIABLES ESTADÍSTICAS SOBRE INTUICIONES ACERCA DE LAS SECUENCIAS ALEATORIAS

El fichero MONEDAS contiene datos sobre seis variables estadísticas correspondientes a un proyecto de estudio de las intuiciones de las personas acerca de la distribución de secuencias aleatorias en el experimento de lanzar una moneda 20 veces.

En dicho experimento hemos pedido a un grupo de estudiantes simular los resultados del lanzamiento de una moneda 20 veces. Esta es, por ejemplo, una de las secuencias previstas:
CC+C+++CC+CCC++C+CC+

A continuación se ha pedido hacer realmente el experimento y anotar la secuencia de caras y cruces obtenidas. Esta es, por ejemplo, una de las secuencias obtenidas:
CCC++CCC+++++CCCC++C

Las seis variables registradas en el fichero MONEDAS son las siguientes:

carassimu:	Número de caras en la serie de datos simulados
nrachasimu:	Número de rachas en la serie de datos simulados
lrachasimu:	Longitud de la racha más larga en la serie simulada
carasreal:	Número de caras en la serie de datos reales
nrachasrea:	Número de rachas en los datos reales
lrachasrea:	Longitud de la racha más larga en los datos reales

Cuestiones:

La pregunta de investigación se formula de la siguiente manera:

¿Son correctas las intuiciones de las personas sobre el comportamiento de las secuencias aleatorias en el lanzamiento de una moneda? ¿Qué tipo de sesgos existen?

O de manera equivalente,

¿Es posible discriminar, mediante el análisis estadístico de las series SIMULADAS, qué grupo de series son las simuladas o las reales?

Utiliza los datos de las seis variables del fichero MONEDAS para responder a estas cuestiones. Aplica las técnicas de análisis de datos que consideres pertinentes.

BIBLIOGRAFÍA

Batanero, C. (1998). Recursos para la educación estadística en Internet. *UNO, 15,* 13-26.

Batanero, C. (1999). Taller sobre analisis exploratorio de datos en la enseñanza secundaria. *Actas de la Conferencia Internacional "Experiências e Expectativas do Ensino de Estatística - Desafios para o Século XXI"*. Florianópolis, Santa Catarina, Brasil - 20 a 23 de Setembro de 1999.

Batanero, C. (1998). Recursos en Internet para la Educación Estadística. *UNO, 15,* 13-25.

Batanero, C. (2000). Significado y comprensión de las medidas de tendencia central. *UNO,* 25, 41-58.

Batanero, C. (2000). Controversies around significance tests. *Mathematical Thinking and Learning, 2*(1-2), 75-98.

Batanero, C., Estepa, A. y Godino, J. D. (1991). Análisis exploratorio de datos: sus posibilidades en la enseñanza secundaria. *Suma, nº 9, 1991: 25-31.*

Batanero, C., Godino, J. D. Green, D. R., Holmes, P. y Vallecillos, A. (1994). Errores y dificultades en la comprensión de los conceptos estadísticos elementales. [Errors and difficulties in understanding elementary statistical concepts. *Internation Journal of Mathematics Education in Science and Technology,* 25(4), 527-547]

Batanero, C. y Serrano, L. (1995). Aleatoriedad, sus significados e implicaciones educativas. *UNO,* 15-28.

Cabriá, S. (1994). *Filosofía de la estadística.* Servicio de Publicaciones de la Universidad de Valencia.

Cobo, B. y Batanero, C. (2000). La mediana en la educación secundaria obligatoria: ¿Un concepto sencillo? *UNO* 23, 85-96.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children.* Dordrecht: Reidel.

Godino, J. D. (1995). ¿Qué aportan los ordenadores al aprendizaje y la enseñanza de la estadística? *UNO,* 5, 45-56.

Godino, J. D., Batanero, C. y Cañizares, M. J. (1987). *Azar y probabilidad. Fundamentos didácticos y propuestas curriculares.* Madrid: Síntesis.

Godino,J. D., Batanero, C., Cañizares, M. J. y Vallecillos, A. (1998). Recursos para el estudio de los fenómenos estocásticos.

Moore, D. S. (1998). *Estadística aplicada básica*. Barcelona: Antoni Bosch, editor.

Sánchez-Cobo, F. T., Estepa, A. y Batanero, C. (2000). Actividades de traducción en la estimación de la correlación. *Enseñanza de las Ciencias,* 18(2). 297-310.

Spiegel, M. R (1990). *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.

Vallecillos, A., y Batanero, C. (1997). Aprendizaje y enseñanza del contraste de hipótesis: Concepciones y errores. *Enseñanza de las Ciencias, 15* (2), 180-197.