FUNDAMENTOS CONCEPTUALES DE LAS PRINCIPALES PRUEBAS DE SIGNIFICACIÓN ESTADÍSTICA EN EL AMBITO EDUCATIVO

AUTORES:

Clemente Rodríguez Sabiote
José Gutiérrez Pérez
Teresa Pozo Llorente

© Los autores
Edita: Grupo Editorial Universitario
ISBN:
Depósito Legal:
Imprime: Lozano Impresores S.L.L.
Distribuye: Grupo Editorial Universitario
Telf: 958 800580 Fax: 958 291615
http://www.editorial-geu.com
E-mail: grupoeditorial@terra.es

INDICE

PRIMERA PARTE: Bases conceptuales de los principales contrastes de hipótesis

El método científico se caracteriza por:
a) una cuidadosa y precisa clasificación de hechos
y observaciones, debidamente relacionadas e hiladas,
b) el descubrimiento de leyes científicas por
medio de la imaginación creativa,
c) el sentido de la autocrítica, como piedra angular de la
potencialidad innata que poseen todas las mentes para ejercitarla
(Pearson, 1911)

1. INTRODUCCIÓN

La Estadística es una herramienta de reconocido valor al servicio de la investigación educativa. Mediante los procedimientos estadísticos se construyen modelos de interpretación del mundo, de los fenómenos naturales, sociales y también educativos, basándonos en los principios en que se asienta la lógica de su método y las reglas de su lenguaje. La Estadística trata de formular generalizaciones con el mayor alcance posible dentro unos márgenes de error acordados. En cuanto a la lógica de su método cabe afirmar que las cosas no son verdaderas o falsas de forma categórica, pues nunca podemos afirmar con rotundidad que nuestras conclusiones obtenidas en una investigación son correctas al 100%; hacemos afirmaciones y sacamos conclusiones tras el análisis de los datos asumiendo el riesgo de equivocarnos 1 de cada 100 veces o 1 de cada 20 veces. Evidentemente la posibilidad de equivocarse en el primer caso es menor que en el segundo, la tolerancia del error es mayor a medida que nuestras conclusiones cabe la posibilidad de que sean falsas con un mayor margen de error.

En cuanto a las reglas de su lenguaje, decimos que las cosas son probablemente ciertas y, por tanto, las conclusiones a las que se llega en las investigaciones se basan en afirmaciones probabilísticas y nunca en verdades taxativas. Estas afirmaciones permiten establecer generalizaciones, asumiendo un riesgo de error más tolerante o más exigente, sobre el total de la población objeto de estudio; aun a sabiendas de que los datos se han recogido no sobre el total sino sobre muestras representativas del total a las que hemos aplicado las reglas y que con las licencias del lenguaje hacemos inferencias y decimos que es muy probable que esas afirmaciones sean válidas también para toda la población.

La Estadística Inferencial aplica estos modelos de razonamiento lógico y estas reglas de lenguaje para formular conclusiones generalizables a la población mediante el establecimiento de conjeturas e hipótesis basadas en pruebas matemáticas de distinta naturaleza que hacen posible aceptarlas como afirmaciones válidas dentro de unos márgenes de confianza.

Cuando el interés del análisis de datos se centra en la verificación de hipótesis a fin de denotar posibles efectos diferenciales y, además, en las relaciones de interdependencia y dependencia que mantienen un conjunto de variables podemos hablar de análisis confirmatorio. Precisamente uno y otro objetivo dan lugar a los dos grandes bloques del análisis confirmatorio: análisis inferencial y multivariante. En este cuaderno de tratarán sólo diversos aspectos del primero (análisis inferencial) haciendo hincapié en las principales pruebas de tipo paramétrico y no paramétrico. Por tanto, se trata de un material de apoyo que complementa, por una parte los contenidos de análisis de datos que se desarrollan de manera manual y, por otra, el primer cuaderno dedicado al análisis descriptivo.

2. ANÁLISIS INFERENCIAL

2.1. *Aspectos generales del contraste de hipótesis*

El propósito fundamental de los análisis estadísticos de tipo inferencial es el conocimiento de poblaciones a partir del estudio de muestras o subconjuntos representativos y suficientes de dichas poblaciones (Fernández, García, Fuentes y Asensio, 1992; Tejedor y Etxeberria, 2006).

En este tipo de análisis, amén de la *probabilidad,* el *muestreo* y la *estimación de parámetros* (de tipo puntual e interval) destaca el importante papel que juegan los *contrastes de hipótesis*. Para Martín (2001:9) los términos “contraste de hipótesis” son sinónimos de “pruebas de significación” y se refieren a los procedimientos estadísticos mediante los cuales aceptamos o rechazamos una hipótesis nula (Ho) lo que automáticamente nos habilita para rechazar o aceptar otra hipótesis denominada hipótesis alternativa (H1). Mientras la primera postula la ausencia de diferencias estadísticamente entre dos medidas o más (las que existen se deben al azar), la segunda postula todo lo contrario, o sea, la existencia de diferencias estadísticamente significativas entre dos o más medidas.

Teniendo en cuenta que una hipótesis es una relación potencial entre dos o más variables, por ejemplo, "*los alumnos que utilizan calculadora de bolsillo obtienen un mayor rendimiento académico que aquellos que no lo hacen*", podemos afirmar que el contraste de hipótesis es el procedimiento que nos permite verificar y confirmar si esa relación potencial es verdadera o no y con qué margen de error. A continuación explicitamos otro ejemplo más actual para tratar de comprender mejor el concepto de contraste de hipótesis.

Por ejemplo, un estudiante de doctorado ha recogido datos sobre el fenómeno del "botellón" en Granada. Le interesa, tomando como objeto de estudio varias Facultades, determinar el grado de acuerdo que tienen los estudiantes universitarios sobre este tipo de eventos tomando en consideración diferentes variables de agrupación: titulación, curso… En este caso, resultaría útil comparar si la variable opinión sobre este tipo de eventos resulta más o menos importante dependiendo de la titulación a la que pertenece el alumnado, el curso en el que está…

Foto extraída de: www.ideal.es/granada/Media/200703/05/botellon

Siguiendo un procedimiento sistemático de muestreo en las diferentes Facultades, el estudiante de doctorado podrá obtener unas conclusiones sobre la participación y la opinión que podrá generalizar al conjunto de la población estudiantil de universitarios granadinos. Para ello deberá establecer una serie de hipótesis que finalmente le han de permitir establecer sus conclusiones con un mayor grado de precisión.

En términos generales, sus formulaciones hipotéticas sobre el botellón pueden establecerse en los términos siguientes: Los universitarios están a favor del botellón, los universitarios están en contra del botellón. Pero los matices sobre esta formulación pueden expresarse con más precisión: Hay diferencias entre los alumnos de la Facultad A y los de la Facultad B. De igual forma se pueden establecer aproximaciones al tema teniendo en cuenta la variable sexo: No hay diferencias en la opinión de los estudiantes y estudiantas universitarios sobre el fenómeno botellón. O bien en función de la variable curso, lugar de residencia (Granada, no Granada; pueblo, ciudad).

Un análisis exhaustivo de esta serie de variables nos permitiría entender el fenómeno más ampliamente, aunque para generalizar los resultados a toda la población universitaria deberíamos contar con una muestra de todo el país.

Por otra parte, existe un consenso más o menos extendido a la hora de formular dichas hipótesis, que se resume de la siguiente forma.

Hipótesis nula (Ho): no se postulan diferencias estadísticamente significativas entre las medidas y si las hay se deben al azar.

- Bilateral:
  Contraste con una muestra ⇒ (Ho): $\mu = k$
  Contraste con más de una muestra ⇒ (Ho): $\mu1 = \mu2$ o también $\mu1 - \mu2 = 0$

- Unilateral:
  Contraste con una muestra ⇒ (Ho): $\mu \geq k$ o también (Ho): $\mu \leq k$
  Contraste con más de una muestra ⇒ Ho): $\mu1 \geq \mu2$ y también H1): $\mu1 \leq \mu2$

Hipótesis alternativa (H1):

- Bilateral: se postulan diferencias estadísticamente significativas entre las medidas, pero se desconoce a favor de cual de ellas:
  Contraste con una muestra ⇒ (H1): $\mu \neq k$
  Contraste con más de una muestra ⇒ (H1): $\mu1 \neq \mu2$ o también $\mu1 - \mu2 \neq 0$

- Unilateral: se postulan diferencias estadísticamente significativas entre las medidas y, además, se contemplan a favor de alguna de ellas:
  Contraste con una muestra: (H1): $\mu < k$ o también (H1): $\mu > k$
  Contraste con más de una muestra ⇒ H1): $\mu1 < \mu2$ o también $\mu1 > \mu2$

Cuadro 1. Formulación convencional de las hipótesis estadísticas

En el desarrollo del proceso de contraste de hipótesis es igualmente relevante el nivel de significación, error tipo I o α (rechazar una Ho cuanto ésta es verdadera) que se asume en el mismo, así como el nivel de confianza (1-α). Los valores habituales asumidos para los errores son el 10%, menos usado y, sobre todo, 5% y 1%, siendo por ende, los niveles de confianza del 90%, 95% y 99%.

También destacamos el error tipo II o β (no rechazar una Ho cuando ésta es falsa) y la potencia de contraste (1-β). Todos estos aspectos quedan esquemáticamente reflejados en la siguiente tabla:

Decisiones respecto a Ho	H_0 cierta	H_0 falsa
H_0 rechazada	Error tipo I (α)	Decisión correcta *Potencia de contraste* (1-β)
H_0 no rechazada	Decisión correcta *Nivel de confianza* (1-α)	Error tipo II (β)

Tabla 1. Situaciones de un contraste de hipótesis y posibles decisiones

Como podemos apreciar en la tabla anterior (tabla 1) el proceso final del contraste de hipótesis es un resultado que sirve para aceptar o rechazar la hipótesis nula con un cierto grado de error. Este resultado se denomina valor *empírico o teórico* y se compara con un valor *crítico o tabular* (valores de tabla) asumiendo un determinado nivel de error (generalmente 5% o α = 0.05) y en algunas ocasiones unos determinados grados de libertad. La norma general, es que si el valor empírico de la prueba calculada (el que se produce como resultado de aplicar una expresión matemática) es menor que el crítico se acepta la hipótesis nula, ya que entraría dentro de la región de aceptación de la misma.

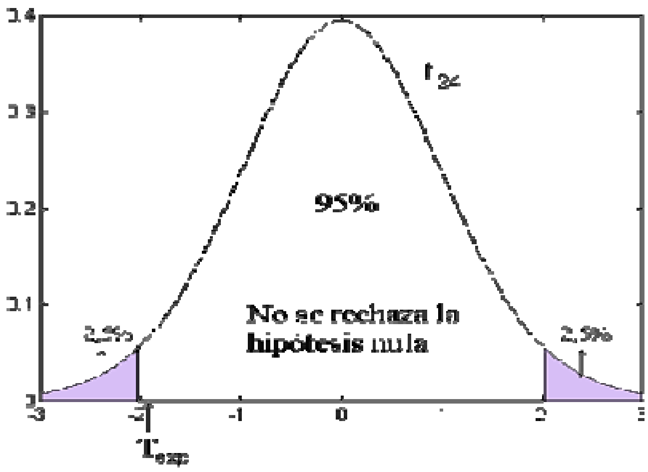
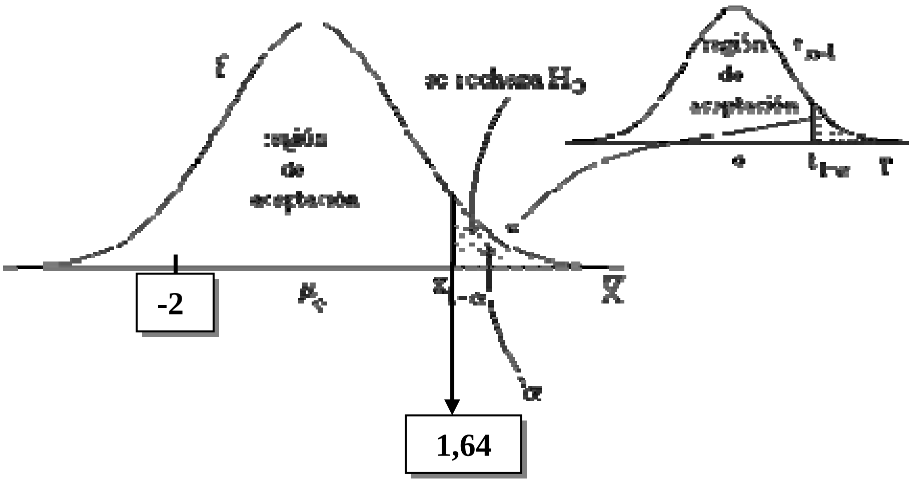


Gráfico 2. Zonas de aceptación de Ho con un α = 0.05/2, es decir, bilateral.

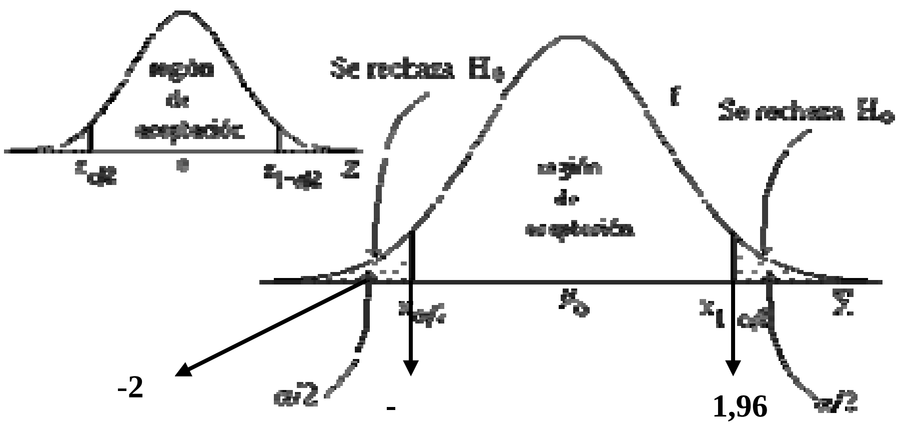
Sin embargo esta “receta” sólo funciona cuando el contraste es unilateral, es decir, cuando el error α asumido se sitúa en una sola cola de la curva, exactamente la derecha. Por esa razón, es posible encontrar numeradores de expresiones de cálculo de contrastes de hipótesis donde la diferencia entre parámetro y estadístico (“μ” media de población – media de muestra “x”) o entre dos medias muestrales ($x_1 – x_2$) se muestra en valor absoluto a fin de evitar una situación como la que vamos a describir a continuación.

Ejemplo 1: Una educadora ha comparado el rendimiento en matemáticas de los grupos de mañana y tarde. Para ello ha utilizado una prueba de contraste de hipótesis determinada (por ejemplo "T" de Student para grupos grandes e independientes) habiendo obtenido un resultado Z (empírica) = -2.

Si el contraste es unilateral derecho asumiendo un α = *0.05 (5%)* el criterio *Valor empírico* < *Valor crítico* sería válido ya que:



Evidentemente Z (empírica) es menor que Z (crítica) razón por la cual se acepta Ho. Pero imaginemos que el contraste fuese bilateral (con dos colas). En ese caso ocurriría lo siguiente:



Así pues, aunque Z (empírica) es menor que ambas Zetas (críticas) no se puede aceptar Ho, ya que –2 cae fuera del área de aceptación de la hipótesis nula. Ello no hubiera ocurrido si en el numerador de la expresión utilizada en el contraste se hubiese utilizado valor absoluto en la diferencia de medias.

En conclusión se pueden utilizar dos estrategias para una correcta decisión en una situación de contraste bilateral:

1) Utilizar en el numerador de la expresión de diferencias entre medidas el valor absoluto.
2) Considerar que todo valor empírico que se salga fuera de la región de aceptación de la hipótesis nula, bien por encima o por debajo, implicará la aceptación de la hipótesis alternativa.

Como puede apreciarse se han manejado distintos valores de puntuaciones zetas críticas. Tomando como referencia la lateralidad de la prueba (unilateral o bilateral), así como los distintos riesgos o alfas habitualmente asumidos (10%, 5% y 1%), podemos establecer diferentes valores críticos para las puntuaciones zeta:

	Riesgo α		
Lateralidad	10%	5%	1%
Contraste unilateral	*1,28*	*1,64*	*2,33*
Contraste bilateral	*1,64*	*1,96*	*2,58*

Tabla 2. Valores críticos para las puntuaciones zeta, según la lateralidad y el error alfa asumido.

2.2. *Pruebas de significación estadística*

Existen dos grandes grupos de pruebas de significación estadística, el referido a las paramétricas y el relacionado con las no paramétricas con rasgos distintivos que las caracterizan. Los más importantes son el cumplimiento o incumplimiento de determinados supuestos (normalidad, homoscedasticidad e independencia), así como el nivel de escala de medida (ordinal, nominal, intervalo o de razón) de las variables y, finalmente, el número de sujetos que conforman el estudio (muestra).

Las pruebas de tipo *paramétrico* están sometidas a determinadas condiciones de aplicación, normalmente: normalidad, homoscedasticidad e independencia (Tejedor, 1999).

Con Ximénez y San Martín (2000: 31) entendemos por *normalidad* el ajuste de los datos, en mayor o menor medida, a la curva normal; por *independencia* el que las “n” observaciones hayan sido aleatoriamente extraídas y sean independientes entre sí; y por *homoscedasticidad* que las varianzas de las distribuciones intervinientes sean homogéneas.

Las pruebas de tipo *no paramétrico* por el contrario no están sometidas a determinadas condiciones de aplicación y son, pues, adecuadas cuando se incumple alguno de los criterios previstos para las pruebas de significación de tipo paramétrico.

Antes de proponer un catálogo sobre algunos de los diferentes tipos de pruebas de significación que podemos encontrar hacemos referencia a las distintas técnicas y estrategias de verificación de los supuestos paramétricos. Con Tejedor (1999) contemplamos, entre otros, los siguientes:

SUPUESTO	***TÉCNICA/ESTRATEGIA***
Normalidad	❑ Representaciones gráficas: histograma con curva normal, gráfico P-P o de proporciones, gráfico Q-Q o de cuantiles ❑ Valores de asimetría y apuntamiento ❑ Constraste de Shapiro y Wilk ❑ Contraste de χ^2 ❑ Contraste de Kolmogorov-Smirnov con corrección de Lilliefors
Homoscedasticidad	❑ Contraste de Bartlett ❑ Contraste de Lehman ❑ Contraste de Hartley ❑ Contraste de Cochran ❑ Contraste de Levene ❑ Prueba con los logaritmos de las cuasivarianzas (para ANOVA)
Independencia	❑ Coeficiente de correlación serial de separación 1 (prueba de autocorrelación) ❑ Contraste de rachas

Tabla 3. Técnicas y estrategias para verificar los distintos supuestos paramétricos

Finalmente, proponemos una serie de pasos para implementar una prueba de significación estadística en el campo educativo, bien sea de forma manual o informatizada:

1. *Exposición (formulación) de las hipótesis estadísticas: nula y alternativa*.

2. *Establecimiento de un nivel de significación alfa o error*.

3. *Cálculo de la probabilidad de que nuestros resultados puedan haberse obtenido bajo la hipótesis nula (Ho)*:

3.1. *Selección de la prueba estadística adecuada, dependiendo de la escala de medida de las variables del estudio, el número de participantes en el mismo*...

3.2. *Ejecución de la prueba, bien mediante estrategia manual o informatizada*.

4. *Toma de decisiones teniendo en cuenta que*:

4.1. *Si el valor obtenido tras la aplicación de la prueba se encuentra localizado en la región de aceptación de la hipótesis nula se acepta dicha hipótesis, si por el contrario cae fuera de dicha región, bien por debajo o por encima se acepta la hipótesis alternativa*.

4.2. *Si la implementación es informatizada podemos fijarnos en el valor obtenido tras la aplicación (valor empírico o teórico) y compararlo con los valores críticos o tabulares. Una estrategia más rápida consiste en centrar nuestro interés en el valor-p de probabilidad obtenido en la salida del programa informático utilizado: la fórmula usual es que todo valor con una p $\leq$ 0,05 implica la aceptación de la hipótesis alternativa, mientras una p > 0,05 el rechazo de la misma, o lo que es lo mismo, la aceptación de la nula*.

5. *Conclusiones de tipo estadístico y educativo*.

Dichos pasos pueden resumirse a través del siguiente cuadro organizado a modo de mapa conceptual.

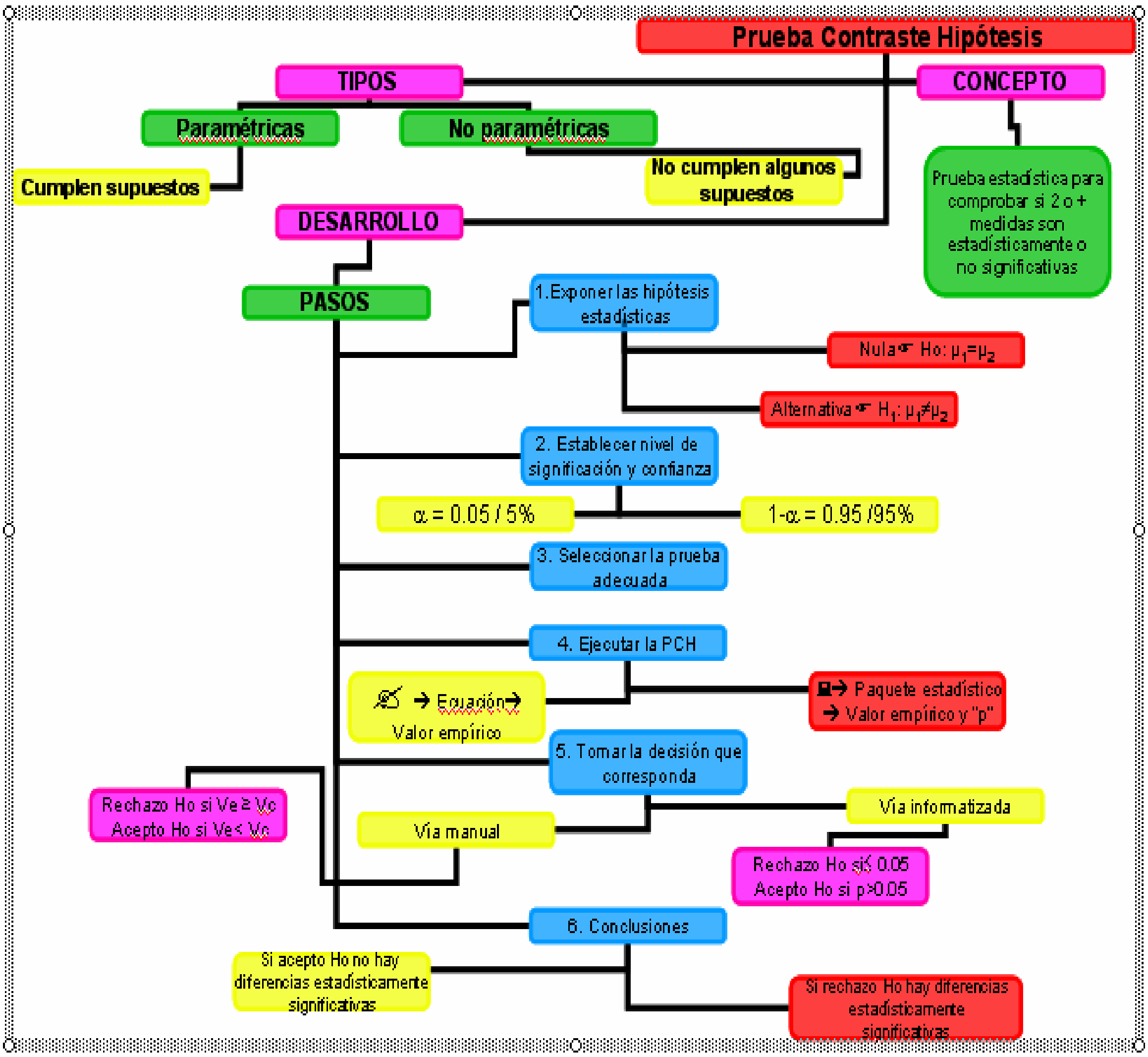


Figura 1. Proceso de desarrollo de un contraste de hipótesis. Elaboración propia.

Obtener una probabilidad $p \leq 0,05$ significa que tenemos un 5% de probabilidades de error en las conclusiones, por lo cual la probabilidad que equivocarnos es baja. En definitiva, y antes de pasar a los distintos tipos de pruebas de contrastes de hipótesis, podemos afirmar que la significación estadística (SE) hace referencia a la cuestión tendente a determinar estadísticamente si un valor o resultado obtenido de una muestra es poco probable, de modo que no puede explicarse por la fluctuaciones propias de esta muestra en cuestión. En este caso, las conclusiones pueden ser extensibles a la población de la cual deriva la muestra, dando el basamento de rechazo de la hipótesis nula.

Las pruebas de significación estadística que presentamos en la presente obra son algunas de las que mostramos en la siguiente tabla (tabla 4) no sin antes advertir al lector que en la literatura sobre análisis de datos se contemplan muchas más técnicas obviadas por razones de espacio. Para una información más detallada, y en relación al campo educativo, pueden consultarse las obras de Gil, Rodríguez y García (1995), así como la más reciente de Tejedor y Etxeberria (2006).

PRUEBAS PARAMÉTRICAS	Pruebas relacionadas con la media de una población • Prueba Z (N>30) • Prueba T (N≤30) Pruebas relacionadas con la igualdad de medias de dos grupos • Prueba Z (N>30) • Prueba T (N≤30) Pruebas relacionadas con la igualdad de varianza de dos o más grupos • ANOVA simple
PRUEBAS NO PARAMÉTRICAS	Para una sola muestra • Prueba de Kolmogorov-Smirnov • Prueba de Ji cuadrado (χ^2) • Prueba binomial Para dos muestras • Prueba U de Mann-Whitney (muestras independientes) • Prueba W de Wilcoxon (muestras relacionadas)

Tabla 4. Principales pruebas de significación estadística

2.2.1. *Principales pruebas de contraste paramétricas*

2.2.1.1. *Pruebas relacionadas con la igualdad de medias de dos grupos*

Cuando se desea determinar si las medias teóricas de dos grupos son iguales, o por el contrario diferentes, en definitiva, si las medias muestrales de ambos grupos difieren estadísticamente entre sí o no lo hacen podemos utilizar los contrastes de hipótesis paramétricos para dos grupos materializados en las diversas tipologías de la prueba “t”.

Suele ser ésta, pues, una situación más habitual de lo que a priori podemos pensar en el campo de la investigación social en general y educativa en particular. Ahora bien, se manejan diversas expresiones dependiendo de algunos aspectos:

1. Varianzas iguales o diferentes entre los grupos.
2. Consideración de muestras independientes o relacionadas.
3. Número de participantes que constituyen el estudio ($N \leq 30$ ó $N > 30$).

Los aspectos uno y tres no necesitan mayor aclaración, sin embargo, el aspecto dos necesita de una aproximación conceptual. En este sentido, para Martín (2001:38) las muestras *independientes, sin aparear o no relacionadas* son aquellas en las que los datos de una muestra no se pueden relacionar con lo de la otra. Se trata, por tanto, de dos conjuntos de datos independientes entre sí y cuyos tamaños de muestras pueden ser diferentes. Por otra parte, las muestras *apareadas, relacionadas o dependientes* son aquellas en las que cada dato de una muestra se puede asociar de manera unívoca con otro de otro grupo debido, fundamentalmente, a que las observaciones de ambos se realizan sobre las mismas unidades muestrales.

Así pues, en la literatura existente se contemplan una gran variedad de expresiones matemáticas para determinar los valores empíricos de los contrastes de hipótesis. Asumiendo la homogeneidad de las muestras consideramos las dos ecuaciones que habitualmente suelen utilizarse en diferentes situaciones de investigación. Las expresiones consideradas son:

- Para un N > 30 participantes y muestras independientes:

$$Z = \frac{x_1 - x_2}{\sqrt{S^2{}_1/N_1 + S^2{}_2/N_2}}$$

- Para un N ≤ 30 participantes y muestras relacionadas:

$$t = \frac{x_1 - x_2}{Sd / \sqrt{N}}$$

donde:

x1: media muestral del grupo 1
x2: media muestral del grupo 2
S²1: cuasivarianza o varianza poblacional del grupo 1
S²2: cuasivarianza o varianza poblacional del grupo 2
N1: Número de participantes del grupo 1
N2: Número de participantes del grupo 2
N: Número de participantes del estudio (N1+N2)
Sd: Cuasidesviación o desviación típica poblacional de la diferencia entre puntuaciones del grupo 1 y 2.

A) PRUEBA T PARA MUESTRAS INDEPENDIENTES

Ejemplo 2 **(PRUEBA T PARA MUESTRAS INDEPENDIENTES)**: *Un educador y su grupo de investigación están interesados en verificar la superioridad del método de lectura silábico frente al fonológico. Para ello enseñan a 16 alumnos de una clase con el primer método (silábico) y a otros 16 de otra clase con el segundo (fonológico). Tras un periodo prudencial miden la variable dependiente, rendimiento lector, habiéndose obtenido los siguientes resultados:*

M. silábico (1)	*3,3,4,4,5,5,5,6,6,7,7,8,8,9,9,10*
M. fonológico (2)	*2,2,3,3,4,5,6,6,6,6,7,7,7,7,7,7*

Desarrollo y resolución del ejemplo:

1. Planteamiento de las hipótesis estadísticas:

❖ Contraste bilateral:
Ho: $\mu 1 = \mu 2$ ó también $\mu 1 - \mu 2 = 0$
H1: $\mu 1 \neq \mu 2$ ó también $\mu 1 - \mu 2 \neq 0$

❖ Contraste unilateral:
Ho: $\mu 1 \leq \mu 2$ ó también $\mu 1 - \mu 2 \leq 0$
H1: $\mu 1 > \mu 2$ ó también $\mu 1 - \mu 2 > 0$

2. Selección de la prueba adecuada. Considerando las varianzas de ambos grupos homogéneas y tratándose de dos grupos de sujetos independientes, cuya suma de elementos excede de 30 (N>30) utilizamos la expresión:

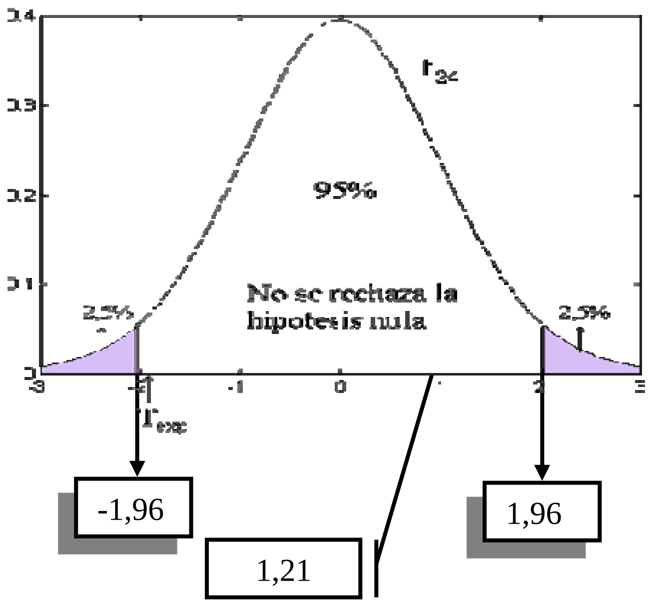
$$Z = \frac{x_1 - x_2}{\sqrt{S^2{}_1/N_1 + S^2{}_2/N_2}}$$

Si sustituimos cada miembro de la expresión por su valor tenemos que:

$$Z = \frac{6,18 - 5,31}{\sqrt{4,791/16 + 3,53/16}} \Rightarrow Z = \frac{0,87}{\sqrt{0.29 + 0,22}} = 1,21$$

3. <u>Interpretación y decisión</u>

Asumiendo un error α = 0,05 bilateral y, por tanto, unos valores críticos de zeta ±1,96 podemos apreciar como el valor de Ze (1,21) se sitúa en la región de aceptación de Ho. Cualquier lateralidad supuesta, así como errores asumidos darían lugar a una decisión similar a la tomada.



4. <u>Implicaciones estadísticas y educativas</u>

Ante los resultados obtenidos podemos afirmar que no existen diferencias estadísticamente significativas entre los métodos de lectura silábico y fonológico. Por tanto, el educador puede abordar la enseñanza de la lectura desde cualquiera de los dos métodos ya que ninguno se ha mostrado significativamente superior. Ello no implica, desde luego, que una réplica de este estudio con diferentes sujetos ofrezca resultados distintos a los obtenidos.

B) PRUEBA T PARA MUESTRAS RELACIONADAS

Ejemplo 3 **(PRUEBA T PARA MUESTRAS RELACIONADAS)**: *Un educador ha medido el número de palabras correctamente leídas (de 10 posibles) por 10 alumnos con n.e.e. (necesidades educativas especiales) después de haber sido instruidos mediante una ACI (Adaptación Curricular Individual). Estos son los resultados*

Antes de ACI	Después de ACI	D
3	5	-2
3	5	-2
5	6	-1
4	7	-3
3	5	-2
4	8	-4
5	7	-2
4	6	-2
3	7	-4
3	8	-5
Xa = 3,7	Xd = 6,4	Sd =1,25

<u>Desarrollo y resolución del ejemplo</u>:

1. <u>Planteamiento de las hipótesis estadísticas</u>:

❖ Contraste bilateral:
Ho: $\mu a = \mu d$ o también $\mu a - \mu d = 0$
H1: $\mu a \neq \mu d$ o también $\mu a - \mu d \neq 0$

❖ Contraste unilateral:
Ho: $\mu a \leq \mu d$ o también $\mu a - \mu d \leq 0$
H1: $\mu a > \mu d$ o también $\mu a - \mu d > 0$

2. <u>Selección de la prueba adecuada</u>. Considerando las varianzas de ambos grupos homogéneas y tratándose de dos grupos relacionados, ya que son los mismos sujetos los que se someten a diferentes condiciones experimentales o niveles (no ACI/sí ACI), comparándose consigo mismos y teniendo en cuenta que el número de elementos que conforman el estudio es inferior a 30 (N<30) utilizamos la expresión:

$$t = \frac{X_a - X_d}{Sd / \sqrt{N}}$$

Si sustituimos cada miembro de la expresión por su valor tenemos que:

$$t = \frac{3{,}7 - 6{,}4}{1{,}25 / \sqrt{10}} \Longrightarrow t = \frac{-2{,}7}{0{,}39} = -6{,}92$$

3. <u>Interpretación y decisión</u>

Asumiendo un error α = 0,05 bilateral y unos grados de libertad (nº de parejas-1), es decir, 10-1 = 9 situamos la región de aceptación de Ho entre las puntuaciones t = [-2,26, 2,26] . Claramente el valor de te está situado fuera de la región de aceptación de la hipótesis nula (Ho), razón por la cual la decisión es aceptar H1. A colación de la determinación de los grados de libertad (gl) debemos recordar que éstos se calculan de tres formas diferentes dependiendo de las características del tipo de contraste:

a) Si el contraste es de 1 sola muestra en comparación con un parámetro poblacional:

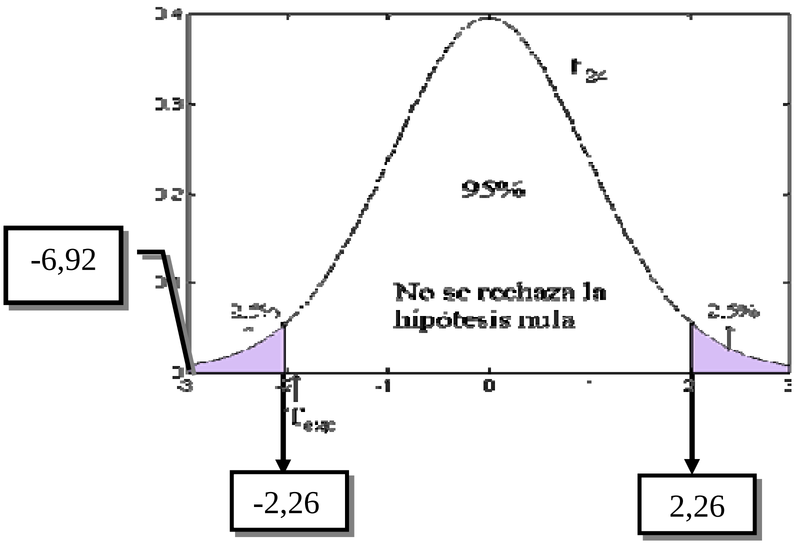
Grados de libertad = N – 1
donde N: número de participantes del estudio

b) Si el contraste es de 2 muestras y éstas son independientes:

Grados de libertad = N1 + N2 – 2
donde N1 y N2 son el número de participantes de uno y otro grupo

c) Si el contraste es de 2 muestras y éstas son relacionadas:

Grados de libertad = Número de parejas – 1
donde el número de parejas está constituido por la comparación de cada sujeto consigo mismo, así en una muestra de 5 sujetos el número de parejas sería de 5



4. <u>Implicaciones estadísticas y educativas</u>

Ante los resultados obtenidos podemos afirmar que existen diferencias estadísticamente significativas entre la capacidad lectora de los niños antes ser instruidos mediante una ACI y tras ser instruidos a través de la misma. Desde estas coordenadas, los autores del estudio están en condiciones de afirmar que las adaptaciones curriculares individuales a niños con n.e.e. parecen incidir en una mejora de la capacidad lectora.

2.2.1.2. *Pruebas relacionadas con la igualdad de varianzas de dos o más grupos*

A) *EL ANOVA simple*

También hay situaciones en que el número de muestras y/o grupos a comparar exceden de dos. En este caso, la opción más válida es el cálculo del ANOVA o ANVA (Análisis de la varianza). En esencia, el análisis de varianza intenta determinar las variaciones que se generan entre los participantes u observaciones de cada grupo (fuente de variación entre o inter) y entre los sujetos de un mismo grupo y las achacables al error (fuente de variación intra o de error). Puede consultarse la obra de Arnal y otros (1994: 107-110) para un entendimiento ameno y pedagógico del fundamento del ANOVA a través de un ejemplo de tipo experimental.

El ANOVA, pues, es una prueba semejante a las pruebas “t” y “z” en cuanto a la práctica, pero la comparación entre grupos no se basa en las diferencias entre las medias, sino en la varianza de la variable numérica “*y*” o variable dependiente en cada grupo (nivel) de la variable categórica “*x*” o variable independiente.

Si las diferencias entre las varianzas de cada grupo (fuente de variación inter) son mayores que las intragrupales (fuente de variación intra), seguramente se reportarán diferencias estadísticamente significativas entre los grupos que no son debidas al azar. Así, mientras en las pruebas de significación para dos grupos obtenemos valores empíricos “t” y “z” en el ANOVA la cifra estadística obtenida se denomina razón F de Snedecor. Al ser F el resultado de dividir la media cuadrática inter entre la media cuadrática intra cuanto mayor sea el dividendo mayor, por tanto, será el cociente y mayor, por ende, la probabilidad de que existan diferencias estadísticamente significativas.

Cabe preguntarse entonces, porque cuando existen más de dos grupos en vez de utilizar el ANOVA, aparentemente no sería más lógico e intuitivo comparar todas las posibles combinaciones tomadas de dos en dos. Supongamos, en este sentido, que un educador está interesado en denotar la posible superioridad de un programa de técnicas de estudio sobre otros dos programas. Para ello implementa cada programa en un grupo diferente. En este caso la *variable independiente* sería *programa de técnicas de estudio* con tres niveles: A, B y C, mientras la *dependiente* el *rendimiento escolar en Historia*.

Si no se contempla el ANOVA como técnica de contraste, la manera lógica de resolver esta cuestión sería comparar cada grupo con el resto por binomios, o lo que es lo mismo combinar n- elementos (3 grupos) tomados dos a dos. Si tenemos en cuenta la expresión matemática:

$$C\ n(m) = \frac{N!}{m * (N-m)!}$$

donde:

m: número de elementos que se combinan
N!: factorial de n-elementos
N y/o n: número de elementos

tomando como ejemplo el caso de los programas de técnicas de estudio y sustituyendo cada elemento por su valor en la expresión anterior tenemos que:

$$C_{3(2)} = \frac{3!}{2 * (3-2)!} \Rightarrow \frac{3*2*1}{2* (1)!} = 3$$

Podemos determinar que son tres las combinaciones necesarias para denotar las posibles diferencias entre los tres grupos tomados de dos a dos, ya que las otras tres son imagen de la matriz resultante y, por tanto, idénticas combinaciones. Gráficamente podría mostrarse de la siguiente forma:

	A	B	C
A	-	A*B	A*C
B		-	B*C
C			-

En este caso en las tres comparaciones podrían utilizarse pruebas “t” o “z” dependiendo del número de participantes que constituyen el estudio. Sin embargo, existen razones de peso para descartar esta posibilidad. Abraira y Pérez (1996) y Tejedor (1999) destacan las siguientes:

1. Ya que se realizan varios contrastes de hipótesis simultáneos e independientes la probabilidad de encontrar alguno estadísticamente significativo sin serlo verdaderamente (comisión del error tipo I o alfa) aumentaría.

2. Es difícil interpretar la verdadera influencia de la variable que actúa como factor de clasificación (variable independiente) ya que se generan diferentes niveles de significación (p) resultante de las comparaciones entre sus subgrupos (niveles de la variable independiente).

3. Cuando el número de niveles de la variable independiente es mayor o igual a 5 el número de comparaciones se dispara convirtiéndose en una ingente cantidad de cruces por parejas donde es difícil discernir cuáles de estos contrastes son estadísticamente significativos y cuáles no lo son.

Así pues, el uso del ANOVA resulta una estrategia claramente válida para minimizar en gran medida los inconvenientes de los múltiples contrastes por parejas por dos razones fundamentales:

1. Simplifica todas las comparaciones posibles entre los niveles de la variable independiente a un solo valor F asociado a una probabilidad (p).

2. Especifica entre qué niveles de la variable independiente se reportan diferencias estadísticamente significativas mediante los contrastes *post-hoc* por parejas a través de diferentes estadísticos: por ejemplo los tests de Scheffé, Tukey...

En este capítulo se tratará profundamente el ANOVA simple o de un factor (grupos independientes), el más sencillo de todos los que se contemplan en la literatura sobre análisis de datos. No obstante, destacamos que además existen otros tipos de ANOVA. Fernández y otros (1992:130) contemplan los siguientes:

I. *ANOVAS DE UN FACTOR O VARIABLE INDEPENDIENTE* 1. ANOVA simple (grupos independientes) 2. ANOVA simple de medidas repetidas MR (grupos relacionados) 3. ANOVA simple con un factor de bloqueo 4. ANOVA simple con medidas repetidas
II. *ANOVAS DE DOS O MÁS FACTORES O VARIABLES INDEPENDIENTES* 1. ANOVA FACTORIAL (DISEÑOS COMPLETOS) 1.1. CON DOS FACTORES 1.1.1. Dos factores 1.1.2. Dos factores + medidas repetidas 1.1.3. Dos factores + una variable control por emparejamiento 1.1.4. Dos factores + una variable de bloqueo 1.2. TRES FACTORES O MÁS 2. ANOVA EN DISEÑOS INCOMPLETOS 2.1. Diseños en cuadrado latino 2.2. Diseños jerárquicos 2.3. Diseños en cuadrado grecolatino

Tabla 5. Tipología de ANOVAS

Debe tenerse en cuenta, además, que la aplicación del ANOVA precisa del cumplimiento de tres supuestos básicos de carácter paramétrico: homocedasticidad, normalidad e independencia (consultar tabla nº3)

Finalmente, y antes de proponer y desarrollar un ejemplo de ANOVA simple para grupos independientes, mostramos los diferentes apartados que constituyen este tipo de ANOVA, así como las expresiones que se utilizan para su cálculo:

Fuentes de variación (FV)	Sumas de cuadrados (SC)	Grados de libertad (GL)	Medias cuadráticas (MC)	Valor de F
Entre grupos	SCe	t-1	MCe	MCe / MCi
Intragrupos o Error	SCi	N-t	MCi	
Total	SCt	N-1		

siendo:

$$SCe = \sum \frac{(\Sigma xi)^2}{N} - C$$

SCi = SCt – SCe

SCt = Σxi^2 – C

C: término de corrección que a su vez se define como:

$$C = \frac{(\Sigma xi)^2}{N}$$

t-1 (grados de libertad entre): número de grupos/tratamientos –1
N-t (grados de libertad intra): número de participantes total – menos el número de grupos/tratamientos
N-1 (grados de libertad del total): número de participantes – 1

MCe (media cuadrática entre) = SCe / t-1
MCi (media cuadrática intra) = SCi / N-t

Ejemplo 4 **(ANOVA simple):** *Un profesor está interesado en verificar si el uso de diferentes metodologías de enseñanza (M1, M2, M3) resulta determinante en la consecución de un mayor o menor rendimiento escolar en la asignatura de inglés. Así pues, durante un trimestre implementa las tres metodologías contempladas en tres diferentes grupos de alumnos midiendo el rendimiento académico en inglés al final del trimestre en una prueba final. Estos han sido los resultados*

	M1	M1	M2	M2	M3	M3	TOTAL
	A	4	F	5	K	7	
	B	3	G	4	L	8	
	C	4	H	5	M	9	
	D	3	I	5	N	8	
	E	3	J	6	O	9	
Σxi		17		25		41	83
Σxi²		59		127		339	525
Media		3,4 (17/5)		5 (25/5)		8,2 (41/5)	

Asumiendo un nivel de significación α = 0,05 bilateral puede decirse que existe alguna metodología que influye decisivamente en el rendimiento en inglés, o lo que es lo mismo ¿se reportan diferencias estadísticamente significativas entre los tres grupos que utilizan metodologías de enseñanza diferentes, tomando como variable dependiente el desempeño en inglés?

Desarrollo y resolución del problema

1. Planteamiento de las hipótesis estadísticas:

Ho: xm1 = xm2 = xm3 ó también σ^2M = 0

H1: xm1 $\neq$ xm2 $\neq$ xm3 ó también σ^2M $\neq$ 0

2. Selección de la prueba adecuada y cálculo de términos

Las características de la situación planteada (comparación de tres grupos con una sola variable independiente) aconsejan la utilización del ANOVA, aunque su uso esté condicionado al cumplimiento de determinados supuestos paramétricos. Asumiendo el cumplimiento de estos supuestos desarrollamos diferentes cálculos de diversos términos

a) El primer elemento es el término de corrección (C) que se define como:

$$C = \frac{(\Sigma xi)^2}{N}$$

Por tanto, el término de corrección sustituyendo cada elemento por su valor es:

$$C = \frac{(83)^2}{15} = \mathbf{459{,}266}$$

b) Suma de cuadrados total

SCt = 525 – 459,266 = **65,73**

c) Suma de cuadrados entre

SCe = [$(17)^2/5 + (25)^2/5 + (41)^2/5$] - 459,266 = **59,73**

a) Suma de cuadrados intra

SCi = 65,73 – 59,73 = **6**

e) Valor de la razón F

F = 29,85 / 0,5 = 59,73

f) Asociación de la magnitud del ANOVA

Para tratar de determinar la varianza explicada por el modelo, es decir, en qué porcentaje el rendimiento en inglés está determinado por el uso de una u otra metodología de enseñanza se utiliza el coeficiente de asociación η y su correlato para determinar el porcentaje de varianza η^2.

Ambos se definen como:

$$\eta = \sqrt{SCe/ SCt}$$

$$\eta^2 = SCe / SCt$$

Sustituyendo cada elemento de ambas expresiones por su valor tenemos que:

$$\eta = \sqrt{59{,}73/65{,}73} = 0{,}95$$

$$\eta^2 = SCe / SCt = 59{,}73 / 65{,}73 = 0{,}90 * 100 = 90\%$$

Como puede apreciarse el rendimiento en inglés está determinado por las metodologías de enseñanza en un 90%, un porcentaje altísimo que a fe de ser sinceros pocas veces se produce en la realidad educativa. Un resumen de todo el proceso de cálculo puede apreciarse en la siguiente tabla:

Fuentes de variación	Sumas de cuadrados (SC)	Grados de libertad (GL)	Medias cuadráticas	Valor de F
Entre grupos	59,73	3-1	29,86	59,73
Intragrupos o Error	6	15-3	0,5	
Total	65,73	15-1		

3. <u>Interpretación y decisiones</u>

Aunque todos los valores calculados son importantes, el valor decisivo para interpretar un ANOVA desarrollado manualmente es la razón F. Dicho valor es el estadístico empírico o teórico que hay que comparar con una razón F crítica o tabular. Esta segunda razón viene determinada por los grados de libertad *entre e intra* y el valor alfa que se asuma. Los grados de libertad *entre e intra* actúan como numerador y denominador respectivamente, de tal forma que el proceso de búsqueda y comparación entre razones F es, para el caso que nos ocupa, el siguiente:

F empírico = 59,73 mientras F crítico (gl m/n, α/2) $\left\{ \begin{array}{l} m = gl/entre \\ n = gl/intra \end{array} \right.$

F empírico = 59,73 mientras F crítico (2/12, 0,05/2), o sea, 5,09

Evidentemente Femp. > Fcrít., motivo por el cual se acepta H_1; ello implica que se reportan diferencias estadísticamente significativas entre los tres tipos de metodologías de enseñanza del inglés. En este sentido, la observación de los resultados obtenidos a nivel de medias resulta elocuente, ya que aquellos alumnos instruidos con la metodología 3 (M3) han alcanzado un rendimiento académico en inglés muy superior a los instruidos mediante las metodologías M2 y M3.

Además de verificar que existen diferencias estadísticamente significativas entre los grupos resulta conveniente determinar entre qué grupos se producen éstas. Para ello pueden utilizarse diferentes estrategias *post-hoc*

2.2.2. Principales pruebas de contraste no paramétricas

2.2.2.1. Para una muestra

A) *Prueba de χ^2 para una muestra*

El test de χ^2 es una prueba de bondad de ajuste y como tal intenta determinar el ajuste entre las distribuciones de frecuencias observadas y esperadas de los diferentes niveles categoriales de una variable. Desde estas coordenadas, podemos afirmar que esta prueba se aplica cuando la variable objeto de estudio está medida en una escala de tipo categorial o nominal. El mecanismo de análisis, por tanto, es parecido al de otras pruebas de similares características y parte de la siguiente expresión:

$$\chi^2 = \Sigma \frac{(fo-fe)^2}{fe}$$

donde:

fo: frecuencia observada de cada categoría
fe: frecuencia esperada de cada categoría

definiéndose fe como:

fe = suma de frecuencias de todas las categorías / número de categorías

En cuanto a las hipótesis estadísticas se plantean de la siguiente forma:

Ho: fo = fe
H1 : fo $\neq$ fe

Debemos tener en cuenta, además, que su uso está determinado por el cumplimiento de algunos supuestos:

1. La variable estudiada debe estar medida en escala nominal.
2. La variable estudiada debe ser discreta, o sea, sólo admite valores enteros no fraccionarios.
3. La variable debe ser como mínimo dicotómica o admitir un determinado número de categorías.
4. La presencia de frecuencias esperadas < 5 en un 20% de los casos puede distorsionar el resultado obtenido en un contraste χ^2, razón por la cual se desaconseja su aplicación para tal supuesto (Pick y López, 1994; Seoane, 1983 y Siegel, 1991).

Ejemplo 5 (**Prueba de χ^2 para una muestra)**: *Dos investigadores pretenden determinar si entre las tres posibles respuestas a un ítem: buena, mala y regular se reportan diferencias estadísticamente significativas en una encuesta pasada a 15 participantes. Los resultados han sido los siguientes:*

Buena (9 respuestas)
Mala (3 respuestas)
Regular (3 respuestas)

Desarrollo del problema

1. Planteamiento de las hipótesis estadísticas

Ho: fo = fe
H1 : fo $\neq$ fe

2. Cálculo de la prueba

Las características del ejemplo planteado (variable nominal y que sólo admite valores enteros) aconsejan el uso de la prueba χ^2. Aplicando su expresión de cálculo y sustituyendo cada miembro de la expresión por su valor tenemos que:

$$\chi^2 = \sum \quad \frac{(9-5)^2}{5} \quad \frac{(3-5)^2}{5} \quad \frac{(3-5)^2}{5} \quad = 4,8$$

teniendo en cuenta que las fe = 15 /3 = 5

3. Interpretación y decisión

Tomando el nivel de significación habitual, es decir, α = 0,05 unilateral, por ejemplo, y 2 grados de libertad (número de categorías – 1) tenemos que el valor de χ^2crit.= 5,99. Ya que χ^2 emp < χ^2crit, es decir, 4,8 < 5,99 aceptamos la hipótesis nula. Desde estas coordenadas, podemos afirmar que a pesar de que la alternativa más respondida ha sido *buena* (60%) frente al 20% de las otras dos alternativas *mala* y *regular* no podemos afirmar que esta diferencia sea estadísticamente significativa y, por ende, a tener en cuenta para interpretaciones posteriores.

b) Prueba de Kolmogorov-Smirnov para una muestra

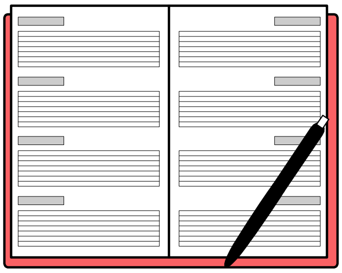
Esta prueba resulta útil para determinar el grado de acuerdo entre la distribución de un conjunto de valores de una muestra (valores observados) y alguna distribución teórica (valores esperados). Por tanto, y al igual que la técnica Ji cuadrado, se utiliza como prueba de bondad de ajuste, es decir, para decidir acerca de cuándo un conjunto de observaciones se ajusta a una distribución de probabilidad dada. La comparación se lleva a cabo a través de las distribuciones acumulativas teórica Fo (xi) y la observada Sn (xi) estableciéndose las siguientes hipótesis estadísticas a nivel bilateral:

Ho: Fo = Sn
H1: Fo $\neq$ Sn

En este sentido, esperamos que las diferencias entre Fo (xi) y Sn (xi) de cada una de las categorías comparadas sea pequeña y esté dentro de los límites de los errores aleatorios. Es por ello que para denotar si existen diferencias estadísticamente significativas entre lo observado y lo teórico se toma como valor criterio la diferencia máxima resultante, es decir:

$D = \max | Sn(xi) – Fo(xi) |$

*Ejemplo 6 (***Prueba de Kolmogorov-Smirnov para una muestra):** *Las respuestas de 25 participantes acerca de su opinión sobre la nueva Ley de Calidad de la Educación (LOCE) medida mediante escala tipo Likert con formato de 5 categorías: 1> muy en desacuerdo hasta 5> muy de acuerdo es la siguiente:*



Categorías de respuesta	Frecuencia
1	4
2	3
3	2
4	8
5	8

Asumiendo un nivel de significación de 0,05 bilateral, ¿podemos afirmar que las respuestas son unánimes o por el contrario diferentes?

Desarrollo del problema

1. Planteamiento de las hipótesis estadísticas

Ho: Fo = Sn

H_1: Fo $\neq$ Sn

2. Elaboración de la tabla de frecuencias observadas y esperadas y sus derivaciones

Categorías	OBSERVADAS	TEÓRICAS	OBS.ACUM.	TEÓ.ACUM.	Sn (xi)	Fo (xi)	Sn - Fo
1	4	5	4	5	4/25 (0,16)	5/25 (0,2)	- 0,04
2	3	5	7	10	7/25 (0,28)	10/25 (0,4)	-0,12
3	2	5	9	15	9/25 (0,36)	15/25 (0,6)	-0,24
4	8	5	17	20	17/25 (0,68)	20/25 (0,8)	-0,12
5	8	5	25	25	25/25 (1)	25/25 (1)	0

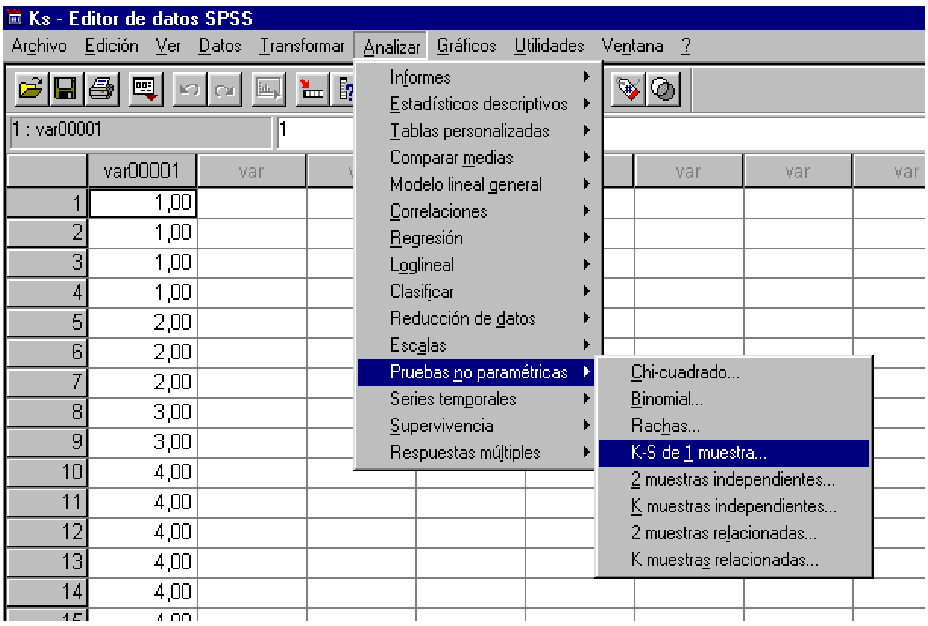
3. Interpretación y decisión

La diferencia máxima l Sn(xi) – Fo (xi) l es l0,24l . Asumiendo el alfa anteriormente contemplado (0,05/2) con un N =25 la tabla de valores críticos para la prueba K-S arroja un valor de 0,29.

Así pues, tenemos que el K-S emp. < K-S crít., es decir, que l0,24l < 0,29 razón por la cual aceptamos Ho y podemos afirmar que existe una cierta unanimidad en las respuestas efectuadas.

4. Cálculo informatizado del ejemplo

Como haremos en casos posteriores el primer paso una vez creada la plantilla de datos es marcar la opción *analizar* y dentro de ella *pruebas no* paramétricas para finalmente *marcar K-S de 1 muestra.*



Los resultados suministrados por paquete estadístico SPSS 11.0 son los siguientes:

Prueba de Kolmogorov-Smirnov para una muestra

		VAR00001
N		25
Parámetros normales	Media	3,5200
	Desviación típica	1,4754
Diferencias más extremas	Absoluta	,268
	Positiva	,158
	Negativa	-,268
Z de Kolmogorov-Smirnov		1,338
Sig. asintót. (bilateral)		,056

Los resultados arrojados por el programa SPSS no son exactamente iguales a los hallados manualmente, ya que –0,24 y –0,26 (diferencias máximas manual e informatizada respectivamente) distan dos centésimas, sin embargo, estas diferencias son atribuibles a la exactitud con la que opera el programa estadístico (4 ó más decimales). La decisión en ambos casos es aceptar la hipótesis nula, en el caso manual por que el valor empírico es menor que el crítico y en el informatizado por que el p-valor es > 0,05, exactamente 0,056.

d) <u>Prueba Binomial</u>

Cuando el número de categorías a comparar es exclusivamente dicotómico, por ejemplo: acuerdo/desacuerdo; acierto/fallo... es aconsejable la utilización de la prueba binomial como sustituta de la prueba de χ^2. En este caso, la situación de estudio se caracteriza por que:

1. Cada una de las "n" observaciones se clasifican en dos categorías exclusivamente.
2. Las "n" observaciones son independientes.
3. La probabilidad de pertenecer a una categoría es constante.

Las expresiones que se utilizan para su cálculo están determinadas por el tamaño de la muestra. De esta forma contemplamos dos opciones:

- Para un $N \leq 25$:

$$P = \binom{18}{8} \; 0{,}5^{x} * \;^{N-x}$$

- Para un N > 25:

donde:

$$Z = \frac{(x \pm 0{,}5) - N*p}{\sqrt{N* p * q}}$$

P: probabilidad de un caso individual

N: tamaño muestral

x: número de casos en la categoría P de la variable dicotómica

p: probabilidad de *x* conforme a hipótesis nula

q: 1– p

En cuanto a las hipótesis estadísticas el planteamiento bilateral sería:

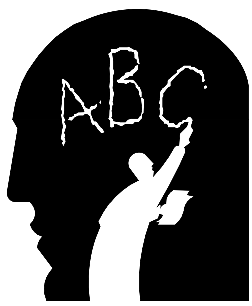
Ho: p = 0,5 o también p = po

H_1: p ≠ 0,5 o también p ≠ po

Ejemplo 7 **(Prueba Binomial):** *Un investigador ha clasificado a 18 niños con dislexia después de un programa de actuación en dos grupos:*

a) Los que han mejorado: 8 niños
b) Los que no han mejorado: 10 niños

Tomando como referencia un nivel de significación 0,05 bilateral y p y q = 0,5 cada una respectivamente, ¿podemos afirmar que el programa de actuación ha resultado significativamente efectivo?



Desarrollo del problema

1. <u>Planteamiento de las hipótesis estadísticas</u>

Ho: $p = 0{,}5$
H_1: $p \neq 0{,}5$

2. <u>Aplicación de la expresión de probabilidades binomiales</u>

$$P = \binom{18}{8} 0{,}5^{8} * 0{,}5^{10} + \ldots \binom{18}{0} 0{,}5^{0} * 0{,}5^{18} =$$

3. <u>Interpretación y decisión</u>

Comparando esta probabilidad con la mitad de α (dado que el contraste es bilateral) comprobamos que: P emp. > $\alpha/2 \Rightarrow 0{,}4073 > 0{,}025$. Por tanto, se debe aceptar la hipótesis nula pudiendo ocurrir que la proporción de niños que han mejorado tras el programa sea igual a 0,5. Desde estas coordenadas, podemos afirmar que dicho programa, pues, posee una efectividad que no ha resultado estadísticamente superior y, por ende, puede ser considerarse cuestionable.

4. <u>Cálculo informatizado del ejemplo</u>

Sustituyendo el tipo de prueba por Binomial tenemos que:

Prueba binomial

		Categoría	N	Proporción observada	Prop. de prueba	Sig. exacta (bilateral)
EFECTO TRATAMIENTO	Grupo 1	MEJORAN	8	,44	,50	,815
	Grupo 2	NO MEJORAN	10	,56		
	Total		18	1,00		

Al igual que ha sucedido en la prueba Ji cuadrado los resultados manuales son coincidentes con los informatizados. En los dos casos el p-valor es > 0,025 (contraste bilateral) ó > 0,05 (en caso que fuese un contraste unilateral), razón por la que se acepta la hipótesis nula. En cuanto a la probabilidad hallada manualmente 0,4073 si la multiplicamos por 2 dará como resultado una probabilidad aproximada de 0,815 la reseñada por el programa por haber operado con un contraste bilateral.

2.2.2.2. Para dos muestras

e) Tablas de contingencia[1] o χ^2 para dos muestras

Son muchas las ocasiones en que un investigador/a está interesado/a en verificar la relación que caracteriza a dos variables medidas en escala nominal o categórica. El tipo de análisis más apropiado para estos casos es el cálculo de una tabla de contingencia bidimensional (AxB), también conocida con las denominaciones de tabla de dos vías, tabla cruzada o *crosstabs* en su acepción anglosajona.

b1) Elementos básicos de una tabla de contingencia

En esencia una TCB consiste en una representación de la potencial correspondencia que pudiera existir entre dos elementos nominales (A,B) y los niveles (I,J) que los constituyen, es decir, asociar a cada elemento (A,B) de IxJ un número no negativo K (A,B). Si todos los valores K (A,B) son enteros se trata de una correspondencia estadística, ya que los números indican cuántas veces se presenta el elemento (A,B), aunque otras representaciones son también posibles mediante probabilidades (proporciones) y porcentajes.

Para un mayor y mejor entendimiento le proponemos el siguiente ejemplo:

---

[1] De ahora en adelante TCB.

Ejemplo 7 **(Tablas de contingencia bidimensionales):**
Sean las siguientes opiniones a favor y en contra de la nueva ley de calidad sobre la educación (LOCE) de 50 maestros (30 mujeres y 20 hombres).

	Opinión			
Género		A favor	En contra	Marginal
	Mujer	20	10	30
	Hombre	15	5	20
	Marginal	35	15	50

	A			
B		J1	J2	Marginal
	I1	Ni1j1	Ni1j2	Ni1
	I2	Ni2j1	Nij2	Ni2
	Marginal	Nj1	Nj2	N

	Opinión			
Género		A favor	En contra	Marginal
	Mujer	40%	20%	60%
	Hombre	30%	10%	40%
	Marginal	70%	30%	100%

	Opinión			
Género		A favor	En contra	Marginal
	Mujer	0,4	0,2	0,6
	Hombre	0,3	0,1	0,4
	Marginal	0,7	0,3	1

A la vista de las anteriores tablas en las que, por cierto, se muestran los diversos tipos de representación de esta técnica, es decir: notación estadística, frecuencias, porcentajes y proporciones, destacamos los cuatro elementos de toda tabla de contingencia:

1. Las frecuencias observadas.
2. Las frecuencias esperadas.
3. Las frecuencias marginales.
4. Los grados de libertad.

1. Frecuencias observadas

Las frecuencias observadas (fo) se definen como el número de veces que se presenta en una muestra cada combinación de niveles (I,J) de las variables (A,B). Así, retomando el ejemplo anterior existen cuatro frecuencias observadas, las correspondientes al cruce de (A,B) con los niveles (IxJ), o sea, 2x2. En este caso la primera combinación “Ni1j1” estaría representada por el valor 20 y así sucesivamente.

2. Frecuencias esperadas

Las frecuencias esperadas son el resultado de dividir el producto de cada total marginal de fila y columna por el total de las frecuencias observadas. Siguiendo con el ejemplo anterior tendríamos que:

	Opinión			
		A favor	En contra	Marginal
Género	Mujer	35*30/50 (21)	15*30/50 (9)	30
	Hombre	35*20/50 (14)	15*20/50 (6)	20
	Marginal	35	15	50

3. Frecuencias marginales

Las frecuencias marginales (fm) son la suma por filas (I) y columnas (J) de las frecuencias observadas (fo). Obviamente el sumatorio, en este caso, las cuatro frecuencias observadas debe ser igual a “N” o número total de frecuencias observadas.

	Opinión			
		A favor	En contra	Marginal
Género	Mujer	20	10	**30**
	Hombre	15	5	**20**
	Marginal	**35**	**15**	**50**

4. Los grados de libertad

Se definen como el producto del número de filas menos uno por el número de columnas menos uno, o sea, GL = [(I-1) * (J-1)]. En nuestro caso, pues, los grados de libertad GL = [(2-1) * (2-1)] = 1.

b2) Objetivos fundamentales de una tabla de contingencia

Como ya adelantamos al principio el objetivo básico de una TCB es denotar si la relación que existe entre dos variables nominales (A,B) es o no estadísticamente significativa, es decir, si las diferentes condiciones (IxJ) determinan o no un comportamiento diferencial al combinarse. Siguiendo con el ejemplo anterior, podemos denotar si la condición de género (hombre vs mujer) puede o no ser una variable determinante sobre las opiniones favorable vs desfavorable hacia la nueva ley de calidad de la educación (LOCE). La consecución de este objetivo implica la implementación de dos estrategias diferentes, pero estrechamente relacionadas:

a) Contraste de la independencia entre las variables nominales propuestas.
b) Determinación del grado de asociación que hay entre dichas variables categoriales.

b2.1) Independencia en TCB

Para denotar la independencia entre variables nominales en una TCB pueden calcularse diferentes colecciones de pruebas de contraste de hipótesis, por ejemplo: de bondad de ajuste, asintóticos exactos o para tablas cuadradas generadas por datos dependientes. Por ello, sin algún lector desea profundizar en las diferentes pruebas para verificar la independencia entre dos variables nominales puede consultar la obra de Aguilera del Pino (2001). Por nuestra parte destacamos la posibilidad de utilizar el contraste chi cuadrado de bondad de ajuste, ya suficientemente explicitado, aunque para una muestra en un apartado anterior (aptado. 2).

b2.2) Grado de asociación en una TCB

Al igual que se puede contrastar la independencia entre dos variables nominales, también podemos determinar el grado de asociación entre ambas. De hecho, el rechazo de independencia entre dos variables categoriales y/o ordinales es el punto de partida para indagaciones posteriores más profundas de cara de determinar la intensidad de la asociación, así como su dirección. Con Aguilera del Pino (2001) destacamos las principales propiedades de estas medidas de asociación:

a) Interpretabilidad: Resulta importante establecer unos criterios de etiquetado que faciliten la interpretabilidad de los resultados de modo que un extremo indique asociación nula (independencia) y el otro asociación perfecta. Generalmente se suele estandarizar las medidas entre 0 y 1 o entre -1 y 1 para dirigir la asociación, aunque la interpretación de los extremos no es generalmente la misma para todas ellas.

b) Simetría: Se dice que una medida de asociación es simétrica cuando su valor resulta invariable independientemente del factor (de los dos posibles) que actúe como criterio o explicación.

c) Invarianza: Las medidas de asociación, además, deben permanecer invariantes frente a cambos de escala de medida sobre filas y columnas.

En cuanto a los tipos de medidas de asociación se barajan gran cantidad de coeficientes. Para una mayor información puede consultarse la obra de Aguilera del Pino (2001).

b3) Proceso de cálculo y ejemplo práctico de una TCB

Para explicitar los pasos que constituyen el proceso de cálculo de una TCB tomaremos como referencia el ejemplo propuesto al principio del capítulo. Así pues, los pasos que guían la determinación de una TCB son:

A) Elaboración de la TCB tomando como base el enunciado del caso o problema. En nuestro caso, la TCB incluye, al menos frecuencias observadas, marginales y total:

Género	Opinión			
		A favor	En contra	Marginal
	Mujer	20	10	**30**
	Hombre	15	5	**20**
	Marginal	**35**	**15**	**50**

B) Determinación de las frecuencias esperadas y los grados de libertad

B	A			
		J_1	J_2	Marginal
	I_1	a*c / t	b*c / t	
	I_2	a*d/ t	b*d /t	
	Marginal			

Género	Opinión			
		A favor	En contra	Marginal
	Mujer	35*30/50 (**21**)	15*30/50 (**9**)	
	Hombre	35*20/50 (**14**)	15*20 (**6**)	
	Marginal			

C) Aplicación de estrategias de cálculo a fin de verificar el grado de relación existente entre las dos variables (A,B) y sus diferentes niveles (IxJ).

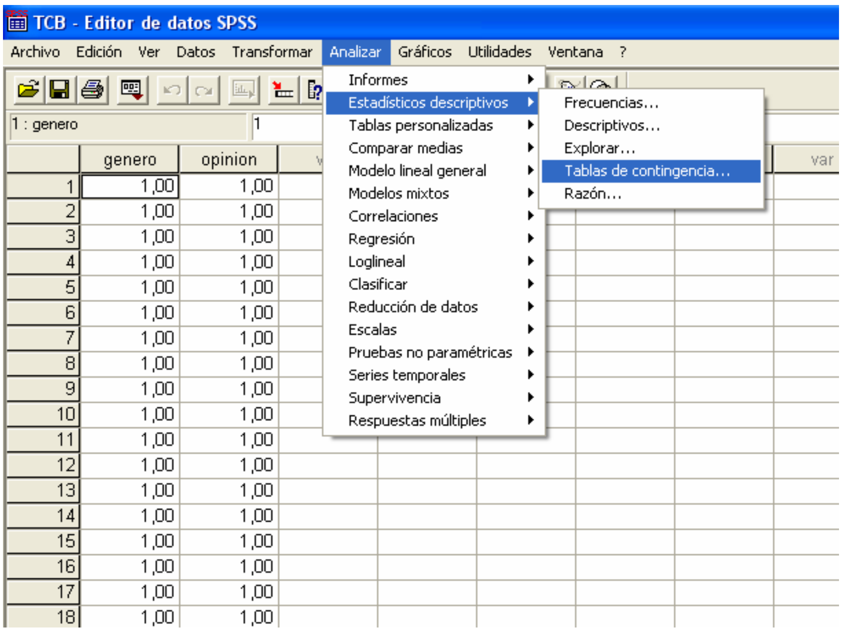
Para denotar la posible relación existente las variables género (hombre vs mujer) y opinión hacia la nueva ley de calidad sobre la educación (LOCE) (a favor vs en contra) se pueden utilizar diversas pruebas de contraste de hipótesis, así como medidas de asociación. Como anteriormente, le proponemos la clásica expresión de chi cuadrado, cuya sustitución de los valores numéricos en la expresión arroja un resultado final:

$$\chi^2 = \sum \frac{(20-21)^2}{21} + \frac{(15-14)^2}{14} \ldots\ldots = 0{,}39$$

Antes de implementar el mismo cálculo mediante SPSS, trataremos de interpretar el resultado que procede del desarrollo manual. Como podemos apreciar el valor de χ^2empírico/teórico = 0,39 que asociado a unos grados de libertad, en nuestro caso (2-1)*(2-1) = 1, y a un nivel de significación del 5% daría como resultado un χ^2crítico/tabular = 3,84. Aplicando el criterio de que todo χ^2empírico/teórico > χ^2crítico/tabular implica aceptación de la hipótesis alternativa o el contrario χ^2empírico/teórico $\leq$ χ^2crítico/tabular la aceptación de la hipótesis nula obtendríamos finalmente la conclusión de que no se reportan diferencias estadísticamente significativas en la condición de ser hombre vs mujer y estar a favor o en contra de la LOCE.

Mediante el programa estadístico SPSS deberíamos para llegar a resultados similares al poner en práctica los siguientes pasos:

1º Activar la función **analizar**, después **estadísticos descriptivos** y finalmente **tablas de contingencia**



2º Una vez en **tablas de contingencia** colocamos la variable **género** como **filas** y la variable **opinión** como **columnas.** Después activamos **estadísticos** y una vez desplegada la pantalla de la derecha marcamos **Chi cuadrado**, y por ejemplo, tres medidas de asociación más, como son el **Coeficiente de Contingencia, Phi y V de Cramer**. Finalmente, activamos la opción **continuar** y obtenemos los siguientes resultados

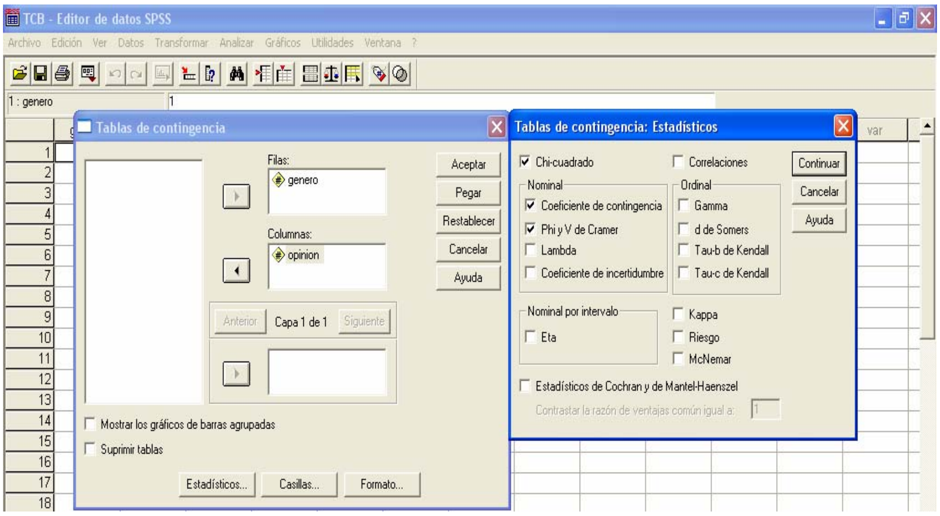


Tabla de contingencia GENERO * OPINION

		OPINIÓN		Marginal
		a favor	en contra	
GÉNERO	hombre	20	10	30
	mujer	15	5	20
Marginal		35	15	50

Pruebas de chi-cuadrado

Estadístico	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson*	,397	1	,529
N de casos válidos	50		

* 0 casillas (0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 6.

Medidas simétricas

Tipos	Medidas de asociación	Valor	Sig. aproximada
Nominal x Nominal	Phi	,089	,529
	V de Cramer	,089	,529
	Coeficiente de contingencia	,089	,529
N de casos válidos		50	

* Asumiendo la hipótesis alternativa.
** Empleando el error típico asintótico basado en la hipótesis nula.

Puede apreciarse como los resultados de la prueba de Chi cuadrado son idénticos a los obtenidos manualmente. Además, podemos apreciar como la significación asintótica (bilateral) es p > 0,05, exactamente 0,529 lo que implica la aceptación de la hipótesis nula. Además se han calculado tres medidas más asociación que como puede observarse no han resultado significativas. Le proponemos como tarea práctica calcular manualmente estos tres valores de asociación, cuyas expresiones son las siguientes:

$$\phi = \sqrt{\chi^2/N}$$

$$C = \sqrt{\chi^2 / \chi^2+N}$$

$$V = \sqrt{\phi^2 / \text{mín }(I-1),(J-1)}$$

f) ***Prueba U de Mann-Whitney (muestras independientes)***

Está considerada como una de las pruebas más potentes dentro del contexto no paramétrico y la alternativa ideal a la prueba “t” cuando ésta por las características del estudio no pueda realizarse. Las variables han de estar medidas en escala de intervalo u ordinal contemplándose dos procedimientos de cálculo, según el tamaño de la muestra:

- Para muestras pequeñas (tamaño de, al menos un grupo $\leq$ 8 participantes)

$$U_1 = n_1 * n_2 + \left[n_1*(n_1+1)/ 2\right] - R_1$$

$$U_2 = n_1 * n_2 + \left[n_2*(n_2+1)/ 2\right] - R_2$$

donde:

n_1 y n_2: número de sujetos de cada grupo
R_1 y R_2: suma de rangos correspondientes a cada grupo

- Para muestra grandes (tamaño de cada grupo > 8 participantes)

$$Z = \frac{U - (n_1 * n_2) / 2}{\sqrt{n_1 * n_2 * (n_1 + n_2 + 1) / 12}}$$

donde:

U: menor valor de las dos U calculadas

n1 y n2: número de sujetos de cada grupo

Secuencia de cálculo

1. Se ordenan conjuntamente de menor a mayor las puntuaciones de ambas muestras.
2. Se asigna un rango de orden a cada puntuación.
3. Se suman los rangos de cada muestra.
4. Se calculan los valores U de cada muestra tomándose el menor de ellos.

Veamos a través de un ejemplo el desarrollo de cálculo de la U de Mann-Whitney:

Ejemplo 8 ***(Prueba de U de Mann-Whitney)***:
Se ha pasado un determinado test de rendimiento a dos grupos de clase de 2º de bachillerato (grupo de mañana y de tarde). Los resultados obtenidos para ambos grupos son los siguientes:

Grupo de mañana	Grupo de tarde	Rango de mañana	Rango de tarde
14	18	11,5º	15º
12	16	9º	14º
13	15	10º	13º
10	14	8º	11,5º
7	19	5,5º	16º
6	7	3,5º	5,5º
4	8	2º	7º
	6		3,5º
	3		1º
Media = 9,42	Media = 11,77	Σ Rm = 49,5	Σ Rt = 86,5

Antes estos resultados, ¿podemos afirmar que alguno de los dos grupos ha obtenido un mayor rendimiento estadísticamente significativo que el otro, o por el contrario las diferencias entre ambos pueden atribuirse al azar?

Como tarea previa al cálculo de la prueba le proponemos la determinación del rango 3,5º como ejemplo para el resto de rangos:

Ya que los rangos 1º y 2º están ocupados por los valores 3 y 4 el siguiente valor en orden ascendente es 6. Al haber dos 6 debemos interpolar: 3º + 4º/2 = 3,5º. El resto de rangos se halla siguiendo semejante lógica. De esta forma, al haber dos 7 se vuelve a interpolar de la siguiente forma 5º + 6º/2 = 5,5º y así sucesivamente hasta calcular todos los rangos.

Desarrollo y resolución del problema:

1. Planteamiento de las hipótesis estadísticas:

Ho: $\mu_m = \mu_t$ ó también $\mu_m - \mu_t = 0$

H_1: $\mu_m \neq \mu_t$ ó también $\mu_m - \mu_t \neq 0$

2. Cálculo de la prueba

$U_1 = 7 * 9 + \left[7 * (7+1)/ 2\right] - 49{,}5 = 41{,}5$

$U_2 = 7 * 9 + \left[9 * (9+1)/ 2\right] - 86{,}5 = 21{,}5$

3. Interpretación y decisión

De los dos valores U se toma el menor, $U_2 = 21{,}5$, que se compara con el valor crítico de U en tabla. Para $n_1 = 7$ y $n_2 = 9$ y un nivel de significación de 0,05 $\Rightarrow$ U = 16. De esta forma, ya que $U_{emp.} > U_{crit.}$ se acepta Ho. Llamamos la atención del lector sobre la decisión tomada, que como se habrá apreciado posee una lógica totalmente opuesta hasta la ahora argumentada. Ello se debe a que en las tablas de U se ofrece el valor crítico por la izquierda de la curva, luego la región de rechazo estará formada por los valores de U menores o iguales al valor crítico. En este caso, pues, la decisión de aceptar Ho implica afirmar que el rendimiento demostrado por los alumnos de tarde, aunque superior (11,77 > 9,42) a nivel de medias, no es estadísticamente superior al de los alumnos de mañana. Las diferencias entre ambos grupos se deben, en este caso, al azar.

4. <u>Cálculo informatizado del ejemplo</u>

Sustituyendo el tipo de prueba por 2 muestras independientes, prueba U de Mann-Whitney tenemos que:

Rangos

	grupo	N	Rango promedio	Suma de rangos
rendimiento	mañana	7	7,07	49,50
	tarde	9	9,61	86,50
	Total	16		

Estadísticos de contraste**

	rendimiento
U de Mann-Whitney	21,500
Z	-1,061
Sig. asintót. (bilateral)	,289
Sig. exacta [2*(Sig. unilateral)]	,299*

* No corregidos para los empates.
** Variable de agrupación: grupo

Los resultados manuales e informatizados, igual que en los ejemplos anteriores, son enteramente coincidentes. Los valores críticos y el de los p-valores hallados determinan la aceptación de la hipótesis nula.

2. <u>Prueba de Wilcoxon (muestras dependientes)</u>

Al igual que la prueba U de Mann-Whitney, ésta es la más potente de su categoría y una buena alternativa de la prueba T para dos grupos relacionados. Además de considerar el sentido de las diferencias de las puntuaciones contempla también la magnitud de las mismas. El procedimiento de cálculo que sigue es el siguiente:

1. Se calculan las diferencias entre los dos pares de puntuaciones de cada sujeto/objeto.

2. Se ordenan estas diferencias en sentido creciente de su valor absoluto a través de la asignación de rangos que mantendrán el mismo signo de la diferencia de referencia. Por esa razón se crearán dos columnas de rangos: R+ y R–

3. Se suman las dos columnas de R+ y R–

4. Se toma como valor criterio o empírico el rango de menor cuantía.

Para el cálculo de la prueba se manejan dos expresiones dependiendo del tamaño muestral:

- Para muestras pequeñas ($N \leq 25$)

$$W = [\ S\ (+)] - [\ S\ (-)]$$

donde:

S (+): suma de rangos con signo positivo

S (-): suma de rangos con signo negativo

- Muestras grandes ($N > 25$)

$$Z = \frac{W - n\ (n+1)\ /4}{\sqrt{n * (n+1) * (2n+1)/\ 24}}$$

donde:

W: Valor mínimo de S(+) y S(-)

N: número de diferencias entre los pares de observaciones cuyo valor sea diferente de 0, o lo que es lo mismo el número de rangos cuyas diferencias no sean = 0

*Ejemplo 9 (**Prueba W de Wilcoxon):** Un investigador ha medido el rendimiento matemático de 10 niños antes de un programa sobre el uso de la calculadora de bolsillo. Tras implementarse este programa durante dos meses vuelve a medir el rendimiento matemático obteniéndose los siguientes resultados:*

Antes programa	Después programa	Antes – Después	R(+)	R(-)
4	8	- 4		8,5º
3	7	- 4		8,5º
6	5	1	2º	
2	6	- 4		8,5º
5	4	1	2º	
4	5	-1		2º
3	7	- 4		8,5º
3	6	- 3		5,5º
1	3	- 2		4º
2	5	- 3		5,5º
Media = 3,4	Media = 5,6		Σ R(+) = 4	Σ R(-) = 51

Tomando como referencia un nivel de significación de 0,05 unilateral, ¿podemos afirmar que el programa con calculadora de bolsillo ha influido significativamente en la consecución de un mayor rendimiento matemático?

Cálculo y desarrollo del ejemplo

1. Planteamiento de las hipótesis estadísticas

Ho: $\mu_a = \mu_d$ ó también $\mu_a - \mu_d = 0$
H_1: $\mu_a \neq \mu_d$ ó también $\mu_a - \mu_d \neq 0$

2. Cálculo de la prueba y decisión e interpretación de la misma

De la suma de los dos rangos, el Ra o R(+) es evidentemente menor que el Rd o R(-). Por tanto, es Ra = 4 el sumatorio de rangos que se toma como criterio. El valor de Wilcoxon crítico para un α = 0,05 unilateral es 11. Como R(+) (valor empírico) = 4 < 11 (valor crítico) siguiendo con la misma lógica de la prueba de U de Mann-Whitney rechazamos Ho y, por ende, aceptamos automáticamente H_1. Ello implica que podemos afirmar que el programa de calculadora de bolsillo aplicado a los diez niños objeto del estudio ha resultado significativamente efectivo, ya que ha generado diferencias estadísticamente significativas en el rendimiento matemático medido a los diez niños antes (media = 3,4) y después (media = 5,6) del programa.

3. Cálculo informatizado del ejemplo

Sustituyendo el tipo de prueba por 2 muestras relacionadas, prueba de Wilcoxon tenemos que:

		N	Rango promedio	Suma de rangos
DESPUES - ANTES	Rangos positivos	2*	2	4
	Rangos negativos	8**	6,37	51
	Empates	0***		
	Total	10		

* DESPUES < ANTES
** DESPUES > ANTES
*** ANTES = DESPUES

Estadísticos de contraste**

	DESPUES - ANTES
Z	-2,325*
Sig. asintót. (bilateral)	,020

* Basado en los rangos positivos.
** Prueba de los rangos con signo de Wilcoxon

Finalmente, también los resultados manuales han sido exactamente iguales a los informatizados. En ambos casos determinan la aceptación de la hipótesis alternativa.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

Abraira, V. y Pérez, A. (1996). *Métodos multivariantes en bioestadística*. Madrid. CERA.

Aguilera del Pino, A.M. (2001). *Tablas de contingencia bidimensionales*. Madrid. La Muralla.

Arnal, J. y otros (1994). *Investigación educativa. Fundamentos y metodología*. Barcelona. Labor.

Fernández, M.J.; García, J.M.; Fuentes, A. y Asensio, A. (1992). *Resolución de problemas de estadística aplicada a las Ciencias Sociales. Guía práctica para profesores y alumnos*. Madrid. Síntesis.

Gil, J.; Rodríguez, G. y García, E. (1995). *Estadística básica aplicada a las Ciencias de la Educación.* Sevilla. Kronos.

Martín, Q. (2001). *Contrastes de hipótesis*. Madrid. La Muralla.

Pick, S. y López, A.L. (1994) (5ª edición). *Cómo investigar en ciencias sociales*. México. Trillas.

Seone, J. (1983): *Psicología Matemática I*. Madrid. UNED.

Siegel, S. (1991): *Estadística no paramétrica aplicada a las ciencias de la conducta.* México. Trillas.

Tejedor, F.J. (1999). *Análisis de varianza*. Madrid. La Muralla.

Tejedor, F.J. y Etxeberria, J. (2006). *Análisis inferencial de datos en educación*. Madrid. La Muralla.

Ximénez, C. y San Martín, R (2000). *Análisis de varianza con medidas repetidas*. Madrid. La Muralla.