

LOS MÉTODOS PARAMÉTRICOS

Objetivo: El alumno al terminar el aprendizaje del presente capítulo realizará adecuadamente una inferencia estadística de la media muestral y a la media poblacional, a través de los métodos paramétricos.

LOS MÉTODOS PARAMÉTRICOS

La inferencia estadística es el conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por una muestra, cual es el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad.

Los métodos paramétricos de la inferencia estadística se pueden dividir, básicamente, en dos: métodos de estimación de parámetros y métodos de contraste de hipótesis. Ambos métodos se basan en el conocimiento teórico de la distribución de probabilidad del estadístico muestral que se utiliza como estimador de un parámetro.

La estimación de parámetros consiste en asignar un valor concreto al parámetro o parámetros que caracterizan la distribución de probabilidad de la población. Cuando se estima un parámetro poblacional, aunque el estimador que se utiliza posea todas las propiedades deseables, se comete un error de estimación que es la diferencia entre la estimación y el verdadero valor del parámetro. El error de estimación es desconocido por lo cual es imposible saber en cada caso cual ha sido la magnitud o el signo del error; para valorar el grado de precisión asociado con una estimación puntual se parte de dicha estimación para construir un intervalo de confianza. En síntesis, un intervalo de confianza está formado por un conjunto de valores numéricos tal que la probabilidad de que éste contenga al verdadero valor del parámetro puede fijarse tan grande como se quiera. Esta probabilidad se denomina grado de confianza del intervalo, y la amplitud de éste constituye una medida del grado de precisión con el que se estima el parámetro.

Los métodos de contraste de hipótesis tienen como objetivo comprobar si determinado supuesto referido a un parámetro poblacional, o a parámetros análogos de dos o más poblaciones, es compatible con la evidencia empírica contenida en la muestra. Los supuestos que se establecen respecto a los parámetros se llaman hipótesis paramétricas. Para cualquier hipótesis paramétrica, el contraste se basa en establecer

un criterio de decisión, que depende en cada caso de la naturaleza de la población, de la distribución de probabilidad del estimador de dicho parámetro y del control que se desea fijar a priori sobre la probabilidad de rechazar la hipótesis contrastada en el caso de ser ésta cierta.

En todo contraste intervienen dos hipótesis. La hipótesis nula (H_0) es aquella que recoge el supuesto de que el parámetro toma un valor determinado y es la que soporta la carga de la prueba. La decisión de rechazar la hipótesis nula, que en principio se considera cierta, está en función de que sea o no compatible con la evidencia empírica contenida en la muestra. El contraste clásico permite controlar a priori la probabilidad de cometer el error de rechazar la hipótesis nula siendo ésta cierta; dicha probabilidad se llama nivel de significación del contraste (α) y suele fijarse en el 1%, 5% o 10%.

La proposición contraria a la hipótesis nula recibe el nombre de hipótesis alternativa (H_1) y suele presentar un cierto grado de indefinición: si la hipótesis alternativa se formula simplemente como 'la hipótesis nula no es cierta', el contraste es bilateral o a dos colas; por el contrario, cuando se indica el sentido de la diferencia, el contraste es unilateral o a una sola cola.

Cuando se realiza un contraste con el SPSS no se fija el nivel de significación deseado, el programa calcula el valor-p o significación asintótica, que es la probabilidad de que el estadístico de prueba tome un valor igual o superior al muestral bajo el supuesto de que la hipótesis nula es cierta. Por tanto, si el valor-p es menor o igual que el nivel de significación deseado se rechazará H_0 . Un valor-p próximo a cero indica que se rechazará la H_0 para cualquier nivel de significación.

MEDIA POBLACIONAL

Para poblaciones normales o aproximadamente normales el intervalo de confianza para la media poblacional está centrado en la media muestral; siendo sus límites, superior e inferior, donde es el valor crítico correspondiente al grado de confianza de la distribución t de Student con $n-1$ grados de libertad. En la práctica si n es moderadamente grande el valor crítico es igual a 1,64 para un intervalo del 90% de confianza, 1,96 para el 95%, o 2,58 para el 99%.

Para obtener el intervalo de confianza para la media la secuencia es:

Analizar

Estadísticos Descriptivos

Explorar

DE MEDIAS POBLACIONALES $\mu_1 - \mu_2$

En ocasiones interesa definir un intervalo de valores tal que permita establecer cuales son los valores mínimo y máximo aceptables para la diferencia entre las medias de dos poblaciones. Pueden darse dos situaciones según las muestras sean o no independientes; siendo en ambos casos condición necesaria que las poblaciones de origen sean normales o aproximadamente normales:

- MUESTRAS INDEPENDIENTES

Si puede suponerse que las varianzas de ambas poblaciones son iguales, el intervalo de confianza para la diferencia de medias poblacionales está centrado en la diferencia de las medias muestrales, siendo sus límites superior e inferior:

$t^{\alpha/2}$ es el valor crítico correspondiente al grado de confianza $1-\alpha$ de la distribución t de

Student con $n_1 + n_2 - 2$ grados de libertad y $S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$ es una estimación de la desviación típica común a ambas poblaciones obtenida a partir de las varianzas de las dos muestras. En la práctica si n_1 y n_2 son moderadamente grandes, el valor crítico

$t^{\alpha/2}$ se aproxima, como ya se ha visto anteriormente, a los valores de la distribución normal.

Si las varianzas poblacionales no pueden suponerse iguales los límites del intervalo de confianza son:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

El valor crítico $t^{\alpha/2}$ corresponde a una distribución t cuyos grados de libertad se calculan en base a ambos tamaños muestrales y a las desviaciones típicas de cada grupo según la corrección propuesta por Dixon y Massey:

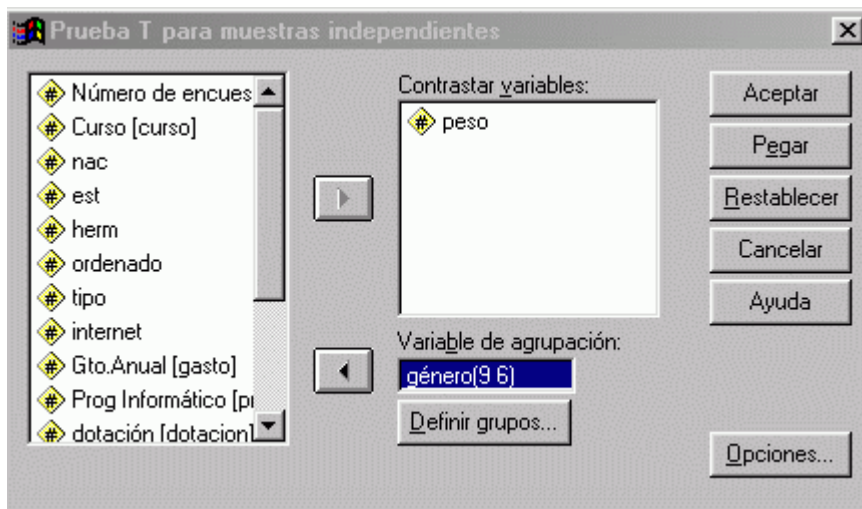
$$g.l. = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 \left(\frac{1}{n+1}\right) + \left(\frac{S_2^2}{n_2}\right)^2 \left(\frac{1}{n+1}\right)} - 2$$

Para obtener el intervalo de confianza en ambos casos la secuencia es:

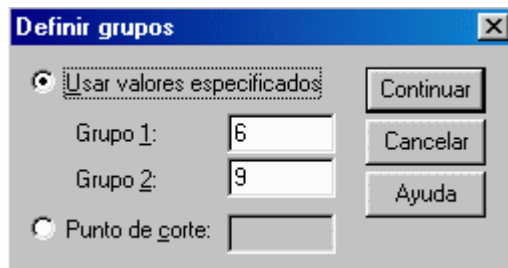
Analizar

Comparar medias

Prueba T para muestras independientes



En el cuadro de diálogo hay que seleccionar en *Contrastar variables* la variable objeto de análisis e indicar la *Variable de agrupación* junto con el criterio para *Definir grupos* (las dos poblaciones).



Los grupos pueden definirse en función de una variable cuantitativa o de una cualitativa. Si la variable de agrupación presenta sólo dos valores o modalidades, entonces se debe seleccionar *Usar valores especificados* e indicar la modalidad que define el grupo 1 y la del grupo 2. Si la variable tiene más de 2 valores o modalidades se elige la opción *Punto de corte* indicando el valor de la variable que induce una

partición en dos grupos, uno de los cuales estará formado por todos los casos con valores menores que el especificado y el otro por el resto de casos.

Al aceptar se obtienen:

- Resultados de la prueba de Levene para contrastar la igualdad de varianzas ($H_0: \sigma_1^2 = \sigma_2^2$) *
- resultados de la prueba T para contrastar la igualdad de medias ($H_0: \mu_1 = \mu_2$)
- intervalo de confianza para la diferencia de medias al 95% por defecto.

Si se quiere cambiar el grado de confianza del intervalo, antes de aceptar hay que modificarlo con el botón *Opciones*.

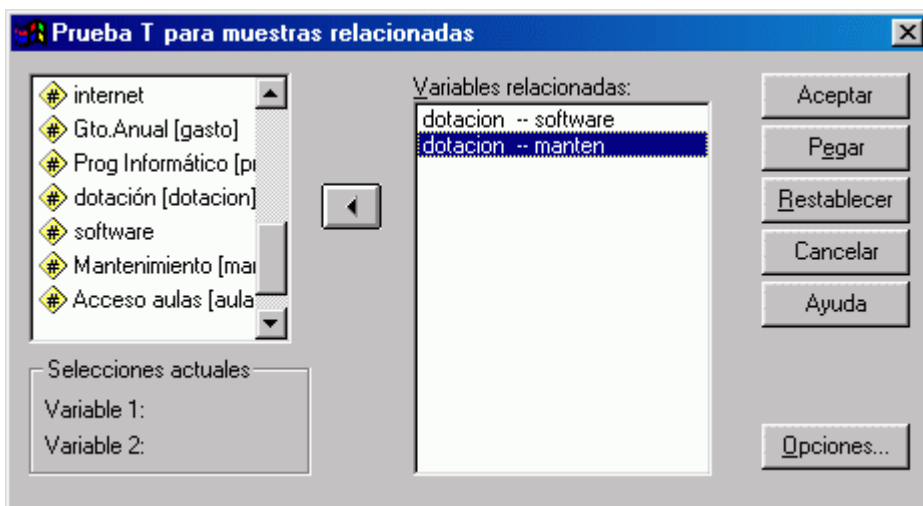
- MUESTRAS DEPENDIENTES. En este caso las muestras están formadas por parejas de valores, uno de cada población y el estadístico se obtiene a partir de las diferencias de los valores de las dos variables correspondientes a cada caso o d_i que se define como $d_i = x_i - y_i$.

Para contrastar la hipótesis de igualdad de medias y obtener el intervalo de confianza la secuencia es:

Analizar

Comparar medias

Prueba T para muestras independientes



Las variables se deben seleccionar por parejas. Haciendo clic sobre las variables de la lista aparecen sus nombres en el cuadro *Selecciones actuales*; una vez seleccionadas las dos variables se trasladan al recuadro *Variables relacionadas* de la forma habitual. En cada sesión se pueden seleccionar tantos pares de variables como medias se quieran comparar.

Al aceptar se obtienen para cada par de variables, los siguientes resultados:

- Media, desviación tipo y error típico de la media de las *d_i*.
- Intervalo de confianza, por defecto al 95%, para la diferencia de medias poblacionales que viene dado por:

$(\bar{d} \pm t_{\alpha/2} \frac{S_d}{\sqrt{n}})$ donde $t_{\alpha/2}$ es el valor de la distribución t de Student con n-1 grados de libertad que deja por encima una probabilidad de $\alpha/2$.

- Estadístico t del contraste $H_0: \mu_1 = \mu_2$ y $H_1: \mu_1 \neq \mu_2$ (contraste de dos colas).

EJEMPLO

Ejemplo 1

Con los datos de una encuesta obtener la estimación puntual y los intervalos de confianza del 95 y del 99% para la media de la población de la variable *Coste*.

En el cuadro de diálogo *Explorar*, que se obtiene con la secuencia *Analizar > Estadísticos descriptivos > Explorar*, se selecciona como variable dependiente la variable *Coste*. En Estadísticos comprobamos que está activada la opción *Descriptivos* y que el intervalo para la media definido es el del 95%.

Al aceptar se obtiene el siguiente cuadro de resultados:

Descriptivos			Estadístico	Error tip.
COSTE	Media		5236,40	365,97
	Intervalo de confianza para la media al 95%	Limite inferior	4511,34	
		Limite superior	5961,46	
	Media recortada al 5%		4813,81	
	Mediana		4000,00	
	Varianza		15268788	
	Desv. tip.		3907,53	
	Mínimo		0	
	Máximo		20000	
	Rango		20000	
	Amplitud intercuartil		2387,50	
	Asimetría		2,076	,226
	Curtosis		4,836	,449

La estimación puntual del valor esperado del coste es 5236,40 Pta. Esta estimación tiene un error típico de 365,97. Los límites inferior y superior del intervalo de confianza del 95% son 4511,34 y 5951,46, respectivamente. Este resultado se interpreta como que de los intervalos obtenidos con este método el 95% contendrán el verdadero valor esperado del coste. Una medida del grado de precisión con el que se está estimando el valor esperado es la amplitud del intervalo, que en este caso es igual a 1450,12 y la mitad de la amplitud, que es 725,06, es el error máximo de estimación que puede

garantizarse con una probabilidad de 0,95. Este error máximo es igual a $t_{\alpha/2} \frac{S}{\sqrt{n}}$, donde $t_{\alpha/2}$, es el valor crítico para $\alpha=0,05$ de la distribución t e Student, en este caso con 113 grados de libertad, y $\frac{S}{\sqrt{n}}$ es el error típico de la estimación.

Para obtener el intervalo del 99% de confianza modificamos el valor del grado de confianza en el cuadro Explorar: Estadísticos fijándolo en el 99%.

Descriptivos			Estadístico	Error tip.
COSTE	Media		5236,40	365,97
	Intervalo de confianza para la media al 99%	Limite inferior	4277,54	
		Limite superior	6195,27	
	Media recortada al 5%		4813,81	
	Mediana		4000,00	
	Varianza		15268788	
	Desv. tip.		3907,53	
	Mínimo		0	
	Máximo		20000	
	Rango		20000	
	Amplitud intercuartil		2387,50	
	Asimetría		2,076	,226
	Curtosis		4,836	,449

Los límites del intervalo de confianza del 99% son 4277,54 y 6195,27; la confianza de que este intervalo contenga el verdadero valor esperado del coste es 0,99. La amplitud de este intervalo es 2217,73 que es mayor que la amplitud del intervalo del 95%, por lo tanto, 1108,865, es el error máximo de estimación que puede garantizarse con una probabilidad de 0,99. Como puede verse, a medida que aumenta el grado de confianza del intervalo disminuye la precisión de la estimación.

Ejemplo 2.

Para la misma variable Coste verificar si se puede aceptar el supuesto de que el valor esperado del Coste es superior a 6000.

Con la secuencia *Analizar > Comparar medias > Prueba T para una media* se abre el cuadro de diálogo *Prueba T para una muestra* en el cual se selecciona la variable Coste y se indica como *Valor de prueba* 6000. Esto quiere decir que las hipótesis que se están contrastando son $H_0:\mu = 6000$ frente a $H_1:\mu > 6000$. Se trata por tanto de un contraste a una sola cola.

Prueba para una muestra						
Valor de prueba = 6000						
	t	gl	Sig. (bilateral)	Diferencia de medias	Intervalo de confianza para la diferencia	
					Inferior	Superior
COSTE	-2,086	113	,039	-763,60	-1488,66	-38,54

El estadístico de prueba $t = \frac{\bar{X} - 6000}{S/\sqrt{n}}$ toma el valor $t = -2,086$, que en las tablas de la distribución t de Student con 113 grados de libertad deja por debajo un área de 0,0195. Esto quiere decir que se rechaza la hipótesis nula en favor de la alternativa para niveles de significación superiores a 0,0195. Obsérvese que 0,0195 es la mitad del nivel de significación para la prueba de dos colas que aparece en el cuadro de resultados.

Por otra parte, si las hipótesis hubieran sido $H_0: \mu = 6000, H_1: \mu \neq 6000$ se rechazaría la hipótesis nula en favor de la alternativa para niveles de significación superiores a 0,039. El intervalo del 95% de confianza para la media calculado en el apartado anterior no contenía el valor 6000; lo que equivale a decir que para un nivel de significación del 5% se rechaza la hipótesis nula. Por el contrario, el intervalo del 99% contenía el valor 6000 y, por lo tanto, para un nivel de significación del 1% no se rechazaría la hipótesis nula.

Ejemplo 3.

Verificar si existe diferencia significativa entre el coste esperado en transporte de los alumnos que viven en Barcelona y el de los que viven fuera.

Con la secuencia *Analizar > Comparar medias > Prueba T para muestras independientes* se abre el cuadro de diálogo *Prueba T para muestras independientes* en el cual se selecciona la variable *Coste* y se indica como *Variable de agrupación Resid.* En la opción *Definir grupos* se asigna al *Grupo 1* el valor 1 (vive en Lima) y al *Grupo 2* el valor 2 (no vive en Lima). Aceptando se obtienen entre otros los siguientes resultados:

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias		
		F	Sig.	t	gl	Sig. (bilateral)
COSTE	Se han asumido varianzas iguales	37,671	,000	-4,459	112	,000
	No se han asumido varianzas iguales			-3,750	50,997	,000

Las hipótesis que se están contrastando son frente $H_0: \mu_1 = \mu_2$ frente a $H_1: \mu_1 \neq \mu_2$. Para realizar este contraste previamente se debe comprobar si es aceptable la hipótesis de varianzas poblacionales iguales para los dos grupos $H_0: \sigma_1^2 = \sigma_2^2$. El estadístico F de la prueba de Levene* no permite aceptar la igualdad de varianzas poblacionales, por lo cual el valor del estadístico de prueba es $t = -3,750$ que para cualquier nivel de significación lleva a rechazar la hipótesis de igualdad de medias. El signo negativo del estadístico t indica que el coste del transporte es significativamente superior para los que viven fuera de Barcelona.

Ejemplo 4.

Con los datos de una encuesta contrastar si existe diferencia significativa entre las puntuaciones medias asignadas en cuanto a una variable cualitativa .

Las puntuaciones que se quiere comparar han sido generadas dos a dos por los mismos individuos; se trata por tanto del caso de muestras relacionadas. Las hipótesis

que se contrastan son $H_0: \mu_1 = \mu_2$ frente a $H_1: \mu_1 \neq \mu_2$.

Con la secuencia *Analizar > Comparar medias > Prueba T para muestras relacionadas* se abre el cuadro de diálogo en el cual se selecciona la pareja de variables. Al aceptar se obtienen los siguientes resultados:

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación tip.	Error tip. de la media	Intervalo de confianza para la diferencia				
				Inferior	Superior			
dotación - SOFTWARE	-1,12	1,95	,19	-1,50	-,75	-5,93	105	,000

El análisis sólo ha considerado los casos que no presentan ningún valor missing en el par de puntuaciones, quedando únicamente 106 casos válidos de los 114.

El promedio de las diferencias entre las puntuaciones asignadas a la dotación y al software es de -1,12 con un error típico igual a 0,19. El estadístico de prueba t es igual a -5,93 y se distribuye según una t de Student con 105 grados de libertad. Con este valor de t se rechaza la hipótesis nula para cualquier nivel de significación. Los resultados proporcionan también el intervalo de confianza para la diferencia de las dos medias poblacionales con el 95% de nivel de confianza. Como puede observarse el

intervalo no contiene el valor 0, de lo que se deduce también que no se puede aceptar que las puntuaciones medias sean significativamente iguales.

PREGUNTAS

- **¿Cómo se infiere a la Media Poblacional?**
- **¿Cómo se elabora un análisis de datos con la T de student Pareada y para muestras independientes?**

LOS MÉTODOS NO PARAMÉTRICOS

Objetivo: El alumno al terminar el aprendizaje del presente capítulo realizará adecuadamente una inferencia estadística de la proporción poblacional, a través de los métodos no paramétricos.

PROPORCIÓN POBLACIONAL π

En poblaciones dicotómicas con una proporción π de éxitos el estimador puntual del parámetro π es la proporción muestral de éxitos, p , que coincide con la media de la muestra cuando se codifica como 1 la característica que se considera como éxito y 0 la que se considera no éxito. A partir de un tamaño muestral moderadamente grande el estadístico p tiene una distribución aproximadamente normal. El intervalo de confianza para la proporción poblacional está centrado en la proporción muestral;

siendo sus límites superior e inferior $p \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$ donde $z_{\alpha/2}$ es el valor crítico correspondiente al grado de confianza $1-\alpha$ de la distribución normal tipificada y $\sqrt{\frac{p(1-p)}{n}}$ es el error típico de la proporción.

Para obtener el intervalo de confianza y contrastar hipótesis sobre la proporción una alternativa consiste en tratar a la proporción π como la media poblacional de una variable dicotómica codificada como se ha descrito anteriormente (éxito=1, no éxito=0) y la secuencia es:

- Para el intervalo de confianza:

Analizar

Estadísticos Descriptivos

Explorar

$$H_0: \pi = \pi_0:$$

- Para contrastar la hipótesis nula

Analizar

Comparar medias

Prueba T para una muestra

Utilizando este criterio los resultados numéricos no coinciden exactamente con los que se obtendrían aplicando la expresión del error típico de la proporción; no obstante, la

discrepancia es despreciable si el número de observaciones es suficientemente grande.

Otras alternativas para realizar este contraste son de naturaleza no paramétrica.

PRUEBA BINOMIAL

La prueba binomial prueba binomial analiza variables dicotómicas y compara las frecuencias observadas en cada categoría con las que cabría esperar según una distribución binomial de parámetro π especificado en la hipótesis nula. El nivel de significación crítico de esta prueba indica la probabilidad de obtener una discrepancia igual o superior a la observada a partir de la muestra si la distribución es la postulada por la hipótesis nula.

El nivel de significación crítico (bilateral) de este contraste debe interpretarse como:

$$P(|X - n\pi_0| \geq X_0 / X \sim B(n, \pi_0))$$

el número de éxitos en la muestra.

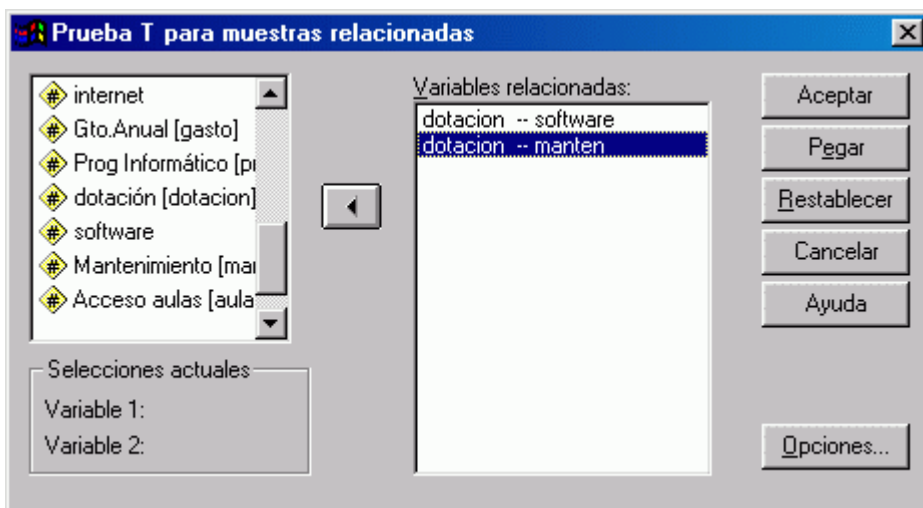
Cuando n es suficientemente grande se calcula esta probabilidad aproximando la distribución binomial a la normal con corrección de continuidad.

La secuencia para realizar este contraste es:

Analizar

Pruebas no paramétricas

Binomial



Se especifica la variable en *Contrastar variables*. Se indica la proporción postulada en la hipótesis nula en *Contrastar proporción*. Si la variable es dicotómica se mantiene activada la opción *Obtener de los datos*. Si la variable no es dicotómica en *Punto de corte* se indica el valor de corte de forma que los inferiores o iguales se agrupan en la primera categoría y el resto en la otra.

El botón *Opciones* permite obtener estadísticos de resumen y modificar el tratamiento de los valores missing.

PRUEBA CHI-CUADRADO

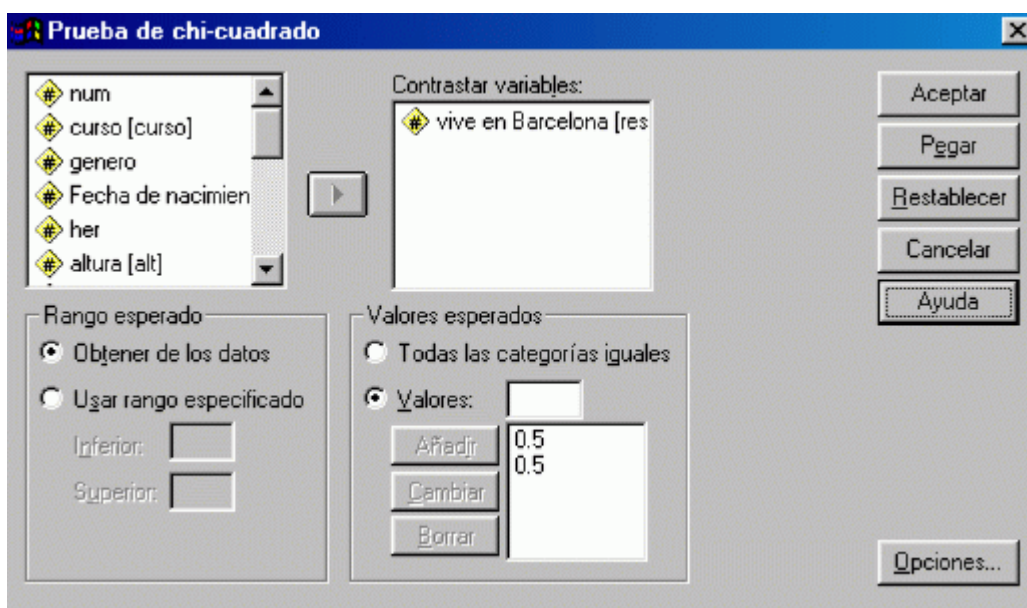
Otra alternativa de naturaleza no paramétrica se basa en el valor del estadístico Chi-cuadrado. Esta prueba compara la frecuencia observada para cada categoría (O_i) con la frecuencia teórica o esperada (E_i) bajo el supuesto de que la proporción de éxitos es la postulada por la hipótesis nula. Para una muestra de tamaño n la frecuencia esperada se calcula: $E_0 = \pi_0 n$ para los éxitos y $E_1 = (1 - \pi_0)n$ para los no éxitos. El estadístico de prueba Chi-cuadrado se define $\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$ y presenta una distribución Chi-cuadrado con 1 grado de libertad.

La secuencia es:

Analizar

Pruebas no paramétricas

Chi-cuadrado



En el cuadro de diálogo se indica, en la casilla *Contrastar variables* la variable sobre la que se realiza el contraste; obsérvese que en *Rango esperado* está activada la opción *Obtener de los datos*, que toma tantas categorías como valores distintos presenta la variable. Si la hipótesis nula es $H_0: \pi = 0,5$ los valores esperados serán *Todas las categorías iguales*, en cualquier otro caso se pueden introducir los valores (frecuencias esperadas absolutas o relativas) con la opción *Valores*.

EJEMPLO

Con los datos de una encuesta estimar la proporción mediante un intervalo de confianza del 90% y contrastar la hipótesis de que son mayoría.

El primer paso es comprobar si la variable Resid está correctamente codificada, es decir, presenta valor 1 cuando el alumno es residente en Barcelona y 0 en otro caso. En la base de datos se ve que la codificación de la variable Resid no es la adecuada, por lo tanto, en primer lugar, habrá que recodificar la variable. La secuencia es *Transformar > Recodificar > En la misma variable*. En el cuadro de diálogo se selecciona la variable Resid y se definen los *Valores antiguos y nuevos* asignando al *valor antiguo 2 el valor nuevo 0*. En el editor de datos se puede etiquetar el valor 0 como 'no' haciendo doble clic sobre el título de la variable Resid.

- Una aproximación del intervalo de confianza se obtiene con la secuencia *Analizar > Estadísticos descriptivos > Explorar*. En el cuadro de diálogo se selecciona como variable dependiente la variable Resid y en *Estadísticos* se modifica el grado de confianza fijándolo en el 90%. Los resultados que se obtienen al aceptar son:

Descriptivos

			Estadístico	Error típ.
vive en Barcelona	Media		,61	4,58E-02
	Intervalo de confianza para la media al 90%	Límite inferior	,54	
		Límite superior	,69	
	Media recortada al 5%		,63	
	Mediana		1,00	
	Varianza		,239	
	Desv. típ.		,49	
	Mínimo		0	
	Máximo		1	
	Rango		1	
	Amplitud intercuartil		1,00	
	Asimetría		-,475	,226
	Curtosis		-1,807	,449

Como se observa la media igual a 0,61 indica que el 61% de los elementos de la muestra residen en Barcelona. A partir de esta estimación puntual de π y de la estimación de su error típico se obtiene el intervalo de confianza al 90% cuyo límites inferior y superior son 0,54 y 0,69, respectivamente.

- Para contrastar las hipótesis $H_0:\pi = 0,5$ y $H_1:\pi > 0,5$ considerando la proporción como media la secuencia es *Analizar > Comparar medias > Prueba T para una muestra*. Se selecciona la variable Resid y se indica como *Valor de prueba* 0,5, los resultados del contraste se recogen en el siguiente cuadro:

Prueba para una muestra

	Valor de prueba = 0.5					
	t	gl	Sig. (bilateral)	Diferencia de medias	Intervalo de confianza para la diferencia	
					Inferior	Superior
vive en Barcelona	2,490	113	,014	,11	2,33E-02	,20

Como el contraste es a una sola cola el nivel de significación crítico es la mitad de 0,014, es decir, 0,007. Para niveles de significación superiores a 0,007 se rechaza la hipótesis nula. Así pues, para los niveles de significación habituales se acepta que la mayoría de alumnos viven en Barcelona ($\pi > 0,5$).

- Utilizando la prueba binomial para contrastar las hipótesis $H_0:\pi = 0,5$ y $H_1:\pi > 0,5$ la secuencia es *Analizar > Pruebas no paramétricas > Binomial*.

Prueba binomial

		Categoría	N	Proporción observada	Prop. de prueba	Sig. asintót (bilateral)
vive en Barcelona	Grupo 1	si	70	,61	,50	,019 ^a
	Grupo 2	no	44	,39		
	Total		114	1,00		

a. Basado en la aproximación Z.

Si la hipótesis nula es cierta, el número de residentes observado (70) proviene de una población binomial de parámetros $n=114$ y $\pi = 0,5$ y por tanto con valor esperado 57. En tal caso:

$$P(|X - 57| \geq 70) = P(X \leq 13) + P(X \geq 70) = 0,019$$

Por lo tanto, $P(X \geq 70) = 0,0095$ que es el nivel de significación crítico del contraste a cola superior, en consecuencia se rechaza la hipótesis nula.

- Para contrastar las hipótesis $H_0:\pi = 0,5$ y $H_1:\pi > 0,5$ aplicando la prueba Chi-cuadrado la secuencia es *Analizar > Pruebas no paramétricas > Chi-cuadrado*.

vive en Barcelona

	N observado	N esperado	Residual
si	70	57,0	13,0
no	44	57,0	-13,0
Total	114		

Estadísticos de contraste

	vive en Barcelona
Chi-cuadrado ^a	5,930
gl	1
Sig. asintót.	,015

a. 0 casillas (.0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 57,0.

Como el nivel de significación asintótico es 0,015 la hipótesis nula no se rechaza para los niveles de significación inferiores al 1,5 %.

DIFERENCIA DE PROPORCIONES $\pi_1 - \pi_2$

El estadístico, de prueba que permite contrastar $H_0:\pi_1 = \pi_2$ frente a $H_1:\pi_1 \neq \pi_2$ a

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}$$

partir de dos muestras aleatorias e independientes es Z siendo p la estimación de π obtenida del total de observaciones.

Si se consideran las proporciones como medias y se aplica la prueba t utilizada para comparar medias poblacionales los resultados no son fiables ya que la estimación del error típico que realiza el programa no coincide con la del estadístico de prueba. Para resolver el problema con el programa SPSS se deberá cruzar la variable analizada con

la que define los grupos (obtener la tabla de contingencia) y realizar el contraste de independencia Chi-cuadrado.

El estadístico de prueba Chi-cuadrado se define:
$$\chi^2 = \sum_{j=1}^2 \sum_{i=1}^2 \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
 y presenta una distribución Chi-cuadrado con $(I-1)(J-1)$ grados de libertad. Las E_{ij} se calculan multiplicando las frecuencias marginales y dividiendo el producto por n . Estas E_{ij} son estimaciones de las frecuencias absolutas que cabría esperar en cada casilla bajo el supuesto de que la proporción de éxitos es la obtenida a partir del total de observaciones muestrales sin considerar diferencias entre los dos grupos.

La secuencia es:

Analizar

Estadísticos Descriptivos

Tablas de contingencia

En el cuadro de diálogo se indica la variable que se quiere contrastar (filas), la variable que define los dos grupos (columnas) y se selecciona la opción Chi-cuadrado en Estadísticos.

EJEMPLO

*Con referencia a la encuesta **Enctrans.sav** se quiere comprobar si la proporción de alumnos con vehículo difiere significativamente entre los grupos definidos según el género.*

La hipótesis nula del contraste es $H_0: \pi_1 = \pi_2$; siendo π_1 la proporción poblacional de hombres con vehículo y π_2 la proporción poblacional de mujeres con vehículo.

Con la secuencia *Analizar > Estadísticos Descriptivos > Tablas de contingencia* se accede al cuadro de diálogo donde se indica que la variable a contrastar es Vehículo y que la variable de agrupación es el Género, y se selecciona la opción *Chi-cuadrado en Estadísticos*. Al aceptar se obtiene el siguiente cuadro de resultados.

Tabla de contingencia Tiene vehículo? * GENERO

			GENERO		Total
			hombre	mujer	
Tiene vehículo?	No	Recuento	33	42	75
		Frecuencia esperada	35,5	39,5	75,0
		% del total	28,9%	36,8%	65,8%
	Sí	Recuento	21	18	39
		Frecuencia esperada	18,5	20,5	39,0
		% del total	18,4%	15,8%	34,2%
Total	Recuento	54	60	114	
	Frecuencia esperada	54,0	60,0	114,0	
	% del total	47,4%	52,6%	100,0%	

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	,998 ^b	1	,318		
Corrección de continuidad	,642	1	,423		
Razón de verosimilitud	,998	1	,318		
Estadístico exacto de Fisher				,331	,212
Asociación lineal por lineal	,989	1	,320		
N de casos válidos	114				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 18,47.

Si es cierto que la proporción de propietarios de vehículo es la misma en los dos grupos, $\pi_1 = \pi_2 = \pi$, la estimación de π es la proporción de propietarios de vehículo para el total de alumnos de la muestra, es decir, $39/114=0,3421$. La frecuencia esperada de hombres con vehículo se obtendrá multiplicando esta proporción por el total de hombres en la muestra, o sea, $0,3421 \cdot 54=18,5$; y de la misma forma se obtendrá la frecuencia esperada de mujeres con vehículo: $0,3421 \cdot 60=20,5$ (véase que estas frecuencias esperadas coinciden con las que cabría esperar en el caso de que las variables Género y Vehículo fueran independientes).

El estadístico Chi-cuadrado toma el valor 0,998 y el nivel de significación crítico es 0,318, por lo tanto no se rechaza la hipótesis nula para los niveles de significación habituales y se puede aceptar que no hay diferencia entre la proporción de hombres y mujeres propietarios de vehículos

PREGUNTAS

- **¿Cómo y para que se emplean las pruebas no paramétricas?**
- **¿Cómo se elabora un análisis de datos con Chi²?**

