Cuat. II Un. 2 y 3

Manuel García Ferrando

Socioestadística

Introducción a la estadística en sociología

Alianza
Editorial

Capítulo 6
PRUEBAS DE DECISIÓN ESTADÍSTICA PARA EL CASO DE UNA SOLA MUESTRA

6.1. INTRODUCCIÓN

En el presente capítulo nos vamos a ocupar del estudio de aquellas pruebas de decisión estadística, paramétricas y aparamétricas, que sólo requieren la extracción de una muestra. Básicamente, tales pruebas estadísticas nos informan acerca de si la muestra bajo estudio pertenece a una población determinada. Las pruebas de decisión estadística para una sola muestra suelen medir la bondad del ajuste (en inglés, *goodness-of-fit*). En el caso típico se extrae una muestra aleatoria y, a continuación, se somete a prueba la hipótesis de que dicha muestra se ha extraído de una población que presenta una distribución específica. Como destaca Siegel (1956, pág. 35), las pruebas de decisión estadística para una sola muestra permiten responder preguntas como las que siguen: ¿existe una diferencia significativa de posición (tendencia central) entre la muestra y la población?; ¿existe una diferencia significativa entre las frecuencias observadas y las frecuencias que cabría esperar en base a algún principio?; ¿existe una diferencia significativa entre las proporciones observadas y las proporciones esperadas?; ¿está justificado considerar que una muestra concreta pertenece a una población con una forma determinada? (por ejemplo, normal); ¿está justificado considerar que una muestra concreta es una muestra aleatoria de una población conocida? Como vemos, se trata, en último término, de contrastar los valores observados de una sola variable en una muestra en relación a los valores que toma dicha variable en la población. Estas son, básicamente, las preguntas que cabe formularse cuando se trabaja con una sola muestra, y que pueden ser respondidas mediante el empleo de las correspondientes pruebas de decisión estadística. En primer lugar estudiaremos la distribución probabilística binomial y la correspondiente prueba binomial y, a continuación, la prueba del chi-cuadrado para una sola muestra. Las pruebas estadísticas de comparación de una proporción y media observadas a una proporción y media teóricas, respectivamente, serán igualmente tratadas en las páginas siguientes, así como la distribución t de

Student y la estimación por intervalo de una media. Con esto no pretendemos ofrecer una panorámica exhaustiva de esta parcela de la estadística inferencial. Otros libros de estadística citados en el apartado bibliográfico contienen un mayor número de pruebas, y a ellos remitimos al lector interesado en el estudio más detallado de las mismas. Con todo, las pruebas que se estudian en el presente capítulo se encuentran entre las más conocidas y utilizadas en la investigación empírica sociológica, y las hemos considerado suficientes para que el estudiante de sociología se familiarice con la lógica de las pruebas de decisión estadística en el trabajo sociológico, sin que se sienta abrumado por una diversidad de técnicas estadísticas que pueden distraer su atención de los temas básicos, y relativamente sencillos, de la lógica del contraste y verificación de hipótesis para el caso de una sola muestra y una sola variable.

6.2. La distribución binomial. La prueba binomial

Existen poblaciones que pueden considerarse que están formadas tan sólo por dos categorías. Así, por ejemplo, hombre y mujer, rural y urbano, éxito y fracaso, miembro y no miembro, alfabeto y no alfabeto, soltero y casado, religioso y no religioso, etc. Para tales casos, cada observación que se realice a partir de la correspondiente población pertenecerá a una u otra de las dos categorías discretas. Resulta obvio que, una vez conocida la proporción P de casos que pertenecen a una de las categorías, conoceremos automáticamente la proporción de casos que pertenecen a la segunda categoría, proporción que valdrá $1-P$. Habitualmente, para representar la expresión $1-P$ se utiliza el símbolo Q ($Q=1-P$).

Al extraer una muestra aleatoria de una población de este tipo, no cabe esperar que las proporciones respectivas de casos pertenecientes a ambas categorías en la muestra sean exactamente P y Q. Los efectos del azar del muestreo impedirán habitualmente que los valores de las muestras sean exactamente los valores P y Q de la población. Así, por ejemplo, podemos conocer por medio del Censo de Población que la proporción de solteros y la de casados en la población adulta de una región determinada es 35 y 65 por 100, respectivamente. Pero si extraemos una muestra aleatoria de la población adulta de dicha región, los solteros y los casados pueden representar, por ejemplo, el 32 y el 68 por 100, respectivamente, o incluso el 38 y el 62 por 100, respectivamente. Como se ha dicho anteriormente, el azar que se introduce al extraer una muestra aleatoria es el responsable de la aparición de tales diferencias entre los valores de la población y los valores observados.

Pues bien, la *distribución binomial* es la distribución muestral de las proporciones que se pueden observar en muestras aleatorias extraídas de una población que se caracteriza por estar compuesta por dos categorías de casos o miembros. Al ser una distribución muestral, la distribución binomial ofrece los diversos valores que pueden ocurrir bajo H_0, siendo en este caso H_0 la hipótesis de que el valor de la población es P. Por tanto, cuando las puntuaciones o valores obtenidos en una investigación se pueden distribuir en dos categorías, se puede utilizar la distribución binomial para contrastar H_0. La prueba binomial, al ser una prueba que mide la bondad del ajuste, nos dice si cabe esperar que las proporciones (o frecuencias) que se observan en una muestra pueden pertenecer a una población que tiene un valor específico de P.

Veamos ahora cómo se opera empíricamente con la prueba binomial. Mediante la combinación de la regla de la multiplicación de probabilidades y de la fórmula que expresa las combinaciones de m elementos tomados de n en n, se puede demostrar que la probabilidad de obtener x objetos en una categoría y $N-x$ objetos en la otra categoría viene dada por la fórmula:

$$p(x) = \binom{N}{x} P^x Q^{N-x} \qquad [6.1]$$

en donde P es la proporción de casos que pertenecen a una categoría; Q la proporción de casos que pertenecen a la segunda categoría, y:

$$\binom{N}{x} = \frac{N!}{x!\,(N-x)!}$$

La utilización de la fórmula [6.1] es bien sencilla. Supongamos que lanzamos un dado cuatro veces y deseamos saber cuál es la probabilidad exacta de que en dos de los lanzamientos salga el «cinco». En este caso, N es el número de lanzamientos, esto es, 4; x es el número de «cincos», que es 2; P es la proporción esperada de obtener un cinco, que es $\frac{1}{6}$ (ya que se supone que cada una de las seis caras del dado pueden salir con la misma probabilidad), y $Q = 1 - P = 1 - \frac{1}{6} = \frac{5}{6}$. Ya con estos datos podemos calcular la probabilidad de que salga dos veces el «cinco» en cuatro lanzamientos de un dado, mediante la aplicación de la fórmula [6.1]:

$$p(2) = \frac{4!}{2!\,2!}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 = 0{,}11$$

Así, pues, la probabilidad de obtener exactamente dos veces el «cinco» en cuatro tiradas de un dado es $p = 0{,}11$.

Ahora bien, en la realidad de la investigación empírica, la pregunta que formularemos no es habitualmente la que se refiere a la probabilidad de obtener *exactamente* los valores que fueron observados, sino la siguiente pregunta: ¿cuál es la probabilidad de obtener los valores observados o incluso valores más extremos? Para responder a preguntas de este tipo recurrimos a la distribución muestral binomial, que viene dada por la expresión:

$$\sum_{i=0}^{x} \binom{N}{i} P^i Q^{N-i} \qquad [6.2]$$

que recoge la suma de la probabilidad del valor observado y las probabilidades de los valores más extremos.

Para continuar con el ejemplo anterior, supongamos que deseamos conocer la probabilidad de obtener *dos veces o menos* el «cinco» cuando lanzamos un dado cuatro veces. De nuevo es $N=4$, $x=2$, $P=1/6$ y $Q=5/6$; pero ahora se trata de calcular la probabilidad de obtener dos o menos veces el «cinco», esto es, $p \leq 2$. La probabilidad de obtener cero veces el «cinco» es $p\,(0)$; la probabilidad de obtener una vez un «cinco es $p\,(1)$, y la probabilidad de obtener dos veces el «cinco» es $p\,(2)$. Pues bien, aplicando el sumatorio [6.2] a las anteriores probabilidades, tenemos que:

$$p\,(x \leq 2) = p\,(0) + p\,(1) + p\,(2)$$

En otras palabras, que la probabilidad de obtener dos veces o menos el «cinco» es igual a la suma de las tres probabilidades señaladas. Si ahora utilizamos la fórmula [6.1] para calcular cada una de las tres probabilidades que aparecen en la parte derecha de la anterior expresión, obtenemos los siguientes resultados:

$$p\,(0) = \frac{4!}{0!\,4!}\left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^4 = 0{,}48$$

$$p\,(1) = \frac{4!}{1!\,3!}\left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^3 = 0{,}39$$

$$p\,(2) = \frac{4!}{2!\,2!}\left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2$$

y, por tanto:

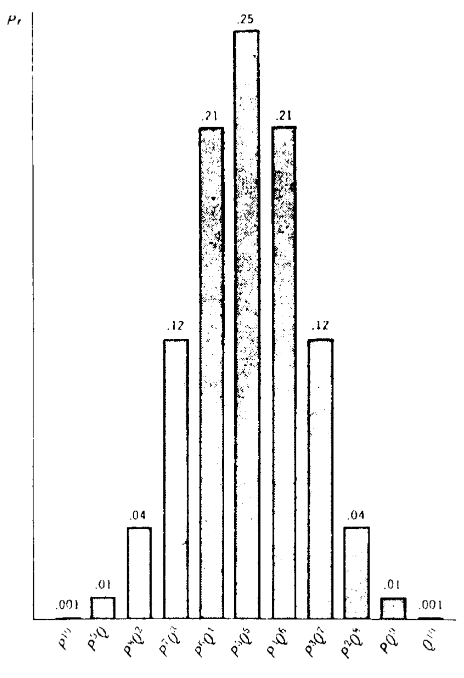
$$p\,(x \leqslant 2) = p\,(0) + p\,(1) + p\,(2) = 0{,}48 + 0{,}39 + 0{,}11 = 0{,}98$$

De este modo, pues, hemos determinado que la probabilidad, bajo H_0, de obtener dos veces o menos un «cinco» al lanzar un dado cuatro veces es $p=0{,}98$.

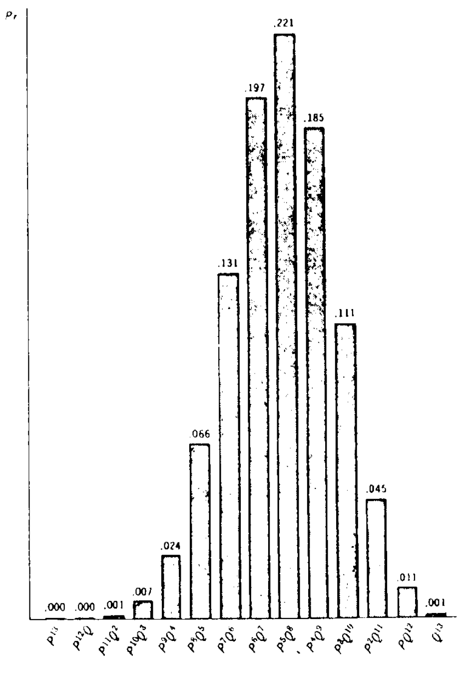
Ocupémonos ahora de estudiar algunas características (la tendencia central, la variación y la forma) de la distribución muestral binomial. Al tratarse de una distribución exacta, el N y el P en que se basa la distribución son parámetros. Dados estos parámetros es posible calcular una medida de la posición o tendencia central, que es también un parámetro. La media de una distribución binomial viene dada por la fórmula $\mu_B = NP$, en donde N es el tamaño de cada una de las muestras de la distribución y P es la proporción de resultados favorables. De este modo sencillo, podemos calcular la media como una medida de la posición o tendencia central de la distribución binomial.

También es posible calcular sencillamente otro parámetro como medida de la variación de la distribución binomial. La fórmula para el error típico de la distribución binomial es la siguiente: $\sigma_B = \sqrt{NPQ}$, en donde N y P tienen el mismo significado que en el caso anterior y, como ya sabemos, $Q = 1 - P$. Este error típico mide la variación de las frecuencias muestrales de resultados favorables alrededor de la media de la distribución muestral.

Por lo que se refiere a la forma de la distribución binominal, ésta depende de los valores que tomen N y P. Cuando $P = Q = 0{,}5$, la distribución será simétrica, y cuando N tiende al infinito, la distribución binomial tiende a aproximarse a la distribución normal. Incluso cuando P no es exactamente igual a 0,5 pero N es suficientemente grande, la distribución binomial tiende a parecerse a la distribución normal. Las siguientes figuras muestran la forma de la distribución binomial para el caso de igualdad de P y Q, forma simétrica, y para el caso de P ligeramente diferente de Q, forma ligeramente asimétrica:



Distribución binomial simétrica con $P = Q = 0.5$ y $N = 10$.



Distribución binomial ligeramente asimétrica con $P = 0.4$, $Q = 0.6$ y $N = 13$.

Cuando la distribución normal se aproxima a la distribución binomial, el error típico de la distribución binomial admite una interpretación similar al de la desviación típica. Esto es, aproximadamente el 68 por 100 de las frecuencias de la muestra quedan dentro del intervalo señalado por una unidad de error típico de la media, el 95 por 100 dentro de dos unidades de error típico y, aproximadamente, todas las frecuencias de la muestra quedan dentro de tres unidades de error típico. Ahora bien, cuando la distribución binomial se aleja de la normalidad, ya no es posible interpretar de este modo el error típico.

Tal como se ha señalado anteriormente, se han calculado tablas con las probabilidades asociadas para el caso de la prueba binomial, a las que se puede referir el investigador cuando está trabajando con la distribución binomial. La tabla C del apéndice recoge las probabilidades unilaterales o de una sola cola asociadas con la ocurrencia de diversos valores tan extremos como x bajo la hipótesis nula de $P=Q=1/2$. A efectos del uso de la tabla C, se hace coincidir x con la frecuencia observada más pequeña. El uso de la tabla C ya no hace necesaria la utilización de la fórmula [6.2], sobre todo cuando el valor de N es menor de 25. Sin embargo, cuando $P=Q$ y N es mayor de 25, hay que recurrir a la fórmula [6.2], ya que en tal caso no se puede utilizar la tabla C.

La tabla C contiene las probabilidades asociadas con la ocurrencia de diversos valores de frecuencia observados para diversos tamaños N de muestra (entre 5 y 25). Su utilización es bien sencilla. Supongamos que observamos que 6 casos pertenecen a una categoría y 4 pertenecen a otra, con lo que $N=10$ y $x=4$. La tabla C pone de manifiesto que la probabilidad unilateral de ocurrencia bajo H_0 de $x=4$ o menos, cuando $N=10$, es $p=0{,}377$.

Cuando no se conoce la dirección de la diferencia no se puede utilizar una prueba unilateral o de una sola cola y, por tanto, no se puede utilizar directamente la tabla C. Para una prueba bilateral o de dos colas hay que multiplicar por dos el valor de la probabilidad que ofrece dicha tabla. Con los mismos datos que en el ejemplo anterior, pero para el caso de una prueba bilateral o de dos colas, la probabilidad asociada con la ocurrencia bajo H_0 de un valor tal que x, la probabilidad es $p=2\,(0{,}377)=0{,}754$.

6.2.1. *Ejemplo del uso de la prueba binomial*

Veamos ahora, a través de otro ejemplo hipotético, el uso de la prueba binomial de decisión estadística para el caso en que $P=Q=0{,}5$. Los efectos de las películas con contenido violento sobre la población juvenil, es motivo de preocupación para muchos educadores y científicos sociales. La utilización de diversas fuentes de datos condujo a unos investigadores sociales a dividir en dos partes iguales a la población escolar de nivel primario de una comunidad, según que aceptasen o rechazasen la presencia de la violencia en la vida cotidiana. En un intento por encontrar métodos que facilitasen el incremento de las pautas de rechazo de la violencia, se pensó que la difusión en el colegio de películas de contenido artístico y científico podía ayudar en este sentido.

Para comprobar los efectos de una mayor exposición a este último tipo de películas, se eligió una muestra aleatoria de 20 niños a los que periódicamente se les hizo ver en el colegio películas de contenido artístico y científico durante un trimestre. Al finalizar el referido período, se les volvió a aplicar al grupo de 20 niños las mismas pruebas actitudinales sobre la aceptación-rechazo de la violencia que se habían aplicado con anterioridad al conjunto de la población escolar. La predicción realizada por el equipo de investigadores se formuló en el sentido de que el grupo de niños que habían sido sometidos en clase a las proyecciones y comentarios de películas de contenido artístico y científico, manifestarían un mayor rechazo de la violencia que el resto de sus compañeros. Los pasos seguidos para someter a la prueba binomial de decisión estadística los resultados del experimento fueron los siguientes:

a) *Hipótesis estadística.* La hipótesis nula será $H_0: p=P=0{,}5$. Esto es, la probabilidad de encontrar que un niño rechace la presencia de la violencia en la vida cotidiana es idéntica a la probabilidad de encontrar a uno que la acepte; cualquier diferencia que se observe en las frecuencias de los resultados de las pruebas actitudinales es de tal magnitud que puede esperarse en una muestra perteneciente a una población de resultados posibles bajo H_0. La hipótesis alternativa, de carácter unilateral, se formula como $H_1: p>P$.

b) *Prueba estadística.* Se elige la prueba binomial porque los datos pertenecen a dos categorías discretas y el diseño de la investigación es del tipo de una sola muestra. Dado que la muestra de niños se eligió aleatoriamente, no existe razón para suponer que dichos niños tuvieran actitudes previas hacia la violencia diferentes a las del resto de los niños que componen la población escolar estudiada, bajo H_0, con lo que $P=Q=0{,}5$.

c) *Nivel de significación.* Se estableció que $\alpha=0{,}05$. El número de casos $N=20$.

d) *Distribución muestral.* La distribución muestral viene dada por la fórmula [6.2], pero como N es menor de 25, y dado que $P=Q=0{,}5$, se puede utilizar la tabla *C*, que contiene las probabilidades asociadas con la ocurrencia bajo H_0 de valores observados tan pequeños como α, y que, por tanto, nos evita la necesidad de utilizar la anterior fórmula para calcular la distribución muestral para esta prueba.

e) *Región de rechazo.* La región de rechazo consiste en todos los valores de x que son tan pequeños que la probabilidad asociada con la ocurrencia bajo H_0 es igual o menor que $\alpha=0{,}05$. Como la dirección de la diferencia se ha establecido con antelación ($p>P$), la región de rechazo es de una sola cola o unilateral.

f) *Decisión.* En el experimento, 15 niños dieron resultados en las pruebas actitudinales de rechazo de la violencia, y 5 niños dieron resultados de aceptación en tales pruebas.

Así, pues, $N=20$, x=frecuencia menor=5. La tabla C pone de manifiesto que, para $N=20$, la probabilidad asociada con $x\leq 5$ es $p=0{,}021$. Como esta probabilidad es menor que $\alpha=0{,}05$, la decisión tomada por el equipo de investigadores fue la de rechazar H_0 en favor de H_1. La conclusión es, pues, que $p>P$ o, en otras palabras, que los niños que ven con frecuencia películas de contenido artístico y científico en el colegio, tienden a rechazar los actos violentos en la vida cotidiana en mayor proporción que los niños que siguen la programación habitual de los medios de comunicación de masas.

Tal como se ha señalado anteriormente, la tabla C sólo puede utilizarse cuando N vale 25 o menos. Esto quiere decir que, para valores superiores a 25, hay que recurrir a la fórmula [6.2] de la distribución muestral binomial. Ahora bien, se puede evitar esto recordando que la distribución binomial se aproxima a la normalidad cuando $P=Q=0{,}5$ y para valores suficientemente grandes de N. En tal caso, ya vimos anteriormente que la media $\mu_B=NP$ y el error tipo $\sigma_B=\sqrt{NPQ}$, con lo que la hipótesis nula H_0 puede ser sometida a prueba por medio de la expresión, referente a la distribución normal:

$$z=\frac{x-\mu_B}{\sigma_B}=\frac{x-NP}{\sqrt{NPQ}} \qquad [6.3]$$

donde z se distribuye de una forma aproximadamente normal con media cero y varianza la unidad.

De acuerdo con Siegel (*op. cit.*, págs. 40-41), esta aproximación se mejora notablemente si se incorpora una corrección para la continuidad. Tal corrección es necesaria, ya que, como se recordará, la distribución normal se basa en variables continuas, mientras que la distribución binomial se basa en variables discretas. Dicha corrección para la continuidad se realiza considerando la frecuencia observada x de la fórmula [6.3] como que ocupa un intervalo en el que el límite inferior se encuentra media unidad por debajo de la frecuencia observada, mientras que el límite superior se encuentra media unidad por encima de la frecuencia observada. La corrección para la continuidad consistirá en reducir en 0,5 la diferencia entre el valor observado de x y el valor esperado, $\mu_B=NP$.

Por tanto, cuando $x<\mu_B$ se añade 0,5 a x, mientras que cuando $x>\mu_B$ restamos 0,5 a x. Esto es, que la diferencia observada se reduce en 0,5. Entonces, el cálculo de z responde a la expresión:

$$z=\frac{(x\pm 0{,}5)-NP}{\sqrt{NPQ}} \qquad [6.4]$$

utilizándose $x+0{,}5$ cuando $x<NP$, y $x-0{,}5$ cuando $x>NP$. De este modo, se puede considerar que el valor de z, obtenido mediante la aplicación de la fórmula [6.4], se distribuye normalmente con media cero y varianza la unidad y, por tanto, se puede determinar la significación de un valor obtenido de z mediante la referencia a la tabla B del apéndice, que recoge la ley normal. Esto es, que la tabla B ofrece la probabilidad unilateral asociada con la ocurrencia, bajo H_0, de valores tan extremos como el z observado. Si se requiriese una prueba bilateral habría que multiplicar por dos la probabilidad p que ofrece la tabla B.

Para ver directamente el funcionamiento de la expresión [6.4] podemos aplicarla a los datos del ejemplo anterior. Recordemos que $N=20$, $x=5$ y $P=Q=0{,}5$. Para estos datos, $NP=20\,(0{,}5)=10$, y, por tanto, $x<NP$, ya que $5<10$, y, por tanto, la fórmula [6.4] queda así:

$$z=\frac{(5+0{,}5)-(20)\,(0{,}5)}{\sqrt{20\,(0{,}5)\,(0{,}5)}}=-2{,}01$$

La tabla B pone de manifiesto que un $z=-2{,}01$ tiene una probabilidad unilateral asociada con su ocurrencia bajo H_0 de $p=0{,}022$, que es prácticamente la misma probabilidad que se encontró anteriormente cuando se utilizó la tabla C de probabilidades exactas.

6.3. La prueba de chi-cuadrado (χ^2) para una sola muestra

En la investigación social, el sociólogo se interesa con frecuencia por el número de personas, objetos o respuestas que pertenecen a varias categorías. Así, por ejemplo, se puede clasificar a un grupo de personas según la preferencia ideológica de cada uno de sus miembros, medida mediante una escala izquierda-derecha de preferencia política, y el sociólogo puede predecir que ciertas posiciones de la escala serán más frecuentes que otras. También se puede caracterizar a un grupo de entrevistados según el «grado de acuerdo» (mucho, bastante, poco, nada) manifestado ante cierto tipo de opinión, y el sociólogo puede contrastar la hipótesis de que la frecuencia de las respuestas obtenidas para cada categoría serán diferentes.

Cuando se tienen datos de este tipo está aconsejado utilizar la prueba de chi-cuadrado (χ^2). La técnica χ^2 es del tipo de las que miden la bondad del ajuste, cuando se dispone del *número observado* de personas, objetos o respuestas que pertenecen a cada categoría y del *número esperado* basado en la hipótesis nula. La prueba de x^2 mide la existencia o no de una diferencia significativa entre ambos tipos de números o frecuencias.

Con el fin de poder comparar los valores observados con los valores esperados es preciso establecer qué frecuencias cabe esperar. La hipótesis nula se formula de modo que establece la proporción de personas, objetos o respuestas que pertenecen a cada una de las categorías en la población supuesta. Con lo que se pueden deducir de la hipótesis nula las frecuencias esperadas. Mediante la técnica de x^2 se puede probar si las frecuencias observadas se asemejan suficientemente a las frecuencias esperadas como para suponer que han ocurrido bajo H_0. La expresión algebraica que permite probar la hipótesis nula es la siguiente:

$$x^2 = \sum_{i=1}^{k} \frac{(0_i - E_i)^2}{E_i} \qquad [6.5]$$

en donde 0_i es el número observado de casos pertenecientes a la categoría i; E_i es el número esperado de casos pertenecientes a la categoría i, bajo H_0, y el sumatorio $\sum_{i=1}^{k}$ representa la suma de la expresión algebraica referida para todas las k categorías.

La interpretación inmediata de la expresión [6.5] es bien sencilla. Si las frecuencias observadas y esperadas se asemejan, las diferencias $(0_i - E_i)$ serán pequeñas y, consiguientemente, el valor de x^2 será pequeño. Por el contrario, si los valores se alejan entre sí, las diferencias serán grandes y, por tanto, el valor de x^2 también será elevado. Por ello, y hablando en términos generales, se puede afirmar que cuanto mayor sea el valor de x^2, mayor será la probabilidad de que las frecuencias observadas no provengan de la población en la que se basa la hipótesis nula.

Obsérvese otro aspecto, pero ahora de carácter terminológico. Con el fin de evitar confusiones, algunos autores (por ejemplo, Blalock, 1960, pág. 212, y Siegel, *op. cit.*, pág. 43) utilizan, al igual que hacemos aquí, el símbolo x^2 para el número que se calcula utilizando la fórmula [6.5] al realizar la prueba de x^2, mientras que la expresión «chi-cuadrado» se refiere a una variable aleatoria que se distribuye según lo hace la distribución muestral de chi-cuadrado, algunos de cuyos valores se contienen en la tabla D del apéndice.

Se puede demostrar que la distribución muestral de x^2, bajo H_0, tal como se calcula a partir de la fórmula [6.5], sigue la distribución chi-cuadrado con $df = k - 1$ grados de libertad (más adelante estudiaremos el significado de esta expresión). Como hemos dicho anteriormente, la tabla D que hemos reproducido en el apéndice pertenece a la distribución muestral del chi-cuadrado, y contiene ciertos valores críticos. Encabezando cada columna de dicha tabla aparecen las probabilidades bilaterales asociadas de ocurrencia, bajo H_0. Así, los valores que aparecen en cada columna serán los valores de chi-cuadrado que tienen la probabilidad asociada de ocurrencia, bajo H_0, dada en el encabezamiento de cada columna.

Veamos ahora el significado de los *grados de libertad, df*. Para cada valor de df existe un valor diferente de chi-cuadrado. El valor de df refleja el número de observaciones que pueden variar libremente después de haber establecido determinadas restricciones inherentes a la propia naturaleza de los datos. Así, por ejemplo, si los datos correspondientes a 30 casos se clasifican en dos categorías, tan pronto como sepamos que en una categoría hay 18 casos, sabremos de inmediato que en la segunda categoría habrán los 12 casos restantes. En este ejemplo, $df=1$, ya que, al disponer de dos categorías para un valor fijo de N, tan pronto como conozcamos el número de casos en una categoría se pueden determinar a continuación los casos pertenecientes a la segunda de las categorías. En general, y para el caso de una sola muestra, cuando la hipótesis nula H_0 especifica claramente el número de observaciones esperadas, los grados de libertad vendrán dados por la expresión $df=k-1$, en donde k representa el número de categorías que entran en la clasificación.

El uso del valor de x^2 para contrastar una hipótesis en el caso de una sola muestra (y una sola variable) es bien sencillo. En cada una de las k celdillas se colocan las frecuencias esperadas y las observadas en las muestras. Si la hipótesis nula se formula como que la proporción de casos en cada categoría es la misma, entonces $E_i=N/K$. Una vez conocidos los valores de E_i y 0_i se calcula el valor de x^2 mediante la expresión [6.5], y la significación del valor obtenido se determinará mediante el uso de la tabla D. Si la probabilidad asociada con la ocurrencia, bajo H_0, del valor obtenido de x^2 para $df=k-1$ es igual o menor que el valor previamente asignado de α, entonces se puede rechazar H_0. Si el valor de x^2 es mayor, entonces no se rechaza H_0.

6.3.1. *Ejemplo del uso de la prueba de x^2*

Algunos autores mantienen que el reclutamiento del profesorado universitario se hace preferentemente en determinadas clases o estratos sociales, y no en otras. Con el fin de comprobar esta hipótesis, se realizó un estudio entre una muestra del profesorado que había accedido a la categoría de profesor numerario durante el curso 1979-1980 en las universidades españolas. Se eligieron 180 profesores de una forma aleatoria y se clasificaron, según su origen social, en seis categorías, que representaban otros tantos estratos sociales. Estos estratos se construyeron de tal manera (en función del nivel de educación y renta y tipo de profesión del padre) que la población activa nacional quedaba distribuida en partes prácticamente iguales en cada uno de ellos.

La *hipótesis nula* H_0 se formuló del siguiente modo: no existen diferencias en el número esperado de profesores pertenecientes a los diversos estratos sociales considerados, y cualquier diferencia observada se debe a las fluctuaciones al azar que cabe esperar de una muestra aleatoria. Así, pues, $f_1=f_2=f_3=f_4=f_5=f_6$. La hipótesis alternativa H_1 será que las frecuencias f_i son diferentes.

La *prueba estadística* que se elegirá será la prueba de x^2, ya que vamos a comparar datos de una muestra con una supuesta población, con lo que la técnica del x^2 permitirá comparar las frecuencias observadas con las frecuencias esperadas en categorías discretas.

El *nivel de significación* lo establecemos en $\alpha=0{,}01$, siendo $N=180$ profesores.

La *distribución muestral* de x^2, tal como se calcula a partir de la fórmula [6.5], sigue la distribución de chi-cuadrado con $df=k-1$.

La *región de rechazo* permitirá rechazar H_0 si el valor observado de x^2 es tal que la probabilidad asociada con su ocurrencia, bajo H_0, para $df=6-1=5$ es igual o menor que $\alpha=0{,}01$.

La *decisión* se tomará ahora a la vista de los resultados. Los 180 profesores de la muestra se distribuyeron en los seis estratos sociales de la forma siguiente:

Estrato social	*1*	*2*	*3*	*4*	*5*	*6*	*Total*
Valores esperados	30	30	30	30	30	30	
Valores observados	33	28	35	24	35	25	180

En esta tabla se han incluido, en la misma celdilla correspondiente a cada estrato social, los valores esperados (que para este caso son $E_i=N/k=180/6=30$) y los valores observados, esto es, la distribución de los profesores según el estrato social de pertenencia. Ya con estos datos, el cálculo de x^2 es inmediato:

$$x^2=\sum_{i=1}^{K}\frac{(0_i-E_i)^2}{E_i}=$$

$$=\frac{(33-30)^2}{30}+\frac{(28-30)^2}{30}+\frac{(35-30)^2}{30}+\frac{(24-30)^2}{30}+\frac{(35-30)^2}{30}+\frac{(25-30)^2}{30}=$$

$$=\frac{9}{30}+\frac{4}{30}+\frac{25}{30}+\frac{36}{30}+\frac{25}{30}+\frac{25}{30}=4{,}13$$

En la tabla D se puede observar que $\chi^2\geqslant 4{,}13$ para $df=5$ tiene una probabilidad de ocurrencia que se encuentra entre $p=0{,}50$ y $p=0{,}30$.

Pero en tanto que esta probabilidad es claramente superior que el nivel de significación previamente establecido, $\alpha=0{,}01$, no podemos rechazar H_0 para dicho nivel de significación. En conclusión, pues, habrá que obtener más datos, y para un período más amplio, antes de tomar una decisión definitiva en relación a la hipótesis alternativa H_1, esto es, que los profesores universitarios provienen preferentemente de determinados estratos sociales.

Algunos autores (por ejemplo, Cochran, 1954) señalan diversos requisitos que deben cumplir los valores de las frecuencias esperadas para poder calcular x^2. Cuando $K=2$ y, por tanto, $df=1$, cada frecuencia esperada debe ser al menos 5. Cuando $K>2$ y, por tanto, $df>1$, la prueba x^2 para una sola muestra no debe utilizarse cuando más del 20 por 100 de las frecuencias esperadas sean menores de 5, o cuando cualquier frecuencia esperada sea menor de 1. En tales casos se puede superar este obstáculo recombinando categorías de tal forma que las frecuencias esperadas ofrezcan valores más altos. Por supuesto, cuando se recombinen categorías hay que tener cuidado de que las categorías que se sumen tengan un significado similar.

6.4. Distribuciones muestrales de las medias

En los ejemplos utilizados para glosar el funcionamiento de la distribución (y prueba) binomial y de la distribución (y prueba) del chi-cuadrado, los estadísticos observados eran frecuencias o proporciones. Sin embargo, muchas veces el interés del sociólogo se dirigirá al estudio de las medias. Así, puede desear estudiar la media de ingresos de un colectivo profesional, o la media de años de escolaridad de un grupo social. La media es un estadístico utilizado con mucha frecuencia porque ofrece la mayor información sobre la tendencia central de una distribución de puntuaciones relativamente simétrica.

Cuando el sociólogo calcula una media de una muestra trata de generalizar a la población de donde proviene la muestra. Así, al calcular los ingresos medios de una muestra de trabajadores, trata de formular una generalización sobre la media de los ingresos de la población trabajadora de la que extrajo la muestra. Pero para realizar tal generalización necesita conocer la distribución muestral de las medias.

Como sabemos, al calcular las medias de todas las muestras del mismo tamaño extraídas de una población se obtiene una distribución muestral de las medias. La medida del error muestral que indica la magnitud de las desviaciones de los estadísticos de la muestra alrededor de sus respectivos parámetros se denomina *error típico*. Pues bien, el error típico de la media es una medida de la variabilidad de las medias de las muestras, alrededor de la media de la población. Fijémonos en que, mientras la desviación típica mide la variabilidad de los valores alrededor de su media, el error típico de la media mide la variabilidad de las medias muestrales alrededor de la media de la población.

A partir de la fórmula de la desviación típica de la población, que es $\sigma = \sqrt{\Sigma (X-\mu)^2/N}$, se sustituye $\overline{X}$ por X y N_s por N para convertirla en la fórmula del error típico de la media, que será $\sigma_{\overline{X}} = \sqrt{\Sigma (\overline{X}-\mu)^2/N_s}$, en donde $\overline{X}$ es la media muestral, μ es la media de la población, N_s es el número de muestras y $\sigma_{\overline{X}}$ es el error típico de la media.

El valor del error típico se puede interpretar de la misma forma que la desviación típica, si la distribución muestral es normal o casi normal. Dada una distribución muestral normal, alrededor del 68 por 100 de las medias muestrales en la distribución de la muestra quedan dentro de una unidad de error típico de la media de la población.

Como se recordará, el teorema del límite central y la ley de los grandes números pueden considerarse una extensión de las propiedades de las medias muestrales anteriormente señaladas. La utilización directa del teorema del límite central y, lo que es mejor, de la ley de los grandes números puede servir para elaborar pruebas de decisión estadística muy sencillas. Para poner de manifiesto el funcionamiento del proceso de decisión estadística con datos de intervalo, comenzaremos con un modelo muy sencillo a través de un ejemplo sociológico.

6.4.1. *Prueba para la media de una población, cuando se conoce la desviación típica σ*

En una consulta que realizó una revista española de gran tirada entre su público lector femenino se encontró que el 32 por 100 de las 300 mujeres casadas que respondieron a la encuesta promovida por la revista afirmaba que mantenía relaciones sexuales extramatrimoniales. Dado que este porcentaje parece un tanto elevado, dado el tipo de valores sociales predominantes en la sociedad española, el sociólogo que supervisó la encuesta sospechó que las mujeres que habían respondido a la encuesta pertenecían a grupos sociales muy concretos —sobre todo, clase media-media y media-alta, de tipo urbano—, por lo que no podían considerarse representativas de la población femenina española. Para confirmar esta sospecha, el sociólogo disponía de algunos datos referentes a la situación socioeconómica de las mujeres que habían respondido a la encuesta y de los mismos datos referentes a la población en general. En concreto, sabía que la media de los ingresos familiares de las mujeres casadas que habían contestado a la encuesta era de 70.000 pesetas mensuales, mientras que la media mensual de los ingresos familiares de las familias españolas se situaba, en el momento de realizar el estudio, en 60.000 pesetas, con una desviación típica de 20.000 pesetas. A partir de estos datos, ¿cómo se puede comprobar que las mujeres que habían respondido a la encuesta constituyen una muestra sesgada y, por tanto, no representativa de la población femenina general?

Para hacerlo, el investigador recurrió a la ley de los grandes números, para lo cual tuvo que realizar algunos supuestos previos. En primer lugar hay que asumir que la muestra es aleatoria. En realidad, en esto consiste la prueba, ya que se desea saber si las mujeres que responden a la encuesta se puede o no considerar que constituyen una muestra aleatoria de la población femenina. También habrá que suponer que los datos referentes a la población general son exactos, ya que si no lo fueran no se podría realizar la prueba. Así, pues, la *hipótesis nula* H_0 es que se trata de una muestra aleatoria. El resto de los supuestos realizados acerca de la población constituyen el modelo estadístico. La hipótesis alternativa H_1 será que se trata de una muestra sesgada y que, por tanto, no ha sido extraída aleatoriamente de la población.

Dado que N es suficientemente grande, 300, podemos soslayar el supuesto de la normalidad de la población —que sería necesario si deseáramos utilizar el teorema del límite central— y pasar directamente a utilizar la ley de los grandes números. Además, hay que asumir que la media μ y la desviación típica σ de los datos de la población general son datos de intervalo, como de hecho así es, ya que representan unidades monetarias. Por tanto, tenemos los siguientes supuestos:

Nivel de medición: Escala de intervalo.

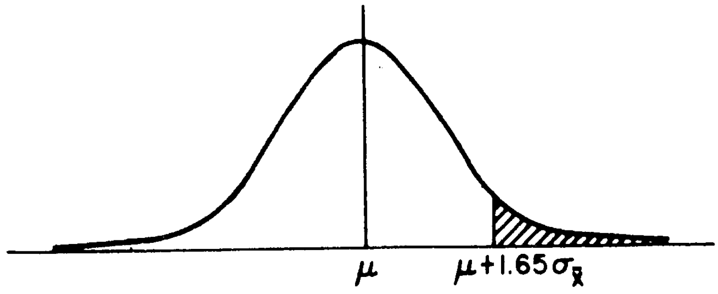
Modelo: Población normal. $\mu = 60.000$ pesetas; $\sigma = 20.000$ pesetas.

Hipótesis nula: Muestreo aleatorio.

La obtención de la *distribución muestral* es también sencilla, ya que en realidad está calculada con anterioridad. En efecto, como se sabe que la distribución muestral de las medias muestrales es normal o aproximadamente normal, se puede utilizar directamente la tabla normal.

El investigador eligió como *nivel de significación* $\alpha = 0{,}05$. Además, decidió utilizar una prueba unilateral o de una sola cola, ya que la dirección del sesgo ya ha sido establecida con anterioridad. Dado que la media de la muestra, 70.000 pesetas, es claramente superior a la media de la población, que es 60.000 pesetas, parece bien fundamentada la sospecha de que se encuentran sobrerrepresentadas en la muestra las mujeres de clase media y alta.

Una vez realizada la elección del nivel 0,05, y de una prueba unilateral, la región crítica o de rechazo viene determinada por la tabla normal. Dado que sólo el 5 por 100 del área de la curva normal se encuentra a la derecha de una ordenada que es 1,65 unidades de desviación típica mayor que la media, como se observa en la figura:



se puede saber ya que si el resultado obtenido es más de 1,65 unidades de desviación típica superior a la media μ que la hipótesis nula debe ser rechazada.

Realicemos ahora el cálculo de la prueba estadística. Se sabe que, si todos los supuestos son correctos, la distribución muestral de las medias $\bar{X}$ se distribuirá normalmente con una media μ y una varianza σ^2/N, es decir, $Nor\ (\mu, \sigma^2/N)$. Para los datos de que disponemos:

$$\mu = 60.000 \text{ ptas.}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{20.000}{\sqrt{300}} = 1.153 \text{ ptas.}$$

Ahora bien, para poder utilizar la tabla normal es preciso convertir los datos anteriores en puntuaciones típicas, esto es, obtener un estadístico z que tenga de media 0 y de varianza la unidad, Nor (0,1). Recordemos que, al estudiar la curva normal, utilizamos la fórmula:

$$z = \frac{X - \bar{X}}{s}$$

que es aplicable a la distribución de una muestra que tenga como media $\bar{X}$ y como varianza s^2, pero no resulta aplicable a una distribución muestral.

Recapitulemos lo que hemos hecho hasta ahora. En primer lugar formulamos una serie de supuestos con objeto de obtener una distribución muestral que nos va a permitir saber cuál es la probabilidad de una media $\bar{X}$ dada si los supuestos son verdaderos. A partir de la muestra, el investigador obtiene un solo valor de $\bar{X}$ y, a continuación, utilizará la distribución muestral teórica con el fin de evaluar la probabilidad de obtener un resultado tan poco corriente o más poco corriente que el valor particular de $\bar{X}$. La distribución muestral que utiliza viene dada, en realidad, por la tabla normal. En esta distribución, cada «caso» es un valor $\bar{X}$, la media es μ y la desviación típica es $\sigma/\sqrt{N}$. De este modo,

$\overline{X}$ reemplaza a X, μ reemplaza a $\overline{X}$ y $\sigma/\sqrt{N}$ reemplaza a s en la anterior fórmula de z, con lo que dicha expresión se puede escribir del siguiente modo:

$$z=\frac{\overline{X}-\mu}{\sigma/\sqrt{N}}=\frac{70.000-60.000}{1.153}=8{,}67$$

Es decir, que la media muestral es 8,67 unidades de desviación típica mayor que la media de la población.

La *decisión* no admite dudas. Dado que $\overline{X}$ se desvía con respecto a la media μ asumida en más de 1,65 unidades de desviación típica, la hipótesis nula H_0 debe rechazarse al nivel $\alpha=0{,}05$. En otras palabras, las lectoras de la revista que habían respondido a la encuesta no constituyen una muestra aleatoria y representativa de la población general femenina, sino que se trata de un grupo concreto y no representativo del conjunto de dicha población.

6.4.2. *La distribución t de Student*

No siempre podemos operar de la forma que lo hemos hecho en la sección anterior, por la sencilla razón de que se desconoce el valor de la desviación típica σ. Una posible solución consiste en sustituir la desviación típica de la población σ por la desviación típica de la muestra s. En la fórmula de z, el cociente $\sigma/\sqrt{N}$ se puede sustituir simplemente por $s/\sqrt{N}$ y, dado que s se puede calcular a partir de los datos de la muestra, ya se puede obtener el valor de z. Ahora bien, esta sustitución ofrece resultados razonables cuando el tamaño N de la muestra es suficientemente grande. Cuando N es pequeño, los resultados aparecen distorsionados.

Con el fin de obviar esta dificultad se puede utilizar una prueba estadística alternativa, que tiene una distribución muestral conocida, llamada la distribución t de Student. Tal distribución fue introducida por el matemático irlandés W. S. Gosset (1876-1977), quien fue el que descubrió que, para tamaños pequeños de N, la utilización de la desviación típica de la muestra s ofrece una distribución muestral de las medias que no es normal. Gosset publicó sus investigaciones en 1908, bajo el seudónimo de «Student», y con este nombre ha pasado a la historia de la estadística moderna.

La distribución muestral t de Student responde a la siguiente expresión:

$$t=\frac{\overline{X}-\mu}{s/\sqrt{N-1}} \qquad [6.6]$$