Tema 5

Inferencia Estadística.

5.1 Introducción a la Inferencia Estadística

La palabra inferir significa extraer consecuencias, o deducir un conocimiento a partir de otro. La Inferencia Estadística es la parte de la estadística que se encarga de deducir características de la población a partir de los resultados obtenidos en muestras de esta población. Las decisiones se basan en la información contenida en muestras extraídas de ella.

En muchas circunstancias hay que tomar decisiones basándose sólo en la información contenida en una muestra: Un gerente de Control de Calidad debe determinar si un proceso funciona correctamente. Para ello, cada cierto tiempo, analiza la calidad de una pequeña cantidad de productos fabricados por este proceso. Con esta información debe decidir si continúa fabricando nuevas piezas, o si debe realizar algún ajuste o reparación de la maquinaria de la fábrica antes de continuar el proceso de fabricación. Un gerente de Marketing debe determinar si una nueva estrategia de mercado aumentará las ventas. Para ello se basará fundamentalmente en encuestas realizadas a unos cuantos clientes potenciales, etc. Para adoptar estas decisiones se toma toda la información posible de la muestra seleccionada y se estudia, en términos de probabilidad, el grado de fiabilidad de las decisiones adoptadas.

Podemos distinguir de modo general dos grandes métodos dentro de la Inferencia Estadística:

Métodos Paramétricos.- Se supone que los datos provienen de una familia de distribuciones conocida (Normal, Poisson,...) y que lo único que se desconoce es el valor concreto de alguno de los parámetros que la definen (μ y σ para la Normal, λ para la Poisson,...).

Se pueden hacer inferencias acerca de los parámetros poblacionales de dos maneras. Dando valores aproximados para los parámetros (Estimación) o tomando decisiones con respecto a ellos (Contrastes de Hipótesis).

Métodos No Paramétricos.- No suponen conocida la distribución, y solamente suponen hipótesis muy generales respecto a las mismas. Estos métodos se aplican en los tests de bondad de ajuste, que prueban la adecuación de los datos a ciertos modelos de distribuciones teóricas, los test de independencia, etc.

Evidentemente, las conclusiones que obtengamos y que generalizaremos para toda la población dependerán de los valores concretos que se hayan observado en la muestra. Muchas personas manifiestan su desconfianza y su recelo sobre las conclusiones obtenidas con métodos estadísticos, debido, entre otras causas, a que estas conclusiones dependen de la muestra extraída, y que las muestras presentan fluctuaciones aleatorias. Sin embargo, en la vida cotidiana, nuestras opiniones y nuestros comportamientos se basan en generalizaciones que hacemos a partir de muestras. Así, es muy frecuente que manifestemos que los productos de una determinada marca son mejores que los de la competencia. Dicha afirmación no la hacemos, evidentemente, tras un análisis exhaustivo de todos los productos de una y otra marca, sino basándonos en nuestra propia experiencia personal, que es claramente muy limitada. Es decir, generalizamos a partir de observaciones realizadas en muestras pequeñas.

5.2 Tipos de estimación

Cuando se desean estimar los parámetros de la población a partir de los de la muestra se consideran dos formas de realizar dicha estimación.

Estimación puntual.- En la estimación puntual damos un solo punto como valor estimado del parámetro. Por ejemplo, si queremos estimar la altura media, μ, de los varones españoles de 20 años, obtendremos una muestra aleatoria de cierto tamaño de esta población, hallaremos la altura media de las personas seleccionadas en esta muestra y diremos que este valor, el de la media muestral, es una estimación puntual de la altura media de la población de varones de 20 años.

Estimación por intervalos.- En realidad, cuando realizamos una estimación puntual, nos damos cuenta que es muy difícil que ésta estimación sea realmente el verdadero valor del parámetro desconocido. Tendremos más oportunidades de acertar si indicamos que el parámetro desconocido pertenece a un cierto intervalo. En el ejemplo de la altura media de los varones de 20 años, si la media muestral resultara 1.75 m., podríamos decidir manifestar que la media verdadera pertenece al intervalo $(1.75 - 0.05, 1.75 + 0.05)$. El intervalo en el que se afirma que se encuentra el parámetro poblacional se denomina *intervalo de confianza*. Tampoco en este caso podemos estar seguros de que el valor real pertenezca a dicho intervalo. Por este motivo suele decirse que el valor real del parámetro pertenece a dicho intervalo con un cierto "grado de confianza". La cuantificación de la confianza que se tiene en que el parámetro desconocido esté verdaderamente en el intervalo dado se denomina *grado de confianza* y es una medida relacionada con la función de distribución de probabilidad del parámetro en estudio.

5.3 Estadísticos y Estimadores

Un *estadístico* es una función de los elementos de la muestra. Si tenemos una población en la que estamos observando una característica que se distribuye según una variable aleatoria X, y consideramos una muestra aleatoria simple de tamaño n

$$x_1, x_2, \ldots, x_n$$

podemos calcular el siguiente estadístico $\overline{X}$:

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Evidentemente, el valor del estadístico dependerá de los valores que hayan tomado los elementos de la muestra. Si repetimos el experimento de tomar una muestra y calculamos de nuevo el valor del mismo estadístico, obtendremos, por lo general, otro valor distinto. Tenemos por tanto que el estadístico es una variable aleatoria. La distribución que seguirá dicha variable aleatoria dependerá de la distribución de la variable X. En determinados casos podremos calcular la distribución del estadístico.

Un *estimador* de un parámetro poblacional es un estadístico que se utiliza para obtener un valor aproximado de ese determinado parámetro de la población. Por ejemplo, la media muestral es el estadístico que suele usarse más frecuentemente para estimar la media poblacional. Entonces, la media muestral es un estimador de la media poblacional. La mediana y la moda son también estimadores de la media poblacional. Para indicar que T es un estimador del parámetro poblacional θ se indicará

$$T = \hat{\theta}$$

El valor que toma este estimador en la muestra concreta que estamos considerando es una *estimación* del parámetro desconocido.

5.4 Propiedades de los estimadores

¿Cómo se eligen estimadores para la media y la varianza de la población? Más en general, ¿cómo se eligen los estimadores de los parámetros poblacionales? Es normal que exijamos que los estimadores, al menos en promedio, se parezcan al parámetro que quiere estimarse. También es conveniente que no fluctúen demasiado con las distintas muestras y que mejoren si aumentamos el tamaño de ésta. Estas condiciones son las qué están formuladas en las siguientes definiciones.

Centrado o insesgado

Una de las propiedades que con más frecuencia se le exige a los estimadores es que sean *insesgados*. Decimos que T es un estimador centrado o insesgado del parámetro θ si para cualquier tamaño muestral se cumple que

$$E(T) = \theta$$

Eficiencia de dos estimadores

Si tenemos dos estimadores T_1 y T_2 de un parámetro θ, decimos que T_1 es más eficiente que T_2 si para cualquier tamaño muestral se verifica que

$$\mathrm{Var}(T_1) \leq \mathrm{Var}(T_2)$$

Entre dos estimadores posibles sería preferible el más eficiente.

Consistencia de un estimador

Diremos que un estimador es consistente si cumple:

$$\lim_{n\to\infty} E(T) = \theta \qquad \text{y} \qquad \lim_{n\to\infty} \mathrm{Var}(T) = 0$$

donde n es el tamaño de la muestra.

5.5 Estimadores insesgados de la media y la varianza

Para una variable aleatoria X, con media μ y varianza σ^2, un estimador insesgado de la media poblacional es

$$\hat{\mu} = \overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

y un estimador insesgado de la varianza es

$$\widehat{\sigma^2} = S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Es decir, la media muestral es un estimador insesgado de la media poblacional y la cuasivarianza muestral es un estimador insesgado de la varianza poblacional.

5.6 Distribución de los estadísticos muestrales

Cuando se realiza una estimación por medio de intervalos de confianza se da un *grado de confianza.* Este valor se basa en la proporción de muestras en las que el parámetro que se desea estimar quedaría dentro del intervalo de confianza dado. Para calcular esta proporción es necesario conocer la distribución del estimador en el muestreo. Con el propósito de conocer el grado de confianza asociado a las estimaciones por intervalo de la media y de la varianza poblacionales son útiles los siguientes teoremas.

Teorema 2 *En el caso de que la variable X sea normal, podemos afirmar que*

$$\overline{X} \in N(\mu, \sigma/\sqrt{n})$$

y que

$$\frac{(n-1)S_c^2}{\sigma^2} \in \chi^2_{n-1}$$

esto es, una chi-cuadrado con n-1 grados de libertad.

Nota: S_c es la cuasidesviación muestral.

Si la población no es normal y n es grande podemos afirmar, por el Teorema Central del Límite que

$$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$$

A efectos prácticos se emplea esta aproximación si $n \geq 30$.

Teorema 3 *Si X es una variable aleatoria $N(\mu, \sigma)$*

$$T = \frac{\overline{X} - \mu}{S_c/\sqrt{n}}$$

se distribuye según una t de Student con n-1 grados de libertad.

Si la X no sigue una distribución normal, este resultado puede aplicarse a efectos prácticos si el número de elementos de la muestra es mayor o igual que 60.

5.7 Intervalos de confianza para la media

En este epígrafe consideraremos únicamente poblaciones normales

5.7.1 Con varianza conocida

Supongamos que la población se distribuye según una variable

$$X \in N(\mu, \sigma)$$

y que la desviación típica σ es conocida.

Si tomamos muestras de tamaño n y calculamos el estadístico $\overline{X}$, se tiene que segun el teorema 17.3

$$\overline{X} \in N(\mu, \sigma/\sqrt{n})$$

Un *intervalo de confianza* para la media poblacional, centrado en la media muestral, $\left(\overline{X} - b, \overline{X} + b\right)$, con una *grado de confianza* $100(1-\alpha)\%$, es un intervalo que cumple la siguiente condición:

$$P(\overline{X} - b < \mu < \overline{X} + b) = 1 - \alpha$$

lo que quiere decir que la media poblacional estaría contenida en el intervalo obtenido a partir de la muestra, en el $100(1-\alpha)\%$ de las muestras. El valor de α se conoce con el nombre de *nivel de significación.*

Para obtener dicho intervalo se tienen en cuenta las siguientes consideraciones:

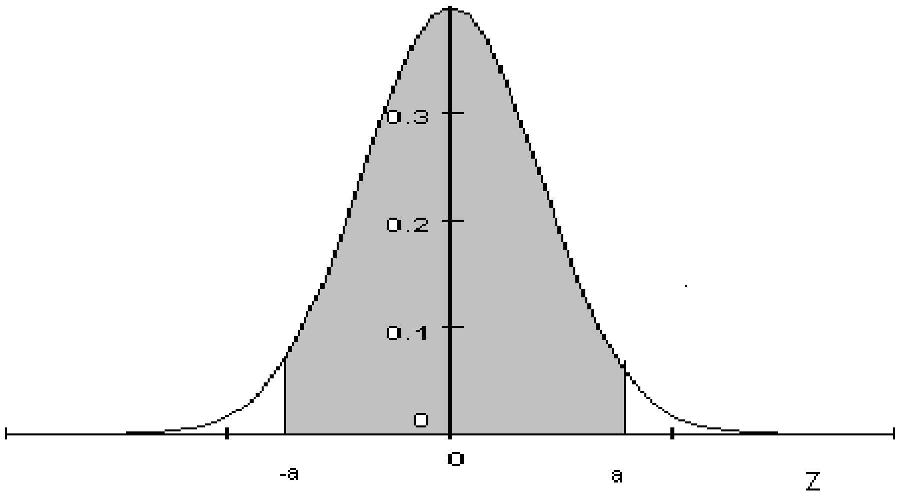
Tipificando la variable $\overline{X}$, se obtiene

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \in N(0,1)$$

Buscamos un valor a que cumpla

$$P(-a < Z < a) = 1 - \alpha$$

Gráficamente, considerando la función de densidad de la $N(0,1)$, tendríamos la siguiente situación:



La zona gris de la figura debería tener un área $1-\alpha$, así que cada una de las zonas laterales tiene un área $\frac{\alpha}{2}$,

Al ser $Z \in N(0,1)$, el valor de a se busca en las tablas de la normal estándar, considerando que ha de cumplir:

$$P(Z < a) = 1 - \alpha/2$$

Sustituyendo a por el valor hallado en las tabla, que denotamos[1] por $z_{1-\alpha/2}$, y teniendo en cuenta la simetría de la función de densidad, se obtiene:

$$P(-z_{1-\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}) = 1-\alpha$$

Despejando μ en ambas desigualdades, se obtiene:

$$P(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}) = 1-\alpha$$

Así que el intervalo de confianza para la media poblacional μ, al nivel de significación α, será:

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2},\ \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2})$$

Ejemplo 31 *Calculemos un intervalo de confianza al 90%, para la media de una población normal a partir de la siguiente muestra extraída de ella.*

$6.411, 4.324, 5.2825, 3.2689, 3.705, 4.9736, 1.8977, 4.4681, 3.7961, 4.9666$

Suponemos conocida la desviación típica de la población $\sigma = 1.2$.

[1] Frecuentemente se usa también la notación $z_{\alpha/2}$

Obtenemos la media muestral que resulta ser: $\overline{X} = 4.3094$

El intervalo de confianza para la media de la población, con una confianza del 90%, es:

$(4.3094 - \frac{1.2}{\sqrt{10}}1.6449,\ 4.3094 + \frac{1.2}{\sqrt{10}}1.6449) = (3.6852,\ 4.9336)$

El valor de $z_{1-\alpha/2} = z_{1-0.10/2} = z_{0.95}$, se ha calculado con la condición: $P(z < z_{0.95}) = 0.95$. Mirando en la tabla de la N(0,1) resuta que el valor de $z_{0.95} = 1.6449$.

5.7.2 Con varianza desconocida

Supongamos que la población se distribuye según una variable

$$X \in N(\mu, \sigma)$$

con σ desconocida. Si tomamos una muestra de tamaño n, y calculamos el estadístico $\overline{X}$, se tiene que

$$\overline{X} \in N(\mu, \sigma/\sqrt{n})$$

de donde, tipificando, se obtiene

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \in N(0,1)$$

Pero ahora, al no conocer el valor de σ no podemos calcular exactamente el valor de Z.

Estimando el valor de σ por su estimador S_c, se obtiene el estadístico

$$T = \frac{\overline{X} - \mu}{S_c/\sqrt{n}} \tag{5.1}$$

que ya no se distribuiría según una normal sino que, según el teorema 3, se distribuye como una t de Student con $n-1$ grados de libertad.

Si queremos obtener un intervalo que tenga una probabilidad $1-\alpha$ de contener al parámetro μ, esto equivaldría a

$$P(-t_{1-\alpha/2} < T < t_{1-\alpha/2}) = 1 - \alpha \tag{5.2}$$

Al ser T una t de Student con $n-1$ grados de libertad, el valor de $t_{1-\alpha/2}$ se busca en las tablas de la citada distribución. Considerando que

$$P(T < t_{1-\alpha/2}) = 1 - \alpha/2$$

y sustituyendo la expresión 5.1 de T en la 5.2, se obtiene:

$$P(-t_{1-\alpha/2} < \frac{\overline{X} - \mu}{S_c/\sqrt{n}} < t_{1-\alpha/2}) = 1 - \alpha$$

Despejando μ en la doble desigualdad deducimos que:

$$P(\overline{X} - \frac{S_c}{\sqrt{n}} t_{1-\alpha/2} < \mu < \overline{X} + \frac{S_c}{\sqrt{n}} t_{1-\alpha/2}) = 1 - \alpha$$

Así que el intervalo de confianza para la media poblacional μ con una confianza $100(1-\alpha)\%$ es:

$$(\overline{X} - \frac{S_c}{\sqrt{n}} t_{1-\alpha/2},\ \overline{X} + \frac{S_c}{\sqrt{n}} t_{1-\alpha/2}) \tag{5.3}$$

Ejemplo 32 *Hallar un intervalo de confianza al 95% para la media basado en la misma muestra del ejemplo anterior:*

$6.411, 4.324, 5.2825, 3.2689, 3.705, 4.9736, 1.8977, 4.4681, 3.7961, 4.9666$

Se supone ahora que la desviación típica de la población de la que procede dicha muestra es desconocida.

Calculamos en primer lugar la media y la cuasidesviación de la población:

$$\overline{X} = 4.3094 \quad S_c = 1.2378$$

El valor de $t_{1-\alpha/2} = t_{1-0.05/2} = t_{0.975}$ se ha calculado con la condición: $P(t < t_{0.975}) = 0.975$. Por tanto, mirando en la tabla de la t de Student con con 9 grados de libertad obtenemos que su valor es: $t_{0.975} = 2.2622$.

Sustituyendo en la expresión 5.3, los valores obtenidos y teniendo en cuenta que el número de elementos de la muestra es 10, obtenemos el siguiente intervalo de confianza para la media poblacional:

$$(4.3094 - \frac{1.2378}{\sqrt{10}} 2.2622,\ 4.3094 + \frac{1.2378}{\sqrt{10}} 2.2622) = (3.4239, 5.1949)$$

5.8 Intervalos de confianza para la varianza

Los intervalos de confianza para la varianza se basan en la distribución del estadístico $\frac{(n-1)S_c^2}{\sigma^2}$ que según el teorema 17.3, sigue una distribución chi-cuadrado con $n-1$ grados de libertad, si la variable de partida es normal..

Para obtener un intervalo de confianza partimos de la relación

$$P\left(a < \frac{(n-1)S_c^2}{\sigma^2} < b\right) = 1 - \alpha$$

Considerando la siguiente gráfica de la distribución chi-cuadrado, asignamos al área central de color gris el valor $1-\alpha$, de modo que

$$P\left(\chi^2_{n-1} < a\right) = \tfrac{\alpha}{2} \qquad \text{y} \qquad P\left(\chi^2_{n-1} < b\right) = 1 - \tfrac{\alpha}{2}$$

Por tanto, si denotamos por a y b los valores de chi-cuadradocon $n-1$ grados de libertad. que dejan delante las areas $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$ respectivamente:

$$a = \chi^2_{n-1,\frac{\alpha}{2}} \qquad y \qquad b = \chi^2_{n-1,1-\frac{\alpha}{2}}$$

obtenemos:

$$\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)S_c^2}{\sigma^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}$$

Despejando σ^2

$$\frac{(n-1)\,S_c^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)\,S_c^2}{\chi^2_{n-1,\frac{\alpha}{2}}}$$

De ello deducimos que un intervalo de confianza para la varianza poblacional σ^2, con nivel de confianza $(1-\alpha)\%$ es

$$\left(\frac{(n-1)\,S_c^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}},\ \frac{(n-1)\,S_c^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) \tag{5.4}$$

y para la desviación típica

$$\left(S_c\sqrt{\frac{n-1}{\chi^2_{n-1,1-\frac{\alpha}{2}}}},\ S_c\sqrt{\frac{n-1}{\chi^2_{n-1,\frac{\alpha}{2}}}}\right)$$

Ejemplo 33 *Hallar un intervalo de confianza para la desviación típica de la población, usando la información de la muestra de los dos ejemplos anteriores.*

$$\left(1.2378\sqrt{\frac{9}{19.023}},\ 1.2378\sqrt{\frac{9}{2.7004}}\right) = (0.8514, 2.2597)$$

siendo los denominadores:

$\chi^2_{n-1,\frac{\alpha}{2}} = \chi^2_{10-1,\frac{0.05}{2}} = \chi^2_{9,0.025} = 2.7004$

$\chi^2_{n-1,1-\frac{\alpha}{2}} = \chi^2_{10-1,1-\frac{0.05}{2}} = \chi^2_{9,0.975} = 19.023$

Estos valores se obtienen a partir de la tabla de la distribución Chi-cuadrado.

5.9 Contrastes de hipótesis

5.9.1 Introducción:

Proponemos, como paso previo, un ejemplo sencillo que nos ayudará a entender los conceptos fundamentales que se ponen en juego a la hora de construir un contraste de hipótesis (también llamado prueba de hipótesis o test de hipótesis).

Supongamos que tratamos de comprobar si una moneda está bien construida, es decir si cara y cruz tienen la misma probabilidad de aparición. Para ello realizamos una prueba con esta moneda consistente en arrojarla 50 veces. Si esta experiencia ha dado como resultado 49 veces cara y 1 vez cruz. ¿Qué concluiríamos? Seguramente no aceptaríamos que la moneda esté bien construida, a pesar de que el resultado obtenido es posible para una moneda bien construida. Incluso sería posible obtener 50 caras. Entonces ¿por qué la moneda no nos parece aceptable? Intuitivamente percibimos que el resultado que hemos obtenido sería “demasiado raro” para una moneda simétrica. Esperábamos que el resultado estuviera más cerca del promedio: 25 caras y 25 cruces.

Antes de tomar una decisión deberíamos adoptar un criterio razonable para aceptar o rechazar la simetría de la moneda. ¿Qué diferencia con respecto al promedio estaríamos dispuestos a aceptar? ¿Qué número de caras y cruces nos parecerían valores “demasiado raros”?

La respuesta sólo podemos darla en términos de probabilidad, ya que sabemos que los resultados dados en el ejemplo, y considerados “raros”, son posibles. Es decir, debemos aceptar un cierto riesgo: El de catalogar como mal construida una moneda aceptable.

Supongamos que decidimos no admitir que la moneda es simétrica cuando, como en el ejemplo, la diferencia en número de caras o cruces con respecto al valor medio, 25, sea mayor o igual que 24. ¿Cuál sería la probabilidad de declararla erróneamente como defectuosa?

Rechazaríamos una moneda bien construida si se diera uno de los siguientes sucesos:

a) si salen 49 caras y 1 cruz.

b) si salen 50 caras.

c) si salen 49 cruces y una cara.

d) si salen 50 cruces.

La probabilidad de rechazar erróneamente la moneda sería entonces:

$\binom{50}{1}0.5^{49}\times0.5+0.5^{50}+\binom{50}{1}0.5^{49}\times0.5+0.5^{50}=2\times(\binom{50}{1}0.5^{49}\times0.5+0.5^{50})=$

.

$=9.0594\times10^{-14}$

Si la moneda fuera buena este suceso es muy poco probable. Por eso cuando obtenemos ese resultado, 49 caras y una cruz, pensamos que ha ocurrido algo "demasiado raro" para una moneda bien balanceada. Pensamos que lo normal sería desconfiar y decidimos que la moneda es defectuosa.

Con el objeto de ayudarnos a decidir cuando el suceso nos parece "muy poco probable" obtenemos la probabilidad de los sucesos siguientes:

a) Que el número de caras o de cruces sea menor o igual que cinco

$$2\sum_{i=0}^{5}\left(\binom{50}{i}0.5^{i}\times0.5^{50-i}\right)=4.2099\times10^{-9}$$

b) Que el número de caras o de cruces sea menor o igual que diez

$$2\sum_{i=0}^{10}\left(\binom{50}{i}0.5^{i}\times0.5^{50-i}\right)=2.3861\times10^{-5}$$

c) Que el número de caras o de cruces sea menor o igual que quince

$$2\sum_{i=0}^{15}\left(\binom{50}{i}0.5^{i}\times0.5^{50-i}\right)=6.6004\times10^{-3}$$

d) Que el número de caras o de cruces sea menor o igual que diecisiete.

$$2\sum_{i=0}^{17}\left(\binom{50}{i}0.5^{i}\times0.5^{50-i}\right)=3.2839\times10^{-2}$$

Naturalmente que la decisión depende del riesgo que queramos correr (rechazando una moneda buena). Hay que resaltar que si disminuye el riesgo de rechazar una moneda buena, aumenta el riesgo de aceptar como buena una moneda mal equilibrada, por lo que debemos llegar a un compromiso entre ambas opciones.

Llamamos nivel de significación, α, a la probabilidad de rechazar una moneda aceptable. Así, si aceptamos para α el valor 0.032839, que corresponde al último suceso considerado, admitiríamos la moneda como buena si el número de caras o de cruces estuviera entre 17 y 33. Este intervalo, [17, 33], sería la *región de aceptación* de la prueba y el resto de los valores formarían la *región de rechazo.* Los valores de separación, 17 y 33, recibirían el nombre de *valores críticos* para esta prueba.

En resumen: Se pretende diseñar una prueba que nos permita dar un criterio para aceptar una hipótesis de partida: La moneda está bien construida, que se conoce con el nombre de *Hipótesis nula,* o rechazarla y aceptar la hipótesis contraria: la moneda no está bien construida, que se suele conocer con el nombre de *Hipótesis alternativa.*

El modo de actuar para aceptar o rechazar la moneda en cuestión podría ser el siguiente: La prueba consistiría en arrojar la moneda 50 veces. Se aceptaría que la moneda es buena si el número de caras o de cruces estuviera entre 17 y 33 y se rechazaría en caso contrario. El nivel de significación de la prueba (probabilidad de rechazar la hipótesis nula siendo verdadera) sería $\alpha = 0.032839$.

5.9.2 Conceptos generales

Una *hipótesis estadística* es una proposición referente a una o varias poblaciones. A menudo se refieren a su distribución de probabilidad o al valor de sus parámetros. Por ejemplo: La distribución de probabilidad de la moneda es uniforme discreta, el valor de la media de una variable aleatoria es 34, etc.

Un *test, prueba o contraste de hipótesis* es un conjunto de reglas para decidir cual de dos hipótesis, H_0,(*Hipótesis nula*) o H_1,(*hipótesis alternativa*), debe aceptarse en base a la información obtenida con una muestra. Una de estas hipótesis es la negación de la otra. Por lo general se supone que la hipótesis nula es algo que se ha admitido durante un cierto periodo de tiempo, que está vigente y que se mantendrá, salvo que haya pruebas que favorezcan claramente a la hipótesis alternativa.

Para realizar un test de hipótesis hay que decidir un valor para α *(nivel de significación o probabilidad de error tipo I*), que es la probabilidad de rechazar la hipótesis nula siendo cierta.

Para realizar el test se supone que se cumple la hipótesis nula, H_0. La decisión se basa en un estadístico (*Estadístico de contraste*) que comprime la información relevante de la muestra), y que, si se cumple la hipótesis nula, se rige por una ley de probabilidad conocida Para realizar la decisión se obtiene un intervalo de aceptación para el estadístico de contraste (*región de aceptación*). La región complementaria, formada por los valores que no pertenecen a la región de aceptación, suele llamarse *Región crítica* o de rechazo.

Si la muestra da al estadístico un valor dentro de la región de aceptación se acepta H_0, y se dice que el test es *estadísticamente no significativo.* En caso contrario se acepta H_1.y se dice que el test es *estadísticamente significativo.*

En un test de hipótesis pueden cometerse dos tipos de errores:

	H_0 es cierta	H_0 es falsa
Se acepta H_0	Acierto	Error tipo II
Se rechaza H_0	Error tipo I	Acierto

La decisión por H_1 viene acompañada de una probabilidad de error dado por α (Probabilidad de cometer un error tipo I o *nivel de significación* del test) que es la probabilidad de decidirse por la hipótesis alternativa siendo cierta la hipótesis nula.

La decisión por H_0 viene acompañada de una probabilidad de error dado por β (Probabilidad de cometer el error tipo II) que es la probabilidad de decidirse por la hipótesis nula siendo cierta la alternativa. Este error no esta controlado de antemano, por eso la decisión de aceptar H_0 no es de fiar. Se denomina *Potencia* del test al valor de $1-\beta$ (probabilidad de decidirse por la hipótesis alternativa, si es cierta).

Un test de hipótesis lleva asociado un intervalo de confianza para los parámetros implicados.

Para aclarar estos conceptos teóricos proponemos el siguiente ejemplo:

Ejemplo 34 *Se pretende diseñar una prueba de hipótesis con una muestra de 74 automóviles para comprobar su capacidad de frenado. Para ello se medirá en todos ellos la distancia de frenado si el automóvil parte de una velocidad inicial de 100 Km/h. Se quiere saber si, tras un frenazo brusco, la distancia media recorrida antes de pararse es de 110 metros. Se supone que la distancia de frenado sigue una distribución normal con desviación típica conocida* $\sigma = 3$ *m.*

Consideramos las dos hipótesis:
$H_0 \equiv$ La media de la distribución es 110 m.
$H_1 \equiv$ La media de la distribución no es 110 m.

Para decidir el estadístico de contraste suponemos, momentáneamente, que la hipótesis nula es cierta (la media de la distribución es 110 m). Por tanto se admite que la distancia de frenado sigue una distribución $N(110, 3)$. Bajo está hipótesis y según el teorema 17.3, el estadístico

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - 110}{3/\sqrt{74}} \in N(0,1)$$

Seleccionamos este estadístico como estadístico de contraste. Ahora debemos seleccionar una región de aceptación o de no rechazo para este estadístico. Tomamos como región de aceptación la que contiene los valores más plausibles para este estadístico, que parece que deberían ser los valores más cercanos a su media, que coincide con la parte central de su distribución, el intervalo $(-a, a)$, que tomamos como región de aceptación. Los puntos de separación entre las regiones de aceptación y de rechazo, $-a$ y a, se suelen llamar puntos críticos. Si imponemos que el error tipo I o nivel de significación del test sea α (probabilidad de rechazar la hipótesism nula siendo cierta), entonces la probabilidad de aceptar la hipótesis nula sería

$$P(-a < Z < a) = 1 - \alpha \qquad \Longrightarrow a = z_{1-\frac{\alpha}{2}}$$

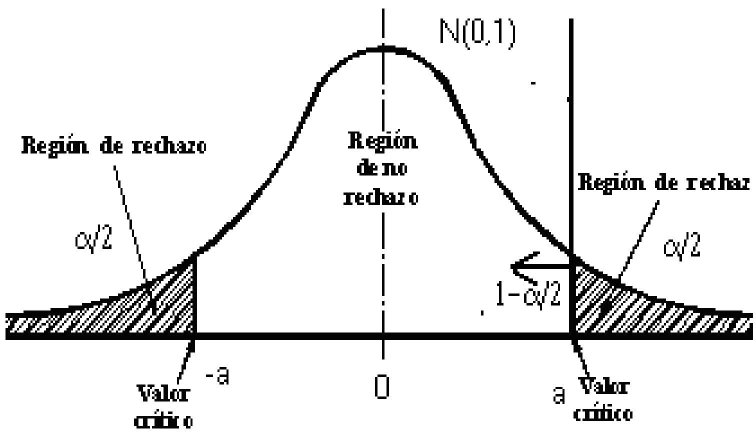


Figura 5.1: Test de dos colas.

Si tomamos para α el valor estándar 0.05, obtenemos a de la condición

$$P(Z < a) = 1 - \alpha/2 = 1 - 0.05/2 = 0.975$$

El valor de a correspondiente lo obtenemos de la tabla de la N(0,1), que resulta 1.96, lo que nos da una región de aceptación para Z

$$-1.96 < \frac{\overline{X} - 110}{3/\sqrt{74}} < 1.96$$

La prueba de hipótesis a realizar es la siguiente: Se calcula la media de las distancias recorridas por los 74 automóviles y luego el valor de $Z = \frac{\overline{X}-110}{3/\sqrt{74}}$. Si el valor resultante queda dentro del intervalo de aceptación se aceptaría la hipótesis H_0 (la media de la distribución es 110 m.) con un nivel de significación $\alpha = 0.05$. Si quedara fuera de este intervalo se rechazaría esta hipótesis, ya que la muestra no apoyaría suficientemente la hipótesis nula.

Supongamos ahora que hemos realizado efectivamente la prueba a los 74 automóviles y hemos obtenido las siguientes distancias de frenado.

Distancias	98	102	105	113	123	126	
Num. de autos	15	10	12	8	16	13	Total 74

¿Se acepta la hipótesis de que la distancia media de frenado es de 110 m, con un nivel de significación $\alpha = 0.05$?

En este caso el valor de $\overline{X}$ de la muestra es

$$\overline{X} = \frac{15 * 98 + 10 * 102 + 12 * 105 + 8 * 113 + 16 * 123 + 13 * 126}{15 + 10 + 12 + 8 + 16 + 13} = 111.62$$

así que

$$Z = \frac{\overline{X} - 110}{3/\sqrt{74}} = \frac{111.62 - 110}{3/\sqrt{74}} = 4.65$$

Como este valor no entra dentro de la región de aceptación $(-1.96, 1.96)$ nos decidimos por la hipótesis alternativa H_1. Concluimos que la media de frenado no es 110 m.

Este test de hipótesis lleva asociado un intervalo de confianza para la media. El intervalo de confianza para la media al 95%, sería

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2})$$

$$(111.62 - \frac{3}{\sqrt{74}} 1.96, 111.62 + \frac{3}{\sqrt{74}} 1.96) = (110.94, 112.3)$$

Dicho intervalo no contiene el valor 110 m que es el valor que estábamos probando para la media de las distancias de frenado. Concluimos que la media no es 110m. Esta comprobación por medio de un intervalo de confianza es totalmente equivalente a la realizada con el contraste de hipótesis.

En resumen, una prueba de hipótesis consta de los siguientes pasos:

1. Concretar la hipótesis nula H_0.

2. Formular una hipótesis alternativa H_1.

3. Decidir el estadístico de contraste (con función de distribución conocida si se verifica H_0).
4. Fijar el nivel de significación deseado α. Usar este valor para construir las regiones de aceptación y rechazo del estadístico de contraste.
5. Calcular el valor del estadístico a partir de la muestra.
6. Si el valor del estadístico pertenece a la región crítica o de rechazo, entonces rechazar H_0. En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar H_0.

Los contrastes de hipótesis, atendiendo al tipo de hipótesis alternativa, pueden ser bilaterales, como el correspondiente a la figura 5.1 de la página 173, o unilateral como el que corresponde a la figura 5.2 de la página 177.

5.10 Prueba de hipótesis para la media

5.10.1 Poblaciones normales

Para las pruebas de hipótesis que se describen en este epígrafe, se supone que las muestras proceden de poblaciones normalmente distribuidas.

Test de dos colas con varianza conocida

1. $H_0 \equiv$ la media de la distribución es $\mu = \mu_0$
2. $H_1 \equiv$ la media de la distribución , $\mu \neq \mu_0$
3. Usamos como estadístico de contraste
   $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$, que sigue una distribución $N(0,1)$ (n es el tamaño de la muestra, $\overline{X}$ es la media de la muestra)
4. Si el nivel de significación es α. La región de aceptación es
$$-z_{1-\alpha/2} < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2} \tag{5.5}$$
5. Calcular $\overline{X}$ a partir de la muestra y evaluar $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$.
6. Si el valor de $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ no cumple la relación 5.5, rechazar $\mu = \mu_0$, y por tanto aceptar $\mu \neq \mu_0$. En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar que la media sea μ_0.

Test de dos colas con varianza desconocida

1. $H_0 \equiv$ la media de la distribución es $\mu = \mu_0$
2. $H_1 \equiv$ la media de la distribución , $\mu \neq \mu_0$
3. Usamos como estadístico de contraste
   $T_{n-1} = \dfrac{\overline{X} - \mu_0}{S_c/\sqrt{n}}$, que sigue una distribución t de Student con $n-1$ grados de libertad (n es el tamaño de la muestra, $\overline{X}$ es la media de la muestra y S_c es la cuasidesviación de la muestra)
4. Si el nivel de significación es α. La región de aceptación es
   $$-t_{n-1,\,1-\alpha/2} < \frac{\overline{X} - \mu_0}{S_c/\sqrt{n}} < t_{n-1,\,1-\alpha/2} \tag{5.6}$$
5. Calcular $\overline{X}$ y S_c a partir de la muestra y evaluar $\frac{\overline{X}-\mu_0}{S_c/\sqrt{n}}$.
6. Si el valor de $\frac{\overline{X}-\mu_0}{S_c/\sqrt{n}}$ no cumple la relación 5.6, rechazar $\mu = \mu_0$, y por tanto aceptar $\mu \neq \mu_0$. En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar que la media sea μ_0.

Test de hipótesis con una cola

Cola de la izquierda con varianza conocida Este test compara las dos hipótesis siguientes:

1. $H_0 \equiv$ la media de la distribución es $\mu = \mu_0$
2. $H_1 \equiv$ la media de la distribución es $\mu < \mu_0$
3. Usamos como estadístico de contraste
   $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$, que sigue una distribución $N(0,1)$ (n es el tamaño de la muestra, $\overline{X}$ es la media de la muestra)
4. Si el nivel de significación es α. La región de aceptación es
   $$-z_{1-\alpha} < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \tag{5.7}$$
5. Calcular $\overline{X}$ a partir de la muestra y evaluar $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$.

6. Si el valor de $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ no cumple la relación 5.7, rechazar $\mu = \mu_0$, y por tanto aceptar $\mu < \mu_0$. En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar que la media sea μ_0.

Ejemplo 35 *Supongamos que en el ejemplo de los automóviles enunciado en la página 172 queremos contrastar, al 95% de confianza, las dos hipótesis siguientes: a) $H_0 \equiv$ la distancia media de frenado es 112.5 metros. b)$H_1 \equiv$ la distancia media de frenado es menor que 112.5 metros.*

Calculamos en primer lugar el valor del estadístico:

$$Z = \frac{\overline{X} - 112.5}{3/\sqrt{74}} = \frac{111.62 - 112.5}{3/\sqrt{74}} = -\,2.\,523\,3$$

La región de aceptación es:

$$-z_{1-\alpha} < Z \tag{5.8}$$

Ahora $-z_{1-\alpha} = -z_{0.95} = -1.\,644\,9$. Como el valor de Z no pertenece a la región de aceptación, porque pertenece a la cola izquierda, se rechaza la hipótesis nula con una confianza del 95% y se acepta la alternativa: La distancia media de frenado es menor que 112.5 metros.

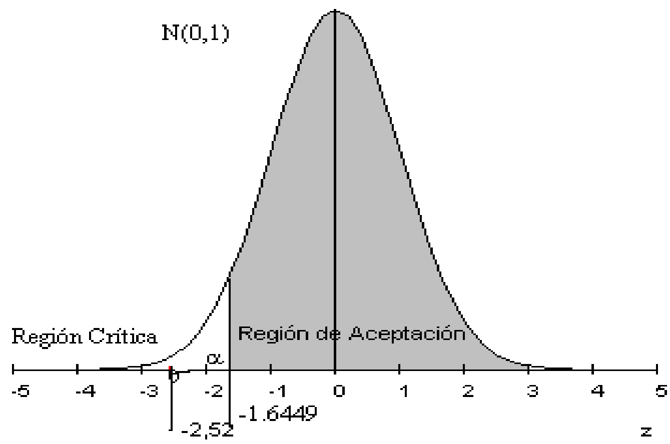


Figura 5.2: Test de una cola.

Test de la cola izquierda con varianza desconocida

1. $H_0 \equiv$ la media de la distribución es $\mu = \mu_0$

2. $H_1 \equiv$ la media de la distribución es $\mu < \mu_0$
3. Usamos como estadístico de contraste
   $T_{n-1} = \dfrac{\overline{X} - \mu_0}{S_c/\sqrt{n}}$, que sigue una distribución t de Student con $n-1$ grados de libertad (n es el tamaño de la muestra, $\overline{X}$ es la media de la muestra y S_c es la cuasidesviación de la muestra)
4. Si el nivel de significación es α. La región de aceptación es
$$-T_{n-1,\,1-\alpha} < \frac{\overline{X} - \mu_0}{S_c/\sqrt{n}} \tag{5.9}$$
5. Calcular $\overline{X}$ y S_c a partir de la muestra y evaluar $\frac{\overline{X}-\mu_0}{S_c/\sqrt{n}}$.
6. Si el valor de $\frac{\overline{X}-\mu_0}{S_c/\sqrt{n}}$ no cumple la relación 5.9, rechazar $\mu = \mu_0$, y por tanto aceptar $\mu < \mu_0$. En caso contrario, lo que se puede afirmar es que no hay suficiente evidencia para rechazar que la media sea μ_0.

Test de la cola derecha

Sirve para comparar la hipótesis $\mu = \mu_0$ contra la alternativa $\mu > \mu_0$. Se realizan con los mismos estadísticos que los de la cola izquierda con las siguientes modificaciones en el punto 4:

a) Si la varianza es conocida la región de aceptación de la hipótesis nula, $\mu = \mu_0$, es:

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha}$$

b) Si la varianza fuera desconocida la región de aceptación sería:

$$\frac{\overline{X} - \mu_0}{S_c/\sqrt{n}} < t_{n-1,\,1-\alpha}$$

5.11 Distribuciónes no normales

¿Como se reealizan las pruebas de hipótesis si las poblaciones no son normales?

Los procedimientos anteriores se pueden aplicar para contrastar las hipótesis sobre las medias de variables aleatorias que se rijan por una función de distribución arbitraria siempre que las muestras utilizadas para realizar los contrastes sean suficientemente grandes.

a) Cuando se conoce σ basta con $n > 30$.

b) Si no se conoce σ se debe cumplir $n > 100$.

5.12 Prueba de hipótesis para una proporción

Se supone una población cuyos elementos son susceptibles de tomar dos caracteres cualitativos. Considérese, por ejemplo, en una ciudad los habitantes adultos que están desempleados y los que no lo están. Se desea contrastar la hipótesis H_0 de que la proporción de elementos con una de esas características es $p = p_0$ contra la hipótesis alternativa de que $p \neq p_0$. Para ello se usa una muestra de n elementos y se cuenta el número de ellos x que posea la citada característica. Así, si se trata el ejemplo de los habitantes de la citada ciudad se seleccionaría una muestra de n personas adultas y se contaría el número de entre ellas, x, que estén desempleadas. Tomamos este valor como el estadístico de contraste. Si se cumple la hipótesis nula, la variable aleatoria x se distribuye como una binomial $B(n, p_0)$. Según el teorema central del límite esta distribución puede aproximarse con una distribución $N(np_0, \sqrt{np_0q_0})$ y la aproximación es razonable si $np_0,\ nq_0 = n(1-p_0) > 5$ y $p_0,\ q_0 = 1 - p_0 > 0.05$.

$$P(np_0 - z_{1-\frac{\alpha}{2}}\sqrt{np_0q_0} < x < np_0 - z_{1-\frac{\alpha}{2}}\sqrt{np_0q_0}) = 1 - \alpha$$

Por tanto el intervalo de aceptación para x es

$$(np_0 - z\sqrt{np_0q_0},\ np_0 - z_{1-\frac{\alpha}{2}}\sqrt{np_0q_0}) \tag{5.10}$$

Luego la prueba estadística que corresponde es la siguiente: Se cuenta el número de elementos, x, que posee la característica en cuestión. Si este valor quedara fuera del intervalo de aceptación dado en 5.10 se rechazaría la hipótesis nula y se concluiría que $p \neq p_0$. Si queda dentro del intervalo de aceptación se concluiría que no hay bastante evidencia para rechazar la hipótesis nula.

Otra forma de realizar este test es usar la proporción muestral $\frac{x}{n}$ como estadístico de contraste. Sin más que dividir por n las relaciones anteriores concluimos que el intervalo de aceptación para la proporción muestral $\frac{x}{n}$ es

$$\left(p_0 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_0q_0}{n}},\ p_0 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_0q_0}{n}}\right)$$

Ejemplo 36 *Constrastar la hipótesis de que la proporción de personas de una población con RH negativo es del 15% contra la hipótesis alternativa de que no es el 15%, con una confianza del 95%, sabiendo que se ha analizado la sangre de 400 personas elegidas aleatoriamente de esta población, obteniéndose que 72 de ellas tenía el RH negativo.*

Considerando como estadístico de contraste $x = \mathrm{n}^o$ de personas con Rh negativo, que si se cumple la hipótesis nula se distribuye aproximadamente como una

$$N(400 \times 0.15, \sqrt{400 \times 0.15 \times 0.85})$$

La aproximación es razonable, ya que:

$$np_0 = 400 \times 0.15 = 60 > 5 \;\text{ y }\; nq_0 = 400 \times 0.85 = 340 > 5$$
$$p_0 = 0.15 > 0.05 \quad y \;\; q_0 = 0.85 > 0.05$$

El intervalo de aceptación de la hipótesis nula, viene dado por la expresión 5.10, que en este caso tomaría la forma:
$\left(72 - 1.96\sqrt{400 \times 0.15 \times 0.85},\ 72 + 1.96\sqrt{400 \times 0.15 \times 0.85}\right) = (58.003, 85.997)$

Por lo tanto, con una confianza de 95%, se acepta la hipótesis nula de que el porcentaje de personas con RH negativo es de 15%, ya que el valor experimental, 72 personas, pertenece al intervalo de aceptación del test.

5.13 Prueba de hipótesis para la varianza

Frecuentemente, cuando se analizan variables cuantitativas, es importante sacar conclusiones en cuanto a la mayor o menor dispersión de la variable tratada. En este caso, lo que interesa es llegar a conclusiones acerca de la desviación estándar o la varianza de la población.

Si se intenta llegar a conclusiones en cuanto a la dispersión de la población, primero se debe determinar qué prueba estadística se puede emplear para representar la distribución de la dispersión de los datos de la muestra: Si la variable tiene distribución normal, según el teorema 17.3, el estadístico $\frac{(n-1)S_c^2}{\sigma^2}$ sigue una distribución chi-cuadrado con $(n-1)$ grados de libertad.

Suponiendo que la población de partida es normal de varianza desconocida σ^2, se formulan las siguientes hipótesis:

$H_0 : \sigma^2 = \sigma_0^2$.

$H_1 : \sigma^2 \neq \sigma_0^2$.

A partir de una muestra de n elementos se halla una estimación del estadístico $\frac{(n-1)S_c^2}{\sigma_0^2}$ Podemos tomar como intervalo de aceptación de la hipótesis nula, relativa a este estadístico, el intervalo de confianza dado en la expresión 5.4 de la página 168. Es decir, no rechazamos que $\sigma^2 = \sigma$, para un nivel de significación α, si se cumple que

$$\frac{(n-1)\,S_c^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma_0^2 < \frac{(n-1)\,S_c^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \tag{5.11}$$

Las pruebas referidas a varianzas son generalmente de cola superior, es decir que se contrasta la hipótesis nula, $\sigma^2 = \sigma_0^2$, contra la alternativa $\sigma^2 > \sigma_0^2$, ya que normalmente la preocupación está en el hecho de que la varianza pueda ser demasiado grande. En este caso se tomará como intervalo de aceptación al nivel de significación α :

$$\frac{(n-1)\, S_c^2}{\chi^2_{n-1,1-\alpha}} < \sigma_0^2$$

5.14 Prueba de bondad de ajuste

Los contrastes de hipótesis que hemos tratado hasta el momento son contrastes paramétricos: Se supone que la muestra pertenece a una población que se distribuye con arreglo a una cierta distribución conocida y se trata de estimar algunos parámetros de ésta. En esta ocasión, estudiaremos un test no paramétrico: No se supone que los datos de la muestra procedan de ninguna distribución concreta. Al contrario, nos preguntamos cuál podrá ser la distribución de la población de donde se han extraído éstos.

Una forma de decidirse por un modelo concreto es aplicar el *test chi-cuadrado de bondad de ajuste a distribuciones*. Este test realiza una comparación entre la distribución muestral y el modelo de distribución teórica a la que queremos probar si se ajustan los datos de la muestra. Puede aplicarse de la forma siguiente: Si se dispone de una muestra de n elementos (al menos 20, aunque este número puede variar según la precisión deseada), se agrupan en k clases, como en los histogramas de frecuencias. Si n_i es la frecuencia observada en la clase i y p_i la probabilidad que correspondería a este intervalo en la distribución teórica que deseamos contrastar, np_i sería el número de elementos que teóricamente debería caer en esta clase.

Para aceptar la validez del test suele exigirse que el valor asignado a cada clase teórica (np_i) sea al menos 5.

Si se cumplen estas restricciones, y se cumple la hipótesis nula de que los datos proceden de la distribución teórica que estemos probando, el estadístico

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \tag{5.12}$$

sigue (aproximadamente) una distribución χ^2 con $k-p-1$ grados de libertad, donde p es el número de parámetros de la distribución propuesta que se han estimado a partir de la muestra. El valor de χ^2 dado por la expresión 10.24 no puede ser negativo. Será nulo si hay un acuerdo perfecto entre los valores experimentales y teóricos. Es obvio que los valores experimentales se aproximan mejor a los teóricos cuanto más cerca de cero esté el valor χ^2, esto es, cuanto menor sea la discrepancia entre los valores empíricos y los teóricos.

Rechazamos la distribución propuesta, al nivel de significación α, si el estadístico es demasiado grande: $\chi^2 > \chi^2_{k-p-1,1-\alpha}$, siendo este último valor el de chi-cuadrado con $n-p-1$ grados de libertad que deja a la derecha un área α $\left(P(\chi^2 < \chi^2_{k-p-1,1-\alpha}) = 1-\alpha\right)$. La gráfica de la fígura 5.3, muestra un esquema de la regiones de aceptación y de rechazo de este tipo de tests.

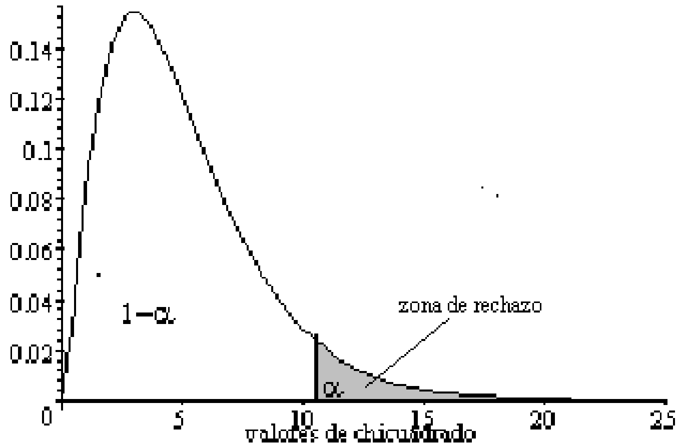


Figura 5.3: Test Chi-Cuadrado.

Si el número de clases, k, es menor o igual que 4, es preciso realizar en el estadístico χ^2 la corrección siguiente:

$$\chi^2 = \sum_{i=1}^{k} \frac{(|n_i - np_i| - 0.5)^2}{np_i}$$

Si el número de elementos no es suficiente para realizar este test puede emplearse el test de bondad de ajuste de *Kolmogorov-Smirnov* que, como el test chi-cuadrado, puede encontrarse implementado en la mayoría de los paquetes estadísticos.

Ejemplo 37 *Se desea determinar si el número de errores cometen los secretarios de una cierta oficina en las hojas mecanografiadas sigue una distribución de Poisson con un número medio de 3 errores por hoja. Para ello se han contado los errores de 440 hojas seleccionadas al azar. Los valores observados se especifican en la siguiente tabla.*

Errores por hoja	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*
Frec.observada	*18*	*53*	*103*	*107*	*82*	*46*	*18*	*10*	*2*	*1*	*440*

A continuación mostramos la tabla que compara las frecuencias observadas con las frecuencia esperadas o teóricas si se supone que la distribución fuese realmente una Poisson de parámetro 3.

N. de errores	n_i Frecuencias observadas	$p_i = \frac{3^i e^{-3}}{i!}$	$440 \times p_i$ Frecuencias esperadas
0	18	4.9787×10^{-2}	21.906
1	53	0.14936	65.719
2	103	0.22404	98.578
3	107	0.22404	98.578
4	82	0.16803	73.934
5	46	0.10082	44.36
6	18	5.0409×10^{-2}	22.18
7	10	2.1604×10^{-2}	9.5058
8	2	8.1015×10^{-3}	3.5647
9	1	2.7005×10^{-3}	1.1882

En la tabla de np_i las dos últimas casillas tienen valores menores que 5. Sumamos entonces sus valores a la antepenúltima, para que no haya ninguna con valores menores que 5, teniendo ahora las dos últimas columnas el aspecto siguiente: Ahora sólo hay 8 clases, es decir $k = 8$.

N. de errores	n_i Frecuencias observadas	$440 \times p_i$ Frecuencias esperadas
0	18	21.906
1	53	65.719
2	103	98.578
3	107	98.578
4	82	73.934
5	46	44.36
6	18	22.18
7 o más	13	14.2587

$$\chi^2 = \left[\frac{3.906^2}{21.906} + \frac{12.719^2}{65.719} + \frac{4.422^2}{98.578} + \frac{8.422}{98.578} + \frac{8.066^2}{73.934} + \frac{1.64^2}{44.36} + \frac{4.18^2}{22.18} + \frac{1.2587^2}{14.2587}\right]$$

$=$

$= 5.2813$

Como $\chi^2 = 5.2813$ es menor que $\chi^2_{k-p-1,1-\alpha} = \chi^2_{8-0-1,\ 0.95} = \chi^2_{7,\ 0.95} = 14.0671$, no se puede rechazar la hipótesis de que la distribución sea de Poisson al nivel de significación 0.05. Es decir que hay un buen acuerdo entre los datos experimentales y los teóricos que proceden de la Poisson de media 3.

Nota: Hemos tomado para p el valor 0, ya que es el único parámetro de la distribución de Poisson λ, número medio de errores por hoja. Este parámetro no se ha estimado a partir de la muestra, ya que venía dado como dato en el enunciado del ejemplo ($\lambda = 3$).

5.15 Contrastes de hipótesis para dos poblaciones

5.15.1 Test para comparar la igualdad entre las varianzas de dos poblaciones

Muestras independientes

Queremos contrastar la hipótesis nula de que $\sigma_1^2 = \sigma_2^2$

Si partimos de dos muestras independientes de tamaño n_1 y n_2 procedentes de dos distribuciones normales, el estadístico de contraste usado suele ser:

$$F = \frac{S_1^2}{S_2^2}$$

cociente entre las cuasivarianzas de estas dos muestras, que se distribuye como una F de Fisher-Snedecor con $n_1 - 1$, $n_2 - 1$ grados de libertad si se cumple la hipótesis nula de igualdad entre las varianzas poblacionales. Para contrastar, por ejemplo, la hipótesis nula $\sigma_1^2 = \sigma_2^2$, contra la alternativa $\sigma_1^2 \neq \sigma_2^2$, la región de aceptación de la hipótesis nula corresponde al intervalo

$$F_{\frac{\alpha}{2}} < \frac{S_1^2}{S_2^2} < F_{1-\frac{\alpha}{2}}$$

siendo F_p el percentil p de la distribución F de Fisher-Snedecor con $n_1 - 1$, $n_2 - 1$ grados de libertad cumpliendo

$$P(F < F_p) = p.$$

Ejemplo 38 *Para contrastar la igualdad entre las varianzas de dos poblaciones normales consideramos el test cuya hipótesis nula es que ambas varianzas son iguales. La hipótesis alternativa supone que las varianzas son distintas. Para realizar el test se parte de dos muestras, una de cada población. Los datos sobre el número de elementos, la cuasidesviación y la media de las muestras de cada población son*

	número de elementos	*media muestral*	*cuasidesviación*
Muestra 1	$n_1 = 10$	$\bar{X}_1 = 6$	$S_1 = 0.1$
Muestra 2	$n_2 = 8$	$\bar{X}_2 = 5.2$	$S_2 = 0.3$

Para realizar el test se calcula el valor muestral del estadístico de contraste $F = \frac{S_1^2}{S_1^2} = \frac{0.1^2}{0.3^2} = 0.111\,11$

Se concluirá, al nivel 0.05, que puede admitirse la igualdad de las varianzas si este valor, $F = 0.11111$, no es demasiado grande ni demasiado pequeño, es decir si está dentro de la región de aceptación del test:

$$F_{\frac{0.05}{2}} < F < F_{1-\frac{0.05}{2}} \tag{5.13}$$

donde se toma el valor de F con $9, 7$ grados de libertad.

Con un programa de ordenador, se ha obtenido que
$F_{\frac{0.05}{2}} = F_{0.025} = 0.\,238\,26$
$F_{1-\frac{0.05}{2}} = F_{0.975} = 4.\,823\,2$

Como no se cumple la relación 5.13, concluimos que no puede aceptarse la igualdad entre las varianzas.

Muestras pareadas

Las muestras pareadas son datos formados por pares de observaciones que guardan algún tipo de relación entre sí. Por ejemplo: Se desea comparar la eficacia de un plan de adelgazamiento. Con este objetivo se han pesado n personas antes de seguir el citado plan y después de un mes de haberlo seguido. En este caso no esperamos que las observaciones sobre el peso realizadas antes y después de seguir el plan de adelgazamiento no serán independientes, ya que son medidas del peso de la misma persona.

El estadístico de contraste para la hipótesis nula de igualdad de las varianzas es, para el caso de muestras pareadas:

$$T = \frac{|F-1|}{2}\sqrt{\frac{n-2}{F(1-r^2)}}$$

siendo $F = \frac{S_1^2}{S_2^2}$, r el coeficiente de correlación entre ambas variables y n el número de pares de datos. Este estadístico se distribuye como una t de Student con $n-2$ grados de libertad si se cumple la hipótesis de igualdad entre las varianzas y las poblaciones de partida son normales .

La región de aceptación al nivel α es $T < t_{1-\alpha}$ donde t_p es el valor de la t de Student con $n-2$ grados de libertad que cumple :

$$P(t < t_p) = p.$$

Ejemplo 39 *La siguiente tabla nos da el peso en Kg. de 9 personas antes y después de haber seguido una misma dieta de adelgazamiento:*

	Antes	*Después*
Persona 1	*155*	*142*
Persona 2	*147*	*139*
Persona 3	*123*	*110*
Persona 4	*107*	*100*
Persona 5	*105*	*96*
Persona 6	*93*	*87*
Persona 7	*100*	*95*
Persona 8	*123*	*110*
Persona 9	*106*	*93*

Comparar las varianzas de las poblaciones de procedencia de estas muestras (peso antes y después de la dieta).

$$F = \frac{S_1^2}{S_2^2} = \frac{455.25}{396} = 1.1496$$
$$r = 0.9829$$
$$n = 9$$
$$T = \frac{|F-1|}{2}\sqrt{\frac{n-2}{F(1-r^2)}} = \frac{|1.1496-1|}{2}\sqrt{\frac{9-2}{1.1496(1-0.9829^2)}} = 1.0024$$

Este estadístico se distribuye como una t de Student con 7 grados de libertad si se cumple la hipótesis de igualdad entre las varianzas y las poblaciones son normales .

La región de aceptación al nivel 0.05 es $(0,\ t_{7,\ 0.95}) = (0,\ 1.8946)$ que contiene el valor obtenido de la muestra $T = 1.0024$, por lo que no podemos rechazar la igualdad entre las varianzas.

El resultado del test no depende del orden de las muestras. Si tomamos el orden contrario obtenemos

$$F = \frac{S_2^2}{S_1^2} = \frac{396}{455.25} = 0.86985$$

y el estadístico T toma exactamente el mismo valor:

$$T = \frac{|0.86987-1|}{2}\sqrt{\frac{9-2}{0.86987(1-0.9829^2)}} = 1.0024$$

5.15.2 Comparación entre las medias de dos poblaciones normales

Muestras independientes (varianzas conocidas)

Para comparan las medias de dos poblaciones, de medias y desviaciones típicas μ_1, σ_1 y μ_2, σ_2 respectivamente, conociendo los valores de σ_1 y σ_2, por medio de sendas muestras independientes de tamaños n_1, n_2 y con medias muestrales $\overline{X}_1$ y $\overline{X}_2$, se contrastan las hipótesis $\mathrm{H}_0 \equiv \mu_1 - \mu_2 = d$ y $\mathrm{H}_1 \equiv \mu_1 - \mu_2 \neq d$ usando el estadístico de contraste

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

El estadístico Z se distribuye con arreglo a una distribución normal estándar. La región de aceptación de la hipótesis nula, $\mu_1 - \mu_2 = d$, con una confianza de $100(1-\alpha)\%$ es

$$-z_{1-\alpha/2} < \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{1-\alpha/2}$$

donde el valor de $z_{1-\alpha/2}$ se calcula en la tabla de la distribución N(0,1). Es el valor que cumple:

$$P(Z < z_{1-\alpha/2}) = 1 - \alpha/2$$

Muestras independientes (varianzas desconocidas e iguales).

En este caso se comparan las medias de dos poblaciones, de medias y desviaciones típicas μ_1, σ y μ_2, σ respectivamente, por medio de sendas muestras independientes de tamaños n_1, n_2 y con medias y cuasidesviaciones muestrales $\overline{X}_1$, S_1, y $\overline{X}_2$, S_2. Considerando varianzas poblacionales desconocidas, pero iguales, se contrastan las hipótesis $\mathrm{H}_0 \equiv \mu_1 - \mu_2 = d$ y $\mathrm{H}_1 \equiv \mu_1 - \mu_2 \neq d$ usando el estadístico de contraste

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ siendo } S^2 = \frac{S_1^2(n_1 - 1) + S_{21}^2(n_2 - 1)}{n_1 + n_2 - 2}$$

El estadístico T se distribuye en esta ocasión como una t de Student de $n_1 + n_2 - 2$ grados de libertad. La región de aceptación de la hipótesis nula, $\mu_1 - \mu_2 = d$, con una confianza de $100(1-\alpha)\%$ es

$$-t_{1-\alpha/2} < \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{1-\alpha/2}$$

donde el valor de $T_{1-\alpha/2}$ es el de la t de Student con $n_1 + n_2 - 2$ grados de libertad que cumple:

$$P(t < t_{1-\alpha/2}) = 1 - \alpha/2$$

Muestras independientes (varianzas desconocidas y desiguales)

Si las varianzas no pueden suponerse iguales se usa el estadístico

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

que se supone distribuida como una t de Student con los grados de libertad dados por el número entero más próximo al resultado que se obtenga en la expresión:

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1-1}\right)^2 + \left(\frac{S_2^2}{n_2-1}\right)^2} \tag{5.14}$$

Ejemplo 40 *Disponemos de dos muestras procedentes de poblaciones normales, independientes y con los siguientes datos:*

	número de elementos	*media muestral*	*cuasidesviación*
Muestra 1	$n_1 = 10$	$\bar{X}_1 = 6$	$S_1 = 0.1$
Muestra 2	$n_2 = 8$	$\bar{X}_2 = 5.2$	$S_2 = 0.3$

Se desea emplear estos datos para contrastar si son iguales las medias de las poblaciones de partida.

En estos contrastes debe estudiarse en primer lugar si puede admitirse la igualdad de las varianzas. Para ello puede usarse el test dado en la sección 5.15.1. Este ejemplo ya se ha analizado en dicha sección habiendo concluido que, al nivel $\alpha = 0.05$, las varianzas se consideraban diferentes.

Constratemos ahora la igualdad entre las medias de ambas distribuciones, al 95% de confianza. Para ello se calcula el valor del estadístico de contraste:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(5.2-6)-0}{\sqrt{\frac{0.01}{10} + \frac{0.09}{8}}} = -7.228\,1$$

que se compara con la t de Student cuyos grados de libertad se calculan con la fórmula dada por la expresión 5.14.

$$\frac{\left(\frac{0.01}{10} + \frac{0.09}{8}\right)^2}{\left(\frac{0.01}{9}\right)^2 + \left(\frac{0.09}{7}\right)^2} = \frac{1.500\,6 \times 10^{-4}}{1.665\,4 \times 10^{-4}} = 0.901\,06 \approx 1$$

Como $t_{1,1-0.05/2} = t_{1,0.975} = 12.706$ y $-12.706 < -7.228\,1 < 12.706$, no se puede rechazar la igualdad entre las medias al 95% de confianza.

Muestras pareadas

Como ya hemos aclarado en el apartado 5.15.1 (página 185), las muestras pareadas son datos formados por pares de observaciones que guardan algún tipo de relación entre sí. En aquel caso se probaba la igualdad entre las varianzas. En este apartado contrastaremos los valores medios de ambas muestras. Por ejemplo: Se desea comparar la eficacia de un plan de adelgazamiento. Con este objetivo se han pesado n personas antes de seguir el citado plan y después de un mes de haberlo seguido. Para cada una de estas personas calculamos la diferencia, $d_i =$ peso de la persona i antes de la dieta $-$ peso de la persona i después de la dieta. Si suponemos que esta variable sigue una distribución normal, podemos contrastar la hipótesis de que la diferencia de los promedios de los pesos antes y después es $\mu_1 - \mu_2 = d$ contra la alternativa de que esta diferencia es distinta de d.

El estadístico de contraste es el correspondiente a contrastar la media de la variable d_i. Suponiendo varianza desconocida y usando el criterio dado en la expresión 5.6 de la página 176. obtenemos la región de aceptación para la hipótesis $\mu_1 - \mu_2 = d$

$$-t_{n-1,\,1-\alpha/2} < \frac{\overline{d_i} - d}{S_c/\sqrt{n}} < t_{n-1,\,1-\alpha/2} \tag{5.15}$$

Siendo $\overline{d_i}$ la media de las diferencias de pesos de las n personas y S_c la cuasidesviación típica de las diferencias de pesos d_i.

Ejemplo 41 *Deseamos comprobar si la eficacia de la dieta de adelgazamiento (ejemplo de la página 185) se refleja en una pérdida de peso de 7 Kg por término medio, o si por el contrario, los datos parecen apoyar la alternativa de que la pérdida media de peso ha sido mayor que 7 Kg.*

Las hipótesis a contrastar son: $\mathrm{H}_0 : d = 7$, $\mathrm{H}_1 : d > 7$

En este caso usamos el test de la cola derecha especificado en la sección 5.10.1. Si el nivel de significación es 0.05, la región de aceptación de la hipótesis nula es ahora

$$\frac{\overline{d_i} - d}{S_c/\sqrt{n}} < t_{9-1,\,1-0.05} = 1.8595$$

En el ejemplo de la página 185, las diferencias de peso son los valores indicados en la columna d_i. Se supone que esta variable se distribuye normalmente.

	Antes	Después	d_i
Persona 1	155	142	13
Persona 2	147	139	8
Persona 3	123	110	13
Persona 4	107	100	7
Persona 5	105	96	9
Persona 6	93	87	6
Persona 7	100	95	5
Persona 8	123	110	13
Persona 9	106	93	13

El valor del estadístico de contraste para la muestra cumple:

$$\frac{\overline{d_i} - d}{S_c/\sqrt{n}} = \frac{9.66667 - 7}{3.3541/\sqrt{9}} = 2.3851$$

Este valor está en la zona de rechazo. Por tanto, entre las dos hipótesis se decide aceptar la hipótesis alternativa y se concluye que la dieta de adelgazamiento disminuye por término medio más de 7 Kg. de peso al cabo de un mes.

5.15.3 Test para la diferencia entre dos proporciones

Un ejemplo de este tipo de test se daría si quisiéramos comparar la proporción de parados en Andalucía con la proporción de parados en la Comunidad Europea. Para ello se seleccionarían aleatoriamente una muestra de personas en Andalucía y otra en Europa.

Sean $\widehat{p_1}$ la proporción de parados detectados en la muestra, de n_1 personas seleccionadas en Andalucía y $\widehat{p_2}$ la proporción de parados detectados en la muestra, de n_2 personas seleccionadas en toda la Comunidad Europea.

Para contrastar la hipótesis de que la proporciones verdaderas, p_1 y p_2 difieren en $\mu = p_1 - p_2$ usamos el estadístico

$$Z = \frac{(\widehat{p_1} - \widehat{p_2}) - \mu}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

que supondremos distribuida como una $N(0,1)$ si el tamaño de la muestra es amplio.

5.15.4 Test de independencia de dos variables cualitativas

El test chi-cuadrado de bondad de ajuste, ya tratado en la sección 5.14, puede aplicarse para contrastar la hipótesis de que dos variables de clasificación de una población sean independientes contra la hipótesis alternativa de que no lo sean. En este caso no se suponen conocidas las distribuciones de laa muestras, ni se trata de determinar ningun parámetro. Se trata de un test no paramétrico.

Los datos de la muestra en que se va a basar el test se presentan en una tabla de contingencia.

Ejemplo 42 *Se desea estudiar si el nivel de estudios realizados por las personas guarda alguna relación con su preferencia por pasar sus vacaciones en el campo o en la playa. Para ello se han entrevistado a 500 personas y se han clasificado con arreglo a estos criterios en la forma indicada en la siguiente tabla de contingencia*

OBSERVACIONES	*Primarios*	*Medios*	*Universitarios*
C =Campo	*33*	*30*	*145*
Pl = Playa	*60*	*25*	*355*
	93	*55*	*Total=500*

Realizamos un test de hipótesis tipo chi-cuadrado similar al descrito en el apartado 5.14. Consideraremos la hipótesis nula consistente en que ambas clasificaciones (estudios y preferencia para las vacaciones) no guardan relación alguna, es decir que los sucesos correspondientes a una y otra son independientes: Por ejemplo los sucesos consistentes en "haber realizado estudios primarios (Pr)" y "preferir el campo (C)" son independientes. La hipótesis alternativa es que ambas clasificaciones guardan algún tipo de relación o que son dependientes.

Nombramos cada suceso con sus iniciales, estimando la probabilidad de los sucesos por su frecuencia relativa y teniendo en cuenta su independencia (hipótesis nula) tendremos que, por ejemplo, para el suceso correspondiente a la primera casilla:

$$P(\mathrm{Pr} \cap C) = P(\mathrm{Pr}) \times P(C) = \frac{93}{500} \times \frac{145}{500} = 0.053\,94$$

Entonces, los valores esperados, np_i, para cada una de las 6 clases serían:

$$500 \times P(\mathrm{Pr} \cap C) = 500 \times P(\mathrm{Pr}) \times P(C) = 500 \times \frac{93}{500} \times \frac{145}{500} = 26.\,97$$
$$500 \times P(M \cap C) = 500 \times P(M) \times P(C) = 500 \times \frac{352}{500} \times \frac{145}{500} = 102.\,08$$
$$500 \times P(U \cap C) = 500 \times P(\mathrm{Pr}) \times P(C) = 500 \times \frac{55}{500} \times \frac{145}{500} = 15.\,95$$
$$500 \times P(\mathrm{Pr} \cap Pl) = 500 \times P(\mathrm{Pr}) \times P(C) = 500 \times \frac{93}{500} \times \frac{355}{500} = 66.\,03$$
$$500 \times P(M \cap Pl) = 500 \times P(\mathrm{Pr}) \times P(C) = 500 \times \frac{352}{500} \times \frac{355}{500} = 249.\,92$$
$$500 \times P(U \cap Pl) = 500 \times P(\mathrm{Pr}) \times P(C) = 500 \times \frac{55}{500} \times \frac{355}{500} = 39.\,05$$

En la siguiente tabla se resumen estos valores:

ESPERADOS	Primarios	Medios	Universitarios	
C =Campo	26.97	102.08	15.95	145
Pl = Playa	66.03	249.92	39.05	355
	93	352	55	Total=500

Procedemos ahora a calcular el valor de $\chi^2_{\exp} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$, que en este caso se compara con la χ^2, cuyos grados de libertad se calculan de la siguiente forma: Si f y c son el número de características de cada criterio de clasificación, en este caso $f = 2$ preferencias para las vacaciones y $c = 3$ tipos de niveles de estudios, entonces el número de grados de libertad se calculan como $(f-1)(c-1) = 1 \times 2 = 2$.

$\chi^2_{\exp} = \frac{(33-26.97)^2}{26.97} + \frac{(82-102.08)^2}{102.08} + \frac{(30-15.95)^2}{15.95} + \frac{(60-66.03)^2}{66.03} + \frac{(270-249.92)^2}{249.92} + \frac{(25-39.05)^2}{39.05} =$
$= 24.894$

La región crítica, o de rechazo, es

$$\chi^2_{\exp} > \chi^2_{(f-1)(c-1),\ 1-\alpha}$$

$$\chi^2_{(f-1)(c-1),\ 1-\alpha} = \chi^2_{2,\ 0.95;} = 5.9915$$

Como el valor experimental, 24.894, es mayor que el teórico, 5.9915, se rechaza la hipótesis nula, y se concluye que la muestra parece indicar que hay alguna relación entre el nivel de estudios y las preferencias sobre el lugar de vacaciones.

En el caso de las tablas con 4 clases hay que realizar una corrección en la fórmula empleada para la chi-cuadrado, ya indicada en el apartado 5.14 de la página 181, tal como se muestra en el ejemplo siguiente:

Ejemplo 43 *Supongamos que se han seleccionado 90 personas adultas y se han clasificado por dos características: sexo y hábito de fumar. Los resultados de la clasificación están resumidos en la siguiente tabla:*

OBSERVACIONES	*Hombre*	*Mujer*	
Fuma	*15*	*29*	*44*
No fuma	*27*	*19*	*46*
	42	*48*	*90*

Se desea saber si estos datos apoyan la hipótesis nula de que no hay relación entre el sexo y el hábito de fumar o más bien apoyan la hipótesis contraria de que el sexo de una persona influye de alguna forma en el hábito de fumar. Los valores teóricos obtenidos bajo la hipótesis nula de independencia son:

$$90 \times P(H \cap F) = P(H)\ P(F) = 90 \times \frac{42}{90}\frac{44}{90} = 20.533$$
$$90 \times P(M \cap F) = P(M)\ P(F) = 90 \times \frac{48}{90}\frac{44}{90} = 23.467$$
$$90 \times P(H \cap F') = P(H)\ P(F') = 90 \times \frac{42}{90}\frac{46}{90} = 21.467$$
$$90 \times P(M \cap F') = P(M)\ P(F') = 90 \times \frac{48}{90}\frac{46}{90} = 24.533$$

ESPERADOS	Hombre	Mujer	
Fuma	20.534	23.467	44
No fuma	21.467	24.533	46
	42	48	90

Usando la corrección adecuada para este caso

$$\chi^2_{\text{exp}} = \sum_{i=1}^{k} \frac{(|n_i - np_i| - 0.5)^2}{np_i} \tag{5.16}$$

$\chi^2_{\text{exp}} = \frac{(|15-20.534|-0.5)^2}{20.534} + \frac{(|29-23.467|-0.5)^2}{23.467} + \frac{(|27-21.467|-0.5)^2}{21.467} + \frac{(|19-24.533|-0.5)^2}{24.533} =$
$= 4.526\,1$

Los grados de libertad se calculan como el producto del número de niveles de una de las caracteristicas menos uno y el número de niveles de la otra característica menos uno, $(2-1)(2-1) = 1$. El valor de la chi-cuadrado con 1 grado de libertad correspondiente a una probabilidad 0.95 es $\chi^2_{1,\ 0.95} = 3.841\,5$.

En este caso la chi-cuadrado experimetal, 4.5261, supera al valor teórico, 3.8415. Se rechaza la hipótesis de independencia. Concluimos por tanto que el sexo tiene alguna relación con el hábito de fumar. Los datos parecen apoyar la idea de que las mujeres fuman más que los hombres.

5.16 EJERCICIOS PROPUESTOS

Ejercicio 62 *La media de una muestra de 36 elementos de una distribución normal es 4.1. Hallar un intervalo de confianza al 95% para la media. (La desviación típica de la población es 3).*

Ejercicio 63 *Se ha repetido un experimento físico 9 veces obteniendose una media de los valores medidos de 42.319 y una cuasi-desviación típica de 5.0 Estimar el valor real de la magnitud con una confianza del 95 por 100*

Ejercicio 64 *Para probar si una moneda es defectuosa (la cara y la cruz no tienen la misma probabilidad) se recurre al siguiente ensayo. Se tira la moneda 100 veces y se declara defectuosa si el número de caras es un número fuera del intervalo [40, 60].*

1. *Calcular la probabilidad de declarar la moneda como defectuosa una moneda correcta (error tipo I del test de hipótesis)*

2. *Calcular la probabilidad de declararla correcta si la probabilidad de sacar cara fuera: a) 0.6, b) 0.65, c) 0.70, d) 0.80.*

Ejercicio 65 *Diseñar una prueba de hipótesis* (*al* 95% *de confianza) para la longitud media de una serie de tornillos basada en muestras de 9 elementos, que permita rechazar los lotes cuya longitud media no sea 5 mm. La longitud de estos tornillos se distribuye según una normal de desviación típica* $\sigma = 2\,mm$.

Ejercicio 66 *Un vendedor de bandas elasticas afirma que resisten un estiramiento promedio de 180Kg. Se ha hecho una prueba con 5 de estas banda observandose una resistencia promedio de 169.51Kg. con una cuasi desviación de 5.7 kg*

1. *¿Se rechazaría al 99% de confianza la media de resistencia indicada por el vendedor.*

2. *¿Cual es la región de rechazo para la resistencia promedio de la muestra? ¿Y el valor Crítico?*

Ejercicio 67 *Un tipo de botes de pintura esta declarada como apta para pintar un promedio de 80* m^2 *con una desviación típica de 8.4* m^2*.Se desea comprobar si puede aceptarse este valor promedio. Con este objetivo se ha decidido probar 100 de estos botes y rechazar la pintura si el promedio de superficie pintada resultará menor que 78* m^2 *Se aceptará el valor de la desviación típica.*

1. *Calcular el nivel de confianza y la significación de esta prueba.*

2. *Si la pintura pintara ralmente un promedio de 79* m^2 *cual sería la probabilidad de no rechazar la media indicada por el fabricante.*

3. *¿Y si el promedio fuera de 75* m^2

Ejercicio 68 *Un vendedor de neumáticos dice que la vida media de sus neumáticos es de 28000 Km. Admitiendo para la desviación típica el valor 1348 Km. diseñar un test de hipótesis al 99% de confianza, basado en muestras de 40 elementos que permita contrastar la hipótesis nula de ser* $\mu = 28000Km$ *usando como hípótesis alternativa* $\mu < 28000Km$

Ejercicio 69 *Si de un total de 100 personas entrevistadas 36 han afirmado que conocen una cierta marca de detergente*

1. *Hallar un intervalo de confianza al 95% para la proporción real de personas que conocen este detergente.*

2. *¿Cuantas personas se precisan entrevistar para que el intervalo de confianza para la proporción tenga una amplitud de 0.1?*

Ejercicio 70 *Se desea saber la proporción de personas de una gran ciudad que encuentran adecuado el transporte público. ¿ Cuántas personas hay que entrevistar si se desea estimar esta proporción con un intervalo de confianza de 95% y un error de precisión menor del 6%?.*

Ejercicio 71 *Encuestadas 267 personas ha resultado que 114 de ellas encuentran satisfactorio el transporte público. Dar un intervalo de confianza para la proporción de personas que encuentran satisfactorio este tipo de transporte.(95% de confianza)*

Ejercicio 72 *32 medidas del punto de ebullución del azufre tienen una cuasidesviación de 0.83 grados. Calcular un intervalo de confianza para la varianza con una confianza del 98%*

Ejercicio 73 *Las piezas de una maquina deben ser del mismo tamaño, por eso se exige que la desviación típica de la población sea 0.05 mm. Diseñar un test al 95% de confianza para contrastar la hipótesis de que* $\sigma = 0.05$ *mm. con muestras de 15 elementos*

Ejercicio 74 *Se ha llevado a cabo un estudio para determinar si hay diferencia entre el tiempo que tardan los hombres y las mujeres en hacer determinada maniobra en una línea de ensamble. Los valores obtenidos en el estudio se resumen en la siguiente tabla*

	Nº de elementos	*media muestral*	*Varianza poblacional*
hombres	*50*	*42 seg.*	*18* seg^2
mujeres	*50*	*38 seg*	*14* seg^2

¿Es significativa la diferencia de rendimiento entre hombres y mujeres?

Ejercicio 75 *Un fabricante asegura que sus fusibles, con una sobrecarga del 20%, se fundiran por promedio al cabo de 12.40 min. Una muestra de 20 fusibles se sobrecarga un 20%, obteniendose una media de 10.63 y una cuasidesviación de 2.48 min. ¿Confirma la muestra la afirmación del fabricante para el promedio?*

Ejercicio 76 *Se han recogido muestras de aire para estudiar su contaminación, obteniendose las siguientes cantidades de impurezas en* $\frac{Kg}{m^3}$

$$2.2;\ 1.8;\ 3.1;\ 2.0;\ 2.4;\ 2.0;\ 2.1;\ 1.2$$

Dad un intervalo de confianza al 95% para la media de impurezas contenidas en el aire

Ejercicio 77 *El director de un colegio quiere saber el tiempo medio que tardan los alumnos en cambiar de clase, con una confianza del 99% y un error que no sobrepase 0.25 mininutos. Si se puede suponer que el valor de* σ *es 1.40 minutos, ¿Cuál debe ser el tamaño de la muestra?*

Ejercicio 78 *Se realizó un muestreo para decidir si los sueldos de los peones de albañil de una ciudad A y de otra B son iguales por promedio o no. Para ello se consulto a 100 peones de la ciudad A y a 150 de la ciudad B. Analizadas la respuestas realizadas por dichos operarios se determino que la media de los sueldos de los 100 operarios de la ciudad A era de 760 € y la de los 150 empleados de ciudad B era de 720 €. Suponiendo que la desviación típica poblacional de los sueldos de A es 12€ y la de B 9 €, decidir si el sueldo medio en ambas ciudades es igual o distinto.*

Ejercicio 79 *Se desea comparar el gasto medio mensual en alimentación entre las familias de dos barrios. Para ello se seleccionaron 20 familias de cada barrio, observando sus gastos mensuales en alimentación. Se determino la media y las cuasidesviaciones típicas, obteniéndose los siguientes resultados muestrales:* $\left(\overline{X_1} = 200,\ S_1 = 20 \quad , \quad \overline{X_2} = 175,\ S_2 = 17\right)$. *Suponiendo que los gastos se distribuyen normalmente decidir sobre la cuestión planteada. Los gastos medios en alimentación entre ambos barrios, ¿pueden considerarse iguales?*

Ejercicio 80 *Mendel sembró 532 plantas de guisantes usando semillas del mismo tipo y los frutos resultantes los clasificó atendiendo al color en: verde, verde amarillento y amarillo y atendiendo a la forma: redondo, levemente rugoso y rugoso. Obtuvo los siguientes datos:*

	Verde	*Verde-Amarillo*	*Amarillo*	
Redondo	*35*	*68*	*38*	*141*
Levemente Rugoso	*67*	*138*	*60*	*265*
Rugoso	*30*	*68*	*28*	*126*
	132	*274*	*126*	*532*

¿Había alguna relación de dependencia entre la forma y el color de esos guisantes?