

Apuntes de ESTADÍSTICA

Inferencia estadística



Sixto Sánchez Merino
Dpto. de Matemática Aplicada
Universidad de Málaga



Mi agradecimiento al profesor Carlos Cerezo Casermeiro y Carlos Guerrero García, por sus correcciones y sugerencias en la elaboración de estos apuntes.

Apuntes de Estadística

©2011, Sixto Sánchez Merino.




Este trabajo está editado con licencia “Creative Commons” del tipo:

Reconocimiento-No comercial-Compartir bajo la misma licencia 3.0 España.

Usted es libre de:

-  copiar, distribuir y comunicar públicamente la obra.
-  hacer obras derivadas.

Bajo las condiciones siguientes:

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

Capítulo 7

Inferencia estadística

Cuando queremos obtener información sobre una población y disponemos de los datos de todos los individuos (censo), entonces podemos utilizar la estadística descriptiva que tiene como objeto el estudio de un conjunto de elementos con alguna característica común a todos ellos.

Sin embargo, cuando no podemos tener acceso a los datos de todos los individuos, utilizaremos la inferencia estadística que tiene por objeto extraer conclusiones de la totalidad de la población, a partir de los datos de una muestra de ella.

Los dos problemas fundamentales que estudia la inferencia estadística son el “problema de la estimación” y el “problema del contraste de hipótesis”. Cuando se conoce la distribución que sigue la variable aleatoria objeto de estudio y sólo tenemos que estimar los parámetros que la determinan, estamos ante un problema de inferencia estadística paramétrica; por el contrario, cuando no se conoce la distribución que sigue la variable aleatoria objeto de estudio, estamos ante un problema de inferencia estadística no paramétrica.

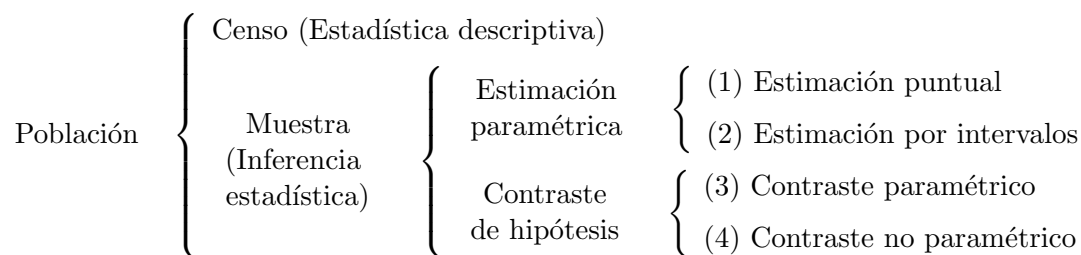
En todos los problemas que estudia la inferencia estadística juega un papel fundamental la “teoría de muestras” que estudia las técnicas y procedimientos que debemos emplear para que las muestras sean representativas de la población que pretendemos estudiar, de forma que los errores en la determinación de los parámetros de la población objeto de estudio sean mínimos.

7.1. Inferencia estadística

La Inferencia Estadística es la parte de la estadística matemática que se encarga del estudio de los métodos para la obtención del modelo de probabilidad (forma funcional y parámetros que determinan la función de distribución) que sigue una variable aleatoria de una determinada población, a través de una muestra (parte de la población) obtenida de la misma.

En la inferencia estadística se distinguen básicamente dos tipos de objetivos:

1. **Inferencia paramétrica:** Deducir características (parámetros) de la población a partir de los datos de una muestra.
2. **Contraste de Hipótesis:** Analizar la concordancia o no de los resultados muestrales con determinadas hipótesis sobre la población.



En este tema estudiaremos algunos problemas tanto de inferencia paramétrica (1, 2 y 3) como de inferencia no paramétrica (4). En inferencia estadística paramétrica nos vamos a limitar a problemas donde la variable aleatoria objeto de estudio sigue una distribución binomial, Poisson o normal, y nuestro objetivo será tratar de estimar los parámetros que la determinan, es decir, el parámetro p de la binomial, el parámetro λ de la Poisson, y los parámetros μ y σ de la normal. En los problemas de estimación no paramétrica nos limitaremos al estudio de la bondad de un ajuste, la homogeneidad de varias muestras y la independencia de caracteres, como aplicaciones de la χ^2 .

7.1.1. Teoría de muestras

En la práctica, suele ocurrir con frecuencia que no es posible estudiar todos los elementos de la población, por distintas razones:

- Si el número de elementos de la población es muy elevado, el estudio llevaría tanto tiempo que sería impracticable o económicamente inviable.
- El estudio puede implicar la destrucción del elemento objeto de estudio. Por ejemplo, estudiar la vida media de una partida de bombillas, o la tensión de rotura de cables.
- Los elementos pueden existir conceptualmente, pero no en la realidad. Por ejemplo, la proporción de piezas defectuosas que producirá una máquina.

En estas ocasiones, lo que se hace es seleccionar una muestra de la población, de manera que, de la observación del comportamiento individual de cada uno de los elementos, se puedan obtener unas leyes generales de comportamiento de tipo promedio o de tipo predominante para todos los elementos de la población.

La teoría de muestras estudia los procedimientos para tomar muestras de manera apropiada, es decir, las muestras tienen que ser representativas de la población. Y para conseguirlo, se deben cumplir dos principios básicos:

1. Independencia en la selección de los individuos que forman la muestra
2. Que todos los individuos tengan la misma probabilidad de ser incluidos en la muestra

Para conseguir estos objetivos se emplean distintas técnicas de muestreo. De los distintos métodos que existen para la obtención de muestras, destacamos tres de los más utilizados:

- *Muestreo aleatorio simple*. Se eligen al azar los elementos para garantizar que todos los individuos de la población tienen la misma oportunidad de ser incluidos en dicha muestra. Puede ser de dos tipos: con o sin reposición.
- *Muestreo estratificado*. Los elementos de la población se dividen en clases o estratos. La muestra se toma asignando un número o cuota de miembros a cada estrato (proporcional a su tamaño relativo o su variabilidad) y escogiendo los elementos por muestreo aleatorio simple dentro del estrato.
- *Muestreo sistemático*. Los elementos de la población están ordenados en listas. Se divide la población en tantas partes como el tamaño muestral y se elige al azar un número de orden. La muestra se obtiene tomando el elemento que ocupa ese número de orden en cada parte de la población.

En adelante, en los problemas de inferencia estadística consideraremos que las muestras son suficientemente representativas para inferir o estimar las características poblacionales.

Si consideramos una muestra de tamaño n representativa de la población, puesto que los n elementos que integran la muestra son elegidos aleatoriamente, es evidente que sus medidas o características son, a su vez, variables aleatorias, ya que dependen de los valores aleatorios de los valores muestrales tomados al azar.

Por tanto, una *muestra* es un vector aleatorio $(X_1, X_2, \dots, X_n) \in E^n$, que tendrá asociada una probabilidad de ser elegido.

Llamaremos *estadístico* a una función $F : E^n \rightarrow \mathbb{R}$, es decir, una “fórmula” de las variables que transforma los valores tomados de la muestra en un número real. Además, a la distribución de F se le llama distribución del estadístico en el muestreo. Por ejemplo, la función

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

es un estadístico que permitirá obtener la media de los valores muestrales, cuando dispongamos de los datos de la muestra.

7.2. Estimación paramétrica

Cuando se realiza una afirmación acerca de los parámetros de la población en estudio, basándose en la información contenida en la muestra se dice que realizamos una estimación puntual pero si señalamos un intervalo de valores dentro del cual se tiene confianza de que esté el valor del parámetro decimos que estamos realizando una estimación por intervalos.

7.2.1. Estimación puntual

El proceso de estimación puntual utiliza un estadístico, que llamaremos *estimador puntual*, para obtener algún parámetro de la población. Como estadístico que es, el estimador puntual es una variable aleatoria que tiene una distribución en el muestreo que depende, en general, del parámetro en cuestión.

Se utilizan dos criterios esenciales para medir la bondad del estimador:

- a) Que sea *centrado* o *insesgado*, es decir, que su media coincida con el parámetro a estimar.
- b) Que sea de *mínima varianza* o que tenga la menor varianza entre todos los estimadores del parámetro.

Si verifica las dos condiciones diremos que el estimador es *eficiente*. A continuación, relacionamos los estadísticos eficientes más usuales, así como su distribución de probabilidad que nos permitirá obtener los intervalos de confianza. Para todos ellos, consideraremos que la muestra de tamaño n es $\{x_1, x_2, \dots, x_n\}$.

La proporción muestral en una distribución binomial

La proporción muestral del suceso E

$$\hat{p} = \frac{\text{frecuencia absoluta del suceso } E}{n}$$

estima la proporción p de la población que presenta una determinada característica E (éxito) frente a los que no la presentan F (fracaso). Las propiedades más importantes son:

1. El estimador es insesgado, es decir, la distribución en el muestreo de \hat{p} tiene de media p .
2. El estimador es de varianza mínima igual a $\frac{p \cdot q}{n}$ con $q = 1 - p$.
3. Para valores grandes del tamaño de la muestra (en la práctica $n > 30$), la proporción muestral \hat{p} se distribuye según una distribución normal:

$$\text{Si } n > 30 \quad \text{entonces} \quad \hat{p} \rightsquigarrow N\left(p, \sqrt{\frac{pq}{n}}\right) \iff \frac{\hat{p} - p}{\sqrt{pq/n}} \rightsquigarrow N(0, 1)$$

La media muestral en una distribución de Poisson

La media muestral

$$\hat{\lambda} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

es un estimador puntual del parámetro λ de una población cuya característica estudiada sigue una distribución de Poisson de parámetro λ (= media de la población). Las propiedades más importantes son:

1. El estimador es insesgado, es decir, la distribución en el muestreo de $\hat{\lambda}$ tiene de media λ .
2. El estimador es de varianza mínima.
3. Si el tamaño de la muestra es suficientemente grande, el estimador $\hat{\lambda}$ se distribuye según una distribución normal:

$$\hat{\lambda} \rightsquigarrow N\left(\lambda, \sqrt{\frac{\lambda}{n}}\right)$$

La Cuasivarianza muestral en una distribución normal

La cuasivarianza o varianza muestral

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 f_i = \frac{n}{n-1} \cdot \left(\sum_{i=1}^n x_i^2 f_i - \bar{x}^2 \right)$$

es un estimador de la varianza σ^2 de una población cuya característica en estudio sigue una distribución normal $N(\mu, \sigma)$. Las propiedades más importantes son:

1. El estimador es insesgado, es decir, $E(s^2) = \sigma^2$.
2. El estimador es de varianza mínima.
3. La variable $\frac{(n-1)s^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2$

La media muestral en una distribución normal

La media muestral

$$\hat{\mu} = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$$

es un estimador de la media μ de una población cuya característica en estudio sigue una distribución normal $N(\mu, \sigma)$. Las propiedades más importantes son:

1. El estimador es insesgado, es decir, $E(\bar{x}) = \mu$.
2. El estimador es de varianza mínima.
3. Para valores grandes del tamaño de la muestra (en la práctica $n > 30$), la media muestral \bar{x} se distribuye según una distribución normal que depende del tamaño N_p de la población:

$$\text{Si } n > 30 \quad \text{entonces} \quad \bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N_p - n}{N_p - 1}}\right)$$

4. Si la población es infinita o el muestreo es con reposición, la segunda raíz vale 1, es decir,

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

lo que permite considerar las siguientes tipificaciones del estimador de la media:

- Si σ es conocido entonces $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1)$ pues $\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- Si σ es desconocido y $n > 30$ entonces

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \rightsquigarrow N(0, 1) \quad \text{pues} \quad \bar{x} \rightsquigarrow N\left(\mu, \frac{s}{\sqrt{n}}\right)$$

- Si σ es desconocido y $n \leq 30$ entonces $z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \rightsquigarrow t_{n-1}$

7.2.2. Estimación por intervalos

En la práctica, no sólo interesa dar una estimación puntual de un parámetro θ sino un intervalo de valores dentro del cual se tiene confianza de que esté el estimador $\hat{\theta}$ del parámetro. Por tanto, lo que buscamos es un estimador denominado “estimador por intervalo” compuesto de una pareja de estadísticos L_i (límite inferior) y L_s (límite superior) tales que

$$P(L_i \leq \hat{\theta} \leq L_s) = 1 - \alpha \quad \text{con} \quad 0 < \alpha < 1$$

donde $1 - \alpha$ se llama **nivel de confianza** y α se denomina **nivel de significación**. Es decir, llamamos **intervalo de confianza** para el parámetro θ con nivel de confianza $1 - \alpha$, a una expresión del tipo $L_i \leq \hat{\theta} \leq L_s$ donde los límites L_i y L_s dependen de la muestra y se calculan de manera tal que si construimos muchos intervalos, cada vez con distintos valores muestrales, el $100(1 - \alpha)\%$ de ellos contendrán el verdadero valor del parámetro.

Sin embargo, cuando tenemos el intervalo de confianza de una muestra concreta, o este intervalo pertenece al $100(1 - \alpha)\%$ de los que contienen al parámetro y, por lo tanto, el parámetro está en el intervalo con probabilidad 1; o bien, este intervalo pertenece al $100\alpha\%$ de los que no contienen al parámetro y, por lo tanto, el parámetro está en el intervalo con probabilidad 0. Pero como difícilmente se llegará a saber con exactitud si el intervalo concreto es de uno u otro tipo, entonces el nivel de confianza $100(1 - \alpha)\%$ nos determinará una medida de la bondad del intervalo.

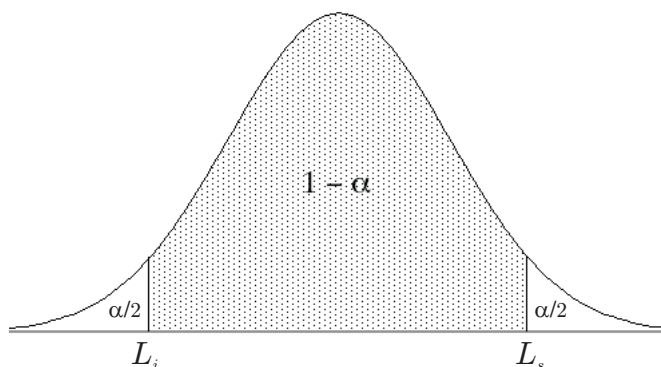


Figura 7.1: Intervalo de confianza

La amplitud del intervalo está íntimamente relacionada con los niveles de confianza y significación. Si la amplitud del intervalo es pequeña entonces la afirmación de que el parámetro pertenece al intervalo tiene gran significación (α es grande) pero ofrece poca confianza ($1 - \alpha$ es pequeña). Pero si la amplitud del intervalo es grande entonces la afirmación de que el parámetro pertenece al intervalo tiene menor significación (α es pequeño) aunque ofrece mucha confianza ($1 - \alpha$ es grande). Por ejemplo, la afirmación “la altura media de una población está entre 1’69 y 1’71 metros” con $\alpha = 0’25$ es más significativa que la afirmación “la altura media de una población está entre 1’60 y 1’80 metros” con $\alpha = 0’01$, aunque esta última afirmación ofrece más confianza $1 - \alpha = 0’99$ que la primera $1 - \alpha = 0’75$.

Las tablas del anexo presentan los principales intervalos de confianza para los parámetros μ y σ de la distribución normal $N(\mu, \sigma)$, el parámetro p de la distribución binomial $B(n, p)$, y el parámetro λ de la distribución de Poisson $P(\lambda)$. Si no se especifica o se deduce lo contrario,

supondremos que la distribuciones consideradas son de tipo normal, y que el nivel de confianza es del 95 %.

Ejemplo 7.1 *Obtener dos intervalos de confianza, uno al 99 % y otro al 95 %, para el consumo medio de combustible de un determinado tipo de coche, sabiendo que los consumos observados en 5 ensayos fueron 5'2, 4'3, 5'1, 4'7 y 4'9.*

En primer lugar, suponemos (puesto que ni se dice, ni se deduce lo contrario) que el consumo medio de gasolina de ese determinado tipo de vehículo sigue una distribución normal $N(\mu, \sigma)$ con μ y σ desconocidos.

En este caso, como el tamaño de la muestra es pequeño ($n \leq 30$) y σ es desconocido entonces

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \rightsquigarrow t_{n-1}$$

lo que nos permite determinar los extremos del intervalo de confianza para μ que resulta ser

$$I = \left[\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right]$$

A partir de la muestra, se obtiene que $n = 5$, $\bar{x} = 4'84$ y $s = 0'358$ y si el nivel de significación es $\alpha = 0'01$ (Nivel de confianza del 99 %) entonces $t_{\alpha/2, n-1} = 4'604$. Con estos datos ya podemos obtener el intervalo de confianza al 99 %, que resulta ser

$$\left[4'84 \pm 4'604 \cdot \frac{0'358}{\sqrt{5}} \right] = [4'103, 5'577]$$

Si el nivel de significación es $\alpha = 0'05$ (Nivel de confianza del 95 %) entonces $t_{\alpha/2, n-1} = 2'776$ y con este dato ya podemos obtener el intervalo de confianza al 95 %, que resulta ser

$$\left[4'84 \pm 2'776 \cdot \frac{0'358}{\sqrt{5}} \right] = [4'396, 5'284]$$

Obsérvese que al disminuir el nivel de confianza, también disminuye la amplitud del intervalo pues se pierde confianza de que el parámetro esté en el intervalo, aunque se gana significación pues la región precisa más el rango de posibles valores del parámetro. Como se puede observar en la fórmula, otra forma de reducir la amplitud del intervalo es aumentar el tamaño de la muestra.

□

7.3. Contraste de Hipótesis

Otro objetivo fundamental de la Teoría de Muestras, es confirmar o rechazar hipótesis sobre un parámetro poblacional, mediante el empleo de muestras. Es decir, contrastar una hipótesis estadísticamente es juzgar si cierta propiedad supuesta para cierta población es compatible con lo observado en una muestra de ella.

Supongamos que el parámetro de la población, que es objeto de estudio, es θ . El procedimiento que se sigue para contratar un valor de θ es el siguiente. En primer lugar, se establece a priori, antes de tomar la muestra, la hipótesis que queremos contrastar, es decir, la suposición

que queremos ver si se cumple o no. Esta hipótesis es una igualdad referida al parámetro θ , se denomina *hipótesis nula*, se denota por H_0 y será rechazada o no a la vista de los datos de la muestra.

En segundo lugar, se establece, también previamente, la llamada *hipótesis alternativa* que se denota por H_a y que será admitida cuando H_0 sea rechazada. La hipótesis alternativa puede ser de dos tipos: de tipo desigualdad “mayor que” ($>$) o “menor que” ($<$), y de tipo negación (\neq). Como veremos, cada uno de estos tipos de hipótesis dan lugar a un tipo de contraste (unilateral y bilateral, respectivamente)

En tercer lugar, se define un estadístico $\hat{\theta}$ relacionado con la hipótesis que queremos contrastar. Por ello, $\hat{\theta}$ se denomina *estadístico de contraste*. La distribución de probabilidad de este estadístico es la que nos permitirá establecer el criterio de aceptación o rechazo de la hipótesis.

A continuación, suponiendo que H_0 es verdadera, se calculan dos regiones complementarias: la *región de aceptación* y la *región crítica* (R) o *región de rechazo* de la hipótesis nula. Para establecer estas regiones se fija un valor de probabilidad α (suficientemente pequeño) que denominaremos *nivel de significación* y que representa la probabilidad de que el estadístico de contraste tome un valor en la región crítica.

Por último, a partir de los valores de la muestra, calculamos el valor $\hat{\theta}_0$ que toma el estadístico para esos valores y tomamos la decisión final con el siguiente criterio:

- Si $\hat{\theta}_0 \in R$ entonces rechazamos H_0 y aceptamos H_a .
- Si $\hat{\theta}_0 \notin R$ entonces no podemos rechazar H_0 .

Obsérvese que, en el segundo supuesto, no rechazamos la hipótesis nula. Sin embargo, eso no quiere decir que podamos afirmar que sea H_0 sea cierta, aunque tampoco podemos descartarlo y, por lo tanto, admitimos que H_0 es cierta, por una cuestión de simplicidad.

La decisión de rechazar o no la hipótesis nula está basada en los datos de la muestra y, por lo tanto, podemos cometer dos tipos de errores:

1. Error de tipo I: Rechazar H_0 cuando es cierta. La probabilidad de cometer este error es lo que hemos denominado nivel de significación (α).

$$\alpha = P(\text{rechazar } H_0 \mid H_0 \text{ es cierta}) = P(\text{aceptar } H_a \mid H_0 \text{ es cierta})$$

2. Error de tipo II: No rechazar H_0 cuando es falsa. La probabilidad de cometer este error se denota con la letra β .

$$\beta = P(\text{no rechazar } H_0 \mid H_0 \text{ es falsa})$$

Estos errores están íntimamente relacionados pues cuando α decrece entonces β crece, y no es posible encontrar contrastes que permitan simultáneamente hacer ambos errores tan pequeños como queramos. Por lo tanto, será necesario destacar una de las hipótesis de manera que no será rechazada salvo que su falsedad se haga muy evidente. En los contrastes, la hipótesis considerada es H_0 que sólo será rechazada cuando la evidencia de su falsedad supere el $100(1 - \alpha)\%$, que denominamos, *nivel de confianza*.

Al tomar un valor de α pequeño tendremos que β se aproxima a uno. Lo ideal a la hora de definir un contraste es encontrar un compromiso satisfactorio entre α y β , aunque siempre, a favor de H_0 . Denominamos *potencia del contraste* a la cantidad $1 - \beta$, es decir:

$$1 - \beta = P(\text{rechazar } H_0 \mid H_0 \text{ es falsa})$$

En la siguiente tabla se recogen las distintas situaciones que se pueden dar en función de la decisión que tomemos y con las probabilidades correspondientes:

	no rechazar H_0	rechazar H_0
H_0 es cierta	Acierto $1 - \alpha$	Error tipo I α
H_0 es falsa	Error tipo II β	Acierto $1 - \beta$

En muchos casos resulta indiferente qué hipótesis se considera la nula y cual la alternativa. Sin embargo, cuando la decisión que tomemos tenga graves consecuencias, entonces tomaremos como hipótesis nula la más desfavorable, es decir, aquella cuyas consecuencias por rechazarla cuando es cierta son más graves que las de aceptarla cuando sea falsa.

Por ejemplo, pensemos que tenemos que decidir si un acusado es inocente o culpable, si un paciente mejora o empeora ante un tratamiento, o si un vehículo de pasajeros tendrá o no un accidente. En estos ejemplos debemos tomar como hipótesis nula que el acusado es inocente, que el enfermo empeora o que el vehículo tendrá un accidente, pues en todas ellas, es más grave rechazarla cuando es cierta que admitirla siendo falsa.

En estos casos, hay que elegir la hipótesis nula a menos que la evidencia a favor de la hipótesis alternativa sea muy significativa. Es decir, sólo se aceptará la hipótesis alternativa para α próximo a cero, aunque para ellos sea necesario que β sea próximo a uno, ya que las consecuencias del error tipo I (condenar a un inocente, creer equivocadamente que el enfermo mejora ante el tratamiento, o pensar erróneamente que el vehículo no tendrá un accidente), son más graves que las del error de tipo II (liberar a un culpable, creer equivocadamente que el enfermo empeora, o pensar erróneamente que el vehículo tendrá un accidente).

Y ahora, veamos un ejemplo que pone de manifiesto tanto los conceptos y reflexiones que hemos planteado, como el procedimiento que se sigue en el contraste de hipótesis de un problema estadístico.

Ejemplo 7.2 *Consideremos un proceso de fabricación que en condiciones correctas produce componentes cuya resistencia eléctrica se distribuye normalmente con media 20 Ohm y desviación típica 0'5 Ohm. A veces, y de forma imprevisible, el proceso se desajusta, produciendo un aumento o disminución de la resistencia media de los componentes, pero sin variar la desviación típica. Para contrastar si el proceso funciona correctamente se toma una muestra de cinco unidades midiendo su resistencia, resultando 18'4, 19'2, 20'3, 19'5 y 20'1. ¿Podríamos concluir con estos datos que el proceso está desajustado?*

El problema nos dice que la distribución de probabilidad de la resistencia eléctrica de un componente es de tipo normal y el parámetro objeto de estudio es la media μ . Por lo tanto, en primer lugar, elegimos las siguientes hipótesis nula y alternativa que nos permitirán responder

a la pregunta

$$\begin{cases} H_0 & : \mu = 20 \\ H_a & : \mu \neq 20 \end{cases}$$

En segundo lugar, elegimos el estadístico de contraste \bar{x} asociado a nuestro parámetro μ y cuya distribución de probabilidad es:

$$\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{o bien} \quad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1)$$

En tercer lugar, fijamos el nivel de significación $\alpha = 0'05$ (valor por defecto) y suponiendo que H_0 es verdadera ($\mu = 20$), se calculan la región de aceptación (con probabilidad $1 - \alpha$) y la región crítica o de rechazo (con probabilidad α) a partir de la distribución de probabilidad del estadístico, de la siguiente manera:

$$1 - \alpha = P(\text{"región de aceptación"}) = P(-z_{\alpha/2} \leq \frac{\bar{x} - 20}{0'5/\sqrt{n}} \leq z_{\alpha/2})$$

y, por lo tanto, la región de aceptación para nuestro estadístico de contraste z es el intervalo $[-1'96, 1'96]$. De manera que la región crítica o de rechazo es su complementario, es decir, $(-\infty, -1'96) \cup (1'96, \infty)$ que se obtendría así:

$$\alpha = P(\text{"región crítica"}) = P\left(\left|\frac{\bar{x} - 20}{0'5/\sqrt{n}}\right| > z_{\alpha/2}\right)$$

Por último, a partir de los valores de la muestra, calculamos el valor del estadístico y tomamos la decisión final. Como $n = 5$ y $\bar{x} = 19'5$ entonces el estadístico de contraste $z = 2'236$ pertenece a la región crítica y, por lo tanto, rechazamos la hipótesis nula y aceptamos la hipótesis alternativa ($\mu \neq 20$), es decir, tendremos que suponer que el proceso se ha desajustado. \square

Obsérvese que, en este ejemplo, la región de rechazo estaba constituida por la unión de dos intervalos. Por esta razón, este tipo de contrastes se denominan bilaterales y se producen cuando la hipótesis alternativa es la negación de la hipótesis nula, es decir, cuando la hipótesis nula es de tipo “=” y la alternativa es de tipo “ \neq ”. Veamos ahora un ejemplo de contraste unilateral

Ejemplo 7.3 Con los datos del ejemplo 7.2, ¿podemos concluir que el proceso se ha desajustado por exceso?

En este caso, las hipótesis nula y alternativa que nos permitirán responder a la pregunta son

$$\begin{cases} H_0 & : \mu = 20 \\ H_a & : \mu > 20 \end{cases}$$

Elegimos el estadístico de contraste \bar{x} asociado a nuestro parámetro μ y cuya distribución de probabilidad es:

$$\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{o bien} \quad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1)$$

Si fijamos el nivel de significación $\alpha = 0'05$ y suponemos que H_0 es verdadera ($\mu = 20$) entonces se puede calcular la región crítica o de rechazo (con probabilidad α) a partir de la distribución de probabilidad del estadístico, de la siguiente manera:

$$\alpha = P(\text{"región crítica"}) = P\left(\frac{\bar{x} - 20}{0'5/\sqrt{n}} > z_{\alpha}\right)$$

y, por lo tanto, la región de rechazo es $(1'645, \infty)$. Como $n = 5$ y $\bar{x} = 19'5$ entonces el estadístico de contraste $z = -2'236$ no pertenece a la región crítica y, por lo tanto, no podemos rechazar la hipótesis nula. Eso no significa que debamos aceptarla, aunque en estos casos, es bastante común rechazar la hipótesis alternativa, de manera que afirmaríamos que la media no ha aumentado, es decir, que $\mu \leq 20$. \square

Las tablas del anexo presentan los principales contrastes de hipótesis para los parámetros μ y σ de la distribución normal, y el parámetro p de la distribución binomial. Para cada uno de ellos, se presentan las regiones críticas o de rechazo, de los distintos contrastes unilaterales y bilaterales. Si no se especifica o se deduce lo contrario, supondremos que la distribuciones consideradas son de tipo normal, y que el nivel de confianza es del 95 %.

7.4. Inferencia no paramétrica

Por lo general, para estudiar un carácter en una población, se examina solamente una muestra tomada de la población. Cualquiera que sea la población teórica que se considere, siempre existirán desviaciones entre la distribución teórica y la distribución empírica u observada. El problema consiste, por tanto, en saber en qué medida estas desviaciones son debidas a:

1. *El azar.* Estas diferencias tienden a desaparecer si el número de observaciones (tamaño de la muestra) es suficientemente grande.
2. *Tomar una distribución teórica inadecuada.*

En este último caso, la distribución χ^2 de Pearson se puede aplicar para ver si un conjunto de datos observados coincide o no con un conjunto de datos esperados.

A continuación se enumeran las principales aplicaciones de la χ^2 . En cada una de ella se trata de contrastar si una cierta hipótesis H_0 es coherente con los datos obtenidos en la muestra.

1. *Bondad de ajuste:* Se trata de determinar si la hipótesis sobre el tipo de distribución teórica (binomial, poisson, normal, etc.) que rige un experimento es consistente con los datos que aparecen en la muestra.
2. *Contraste de homogeneidad de varias muestras:* Se trata de contrastar si varias muestras con un mismo carácter han sido o no tomadas de una misma población.
3. *Contraste de dependencia o independencia de caracteres:* Se trata de comparar si dos o más distribuciones empíricas son comparables a una misma distribución teórica. Y esto se utiliza para ver si dos caracteres son o no independientes.

En todos los casos se realiza el test de la χ^2 que consiste en lo siguiente: Supongamos que al tomar una muestra los posibles sucesos x_1, x_2, \dots, x_k se presentan con frecuencias o_1, o_2, \dots, o_k , llamadas frecuencias observadas, y que según las leyes de la probabilidad, se esperaba que apareciesen con frecuencias e_1, e_2, \dots, e_k , llamadas frecuencias esperadas o teóricas. Una medida de la discrepancia entre las frecuencias esperadas y las observadas viene proporcionada por el estadístico $\hat{\chi}^2$ definido por

$$\hat{\chi}^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_k - e_k)^2}{e_k}$$

de manera que si $\hat{\chi}^2 = 0$ entonces las frecuencias observadas y teóricas coinciden completamente, mientras que si $\hat{\chi}^2 > 0$, estas frecuencias no coinciden exactamente. A mayor valor de $\hat{\chi}^2$ mayor discrepancia entre las frecuencias esperadas y las observadas.

Para contrastar si las frecuencias observadas difieren significativamente de las esperadas utilizaremos que la distribución del estadístico $\hat{\chi}^2$ se aproxima muy bien si $k \geq 5$ y $e_i \geq 5$ por la distribución χ_v^2 . El número de grados de libertad viene dado por $v = k - 1 - m$, siendo m el número de parámetros de la población que ha sido necesario estimar, a partir de estadísticos de la muestra, para poder calcular las frecuencias teóricas.

7.4.1. Bondad de ajuste. Tabla de contingencia

Consideremos en una población el carácter X que admite las modalidades x_1, \dots, x_k excluyentes (o una variable continua y dividimos el recorrido en k clases). Se toma una muestra de tamaño n de la población, siendo o_i el número de elementos que presentan la modalidad x_i (frecuencia observada de x_i). Si denotamos por p_i la probabilidad que teóricamente asignamos a la modalidad x_i , entonces las frecuencias esperadas para cada x_i serán $e_i = n \cdot p_i$ con $i = 1, \dots, k$.

Con estos datos podemos construir la siguiente tabla

X	x_1	x_2	\dots	x_i	\dots	x_k
Frec. Observada	o_1	o_2	\dots	o_i	\dots	o_k
Frec. Esperada	e_1	e_2	\dots	e_i	\dots	e_k

que recibe el nombre de *tabla de contingencia* $1 \times K$ y cuyos elementos verifican:

$$\sum_{i=1}^k o_i = n \quad , \quad \sum_{i=1}^k p_i = 1 \quad \text{y} \quad \sum_{i=1}^k e_i = n$$

Ahora consideramos la hipótesis H_0 que consiste en suponer que la distribución teórica escogida representa bien a la distribución empírica y que, por tanto, las desviaciones entre las frecuencias observadas y las teóricas son debidas al azar. Veamos en qué condiciones podemos aceptar o rechazar la hipótesis H_0 . Para ello, definimos el estadístico

$$\hat{\chi}^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \dots + \frac{(o_k - e_k)^2}{e_k} = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

que sigue aproximadamente una distribución χ^2 de Pearson con $k - 1$ grados de libertad si no existe diferencia significativa entre las frecuencias observadas y las teóricas. Así pues, a un nivel de significación α , tenemos que:

$$\begin{cases} \text{Si } \hat{\chi}^2 < \chi_{\alpha; k-1}^2 & \text{se acepta la hipótesis a nivel } \alpha. \\ \text{Si } \hat{\chi}^2 \geq \chi_{\alpha; k-1}^2 & \text{se rechaza la hipótesis a nivel } \alpha. \end{cases}$$

Para calcular el estadístico $\hat{\chi}^2$ podemos utilizar la siguiente igualdad:

$$\sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{o_i^2}{e_i} - n$$

En el test de la χ^2 hay que hacer las siguientes consideraciones:

1. Si la distribución que queremos ajustar es continua, determinaremos, siempre que sea posible, k clases excluyentes, con $k \geq 5$, que determinarán las modalidades de la variable.
2. Si hay alguna modalidad que tenga alguna frecuencia esperada menor que cinco se agrupan dos o más modalidades contiguas en una sola hasta lograr que la nueva frecuencia sea mayor o igual que cinco.
3. Si para obtener las frecuencias esperadas, necesitamos calcular m parámetros de la distribución teórica entonces los grados de libertad de la distribución χ^2 son $k - m - 1$.
4. Si el estadístico $\hat{\chi}^2$ es demasiado próximo a cero, debe mirarse con suspicacia el experimento, pues es raro que las frecuencias observadas coincidan demasiado bien con las frecuencias esperadas. Para estudiar estas situaciones podemos examinar si el valor de $\hat{\chi}^2$ es menor que $\chi_{0'95;v}^2$ ó $\chi_{0'99;v}^2$, en cuyo caso decidimos que el acuerdo es demasiado bueno al nivel de significación 0'05 ó 0'01 respectivamente.

Ejemplo 7.4 La siguiente tabla contiene las notas (sobre 100) que han obtenido los alumnos en Estadística en los últimos 5 años clasificadas en rangos de 10 puntos:

Rango Nota	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frecuencia	2	30	80	145	250	245	140	75	28	5

¿Se puede afirmar al 95% que las distribución de las notas es de tipo normal?

El problema nos plantea si la distribución de los datos corresponde a una distribución normal $N(\mu, \sigma)$. En primer lugar, utilizaremos la muestra para determinar los parámetros de la distribución normal, mediante estimación puntual.

$$\hat{\mu} = \bar{x} = 49'84 \quad \text{y} \quad \hat{\sigma} = s = 16'088$$

En segundo lugar, como el propio enunciado ya establece las modalidades, construimos la tabla de contingencia 1×10

x_i	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
o_i	2	30	80	145	250	245	140	75	28	5
e_i	6'64	25'18	76'94	161'64	233'57	232'18	158'77	74'67	24'14	6'28

a partir de los valores de probabilidad p_i obtenidos de la distribución normal $N(49'84, 16'088)$, para cada uno de los intervalos de notas (modalidades), los cuales permiten calcular los valores esperados $e_i = 1000p_i$. Por ejemplo, el valor esperado 25'18 para la modalidad 10-20 se ha obtenido multiplicando 1000 por 0'02518, siendo $0'02518 = P(10 \leq X \leq 20)$ con $X \sim N(49'84, 16'088)$.

Ahora, se calcula el estadístico de contraste

$$\hat{\chi}^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{o_i^2}{e_i} - n = 10'956$$

y se compara con el valor de la χ^2 con 7 grados de libertad ($v = k - m - 1 = 10$ modalidades - 2 parámetros estimados - 1) para el valor por defecto $\alpha = 0'05$:

$$\hat{\chi}^2 = 10'956 < 14'067 = \chi_{7,0'05}^2$$

y, por lo tanto, se acepta que la distribución de las notas es de tipo normal. \square

7.4.2. Contraste de homogeneidad de varias muestras

Una muestra es homogénea cuando todas las observaciones se rigen por la misma distribución de probabilidades. En otro caso se dice que la muestra es heterogénea.

Las causas más importantes por las cuales una muestra no es homogénea son:

- La población es heterogénea respecto a la variable estudiada. Por ejemplo el nivel de renta en una población difiere según se trate de una zona urbana o rural.
- La población es homogénea respecto a la variable del estudio, pero en el proceso de muestreo se producen errores o cambios en el sistema de medida, a consecuencia de lo cual ciertos datos de la muestra son heterogéneos.

El objetivo es determinar si varias muestras de un mismo carácter X han sido o no tomadas de una misma población. Para ello usaremos el test de la χ^2 de Pearson de la siguiente manera:

Supongamos que se tienen k muestras con n_1, n_2, \dots, n_k elementos cada una. Las cuales tienen, respectivamente, o_1, o_2, \dots, o_k elementos con una determinada característica A .

Hacemos la hipótesis H_0 , que consiste en suponer que todas las muestras proceden de la misma población. Bajo esta hipótesis, la proporción p de elementos con la característica A es

$$p = \frac{o_1 + o_2 + \dots + o_k}{n_1 + n_2 + \dots + n_k}$$

y el número de elementos esperados en la muestra que poseen la característica A es:

$$e_i = n_i \cdot p \quad \text{para todo } i = 1, 2, \dots, k.$$

El problema ahora es determinar si la diferencia entre las frecuencias observadas y las esperadas se debe al azar o si se debe a que las muestras no se pueden considerar como procedentes de una misma población. Para ello, definimos el estadístico:

$$\chi_{k-1}^2 = \frac{1}{p(1-p)} \cdot \sum_{i=1}^k \frac{(o_i - e_i)^2}{n_i}$$

que, si H_0 es cierta, sigue aproximadamente una χ^2 con $k - 1$ grados de libertad. El número de grados de libertad es $k - 1$ ya que tenemos $2k$ variables (frecuencias esperadas) y hay que restar $k + 1$ parámetros que hemos obtenido de la muestra (el parámetro p y los k parámetros $n_i - n_i \cdot p$).

Luego, al nivel de significación α podemos establecer:

$$\begin{cases} \text{Si } \chi_{k-1}^2 < \chi_{\alpha; k-1}^2 & \text{Se acepta } H_0 \\ \text{Si } \chi_{k-1}^2 \geq \chi_{\alpha; k-1}^2 & \text{Se rechaza } H_0 \end{cases}$$

De manera análoga podemos contrastar si la frecuencia de un elemento de la población se mantiene constante a lo largo de las extracciones o, lo que es lo mismo, las muestras provienen de una población determinada. Así, en una población binomial se puede contrastar la hipótesis

de que la proporción de elementos con una característica A es constante e igual a p . Entonces el estadístico:

$$\hat{\chi}^2 = \frac{1}{p(1-p)} \cdot \sum_{i=1}^k \frac{(o_i - n_i p)^2}{n_i}$$

sigue aproximadamente una distribución χ^2 con k grados de libertad si la hipótesis es verdadera. Luego si $\chi_k^2 < \chi_{\alpha; k}^2$ se acepta la hipótesis y en otro caso se rechaza a un nivel de significación α .

En una población de Poisson se puede contrastar la hipótesis de que el número medio de elementos con la característica A en cada muestra es constante, es decir:

$$\lambda = \bar{o} = \frac{\sum_{i=1}^k o_i}{k} = \text{constante}$$

entonces, el estadístico

$$\hat{\chi}^2 = \sum_{i=1}^k \frac{(o_i - \bar{o})^2}{\bar{o}} = \frac{1}{\bar{o}} \cdot \sum_{i=1}^k o_i^2 - \sum_{i=1}^k o_i$$

sigue aproximadamente una distribución χ^2 con $k - 1$ grados de libertad si la hipótesis es verdadera.

7.4.3. Contraste de dependencia o independencia de caracteres. Tablas de contingencia $K \times M$

Hasta ahora hemos utilizado el test de la χ^2 para saber si una serie de datos se ajustaban o no a una distribución teórica. Podemos igualmente comparar dos o más distribuciones empíricas entre sí si cada una de ellas es comparable a una misma distribución teórica.

Supongamos que queremos comparar dos caracteres X e Y en una misma población, que admiten las modalidades siguientes: $X = \{x_1, x_2, \dots, x_k\}$ e $Y = \{y_1, y_2, \dots, y_m\}$. Para ello, tomamos una muestra de tamaño n , siendo o_{ij} el número de elementos que presentan la modalidad x_i de X e y_j de Y (frecuencia observada).

Si consideramos la hipótesis H_0 que consiste en suponer que no existen diferencias significativas entre las dos distribuciones empíricas de X e Y , entonces con cada frecuencia observada o_{ij} tenemos una frecuencia teórica o esperada e_{ij} que podemos calcular mediante la expresión

$$e_{ij} = p_{ij} \cdot n = \frac{o_{x_i} \cdot o_{y_j}}{n} \quad \text{para todo } i = 1, 2, \dots, k \text{ y } j = 1, 2, \dots, m$$

siendo p_{ij} las probabilidades de que un elemento tomado de la muestra presente las modalidades x_i de X e y_j de Y , es decir

$$p_{ij} = \frac{o_{x_i}}{n} \cdot \frac{o_{y_j}}{n}$$

Con estos datos podemos construir la siguiente tabla

$X \setminus Y$	$y_1 \dots$	$y_j \dots$	y_m	Frecuencia o_{x_i}
x_1	$o_{11} \dots$	$o_{1j} \dots$	o_{1m}	o_{x_1}
\vdots	\vdots	\vdots	\vdots	\vdots
x_i	$o_{i1} \dots$	$o_{ij} \dots$	o_{im}	o_{x_j}
\vdots	\vdots	\vdots	\vdots	\vdots
x_k	$o_{k1} \dots$	$o_{kj} \dots$	o_{km}	o_{x_k}
Frecuencias o_{y_j}	$o_{y_1} \dots$	$o_{y_j} \dots$	o_{y_m}	n

que recibe el nombre de *tabla de contingencia* $K \times M$ y cuyos elementos verifican:

$$\sum_{i=1}^k \sum_{j=1}^m o_{ij} = n \quad , \quad \sum_{i=1}^k o_{x_i} = \sum_{j=1}^m o_{y_j} = n \quad , \quad \sum_{i=1}^k \sum_{j=1}^m p_{ij} = 1 \quad y \quad \sum_{i=1}^k \sum_{j=1}^m e_{ij} = n$$

Análogamente a los casos anteriores, definimos el estadístico:

$$\hat{\chi}^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^k \sum_{j=1}^m \frac{o_{ij}^2}{e_{ij}} - n$$

que sigue aproximadamente una distribución $\chi^2_{(k-1)(m-1)}$ si es cierta H_0 , con $e_{ij} > 5$, para todo $1 \leq i \leq k, 1 \leq j \leq m$; en otro caso es preciso agrupar filas o columnas contiguas. Así pues, a nivel de significación α podemos contratar la hipótesis H_0 :

$$\begin{cases} \text{Si } \chi^2_{(k-1) \cdot (m-1)} < \chi^2_{\alpha; (k-1) \cdot (m-1)} & \text{se acepta } H_0 \\ \text{Si } \chi^2_{(k-1) \cdot (m-1)} \geq \chi^2_{\alpha; (k-1) \cdot (m-1)} & \text{se rechaza } H_0 \end{cases}$$

Este contraste no paramétrico se utiliza muy frecuentemente para ver si existe o no relación entre los caracteres X e Y , es decir, si son o no independientes. Entonces recibe el nombre de *contraste de independencia de caracteres*:

$$\begin{cases} \text{Si } \chi^2_{(k-1) \cdot (m-1)} < \chi^2_{\alpha; (k-1) \cdot (m-1)} & XyY \text{ son independientes al nivel } \alpha \\ \text{Si } \chi^2_{(k-1) \cdot (m-1)} \geq \chi^2_{\alpha; (k-1) \cdot (m-1)} & XeY \text{ no son independientes al nivel } \alpha \end{cases}$$

Coefficiente de contingencia

Una medida del grado de relación o dependencia entre dos caracteres X e Y en una tabla de contingencia viene dada por el *coeficiente de contingencia* C que se define por

$$C = \sqrt{\frac{\hat{\chi}^2}{\hat{\chi}^2 + n}}$$

A mayor valor de C más alto es el grado de dependencia entre las dos variables X e Y .

7.5. Relación de problemas

1. Un nuevo modelo de automóvil realiza 10 pruebas de consumo con 9 litros de gasolina, obteniéndose: 137'4, 136, 132, 141, 129, 130'8, 140, 129'7, 133 y 136 kilómetros recorridos en cada prueba.
 - a) Utilizar los resultados de las pruebas anteriores para estimar la media y la varianza del consumo de gasolina (suponer que los datos están normalmente distribuidos).
 - b) Estimar un intervalo de confianza para la media de kilómetros recorridos con 9 litros de combustible.
2. Una muestra de 10 medidas de las constantes recuperadoras de muelles para amortiguadores da una media de 15 Nw/mm con desviación típica de 0'2. Encontrar intervalos de confianza al 5 % de la media y de la varianza.
3. El contenido medio en grasa para dos tipos de queso A y B es $\bar{x}_A=33'2\%$ y $s_A=3'4\%$ con $n = 27$ y $\bar{x}_B=35'4\%$ y $s_B=3\%$ con $n = 42$. Se pide:
 - a) Construir un intervalo de confianza al 95 %, para el porcentaje de grasa de los tipos A y B.
 - b) Construir un intervalo de confianza para la diferencia en el contenido en grasa de ambos tipos.
 - c) ¿Se observan diferencias significativas?
4. Lanzamos una moneda 200 veces.
 - a) Halle un intervalo donde se encontrará el número de caras obtenidas con una probabilidad del 99 %, supuesta la moneda equilibrada.
 - b) Si se obtienen 110 veces caras, ¿deberá suponerse al 99 % que está trucada?
 - c) Si se obtiene una proporción de caras del 45 %, ¿cual debe ser el número mínimo de tiradas para rechazar la hipótesis de estar equilibrada?
5. En un sondeo a 500 votantes del barrio A y 300 del barrio B, un candidato resultó preferido por el 43 % de los de A y el 42 % de los de B. Al nivel $\alpha = 5\%$.
 - a) Obtener intervalos de confianza para los resultados esperados en A y en B.
 - b) ¿Puede admitirse que el candidato obtendrá mejores resultados en A que en B?
6. Una compañía aseguradora comprueba que la probabilidad, para determinado grupo de riesgo, de tener un accidente mortal en un periodo del año es de 0'003. Cada accidente provoca un pago fijo de 100 000 euros. Si la compañía tiene 10 000 asegurados.
 - a) Estimar la prima anual que, a un nivel del 1 %, asegure que no se provocarán pérdidas en la empresa.
 - b) Responder a la pregunta anterior si tuviese 100 000 asegurados.
7. Para comprobar si un fármaco es útil en el tratamiento de una enfermedad, de la que datos anteriores nos dan un plazo de recuperación de 34 días con una desviación de 7, tomamos una muestra de 50 pacientes, suministrándosele a 25 de ellos (grupo A) un placebo y a los otros 25 (grupo B) el tratamiento.

El grupo A tuvo un periodo medio de recuperación de 25 días con desviación de 5, mientras el grupo B obtuvo una media de 24 días y desviación 5. Contrastar:

- a) Que el tratamiento es eficaz sobre los métodos anteriores.
 - b) Que su eficacia es psicológica, pues no difiere de los individuos no tratados, que creen que si lo son (grupo A).
8. Un equipo médico sostiene que su tratamiento ha conseguido sanar al 90 % de los pacientes de una enfermedad en 3 días. Realizado un experimento con 400 pacientes sanaron 342 en dicho plazo.
- a) Utilizar el resultado del experimento para estimar la proporción de pacientes que reaccionan favorablemente al tratamiento.
 - b) Contrastar si la hipótesis que sostiene el equipo médico es correcta.
9. Para contrastar el nivel de Matemáticas de los alumnos de dos centros de enseñanza se selecciona un grupo de alumnos de cada centro y se les somete a una prueba de nivel. Las calificaciones obtenidas por los grupos de 40 y 30 alumnos de los centros de enseñanza A y B dan una media de $5'4$ y $5'7$ respectivamente; mientras que las desviaciones típicas resultan ser respectivamente de $1'3$ y $0'9$. Contrastar al nivel del 90 % que no existen diferencias en el nivel de conocimientos de Matemáticas entre ambos centros de enseñanza.
10. Para analizar el efecto de un tratamiento contra la procesionaria del pino, se divide el terreno en 100 parcelas de las que aleatoriamente se tratan 40, obteniéndose en las tratadas 20 árboles atacados de 230 observados, mientras en las no tratadas se observaron 7 atacados de 300 observados. ¿Se puede deducir que el tratamiento es eficaz al 10 %? ¿Y al 5 %?
11. Para estimar el número de castaños de un bosque con 500 km^2 , se seleccionan aleatoriamente 10 parcelas de 1 km^2 cada una, contando exhaustivamente los castaños existentes. Obteniéndose 25, 27, 32, 28, 23, 20, 28, 19, 17 y 20 en cada una de las parcelas.
- a) Hallar un intervalo de confianza al 5 % para el número medio de castaños por km^2 y para el bosque completo.
 - b) Hallar si puede aceptarse que existen más de 15.000 castaños en el bosque (con nivel de significación $\alpha = 0,01$).
 - c) Si en vez de 10 parcelas de 1 km^2 , se hubiesen considerado 100 parcelas de $0'1 \text{ km}^2$, obteniéndose una media de $1/10$ de la anterior y una desviación típica de $1/10$ de la anterior. ¿Se obtiene un intervalo de confianza del número de castaños del bosque más preciso?
12. En una muestra de 50 neumáticos de cierta clase se obtuvo una vida media de 32 000 km. y una desviación estándar (s) de 4 000 km.
- a) ¿Puede afirmar el fabricante que la vida media de esos neumáticos es mayor que 30 000 km.? Establezca y pruebe la hipótesis correspondiente en un nivel del 5 %, suponiendo normalidad.
 - b) ¿Hasta qué número de kilómetros podríamos afirmar que llega la vida media de los neumáticos con el mismo nivel de confianza?

13. Si mediciones simultáneas de una tensión eléctrica por medio de dos tipos diferentes de voltímetros proporcionan las diferencias (en volts) 0'8, 0'2, -0'3, 0'1, 0'0, 0'5, 0'7 y 0'2, ¿puede afirmarse al 4 % que no existen diferencias significativas, en la calibración de los dos tipos de instrumentos?
14. Supóngase que, en un equipo eléctrico alimentado con baterías, es más económico reemplazar todas estas a intervalos fijos que reemplazar cada batería por separado cuando se agota, ello ocurre cuando la desviación estándar de la vida de las mismas sea mayor que cierto límite, esto es, mayor que 5 horas. Plantee y aplique una prueba apropiada, utilizando una muestra de 28 valores de vida con desviación estándar $s = 3'5$ horas y suponiendo normalidad. Tome $\alpha = 6\%$.
15. Suponga que las marcas I y II de focos eléctricos tienen el mismo precio y son de la misma calidad, excepto, tal vez, por su vida útil. Un cliente compró 100 focos de cada marca y comprobó que los focos de la marca I tenían una vida media de 1 120 horas con una desviación estándar de 75 horas; y para los focos de la marca II los valores correspondientes fueron 1 064 y 82 horas, respectivamente. ¿Es significativa la diferencia en media, de la vida útil de los focos de ambas marcas?
16. *Determinación del tamaño muestral.* En los ejercicios anteriores, hemos supuesto conocido el tamaño de la muestra. El problema de determinar el tamaño muestral es crucial ya que un tamaño de la muestra excesivamente elevado puede resultar costoso en tiempo y dinero, sin embargo, si la muestra es demasiado pequeña podemos no encontrar el grado deseado de fiabilidad (la amplitud del intervalo es inversamente proporcional a la raíz cuadrada del tamaño de la muestra). Se trata de despejar la variable n en los estadísticos de los extremos del intervalo de confianza correspondiente
 - a) Consideremos el intervalo de confianza para la media de una distribución normal de varianza conocida. Determinar el tamaño de la muestra que debemos considerar de forma que la diferencia entre la media poblacional y la media muestral sea, en valor absoluto, menor que un cierto error (ϵ) a un determinado nivel de confianza $(1-\alpha)$.
 - b) Una muestra aleatoria de 196 datos extraídos de una población normal de varianza igual a 100, presenta una media muestral de 160.
 - 1) Determinar al 95 % un intervalo de confianza para la media poblacional y señalar la diferencia máxima entre la media muestral y la desconocida media poblacional.
 - 2) Si se quiere tener una confianza del 95 % de que la estimación de la media se encuentra a una distancia de 1'2 de la verdadera media poblacional, ¿debemos tomar más observaciones adicionales? ¿Cuántas?
17. *Nivel crítico.* En los ejercicios anteriores, hemos realizado los cálculos para un determinado nivel de confianza. A partir de los datos de una muestra (tamaño, media, varianza, proporción, etc.) podemos estar interesados en determinar el nivel de confianza crítico a partir del cual se acepta o rechaza una determinada hipótesis sin más que aumentar o disminuir este nivel de confianza. Se trata pues de despejar el valor de α en el contraste de hipótesis correspondiente.

El cálculo del nivel crítico resulta útil para poder manipular el resultado de un determinado contraste, puesto que una vez calculado, se puede fácilmente establecer el nivel de confianza para que los resultados de la muestra respalden una determinada hipótesis.

Ejemplo: Un determinado partido político realiza un sondeo preelectoral y obtiene un 55 % de éxitos en intención de voto. Con estos resultados, ¿cuál es el mínimo nivel de confianza que le permite asegurar que podrá obtener mayoría absoluta, si la consulta se realizó sobre 240 votantes?

18. El 70 % de las bellotas que produce un árbol son comidas por los animales y el resto germina en un 60 %

- a) Hallar un intervalo de confianza con $\alpha = 0'02$ para el número de bellotas germinadas, procedentes de un árbol determinado que produjo 20 000 bellotas.
- b) Un ingeniero agrónomo afirma que en determinado tipo de suelo la proporción de las que germinan es del 75 %. Para ello se dejan caer 100 bellotas impidiendo el acceso de animales, de las cuales germinan 67.
 - 1) Contrastar al nivel $\alpha = 0'01$, que en ese tipo de suelo se produce un aumento del porcentaje de germinación.
 - 2) Contrastar que la hipótesis del ingeniero es cierta.

19. Analizada la operación de montaje de una máquina de un equipo, se observa que puede ser realizada en dos secuencias diferentes A y B. Para evitar la posible influencia del entrenamiento de los operarios, se seleccionaron aleatoriamente 18, que desconocían el proceso de montaje, asignándoles aleatoriamente al aprendizaje del montaje de una u otra secuencia, tras un mes de aprendizaje, se realizaron mediciones obteniéndose los siguientes tiempos de montaje:

Procedimiento A:	32	37	35	28	41	44	35	31	34
Procedimiento B:	35	31	29	25	34	40	27	32	31

- a) Obtener intervalos al nivel de confianza del 99 % para la media del tiempo de montaje por uno y otro método.
 - b) Contrastar al nivel $\alpha = 0'10$ la igualdad de varianzas de ambos métodos.
 - c) De acuerdo al resultado obtenido en el apartado anterior, contrastar al nivel $\alpha = 0'1$ la igualdad de las medias de ambos métodos.
20. Se desea contrastar si la temperatura del agua del mar en Alicante es mayor que en Málaga y para ello se realizaron mediciones cada dos meses durante un año, resultando:

Alicante	14°	16°	18°	21°	22°	14°
Málaga	12°	16°	19°	21°	21°	13°

Realizar el contraste al nivel $\alpha = 0'05$ de significación.

21. Un estudio del precio de los pisos en una ciudad resultó que en el año 1992 se distribuían normalmente con media 100 000 ptas/m² y desviación típica de 8 000 ptas/m².
- a) Estimar el precio mínimo por metro cuadrado que no alcanzan el 25 % de los pisos.
 - b) Si elegimos una muestra al azar de 10 pisos, hallar la probabilidad de que alguno cueste más de 125 000 ptas/m².

- c) En un estudio posterior (1997) se estudian 30 pisos al azar, obteniéndose una media de 105 000 ptas/m², con $s = 10\,000$ ptas/m². Estudiar si es admisible, al nivel de significación $\alpha = 0'1$ que la varianza se ha mantenido.
- d) Analizar si puede admitirse que la media ha aumentado, con las mismas hipótesis del apartado anterior.
22. Se quiere estudiar si la velocidad media de lectura es mayor en ambiente urbano que en rural; para ello, se toma una muestra de 500 personas de tipo urbano, resultando en palabras por minuto:

$$\sum_{i=1}^{500} p_i = 75000 \quad , \quad \sum_{i=1}^{500} p_i^2 = 14'23 \cdot 10^6$$

siendo p_i el número de palabras por minuto del individuo i -ésimo, mientras que para una muestra de 300 personas de ambiente rural, dieron unos resultados de:

$$\sum_{i=1}^{300} p_i = 43500 \quad , \quad \sum_{i=1}^{300} p_i^2 = 6'83 \cdot 10^6$$

Dar un intervalo de confianza para la diferencia de las velocidades medias en ambos ambientes.

23. Un método de depuración de aguas residuales mediante tratamiento con cloro deja un contenido medio de impurezas de $1'48 \text{ mg/m}^3$ con $\sigma = 0'13$.

Un método alternativo con metano produce, mediante muestreo aleatorio simple, los siguientes resultados en dos sectores de una ciudad:

Sector A: Media=1'45 $n_A=10$ $s_A=0'3$

Sector B: Media=1'43 $n_B=20$ $s_B=0'35$

Contrastar si existen diferencias en media y varianza para la muestra total al nivel 0'1 entre el método con metano y el método con cloro.

Sugerencia: Calcular primero la media y cuasivarianza para el total de los 30 datos muestrales

24. La elasticidad del plástico puede variar dependiendo del proceso por el cual se prepara. Para comparar la elasticidad del plástico producido por dos procesos diferentes se tomaron seis muestras extraídas de cada uno de los procesos, obteniéndose los siguientes resultados:

Proceso A:	6'1	9'2	8'7	7'5	9'0	7'3
Proceso B:	9'2	8'1	6'9	7'9	6'5	9'0

- a) Calcular dos intervalos de confianza al 95 %, uno para la elasticidad media y otro para la varianza de los datos obtenidos en el proceso A. Interpretar los resultados.
- b) Cuestión teórica: Deseamos ser más precisos en nuestras afirmaciones sobre la media y varianza de la elasticidad de los plásticos fabricados de acuerdo al proceso A, es decir, queremos ofrecer intervalos de confianza con menor amplitud. ¿Qué dos soluciones se pueden plantear? Razonar la respuesta.
- c) ¿Presentan los datos suficiente evidencia para poder asegurar que existe diferencia entre las elasticidades medias de los dos procesos? Usar $\alpha = 0'05$. Si la respuesta es afirmativa, contrastar qué proceso obtiene un plástico de mayor resistencia.

- d) Obtener un intervalo de confianza al 95 % para la diferencia de las medias de elasticidad de los procesos. A la vista del intervalo calculado, ¿qué respuesta se puede dar a la pregunta del apartado anterior? Comentarla.
25. Los científicos consideran que el benceno es un agente químico que puede causar el cáncer. Diversos estudios han comprobado que la gente que trabaja con benceno durante más de 5 años, tiene 20 veces más probabilidad de contraer leucemia. Como resultado se impuso una Ley para limitar el nivel medio de benceno en el ambiente de trabajo a un máximo de 1 ppm.

Un estudio en una planta productora consiste en tomar 10 muestras del aire en periodos de tiempo regulares (días sucesivos) obteniéndose:

0'95 , 0'97 , 0'90 , 0'88 , 1'00 , 1'05 , 1'18 , 1'13 , 1'15 , 1'09

Estimar si debe aceptarse al 95 % la hipótesis de estar violando el límite medio permitido. ¿Se podría afirmar lo mismo al 80 %? ¿Por qué?

26. El *tiempo de respuesta de un ordenador* se define como el tiempo que un usuario debe esperar mientras el ordenador accede a la información en el disco. Suponga que un centro de datos desea comparar los tiempos de respuesta de dos unidades de disco de ordenador. Se seleccionaron muestras aleatorias independientes de 13 tiempos de respuesta para el Disco 1 y 16 tiempos de respuesta para el Disco 2. A continuación, se presentan los datos registrados en milisegundos:

Disco 1:	59	92	54	102	73	60	73	75	74	84	47	33	61			
Disco 2:	71	38	47	53	63	48	41	68	40	60	44	39	34	75	86	73

Se pide:

- Calcular dos intervalos de confianza al 95 %, uno para el tiempo medio de respuesta y otro para la varianza del tiempo de respuesta del Disco 1. Interpretar los resultados.
 - Cuestión teórica: Deseamos ser más precisos en nuestras afirmaciones sobre la media y varianza del tiempo de respuesta, es decir, queremos ofrecer intervalos de confianza de menor amplitud. ¿Qué dos soluciones se pueden plantear? Razonar la respuesta.
 - Contrastar si podemos considerar que los tiempos medios de respuesta de ambos discos son iguales. Si no lo son, establecer la hipótesis y contrastar cuál de ellos es más rápido.
27. Pensamos que el porcentaje de piezas defectuosas fabricadas por una determinada máquina es del 45 %. Para contrastar nuestra hipótesis se han seleccionado 25 piezas detectándose entre ellas 16 defectuosas.
- ¿Estábamos en lo cierto al 95 %? ¿Y al 99 %?
 - Dar una explicación si las respuestas a los apartados anteriores son distintas.
 - Si alguna respuesta es negativa, proponer una afirmación sobre si el porcentaje real es mayor o menor que lo que pensábamos y contrastar dicha afirmación.

28. En 200 tiradas de una moneda, han salido 115 caras y 85 cruces. Contrastar la hipótesis de que la moneda es buena, con nivel de significación (a) 0'05 y (b) 0'01. Utilice, en primer lugar un contraste paramétrico, y compare los resultados con los que se obtendrían utilizando un contraste no paramétrico.
29. En 120 lanzamientos de un dado las distintas caras del dado han aparecido con frecuencias: 25, 17, 15, 23, 24 y 16. Contrastar al nivel 0'05 que el dado no está trucado.
30. En 360 tiradas de un par de dados, han salido 80 sietes y 30 onces. Al nivel de significación del 0'05 contrastar que los dados no están sesgados.
31. Para contrastar una hipótesis no paramétrica se ha realizado tres veces un mismo experimento. Los valores de χ^2 son 2'37, 2'86 y 3'54 cada uno con un grado de libertad. Verificar que aunque H_0 no se puede rechazar al nivel 0'05 usando un único experimento de los anteriores, sí se puede rechazar cuando se combinan los tres.
32. Se lanzan cinco monedas 1000 veces. Se considera o_i el número de veces que han salido i caras en el experimento, resultando la sucesión

$$o_0 = 38 \quad , \quad o_1 = 144 \quad , \quad o_2 = 342 \quad , \quad o_3 = 287 \quad , \quad o_4 = 164 \quad \text{y} \quad o_5 = 25$$

Ajustar una distribución binomial y contrastar la bondad del ajuste.

33. El número de individuos que poseen los cuatro grupos sanguíneos debe estar en las proporciones $q^2 : p^2 + 2pq : r^2 + 2qr : 2pr$, siendo $p + q + r = 1$. Dadas las frecuencias observadas 180, 360, 132 y 98, verificar la compatibilidad de los resultados con $p = 0'4$, $q = 0'5$ y $r = 0'1$.
34. Las leyes de la herencia de Mendel predicen la aparición de tipos de guisantes en la relación 9 : 3 : 3 : 1 para las clases lisa y amarilla, lisa y verde, arrugada y amarilla, arrugada y verde. En un experimento se obtuvieron, respectivamente, 315, 108, 101 y 32. A un nivel de 0'05, ¿coinciden los datos con la teoría?
35. En un laboratorio se observó el número de partículas α que llegan a una determinada zona procedentes de una sustancia radiactiva en un corto espacio de tiempo, siempre igual, anotándose los resultados en la siguiente tabla:

Número de partículas	0	1	2	3	4	5
Número de periodos de tiempo	120	200	140	20	10	2

- a) Ajuste una distribución de Poisson.
- b) Calcule la probabilidad con que llegan.
- c) Verifique si el ajuste es correcto mediante una χ^2 , con un nivel $\alpha = 0'05$.
36. En un examen de estadística, se obtuvieron las siguientes calificaciones:

60, 70, 90, 85, 90, 50, 75, 90, 80, 70, 60, 75, 75, 75, 80, 60, 65, 60
90, 70, 60, 70, 65, 50, 85, 80, 90, 85, 80, 75, 50, 55, 60, 65, 70, 75

Comprobar si las calificaciones obtenidas se distribuyen según una normal a un nivel 0'05.

37. En un hospital se ensayó la eficacia de cinco medicamentos en un grupo de pacientes, con el objeto de determinar si al final del tratamiento un paciente determinado mejoraba o no. Las observaciones que se encontraron están anotadas en la siguiente tabla:

Tratamientos	A	B	C	D	E	Total
Número de Pacientes	51	54	48	49	48	250
Pacientes mejorados	12	8	10	15	5	50

¿Existe diferencia entre los diferentes medicamentos a un nivel de 0'05?

38. En un experimento con 164 personas resfriadas, se administró un medicamento a la mitad de ellas y a la otra mitad se les dio una píldora de azúcar. Con los datos de la siguiente tabla, verificar la hipótesis de que este medicamento no es mejor que la píldora de azúcar para curar los resfriados.

	Beneficiosa	Perjudicial	Sin efecto
Fármaco	50	10	22
Azúcar	42	12	28

39. Una fábrica de automóviles quiere averiguar si el sexo de sus posibles clientes tiene relación con la preferencia de modelo. Se toma una muestra de dos mil posibles clientes y se clasifican así:

Sexo / Modelo	A	B	C
Mujer	340	400	260
Varón	350	270	380

¿Se puede decir que el sexo influye en el modelo elegido a un nivel $\alpha = 0'01$?

40. Una zapatería se abastece de cuatro fabricantes. Cada zapato es inspeccionado antes de ponerlo en venta. Hay tres defectos diferentes que causarían la devolución al fabricante. En una muestra se encontraron los siguientes defectos:

Fabricante / Defecto	I	II	III
A	17	10	13
B	10	10	10
C	18	15	17
D	15	5	10

¿Se puede decir que los defectos son independientes del fabricante a un nivel $\alpha = 0'01$?

41. En dos ciudades A y B , se observó el color del pelo y de los ojos de sus habitantes, encontrándose las siguientes tablas:

Ojos / Pelo	Rubio	No rubio
Azul	47	23
No Azul	31	93

Ojos / Pelo	Rubio	No rubio
Azul	54	30
No azul	42	80

Se pide:

- Hallar los coeficientes de contingencia de las dos ciudades.
- ¿En cuál de las dos ciudades podemos afirmar que hay mayor dependencia entre el color del pelo y de los ojos?