Bioestadística – Grado en Biología

Curso 2022-2023

Tema 4. Estimación e intervalos de confianza

Profesores: Mª Ángeles Casares de Cal, Fernando Castro Prado, Laura Davila Pena y Pedro Faraldo Roca

Índice

1 Introducción 1

2 Conceptos básicos 1

3 Planteamiento general del problema de inferencia paramétrica 2

4 Estimación puntual de una proporción. Exactitud y precisión de un estimador 2

5 Concepto de intervalo de confianza. Intervalo de confianza para una proporción 4

6 Estimación puntual de la media y la varianza de una población normal 6

6.1 Estimación de la media . . . 7

6.2 Estimación de la varianza cuando la media es conocida . . . 7

6.3 Estimación de la varianza cuando la media es desconocida . . . 7

7 Intervalos de confianza para la media y la varianza de una población normal 8

7.1 Media con varianza conocida . . . 8

7.2 Media con varianza desconocida . . . 8

7.3 Varianza con media conocida . . . 11

7.4 Varianza con media desconocida . . . 11

8 Ejercicios 15

1 Introducción

En el tema 1 hemos estudiado la Estadística Descriptiva, que se dedica al análisis y tratamiento de datos. A partir de ellos, resume, ordena y extrae los aspectos más relevantes de la información que contienen. Sin embargo, los objetivos de la Estadística son más ambiciosos. No nos conformamos con describir unos datos contenidos en una muestra sino que pretendemos sacar conclusiones para la población de la que fueron extraídos. A esta última tarea la llamamos Inferencia Estadística.

Obtendremos las muestras de forma aleatoria y por tanto necesitaremos la Teoría de la Probabilidad para elaborar nuestros argumentos.

En las secciones 2 y 3 se tratan las ideas generales de inferencia. En las secciones 4 y 5 se enfoca el objetivo concreto de este tema, que es la estimación de parámetros y los intervalos de confianza. Para introducir los conceptos se centra en el problema de estimación e intervalo de confianza para una proporción. En las secciones 6 y 7 se aplican estos principios para la estimación e intervalos de confianza relacionadas con la media y la varianza de una población normal.

2 Conceptos básicos

Población. Es el conjunto de individuos que queremos estudiar. Sirva como ejemplo la población de robles de Galicia cuya tasa de supervivencia a los incendios nos interesa conocer. En otros casos (como por ejemplo, al estudiar la probabilidad de tener una suma de diez al lanzar dos dados, o la probabilidad de que se desarrollen microorganimos en un alimento sometido a ciertas condiciones ambientales), se trata de experimentos que pueden producir ciertos resultados. En estas circunstancias no está tan clara la existencia de una población, entendida como conjunto de individuos. En ocasiones se considera como una población infinita, pero nosotros lo llamaremos **patrón probabilístico**. En cualquier caso, el objetivo de la Inferencia Estadística es obtener información sobre una población o un patrón probabilístico, y en adelante usaremos el término población para referirnos indistintamente a uno u otro concepto.

Muestra. Es un subconjunto extraído de la población, al cual podemos observar. Típicamente, múltiples razones nos imposibilitan observar toda la población. Por ese motivo, extraemos una muestra y con ella obtenemos información sobre toda la población. En el caso del patrón probabilístico, la muestra estaría constituída por unas cuantas realizaciones del experimento.

Censo. Es una muestra formada por toda la población, esto es, analizamos a todos y cada uno de los individuos. Es una situación extrema que no trataremos.

Tamaño de la población o de la muestra. Es el número de individuos que los forman, en cada caso.

Cabe hacer una primera distinción, al hablar de Inferencia, según la naturaleza del problema que se plantee:

1. **Inferencia paramétrica:** cuando se conoce la forma de la distribución de probabilidad e interesa averiguar el parámetro o parámetros de los que depende. Por ejemplo, sabemos que la población es Normal e interesa conocer la media μ y la desviación típica σ. A su vez, dentro de la Inferencia Paramétrica vamos a distinguir distintos problemas:

   (a) **Estimación Puntual**. Consiste en aventurar un valor, calculado a partir de la muestra, que esté lo más próximo posible al verdadero parámetro. Por ejemplo, la media muestral puede ser un estimador razonable de la media poblacional y la proporción muestral de la proporción poblacional.

   (b) **Intervalos de Confianza**. Dado que la estimación puntual conlleva un cierto error, construímos un intervalo que con alta probabilidad contenga al parámetro. La amplitud del intervalo nos da idea del margen de error de nuestra estimación.

   (c) **Contrastes de Hipótesis**. Se trata de responder a preguntas muy concretas sobre la población, y se reducen a un problema de decisión sobre la veracidad de ciertas hipótesis. Por ejemplo, nos podemos preguntar si la tasa de supervivencia de los robles (Quercus robur) en zonas incendiadas es inferior al 50%, lo cual justificaría una repoblación de la zona.

2. **Inferencia no Paramétrica:** cuando no se sabe la forma de la distribución poblacional. También se pueden plantear las tareas de estimación, intervalos de confianza y contrastes de hipótesis, aunque las técnicas estadísticas son diferentes.

3 Planteamiento general del problema de inferencia paramétrica

Consideramos un experimento aleatorio sobre el cual medimos una cierta variable aleatoria, que denotaremos por X. Suponemos que la distribución de X, aún siendo desconocida, sigue un modelo como los del tema anterior. Y nos interesa conocer su distribución de probabilidad.
Por ejemplo, provocamos una reacción química y medimos el calor que se desprende X. Nos interesa saber qué valores puede tomar y con qué probabilidades, esto es, su distribución.
Otro ejemplo podría consistir en averiguar la proporción de individuos con cierta característica en una población. En este caso, el experimento consistiría en extraer un individuo al azar, y la variable de interés sería la presencia o ausencia de dicha característica, cuya distribución también queremos conocer.

En el caso del calor desprendido en la reacción, esta variable podría ser normal, y en el caso de la proporción, sería de Bernoulli. En ambos casos, el problema se reduce a averiguar los parámetros (μ y σ para una distribución normal y p para la distribución de Bernoulli).

Para hacer inferencia (en nuestro caso, averiguar los parámetros que desconocemos), repetimos el experimento n veces *en idénticas condiciones y de forma independiente*.
Lo que tendremos, entonces, será una **muestra aleatoria simple** de tamaño n, es decir, n variables

$$X_1, X_2, \cdots, X_n$$

independientes y con la misma distribución que X.

Llamamos **realización muestral** a los valores concretos que toman las n variables aleatorias *después* de la obtención de la muestra.

Emplearemos el término **estadístico** para referirnos a cualquier variable obtenida realizando cierta operación sobre la muestra. Puede ser la suma de las observaciones, la media de la muestra, la varianza de la muestra o cualquier otro cálculo (el valor más frecuente de la muestra, ...). Por ser la muestra aleatoria, el estadístico también es una variable aleatoria y por ello tendrá una cierta distribución, que se denomina **distribución del estadístico en el muestreo**.

Para resolver el problema de estimación puntual, esto es, para aventurar un valor del parámetro poblacional desconocido, escogemos el valor que ha tomado un estadístico calculado sobre nuestra realización muestral. Al estadístico escogido para tal fin le llamamos **estimador** del parámetro. Al valor obtenido con una realización muestral concreta se le llama **estimación**.

Indudablemente el problema ahora radica en elegir un "buen estimador", es decir, una función de la muestra con buenas propiedades.

Es intuitivo y se puede demostrar matemáticamente que, en general, un buen estimador de un parámetro poblacional (la media de una variable en la población en estudio, la proporción de individuos de la población que presentan cierta característica, ...) va a ser el correspondiente parámetro muestral (la media de la muestra, la proporción de individuos que presentan la característica en la muestra,...).

4 Estimación puntual de una proporción. Exactitud y precisión de un estimador

Veamos ahora nuestro primer problema de inferencia paramétrica. Consiste en obtener información sobre la proporción p de individuos con cierta característica en una población (entendida como conjunto de individuos). Un problema semejante es obtener información sobre la probabilidad p de ocurrencia de un suceso, cuando realizamos un experimento aleatorio.

Como ya dijimos en la sección anterior, para hacer inferencia, repetimos el experimento n veces en idénticas condiciones y de manera independiente. La muestra está formada, entonces, por n variables

$X_1, \ldots, X_n$ independientes y con distribución de Bernoulli(p). El parámetro p es la proporción poblacional desconocida.
Como ya habíamos comentado, el estimador razonable es la proporción muestral:

$$\widehat{p} = \frac{\text{número de individuos con la característica en la muestra}}{n} = \frac{X_1 + \cdots + X_n}{n}$$

Observemos que en el numerador tenemos la suma de n variables aleatorias de Bernouilli, y ya hemos visto que eso corresponde a una variable aleatoria con distribución Binomial $B(n, p)$.
Por lo tanto, tenemos que $\widehat{p}$ es una variable aleatoria que toma los valores:
$\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$
con probabilidades que son las mismas que las de la distribución Binomial $B(n, p)$.

Estas probabilidades no las vamos a calcular, lo importante es observar que $\widehat{p}$ podrá tomar distintos valores y lo ideal sería que de esos posibles valores de $\widehat{p}$, los más probables sean los que estén más cerca del parámetro "p" que queremos estimar. Esta información nos la da la distribución de probabilidad de $\widehat{p}$. Esto significa que la distribución de probabilidad de $\widehat{p}$ refleja su calidad como estimador.

En resumen, tenemos que la proporción poblacional p es un parámetro fijo, que en la práctica es desconocido. Por el contrario, su estimador $\widehat{p}$ es una variable aleatoria que puede tomar distintos valores con ciertas probabilidades.

Evaluando la calidad de un estimador

Ahora vamos a ver algunas de las cualidades que debe tener un estimador $\widehat{\theta}$ de un parámetro desconocido θ.

Definición 1

- *Llamamos* ***sesgo*** *de un estimador $\widehat{\theta}$ de un parámetro poblacional θ a*

$$Sesgo\left(\widehat{\theta}\right) = E\left(\widehat{\theta}\right) - \theta$$

Diremos que el estimador es ***insesgado*** *si su sesgo vale cero.*

El término sesgo es propio de la Estadística, mientras que en las ciencias experimentales se emplea el término **exactitud**. Así, un estimador es tanto más exacto cuanto menos sesgo tenga.

Con respecto a $\widehat{p}$, se puede comprobar que:

$$E(\widehat{p}) = p$$

Entonces, $\widehat{p}$ es un estimador insesgado (o exacto) de p.

Definición 2

- *Definimos el* ***error cuadrático medio*** *de un estimador $\widehat{\theta}$ para un parámetro poblacional θ como*

$$E\left[(\widehat{\theta} - \theta)^2\right] = Var(\widehat{\theta}) + \left[Sesgo(\widehat{\theta})\right]^2$$

y diremos que dicho estimador es ***consistente*** *si* $\lim_{n\to\infty} E\left[(\widehat{\theta} - \theta)^2\right] = 0$.

Esto significa que, al aumentar el tamaño de la muestra, la dispersión es más pequeña, y eso hace que el estimador se aproxime al parámetro poblacional, lo cual es una buena propiedad y constituye una justificación fundamental del método estadístico.

Entonces $\widehat{p}$ es un estimador consistente de p pues se puede comprobar que:

$$\lim_{n\to\infty} E\left[(\widehat{p} - p)^2\right] = \lim_{n\to\infty} \text{Var}\,(\widehat{p}) = \lim_{n\to\infty} \frac{p(1-p)}{n} = 0$$

Observemos que si tenemos un estimador insesgado, el interés se centra en su varianza, y por ello es importante conocer su valor (o el de su desviación típica).

La siguiente definición hace referencia precisamente a la desviación típica de un estimador.

Definición 3
*El **error típico** de un estimador $\widehat{\theta}$ para un parámetro poblacional θ es su desviación típica:*

$$\text{error típico}\left(\widehat{\theta}\right) = \text{desviación típica}\left(\widehat{\theta}\right) = \sqrt{Var(\widehat{\theta})}.$$

Es habitual presentar cada estimación acompañada de su error típico, pues sirve de orientación sobre la calidad de la estimación.
En el caso de la proporción muestral, el error típico de $\widehat{p}$ es

$$\text{error típico}\,(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

En las ciencias experimentales se emplea el concepto de **precisión**, la cual mide la dispersión del estimador, bien a través de la varianza o del error típico. Un estimador con un error típico pequeño será un estimador preciso.

El objetivo de todo problema de estimación, es encontrar estimadores con poco sesgo y poca varianza, o dicho de otro modo, estimadores exactos y precisos.

5 Concepto de intervalo de confianza. Intervalo de confianza para una proporción

La estimación puntual resulta incompleta en el siguiente sentido: ¿qué seguridad tenemos de que la estimación obtenida se aproxime al verdadero valor del parámetro? Para poder dar respuesta a esta cuestión construimos intervalos de confianza, que permiten precisar la incertidumbre existente en la estimación.

Definición 4
*Un **intervalo de confianza** es un intervalo construido a partir de la muestra y, por tanto, aleatorio, que contiene al parámetro con una cierta probabilidad, conocida como **nivel de confianza**.*

Sea θ el parámetro desconocido y L_1 y L_2 los extremos del intervalo. Se dice que $[L_1, L_2]$ tiene un nivel de confianza $1-\alpha$, siendo $\alpha \in [0,1]$, si $\mathbf{P}(L_1 \leq \theta \leq L_2) \geq 1-\alpha$.

El nivel de confianza con frecuencia se expresa en porcentaje. Así, un intervalo de confianza del 95% es un intervalo de extremos aleatorios que contiene al parámetro con una probabilidad de 0'95.

Construimos ahora un intervalo de confianza para p, y para ello nos basamos en la proporción muestral, $\widehat{p}$.

Recordemos ahora que la distribución binomial se puede aproximar por la normal cuando n es suficientemente grande, manteniendo p fija.

En el caso que nos ocupa sabemos que el parámetro p es desconocido pero está fijo y, como en cualquier problema de inferencia, el tamaño muestral n debe ser moderado o grande.

Por lo tanto, ya que $\widehat{p}$ sólo consiste en dividir a la binomial por un número real, n, entonces su distribución también se puede aproximar por la normal, con su misma media y su misma desviación típica. Es decir,

$$\frac{\widehat{p}-p}{\sqrt{\frac{p\,(1-p)}{n}}} \approx N(0,1)$$

A esta expresión, *función de la muestra y de los parámetros de la población*, le llamamos **pivote**, * y el método que usamos para construir el intervalo de confianza se llama **método del pivote**.

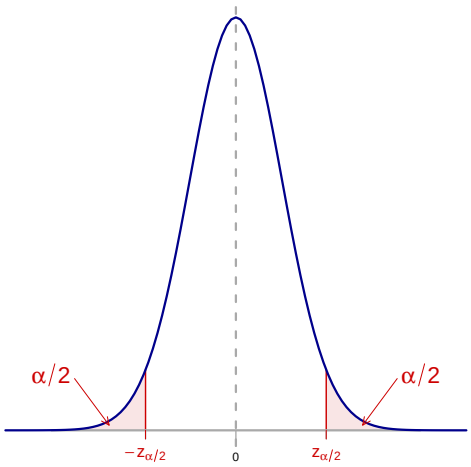
Veamos ahora cómo obtenemos el intervalo de confianza para p.
Sea

$$Z = \frac{\widehat{p}-p}{\sqrt{\frac{p\,(1-p)}{n}}}$$

*En estadística, llamamos "pivote" a una función de las observaciones y de los parámetros, de modo que su distribución de probabilidad no depende de los parámetros desconocidos.

con distribución aproximadamente Normal $N(0,1)$ cuando n es suficientemente grande.
Buscamos un valor $z_{\alpha/2}$ de Z tal que:

$$\mathbf{P}\big[\,|\,Z\,|<z_{\alpha/2}\big]=\mathbf{P}\big[-z_{\alpha/2}<Z<z_{\alpha/2}\big]=1-\alpha$$



(Este valor $z_{\alpha/2}$ se obtiene de la distribución normal estándar, siguiendo los procedimientos vistos en el tema 3.)
Tenemos, entonces:

$$\mathbf{P}\left[-z_{\alpha/2}<\frac{\widehat{p}-p}{\sqrt{\frac{p\,(1-p)}{n}}}<z_{\alpha/2}\right]=1-\alpha$$

Queremos despejar p en esta expresión, y para ello hacemos las siguientes operaciones:

$$\begin{aligned}1-\alpha &= \mathbf{P}\left[-z_{\alpha/2}<\frac{\widehat{p}-p}{\sqrt{\frac{p\,(1-p)}{n}}}<z_{\alpha/2}\right]= \\ &= \mathbf{P}\left[-z_{\alpha/2}\cdot\sqrt{\frac{p\,(1-p)}{n}}<\widehat{p}-p<z_{\alpha/2}\cdot\sqrt{\frac{p\,(1-p)}{n}}\right]= \\ &= \mathbf{P}\left[-\widehat{p}-z_{\alpha/2}\cdot\sqrt{\frac{p\,(1-p)}{n}}<-p<-\widehat{p}+z_{\alpha/2}\cdot\sqrt{\frac{p\,(1-p)}{n}}\right]=\end{aligned}$$

Por lo tanto:

$$\mathbf{P}\left(\widehat{p}-z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}<p<\widehat{p}+z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right)=1-\alpha$$

De la expresión anterior se deduce un intervalo de confianza para p con nivel de confianza $1-\alpha$, que estaría centrado en $\widehat{p}$ y tendría radio $z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}$.

Sin embargo, $\sqrt{\frac{p\,(1-p)}{n}}$, el error típico de $\widehat{p}$, no se puede calcular porque depende de p, y por tanto es desconocido. Por ello, tenemos que tomar una estimación del error típico, sustituyendo p por $\widehat{p}$, esto es:

$$\text{Error Típico}\,(\widehat{p})\simeq\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}$$

y usarla para construir el intervalo de confianza para p:

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}\ ,\ \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}\right)$$

Ejemplo 1
Para conocer el daño ocasionado por una epidemia de roya en las plantaciones de café de una determinada región se seleccionan 500 plantas y se observa que 185 tenían afectadas las hojas por dicha enfermedad. ¿Cómo puede estimarse la proporción de plantas afectadas por la enfermedad? Calcúlese un intervalo de confianza para la proporción al nivel del 95%.

Solución.

Podemos estimar la proporción de plantas atacadas por la roya, p, mediante la proporción muestral, $\widehat{p}$.

En una muestra de tamaño $n = 500$ hay 185 plantas afectadas. Por lo tanto, una estimación de p será:

$$\widehat{p} = \frac{185}{500} = 0'37$$

Para el cálculo de un intervalo de confianza para la proporción al nivel del 95%, a partir de la muestra dada, utilizaremos el pivote:

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}} \approx N(0,1) \qquad \text{(n suficientemente grande)}$$

(donde el denominador contiene una estimación del error típico).

A partir del pivote obtenemos el intervalo de confianza:

$$\left(\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}\ ,\ \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}\right)$$

Calculamos:

$$\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}} = \sqrt{\frac{0'37\cdot(1-0'37)}{500}} = 0'0216$$

Además, necesitamos conocer:

$z_{\alpha/2}$, siendo $\alpha = 0'05$

Ahora bien,

$P[Z \le z_{\alpha/2}] = P[Z \le z_{0'025}] = 0'975$, donde Z es una v.a. $N(0,1)$

(observemos que $z_{\alpha/2}$ es el cuantil $1-\alpha/2$ de la distribución normal)

Y se puede obtener de las tablas de la distribución normal, que

$z_{\alpha/2} = 1'96$

En tal caso, un intervalo de confianza para p del 95% será:

$$\big(0'37 - 1'96\cdot 0'0216\ ,\ 0'37 + 1'96\cdot 0'0216\big) = \big(0'328\ ,\ 0'412\big)$$

□

6 Estimación puntual de la media y la varianza de una población normal

Consideramos ahora el problema de inferencia paramétrica en una población normal. En esta situación disponemos de una muestra aleatoria simple

$$X_1, \ldots, X_n$$

formada por n variables aleatorias independientes y con la misma distribución $N(\mu, \sigma^2)$. El problema de inferencia consiste en averiguar los parámetros μ, media poblacional, y σ^2, varianza poblacional.

6.1 Estimación de la media

Como estimador natural para la media de la población, μ, proponemos la media de la muestra o **media muestral**:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Por la propiedad de aditividad de la distribución normal y dado que $\overline{X}$, la media muestral, es la suma de n variables independientes, entonces la media muestral tiene distribución normal y su media y desviación típica se pueden obtener por las propiedades de la media y la varianza. En conclusión llegamos a:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

De esto se deduce que la media muestral es un estimador insesgado de la media poblacional, que su varianza es la poblacional dividida por n y su error típico es la desviación típica poblacional dividida por $\sqrt{n}$. Por tanto, la dispersión será tanto mayor cuanto mayor sea la de la población y decrece tendiendo a cero cuando el tamaño muestral aumenta. De este modo vemos también que la media muestral es un estimador consistente de la media.

6.2 Estimación de la varianza cuando la media es conocida

Si la media μ es conocida entonces el estimador natural de la varianza sería la media de las desviaciones al cuadrado de los datos muestrales *respecto a la media de la población*:

$$S_\mu^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2$$

Es sencillo obtener que $E(S_\mu^2) = \sigma^2$, esto es, que tenemos un estimador insesgado de la varianza.

6.3 Estimación de la varianza cuando la media es desconocida

Si la media μ es desconocida entonces, para calcular la varianza de la muestra, debemos reemplazar la media de la población por la media de la muestra:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

El hecho de estimar la media hace que $E(S^2)$ ya no sea σ^2, sino que $E(S^2) = \dfrac{(n-1)}{n}\sigma^2$, de modo que la varianza de la muestra es un estimador sesgado, que proporciona estimaciones algo más pequeñas que la verdadera varianza que pretende estimar.

Por este motivo, se corrige la estimación de la varianza, definiendo la **cuasivarianza**, que se calcula así:

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

Lo único que la diferencia de la varianza de la muestra es la sustitución del denominador n por el denominador $(n-1)$. La cuasivarianza es, pues, un estimador alternativo de la varianza de la población. Ahora, $E(S_c^2) = \sigma^2$, esto es, la cuasivarianza es un estimador insesgado de la varianza de la población.

Observemos que aquí solo hemos calculado la media de los estimadores pero no su error típico. Para obtenerlo es necesario recurrir a argumentos de probabilidad algo más complejos y que omitimos aquí.

7 Intervalos de confianza para la media y la varianza de una población normal

7.1 Media con varianza conocida

La distribución de la media muestral permite obtener como pivote

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

y extraer de este pivote un intervalo de confianza para la media cuando la varianza es conocida, de la forma:

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

7.2 Media con varianza desconocida

Podemos construir un pivote para la media cuando la varianza es desconocida de la siguiente manera:

$$\frac{\overline{X}-\mu}{\frac{S_c}{\sqrt{n}}} \sim T_{n-1}$$

donde en el denominador hemos sustituido la varianza poblacional, que es desconocida, por la cuasi-varianza, y mediante T_{n-1} estamos denotando una nueva distribución, que es la **distribución T de Student**, con $(n-1)$ grados de libertad.

Grados de libertad de un conjunto de variables

Los grados de libertad de un conjunto de variables, están dados por el número de valores que pueden ser asignados de forma arbitraria a algunas de estas variables, antes de que el resto de las variables tomen automáticamente un valor, como consecuencia del establecimiento de relaciones entre ellas.

Por ejemplo, si tenemos dos variables independientes, X e Y, los grados de libertad son 2.
Pero si tenemos, además, la siguiente relación entre ellas: $X + Y = 2$,
entonces, los grados de libertad de este conjunto de variables es $2 - 1$ (pues existe una relación entre ellas que hace que al fijar el valor de una de las variables, automáticamente, la otra variable toma el valor establecido por esa relación).

En el caso de la T de Student:

Tenemos n variables $(X_1, X_2, \ldots, X_n)$ independientes (es decir, n grados de libertad).

También tenemos la media muestral $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Pero, además, tenemos la siguiente relación entre las variables:

$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$ *(propiedad de la media muestral, que vimos en el tema 1) y esto hace que los grados de libertad disminuyan en una unidad.*

La distribución T de Student

Omitiremos una definición rigurosa de la T de Student así como un estudio formal de esta distribución. Únicamente indicar que la distribución T de Student surge en este tipo de situaciones, en las cuales es preciso estimar el error típico de un estimador. Esto hace que el pivote siga una distribución muy parecida a la normal, como es la T de Student, pero que produce cuantiles algo más grandes, y por tanto intervalos en general más grandes. Por lo demás, la densidad de una T de Student también es simétrica en torno a cero.

Debemos mencionar también que la distribución T depende de un parámetro, que es el número de grados de libertad. Por lo tanto, para obtener los cuantiles de la T de Student tenemos que proporcionar el número de grados de libertad. En el problema de estimar la media con varianza desconocida, los grados de libertad son $(n-1)$.

Veamos ahora cómo, a partir del pivote

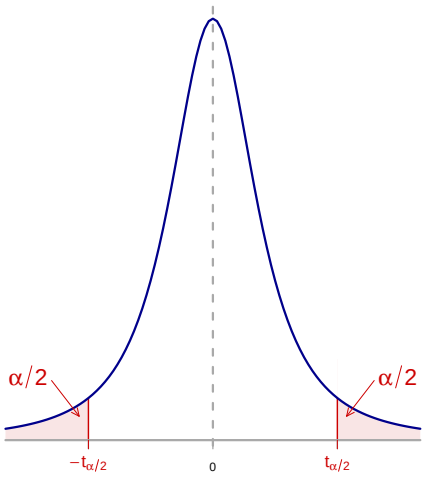
$$T = \frac{\overline{X} - \mu}{\frac{S_c}{\sqrt{n}}}$$

con distribución T de Student con $n-1$ grados de libertad,

podemos obtener un intervalo de confianza para la media (cuando la varianza es desconocida).

Buscamos un valor $t_{\alpha/2}$ de T_{n-1} tal que:

$$\mathbf{P}\big[\mid T \mid < t_{\alpha/2}\big] = \mathbf{P}\big[-t_{\alpha/2} < T < t_{\alpha/2}\big] = 1-\alpha$$



(Este valor $t_{\alpha/2}$ se obtiene de la distribución T de Student con $n-1$ grados de libertad.)

Tenemos, entonces:

$$\mathbf{P}\left[-t_{\alpha/2} < \frac{\overline{X} - \mu}{\frac{S_c}{\sqrt{n}}} < t_{\alpha/2}\right] = 1-\alpha$$

Queremos despejar μ en esta expresión, y para ello hacemos las siguientes operaciones:

$$\begin{aligned} 1-\alpha &= \mathbf{P}\left[-t_{\alpha/2} < \frac{\overline{X} - \mu}{\frac{S_c}{\sqrt{n}}} < t_{\alpha/2}\right] = \\ &= \mathbf{P}\left[-t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}} < \overline{X} - \mu < t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}}\right] = \\ &= \mathbf{P}\left[-\overline{X} - t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}} < -\mu < -\overline{X} + t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}}\right] = \end{aligned}$$

Por lo tanto:

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

Y así tenemos el intervalo de confianza para la media μ (cuando la varianza es desconocida):

$$\left(\overline{X} - t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}} \ , \ \overline{X} + t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}}\right)$$

El precio que tenemos que pagar por no conocer la varianza es que, como $t_{\alpha/2} > z_{\alpha/2}$, el intervalo de confianza para la media con varianza desconocida suele resultar más amplio que el construido con varianza conocida.

Ejemplo 2

Los datos siguientes corresponden a los pesos (en kg) de diez terneros, elegidos aleatoriamente en una explotación agropecuaria.

233 208 304 254 279 287 247 303 194 228

Suponiendo que la población es normal, calcular un intervalo de confianza del 95% para el peso medio de los terneros de esa explotación agropecuaria.

Solución.

Tenemos X: "peso", que es una variable aleatoria normal, con media μ y desviación típica σ, desconocidas.

Podemos estimar el peso medio de los terneros, μ, mediante la media muestral, $\overline{X}$.

Tenemos una muestra de tamaño $n = 10$, por lo tanto, una estimación de μ será:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}\,(233 + 208 + \cdots + 228) = \frac{2537}{10} = 253'7 \text{ kg}$$

Para el cálculo de *un* intervalo de confianza del 95% para la media poblacional, μ, a partir de la muestra dada, necesitamos conocer, además, una estimación para σ:

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 = \frac{1}{9}\left[(233 - 253'7)^2 + (208 - 253'7)^2 + \cdots + (228 - 253'7)^2\right] = 1499'567 \text{ kg}^2$$

$$S_c = \sqrt{1499'567} = 38'724 \text{ kg}$$

El pivote y su distribución son como se indica a continuación:

$$\frac{\overline{X} - \mu}{\frac{S_c}{\sqrt{n}}} \sim T_{n-1}$$

(donde el denominador contiene una estimación del error típico).

A partir del pivote obtenemos el intervalo de confianza:

$$\left(\overline{X} - t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}} \ , \ \overline{X} + t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}}\right)$$

Tenemos que calcular:

$$\frac{S_c}{\sqrt{n}} = \frac{38'724}{\sqrt{10}} = 12'2456$$

Y falta averiguar el valor $t_{\alpha/2}$ de la T de Student con 9 grados de libertad, con $\alpha = 0'05$.

Observemos que dicho valor $t_{\alpha/2}$ es el que deja una probabilidad $\alpha/2$ a su derecha en la distribución T_{n-1} (por lo tanto es el cuantil $1 - \alpha/2$ de la distribución T_{n-1}).

Consultando la tabla de la distribución T de Student con 9 grados de libertad, obtenemos:

$t_{0'025} = 2'2622$

Por lo tanto, un intervalo de confianza para μ al nivel de confianza del 95% será:

$$\left(253'7 - 2'2622 \cdot 12'2456 \ , \ 253'7 + 2'2622 \cdot 12'2456\right) \ = \ \left(225'998 \ , \ 281'401\right)$$

□

7.3 Varianza con media conocida

En este caso el pivote adopta esta forma

$$\frac{nS_\mu^2}{\sigma^2} \sim \chi_n^2$$

siendo χ_n^2 una distribución conocida como **Ji-cuadrado**, con n grados de libertad.

La distribución Ji-cuadrado

De nuevo omitimos los detalles formales sobre la distribución Ji-cuadrado. Sabremos que se trata de la distribución de una variable aleatoria continua, que sólo toma valores positivos, en el intervalo $(0, +\infty)$, y que es asimétrica. En la figura 2 se representa la función de densidad de una distribución χ_k^2, con $k = 6$ grados de libertad.

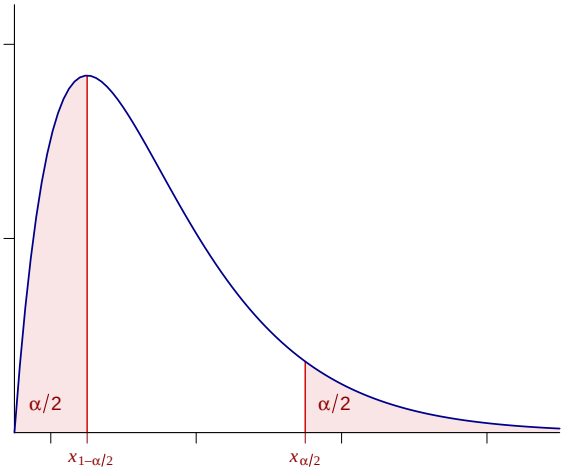


Figura 2. Función de densidad de una variable aleatoria con distribución Ji-cuadrado con m grados de libertad, χ_m^2, junto con los valores $x_{1-\alpha/2}$ y $x_{\alpha/2}$

La distribución Ji-cuadrado surge en problemas de inferencia como éste, en los que se calculan sumas de cuadrados. El número de grados de libertad está relacionado con el número de sumandos, aunque no coincide exactamente pues en ocasiones se estiman parámetros (la media, en el caso que veremos más adelante).

Finalmente, a partir del pivote se obtiene el intervalo de confianza para la varianza con media conocida así:

$$\left(\frac{nS_\mu^2}{x_{\alpha/2}} \ , \ \frac{nS_\mu^2}{x_{1-\alpha/2}} \right)$$

siendo $x_{\alpha/2}$ y $x_{1-\alpha/2}$ los valores que dejan a su derecha probabilidades $\alpha/2$ y $(1 - \alpha/2)$, respectivamente, en la distribución χ_n^2.

7.4 Varianza con media desconocida

Para obtener el intervalo de confianza para la varianza con media desconocida, nos basamos en el pivote:

$$X = \frac{(n-1)S_c^2}{\sigma^2}$$

que tiene distribución Ji-cuadrado con $n - 1$ grados de libertad $\left(\chi_{n-1}^2\right)$.

Buscamos dos valores $x_{1-\alpha/2}$ y $x_{\alpha/2}$ de χ_{n-1}^2 tales que:

$$\mathrm{P}\left[x_{1-\alpha/2} < X < x_{\alpha/2}\right] = 1 - \alpha$$

(Estos valores se obtienen de la distribución χ^2 con $n-1$ grados de libertad.)
Tenemos, entonces:

$$\mathbf{P}\left[x_{1-\alpha/2} < \frac{(n-1)S_c^2}{\sigma^2} < x_{\alpha/2}\right] = 1-\alpha$$

Queremos despejar σ^2 de esta expresión, y para ello hacemos las siguientes operaciones:

$$\begin{aligned} 1-\alpha &= \mathbf{P}\left[x_{1-\alpha/2} < \frac{(n-1)S_c^2}{\sigma^2} < x_{\alpha/2}\right] = \\ &= \mathbf{P}\left[\frac{1}{x_{1-\alpha/2}} > \frac{\sigma^2}{(n-1)S_c^2} > \frac{1}{x_{\alpha/2}}\right] = \\ &= \mathbf{P}\left[\frac{1}{x_{\alpha/2}} < \frac{\sigma^2}{(n-1)S_c^2} < \frac{1}{x_{1-\alpha/2}}\right] = \\ &= \mathbf{P}\left[\frac{1}{x_{\alpha/2}} \cdot (n-1)S_c^2 < \sigma^2 < \frac{1}{x_{1-\alpha/2}} \cdot (n-1)S_c^2\right] \end{aligned}$$

Por lo tanto:

$$\mathbf{P}\left[\frac{1}{x_{\alpha/2}} \cdot (n-1)S_c^2 < \sigma^2 < \frac{1}{x_{1-\alpha/2}} \cdot (n-1)S_c^2\right] = 1-\alpha$$

y así tenemos el intervalo de confianza para la varianza σ^2 (cuando la media es desconocida):

$$\left(\frac{(n-1)S_c^2}{x_{\alpha/2}}\ ,\ \frac{(n-1)S_c^2}{x_{1-\alpha/2}}\right)$$

siendo $x_{1-\alpha/2}$ y $x_{\alpha/2}$ los valores que dejan a su derecha probabilidades $\alpha/2$ y $(1-\alpha/2)$, respectivamente, en la distribución χ^2_{n-1}.

La diferencia respecto al caso anterior consiste en que la varianza se estima mediante S_c^2 en lugar de S_μ^2, y que los grados de libertad de la distribución Ji-cuadrado son $(n-1)$ en lugar de n, como consecuencia de haber tenido que estimar la media para el cálculo del estimador S_c^2.

Ejemplo 3

A continuación se presentan los resultados de la hemoglobina (en g/dl) en 16 animales de laboratorio expuestos a productos químicos perjudiciales:

15'6 14'8 14'4 16'6 13'8 14'0 17'3 17'4 18'6 16'2 14'7 15'7 16'4 13'9 14'8 17'5

Si la población es normal, obtener el intervalo de confianza para la varianza y la desviación típica.

Solución.

Tenemos X: "hemoglobina", que es una variable aleatoria normal, con media μ y desviación típica σ, desconocidas.

Tenemos una muestra de tamaño 16, y para poder calcular un intervalo de confianza para la varianza empezaremos obteniendo la media muestral:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{16}\left(15'6 + 14'8 + \cdots + 17'5\right) = \frac{251'7}{16} = 15'73 \text{ g/dl}$$

Este resultado es una estimación de la media de la hemoglobina.

La cuasivarianza resulta:

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{15}\left[(15'6-15'73)^2 + (14'8-15'73)^2 + \cdots + (17'5-15'73)^2\right] = 2'18 \ \ (\text{g/dl})^2$$

y entonces:

$S_c = \sqrt{2'18} = 1'48$ g/dl

lo cual es una estimación de la desviación típica de la hemoglobina.

Utilizaremos el pivote:

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

obteniendo el intervalo de confianza para la varianza:

$$\left(\frac{(n-1)S_c^2}{\chi_{\alpha/2}} \ , \ \frac{(n-1)S_c^2}{\chi_{1-\alpha/2}}\right)$$

Ahora tenemos que conocer:

$\chi_{1-\alpha/2}$ y $\chi_{\alpha/2}$ con $\alpha = 0'05$

Consultando la tabla de la distribución Ji-cuadrado obtenemos:

$\chi_{0'025} = 27'5$

$\chi_{0'975} = 6'26$

Finalmente, el intervalo de confianza del 95% para la varianza será:

$$\left(\frac{15 \cdot 2'18}{27'5} \ , \ \frac{15 \cdot 2'18}{6'26}\right) = (1'19, 5'22)$$

y el intervalo de confianza del 95% para la desviación típica:

$$\left(\sqrt{\frac{15 \cdot 2'18}{27'49}} \ , \ \sqrt{\frac{15 \cdot 2'18}{6'26}}\right) = \left(\sqrt{1'19}, \sqrt{5'22}\right) = (1'09, 2'28)$$

□

Para terminar este tema, vemos un ejemplo en el cual se calculan estimaciones e intervalos de confianza para la media y la desviación típica en el contexto más habitual, que es cuando no se conoce ni la media ni la desviación típica.

Ejemplo 4

En un estudio sobre el contenido de calcio en alimentos lácteos se obtiene una muestra aleatoria de treinta quesos y se analiza su contenido de calcio (en miligramos por gramo de queso). Se han obtenido una media muestral de 9'5 mg/g y una cuasivarianza de 0'25. Supongamos que el contenido de calcio presenta una distribución normal. Hallar un intervalo de confianza del 99% para la media μ y la desviación típica σ de la población.

Solución.

Tenemos X: "contenido de calcio", que es una variable aleatoria normal, con media μ y desviación típica σ, desconocidas.

La información que tenemos de la muestra es:

$n = 30$

$\overline{x} = 9'5$ mg/dL

$S_c^2 = 0'25$

$S_c = 0'5$ mg/dL

Por lo tanto, una estimación del error típico será:

$$\frac{S_c}{\sqrt{n}} = \frac{0'5}{\sqrt{30}} = 0'09$$

Los intervalos de confianza que nos solicitan son:

Intervalo de confianza del 99% para el contenido medio de calcio

$\left(\overline{X} - t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}} \ , \ \overline{X} + t_{\alpha/2} \cdot \frac{S_c}{\sqrt{n}}\right)$ con un nivel de confianza $1 - \alpha = 0'99$

Necesitamos conocer $t_{\alpha/2} = t_{0'005}$. Para ello empleamos la tabla de la T de Student y obtenemos:

$t_{0'005} = 2'76$

Por lo tanto un intervalo de confianza del 99% para μ será:

$$\left(9'5 - 2'76 \cdot \frac{0'5}{\sqrt{30}}\ ,\ 9'5 + 2'76 \cdot \frac{0'5}{\sqrt{30}}\right)$$
$$= \left(9'5 - 2'76 \cdot 0'09\ ,\ 9'5 + 2'76 \cdot 0'09\right) = \left(9'25\ ,\ 9'75\right)$$

Intervalo de confianza del 99% para la desviación típica del contenido de calcio

$\left(\sqrt{\frac{(n-1)S_c^2}{X_{\alpha/2}}}\ ,\ \sqrt{\frac{(n-1)S_c^2}{X_{1-\alpha/2}}}\right)$ con un nivel de confianza $1 - \alpha = 0'99$

Tenemos que conocer:

$X_{\alpha/2}$ y $X_{1-\alpha/2}$ con $\alpha = 0'01$

Y de la tabla de la distribución Ji-cuadrado se obtiene:

$X_{0'005} = 52'3$

$X_{0'995} = 13'1$

Por lo tanto, un intervalo de confianza del 99% para la varianza del contenido de calcio es:

$$\left(\frac{29 \cdot 0'25}{52'3}\ ,\ \frac{29 \cdot 0'25}{13'1}\right) = (0'139, 0'553)$$

Y un intervalo de confianza del 99% para la desviación típica del contenido de calcio es:

$$\left(\sqrt{0'139}, \sqrt{0'553}\right) = (0'373, 0'744)$$

□

8 Ejercicios

1. El diámetro o la circunferencia son medidas básicas de cualquier árbol pues sirven para conocer, por ejemplo, su volumen y crecimiento. La medida más típica del diámetro de un árbol es *el diámetro a la altura del pecho*[†], que identificamos con el símbolo *"dap"*, y se mide con la corteza y sobre el terreno. A fin de evitar una estimación excesiva del volumen y compensar los errores de medición, las medidas se ajustan en sentido decreciente, por ejemplo, 16'8 cm se convierte en 16 cm.

   En un estudio sobre el diámetro de los árboles "Pinus taeda" se ha medido el "dap" en 600 pinos de esta especie y se ha comprobado que 387 árboles tienen un diámetro mayor de 15 centímetros.

   (a) Estima la proporción de árboles con un diámetro mayor de 15 centímetros.

   (b) Obtén un intervalo de confianza del 90% para la proporción de árboles con un diámetro mayor de 15 centímetros.

2. En una región han registrado las profundidades que tuvieron que alcanzar los pozos hasta obtener agua (en metros):

   21 19 29 30 28 22 26 25 28 22

   (a) Proporciona una estimación de la media.

   (b) Suponiendo que la profundidad tiene distribución normal, construye un intervalo de confianza del 95% para la profundidad media de los pozos.

3. Se obtiene una muestra de 25 semillas del cacao (Theobroma cacao) y resulta un peso medio de dichas semillas de 38 g con una cuasidesviación típica de 3 g.

   Suponiendo que el peso tiene distribución normal, calcula un intervalo de confianza del 95% para el peso medio de los granos del cacao.

4. Los siguientes valores son contenidos de grasa (medidos en porcentaje) de diez muestras de leche de vacas Ayrshire de tres años.

   4'28 3'91 4'09 3'66 4'27 4'16 4'05 4'06 4'67 4'20

   Suponiendo que el contenido de grasa de la leche es una variable aleatoria normal, calcula un intervalo de confianza del 99% para el contenido medio de grasa de la leche de vacas Ayrshire de tres años.

5. En un estudio sobre el pH del terreno en una región, se toman ocho muestras de suelo superficial y los valores obtenidos fueron los siguientes:

   5'56 6'57 6'71 5'19 6'33 5'19 6'25 5'92

   Suponiendo que la variable pH sigue una distribución normal, construye un intervalo de confianza del 95% para la desviación típica del pH del terreno de esa región.

6. Una región devastada por los incendios se ha reforestado con robles y castaños jóvenes. Se sospecha que, debido a una grave sequía durante la siguiente estación, muchos de estos árboles recién plantados se han secado. Para comprobarlo se obtuvo una muestra de 1000 árboles, de los cuales 300 estaban secos.

   Utiliza esta información para estimar la proporción de árboles recién plantados y secos en la población, y obtén un intervalo de confianza del 95% para esta proporción.

[†] El diámetro del árbol a la altura de 1'30 m.