Muestreo y estimación

BENITO J. GONZÁLEZ RODRÍGUEZ (bjglez@ull.es)
DOMINGO HERNÁNDEZ ABREU (dhabreu@ull.es)
MATEO M. JIMÉNEZ PAIZ (mjimenez@ull.es)
M. ISABEL MARRERO RODRÍGUEZ (imarrero@ull.es)
ALEJANDRO SANABRIA GARCÍA (asgarcia@ull.es)

Departamento de Análisis Matemático
Universidad de La Laguna

Índice

1. Introducción

En términos generales, todo estudio estadístico se basa en los siguientes aspectos:

1. Fijar la población: determinar el conjunto de individuos a los que involucra el estudio.
2. Indicar la característica a estudiar (que, en general, es una variable aleatoria).
3. Recopilar información relativa a la característica en ciertos individuos.
4. Extraer conclusiones a partir del estudio.

Ejemplo 1.1. *Son estudios estadísticos:*

i) *Estudio sobre el precio medio de la receta médica por la Seguridad Social en Santa Cruz de Tenerife.*

ii) *Estudio sobre la proporción de hogares de Tenerife con conexión a Internet de banda ancha.*

El siguiente concepto es fundamental en Estadística.

Definición 1.2. Población *es el conjunto de todos los elementos que poseen una determinada característica.*

Por razones de urgencia temporal y ahorro económico, entre otras, a la hora de recopilar información *no* suelen estudiarse todos los individuos de la población.

Definición 1.3. *Se denomina* muestra *a un subconjunto de la población, y* muestreo *al proceso mediante el cual se escoge una muestra de la población. En general, una* muestra de tamaño *n es un grupo de n individuos extraídos de la población.*

Definición 1.4. *Los estudios que involucran a toda la población se denominan* censos *de población.*

En la práctica, los estudios estadísticos se realizan a partir de la información obtenida de ciertas muestras. Las conclusiones que se *infieran* a partir del estudio de muestras pueden contener errores en relación a las conclusiones que se derivarían al estudiar la población entera. La *Inferencia Estadística* trata de la obtención de conclusiones a partir de muestras, controlando el error en dichas conclusiones por medio de técnicas probabilísticas. En general, se desea que las muestras sean lo más representativas de la población posible.

2. Tipos de muestreo

Definición 2.1. *Los muestreos pueden ser de diferentes tipos:*

i) Muestreo aleatorio simple: *es aquel en el cual se eligen al azar n individuos de la muestra; todos los individuos de la población tienen igual probabilidad de ser elegidos.*

ii) Muestreo aleatorio estratificado: *es el caso en el que la población se divide en grupos homogéneos (que presentan características similares) llamados* estratos*, y posteriormente se extrae una muestra aleatoria simple de cada uno.*

iii) Muestreo aleatorio sistemático: *se ordenan numéricamente todos los individuos de la población; se divide el tamaño de la población entre el tamaño de la muestra, resultando un cociente k; finalmente, se elige al azar un elemento de la población, y a partir de él se seleccionan de k en k todos los elementos siguientes.*

iv) Muestreo por conglomerados y áreas: *se divide la población en distintas secciones o conglomerados, es decir, subconjuntos de la población donde la variabilidad de características es similar a la de la población entera; se eligen al azar unas pocas de estas secciones, y se forma la muestra con todos los elementos de las secciones elegidas.*

Ejemplo 2.2. *Supongamos que tenemos* 100 *hogares. Elegir una muestra de* 5 *con muestreo sistemático.*

RESOLUCIÓN. Ordenamos numéricamente los hogares del 1 al 100. El cociente de dividir 100 entre 5 es 20; entonces 20 sería el período. Elegimos al azar un número entre 1 y 20, digamos 16. El hogar con el número 16 sería el primero seleccionado, y los restantes los numerados con 36, 56, 76 y 96. □

Definición 2.3. *Un* parámetro *es una cantidad numérica calculada sobre una población que resume los valores que ésta toma en algún atributo o característica (media, varianza, etc.).*

3. Distribución de distintos estadísticos en el muestreo

La selección de una muestra de una población es un experimento aleatorio. El espacio muestral de este experimento está constituido por todas las posibles muestras del tamaño considerado obtenidas de la población.

Definición 3.1. *Un* estadístico *es una variable aleatoria que asigna un valor numérico a cada muestra. La distribución de esta variable aleatoria se denomina* distribución muestral *del estadístico.*

3.1. Media muestral

Definición 3.2. *Dada una muestra aleatoria $X_1, X_2, ..., X_n$ de tamaño n, la* media muestral *es el estadístico obtenido tomando la media aritmética de los elementos de la muestra. La denotaremos mediante $\overline{X}$:*

$$\overline{X} = \frac{1}{n}\sum_k X_k.$$

Si la variable aleatoria en estudio sigue una distribución normal $N(\mu, \sigma)$ entonces la media muestral $\overline{X}$ sigue una distribución normal $N(\mu, \sigma/\sqrt{n})$, donde n es el tamaño de la muestra. Por otra parte:

Teorema 3.3 (Teorema del Límite Central)**.** *Si el tamaño de la muestra es suficientemente grande $(n \geq 30)$ entonces, para casi todas las poblaciones, la media muestral $\overline{X}$ sigue aproximadamente una distribución normal.*

Luego:

- Si la población de partida es normal, la distribución de las medias muestrales también es normal, cualquiera que sea n.
- Si la población de partida no es normal, la distribución de las medias muestrales es aproximadamente normal cuando $n \geq 30$.

Ejemplo 3.4. *El tiempo que tarda un cajero automático en atender a los clientes es de una media de 3 minutos, con desviación típica de 1.2 minutos. Se observa una muestra de 50 personas. ¿Cuál es la probabilidad de que el tiempo medio de espera supere los 2 minutos?*

RESOLUCIÓN. Sea $X =$'tiempo de espera en el cajero'. Se tiene que $\mu = 3$, $\sigma = 1.2$ y $n = 50$ clientes.

Aunque desconocemos la distribución de la variable aleatoria X, ya que $n \geq 30$ podemos considerar que la variable aleatoria $\overline{X} =$'tiempo medio de espera' sigue una distribución normal

$$N\left(3, \frac{1.2}{\sqrt{50}}\right) = N(3, 0.17).$$

Entonces:

$$P(\overline{X} > 2) = P\left(Z > \frac{2-3}{0.17}\right) = P(Z > -5.88) = P(Z < 5.88) = 1.$$

Esto es, el tiempo medio de espera superará, con casi total seguridad, los 2 minutos. □

3.2. Proporción muestral

Definición 3.5. *Se considera una población de la que se extraen muestras de tamaño $n \geq 30$ y de la que se conoce que la proporción de individuos que presentan una determinada característica es igual a p. La variable aleatoria $\widehat{p}$ de las* proporciones muestrales *es la proporción de individuos de cada muestra que presentan la característica estudiada. Se define como $\widehat{p} = X/n$, donde X es el número de éxitos y n el tamaño de la muestra.*

Se tiene que $\widehat{p}$ sigue una distribución normal

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

Ejemplo 3.6. *Se sabe que el 40 % de los estudiantes de Bachillerato de la provincia de Santa Cruz de Tenerife son aficionados al voley playa femenino. Se elige una muestra de 200 estudiantes. Hallar la probabilidad de que el porcentaje de aficionados de dicha muestra oscile entre el 35 % y el 45 %.*

RESOLUCIÓN. Se tiene que $p = 0.4$ (proporción poblacional) y $n = 200$ (tamaño muestral). Se sigue que:

$$\widehat{p} \sim N\left(0.4, \sqrt{\frac{0.4 \cdot 0.6}{200}}\right) = N(0.4, 0.0346).$$

De aquí:

$$P(0.35 < \widehat{p} < 0.45) = P\left(\frac{0.35-0.4}{0.0346} < Z < \frac{0.45-0.4}{0.0346}\right) = P(-1.45 < Z < 1.45)$$
$$= 2 \cdot 0.9265 - 1 = 0.8530.$$

□

Ejemplo 3.7. *El* 3 *% de las piezas producidas por una máquina son defectuosas. Se toman muestras de* 100 *piezas.*

a) *¿Cuál es la distribución de la proporción de piezas defectuosas en la muestra?*

b) *Hallar la probabilidad de que en una muestra de* 100 *piezas haya menos de* 5 *defectuosas.*

RESOLUCIÓN. *a*) Conforme a lo indicado, $\widehat{p}$ sigue una distribución normal

$$N\left(0.03, \sqrt{\frac{0.03 \cdot 0.97}{100}}\right) \simeq N(0.03, 0.017).$$

b) Por tanto:

$$P\left(\widehat{p} < \frac{5}{100}\right) = P(\widehat{p} < 0.05) = P\left(Z < \frac{0.05 - 0.03}{0.017}\right) = P(Z < 1.18) = 0.8810.$$

□

3.3. Suma muestral

La suma muestral es otro estadístico de interés en determinados estudios. Se trata de estimar la suma de un cierto número de elementos de la población mediante el estudio de la suma de una muestra de ese número de individuos.

Definición 3.8. *Dada una muestra aleatoria* $X_1, X_2, ..., X_n$*, el estadístico* suma muestral *se define como* $T = \sum_k X_k$.

La variable T tiene media $n\mu$ y desviación típica $\sigma\sqrt{n}$, donde μ es la media poblacional, σ la desviación típica poblacional y n el tamaño de la muestra. Si la población es normal, también lo es T. En general, a medida que n crece, la distribución de T se aproxima a una normal $N(n\mu, \sigma\sqrt{n})$.

Ejemplo 3.9. *Se lanza una moneda al aire* 100 *veces; si sale cara le damos el valor* 1*, y si sale cruz, el valor* 0*. Cada lanzamiento es una variable aleatoria independiente que se distribuye según el modelo de Bernoulli, con media* 0.5 *y varianza* 0.25*. Calcular la probabilidad de que en estos* 100 *lanzamientos salgan más de* 60 *caras.*

RESOLUCIÓN. Como $n = 100$ es grande, podemos suponer que la variable aleatoria $T =$'número de caras en 100 lanzamientos', que es la suma muestral de las variables independientes $X_i =$'número de caras en el lanzamiento i-ésimo' $(1 \leq i \leq 100)$, sigue una distribución normal $N(100 \cdot 0.5, \sqrt{100} \cdot 0.25) = N(50, 2.5)$. Por consiguiente:

$$P(T > 60) = P\left(Z > \frac{60-50}{5}\right) = P(Z > 2) = 1 - P(Z < 2) = 1 - 0.9772 = 0.0228.$$

Es decir, la probabilidad de que al tirar 100 veces la moneda salgan más de 60 caras es tan sólo del 2.28 %. □

Ejemplo 3.10. *Un cierto tipo de bombilla eléctrica tiene una duración media de* 1500 *horas, con una desviación típica de* 150 *horas. Se conectan* 3 *bombillas de forma que cuando una se funde, otra sigue alumbrando. Suponiendo que las duraciones se distribuyen normalmente, ¿cuál es la probabilidad de que se tenga luz:*

a) *al menos* 5000 *horas;*

b) *como mucho* 4200 *horas?*

RESOLUCIÓN. En este ejemplo, la variable aleatoria de interés, $T =$'tiempo de iluminación de las 3 bombillas', es suma muestral de las variables aleatorias independientes $X_i =$'tiempo de iluminación de la bombilla i-ésima' $(i = 1, 2, 3)$. Dado que éstas presentan una distribución normal, podemos afirmar que T sigue una distribución normal $N(4500, 150\sqrt{3})$, de lo cual deducimos que

$$P(T > 5000) = P\left(Z > \frac{5000-4500}{150\sqrt{3}}\right) = P(Z > 1.92) = 1 - P(Z < 1.92) = 1 - 0.9726 = 0.0274.$$

Esto responde a *a*). En cuanto a *b*):

$$\begin{aligned} P(T < 4200) &= P\left(Z < \frac{4200-4500}{150\sqrt{3}}\right) = P(Z < -1.15) \\ &= P(Z > 1.15) = 1 - P(Z < 1.15) = 1 - 0.8749 = 0.1251. \end{aligned}$$

□

3.4. Diferencia de medias

Definición 3.11. *La* diferencia de medias *es el estadístico $\overline{X}_1 - \overline{X}_2$, donde las variables $\overline{X}_1$ y $\overline{X}_2$ representan las medias de sendas muestras aleatorias de tamaños n_1 y n_2, respectivamente, seleccionadas de dos poblaciones diferentes y de manera independiente.*

El estadístico $\overline{X}_1 - \overline{X}_2$ sigue una distribución cuya media es $\mu_1 - \mu_2$, con desviación típica

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

A medida que n_1 y n_2 crecen, la distribución de $\overline{X}_1 - \overline{X}_2$ se aproxima a la normal. Si las desviaciones típicas σ_1 y σ_2 son desconocidas se sustituyen por las desviaciones típicas muestrales s_1 y s_2.

Ejemplo 3.12. *En un estudio para comparar los pesos promedio de niños y niñas de sexto grado en una escuela de primaria se usará una muestra aleatoria de 20 niños y otra de 25 niñas. Se sabe que tanto para niños como para niñas los pesos siguen una distribución normal. El promedio de los pesos de todos los niños de sexto grado de esa escuela es de 45 kilogramos y su desviación estándar es de 6.41 kilogramos, mientras que el promedio de los pesos de todas las niñas del sexto grado de esa escuela es de 38.5 kilogramos y su desviación estándar es de 5.55 kilogramos. Si $\overline{X}_1$ representa el promedio de los pesos de una muestra de 20 niños y $\overline{X}_2$ es el promedio de los pesos de una muestra de 25 niñas, encontrar la probabilidad de que el promedio de los pesos de los 20 niños sea, al menos, 9 kilogramos mayor que el de las 25 niñas.*

RESOLUCIÓN. Conforme a los datos del problema, tenemos:

$$\begin{aligned} \mu_1 &= 45 \text{ kg}, & \mu_2 &= 38.5 \text{ kg}, \\ \sigma_1 &= 6.41 \text{ kg}, & \sigma_2 &= 5.55 \text{ kg}, \\ n_1 &= 20 \text{ niños}, & n_2 &= 25 \text{ niñas}. \end{aligned}$$

Sabemos que $\overline{X}_1 - \overline{X}_2$ sigue una distribución normal

$$N\left(25-20, \sqrt{\frac{5.55^2}{25} + \frac{6.41^2}{20}}\right) = N(5, 1.81).$$

Así:

$$P(\overline{X}_1 - \overline{X}_2 > 9) = P\left(Z > \frac{9-(45-38.5)}{\sqrt{\frac{5.55^2}{25}+\frac{6.41^2}{20}}}\right) = P\left(Z > \frac{2.5}{1.81}\right)$$
$$= P(Z > 1.38) = 1 - P(Z < 1.38) = 1 - 0.9162 = 0.0838.$$

□

Ejemplo 3.13. *Un laboratorio farmacéutico fabrica unos comprimidos para la angina de pecho cuya fecha de caducidad han estimado que tiene una media de* 18 *meses con una desviación típica de* 3 *meses. A fin de ampliar el plazo de caducidad han cambiado el sistema de elaboración de estos comprimidos, estimando que con el nuevo método se puede lograr una media de* 24 *meses y una desviación típica de* 3 *meses en la caducidad. Se toma una muestra de* 100 *comprimidos fabricados por el sistema tradicional y* 150 *comprimidos fabricados por el nuevo. Determinar la probabilidad de que la diferencia de medias entre ambas muestras se encuentre entre* 5.5 *y* 6.5 *meses.*

RESOLUCIÓN. En este caso tenemos los siguientes datos:

$\mu_1 = 18$ meses, $\mu_2 = 24$ meses, $\sigma_1 = \sigma_2 = 3$ meses, $n_1 = 100$ comprimidos, $n_2 = 150$ comprimidos.

Sabemos que $\overline{X}_2 - \overline{X}_1$ sigue una distribución normal

$$N\left(24-18, \sqrt{\frac{3^2}{100}+\frac{3^2}{150}}\right) = N(6, 0.39),$$

con lo que

$$P(5.5 < \overline{X}_2 - \overline{X}_1 < 6.5) = P\left(\frac{5.5-6}{0.39} < Z < \frac{6.5-6}{0.39}\right) = P(-1.28 < Z < 1.28)$$
$$= 2\,(P(Z < 1.28) - 0.5) = 2(0.8997 - 0.5) = 0.7994.$$

□

4. Intervalos de confianza

Un parámetro desconocido se puede estimar mediante un valor específico de un estadístico que provenga de alguna muestra aleatoria. A dichos valor y estadístico se les conoce como *estimación puntual* y *estimador puntual* del parámetro, respectivamente. Por ejemplo, un estimador puntual para la media poblacional μ es la media muestral $\overline{X}$, mientras que una estimación puntual para μ será el valor $\bar{x}$ que tome $\overline{X}$ en una muestra aleatoria concreta.

En la práctica es preferible estimar un parámetro mediante un intervalo que con un valor particular del estimador puntual. Así, en muchos procesos de producción sujetos a un control de calidad se establecen intervalos dentro de los cuales los artículos, productos, objetos o medidas se consideran aceptables para salir al mercado.

Definición 4.1. *Un* intervalo de confianza *para un parámetro θ es un conjunto de valores numéricos $IC = (a,b)$ tal que $\theta \in (a,b)$ con una determinada probabilidad, que se denota por $1-\alpha$ y se denomina* nivel de confianza*. El número α se llama* nivel de significación*.*

4.1. Intervalo para la media μ de una población normal $N(\mu,\sigma)$, con desviación típica σ conocida

Supongamos una población normal $N(\mu,\sigma)$ con μ desconocida y σ conocida. Se extrae una muestra de tamaño n y se calcula la media muestral $\overline{X}$ (que, como sabemos, es un estimador puntual para μ). El intervalo de confianza para μ con σ conocida es

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

donde $Z_{\alpha/2}$ es el valor para el cual $P(Z < Z_{\alpha/2}) = 1-\alpha/2$.

Conocidas α, n, $\overline{X}$ y σ, se determina el intervalo de confianza. La media poblacional μ pertenecerá a dicho intervalo con probabilidad $1-\alpha$. El error máximo admisible es $Z_{\alpha/2}\,\sigma/\sqrt{n}$, esto es, la semiamplitud del intervalo.

Ejemplo 4.2. *En un hospital se sabe que la estatura de los recién nacidos se distribuye normalmente, con desviación típica de* 8.9 *centímetros. En una muestra de* 10 *bebés recién nacidos se obtuvieron las siguientes medidas en centímetros:*

$$44, \quad 68, \quad 57, \quad 48, \quad 66, \quad 47, \quad 60, \quad 53, \quad 51, \quad 68.$$

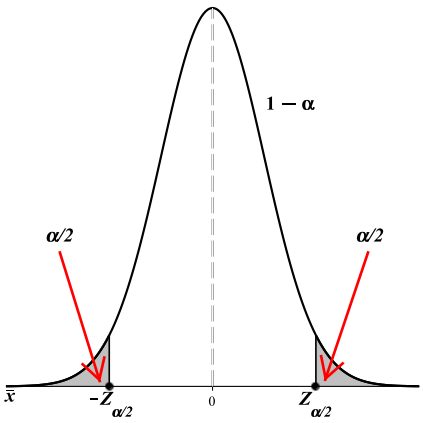

Figura 4.1. Intervalo de confianza para la media muestral.

Encontrar un intervalo de confianza del 90 % *para el peso medio de los recién nacidos.*

RESOLUCIÓN. Tenemos que $1-\alpha = 0.9$, esto es, $\alpha = 0.1$, de donde $Z_{\alpha/2} = Z_{0.05} = 1.645$. Ahora bien, como $\sigma = 8.9$, $n = 10$ y $\overline{X} = 56.2$, sigue que

$$IC = 56.2 \pm 1.645 \cdot \frac{8.9}{\sqrt{10}} = 56.2 \pm 4.6297 = (51.5703, 60.8297),$$

y $\mu \in IC$ al 90 % de confianza (con probabilidad $p \geq 0.9$). □

4.2. Intervalo para la proporción p de una población

En una población se estudia una característica, y se quiere conocer la proporción p de individuos que poseen dicha característica. Se toma una muestra de tamaño $n \geq 30$ y se halla la proporción muestral $\widehat{p}$. La proporción muestral $\widehat{p}$ es una estimación puntual de la proporción poblacional p.

El intervalo de confianza para la proporción poblacional p es

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

De nuevo, $Z_{\alpha/2}$ es el valor para el cual $P(Z < Z_{\alpha/2}) = 1 - \alpha/2$.

La proporción poblacional p pertenecerá a dicho intervalo con probabilidad $1-\alpha$. El error máximo admisible será

$$Z_{\alpha/2} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},$$

esto es, la semiamplitud del intervalo.

Ejemplo 4.3. *Un experimento en un hospital consiste en comprobar si una madre puede distinguir el llanto de su hijo del llanto de los otros niños. Se toma una muestra de* 50 *madres y se observa que* 47 *de ellas distinguen el llanto. Hallar un intervalo de confianza al* 90 % *para la proporción de madres que distinguen el llanto.*

RESOLUCIÓN. Con estos datos podemos afirmar que $1-\alpha = 0.9$ ó, lo que es lo mismo, $\alpha = 0.1$, de donde $Z_{\alpha/2} = Z_{0.05} = 1.645$. Además, $\widehat{p} = 47/50 = 0.94$ y $1-\widehat{p} = 0.06$. Como $n = 50$, tenemos:

$$IC = 0.94 \pm 1.645\sqrt{\frac{0.94\cdot 0.06}{50}} = 0.94 \pm 0.0553 = (0.8847, 0.9953).$$

Así, es posible asegurar al 90 % que entre el 88.47 % y el 99.53 % de las madres reconoce el llanto de su hijo. □

4.3. Determinación del tamaño muestral en la estimación del error

En la construcción de los intervalos de confianza que hemos estudiado se comete un error máximo al aproximar el parámetro igual a

$$E_{\text{máx}} = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{ó} \quad E_{\text{máx}} = Z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Si el tamaño n de la muestra aumenta, entonces el error $E_{\text{máx}}$ tiende a cero. En ocasiones se pide hallar n para que el error máximo sea menor que un cierto umbral E:

$$n \geq \left(\frac{Z_{\alpha/2}\,\sigma}{E}\right)^2 \quad (\mu \text{ desconocida y } \sigma \text{ conocida}),$$

o bien

$$n \geq \left(\frac{Z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})}}{E}\right)^2 \quad (\text{proporción } p \text{ desconocida}).$$

Ejemplo 4.4. *El director de una sucursal desea estimar el tiempo medio de atención a los clientes con una confianza del* 99 % *y con un error máximo de medio minuto. Se sabe que el tiempo medio de atención a los clientes se distribuye normalmente, con desviación típica* 2.6 *minutos. ¿Cuántas personas se deben incluir en el estudio para obtener dicha estimación?*

RESOLUCIÓN. Procediendo como hemos indicado, $\alpha = 0.01$ y

$$n \geq \left(\frac{Z_{\alpha/2}\,\sigma}{E} \right)^2 = \left(\frac{2.575 \cdot 2.6}{0.5} \right)^2 = 179.2921.$$

Por tanto, tomamos $n = 180$ como tamaño de la muestra. □

Ejemplo 4.5. *Se sabe por estudios previos que la proporción de objetos defectuosos en una línea de producción es del orden de* 0.05. *¿De qué tamaño conviene tomar una muestra para tener una confianza del* 95 % *de que la proporción estimada no difiera de la verdadera en más de un* 3 %*?*

RESOLUCIÓN. Igual que en el caso anterior se tiene, con $\alpha = 0.05$:

$$n \geq \left(\frac{Z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})}}{E} \right)^2 = \left(\frac{1.96\sqrt{0.05 \cdot 0.95}}{0.03} \right)^2 = 202.7511.$$

Así pues, tomamos el valor $n = 203$. □

Observación 4.6. *A veces,* $\widehat{p}$ *es desconocida. En tal caso se sustituye* $\sqrt{\widehat{p}(1-\widehat{p})}$ *por su valor máximo, que es* 0.5.

Ejemplo 4.7. *Para estimar la proporción de hogares de una población que tienen ordenador se utiliza una muestra de tamaño n.*

a) ¿Cuál debe ser el mínimo valor de n para garantizar con una confianza del 95 % *que el error en la estimación no sea superior al* 2 %*?*

b) ¿Y si se desea una confianza del 98 % *y un error máximo del* 1 %*?*

RESOLUCIÓN. *a*) Como $\alpha = 0.05$, tenemos que $Z_{\alpha/2} = 1.96$. Si $E_{\text{máx}} = 0.02$ y $\widehat{p}$ es desconocido entonces, en lugar de tomar

$$n \geq \left(\frac{Z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})}}{E_{\text{máx}}} \right)^2,$$

pondríamos

$$n \geq \left(\frac{1.96 \cdot 0.5}{0.02} \right)^2 = 2401,$$

y habría que elegir 2.401 personas.

b) Ahora $\alpha = 0.02$, por lo que $Z_{\alpha/2} = 2.33$. Si el error máximo ha de ser de 0.01 con $\widehat{p}$ desconocido, entonces

$$n \geq \left(\frac{2.33 \cdot 0.5}{0.01}\right)^2 = 13572.23,$$

de donde $n = 13573$, como mínimo. □

4.4. Intervalo de confianza para la suma muestral

Supongamos una población normal que tiene media μ y desviación típica σ. Se estudia una muestra de tamaño n y se quiere determinar un intervalo de confianza para la suma de los elementos de la muestra. En este caso tendremos:

$$IC = (X_1 + X_2 + \ldots + X_n) \pm Z_{\alpha/2}\, \sigma \sqrt{n}.$$

Ejemplo 4.8. *El voltaje medio de las baterías producidas por una compañía es de* 45.1 *voltios y la desviación típica* 0.04 *voltios. Si se conectan cuatro baterías en serie, hallar los intervalos de confianza del* 99 % *para el voltaje total.*

RESOLUCIÓN. Sean X_1, X_2, X_3 y X_4 los voltajes de las cuatro baterías. Como $\mu = 45.1$, $\sigma = 0.04$, $n = 4$, $\alpha = 0.01$ y $Z_{\alpha/2} = 1.96$, se tiene:

$$IC = 4 \cdot 45.1 \pm 1.96 \cdot 0.04\sqrt{4} = 180 \pm 0.1568 = (180.2432, 180.5568).$$

□

4.5. Intervalo de confianza para la diferencia de medias

Supongamos dos poblaciones $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$; de cada una de ellas extraemos una muestra de tamaños n_1 y n_2, respectivamente. Sean $\overline{X}_1$ y $\overline{X}_2$ las medias muestrales respectivas, y $1 - \alpha$ el nivel de confianza.

- Si σ_1 y σ_2 son conocidas, el intervalo de confianza viene dado por

$$IC = (\overline{X}_1 - \overline{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

- Si σ_1 y σ_2 son desconocidas y n_1 y n_2 son grandes (mayores o iguales que 30), el intervalo de confianza viene dado por

$$IC = (\overline{X}_1 - \overline{X}_2) \pm Z_{\alpha/2} \sqrt{\frac{\widehat{s}_1^2}{n_1} + \frac{\widehat{s}_2^2}{n_2}},$$

donde $\widehat{s}_1^2$ y $\widehat{s}_2^2$ son las cuasivarianzas de cada muestra.

Definición 4.9. *La* cuasivarianza muestral *es*

$$\widehat{s}^2 = \frac{n}{n-1} s^2,$$

donde n es el tamaño de la muestra y s la varianza muestral.

Ejemplo 4.10. *Hallar el intervalo de confianza al nivel del* 90% *para la diferencia de los salarios medios de los trabajadores y trabajadoras de una gran empresa en la que se han elegido dos muestras: una de* 40 *hombres y otra de* 35 *mujeres, cuyos salarios medios son* $\overline{X}_1 = 1051$ *euros y* $\overline{X}_2 = 1009$ *euros, sabiendo además que las desviaciones típicas son* $\sigma_1 = 90$ *euros y* $\sigma_2 = 78$ *euros.*

RESOLUCIÓN. Los datos que conocemos son:

$$\begin{aligned} &\overline{X}_1 = 1051, && \overline{X}_2 = 1009, \\ &\sigma_1 = 90, && \sigma_2 = 78, \\ &n_1 = 40, && n_2 = 35, \\ &\alpha = 0.1, && Z_{\alpha/2} = 1.64, \end{aligned}$$

con lo cual:

$$IC = (1051 - 1009) \pm 1.64 \sqrt{\frac{90^2}{40} + \frac{78^2}{35}} = (10.19, 73.81).$$

□

5. Contraste de hipótesis

Se tiene una población en la que se desconoce un parámetro (media poblacional, proporción poblacional, etc.) y se quiere estudiar la falsedad de una afirmación realizada acerca del verdadero valor del parámetro.

En general, se siguen cuatro pasos:

1. Enunciar la *hipótesis nula* H_0: afirmación que se quiere estudiar. La afirmación complementaria se llama *hipótesis alternativa* H_1.

2. Construir la *región de aceptación* o *de no rechazo*: es un intervalo que permitirá decidir sobre la falsedad o no de H_0. Para construirlo se necesita un nivel de significación α. La región complementaria se llama *región crítica* o *de rechazo*.

3. Extracción de una muestra y cálculo del parámetro muestral.

4. Toma de la decisión: *Si el parámetro muestral pertenece a la región de aceptación, entonces no podemos rechazar H_0 a nivel de significación α. Si el parámetro muestral no pertenece a la región de aceptación, entonces debemos rechazar H_0 a nivel de significación α.*

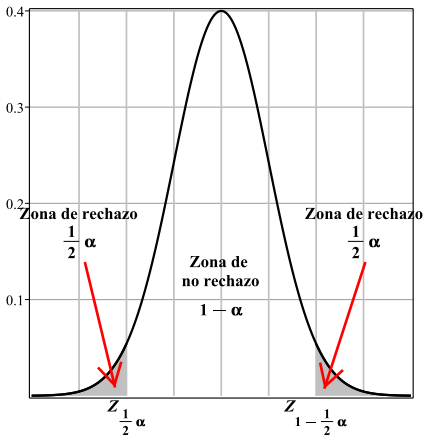

Figura 5.1. Región de aceptación (o de no rechazo) y región crítica (o de rechazo) de H_0.

5.1. Errores de tipo I y tipo II

Definición 5.1. *Si se rechaza una hipótesis cuando debería ser aceptada, se dice que se comete un* error de tipo I. *Si por el contrario, se acepta una hipótesis que debe ser rechazada, se dice que se comete un* error de tipo II. *En cualquiera de los dos casos se comete un error al tomar una decisión equivocada.*

Decisión / Realidad de H_0	Se rechaza H_0	No se rechaza H_0
H_0 es verdadera	Error tipo I	Decisión correcta
H_0 es falsa	Decisión correcta	Error tipo II

Cuadro 5.1. Aparición de los errores de tipos I y II en un test de de hipótesis.

Para que cualquier test de hipótesis o regla de decisión sea bueno, debe diseñarse de forma que minimice los errores de decisión. Esto no es tan sencillo como puede parecer, puesto que para un tamaño de muestra dado, un intento de disminuir un tipo de error va generalmente acompañado de un incremento en el otro tipo de error. En la práctica, un tipo de error puede tener más importancia que el otro, y así se tiende a poner una limitación al error de mayor importancia. La única forma de disminuir al tiempo ambos tipos de error es incrementar el tamaño de la muestra, lo cual no siempre es posible.

5.2. Nivel de significación y p-valor

Definición 5.2. *La probabilidad máxima con la que se puede cometer un error del tipo I en un test de hipótesis se llama* nivel de significación *del test y se denota por medio de la letra* α.

El nivel de significación generalmente se fija antes de la extracción de las muestras, de modo que los resultados obtenidos no influyan en la elección. En la práctica se acostumbra a utilizar niveles de significación del 0.05 ó 0.01, aunque igualmente se pueden emplear otros valores. Si, por ejemplo, se elige un nivel de significación del 0.05, ó 5 %, al diseñar un test de hipótesis, entonces hay aproximadamente 5 ocasiones de cada 100 en que se rechazaría la hipótesis cuando debería ser aceptada, es decir, se tiene un 95 % de confianza en que se tome la decisión adecuada. En tal caso se dice que *la hipótesis ha sido rechazada a nivel de significación del* 0.05, lo que significa que se puede cometer error con una probabilidad de 0.05.

La elección del nivel de significación, tal y como se ha comentado anteriormente, es, en cierta forma, arbitraria. Sin embargo, una vez obtenida la muestra, se puede calcular una cantidad que permite resumir el resultado del experimento de manera objetiva. Esta cantidad es el p-valor, que corresponde al nivel de significación más pequeño que puede ser elegido, para el cual todavía se aceptaría la hipótesis alternativa con las observaciones actuales. En otras palabras:

Definición 5.3. *El valor de* α *para el que se produce un cambio en la decisión se denomina* p-valor *del contraste.*

El p-valor da una medida de cuánto contradice la muestra actual la hipótesis alternativa. Al proporcionar el p-valor, p_v, obtenido con la muestra actual, la decisión se hará de acuerdo a la regla siguiente:

- si $p_v \leq \alpha$, aceptar H_1;
- si $p_v > \alpha$, aceptar H_0.

5.3. Contrastes para la media de una población normal con σ conocida

1. Contraste bilateral:
$$\begin{cases} H_0: & \mu = \mu_0, \\ H_1: & \mu \neq \mu_0. \end{cases}$$
Región de aceptación: $\left(\mu_0 - Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}, \mu_0 + Z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$.

2. Contraste unilateral:
$$\begin{cases} H_0: & \mu \leq \mu_0, \\ H_1: & \mu > \mu_0. \end{cases}$$
Región de aceptación: $\left(-\infty, \mu_0 + Z_{\alpha}\dfrac{\sigma}{\sqrt{n}}\right)$.

3. Contraste unilateral:
$$\begin{cases} H_0: & \mu \geq \mu_0, \\ H_1: & \mu < \mu_0. \end{cases}$$
Región de aceptación: $\left(\mu_0 - Z_{\alpha}\dfrac{\sigma}{\sqrt{n}}, +\infty\right)$.

Ejemplo 5.4. *La longitud de los lápices de una cierta marca se distribuye normalmente con media desconocida y desviación típica de* 0.5 *centímetros. Se toma una muestra de* 50 *lápices y se obtiene una longitud media de* 17.5 *centímetros. ¿Se puede afirmar con una confianza del* 95 % *que la longitud media de todos los lápices es de* 18 *centímetros?*

RESOLUCIÓN. Se tiene:
$$\begin{cases} H_0: & \mu = 18\,(=\mu_0), \\ H_1: & \mu \neq 18. \end{cases}$$

Como $n = 50$, $\sigma = 0.5$, $\overline{X} = 17.5$, $\alpha = 0.05$ y $Z_{\alpha/2} = 1.96$, la región de aceptación es:

$$\mu_0 \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 18 \pm 1.96 \cdot \frac{0.5}{\sqrt{50}} = (17.8614, 18.1386).$$

Y como $17.5 \notin (17.8614, 18.1386)$, podemos rechazar H_0 con una significación del 5 %: la longitud media no es 18 cm con una confianza del 95 %. □

Ejemplo 5.5. *Una universidad afirma que la edad media de sus estudiantes de doctorado es inferior a* 30 *años. Se toma una muestra de* 40 *alumnos y se obtiene una edad media de* 30.5 *años. Se sabe que la edad de los estudiantes tiene una desviación típica de* 2 *años. ¿Se puede aceptar la afirmación de la universidad con una significación del* 10 % *?*

RESOLUCIÓN. En este caso,

$$\begin{cases} H_0: & \mu \leq 30\,(=\mu_0), \\ H_1: & \mu > 30. \end{cases}$$

Teniendo en cuenta que los datos son $n = 40$, $\alpha = 0.1$, $Z_\alpha = 1.28$, $\overline{X} = 30.5$ y $\sigma = 2$, sigue que la región de aceptación es

$$\left(-\infty, \mu_0 + Z_\alpha \frac{\sigma}{\sqrt{n}}\right) = (-\infty, 30.4048)\,.$$

Como $30.5 \notin (-\infty, 30.4048)$, rechazamos H_0 al nivel 10 % de significación: la edad de los estudiantes de doctorado no es inferior a 30 años con una confianza del 90 %. □

Ejemplo 5.6. *En una fábrica de lámparas se garantiza una duración media de* 850 *horas para lámparas de* 60 *watios. Se sabe que el tiempo de vida de las lámparas es normal, con una desviación típica de* 120 *horas. Se toma una caja de* 64 *lámparas y se observa una duración media de* 750 *horas. ¿Será necesario rechazar ese lote de lámparas por no cumplir la garantía con una confianza del* 95 %*? ¿Cuál será la duración media mínima que permite no rechazar el lote de lámparas con el mismo nivel de confianza?*

RESOLUCIÓN. Tenemos:

$$\begin{cases} H_0: & \mu \geq 850\,(=\mu_0), \\ H_1: & \mu < 850. \end{cases}$$

Según los datos: $n = 64$, $\overline{X} = 750$, $\alpha = 0.05$, $Z_\alpha = 1.645$ y $\sigma = 120$. La región de aceptación será:

$$\left(\mu_0 - Z_\alpha \frac{\sigma}{\sqrt{n}}, +\infty\right) = (825.325, +\infty).$$

Como $750 \notin (825.325, +\infty)$, rechazamos la caja de lámparas por no cumplir la garantía de duración con una confianza del 95 %. Para no rechazar el lote de lámparas con el mismo nivel de confianza, la duración media de las lámparas en las cajas debería ser de 826 horas, al menos. □

5.4. Contrastes para una proporción p de una población

1. Contraste bilateral:
$$\begin{cases} H_0: & p = p_0, \\ H_1: & p \neq p_0. \end{cases}$$
Región de aceptación: $\left(p_0 - Z_{\alpha/2}\sqrt{\dfrac{p_0(1-p_0)}{n}}, p_0 + Z_{\alpha/2}\sqrt{\dfrac{p_0(1-p_0)}{n}} \right)$.

2. Contraste unilateral:
$$\begin{cases} H_0: & p \leq p_0, \\ H_1: & p > p_0. \end{cases}$$
Región de aceptación: $\left(-\infty, p_0 + Z_{\alpha}\sqrt{\dfrac{p_0(1-p_0)}{n}} \right)$.

3. Contraste unilateral:
$$\begin{cases} H_0: & p \geq p_0, \\ H_1: & p < p_0. \end{cases}$$
Región de aceptación: $\left(p_0 - Z_{\alpha}\sqrt{\dfrac{p_0(1-p_0)}{n}}, +\infty \right)$.

Ejemplo 5.7. *La policía local de una ciudad afirma que más del* 65 % *de accidentes en fin de semana se deben al exceso de alcohol. Para contrastar esta afirmación se observan* 35 *accidentes y se comprueba que* 24 *de ellos se deben al alcohol. ¿Se puede aceptar la afirmación de la policía local con una confianza del* 99 %*?*

RESOLUCIÓN. Se tiene que
$$\begin{cases} H_0: & p \geq 0.65\,(= p_0), \\ H_1: & p < 0.65. \end{cases}$$
Teniendo en cuenta que en este caso $n = 35$, $\alpha = 0.01$ y $Z_\alpha = 2.33$, la región de aceptación será:
$$\left(p_0 - Z_{\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}, +\infty \right) = (0.4621, +\infty).$$
Como $\widehat{p} = 24/35 = 0.6857 \in (0.4621, +\infty)$, no podemos rechazar la afirmación con un 99 % de confianza. □

Ejemplo 5.8. *En las últimas elecciones el partido gobernante obtuvo un* 54.5 *% de los votos. En una encuesta reciente a* 500 *personas,* 247 *declararon su intención de voto a dicho partido. ¿Se puede afirmar, con una confianza del* 90 *%, que el partido ha perdido popularidad?*

RESOLUCIÓN. En estas condiciones,

$$\begin{cases} H_0: & p \leq 0.545\,(=p_0), \\ H_1: & p > 0.545. \end{cases}$$

Los datos son: $n = 500$, $\alpha = 0.1$ y $Z_\alpha = 1.28$, así que la región de aceptación es

$$\left(-\infty, p_0 + Z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\right) = (-\infty, 0.5735).$$

Como $\widehat{p} = 247/500 = 0.494 \in (-\infty, 0.5735)$, no podemos rechazar H_0 al 90 % de confianza. □

Ejemplo 5.9. *Un profesor afirma que exactamente el* 70 *% de sus alumnos aprueba sus exámenes. Se elige una muestra de* 80 *alumnos, de los que* 50 *han aprobado. ¿Se puede aceptar la afirmación al* 10 *% de significación?*

RESOLUCIÓN. Tenemos que:

$$\begin{cases} H_0: & p = 0.7\,(=p_0), \\ H_1: & p \neq 0.7. \end{cases}$$

Los datos proporcionados son: $n = 80$, $\alpha = 0.1$ y $Z_{\alpha/2} = 1.645$. La región de aceptación es:

$$\left(p_0 - Z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}, p_0 + Z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}\right) = (0.6157, 0.7843).$$

Como $\widehat{p} = 50/80 = 0.625 \in (0.6157, 0.7843)$, no podemos rechazar la afirmación a ese nivel de significación. □

5.5. Contrastes para la diferencia de medias de dos poblaciones con σ_1 y σ_2 conocidas

Sean $\overline{X}_1$ y $\overline{X}_2$ las medias muestrales.

Contraste bilateral:

$$\begin{cases} H_0: & \mu_1 = \mu_2, \\ H_1: & \mu_1 \neq \mu_2. \end{cases}$$

Región de aceptación: $\left(\overline{X}_1 - \overline{X}_2 - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{X}_1 - \overline{X}_2 + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$.

Ejemplo 5.10. *Un fabricante de hilo desea comparar la tensión promedio de su hilo con la de su competidor. Las tensiones de* 100 *hilos de cada marca se observaron bajo condiciones controladas. Las medias y las desviaciones típicas de cada marca fueron las siguientes:*

$$\overline{X}_1 = 110.8, \qquad \overline{X}_2 = 108.2,$$
$$s_1 = 10.2, \qquad s_2 = 12.4.$$

Si se supone que el muestreo se llevó a cabo sobre dos poblaciones normales e independientes, ¿existe alguna razón para creer que hay una diferencia entre las tensiones promedio de ruptura de los hilos? Tómese $\alpha = 0.02$.

RESOLUCIÓN. Estamos ante muestras diferentes donde podemos sustituir σ_j por s_j $(j = 1,2)$, al ser $n \geq 30$. El contraste planteado será:

$$\begin{cases} H_0: & \mu_1 = \mu_2, \\ H_1: & \mu_1 \neq \mu_2. \end{cases}$$

Y la región de aceptación será:

$$\begin{aligned} &\left((\overline{X}_1 - \overline{X}_2) - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{X}_1 - \overline{X}_2) + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) = \\ &= \left(2.6 - 2.33\sqrt{\frac{10.2^2}{100} + \frac{12.4^2}{100}}, 2.6 + 2.33\sqrt{\frac{10.2^2}{100} + \frac{12.4^2}{100}} \right) \\ &= (-1.141, 6.341). \end{aligned}$$

Observamos que el intervalo de confianza contiene al cero, que es lo que postula la hipótesis nula. Por tanto, no podemos rechazar al nivel de confianza del 98 % que no existe diferencia entre ambas medias. □