Introducción a la Inferencia Estadística

Manuel Molina Fernández y Jacinto Martín Jiménez

Presentación basada en contenidos de la asignatura Fundamento Científico del Currículo de Matemáticas en Enseñanza Secundaria II

Máster Formación Profesorado Enseñanza Secundaria

Universidad de Extremadura

Introducción a la Inferencia Estadística

- Principales objetivos de la inferencia estadística.
- Principales procedimientos inferenciales.
- Estimación de parámetros.
- Estimación puntual de parámetros.
- Estimación por intervalos de confianza.
- Intervalos de confianza para la media y para la proporción.
- Contraste de hipótesis.
- Estudio de algunos contrastes de hipótesis.

Principales objetivos de la estadística inferencial

La inferencia estadística (estadística inferencial) es la parte de la estadística que estudia métodos para obtener un conocimiento lo más fiel posible de la distribución de probabilidad de la variable (variables) aleatoria(s) investigada(s) en la población (poblaciones) bajo estudio.

Proporciona los procedimientos estadísticos apropiados para obtener, a partir de la información aportada por la muestra (muestras) seleccionada(s), conclusiones generales (inferencias científicas) sobre la población (poblaciones) objeto de estudio.

Además de su interés teórico, es una de las partes de la estadística que mayor importancia tiene en los estudios de carácter aplicado. En este tipo de estudios, la finalidad principal suele ser obtener, con ciertos márgenes de error, unas conclusiones generales válidas para la población estudiada.

Los métodos que estudia la estadística descriptiva sólo permiten extraer conclusiones para los conjuntos de datos analizados.

Para la obtención de conclusiones generales será necesario recurrir a los métodos que proporciona la inferencia estadística.

Inferencia paramétrica e inferencia no paramétrica

Atendiendo al conocimiento que tenemos sobre la distribución de probabilidad de la variable aleatoria bajo estudio, distinguiremos entre inferencia paramétrica e inferencia no paramétrica.

Inferencia paramétrica

En un problema de inferencia estadística decimos que estamos en situación paramétrica cuando la distribución de probabilidad de la variable aleatoria que estamos estudiando es conocida salvo determinados parámetros que intervienen en su forma funcional. En el momento en que tales parámetros sean conocidos quedará especificada dicha distribución de probabilidad.

Ejemplo 1

Supongamos que los responsables de una empresa dedicada a la fabricación de componentes eléctricas están interesados en realizar un control de calidad. La empresa vende las componentes eléctricas en lotes de 10 unidades, siendo uno de los objetivos del control de calidad el estudio de la variable *número de unidades defectuosas por lote*.

Denotemos por X a dicha variable. Sabemos que es de tipo discreto, siendo sus posibles valores: 0, 1,...,10, y que tiene como distribución de probabilidad la distribución binomial con parámetros 10 y p, siendo p la probabilidad de que una componente eléctrica fabricada por dicha empresa sea defectuosa. Por lo tanto, teniendo en cuenta la distribución de probabilidad del modelo binomial:

$$P(X=k)=\frac{10!}{k!(10-k)!}\,p^k(1-p)^{10-k},\quad k=0,1,\ldots,10.$$

Estamos en situación paramétrica, conocemos la distribución de probabilidad de la variable aleatoria bajo estudio (en este caso la distribución binomial) pero desconocemos el valor del parámetro p que interviene en su forma funcional (sólo sabemos que está entre 0 y 1).

Ejemplo 2

Supongamos que en cierta población se está estudiando la *estatura*. Por estudios previos realizados en dicha población, se sabe que la estatura sigue una distribución normal de probabilidad pero se desconoce su media y su varianza actual.

En consecuencia, si denotamos por X a la estatura en dicha población, su función de densidad será:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < +\infty$$

Estamos de nuevo en situación paramétrica, conocemos la distribución de probabilidad de la variable aleatoria X (en este caso la distribución normal) pero desconocemos los parámetros μ y σ^2 (la media y la varianza) que intervienen en su fucnión de densidad.

Inferencia no paramétrica

En un problema de inferencia estadística decimos que estamos en situación no paramétrica cuando desconocemos la distribución de probabilidad de la variable aleatoria investigada.

Ejemplo 3

Supongamos que en determinada población de pacientes renales, un equipo de nefrólogos está investigando cierta variable bioquímica X cuyos valores están comprendidos entre 150 mg/l y 900 mg/l pero cuyo comportamiento es de tal naturaleza que se desconoce su distribución de probabilidad.

Estamos en situación no paramétrica. Tenemos cierta información sobre X, sabemos que es de tipo continuo y conocemos su espacio muestral, pero desconocemos su distribución de probabilidad.

En situación no paramétrica el desconocimiento sobre la variable aleatoria estudiada es mayor que en situación paramétrica.

Sólo debemos hacer uso de los métodos paramétricos cuando la distribución de probabilidad de la variable estudiada sea conocida. En caso contrario, lo más prudente es utilizar los métodos alternativos (más generales) no paramétricos.

En este tema nos centraremos en algunos métodos inferenciales paramétricos que son estudiados en el bachillerato.

Principales procedimientos inferenciales

- **Procedimientos de estimación**
- **Procedimientos de contraste de hipótesis**

Procedimientos de estimación

Su finalidad es proporcionar métodos apropiados para determinar buenas aproximaciones (estimaciones) a los parámetros de interés de la distribución de probabilidad correspondiente a la variable aleatoria bajo estudio. Proporcionan también métodos para determinar los errores cometidos con dichas estimaciones. La teoría sobre estimación estadística es la parte de la inferencia estadística encargada del estudio de tales procedimientos.

Procedimientos de contraste de hipótesis

Su finalidad es proporcionar las herramientas necesarias para poder decidir, con cierta probabilidad de error, sobre la veracidad o no de determinada afirmación de interés en el estudio a realizar. La teoría sobre contraste de hipótesis es la parte de la inferencia estadística encargada del estudio de tales procedimientos.

En un problema de inferencia estadística es importante saber distinguir si en su resolución hemos de recurrir a procedimientos de estimación, a procedimientos de contraste de hipótesis, o a ambos tipos de procedimientos.

El propio objetivo del estudio a realizar será el que nos indicará si se trata de un problema de estimación o si se trata de un problema de contraste de hipótesis.

Ejemplo 4

Supongamos que los responsables académicos de una población escolar desean conocer la calificación media que obtendrían sus alumnos en una prueba de cultura general calificada entre 0 y 100 puntos.

Se trata de un problema de estimación. Tendríamos que seleccionar una muestra aleatoria de escolares de dicha población, pasarles la prueba de cultura general y, a partir de la información recogida, recurrir a los procedimientos de estimación apropiados que nos permiten estimar la calificación media en esa población de escolares y determinar el error cometido con dicha estimación.

Supongamos que los responsables académicos de la población escolar anterior sospechan que la calificación media que obtendrían sus alumnos en la prueba de cultura general es superior a 60 puntos y desean saber si están en lo cierto.

Se trata de un problema de contraste de hipótesis. Tendríamos que seleccionar una muestra aleatoria de escolares, pasarles la prueba de cultural general y, a partir de la información recogida, haciendo uso del procedimiento apropiado de contraste de hipótesis, decidir (con cierto margen de error) si los responsables académicos de dicha población están o no en lo cierto. Si denotamos por μ a la calificación media en la población, se trataría de decidir si $\mu \leq 60$ (no estarían en lo cierto) o si por el contrario $\mu > 60$ (estarían en lo cierto).

Ejemplo 5

Un laboratorio farmacéutico afirma que ha elaborado un nuevo fármaco que es más eficaz para curar determinada enfermedad cardiovascular que el fármaco que se viene utilizando de forma habitual.

Denotemos por p_N y p_H las probabilidades de que un enfermo con dicha enfermedad cardiovascular se cure con el fármaco nuevo y con el fármaco habitual, respectivamente.

En principio, estamos ante un problema de contraste de hipótesis, pues se trata de decidir si $p_N \leq p_H$ (no estaría en lo cierto el laboratorio) o $p_N > p_H$ (estaría en lo cierto el laboratorio).

Tendríamos que seleccionar dos muestras aleatorias de enfermos con dicha enfermedad cardiovascular. A los enfermos de una muestra tratarlos con el fármaco nuevo y a los enfermos de la otra muestra tratarlos con el fármaco habitual. A partir de la información recogida aplicaríamos el procedimiento de contraste de hipótesis apropiado.

Supongamos que tras aplicar dicho procedimiento aceptamos, con cierto margen de error, que el laboratorio está en lo cierto. Esto no sería motivo suficiente para que las autoridades sanitarias decidan eliminar el fármaco habitual y sustituirlo por el fármaco nuevo.

Sería necesario estimar la diferencia $p_N - p_H$. Teniendo en cuenta la estimación obtenida se adoptaría la decisión correspondiente.

Recordemos que cuando se realiza un estudio experimental hemos de tener en cuenta las siguientes fases:

1. Planificación y diseño de la investigación
2. Recogida de información
3. Análisis de la información recogida
4. Interpretación de resultados y presentación de conclusiones

En la primera fase, es importante seleccionar correctamente los métodos inferenciales a utilizar, distinguiendo claramente si se trata de métodos de estimación o de contraste de hipótesis. Para ello, tendremos que tener en cuenta la situación paramétrica o no paramétrica en la que se desarrolla el estudio y los objetivos que se persiguen.

En las fases segunda y tercera será fundamental aplicar correctamente los métodos inferenciales seleccionados en la fase anterior. Para ello, contaremos con el importante apoyo del software estadístico.

Finalmente, en la cuarta fase hemos de interpretar en sus justos términos los resultados obtenidos y presentarlos de forma clara y sin que puedan inducir a posible confusión o engaño.

Estimación de parámetros

La estimación estadística de parámetros tiene gran importancia práctica. Suele ser frecuente la necesidad de disponer de valores aproximados de ciertas cantidades de interés económico y social. Veamos algunos ejemplos:

- Las estimaciones mensuales del índice de precios al consumo de los artículos de primera necesidad. Son importantes pues intervienen en la revisión de salarios y pensiones.
- Las estimaciones del número medio de pacientes diario. Son importantes con objeto de conocer el número de camas disponibles en los hospitales.
- Las estimaciones de los ingresos medios producidos por ventas en las empresas. Son importantes para la planificación de las políticas comerciales y de creación de empleo en las empresas.

- Sea X una variable aleatoria con función de distribución F_θ dependiente de cierto parámetro desconocido $\theta \in \Theta \subseteq \mathbb{R}$.
- La teoría sobre estimación estadística estudia la metodología para determinar buenas aproximaciones al parámetro θ.
- El proceso se apoya en la observación de una muestra aleatoria simple (m.a.s.) $X_1, \ldots, X_n$ (conjunto de variables aleatorias independientes y todas ellas con la misma función de distribución F_θ). Cada observación concreta de dicha muestra proporciona unos datos $x_1, \ldots, x_n$ que combinados de una forma adecuada determinan una aproximación $T(x_1, \ldots, x_n)$ al parámetro θ.
- La teoría sobre estimación estadística estudia también métodos para valorar el error cometido cuando se aproxima θ a través del valor $T(x_1, \ldots, x_n)$.

Ejemplo 6

En una población de estudiantes de bachillerato se ha comprobado que la estatura se distribuye según un modelo normal de probabilidad. Se tiene interés en conocer (estimar) la estatura media μ en dicha población. Para ello, se dispone de los datos siguientes correspondientes a las estaturas (en centímetros) de 20 estudiantes seleccionados al azar en dicha población:

$$165.6,\ 162.8,\ 181.6,\ 187.5,\ 161.2,\ 154.4,\ 178.1,\ 171.3,\ 158.6,\ 169.1$$

$$167.3,\ 169.1,\ 168.1,\ 173.2,\ 172.8,\ 193.2,\ 178.1,\ 166.2,\ 159.5,\ 153.7$$

Teniendo en cuenta que μ representa la estatura media poblacional (valor que nos informa sobre la centralización de la estatura en esa población) se podrían proponer, entre otros, los siguientes valores aproximados para μ a partir de las 20 estaturas observadas:

- La media aritmética de las estaturas:

$$(165,6 + \cdots + 153,7)/20 = 169.57$$

- La mediana de las estaturas:

$$(168.1 + 169.1)/2 = 168.6$$

- La media aritmética entre las estaturas mínima y máxima:

$$(153.7 + 193.2)/2 = 173.45$$

- La media aritmética entre las estaturas primera y vigésima:

$$(165,6 + 153,7)/2 = 159.65$$

Conceptos básicos

Estimador

Se denomina *estimador* a la función $T(X_1, \ldots, X_n)$ de la m.a.s. $X_1, \ldots, X_n$ que se utiliza para estimar el parámetro desconocido θ. Se deduce que un estimador $T = T(X_1, \ldots, X_n)$ es una variable aleatoria.

En el Ejemplo 6, los estimadores que se habían propuesto, basados en la m.a.s. de tamaño 20: $X_1, \ldots, X_{20}$, fueron:

- $T_1(X_1, \ldots, X_{20}) = \bar{X}_{20} = (X_1 + \cdots + X_{20})/20$
- $T_2(X_1, \ldots, X_{20}) = \tilde{X}_{20}$
- $T_3(X_1, \ldots, X_{20}) = (min\{X_1, \ldots, X_{20}\} + max\{X_1, \ldots, X_{20}\})/2$
- $T_4(X_1, \ldots, X_{20}) = (X_1 + X_{20})/2$

Estimación

Se denomina *estimación* al valor que toma el estimador T para una muestra de datos concreta $x_1, \ldots, x_n$. Cada muestra de datos proporciona una estimación para θ.

En el Ejemplo 6, las estimaciones proporcionadas por los estimadores propuestos para las 20 estaturas obervadas fueron:

- $T_1(165.6, \ldots, 166.2) = (165,6 + \cdots + 153,7)/20 = 169.57$
- $T_2(165.6, \ldots, 166.2) = (168.1 + 169.1)/2 = 168.6$
- $T_3(165.6, \ldots, 166.2) = (153.7 + 193.2)/2 = 173.45$
- $T_4(165.6, \ldots, 166.2) = (165,6 + 153,7)/2 = 159.65$

Error en la estimación

Se denomina *error en la estimación* a la diferencia, en valor absoluto, entre la estimación proporcionada por el estimador T para una muestra concreta de observaciones y el verdadero valor de θ:

$$Error = |T(x_1, \ldots, x_n) - \theta|$$

Puesto que θ es desconocido, no será posible determinar dicho error. En consecuencia, será necesario estudiar métodos que nos permitan valorar la calidad de las estimaciones.

En general, la estimación estadística tiene entre sus objetivos básicos:

- Precisar criterios que permitan comparar estimadores.
- Estudiar propiedades deseables que debe tener un estimador.
- Determinar métodos apropiados para la obtención de buenos estimadores.
- Proporcionar métodos que permitan valorar la calidad de las estimaciones.

Estimación puntual de parámetros

Estimador centrado o insesgado

Decimos que un estimador $T = T(X_1, \ldots, X_n)$ es *centrado o insesgado* para el parámetro θ si el centro de gravedad de su distribución de probabilidad (media) coincide con el verdadero valor del parámetro a estimar, es decir si $E[T] = \theta$.

Si $E[T] - \theta \neq 0$ el estimador T es *no centrado o sesgado* para θ.

- Cuando $E[T] > \theta$ el estimador T es *sesgado a la derecha*. Tiene tendencia a dar estimaciones a la derecha del verdadero valor de θ.
- Cuando $E[T] < \theta$ el estimador T es *sesgado a la izquierda*. Tiene tendencia a dar estimaciones a la izquierda del verdadero valor de θ.

Estimación de los principales parámetros

Estimación de la media

Supongamos que hemos de estimar la media μ de cierta distribución de probabilidad.
Sea $X_1, \ldots, X_n$ una m.a.s. de dicha distribución.

El mejor estimador $T(X_1, \ldots, X_n)$ para μ es la *media muestral*:

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

$E[\bar{X}_n] = \mu$ (es centrado).
$V[\bar{X}_n] = \sigma^2/n$ (σ^2 es la varianza poblacional).

Estimación de la varianza

Supongamos que hemos de estimar la varianza σ^2 de cierta distribución de probabilidad.
Sea $X_1, \ldots, X_n$ una m.a.s. de dicha distribución.
El mejor estimador $T(X_1, \ldots, X_n)$ para σ^2 es la *cuasivarianza muestral*:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

$E[S_n^2] = \sigma^2$ (es centrado).
$s_n^2 = \frac{n-1}{n} S_n^2$ no es centrado para σ^2, $E[s_n^2] = \frac{n-1}{n}\sigma^2$
$E[s_n^2] - \sigma^2 = -\sigma^2/n$ (es sesgado a la izquierda).

Estimación de la proporción

Supongamos que tenemos que estimar la proporción p de cierta característica bajo estudio en la población.
(La distribución de probabilidad subyacente es la Bernoulli: $B(1,p)$).
Sea $X_1,\ldots,X_n$ una m.a.s. de dicha distribución.
El mejor estimador $T(X_1,\ldots,X_n)$ para p es la *proporción muestral*:

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$E[\hat{p}_n] = p$ (es centrado).
$V[\hat{p}_n] = p(1-p)/n.$

Estimación por intervalos de confianza

En ocasiones, en lugar de proporcionar una estimación puntual de un parámetro θ, suele ser más informativo proporcionar un intervalo basado en la m.a.s. $I = [T_1(X_1, \ldots, X_n);\ T_2(X_1, \ldots, X_n)]$ que contenga al verdadero valor del parámetro con una probabilidad prefijada suficientemente alta como para proporcionar una confianza razonable de que $\theta \in I$.

En el ejemplo 6 sobre la estatura en una población de estudiantes de bachillerato, en lugar de afirmar que estimamos que la estatura media en la población es de 169.57 centímetros (estimación que nos proporcionaba la media muestral) sería más preciso afirmar que la estatura media poblacional μ está comprendida entre 164.71 y 174.43 centímetros con una confianza del 95%.

Intervalo de confianza

- Sea $X_1, \ldots, X_n$ una m.a.s. $\quad F_\theta, \quad \theta \in \Theta \subseteq \mathbb{R}$
- Se trata de determinar dos funciones de la m.a.s. que denotaremos como $T_1(X_1, \ldots, X_n)$ y $T_2(X_1, \ldots, X_n)$ tales que:

$$P\left(T_1(X_1, \ldots, X_n) \leq \theta \leq T_2(X_1, \ldots, X_n)\right) = 1 - \alpha$$

  A la probabilidad $1 - \alpha$ (puesta por el experimentador) se le denomina nivel de confianza. Se suele expresar en porcentaje y, en la investigación experimental, oscila entre el 90% y el 99%.

  $[T_1(X_1, \ldots, X_n), T_2(X_1, \ldots, X_n)]$ recibe el nombre de *intervalo de confianza* al nivel $(1 - \alpha)\%$ para el parámetro θ.
- Antes de observar los valores concretos de la muestra, el intervalo $[T_1(X_1, \ldots, X_n), T_2(X_1, \ldots, X_n)]$ es aleatorio. Una vez observados los valores $x_1, \ldots, x_n$, pasa a ser un intervalo fijo.

Intervalos de confianza para la media y para la proporción

Intervalo de confianza para la media (varianza conocida)

- $X_1, \ldots, X_n$ m.a.s $N(\mu, \sigma)$, σ conocida.
- Teniendo en cuenta que $\bar{X}_n \to N(\mu, \frac{\sigma}{\sqrt{n}})$, se deduce:

$$P\left(-z_\alpha \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq +z_\alpha\right) = 1 - \alpha$$

  z_α es el valor en la $N(0,1)$ que deja a su derecha $\alpha/2$.
- Se obtiene como intervalo de confianza al $(1-\alpha)\%$ para μ:

$$\left[\bar{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; \bar{X}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

Ejemplo 7

Se desea estimar, con una confianza del 90%, la intensidad media que circula por una componente de un circuito. Se ha comprobado que la intensidad (en miliamperios) sigue una distribución normal con varianza 144. A partir de una muestra aleatoria de 25 medidas de la intensidad en dicha componente se obtuvo una intensidad media de 85 miliamperios. Con los datos proporcionados vamos a determinar el intervalo de confianza al 90% para la intensidad media que circula por dicha componente:

$$n = 25, \quad \bar{x}_{25} = 85, \quad \sigma = 12 \quad 1 - \alpha = 0.90, \quad z_{0.10} = 1.645$$

$$[85 - (1.645)(12)/\sqrt{25}\ ;\ 85 + (1.645)(12)/\sqrt{25}] = [81.052\ ;\ 88.948]$$

Intervalo de confianza para la media (varianza desconocida)

- $X_1, \ldots, X_n \quad (n \geq 100) \quad$ m.a.s $\quad N(\mu, \sigma), \quad \sigma$ desconocida.
- Daremos como estimación de σ el valor S_n.
- Se deduce como intervalo de confianza al $(1-\alpha)\%$ para μ:

$$\left[\bar{X}_n - z_\alpha \frac{S_n}{\sqrt{n}} \; ; \; \bar{X}_n + z_\alpha \frac{S_n}{\sqrt{n}}\right]$$

Ejemplo 8

Una máquina fabrica cojinetes. Se ha comprobado que el diámetro del cojinete sigue una distribución normal. A partir de una muestra de 120 cojinetes fabricados por la máquina se ha determinado un diámetro medio de 2 centímetros y una cuasidesviación típica de 0.1 centímetros. Determinar el intervalo de confianza para el diámetro medio de los cojinetes producidos por la máquina con una confianza del 95%.

$$n = 120, \quad \bar{x}_{120} = 2, \quad S_{120} = 0.1, \quad 1 - \alpha = 0.95, \quad z_{0.05} = 1.96$$

$$[2 - (1.96)(0.1)/\sqrt{120}\,;\ 2 + (1.96)(0.1)/\sqrt{120}] = [1.982\,;\ 2.018]$$

Intervalo de confianza para la proporción

- $X_1, \ldots, X_n$ m.a.s $B(1,p)$, p desconocida.
- p se estima por $\hat{p}_n$. Teniendo en cuenta la aproximación de la distribución binomial a través de la distribución normal, se demuestra que

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

- Se deduce como intervalo de confianza al $(1-\alpha)\%$ para μ:

$$\left[\hat{p}_n - z_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}};\ \hat{p}_n + z_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

$$n\hat{p}_n > 5, \quad n(1-\hat{p}_n) > 5, \quad \hat{p}_n > 0.05, \quad 1-\hat{p}_n > 0.05$$

Ejemplo 9

Una fábrica produce botellas. Se cargan 1000 botellas en un camión. Se realiza un control de calidad a la carga de dicho camión y se detectan 75 botellas defectuosas. Determinar:

(a) La estimación que daríamos para la proporción de botellas defectuosas producidas en la fábrica.

(b) El intervalo de confianza para la proporción de botellas defectuosas producidas por la fábrica con una confianza del 99%.

Ejemplo 9

(a)

$$\hat{p}_{1000} = 75/1000 = 0.075$$

(b)

$$[0.075 - (2.57)\sqrt{\frac{(0.075)(0.925)}{1000}}; 0.075 + (2.57)\sqrt{\frac{(0.075)(0.925)}{1000}}]$$

$$= [0.054\,; 0.096]$$

Contraste de hipótesis

- Como parte de los objetivos de un estudio estadístico, frecuentemente se plantean determinadas preguntas que se traducen en la formulación de unas hipótesis sobre los parámetros u otras caracteristicas de la población (poblaciones) estudiada(s).

- La teoría sobre contraste de hipótesis, iniciada por Neyman y Pearson en 1940, tiene como objetivo desarrollar la metodología necesaria para decidir, con ciertos niveles de error, la hipótesis que debe aceptarse.

- Un contraste (test) de hipótesis es un procedimiento estadístico que, a partir de la información proporcionada por la muestra(s) aleatoria(s) seleccionada(s), permite aceptar o rechazar una hipótesis previamente formulada en la población (poblaciones) bajo estudio.

Conceptos básicos

- Hipótesis nula e hipótesis alternativa
- Contraste bilateral y contraste unilateral
- Tipos de errores
- Probabilidades asociadas a los tipos de errores
- Prueba de significación

Hipótesis nula e hipótesis alternativa

Con objeto de distinguir entre las dos hipótesis a contrastar, nos referiremos a ellas como:

Hipótesis nula (la denotaremos por H_0).
Hipótesis alternativa (la denotaremos por H_1.)

Usualmente, en la hipótesis alternativa pondremos la sospecha que se trata de confirmar.

Contrate bilateral y contraste unilateral

Dependiendo del objetivo del estudio a realizar, habrá situaciones en las que en la hipótesis alternativa tengamos el signo $\neq$ (no se concreta un sentido) y otras situaciones en las que tengamos los signos $>$ ó $<$ (se concreta un sentido). En el primer caso hablaremos de *contraste de hipótesis bilateral* y en el segundo de *contraste de hipótesis unilateral*.

Tipos de errores

Puesto que hemos de decidir entre las dos hipótesis formuladas a partir de la información proporcionada por la(s) muestra(s) seleccionada(s) en la(s) población (poblaciones) bajo estudio nuestra decisión estará sujeta a ciertos errores. En concreto, podemos cometer dos tipos de errores a los que nos referiremos como *error tipo I* y *error tipo II*.

Error tipo I

Aceptar H_1 cuando en realidad H_0 es cierta.

Error tipo II

Aceptar H_0 cuando en realidad H_1 es cierta.

Probabilidades asociadas a los tipos de errores

Los errores tipo I y tipo II no es posible cometerlos con probabilidad cero dado que la decisión se toma en base a la información proporcionada por muestras y no en base a la información total. Nuestro objetivo será que las probabilidades de cometer tales errores sean lo más próximas a cero que sea posible. Denominaremos:

$$\alpha = P(\textit{Cometer Error Tipo I}) = P(\textit{Aceptar } H_1 \mid \ H_0 \textit{ cierta})$$

$$\beta = P(\textit{Cometer Error Tipo II}) = P(\textit{Aceptar } H_0 \mid \ H_1 \textit{ cierta})$$

Consideremos α, β y el tamaño de la muestra que representaremos por n. De esos tres valores siempre podremos fijar dos y encontrar la regla de decisión apropiada que nos permita llegar a una conclusión.

Cabe pensar en los tres planteamientos siguientes:

- Fijar α y β.
- Fijar α y n.
- Fijar β y el n.

De acuerdo con el primer planteamiento, en el que fijamos α y β, el n no lo controlamos. Se trataría de determinar el tamaño de muestra necesario para garantizar las probabilidades de error prefijadas. Suele ser un planteamiento poco usual en la práctica experimental.

Prueba de significación

De acuerdo con el segundo planteamiento, en el que fijamos n y α, el β no lo controlamos. Se trataría de determinar la mejor regla de decisión posible que, garantizando el α prefijado, nos proporcione el menor β posible. Es el planteamiento mas utilizado en la práctica experimental fundamentalmente por dos razones:

- En la mayoría de las ocasiones el n lo establece el experimentador forzado por los recursos y tiempo disponibles.
- Al experimentador le interesa muy especialmente controlar α (*nivel de significación del test*).

Muchos autores se refieren a este planteamiento, considerado originalmente por R.A. Fisher, como *prueba de significación*.

Resolución práctica de un contraste de hipótesis

Supongamos que una vez formuladas las hipótesis H_0 y H_1, hemos fijado un n determinado y hemos elegido el nivel de significación α que estamos dispuestos a asumir.

La idea directriz en la resolución de un contraste de hipótesis es suponer que H_0 es cierta.

Sólo cambiaremos de opinión si lo observado es muy improbable bajo el supuesto de que H_0 es cierta. En caso contrario no cambiaremos nuestra opinión.

La medida de lo improbable viene determinada por el valor de α prefijado.

Una vez recogidos los datos, el contraste de hipótesis suele resolverse de una forma bastante mecánica. Hay distintas formas de llevar a cabo dicha resolución. La más frecuente consiste en:

- Calcular, a partir de los datos, el denominado *valor experimental del test*. Su cálculo se realiza a partir de una formula estadística que depende de cada situación concreta.
- Determinar, a partir de una distribución de probabilidad teórica, el denominado *valor teórico del test*. Dicho valor depende del α prefijado. La distribución teórica se determina teniendo en cuenta que se está asumiendo que H_0 es cierta.
- Concretar la *regla de decisión* correspondiente a través de la comparación de ambos valores. Dicha regla puede adoptar diferentes formas dependiendo del contraste de hipótesis considerado.

Estudio de algunos contrastes de hipótesis

Contrastes en una población

- Contrastes sobre la media en una población normal.
- Contrastes sobre una proporción.

Contrastes en dos poblaciones

- Contrastes sobre las medias en dos poblaciones normales.
- Contrastes sobre dos proporciones.

Contrastes sobre la media

Sea X una variable aleatoria con distribución de probabilidad $N(\mu, \sigma)$.

Estudiaremos el siguiente contraste de hipótesis sobre μ (considerando los casos σ conocida y σ desconocida):

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_O$$

siendo μ_0 un valor conocido.

$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0 \quad (\sigma \text{ conocida})$

$X_1, \ldots, X_n \quad \text{m.a.s} \quad N(\mu, \sigma), \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to N(\mu, \sigma/\sqrt{n})$

Supuesto que H_0 es cierta: $\bar{X}_n \to N(\mu_0, \sigma/\sqrt{n})$.

Fijado el nivel de significación α:

$$P\left(-z_\alpha \leq \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq +z_\alpha\right) = 1 - \alpha$$

$$V_{exp} = \frac{|\bar{X}_n - \mu_0|}{\sigma/\sqrt{n}}, \quad V_\alpha = z_\alpha$$

Cuando $V_{exp} > z_\alpha$ se acepta H_1 al nivel α.

Cuando $V_{exp} \leq z_\alpha$ se acepta H_0 al nivel α.

Ejemplo 10

En una población de deportistas se ha comprobado que el número de pulsaciones por minuto (en reposo) sigue una distribución normal, siendo la desviación típica 7 pulsaciones por minuto. A partir de una muestra de 50 deportistas seleccionados al azar en dicha población, se ha determinado un numero medio de pulsaciones por minuto de 47. ¿Se podría rechazar, para un nivel de significación de 0.01, la hipótesis de que el número medio de pulsaciones por minuto en esa población es de 45?

$H_0: \mu = 45 \quad H_1: \mu \neq 45$

$n = 50, \quad \bar{x}_{50} = 47, \quad \sigma = 7, \quad \alpha = 0.01$

$V_{exp} = |47 - 45|/(7/\sqrt{50}) = 2.02$

$z_{0.01} = 2.576$

Como $V_{exp} < z_{0.01}$ se acepta H_0.

No hay evidencias para rechazar que el número medio de pulsaciones por minuto en esa población de deportistas es de 45.

Ejemplo 10

Otra forma de resolver el problema sería a través del intervalo de confianza:

$$\left[\bar{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; \bar{X}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

Sustituyendo los correspondientes valores se obtiene el intervalo al 99%:

$$\left[47 - 2.576\frac{7}{\sqrt{50}} \; ; \; 47 + 2.576\frac{7}{\sqrt{50}}\right] = [44.45; 49.55]$$

Teniendo en cuenta que $45 \in [44.5; 49.55]$ se acepta H_0 al nivel 0.05.

$H_0: \mu = \mu_0 \quad H_1: \mu \neq \mu_0 \quad (\sigma \text{ desconocida})$

$X_1, \ldots, X_n$ m.a.s $N(\mu, \sigma)$, $\hat{\sigma}^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$

Supuesto que H_0 es cierta y $n \geq 100$:

$$\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \rightsquigarrow N(0,1)$$

Fijado el nivel de significación α:

$$P\left(-z_\alpha \leq \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \leq +z_\alpha\right) = 1 - \alpha$$

$$V_{exp} = \frac{|\bar{X}_n - \mu_0|}{S_n/\sqrt{n}}, \quad V_\alpha = z_\alpha$$

Cuando $V_{exp} > z_\alpha$ se acepta H_1 al nivel α.

Cuando $V_{exp} \leq z_\alpha$ se acepta H_0 al nivel α.

Contrastes sobre una proporción

Estudiaremos el siguiente contraste:

- $H_0 : \ p = p_0$
- $H_1 : \ p \neq p_0$

siendo p_0 un valor conocido.

$H_0: p = p_0 \quad H_1: p \neq p_0$

$X_1, \ldots, X_n$ m.a.s $B(1,p)$

$np_0 > 5,\ n(1-p_0) > 5,\ p_0 > 0.05,\ 1-p_0 > 0.05$

Si H_0 es cierta:

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N(p_0, \sqrt{p_0(1-p_0)/n})$$

Fijado el nivel de significación α:

$$P\left(-z_\alpha \leq (\hat{p}_n - p_0)/\sqrt{p_0(1-p_0)/n} \leq +z_\alpha\right) = 1-\alpha$$

$V_{exp} = \frac{|\hat{p}_n - p_0|}{\sqrt{p_0(1-p_0)/n}}, \quad V_\alpha = z_\alpha$

Cuando $V_{exp} > z_\alpha$ se acepta H_1 al nivel α.

Cuando $V_{exp} \leq z_\alpha$ se acepta H_0 al nivel α.

Ejemplo 11

A unas elecciones generales se presenta el partido político A. Un comentarista político opina que el partido A obtendrá un 40% de votos. Se pregunta a 250 votantes, seleccionados al azar, sobre su intención de voto. Un total de 132 encuestados manifiestan su intención de votar al partido A. Teniendo en cuenta dicha información ¿se podría rechazar, para un nivel de significación del 0.05, la opinión del comentarista político?

Ejemplo 11

Denotemos por p_A la probabilidad de votar al partido A.

$$H_0 : \ p_A = 0.4 \quad H_1 : \ p_A \neq 0.4$$

$$p_0 = 0.4, \quad 1 - p_0 = 0.6, \quad n = 250, \quad \alpha = 0.05$$

$$\hat{p}_A = 132/250 = 0.528$$

$$V_{exp} = \frac{|0.528 - 0.4|}{\sqrt{0.4(1-0.4)/250}} = 4.13$$

$$V_{0.05} = z_{0.05} = 1.96$$

Como $V_{exp} > V_{0.05}$ se acepta H_1

Para un nivel de significación del 0.05, no lleva razón el comentarista político.

Ejemplo 11

Otra forma de resolver el contraste de hipótesis sería a partir del intervalo de confianza:

$$\left[\hat{p}_n - z_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}};\ \hat{p}_n + z_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

Sustituyendo los correspondientes valores se obtiene el intervalo de confianza al 95%:

$$\left[0.528 - 1.96\sqrt{\frac{0.528(1-0.528)}{250}};\ 0.528 + 1.96\sqrt{\frac{0.528(1-0.528)}{250}}\right] =$$

$$= [0.466; 0.589]$$

Como 0.4 no está incluido en el intervalo se rechaza H_0.

Contrastes en dos poblaciones

Cuando hemos de comparar dos poblaciones, la recogida de la información puede realizarse atendiendo a los siguientes diseños:

- Diseño a través de muestras independientes.
- Diseño a través de muestras relacionadas (apareadas)

Muestras independientes

- Se selecciona una muestra de n_1 unidades de la población 1. Independientemente, se selecciona otra muestra de n_2 unidades de la población 2. Los tamaños muestrales n_1 y n_2 no tienen necesariamente que ser iguales.
- Se determina el valor de la variable estudio X en las unidades de ambas muestras. En consecuencia, tendremos recogidas dos muestras de datos:

$$\{x_1^1, \ldots, x_{n_1}^1\} \quad \text{y} \quad \{x_1^2, \ldots, x_{n_2}^2\}$$

- Para que los contrastes comparativos realizados a través de este diseño sean objetivos, la variable X no deberá estar influenciada por otras variables.

Muestras relacionadas

Cuando X está influenciada por otras variables, será más apropiado proceder de la siguiente forma:

- Se forman n parejas de unidades (la primera unidad de la pareja se selecciona en la población 1 y la segunda unidad se selecciona en la población 2) de tal manera que ambas unidades sean muy similares con respecto a las variables influyentes en X.
- Se determina el valor de X en las dos unidades de cada pareja. En consecuencia, tendremos recogida la muestra de datos:

$$\{(x_1^1, x_1^2), \ldots, (x_n^1, x_n^2)\}$$

- Cuando la pareja está formada por la misma unidad (considerada en dos situaciones diferentes) hablamos de *diseño a través de muestras autoapareadas*.

Ejemplo 12

Supongamos que un equipo de ginecólogos sospecha que determinada variable bioquímica X cambia en valor medio entre las mujeres que están en el tercer mes de gestación y las que están en el séptimo mes de gestación. Para realizar el contraste de hipótesis correspondiente: $H_0 : \mu_3 = \mu_7 \quad H_1 : \mu_3 \neq \mu_7$, se podría diseñar la experiencia de recogida de datos de las siguientes formas:

- *Diseño a través de muestras independientes:*

  Se seleciona una muestra de n_1 mujeres que están en el tercer mes de gestación e, independientemente, otra muestra de n_2 mujeres que están en el séptimo mes de gestación. Se determina el valor de X en ambas muestras y se procede a la resolución del contraste de hipótesis. Es importante comprobar que la variable X no está influenciada por otras. Si tal cosa sucede, la comparación podría resultar no objetiva pues las dos muestras de mujeres seleccionadas podrían ser diferentes en relación a la(s) variable(s) influyente(s).

- *Diseño a través de muestras relacionadas:*

  Supongamos que X está influenciada por la edad. En tal caso, sería más apropiado proceder de la siguiente manera:

  Se seleccionan n parejas de mujeres (la primera en su tercer mes de gestación y la segunda en su séptimo mes) de tal manera que ambas mujeres sean de edades similares. Se determina el valor de X en las mujeres de las n parejas y se procede a la resolución del contraste de hipótesis.

- *Diseño a través de muestras autoapareadas:*

  Se selecciona una muestra de n mujeres que están en su tercer mes de gestación. Se determina en ellas el valor de X. Se espera a que esas mismas mujeres lleguen a su séptimo mes de gestación y se les vuelve a determinar el valor de X. Con los datos recogidos se procede a la resolución del contraste de hipótesis.

Comparación de medias en dos poblaciones

Consideraremos dos muestras aleatorias simples en poblaciones normales. En consecuencia, la variable objeto de estudio X sigue una distribución normal en cada una de las dos poblaciones consideradas, a las que nos referiremos como población 1 y población 2 (las identificaremos a través de los superíndices 1 y 2 respectivamente). Consideraremos la siguiente notación:

$$X_1^1, \ldots, X_{n_1}^1 \quad \text{m.a.s.} \quad N(\mu_1, \sigma_1), \quad X_1^2, \ldots, X_{n_2}^2 \quad \text{m.a.s.} \quad N(\mu_2, \sigma_2)$$

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i^1, \quad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_i^2$$

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i^1 - \bar{X}_1)^2, \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_i^2 - \bar{X}_2)^2$$

Estudiaremos el contraste:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

para el caso de muestras independientes e igualdad de varianzas en ambas poblaciones.

$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$

Consideraremos el caso $\sigma_1^2 = \sigma_2^2 = \sigma^2$, siendo σ^2 un valor conocido. En esta situación, sabemos que:

$$\bar{X}_1 - \bar{X}_2 \to N\left(\mu_1 - \mu_2, \sigma\sqrt{1/n_1 + 1/n_2}\right)$$

Por consiguiente, supuesto que H_0 cierta, fijado el nivel de significación α, se deduce que:

$$P\left(-z_\alpha \leq \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \leq +z_\alpha\right) = 1 - \alpha$$

$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$

$$P\left(-z_\alpha \leq \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \leq +z_\alpha\right) = 1 - \alpha$$

$$V_{exp} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sigma\sqrt{1/n_1 + 1/n_2}}, \qquad V_\alpha = z_\alpha$$

Cuando $V_{exp} > z_\alpha$ se acepta H_1 al nivel α.

Cuando $V_{exp} \leq z_\alpha$ se acepta H_0 al nivel α.

Ejemplo 13

La cantidad de impureza en un lote de una sustancia química se utiliza como medida para evaluar su calidad. Una fábrica tiene dos líneas de producción. La línea 2 ha tenido que ser reparada debido a ciertos problemas. Los responsables de la fábrica desean comprobar que la línea 2 mantiene la misma calidad que la línea 1. Para ello, se selecciona una muestra aleatoria de 25 lotes en cada línea. La cantidad de impureza media obtenida con la muestra de la línea 1 es 3 unidades y con la muestra de la línea 2 es 3.4 unidades. Por estudios previos se sabe que la desviación típica de la cantidad de impureza es 1 unidad en ambas líneas. ¿Se podría aceptar, con un nivel de significación de 0.1, que la cantidad media de impureza en las dos líneas de producción es la misma?

Ejemplo 13

$$H_0: \ \mu_1 = \mu_2 \quad H_1: \ \mu_1 \neq \mu_2$$

$$n_1 = n_2 = 25, \ \bar{X}_1 = 3, \ \bar{X}_2 = 3.4, \ \sigma_1^2 = \sigma_2^2 = \sigma^2 = 1, \ \alpha = 0.1$$

$$V_{exp} = \frac{|\bar{X}_1 - \bar{X}_2|}{\sigma\sqrt{1/n_1 + 1/n_2}} = \frac{|3 - 3.4|}{1\sqrt{1/25 + 1/25}} = 1.414$$

$$\alpha = 0.1, \quad V_{0.1} = z_{0.1} = 1.645$$

Como $V_{exp} < z_{0.1}$ se acepta H_0 para el nivel de significación $\alpha = 0.1$.

Fin de la Presentación