Tema 1

Inferencia estadística. Estimación.

1.1. Introducción a la inferencia estadística

La inferencia estadística trata los métodos mediante los cuales podemos hacer inferencias (extraer determinadas conclusiones o generalizaciones) sobre una población, a partir de la información extraída de una muestra aleatoria de dicha población (como acabamos de repasar en el tema 0).

La inferencia estadística podría dividirse en dos áreas: la estimación y los contrastes de hipótesis. En este tema trataremos la estimación y en el siguiente los contrastes de hipótesis. Veamos algunos ejemplos sencillos como ilustración:

Ejemplo 1.1.: Se evaluó empíricamente un programa utilizando una colección aleatoria de 50 conjuntos de datos de entrada, midiéndose el tiempo de ejecución para cada caso. Se calculó la media y varianza, obteniéndose: $\bar{x} = 65\ ms$ y $s^2 = 36\ ms^2$, respectivamente. La media de dicha muestra puede emplearse para estimar la media de la población entera (todos los tiempos de ejecución para todos los posibles valores de entrada), sin embargo debe quedar claro que NO es la media verdadera de la población. Emplearemos la distribución de muestreo de $\overline{X}$ para tener una idea de la exactitud de la estimación. (**PROBLEMA DE ESTIMACIÓN**).

Ejemplo 1.2.: En el artículo *"Evaluation of Maintenanace software in Real-Time Systems"* (*IEEE Trans. on Comput., 1978*), se pretendía conocerse la cobertura de detección de fallos c de un sistema tolerante a fallos. Para conseguir una estimación de la mencionada proporción se insertaron 200 fallos aleatorios. El mecanismo de recuperación detectó con éxito 178 de estos fallos. La estimación podría ser el cociente del número de fallos detectados entre el total (200). De nuevo, la distribución de muestreo de ese estimador nos proporcionará una idea de la "fiabilidad" de la estimación. (**PROBLEMA DE ESTIMACIÓN**).

Ejemplo 1.3.: A la hora de diseñar un sistema de servicios, resulta adecuado contar con diversas hipótesis de trabajo respecto al número de unidades que se deben servir en un cierto momento. Una estimación demasiado baja conducirá a un servicio inadecuado, mientras que una estimación demasiado alta conduciría a un desperdicio de recursos. Supongamos que a una compañía de distribución de recursos informáticos le interesa el número de usuarios interactivos durante una hora dada y desea comprobarse si el número medio de usuarios se desvía de 110 [1]. Se considera una muestra de 25 horas observadas y se obtiene una media de 112.2 y una desviación típica de 8.4. Se plantearía la hipótesis que el número medio de usuarios es 110 y tras las pruebas oportunas, dicha hipótesis podrá o no podrá ser rechazada. En este ejemplo no se pretende estimar un parámetro, sino decidir sobre una hipótesis. La teoría del muestreo también nos ayudará a determinar la exactitud de nuestra decisión. (**PROBLEMA DE CONTRASTE DE HIPÓTESIS**).

Ejemplo 1.4.: Un proveedor nos suministra una máquina. Este proveedor afirma que la proporción de piezas defectuosas que produce la máquina es 0.001. Decidimos comprobarlo, así que extraemos una muestra aleatoria de 2.000 unidades, de las cuales 15 resultan defectuosas. ¿Es aconsejable creer al proveedor o por el contrario, deberíamos recordarle que "si no quedábamos satisfechos nos devolvía el dinero"? (**PROBLEMA DE CONTRASTE DE HIPÓTESIS**).

1.2. Estimación

Distinguiremos dos tipos:

a) Estimación puntual:

Se trata de estimar un parámetro poblacional mediante un número que lo aproxime. En el ejemplo 1.1. estimamos la media de la población (μ) con la media de una muestra ($\overline{x}$) y en ejemplo 1.2. se emplea la proporción de fallos detectados en la muestra ($\hat{p}$) para obtener una estimación de la proporción (cobertura) real (la de la población completa), p. Sin embargo, no podemos esperar que una estimación puntual coincida exactamente con el parámetro poblacional que pretende estimar, por ello en muchas ocasiones será preferible proporcionar un intervalo que contendrá al parámetro poblacional con un grado razonable de certidumbre.

b) Estimación por intervalos:

Obtendremos intervalos, en los que confiamos que se encuentre el parámetro poblacional a estimar, por ejemplo la media poblacional μ. A estos intervalos se les conoce como intervalos de confianza para el parámetro al $(1 - \alpha) \cdot 100\,\%$, donde $1 - \alpha$ es el grado o nivel de confianza o también intervalos de confianza al nivel de significación α. (α estará entre 0 y 1, valores comunes son: 0.1, 0.05 y 0.01). Cuanto mayor sea $1 - \alpha$ (nivel de confianza), más amplio será el intervalo.

¿Cuál es la interpretación de un intervalo de confianza?

Supongamos que construimos un intervalo de confianza al 95 % para μ, para una serie de

[1] nótese que en realidad la variable es discreta pero que las observaciones se aproximan a una normal

muestras de una población Normal, cada una de ellas formada por, por ejemplo, 20 observaciones. Cada vez tendremos una media muestral ($\overline{x}$) diferente, mientras que μ no varía. Entonces, el 95 % de los intervalos que construyésemos contendrá a μ. Por supuesto, en un experimento concreto sólo disponemos de una muestra (formada por los 20 datos) y esperaremos "con confianza" que nuestra muestra sea una de las del 95 % (¡cuidado!: no tiene sentido hablar de la probabilidad de que μ esté en un intervalo, ya que aunque μ es desconocida, no es una variable aleatoria, sino entraríamos en el campo de la inferencia Bayesiana). Veámoslo gráficamente:

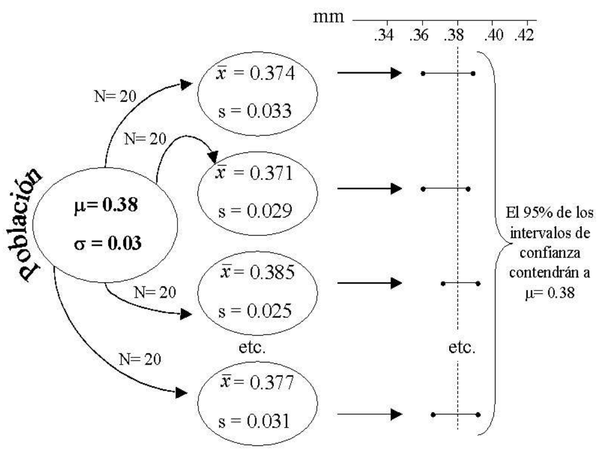

Figura 1.1: El 95 % de los intervalos de confianza contendrán a $\mu = 0.38$. El tamaño muestral considerado cada vez es 20

Si en lugar de 20, el tamaño muestral en cada muestra fuera 5, los intervalos serán más grandes, pero nuevamente el 95 % de los intervalos de confianza contendrán a $\mu = 0.38$, según la siguiente gráfica.

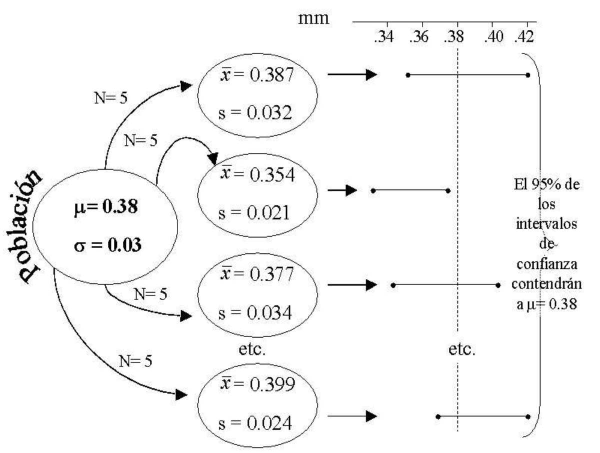

Figura 1.2: El 95 % de los intervalos de confianza contendrán a $\mu = 0.38$. El tamaño muestral considerado cada vez es 5

1.2.1. Estimación puntual

Existen diversos métodos que nos permiten calcular estimadores (estadísticos que se usan para obtener estimaciones puntuales), como son: métodos de máxima verosimilitud, de los momentos, mínimos cuadrados. Nosotros no veremos cómo conseguirlos, ni tampoco en qué propiedades (por ejemplo: si es sesgado o no, eficiencia máxima, consistencia) nos podríamos fijar para elegir un buen estimador. En la bibliografía pueden encontrarse, y en tutorías, por supuesto. Veamos simplemente cómo estimar ciertos parámetros de determinadas distribuciones:

i) Estimador puntual de p, para una Binomial(n,p), n conocido:

$\hat{p} = \frac{X}{n}$ donde X es el número de éxitos que ocurren en las n observaciones.

Ejemplo 1.2.: ¿Cuál sería la estimación de p, la cobertura?

ii) Estimador puntual de μ, para una Normal(μ,σ^2):

$\hat{\mu} = \overline{X}$.

Ejemplo 1.1.: Hacemos el muestreo y $\overline{x} = 65\ ms$

iii) Estimador puntual de σ^2, para una Normal(μ,σ^2):

$S^2 = \frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{N-1}$.

Ejemplo 1.1.: Hacemos el muestreo y $s^2 = 36\ ms^2$

Si en lugar de dividir por $N-1$, hubiésemos dividido por N, habríamos obtenido un estimador sesgado, o sea, $E(S^2) = \sigma^2$, mientras que $E(\frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{N}) = (N-1/N)\sigma^2$.

iv) Estimador puntual del parámetro λ de una Poisson:

$\hat{\lambda} = \overline{X}$.

1.3. Estimación por intervalos

A lo largo de este apartado N denotará el tamaño muestral y α el nivel de significación.

A) Intervalo de confianza para μ, con σ^2 conocida:

[Nota: deducción de los intervalos de confianza, para el resto de casos se haría análogamente:

Sea X_1, X_2, ..., X_N una muestra aleatoria de una población con media μ desconocida y σ^2 conocida. $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{N}}$ es aproximadamente N(0,1) si N es grande (por el teorema central del límite).

Por tanto, P($-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$) = 1 - α, donde $z_{\alpha/2}$ es tal que P($Z \geq z_{\alpha/2}$) = $\alpha/2$.

Por ejemplo, para $\alpha = 0.05$, P($Z \geq 1.96$) = 0.05/2 = 0.025 y P($-1.96 \leq Z \leq 1.96$) = 0.95 $\rightarrow$

P($-1.96 \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{N}} \leq 1.96$) = 0.95 $\rightarrow$ P($-1.96\cdot\sigma/\sqrt{N} \leq \overline{X} - \mu \leq 1.96\cdot\sigma/\sqrt{N}$) = 0.95 $\rightarrow$

P($-1.96\cdot\sigma/\sqrt{N}$ - $\overline{X}$ $\leq$ - $\mu \leq 1.96\cdot\sigma/\sqrt{N}$- $\overline{X}$) = 0.95 $\rightarrow$ P($\overline{X} + 1.96\cdot\sigma/\sqrt{N} \geq \mu \geq \overline{X}$ - $1.96\cdot\sigma/\sqrt{N}$) = 0.95 $\rightarrow$ P($\overline{X}$ - $1.96\cdot\sigma/\sqrt{N} \leq \mu \leq \overline{X} + 1.96\cdot\sigma/\sqrt{N}$) = 0.95].

($\overline{x}$ - $z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$), $\overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$) con P(Z $\geq z_{\alpha/2}$) = $\alpha/2$, $Z \sim$ N(0,1)

B) Intervalo de confianza para μ, con σ^2 desconocida, para Normales:

($\overline{x}$ - $t_{\alpha/2}\frac{s}{\sqrt{N}}$), $\overline{x} + t_{\alpha/2}\frac{s}{\sqrt{N}}$) con P(T $\geq t_{\alpha/2}$) = $\alpha/2$, T es t- Student con $N-1$ grados de libertad

Ejemplo 1.5.: Para acortar el tiempo de transmisión de ciertas imágenes digitales, éstas se comprimen. Se quiere conocer el tiempo de transmisión medio para un algoritmo concreto de compresión, por ello se ha medido el tiempo para 15 imágenes, que consideraremos normal, dando $\bar{x} = 2.35$ segundos y $s = 0.32$ segundos. Encuentra el intervalo de confianza de 95 % para el tiempo medio de transmisión.

C) Intervalo de confianza para μ, con σ^2 desconocida y N grande ($N \geq 30$):

($\overline{x}$ - $z_{\alpha/2}\frac{s}{\sqrt{N}}$), $\overline{x} + z_{\alpha/2}\frac{s}{\sqrt{N}}$) con P(Z $\geq z_{\alpha/2}$) = $\alpha/2$, $Z \sim$ N(0,1)

Observación: *Aun cuando la normalidad no pueda suponerse, si deseamos obtener un intervalo de confianza para μ con la varianza desconocida, si la muestra es grande, podemos usar C). Si la muestra es pequeña, usaremos B) si la distribución es normal.*

Ejemplo 1.1.: Calcula un intervalo de confianza al 95 % para el tiempo medio de ejecución.

[Nota: $z_{0,1} = 1.28$, $z_{0,05} = 1.64$, $z_{0,025} = 1.96$, $z_{0,01} = 2.33$, $z_{0,005} = 2.56$].

[Fíjate que $z_{\alpha/2}$ cumple: P(Z $\geq z_{\alpha/2}$) $= \alpha/2$, $Z \sim$ N(0,1), es decir, la probabilidad que la variable Z sea mayor que $z_{\alpha/2}$ es $\alpha/2$].

Para determinar el tamaño muestral necesario para una precisión determinada, en el caso de la estimación de la media μ a partir de una muestra aleatoria simple, usaremos:

$$N = (\frac{z_{\alpha/2} \cdot \sigma}{Error})^2$$

Cuando σ es desconocida, podemos recurrir a estudios previos o bien a la obtención de una muestra piloto previa, con la que estimaremos σ, mediante s.

Ejemplo 1.5.: ¿Qué tamaño de muestra necesitaríamos si queremos tener 95 % de confianza de que nuestra estimación de μ difiera menos de 0.05? Utiliza que por estudios previos podemos estimar σ por 0.3.

A veces, el interés no está en la estimación de parámetros, sino en *dónde caen las observaciones individuales*. Así pues, debemos distinguir entre intervalos de confianza e intervalos de tolerancia. Para una distribución Normal con media y varianza desconocidas, los límites de tolerancia están dados por $\overline{x} \pm ks$, donde k está determinado de modo que se pueda establecer con una confianza del 100(1 - α) por ciento que los límites contienen al menos una proporción p de la población. En Montgomery (por ejemplo), puedes encontrar las tablas que proporcionan k, con las que calcular estos intervalos de tolerancia.

D) Intervalo de confianza para la diferencia de medias μ_1 - μ_2, con σ_1^2 y σ_2^2 conocidas, para muestras aleatorias independientes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\overline{x}_1 - \overline{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}})$ con $P(Z \geq z_{\alpha/2}) = \alpha/2$, $Z \sim$ N(0,1)

E) Intervalo de confianza para la diferencia de medias μ_1 - μ_2, con σ_1^2 y σ_2^2 desconocidas, para muestras aleatorias independientes y tamaños muestrales grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\overline{x}_1 - \overline{x}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}})$ con $P(Z \geq z_{\alpha/2}) = \alpha/2$, $Z \sim$ N(0,1)

Ejemplo 1.6.: Un ordenador paralelo consiste en elementos de procesamiento, PE, autónomos que comparten una memoria central. Investigadores de la New York University diseñaron una de estas máquinas llamada Ultracomptudora NYU. A fin de evaluar el impacto del retardo introducido por la red sobre el desempeño global de la computadora, los investigadores simularon el tiempo de acceso a la memoria central para instrucciones típicas ejecutadas por una versión paralela de un programa meteorológico de la NASA. Se simularon dos conjuntos de tiempos de acceso, uno procesando con 16 elementos de procesamiento y el otro con 48. Con 16 PE, el tiempo medio de acceso a la memoria central fue de 8.94 microsegundos, mientras que con 48 PE fue de 8.83. Supongamos (información que no se proporcionó) que se simularon 1000 instrucciones para cada una de las dos posibilidades, con desviaciones típicas 3.1 y 3.5 respectivamente. ¿Hay diferencias entre los tiempos de acceso a la memoria central de las instrucciones procesadas con 16 PE y 48 PE? (usemos $\alpha = 0.05$).

[Para el caso de una diferencia entre dos medias, la interpretación del intervalo de confianza

puede extenderse a una compración de las dos medias. De esta manera, por ejemplo, si tenemos gran confianza de que una diferencia μ_1 - μ_2 es positiva, realmente inferiremos que $\mu_1 > \mu_2$ con poco riesgo de caer en un error. Por tanto, en la interpretación de los intervalos de confianza para diferencia de medias nos fijaremos si el cero pertenece al intervalo o no].

F) Intervalo de confianza para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas poblacionales desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2$) (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

($\overline{x}_1$ - $\overline{x}_2 \pm t_{\alpha/2}\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}}\sqrt{\frac{N_1+N_2}{N_1N_2}}$) con P($T \geq t_{\alpha/2}$) = $\alpha/2$, T es t-Student con $N_1 + N_2 - 2$ grados de libertad

Ejemplo 1.7.: Definimos el tiempo de respuesta de un ordenador como el tiempo que un usuario debe esperar mientras el ordenador accede a la información guardada en el disco. Supongamos que un centro de datos desea comparar los tiempos de respuesta medios de sus dos unidades de disco. Se seleccionaron muestras aleatorias independientes de 13 tiempos de respuesta para el disco 1 y 15 tiempos de respuesta para el disco 2; los datos registrados fueron (en milisegundos):

Disco 1: 59, 92, 54, 102, 73, 60, 73, 75, 74, 84, 47, 33, 61 $\rightarrow$ $\overline{x}_1 = 68.2$ $s_1 = 18.6$

Disco 2: 71, 38, 47, 53, 63, 48, 41, 68, 40, 60, 44, 39, 34, 75, 86 $\rightarrow$ $\overline{x}_2 = 53.8$ $s_2 = 15.8$

Calcula el intervalo de confianza de la diferencia de medias al 95 %, asumiendo normalidad e igualdad de varianzas (lo comprobaremos en un apartado posterior). ¿Podemos suponer igualdad de medias poblacionales?

G) Intervalo de confianza para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas poblacionales σ_1^2, σ_2^2 desconocidas y desiguales (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\overline{x}_1$ - $\overline{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}})$ con P$(T \geq t_{\alpha/2}) = \alpha/2$, T es t-student con $\frac{(\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2})^2}{\frac{(s_1^2/N_1)^2}{N_1-1}+\frac{(s_2^2/N_2)^2}{N_2-1}}$ grados de libertad

Ejemplo 1.8.: Un fabricante de ordenadores está desarrollando un nuevo modelo de monitor en color, para lo cual puede utilizar dos tipos de esquemas transistorizados. El fabricante selecciona una muestra de esquemas transistorizados del primer tipo, de tamaño 12, y otra del segundo de tamaño 11. Los datos muestrales respecto a la vida de cada esquema son los siguientes:

Esquma 1: $\overline{x}_1 = 1400$ horas $s_1 = 30$ horas

Esquema 2: $\overline{x}_2 = 1500$ horas $s_2 = 17$ horas

Determina el intervalo de confianza de la diferencia de las medias al 95 % suponiendo que los gastos son normales y $\sigma_1^2 \neq \sigma_2^2$.

H) Intervalo de confianza para la diferencia de medias μ_1 - μ_2 para muestras apareadas. Hay veces que las muestras no son independientes. Pueden ser apareadas como es el caso de tener datos del tipo "antes" y "después", o bien si a cada objeto (u objetos emparejados) se le aplican dos métodos.

$(\overline{d} \pm t_{\alpha/2}\ \frac{s_d}{\sqrt{N}})$ donde $\overline{d}$ es la media de las diferencias y s_d es la desviación típica de las diferencias. Además, P$(T \geq t_{\alpha/2}) = \alpha/2$, T es t-Student con N - 1 grados de libertad, N es el número de objetos (parejas) de que disponemos

Ejemplo 1.9.: Se está investigando la utilidad de dos lenguajes diferentes para mejorar la rapidez de programación. Doce programadores expertos familiarizados con ambos lenguajes, programaron una función estándar en ambos lenguajes y el tiempo que tardaron en minutos fue:

	LENGUAJE 1	LENGUAJE 2	DIFERENCIA (Lenguaje 1 - Lenguaje 2)
Programador 1	17	18	-1
Programador 2	16	14	2
Programador 3	21	19	2
Programador 4	14	11	3
Programador 5	18	23	-5
Programador 6	24	21	3
Programador 7	16	10	6
Programador 8	14	13	1
Programador 9	21	19	2
Programador 10	23	24	-1
Programador 11	13	15	-2
Programador 12	18	20	-2

$\downarrow$

$\bar{d} =$

$s_d =$

Construyamos el intervalo de confianza de la diferencia de medias al 95 %, para comprobar si existe diferencia entre los dos lenguajes.

I) Intervalo de confianza para σ^2 en una población normal:

$(\frac{(N-1)s^2}{\chi^2_{\alpha/2}}, \frac{(N-1)s^2}{\chi^2_{1-\alpha/2}})$ con $P(\chi^2 > \chi^2_{\alpha/2}) = \alpha/2$, χ^2 es chi- cuadrado con $N-1$ grados de libertad

Ejemplo 1.10.: Una queja frecuente de los usuarios de un sistema de ordenadores en red es la gran varianza del tiempo de respuesta. Una importante empresa está pensando en instalar una nueva red entre sus directivos. Con objeto de estudiar el tiempo de respuesta, se observa una muestra aleatoria de 30 tiempos, obteniéndose una varianza muestral de $25ms^2$. Construye un intervalo de confianza de 99 % para la desviación típica de la población muestreada (asume condiciones de normalidad).

J) Intervalo de confianza para el cociente σ_1^2/σ_2^2 de varianzas de dos poblaciones normales independientes:

$(\frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2}}, \frac{s_1^2}{s_2^2}\frac{1}{F_{1-\alpha/2}})$ donde P(F $> F_{\alpha/2}) = \alpha/2$ y F es F de Sndecor con $(N_1-1,\, N_2-1)$ grados de libertad

Ejemplo 1.7.: Construye un intervalo de confianza al 95 % para el cociente de ambas varianzas. ¿Fue apropiado suponer igualdad de varianzas?

[En la interpretación de los intervalos de confianza para cociente de varianzas nos fijaremos si el uno pertenece al intervalo o no].

K) Intervalo de confianza para una proporción p (de una Binomial) cuando N es grande y la proporción no es cercana a cero:

$(\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N}})$, donde P($Z > z_{\alpha/2}) = \alpha/2$ $Z \sim$ N(0,1) y $\hat{p}$ = X /N, $\hat{q}$ = 1 - $\hat{p}$, X = número de éxitos

Ejemplo 1.2.: Calcula un intervalo de confianza al 95 % para p.

La magnitud del error que cometemos al emplear X/N como estimador de p, viene dada por: E = Error = $z_{\alpha/2}\sqrt{\frac{p(1-p)}{N}}$. Esta fórmula nos puede servir para determinar el tamaño muestral necesario para alcanzar un grado de precisión deseado.

$N = p(1-p) \cdot (\frac{z_{\alpha/2}}{E})^2$

Si no dispusiésemos de información acerca del valor de p, por ejemplo en base a una muestra piloto:

$N = p(1-p) \cdot (\frac{z_{\alpha/2}}{E})^2 \leq \frac{1}{4}\ (\frac{z_{\alpha/2}}{E})^2$

Una vez obtenidos los N datos, tendremos un $(1-\alpha)100\,\%$ de confianza que el error no excederá E.

Ejemplo 1.11.: Se está probando la actividad de la CPU. Para la i-ésima observación, consideremos $X_i = 0$ si la CPU no está ocupada y 1 en caso contrario. Supongamos que las observaciones están suficientemente separadas en el tiempo para considerarlas independientes y asumamos qur X es Bernoulli con parámetro p. Queremos estimar la utilización esperada p y deseamos estar al menos 95 % seguros que el error es como mucho de 0.04. ¿Cómo ha de ser de grande la muestra si:

a) no tenemos idea de cuál pueda ser la proporción real?

b) por estudios previos, una estimación preliminar sería 0.8?

L) Intervalo de confianza para una proporción p, si ésta es muy cercana a cero:

$(0, \frac{1}{2N}\chi^2_\alpha)$ con $P(\chi^2 > \chi^2_\alpha) = \alpha$, χ^2 es chi- cuadrado con $2(X+1)$ grados de libertad, $X =$ número de éxitos

✍ **Ejemplo 1.12.:** Durante un mes, se usaron continuamente 2000 componentes y de ellas 4 fallaron. Calcula un intervalo de confianza al 99 % para la probabilidad de que un componente falle en las condiciones establecidas.

M) Intervalo de confianza para la diferencia de dos proporciones, con N_1 y N_2 grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\hat{p_1}$ - $\hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}\hat{q_1}}{N_1} + \frac{\hat{p_2}\hat{q_2}}{N_2}})$, donde P($Z > z_{\alpha/2}) = \alpha/2$ $Z \sim$ N(0,1), $\hat{p_1} = X_1 /N_1$, $\hat{q_1} = 1$ - $\hat{p_1}$, X_1 = número de éxitos en las N_1 pruebas y $\hat{p_2} = X_2 /N_2$, $\hat{q_2} = 1$ - $\hat{p_2}$, X_2 = número de éxitos en las N_2 pruebas

✍ **Ejemplo 1.13.:** Una de las típicas aplicaciones de un sistema experto (sistema informático capaz de resolver problemas partiendo de una base de conocimientos de expertos y una serie de reglas de deducción) es ayudar a aislar un fallo en algún sistema mecánico, electrónico o fisiológico. Supongamos que dos sistemas expertos, A y B, se han desarrollado para analizar datos cardiovasculares y prestar su opinión sobre la condición del paciente. Para comparar los dos sistemas, se introducen datos de 40 condiciones seleccionadas aleatoriamente. Sean X e Y las variables aleatorias correspondientes a los diagnósticos correctos de los sistemas A y B, respectivamente y p_1, p_2 las probabilidades respectivas de éxito. Supongamos que el sistema A realiza 35 diagnósticos correctos, mientras que B 30. Desea determinarse si los sistemas tienen distinto nivel de efectividad, así que determina el intervalo de confianza al 95 % para la diferencia de proporciones.

[En la interpretación de los intervalos de confianza para diferencia de proporciones nos fijaremos si el cero pertenece al intervalo o no].

Problemas del tema 1

En reprografía, podrás encontrar más problemas resueltos.

1. Se sabe que la duración, en horas, de un foco de 75 watts tiene una distribución aproximadamente Normal. Al tomar una muestra aleatoria de 12 focos, se tiene una duración promedio $\overline{x} = 3250$ horas y una desviación $s = \sqrt{1000}$ horas.

   a) Construye un intervalo de confianza del 95 % para la duración promedio

   b) Supóngase que se desea una confianza del 95 % en que el error en la estimación de la duración promedio sea menor que 5 horas. ¿Qué tamaño de muestra debe emplearse para este fin, si por estudios previos se sabe que $\sigma^2 = 990$ horas?

   (Sol. : (3229.9077,3270.09), 153)

2. En un estudio sobre la efectividad del ejercicio físico para la reducción de peso, un grupo de 9 personas participaron en un programa prescrito de ejercicio físico durante 1 mes, obteniéndose:

Persona	Peso antes (Kg.)	Peso después(Kg.)
1	105	98
2	89	86
3	84	85
4	106	105
5	90	88
6	96	93
7	79	75
8	90	85
9	100	96

   Usando el nivel $\alpha = 0.01$, calcula el intervalo de confianza de la diferencia de medias e interprétalo.

   (Sol. : (0.52,5.7), como no contiene al cero sí existe diferencia de peso).

3. Se ha realizado un experimento para comparar las economías en combustible para dos tipos de camiones diesel equipados de forma similar. Se han usado 12 camiones de la marca A y 10 de la marca B en pruebas de velocidad constante de 90 km/h. Si los de la marca A

promedian 16 kilómetros por litro con una desviación estándar de 1 kilómetro por litro y los de la marca B promedian 11 kilómetros por litro con una desviación estándar de 0.8 kilómetros por litro. Calcula un intervalo de confianza al 95 % para la diferencia de medias y determina (razonando porqué) si existe diferencia en el consumo entre estas dos marcas de camiones. (Supón normalidad e igualdad de varianzas).

(Sol. : (4.182, 5.817), existe diferencia porque 0 no pertence al intervalo)

4. En una muestra aleatoria de 500 familias que tienen televisores en una cierta ciudad, se encuentra que 340 están suscritas a un cierto canal. Encuentra un intervalo de confianza de 95 % para la proporción real de familias en esta ciudad suscritas al canal. Determina también el tamaño muestral necesario si queremos tener una confianza de al menos 95 % de que nuestra estimación de p está dentro de 0.02, primero asumiendo la muestra anterior como una muestra preliminar que nos proporciona una primera estimación y en segundo lugar, sin contar con esta información.

   (Sol. : (0.64,0.72), 2090, 2401)

5. Se considera cierto cambio en un proceso de fabricación de partes de componentes. Se toman muestras del procedimiento existente y del nuevo para determinar si éste tiene como resultado una mejoría. Se encuentra que 75 de 1500 artículos del procedimiento actual son defectuosos y 80 de 2000 artículos del procedimiento nuevo también lo son. Encuentra un intervalo de confianza de 90 % para la diferencia real en la fabricación de defectuosos entre el proceso actual y el nuevo, e interprétalo.

   (Sol. : (-0.0017,0.0217), como contiene al cero, no hay razón para creer que el nuevo procedimiento producirá una disminución significativa en la producción de artículos defectuosos comparado con el método existente.)

6. Se investiga la resistencia a la tensión de ruptura de hilo proporcionado por dos fabricantes. Tomamos una muestra de 50 especímenes de prueba provenientes de cada fabricante, obteniéndose como resultados $\overline{x}_1 = 88$ psi y $\overline{x}_2 = 90$ psi con desviaciones respectivas 5 psi y 4 psi. Calcula un intervalo de confianza al 95 % para la diferencia entre las medias de la tensión de ruptura e interprétalo.

   (Sol. : (-3.775,-0.225), como $0 \notin$ al intervalo, existirá diferencia en cuanto a resistencia de los hilos entre ambos fabricantes)

7. Un fabricante de monitores prueba dos diseños de microcircuitos para determinar si producen un flujo de corriente equivalente. Los datos obtenidos son:

Diseño 1:	$n_1 = 21$	$\overline{x}_1 = 24.2$	$s_1^2 = 8$

Diseño 2: $n_2 = 10$ $\overline{x}_2 = 23.9$ $s_2^2 = 25$

Determina si las varianzas son iguales ($\alpha = 0.05$) y tras ello calcula el intervalo de confianza al 95 % correspondiente para la diferencia de medias e interprétalo.

(Sol. : 1 $\notin$ (0.08719,0.9088), con lo cual, asumiremos varianzas distintas. 0 $\in$ (-3.398,3.998), con lo cual no hay razones para asumir flujos medios diferentes.)

8. Las concentraciones de zinc que se sacan del agua en 7 sitios diferentes son: 2.5, 2.4, 2.6, 2.65, 2.76, 2.8, 2.71 gramos por mililitro. Encuentra el intervalo de confianza de 95 % para la concentración media de zinc en el río. ¿Qué tamaño de muestra necesitaríamos si queremos tener 95 % de confianza de que nuestra estimación de μ difiera menos de 0.05? Utiliza que por estudios previos podemos estimar σ por 0.3.

   (Sol. : (2.5,2.76); 139)

9. Además de las especificaciones de peso y perímetro, la FIFA ha estipulado que los balones deben botar 0.5 m de altura cuando se dejan caer a cierta altura. Una empresa juguetera desea estudiar la altura del bote de los balones producidos, para comprobar que la transición diseño a producción en masa se ha llevado a cabo con éxito (diseñar un buen producto y construir prototipos que funcionen es una cosa, otra cosa es la transferencia del diseño a la manufactura). Podríamos obtener una muestra aleatoria de por ejemplo tamaño 60. Calcula un intervalo de confianza al 95 % para la altura media del bote, sabiendo que la media muestral ha sido $\bar{x} = 0.51$m y la varianza muestral es $s^2 = 0.01$.

   (Sol. : (0.485,0.535))

10. Se lleva cabo un expermiento en que se comparan dos tipos de motores, A y B. Se mide el rendimiento en millas por galón de gasolina. Se realizan 50 experimentos con el motor A y 75 con el B. La gasolina que se utiliza y las demás condiciones se mantienen constantes. El rendimiento promedio de gasolina para el motor A es de 36 millas por galón con desviación típica 6, el promedio para el motor B es 42 millas por galón y desviación típica 8. Calcula el intervalo de confianza de 99 % sobre μ_A - μ_B, donde μ_A y μ_B son el rendimiento de gasolina medio poblacional para los motores A y B respectivamente. ¿Podemos suponer que ambas medias poblacionales son iguales?

    (Sol. : (-9.211,-2.789), como no contiene al cero sí existe diferencia).

11. Se desea conocer si dos aleaciones de alumnio tienen o no igual resistencia. Para ello se midió la resistencia a la compresión de 58 especímenes del primer tipo y 27 del segundo, obteniéndose $\overline{x}_1 = 70.7$ y $\overline{x}_2 = 76.13$. Supongamos que se distribuyen normalmente. Sus varianza muestrales son: $s_1^2 = 1{,}8^2$ y $s_2^2 = 2{,}42^2$ (supongamos que $\sigma_1^2 = \sigma_2^2$, compruébalo también, calculando el intervalo de confianza del cociente de varianzas al 95 %). Calcula el intervalo de confianza de la diferencia de medias al 95 %, asumiendo normalidad. ¿Podemos suponer igualdad de medias poblacionales?

    (Sol. : el intervalo de confianza para el cociente de varianzas al 95 % es: (0.256,1.04), como 1 pertenece al intervalo, no hay razón para afirmar que las varianzas sean distintas;

intervalo de confianza de la diferencia de medias al 95 %: (-6.36,-4.5), no podemos suponer igualdad de medias porque el cero no pertenece al intervalo).

12. Queremos estimar la proporción real de unidades defectuosas en un embarque muy grande de azulejos y deseamos estar al menos 95 % seguros que el error es como mucho de 0.04. ¿Cómo ha de ser de grande la muestra si:

    a) no tenemos idea de cuál pueda ser la proporción real?

    b) por estudios previos, sabemos que la proporción real no excede de 0.12?

    (Sol. : 601;254).

13. Para comparar dos tipos de parachoques, seis de cada tipo se instalaron en unos automóviles. Después éstos se lanzaron contra un muro a 20km/h y se determinaron los gastos de las reparaciones (en euros).

    Parachoques 1: 107, 148, 123, 165, 102, 119 → $\overline{x}_1 = 127.33 \quad s_1^2 = 597.867$

    Parachoques 2: 134, 115, 112, 151, 133, 129 → $\overline{x}_2 = 129 \quad s_2^2 = 202$

    Determina el intervalo de confianza de la diferencia de las medias al 95 % suponiendo que los gastos son normales y $\sigma_1^2 \neq \sigma_2^2$.

    (Sol. : (-28.295,24.955), como contiene al cero no existen evidencias para afirmar que exista diferencia entre las medias).

14. Disponemos de dos básculas y desearíamos comprobar si existe diferencia sistemática entre los pesos obtenidos con ambas básculas. Para ello construiremos el intervalo de confianza de la diferencia de medias al 95 %.

	BÁSCULA 1	BÁSCULA 2
Roca 1	11.23	11.27
Roca 2	14.36	14.41
Roca 3	8.33	8.35
Roca 4	10.50	10.52
Roca 5	23.42	23.41
Roca 6	9.15	9.17
Roca 7	13.47	13.52
Roca 8	6.47	6.46
Roca 9	12.4	12.45
Roca 10	19.38	19.35

    (Sol. : (-0.04,0.00051), como cero pertenece al intervalo, no podemos concluir que exista diferencia entre ambas medias).

15. Cinco medidas del contenido de alquitrán de cierta clase de cigarrillos dieron como resultado: 14.5, 14.2, 14.4, 14.3 y 14.6 mg. por cigarrillos. Construye un intervalo de confianza

de 99 % para la desviación típica de la población muestreada (asume condiciones de normalidad).

(Sol. : (0.082,0.6951)).

16. En un estudio para comparar dos líneas de montaje se encontró que: 16 de 200 tractores de la línea 1 necesitaron grandes ajustes antes de embarcarlos, mientras que 14 de 400 los necesitaron en la línea 2. Determina el intervalo de confianza al 95 % para la diferencia de proporciones.

    (Sol. : (0.0033,0.08669), como no contiene al cero sí existe diferencia entre ambas líneas).

Autoevaluación del tema 1

Las soluciones están en reprografía.

Ejercicio 1: La contaminación de metales pesados de varios ecosistemas es una seria amenaza ambiental, en parte debido a la ‘potencial transferencia de sustancias peligrosas a los seres humanos a través de los alimentos. El artículo científico *"Cadmium, Zinc and Total Mercury Levels in the Tiussues of Several Fish Species from La Plata River Estuary, Argentina"* (*Environmental Monitoring and Assessment, 1993*), recoge diversos datos. Para una muestra de 56 peces de la especie *Mugil liza* la concentración media muestral de zinc en el hígado fue 9.15 $\mu g/g$ y la desviación típica muestral fue 1.27 $\mu g/g$ (no confundirse, aquí μ indica micro). Para la especie *Pogonias cromis* se recogieron 61 concentraciones, dando una media de 3.08 y una desviación típica de 1.71.

- Calcula el intervalo de confianza para el verdadero promedio de la concentración de zinc en el hígado para todos los peces de la especie *Mugil liza* en la localidad especificada, usando un nivel de confianza de 95 %.
- Repitelo ahora para la especie *Pogonias cromis*, pero usando un nivel de confianza de 99 %.
- ¿Por qué razones el intervalo de confianza para la *Pogonias cromis* es más amplio que para *Mugil liza*, pese a que el tamaño muestral era mayor?
- Haz una interpretación del primero de los intervalos de confianza calculados.
- ¿Es necesario que la distribución sea Normal para que los dos intervalos obtenidos sean válidos?

Ejercicio 2: Se comparan dos sistemas según su tiempo de respuesta a un comando. Se han tomado 13 datos para el sistema 1, obteniéndose un tiempo de respuesta medio de 682 milisegundos y una desviación típica de 25 milisegundos. Mientras que para el segundo sistema, se toman 10 datos obteniéndose una media de 675 milisegundos y una desviación típica de 28 milisegundos. Suponiendo que los tiempos son normales:

1. Calcula un intervalo de confianza al 95 % para el cociente de varianzas y determina si existe diferencia.
2. Calcula un intervalo de confianza al 95 % para la diferencia de medias y determina si existe diferencia.
3. Calcula un intervalo de confianza al 95 % para la media del sistema 1.

4. Si deseamos que el error de la estimación de la media anterior sea inferior a 2 milisegundos con una confianza del 95 % y teniendo en cuenta que podríamos asumir $\sigma = 25$, ¿cuál será el tamaño muestral requerido?

Ejercicio 3: Según las siguientes salidas, responde a las dos preguntas:

(a) ¿Podemos suponer que las varianzas de las variables Col_1 y Col_2 son distintas? Razona tu respuesta.

```
Comparison of Standard Deviations
---------------------------------

                    Col_1               Col_2
------------------------------------------------------------
Standard deviation  0,918753            0,798088
 Variance           0,844107            0,636944
 Df                  7                    8

     Ratio of Variances = 1,32524

95,0% Confidence Intervals
     Standard deviation of Col_1: [0,607456;1,86991]
     Standard deviation of Col_2: [0,539074;1,52895]
     Ratio of Variances: [0,292641;6,49282]
```

(b) Escribe el intervalo de confianza al 95 % para la diferencia de medias (según lo obtenido en el apartado anterior). ¿Son diferentes? Razona tu respuesta.

```
 Comparison of Means
-------------------

95,0% confidence interval for mean of Col_1: 6,0125 +/- 0,768099   [5,2444,6,7806]
95,0% confidence interval for mean of Col_2: 7,07778 +/- 0,613466   [6,46431,7,69124]
95,0% confidence interval for the difference between the means
   assuming equal variances: -1,06528 +/- 0,887095   [-1,95237,-0,178183]
   not assuming equal variances: -1,06528 +/- 0,900396   [-1,9578,-0,172754]
```