TEMA IV: ESTIMACION

IV.1. Introducción.
IV.2. Estimación puntual. Métodos y propiedades.
IV.2.1. Concepto y definición.
IV.2.2. Propiedades de los estimadores puntuales.
IV.2.3. Métodos de estimación puntual.
IV.2.3.1. Máxima verosimilitud
IV.3. Estimación por intervalos de confianza.
IV.3.1. Introducción.
IV.3.2. Métodos de construcción de intervalos.
IV.3.3. Intervalos de confianza en poblaciones normales.
IV.3.4. Intervalo de confianza para las proporciones.
IV.3.5. Tamaño muestral para una precisión dada.
IV.4.- Problemas.

IV.1. Introducción.

Antes de comenzar con los contenidos específicos de la estimación es conveniente recordar cuál es la situación habitual con la que nos encontramos en un proceso de inferencia.

Nuestro interés se centra en el estudio de una **característica** determinada de una población. Dicha característica puede denotarse por una variable aleatoria X definida sobre la población y con función de densidad f(x). Si conocemos f(x) tendremos caracterizada a la variable aleatoria X y el problema queda resuelto. Sin embargo, la mayoría de las veces no ocurre eso. En principio, es habitual disponer solo de cierta información a priori sobre el comportamiento de la variable aleatoria X (dicha información, normalmente proviene de las conclusiones de un estudio previo de esa variable realizado sobre la misma población u otra diferente).

La función f(x) puede ser totalmente desconocida o bien conocida solamente en su forma, es decir, sabemos qué tipo de distribución es pero nos falta especificarla porque desconocemos el valor de uno o más parámetros de la misma.

Por ejemplo podemos conocer que la variable aleatoria X sólo puede tomar dos valores, éxito o fracaso, y que además la probabilidad de éxito es muy baja. Según esta información podemos suponer que X se distribuye como una Poisson. No obstante, nos falta información para poder definir esa distribución (función de densidad de la variable aleatoria X) puesto que desconocemos el valor de su media.

El objetivo de la inferencia estadística es resolver este problema. Para ello se han de seguir los siguientes pasos:

1. Obtención de la muestra.
2. Cálculo del estadístico que nos interese (utilizando la información muestral).
3. Utilización del estadístico calculado en el apartado previo como estimador del parámetro poblacional que desconocemos.

De lo anterior se deduce que **un estimador no es más que una función de los datos muestrales que se utiliza como valor de un parámetro poblacional**. O dicho de otra manera, un estimador no es más que un estadístico asociado a un valor poblacional. Por ejemplo, podríamos decir que la media muestral es el estimador de la media poblacional. El estadístico media muestral se calcula a través de la siguiente función

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

Cuando se establece una relación entre este estadístico y un parámetro, en este caso concreto la media poblacional, el primero se convierte en estimador del segundo.

El interrogante que surge a continuación es porqué la media muestral es un estimador de la media poblacional, o más concretamente, ¿**qué propiedades tiene el estadístico media muestral para ser considerado un buen estimador de la media poblacional?.**

En el presente tema nos centraremos en el estudio de los siguientes aspectos:

A. Tipos de estimación. Distinguiendo entre **estimación**

puntual, procedimiento mediante el cual se estima el valor del parámetro poblacional basándose en los datos muestrales mediante el uso de un estádistico, y **estimación por intervalos**, en el cual para realizar la estimación del parámetro poblacional se determina un intervalo en el que de forma probable se encuentra el valor del parámetro. Dicho intervalo recibirá el nombre de intervalo de confianza.

B. **Propiedades** que debemos exigir a un **estimador puntual**.

C. Métodos disponibles para la **construcción de estimadores puntuales**.

D. **Métodos** disponibles para la estimación de **intervalos de confianza** de un parámetro poblacional.

E. Determinación del **tamaño de la muestra para que la estimación presente una determinada precisión**.

IV.2. Estimación puntual. Métodos y propiedades.

IV.2.1. Concepto y definición.

Sea X una variable aleatoria y supongamos que la función de densidad de X, f(x), depende de un cierto número de parámetros $\theta=(\theta_1, ..., \theta_k)$ donde k es un número finito.

El problema de la estimación puntual es encontrar un estadístico a partir de una muestra de la variable aleatoria X que estime los valores de θ. A este estadístico lo hemos llamado **estimador**.

Lógicamente, existen muchos posibles estimadores del parámetro θ por lo que es necesario dar algunas propiedades deseables en los mismos que nos ayuden a decidir la elección del estimador adecuado a cada caso concreto.

VII 2.2. Propiedades de interés en los estimadores puntuales.

Para que un estadístico muestral sea un buen estimador de un parámetro poblacional debe ser **insesgado**, **invariante**, **consistente suficiente**, de **varianza mínima** y **eficiente**. En todo caso, todas estas propiedades serían deseables en un estimador. A continuación se desarrollan algunas de estas características.

IV.2.2.1. Insesgadez.

Sea $\hat{q}$ el estimador puntual del parámetro θ. Diremos que $\hat{q}$ es un estimador insesgado de θ si se verifica que

$$E[\hat{q}] = q$$

El significado de esta propiedad es el siguiente: si de una característica de una población se obtuviesen varias muestras de tamaño n, podríamos obtener un estimador con cada una de ellas y la media de los valores de esos estimadores tendería a coincidir con el valor del parámetro. En el caso límite, es decir, cuando obtuviésemos todas las muestras posibles de tamaño n de la población, el valor de la media de todos los estimadores coincidiría exactamente con el valor del parámetro.

Para ver un ejemplo, supongamos que queremos estimar el valor de la media de una variable aleatoria X. Donde X se corresponde con una característica determinada de una población de tamaño N y queremos demostrar que la media muestral es un estimador insesgado de la media poblacional. Para ello debemos seleccionar una muestra de tamaño n, (X_1, X_2, ..., X_n), representativa de la población en estudio. La media muestral se calcula mediante la

expresión ya conocida

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

Llamaremos:

θ: parámetro poblacional que queremos estimar. En este ejemplo $\theta = m_x$ = media de x

$\hat{q}$: estimador del parámetro poblacional que se calcula a partir de los datos muestrales.

En este caso concreto, el estadístico media muestral es el estimador del parámetro media poblacional, es decir, $\hat{q} = \bar{x}$

Para que el estimador cumpla la propiedad de **insesgadez** debe suceder que la esperanza del estadístico muestral (o lo que es lo mismo, la esperanza del estimador) coincida con el valor del parámetro poblacional.

$$E[\hat{q}] = q$$

Esta condición aplicada a nuestro ejemplo se convierte en

$$E[\bar{x}] = m_x$$

Si desarrollamos esa expresión

$$E[\bar{x}] = E[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}E[\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\, n\, m_x = m_x$$

Con esto hemos demostrado que realmente la media muestral es un estimador insesgado de la media poblacional.

En algunos casos podemos encontrarnos con estimadores que no cumplen esta propiedad, son los denominados **estimadores sesgados**. Diremos que $\hat{q}$ es un estimador sesgado de q si se verifica que

$$E[\hat{q}] = q + b(q)$$

donde $b(q)$ es el denominado sesgo, excentricidad o error sistemático.

IV.2.2.2. Consistencia.

Se dice que un estimador es consistente cuando al incrementarse el tamaño de la muestra, el valor del estimador tiende al valor del parámetro. Es decir, $\hat{q}$ (X_1, X_2, ..., X_n), (esta nomenclatura indica que el estimador se obtiene con los datos de la muestra X_1, X_2, ..., X_n) es un estimador consistente de q si se cumple

$$\lim_{n \to \infty} \hat{q}(x_1, x_2, \ldots, x_n) = q$$

Es decir, un estimador es consistente cuando al incrementar el tamaño de la muestra el estimador coincide con el parámetro. Por ejemplo, la media muestral es un estimador consistente de la media poblacional. La demostración es inmediata. La media poblacional es la suma de todos los valores de X_i dividido por el tamaño de la población. Por otra parte, la media muestral es lo mismo pero en vez de sumar todos los valores de la población, solo sumamos los n elementos de la muestra. Si la muestra es muy grande significa que el valor de n tiende a ser igual a N, tamaño de la población. Con lo cual las dos expresiones, la de la media de la muestra y la de la media d la población, son iguales. Por tanto, la media muestral es un estimador consiste de la media poblacional.

IV.2.2.3. Varianza mínima.

Sean $\hat{q}_1 \hat{q}_2, .., \hat{q}_k$ todos los posibles estimadores del parámetro q tal que todos ellos tienen la misma media. Es decir,

$$m_{\hat{q}_1} = m_{\hat{q}_2} = \ldots = m_{\hat{q}_k}$$

Llamaremos estimador de varianza mínima a aquel estimador que tenga menor varianza. Es evidente que esta es una propiedad deseable sobre todo el estimador es insesgado. La razón de ello se fundamenta en el propio concepto de varianza, sobre la cual podemos justificar que al tener menos varianza, la distribución del estimador es menos dispersa, con lo cual el error que cometemos en el proceso de inferencia es más pequeño.

5.2.2.4. Eficiencia.

Se dice que un estimador, $\hat{q}_1$, es más eficiente que otro, $\hat{q}_2$, si se verifica que la varianza del primero es menor o igual que la del segundo para cualquier tamaño muestral. Es decir, $\hat{q}_1$ es más eficiente que $\hat{q}_2$ si y sólo si

$$s^2_{\hat{q}_1} \leq s^2_{\hat{q}_2}$$

Para cualquier valor de n.

Por último se dice que un estimador posee propiedades asintóticas cuando las propiedades mencionadas anteriormente se cumplen si el tamaño de la muestra tiende a infinito. Así se puede hablar de estimadores asintóticamente insesgados, asintóticamente eficientes, etc...

IV.2.3. Métodos de estimación puntual.

Una vez visto el problema que se plantea en cada proceso de estimación, en el sentido de las propiedades que un estimador debe reunir para que se le pueda considerar como

'bueno', queda por resolver la cuestión de cómo proceder a la determinación de los estimadores. Los métodos más conocidos y usados de estimación son el **método de los momentos**, el **método de los mínimos cuadrados** y el **método de la máxima verosimilitud**. Dentro de este apartado sólo desarrollaremos este último, dejando el de mínimos cuadrados para los dos últimos temas de la asignatura.

IV.2.3.2. El método de la máxima verosimilitud.

Antes de empezar con la explicación formal de este método es conveniente que recordemos el concepto de **función de densidad o cuantía conjunta**.

Sean X e Y dos variables aleatorias. Diremos que f(xy) es la función de densidad a de cuantía conjunta si nos da la probabilidad de que suceda cada posible par de modalidades (xy). En el caso e que el alumno no recuerde este concepto le remitimos al tema de variables aleatorias bi-dimensionales de la asignatura Estadística I.

Sea X una variable aleatoria que se distribuye con una función de densidad $f(x; q)$, siendo q uno de los parámetros de los que depende la función de densidad. Sea q desconocido.

Sea X_1, X_2, ..., X_n una muestra aleatoria simple donde X_1 representa al primer elemento de la muestra, X_2 al segundo y así sucesivamente. Obsérvese que hasta ahora hablamos de una muestra genérica de tamaño n y no de una realización muestral. Ello implica que dado que X_1 es cualquier elemento de la población su función de densidad, $f(X_1)$, coincidirá con la función de densidad de la variable aleatoria x, f(X).

Llamaremos distribución conjunta de la muestra y la denotaremos por $f(X_1, X_2, ..., X_n)$ a aquella función que nos da la probabilidad de que se obtenga una muestra determinada.

El método de máxima verosimilitud estima el valor del parámetro q, es decir, $\hat{q}$, como aquel valor que hace que la probabilidad de haber obtenido una determinada realización muestral sea la máxima posible.

Para una muestra obtenida de tamaño n, $(x_1, x_2, ..., x_n)$ se define la **función de verosimilitud,** y que se denota por $L(x_1, x_2, ..., x_n; q)$ a:

$$L(x_1, x_2, ..., x_n; q) = f(x_1, x_2, ..., x_n; q)$$

El método de máxima verosimilitud consiste en elegir como estimador del valor del parámetro poblacional q a aquel que hace máxima la función de verosimilitud.

Una de las ventajas de los métodos de estimación tales como el de máxima verosimilitud es que las propiedades de los estimadores obtenidos mediante los mismos se pueden demostrar de forma genérica, cumpliéndose por tanto par cada caso particular.

Se puede demostrar que los estimadores obtenidos mediante el método de máxima verosimilitud cumplen las siguientes propiedades.

1. Son asintóticamente insesgados. Es decir, al incrementar el tamaño muestral, la media del estimador tiende a ser igual al valor del parámetro.
2. Son asíntóticamente normales. Es decir, la distribución del estimador tiende a tener un comportamiento de

distribución Normal.

3. Son asintóticamente de varianza mínima. Es decir, tiende a ser eficiente.

Deducción de algunos estimadores máximo verosímiles en el muestreo aleatorio simple para poblaciones normales.

A.- Estimación máximo verosímil de la media en poblaciones normales.

Sea X una variable aleatoria normal (m, s^2) en donde desconocemos el valor de la media y de la varianza.

Sea (X_1, X_2, ..., X_n) una muestra aleatoria simple. Su función de verosimilitud se denota L(X_1, X_2, ..., X_n; m,s^2). Como en la muestra aleatoria simple X_1, X_2, ..., X_n son independientes, la función de verosimilitud conjunta se expresa como el producto de las componentes marginales.

$$L(X_1, X_2, ..., X_n; m, s^2) = f(X_1; m, s^2) * f(X_2; m, s^2) * ... * f(X_n; m, s^2)$$

Por seguir la variable aleatoria X una distribución normal de media m y varianza s^2 su función de densidad es la que aparece a continuación

$$f(x) = \frac{1}{\sqrt{2ps}} e^{-\frac{1}{2}\frac{x-m^2}{s^2}}$$

Entonces

$$f(X_1;m,s^2)*f(X_2;m,s^2)*\ldots*f(X_n;m,s^2)=$$
$$\frac{1}{\sqrt{2p}\,s}e^{\frac{-1}{2}\frac{(x_1-m)^2}{s^2}}*\ldots*\frac{1}{\sqrt{2p}\,s}e^{\frac{-1}{2}\frac{(x_n-m)^2}{s^2}}=$$
$$=\left(\frac{1}{\sqrt{2p}\,s}\right)^n e^{\frac{-1}{2s^2}\sum_{i=1}^{n}(x_i-m)^2}$$

En resumen

$$L=\left(\frac{1}{2p\,s^2}\right)^{\frac{n}{2}} e^{\frac{-1}{2s^2}\sum_{i=1}^{n}(x_i-m)^2}$$

Si aplicamos logaritmos neperianos a L nos queda

$$\ln L=\frac{n}{2}\ln\left(\frac{1}{2p\,s^2}\right)+\frac{-1}{2s^2}\sum_{i=1}^{n}(x_i-m)^2\ln e$$
$$=\frac{n}{2}\ln 1-\frac{n}{2}\ln 2p\,s^2-\frac{1}{2s^2}\sum_{i=1}^{n}(x_i-m)^2=$$
$$=-\frac{n}{2}\ln 2p-\frac{n}{2}\ln s^2-\frac{1}{2s^2}\sum_{i=1}^{n}(x_i-m)^2$$

expresión esta última más fácil para derivar y, que, dadas las propiedades de la función logaritmo, sabemos que alcanza los máximos y los mínimos en los mismos puntos que la función L.

Recordemos que para calcular el máximo de una función debemos derivar dicha función e igualarla a cero, debiendo ser su seguda derivada negativa.

En consecuencia, para calcular el estimador máximo verosímil de la media tenemos que calcular la primera derivada del logaritmo de la función de verosimilitud con respecto a la media e igualarla a cero.

$$\frac{d \ln L}{d\mu} = \frac{1}{2\sigma^2} 2 \sum_{i=1}^{n} (x_i - \hat{\mu}) = 0$$

$$\sum_{i=1}^{n} (x_i - \hat{\mu}) = 0$$

$$\sum_{i=1}^{n} x_i = \hat{\mu} \sum_{i=1}^{n} 1$$

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

Según este resultado, el estimador máximo verosímil de la media poblacional de una población normal es la media muestral.

B.- Estimación máximo verosímil de la varianza en

poblaciones normales.

Recordemos que hemos llegado al siguiente resultado

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Para calcular el estimador máximo verosímil de la varianza debemos calcular la primera derivada del logaritmo neperiano de la función de verosimilitud con respecto a la varianza e igualar el resultado a cero.

$$\frac{d\ln L}{d\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\frac{1}{\sigma^4} = 0$$

$$\frac{1}{\hat{\sigma}^2}\left[-\frac{n}{2} + \frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\frac{1}{\hat{\sigma}^2}\right] = 0$$

$$\frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\frac{1}{\hat{\sigma}^2} = \frac{n}{2}$$

$$\hat{\sigma}^2 = \frac{\left[\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right]}{n}$$

El estimador máximo verosímil de de la varianza poblacional de una distribución normal es la varianza muestral.

Como el valor de la media poblacional es desconocido, lo sustituimos por la media muestral que es su estimador máximo verosímil como ya hemos demostrado.

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

Estimación de proporciones.

Sea una población de tamaño N en la cual se quiere estudiar si sus elementos cumplen una determinada condición o no. Llamamos proporción poblacional a

$$P = \frac{k}{N}$$

donde:

k = número de elementos de la población que cumplen la condición que estamos analizando.

N = tamaño de la población.

Sea (X_1, X_2, ..., X_n) una muestra aleatoria simple de individuos pertenecientes a la población en estudio. Podemos definir la proporción muestral como

$$p = \frac{x}{n}$$

donde:

x = número de individuos de la muestra que cumplen la condición analizada.

n = tamaño muestral.

Cada x_i puede tomar dos valores. Uno, si el individuo i posee la característica que estamos estudiando y cero en caso contrario. Es decir X_i es una distribución de Bernoulli de parámetro p.

Dada una muestra aleatoria simple (X_1, X_2, ..., X_n) la

función de verosimilitud viene dada por

$$L(X_1, X_2, \ldots, X_n; p) = P(X_1)P(X_2)\ldots P(X_n) =$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

Si tomamos logaritmo neperiano obtenemos

$$\ln L = \sum_{i=1}^{n} x_i \ln p + (n - \sum_{i=1}^{n} x_i)\ln(1-p)$$

Si calculamos el valor de p que hace máxima esta función

$$\frac{d\ln L}{dp} = \sum_{i=1}^{n} x_i \frac{1}{p} + (n - \sum_{i=1}^{n} x_i)\frac{-1}{1-p} = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{\hat{p}} = \frac{n - \sum_{i=1}^{n} x_i}{1-\hat{p}}$$

$$\frac{1-\hat{p}}{\hat{p}} = \frac{n - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i}$$

$$\frac{1}{\hat{p}} - 1 = \frac{n}{\sum_{i=1}^{n} x_i} - 1$$

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

Es decir, el estimador máximo verosímil de la proporción poblacional es igual a la proporción muestral.

IV.3.- Estimación por intervalos de confianza.

IV.3.1.- Introducción.

El método de estimación por intervalo, como ya se ha dicho, tiene como objetivo estimar un parámetro θ, pero no de forma puntual, sino mediante la determinación de un intervalo (z_1, z_2) de tal manera que la probabilidad de que

el intervalo contenga al parámetro de interés sea alta.

Formalmente, la estimación por intervalo pretende obtener un intervalo (z_1, z_2) de tal manera que la probabilidad de que ese intervalo contenga al parámetro sea igual a $1-\alpha$.

A α se le denomina **nivel de significación,** se determina previamente y desde fuera del proceso de inferencia. Habitualmente α toma los valores 0.1, ó 0.05, ó 0.01. En estos casos hablamos respectivamente de niveles de significación del 10%, 5% y 1%

Al valor $1-\alpha$ se le conoce como **nivel de confianza**. Es decir, si α es igual al 5%, el nivel de confianza es del 95%, esto nos da una medida de la confianza de la estimación por intervalo. Es decir, tenemos una confianza del 95% de que nuestro intervalo contenga al valor del parámetro.

Obsérvese que cuando hablamos del nivel de confianza, cuando definimos el concepto de estimación por intervalo, siempre lo hacemos en términos de la **probabilidad de que nuestro intervalo contenga al parámetro**, y nunca en términos de **la probabilidad de que** θ **caiga dentro del intervalo**. Esto es debido a que lo que es variable es el intervalo, ya que es función de los individuos que forman la muestra. El parámetro es fijo y por tanto la segunda lectura es incorrecta puesto que lo trata como una variable.

Nuestro objetivo, por tanto, puede decirse que es:

$$P\ [\ z_1 < q < z_2\ ] = 1-\alpha$$

Sea un valor alto

Depende de la muestra

Es un valor fijo

El siguiente punto a determinar, dentro de la estimación por intervalos, es la determinación de los valores z_1 y z_2 dado el nivel de significación establecido (α). Las figuras 1 a la 6 muestran la forma en que son

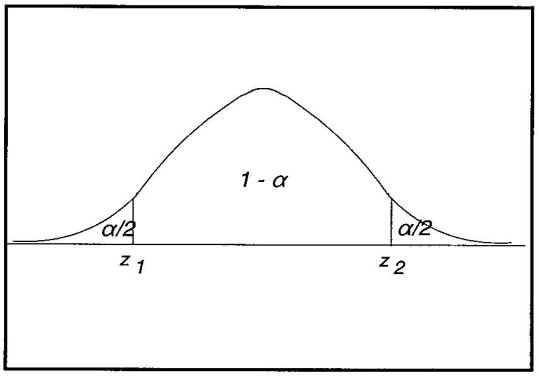

Figura 1

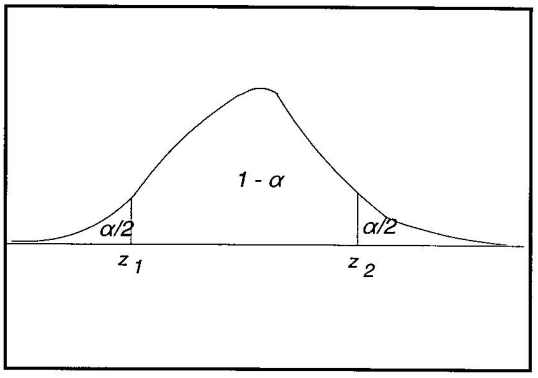

Figura 2

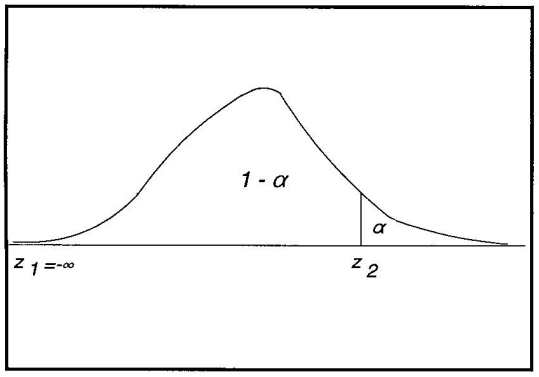

Figura 3

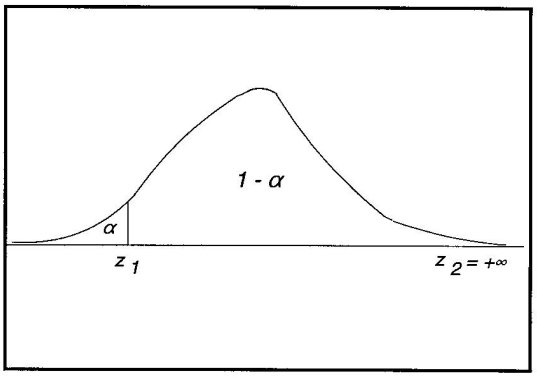

Figura 4

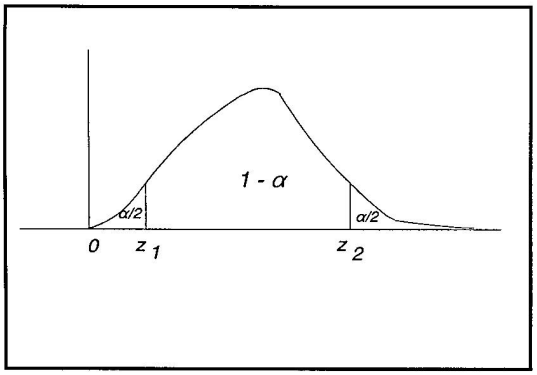

Figura 5

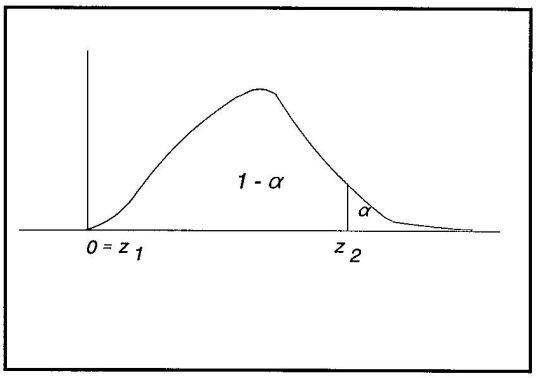

Figura 6

Por último, obsérvese también en estas figuras que la determinación de los valores de z_1 y z_2 exige el conocimiento de la función de densidad de una variable aleatoria que, primero, tenga algo que ver con el parámetro que deseamos estimar (es decir sea una función del mismo), segundo, tenga algo que ver con la realización muestral (esté relacionada con los valores de la muestra), tercero, la función de distribución de esta variable esté perfectamente definida.

Para poder obtener la estimación por intervalo nos vamos a circunscribir solamente al estudio de la estimación por intervalo en el caso de estar trabajando con poblaciones normales, (no debemos olvidar las propiedades que esta distribución tiene, tanto en su propia forma como en los criterios de convergencia a ella) y de proporciones.

IV.3.3.- Intervalos de confianza en poblaciones Normales.

IV.3.3.1.- Intervalo de confianza para la media de una población Normal cuando se conoce el valor de la varianza.

Supongamos que estamos interesados en conocer el comportamiento probabilístico de la estatura de la población de los habitantes de Las Palmas. Por estudios previos realizados en la misma zona geográfica y en otras similares, sabemos que nuestra variable de interés, la estatura de los habitantes de Las Palmas, a partir de ahora la denotaremos por **X**, tiene un comportamiento probabilístico como el de una distribución Normal. También por estudios previos, nosotros conocemos la varianza de la

población en estudio, a partir de ahora σ^2. Nuestro objetivo se centra en llevar a cabo una estimación por intervalo para el valor de la media de la estatura de los habitantes de Las Palmas.

Es decir, partimos de una población en la cual queremos estudiar una de sus características, que denotaremos por **X**. Sabemos que **X** se comporta como una distribución Normal de media μ (**desconocida**) y varianza σ^2 (**conocida**). Nuestro objetivo es obtener una estimación por intervalo de la media poblacional μ para un nivel de significación α.

Para ello obtenemos una muestra de tamaño **n**. Es decir, mediante un proceso de **muestreo aleatorio simple** medimos la estatura de **n** individuos pertenecientes a la población de Las Palmas. Esa muestra la vamos a denotar por $\{x_1, x_2, ..., x_n\}$.

Visto el planteamiento realizado al principio del tema lo que necesitamos es una variable aleatoria que esté en función del parámetro a estimar, μ, que dependa de la muestra y que además tengamos perfectamente definida su función de distribución.

Para el caso que nos ocupa, la estimación por intervalo de la media poblacional, la variable que cumple todas esas premisas o exigencias es la variable aleatoria media muestral. Veámoslo.

Si $\{x_1, x_2, ..., x_n\}$ es una muestra **genérica cualquiera de tamaño n**, esto es no es una realización muestral, sabemos que cada una de ellas, cada x_i se distribuye como una distribución Normal de media μ y varianza σ^2. Además, si la forma de obtener la muestra es mediante el muestreo

aleatorio simple, cada x_i es independiente del resto. Por otra parte, la variable aleatoria **media muestral**, viene dada por

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} * (x_1 + x_2 + \ldots + x_n)$$

es decir, la variable aleatoria **media muestral** es la suma de variables Normales independientes multiplicado por 1/n. Recordando los resultados de temas anteriores, sabemos que la variable **media muestral** se distribuye como una variable Normal cuya media vale

$$m_{\overline{x}} = E[\overline{X}] = \frac{1}{n} * n * E[X] = m$$

y su varianza

$$s^2_{\overline{X}} = E[\overline{X} - E[\overline{X}]]^2 = E[\frac{1}{n}(x_1 + x_2 + \ldots + x_n) - \frac{1}{n}(m + m + \ldots + m)]^2 =$$

$$= \frac{1}{n^2} * E[(x_1 - m) + (x_2 - m) + \ldots + (x_n - m)]^2$$

teniendo en cuenta la independencia entre las x_i podemos seguir desarrollando y obtener

$$s_{\bar{x}}^2 = \frac{1}{n^2} * (s^2 + s^2 + \ldots + s^2) = \frac{s^2}{n}$$

Es decir,

$$\overline{X} \rightarrow N(m, \frac{s^2}{n})$$

Como se puede observar, esta variable depende de la muestra, depende del parámetro a estimar, y si la tipificamos, su distribución está perfectamente definida.

Sea Z la variable tipificada de la **media muestral**. Z viene dada por

$$Z = \frac{\overline{X} - m}{\sqrt{\frac{s^2}{n}}} = \frac{\overline{X} - m}{s/\sqrt{n}}$$

siendo su distribución N(0,1) para cualquier valor de μ. Esto implica que dado el valor de α, la determinación de los puntos a_1 y a_2 viene dado por la figura 7 de forma

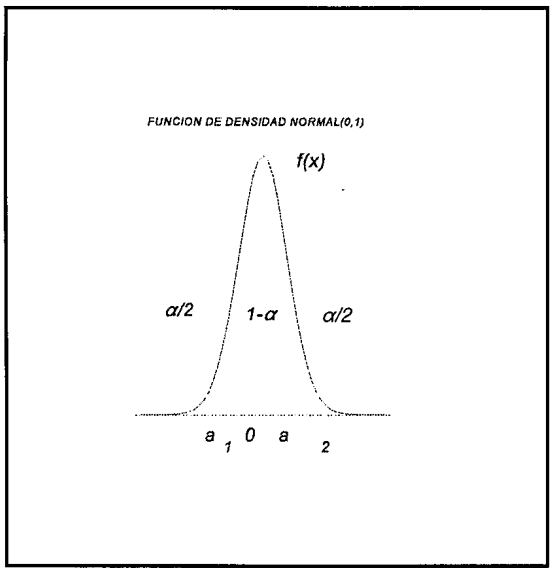

Figura 7

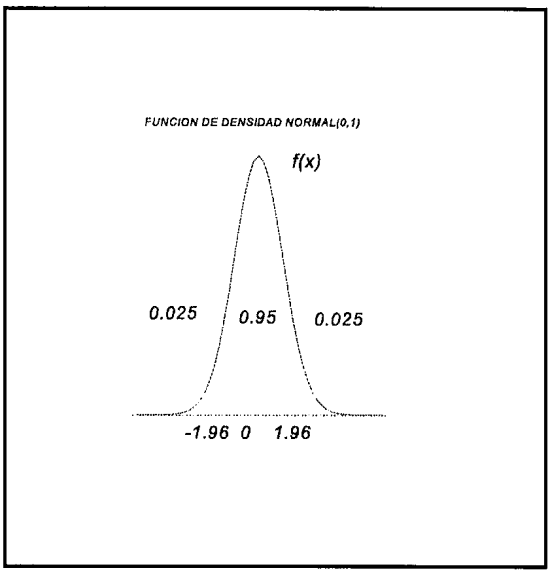

Figura 8

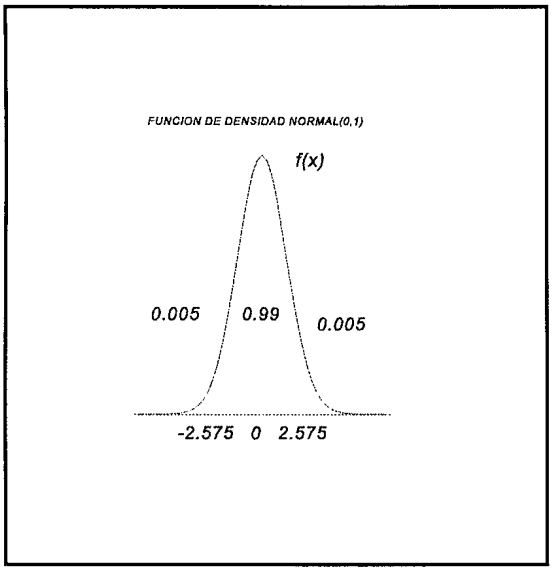

Figura 9

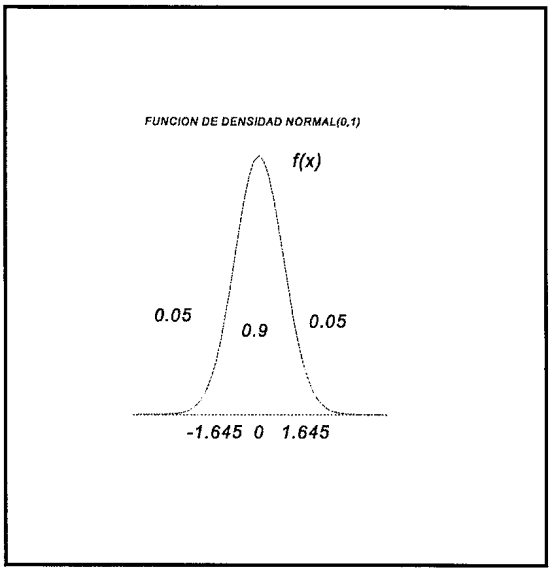

Figura 10

general, y en las figuras 8 a la 10 de forma particular para a=0.05, a=0.01 y a=0.1 respectivamente.

Por ejemplo, para a=0.05 (5%) sabemos que, dado que Z es una variable N(0,1),

$$P[-1.96 < Z < 1.96\ ] = 0.95$$

y teniendo en cuenta la relación entre la variable **media muestral** y la variable Z, obtenemos

$$P[-1.96 < \frac{\overline{X} - m}{s/\sqrt{n}} < 1.96] = 0.95$$

Expresión de la cual podemos despejar μ, obteniendo el intervalo de confianza dado por

$$P[\ \overline{X} - 1.96 * s/\sqrt{n} < m < \overline{X} + 1.96 * s/\sqrt{n}\ ] = 0.95$$

Es decir, **el intervalo**

$$(\ \overline{X} - 1.96 * s/\sqrt{n}, \overline{X} - 1.96 * s/\sqrt{n}\ ]$$

contiene a μ con una probabilidad del 95%

La generalización del intervalo para cualquier valor de α es inmediata. Teniendo como referencia la figura 7, **el intervalo**

$$(\ \overline{X} - a_2 * s/\sqrt{n}, \overline{X} + a_2 * s/\sqrt{n}\ )$$

es la estimación por intervalo que tiene una probabilidad

de 1-α de contener el verdadero valor del parámetro μ.

Un aspecto importante en la estimación por intervalo es **la precisión de la prueba.** Es evidente que si estoy interesado en estimar un parámetro θ mediante una estimación por intervalo, la interpretación del resultado (20,30) y del resultado (-100,200) para un mismo nivel de significación es completamente diferente. La prueba tendrá más **precisión cuanto más pequeño sea el intervalo.** Es decir, podemos decir que el intervalo (-100, 200) contiene a θ con una probabilidad del 95% es un resultado muy impreciso con respecto al otro intervalo, (20,30). Dicho de otra forma, decir que Telde está en Gran Canaria es mucho más preciso que decir que Telde está en España.

Por tanto, la **precisión** en el contexto de la estimación por intervalo tiene que ver con la amplitud del mismo. Una media muy utilizada de precisión es la mitad de la amplitud de la estimación por intervalo.

IV.3.3.2.- Intervalo de confianza para la media de una población Normal cuando no se conoce el valor de la varianza.

El método para obtener este intervalo es el mismo que el descrito para el caso anterior. Nuestro problema ahora es que si nos fijamos en el intervalo anterior, veremos que entre los datos necesarios para poder estimarlo está la desviación típica de la población. Sin embargo, en este caso, nosotros desconocemos este valor. Esto implica que la variable anterior no es suficiente para poder llegar a la estimación por intervalo para el caso concreto que nos ocupa. Necesitamos una nueva variable que nos resuelva el problema.

Sea $\{x_1, x_2, ..., x_n\}$ una realización muestral. Denotemos por S_2 a la varianza de los datos de la muestra, esto es

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

Podemos demostrar que si definimos una variable **U** como **$n*S^2/\sigma^2$, está se distribuye como una chi-cuadrado de (n-1) grados de libertad.** Es decir

$$U \rightarrow c^2_{n-1}$$

Además, por los desarrollos previos, sabemos que la variable **Z** definida como

$$Z = \frac{\overline{X} - m}{s/\sqrt{(n)}}$$

se distribuye como una distribución Normal de media cero y varianza 1, y podemos también demostrar que la variable Z es independiente de la variable U previamente definida.

Con esta información, podemos definir una nueva variable **T** como

$$T = \frac{Z}{\sqrt{\frac{U}{n-1}}}$$

que sabemos se distribuye como una t-student de n-1 grados de libertad.

Desarrollando la expresión anterior obtenemos

$$T = \frac{Z}{\sqrt{\frac{U}{n-1}}} = \frac{\frac{\overline{X} - m}{s/\sqrt{n}}}{\sqrt{\frac{n * S^2 / s^2}{n-1}}} = \frac{\overline{X} - m}{s} * \sqrt{n-1}$$

Obsérvese que T es una variable aleatoria que depende de la muestra, es función del parámetro a estimar y su función de distribución y densidad está perfectamente definida dado el tamaño muestral.

En consecuencia, dado el tamaño muestral y el nivel de significación, los valores t_α que determinan el intervalo que cumple

$$P[-t_\alpha < T < t_\alpha] = 1 - \alpha$$

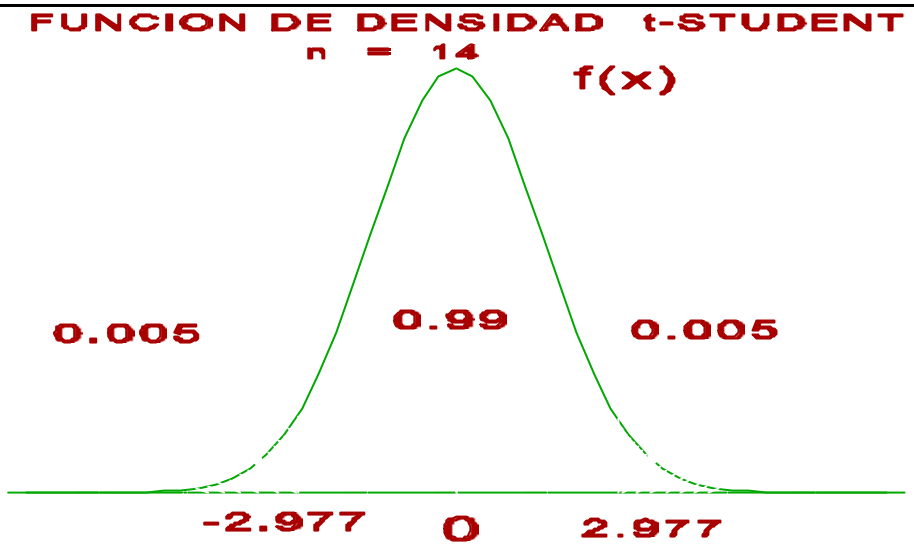

Figura 11

es inmediato. Por ejemplo para un tamaño muestral igual a 15 y para un nivel de significación del 1% el valor de t_α es 2.977. El esquema general se muestra en la figura 12.

Esto implica que

$$P[-2.977 < \frac{\overline{X} - m}{s} * \sqrt{n-1} < 2.977] = 0.99$$

Despejando el valor del parámetro dentro de la doble desigualdad, obtenemos la estimación por intervalo. El resultado viene dado por

$$P[-2.977 < \frac{\overline{X} - m}{s} * \sqrt{n-1} < 2.977] = 0.99$$

$$P[\overline{X} - 2.977 * \frac{S}{\sqrt{n-1}} < m < \overline{X} + 2.977 * \frac{S}{\sqrt{n-1}}] = 0.99$$

sustituyendo adecuadamente por el valor de la media muestral, la varianza muestral y el tamaño muestral, obtendremos la correspondiente estimación por intervalos.

El planteamiento general viene representado gráficamente por la figura 12, y analíticamente por la siguiente expresión

$$P[\overline{X} - t_a * \frac{S}{\sqrt{n-1}} < m < \overline{X} + t_a * \frac{S}{\sqrt{n-1}}] = 1 - a$$

en donde t_α corresponde al valor de una distribución t-Student de n-1 grados de libertad que deja a su izquierda una masa probabilística igual a $\alpha/2$.

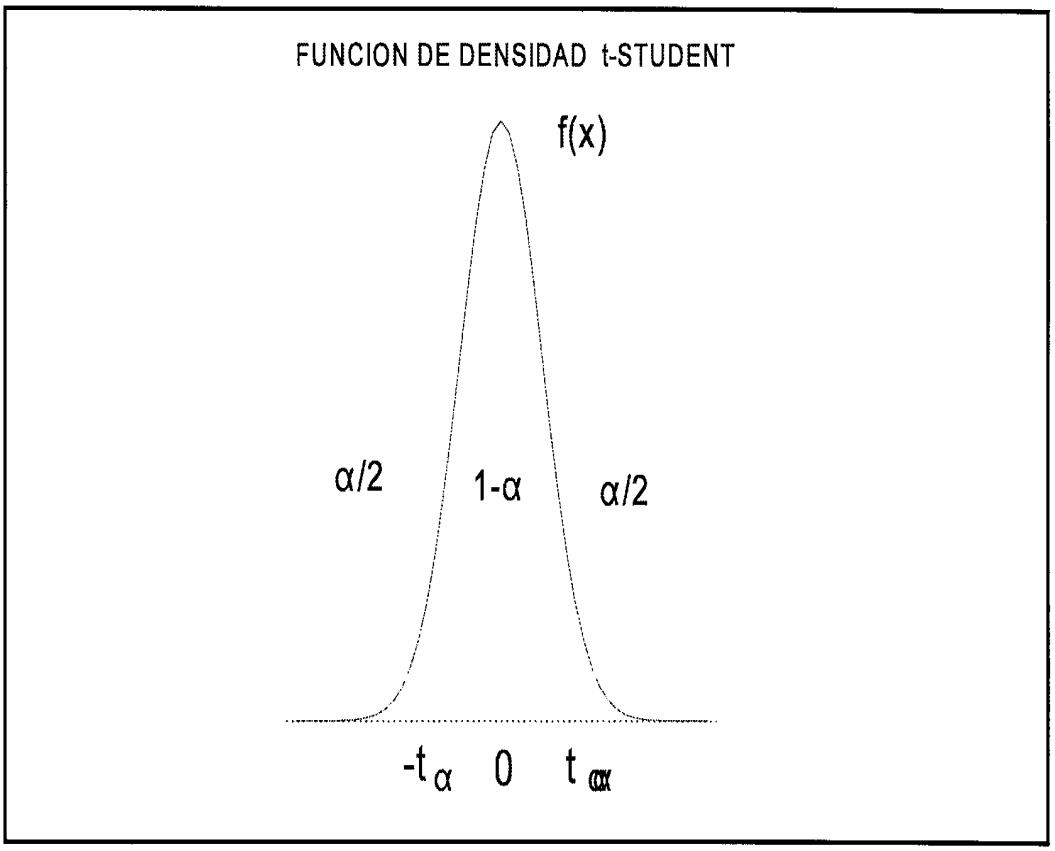

Figura 12

IV.3.3.3.- Intervalo de confianza para la diferencia de medias de poblaciones de medias cuando las varianzas son conocidas.

En este caso nos interesa tener una estimación por intervalo de la diferencia de los valores medios de dos poblaciones normales. El planteamiento lo haremos de forma reducida para que el alumno pueda utilizarlo como práctica.

Partimos de dos poblaciones en donde estamos interesados en estudiar las características X e Y respectivamente de ambas poblaciones. Además sabemos que X se distribuye como una distribución Normal de media μ_x (desconocida) y

desviación típica σ_x (conocida), y la variable Y se distribuye con media μ_y (desconocida) y desviación típica σ_y (conocida).

Nuestro objetivo es obtener una estimación por intervalo para la diferencia de las dos medias.

Como hemos hecho hasta ahora, necesitamos buscar una variable a partir de la cual podamos determinar los valores de z_1 y z_2, extremos del intervalo de confianza.

Para ello obtenemos **dos muestras independientes**. Una muestra aleatoria simple de tamaño **n** de la variable X y una muestra de tamaño **m** de la variable Y. A partir de estas dos muestras, sabemos, por resultados previos,

$$\overline{X} \rightarrow N(m_x, \sigma_x/\sqrt{n})$$

$$\overline{Y} \rightarrow N(m_y, \sigma_y/\sqrt{m})$$

y dado que las dos muestras se han obtenido de forma independiente, sabemos, por resultados previos, que

$$\overline{X} - \overline{Y} \rightarrow N\left(m_x - m_y, \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}\right)$$

En consecuencia, la variable **Z** definida como

$$Z = \frac{(\overline{X} - \overline{Y}) - (m_x - m_y)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

se distribuye como una distribución **N(0,1)**. De esta manera una vez determinado el valor de α, nivel de significación, la determinación del valor $\mathbf{a_2}$ que verifica

$$p[-a_2 < Z < a_2] = 1 - \alpha$$

es inmediato a partir de las tablas de la normal reducida. (Ver figuras de la 7 a la 10 de este tema)

Sustituyendo adecuadamente la variable Z en la expresión del cálculo de probabilidades

$$P[-a_2 < \frac{(\overline{X} - \overline{Y}) - (m_x - m_y)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} < a_2] = 1 - \alpha$$

Despejando en la doble desigualdad el valor de la diferencia de las medias poblacionales, llegamos a la siguiente expresión para la estimación por intervalo para la diferencia de medias para un nivel de significación α

$$P\left[\overline{X} - \overline{Y} - a_2 * \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}} < m_x - m_y < \overline{X} - \overline{Y} + a_2 * \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}\right] = 1 - \alpha$$

en donde $\mathbf{a_2}$ es el valor de la N(0,1) que deja a su derecha una masa probabilística igual a α/2.

V.3.3.4.- Intervalo de confianza para la diferencia de medias de dos poblaciones normales cuando no se conocen sus varianzas.

Se distinguirán dos casos:

Caso 1.- Cuando se dispone de muestras grandes (n, m > 30). En este caso el intervalo de confianza viene dado mediante la siguiente expresión.

$$P\left[\overline{X}-\overline{Y}-a_2*\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}<m_x-m_y<\overline{X}-\overline{Y}+a_2*\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}\right]=1-\alpha$$

en donde $\mathbf{a_2}$ sigue siendo el valor de la N(0,1) que deja a su derecha $\alpha/2$ de masa probabilística, y $\mathbf{S_x^2}$ y $\mathbf{S_y^2}$ son las varianzas muestrales de la muestra de X e Y respectivamente.

Caso 2.- Cuando las varianzas son desconocidas pero supuestamente iguales. En este caso el intervalo de confianza viene dado por

$$P\left[\overline{X}-\overline{Y}-t_2*\sqrt{n\,S_x^2+m\,S_y^2}*\frac{\sqrt{\frac{1}{n}+\frac{1}{m}}}{\sqrt{n+m-2}}<m_x-m_y<\overline{X}-\overline{Y}+t_2*\sqrt{n\,S_x^2+m\,S_y^2}*\frac{\sqrt{\frac{1}{n}+\frac{1}{m}}}{\sqrt{n+m-2}}\right]=1-\alpha$$

en donde $\mathbf{t_2}$ es el valor de la **t-Student** de **n+m-2** grados de libertad que deja a su derecha $\alpha/2$ de masa probabilística.

IV.3.3.5.- Intervalo de confianza para la varianza de una población Normal.

En este caso partimos de una característica X que se comporta de forma normal con media μ y desviación típica σ. Nuestro objetivo es obtener la estimación por intervalo de la varianza poblacional. Para ello obtenemos una muestra aleatoria simple de tamaño n, $\{x_1, x_2, ..., x_n\}$.

Hemos dicho previamente que podríamos demostrar que $Z=(n*S^2/\sigma^2)$ se distribuye como una distribución chi-cuadrado de n-1 grados de libertad. (Recordemos que S^2 es la varianza de la muestra.

Sabiendo como se distribuye esta variable Z, la determinación de los valores **a** y **b** que verifica

$$P[\ a < Z < b\ ] = 1 - \boldsymbol{a}$$

es inmediata a partir de la tabla de la distribución chi-cuadrado.

Desarrollando la expresión anterior, obtenemos el correspondiente intervalo de confianza.

$$P[\ a < \frac{n * S^2}{\boldsymbol{s}^2} < b\ ] = 1 - \boldsymbol{a}$$

$$P[\frac{1}{b} < \frac{\boldsymbol{s}^2}{n * S^2} < \frac{1}{a}] = 1 - \boldsymbol{a}$$

$$P[\frac{n * S^2}{b} < \boldsymbol{s}^2 < \frac{n * S^2}{a}] = 1 - \boldsymbol{a}$$

en donde **a** y **b** son los valores de una distribución chi-cuadrado que cumplen las condiciones establecidas en la figura 13.

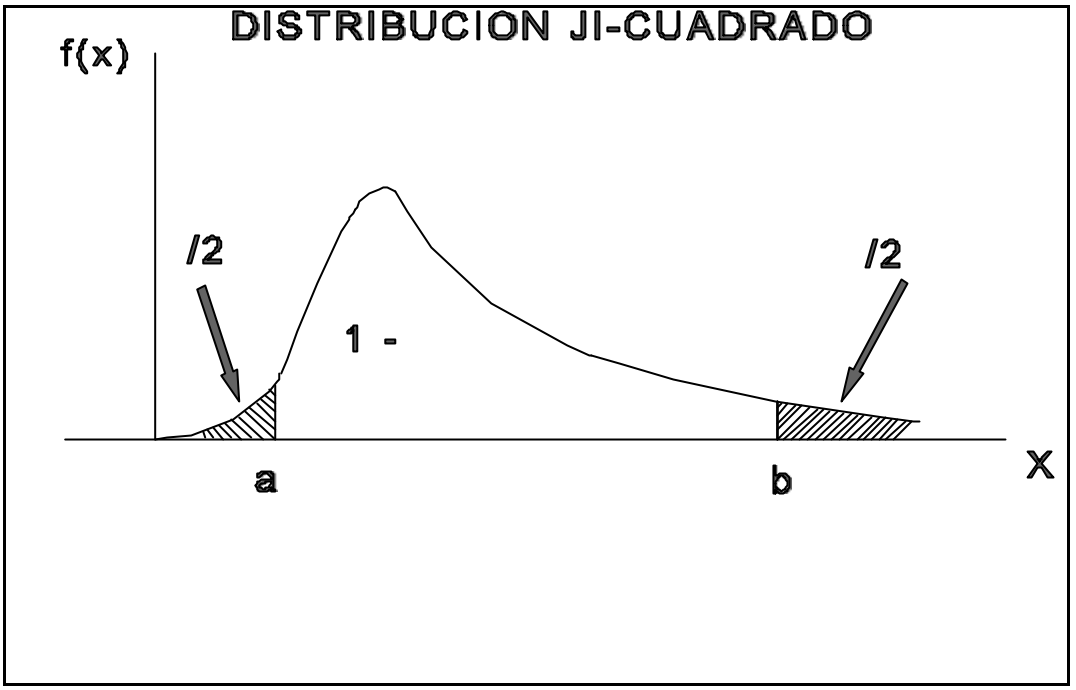

Figura 13

IV.3.4.- Intervalo de confianza para las proporciones

Supongamos una población de tamaño N en la que se quiere estudiar si sus elementos cumplen una determinada condición. Podemos llamar proporción poblacional (P) a la proporción de elementos de la población que cumplen la condición, de manera que

> $P=k/N$, siendo k el número de elementos de la poblaciòn que cumplen la condición que estudiamos.

Tomemos ahora una muestra aleatoria simple de tamaño n (x_1, x_2,....x_n) y llamemos proporción muestral (p) a la proporción de individuos de la muestra que cumplen la condición, de manera que:

> $p=x_i/n$, siendo x_i la variable muestral.

Cada x_i de la muestra puede tomar dos valores: valor 1 si presenta la caracteristica estudiada y valor 0 si no la presenta, por lo que podemos entender que se trata de una distribución de Bernouilli de parámetro p.

Si aplicamos el método de máxima verosimilitud es fácil deducir que el estimador máximoverosimil de la proporción poblacional es igual a la proporción muestral, de manera que

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

En este caso, la media tiene la interpretación de proporción o frecuencia relativa que se observa en la muestra de elementos con la característica estudiada.

Sabemos que

$$E(\hat{p}) = E(\bar{x}) = m = p$$

$$V(\hat{p}) = V(\bar{x}) = \frac{s^2}{n} = \frac{pq}{n} = \frac{p*(1-p)}{n}$$

Si fijamos el nivel de confianza 1-α podemos obtener la constante a_2 como hacemos habitualmente, y el intervalo de confianza para el parámetro P resultante será:

$$(\hat{p} - a_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \hat{p} + a_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

IV.3.5.- TAMAÑO MUESTRAL PARA UNA PRECISION DADA

Nuestro objetivo en este epígrafe es considerar cómo se

puede determinar el tamaño de la muestra para realizar una estimación por intervalos que presente una determinada precisión. Nos limitaremos a los casos más frecuentes en una investigación, es decir para la estimación de una media poblacional (con varianza conocida o desconocida) o de una proporción. En todo caso, si conocemos el estimador por intervalos, el tamaño muestral para una precisión dada lo podemos obtener despejando el valor de n.

IV.3.5.1.- Determinacion del tamaño de la muestra para estimar a media de una poblacion con varianza conocida.

Aplicaremos este caso tanto si la población presenta una distribución Normal como si presenta algún otro tipo de distribución pero el tamaño de la muestra es suficientemente grande como para que converja a una Normal. En ambos casos, sabemos ya que el intervalo de confianza de la media poblacional para el nivel de confianza 1-α viene dado por

$$P(\overline{X} - a_2 * s/\sqrt{n} \leq m \leq \overline{X} + a_2 * s/\sqrt{n}) = 1 - a$$

que puede expresarse también, como sabemos,

$$P(|\overline{x} - m| \leq a_2 * s/\sqrt{n}) = 1 - a$$

donde

$$|\overline{X} - m|$$

es el error (e) que cometemos en el proceso de estimación.

Nosotros podemos situar el error de estimación dentro del límite aceptado, es decir, podemos hacer que

$$e = |\bar{x} - m| = a_2 * s/\sqrt{n}$$

y de aquí

$$n = \frac{a_2^2 * s^2}{e^2}$$

La expresión anterior permite calcular el tamaño de la muestra (n) que habría que tomar, en el supuesto de que la varianza es conocida, dado un nivel de confianza 1-α del que depende el valor a_2 y dado un error máximo (e) que estamos dispuestos a aceptar.

IV.3.5.2.- Determinacion del tamaño de la muestra para estimar media de una poblacion con varianza desconocida.

En este caso, el intervalo de confianza de la media para el nivel e confianza de 1-α sabemos que es

$$P(\overline{X} - t_a * \frac{S}{\sqrt{n-1}} \leq m \leq \overline{X} + t_a * \frac{S}{\sqrt{n-1}}) = 1 - a$$

que puede expresarse también como

$$P(|\overline{X} - m| \leq t_a * \frac{S}{\sqrt{n-1}}) = 1 - a$$

Por el mismo razonamiento del caso anterior, el tamaño muestral viene dado por la expresión

$$n = \frac{t_a^2 * S^2}{e^2} + 1$$

IV.3.5.3. Determinacion del tamaño de la muestra necesario para estimar la proporcion p de una poblacion.

Como en los casos anteriores, el punto de partida es el

intervalo de confianza para el parámetro p para un nivel de confianza de 1-α:

$$P(\hat{p} - a_2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + a_2 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 1 - \boldsymbol{a}$$

donde el error de estimación sería

$$e = |\hat{p} - p|$$

y la probabilidad de cometerlo:

$$P(|\hat{p} - p| \le a_2 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1 - \boldsymbol{a}$$

de donde el tamaño de la muestra sería:

$$n = \frac{a_2^2 * \hat{p}(1-\hat{p})}{e^2}$$

IV.- Problemas

1.- Dada una muestra de tamaño n, demostrar que el estimador de la media poblacional

$$\hat{\boldsymbol{q}} = \frac{1}{n}\sum_{i=1}^{n-1} x_i$$

es un estimador sesgado. Calcular el sesgo.

Solución. Para que el estimador sea insesgado debe cumplirse que la esperanza del estimador sea igual al valor del parámetro.

$$E[\hat{\boldsymbol{q}}] = \boldsymbol{q}$$

$$E[\hat{\boldsymbol{q}}] = E[\frac{1}{n}\sum_{i=1}^{n-1} x_i] = \frac{1}{n} E[\sum_{i=1}^{n-1} x_i] =$$

$$= \frac{1}{n}\sum_{i=1}^{n-1} E(x_i) = \frac{1}{n}(n-1)\, m_x = \frac{(n-1)}{n}\, m_x = m_x - \frac{m_x}{n}$$

Siendo el sesgo del estimador

$$b(\theta) = -\frac{m_x}{n}$$

2.- De una población N(μ,2) se obtiene una muestra de tamaño 4 representada por $\{x_1, x_2, x_3, x_4\}$. Se consideran los siguientes estimadores de μ:

a)

$$\hat{m}_1 = \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{2}X_3$$

b)

$$\hat{m}_2 = \frac{1}{3}X_1 + \frac{1}{6}X_2 + \frac{1}{3}X_3 + \frac{1}{6}x_4$$

c)

$$\hat{m}_3 = \frac{1}{4}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3 + \frac{1}{4}X_4$$

Preguntas:

1.- Determinar la sesgadez o insesgadez de los estimadores.

2.- Calcular la varianza de los estimadores y determinar cual es el de menor varianza.

<u>Solución:</u>

1.- Caso a) sesgado, sesgo=m/4

Caso b) Insesgado

Caso c) Insesgado

2.- Caso a) varianza del estimador = $9s^2/16$

Caso b) varianza del estimador = $10s^2/36$

Caso c) varianza del estimador = $s^2/4$

Por lo tanto el de menor varianza es el caso c), que no es más que la media aritmética de la muestra.

3.- Demostrar que la varianza muestral es un estimador

sesgado de la varianza poblacional.

4.- Demostrar que la covarianza muestral es un estimador insesgado de la varianza poblacional.

5.- Sea X el número de accidentes que se producen en un cruce durante una semana. Sabemos que X se distribuye como una Poisson de parámetro l =2. Se va a realizar un estudio durante cinco semanas y queremos conocer la probabilidad de que en la primera semana se produzcan tres accidentes, en la segunda 1, en la tercera ninguno, en la cuarta dos y en la quinta ninguno. (En realidad en las cinco semanas de estudio vamos a obtener una muestra de la variable aleatoria X de tamaño n=5 y lo que queremos es conocer la probabilidad de una determinada realización muestral).

Nosotros debemos calcular la probabilidad del siguiente suceso:

$$P[X_1 = 3) \cap (x_2 = 1) \cap (x_3 = 0) \cap (x_4 = 2) \cap (x_5 = 0)]$$
$$= P(x_1 = 3) * P(x_2 = 1) * P(x_3 = 0) * P(x_4 = 2) * P(x_5 = 0)$$
$$= \frac{e^{-2}\, 2^3}{3!} * \frac{e^{-2}\, 2^1}{1!} * \frac{e^{-2}\, 2^0}{0!} * \frac{e^{-2}\, 2^2}{2!} * \frac{e^{-2}\, 2^0}{0!} =$$
$$= \frac{e^{-10}\, 2^6}{3!*2!*0!*2\,!*0!} = 0.0002421$$

Imaginemos que ahora sabemos que la variable aleatoria X se distribuye como una Poisson pero nos conocemos el valor de l. En este caso la probabilidad de obtener la muestra anterior vendrá dada por

$$P[X_1 = 3) \cap (x_2 = 1) \cap (x_3 = 0) \cap (x_4 = 2) \cap (x_5 = 0)]$$
$$= P(x_1 = 3) * P(x_2 = 1) * P(x_3 = 0) * P(x_4 = 2) * P(x_5 = 0)$$
$$= \frac{e^{-l}\, l^3}{3!} * \frac{e^{-l}\, l^1}{1!} * \frac{e^{-l}\, l^0}{0!} * \frac{e^{-l}\, l^2}{2!} * \frac{e^{-lamda}\, lamda^0}{0!} =$$
$$= \frac{e^{-5\,l}\, l^{(3+1+0+2+0)}}{3!*2!*0!*2\,!*0!} = \frac{e^{-5\,l}\, l^6}{12}$$

Como se puede observar el valor de la probabilidad depende

del valor del parámetro l. Cuando empleamos el método de la máxima verosimilitud para estimar el valor de l, lo que hacemos es escoger el valor que hacer que la probabilidad de haber obtenido la realización muestral con la que trabajamos sea la máxima posible.

Como ya sabemos la función de verosimilitud viene dada por la función de densidad conjunta. En este caso particular tenemos

$$P[3,1,0,2,0,;\ \lambda\ ] = \frac{e^{-5\lambda}\ \lambda^{6}}{12}$$

$$L[3,1,0,2,0;\ \lambda\ ] = \frac{e^{-5\lambda}\ \lambda^{6}}{12}$$

Para calcular el valor estimado de l debemos calcular el máximo de la función de verosimilitud. Para facilitar los cálculos normalmente se trabaja con el logaritmo neperiano de dicha función. Para este ejemplo tenemos

$$\ln L = -5\ \lambda \ln(e) + 6 \ln\ \lambda - \ln 12$$

$$\frac{d \ln(L)}{d\lambda} = -5 + 6 * \frac{1}{\hat{\lambda}} = 0$$

$$\hat{\lambda} = \frac{6}{5} = 1{,}2$$

$$\frac{d^{2} \ln L}{d\lambda} = \frac{-6}{\lambda^{2}}$$

6.- Sea una población definida por

$$f(x,\theta) = \frac{1}{\theta} x^{-(\frac{1}{\theta}+1)},\ x \geq 1, \theta > 0$$

Se pide obtener el estimador maximoverosímil del parámetro theta.

<u>Solución</u>

La función de verosimilitud para muestras aleatorias simples de tamaño n:

$$L(X;\theta) = f(x_1,\theta).....f(x_n,\theta) =$$
$$= \frac{1}{\theta} x_1^{-(\frac{1}{\theta}+1)}.....\frac{1}{\theta} x_n^{-(\frac{1}{\theta}+1)} =$$
$$= \frac{1}{\theta^n}(x_1....x_n)^{-(\frac{1}{\theta}+1)}$$

Aplicamos logaritmos neperianos

$$lnL(X;\theta) = -nlntheta - (\frac{1}{\theta}+1)\ln(x_1...x_n) =$$
$$-nln\theta - (\frac{1}{\theta}+1)\sum_{i=1}^{n}\ln x_i;$$
$$\frac{d\ln L(X;\theta)}{d\theta} = \frac{-n}{\theta} + \frac{\sum_{i=1}^{n} lnx_i}{\theta^2} = 0$$

por lo que el estimador maximo verosimil será

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\ln x_i}{n}$$

7.- sea X una variable aleatoria con la siguiente función de densidad

$$f(X;a) = \begin{cases} 0 & si \;\; x < 0 \\ a\,e^{-ax} & si \;\; x \geq 0 \end{cases}$$

Se toma una muesta aleatoria simple de tamaño, siendo su realización muestral (1.5, 0.25, 2.75, 3, 4, 0). Calcular el estimador máximo verosimil de a.

8.- Sea X una variable aleatoria cuya función de cuantía viene dada por

$$f(X;\ a,b) = \begin{cases} \dfrac{a+2b-3}{13} & si \quad x = -3 \\ \dfrac{2a+1}{13} & si \quad x = 0 \\ \dfrac{3b}{13} & si \quad x = 1 \\ \dfrac{15-3a-5b}{13} & si \quad x = 2 \\ 0 & si \quad x \neq -3,0,1,2 \end{cases}$$

a. demostrar que es una verdadera función de cuantía.
b. Si tomamos una muestra aleatoria simple con la siguiente realización muestral {-3, 0, 1, 2}. Calcular los estimadores máximo verosímiles de a y b.

9.- Supongamos que tenemos una población que se comporta como una distribución Normal de la cual no conocemos su media, pero si sabemos que su desviación típica es igual a 3. Para poder estudiar esta población, obtenemos una muestra obteniendo los siguientes resultados {18,22,23,17,20, 17,23,14,26}. Obtener la estimación puntual y por intervalos para un nivel de confianza del 95%

Solución:

La estimación puntual por máxima verosimilitud es la media muestral. Si calculamos la media de la muestra, obtenemos que su valor es igual a 20. Esta es la estimación puntual de la media poblacional.

Por otra parte, sabemos que el intervalo de confianza viene dado por

$$(\overline{X} - a_2 * s/\sqrt{n}, \overline{X} + a_2 * s/\sqrt{n})$$

para un nivel de significación genérico α. Para un nivel de significación del 5%, el valor de a_2 es igual a 1.96. El valor de la media muestral es 20, el tamaño muestral 9 y la desviación típica 3. Lo único que tenemos que hacer es sustituir estos valores en la expresión anterior.

$$(20 - 1.96 * 3/\sqrt{9}, 20 + 1.96 * 3/\sqrt{9}) = (18.04, 21.96)$$

Por lo tanto, tenemos una probabilidad del 95% de que el intervalo (18.04, 21.96) contenga al verdadero valor de la media poblacional.

Obsérvese que si nosotros tuviésemos otra muestra, la media muestral puede ser distinta y en consecuencia el intervalo tendría unos extremos distintos.

10.- Dada una población N(30,10),) qué tamaño mínimo debe de tener la muestra para que la estimación de la media difiera del valor real en menos del 30% para un nivel de confianza del 95%.

<u>Solución:</u> n=43

11.- Se desea obtener un intervalo de confianza de la media de una población Normal de parámetros desconocidos a partir de una muestra dada por (5,7,5,3,2,8,11,14,16,18). Con un nivel de significación del 10%

<u>Solución:</u> (5.667, 12.133)

12.- El número diario de piezas fabricadas por la máquina

A en cinco días ha sido {50, 48, 53, 60, 37}, mientras que en esos mismos días la máquina **B** ha hecho {40, 51, 62, 55, 64}. Se pide: a) Construir un intervalo para la diferencia de medias entre las máquinas A y B, suponiendo que ambas máquinas tienen la misma desviación típica. b) Lo mismo, sin suponer la misma desviación típica. c) Determinar cual debía haber sido el tamaño muestral **n** de ambas muestras para que en el caso a) y con el mismo valor de la varianza estimada, la longitud del intervalo para la diferencia de medias con a= 5% fuese de 8 unidades.

Solución:

Caso a) [8.39,17.99]

Caso b) [-14.81,5.21] (solo aplicable a muestras mayores que 30)

Caso c) n > 54 días.

13.- Dada La población N(m,s) calcular el intervalo de confianza para la varianza sabiendo que en una muestra de tamaño 8 hemos obtenido una varianza muestral igual a 9.

Solución: [5.12,33.23]

14.- De una urna con bolas blancas y negras se obtiene una muestra aleatoria simple de tamaño n=140, observándose 80 bolas blancas.

Hallar el intervalo de confianza de la proporción p de bolas blancas de la urna al nivel del 99.7%

Solución

El estimador maximoverosimil de la proporción poblacional es

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{80}{140} = 0{,}57$$

El valor de a_2 para un nivel de o,997 es aproximadamente igual a 3. Por tanto, el intervalo de confianza para p es:

$$(0,57 - 3\sqrt{\frac{0,57(1-0,57)}{140}}\,;0,57 + 3\sqrt{\frac{o,57(1-0,57)}{140}}\,) = [0,444;0,696]$$

15.- Determínese el tamaño de la muestra necesario para estimar la media m de una población con s=4,2. si se quiere tener una confianza del 95% de que el error de estimación se sitúe como mucho en el "0,05

Solucion

Como 1-α es 0,95, en las tablas obtenemos el valor 1,96.

$$n = \frac{(1,96)^2 * (4,2)^2}{(} 0,05)^2 = 27.106$$

Por tanto

16.- Con el fin de estimar la media de una población se definen los siguientes estimadores:

$$q_1 = \frac{1}{n}_(x_1 + x_2 + ... + x_n)$$

$$q_2 = \frac{1}{2}(\frac{x_1 + x_2}{2} + \sum_{i=3}^{n} \frac{x_i}{n-2})$$

Seleccionar el estimador adecuado para obtener una estimación de la media (m), a partir de una muestra de tamaño n=10.

17.- Supongamos que X (n1 de accidentes que ocurren en un determinado punto kilométrico de la carretera nacional 340 por semana) es una variable aleatoria, que se distribuye según un modelo de Poisson de parámetro $l = 2$.

a) ¿Cuál es la probabilidad de obtener la muestra de tamaño 5 (observaciones de 5 semanas consecutivas) para el estudio de esta variable, formada por los valores (

3,1,0,2,0)?.

b) ¿Cuál sería la probabilidad si desconociéramos el valor de l?.

18.- Con las mismas condiciones del ejercicio anterior, y suponiéndo que desconocemos l.

Calcular según el M.M.V. (método de máxima verosimilitud) el estimador de l para dicha muestra.

19.- Sea X una variable aleatoria contínua, con la siguiente función de densidad:

$$f(x;a) = a.e^{-ax} \;\; si \;\; x \geq 0 \; y \; 0 \; para \; el \; resto$$

En el estudio de esta variable se toma una muestra aleatoria simple de tamaño 6, siendo su realización muestral (1.5, 0.25, 2.75, 3, 4, 0).

Se quiere calcular el ESTIMADOR MAXIMO VEROSIMIL de a.

20.- Una variable aleatoria está distribuída uniformemente en el intervalo (0,q).

Si se extrae una muestra de tamaño 4, y se define el estimador de q, como

$$\hat{q} = k_\sum_{i=1}^{4} \frac{x_i}{4}$$

¿Qué valor debe tener k para que el estimador $\hat{q}$ sea insesgado?.

21.- En una muestra de tamaño 2 extraída de una población normal N(m,1) empleamos como estimadores de la media los siguientes estdísticos:

$$\hat{q}_1 = \frac{2}{3}x_1 + \frac{1}{3}x_2$$

$$\hat{q}_2 = \frac{2}{5}x_1 + \frac{4}{5}x_2$$

$$\hat{q}_3 = ESTIMADOR.\ MAXIMO.VEROSIMIL$$

a) ¿Cúal o cúales son insesgados?

b) De los sesgados, ¿cúal es más eficiente?.

22.- Hallar el estimador máximo verosimil de q, para una variable aleatoria que tiene una función de cuantía $f(x;q) = p_q^x$ *donde* $q = 1 - p$, en muestras de extensión n.

23.- Dada la función de densidad $f(x;q) = \frac{e^{-\frac{x^2}{2q^2}}}{q\sqrt{2p}}$, hallar el estimador de máxima verosimilitud para muestras de tamaño n.

24.- Los datos que se dan a continuación son los pesos en gramos del contenido de 16 cajas de cereales que se seleccionan en un proceso de llenado, con el propósito de verificar el peso promedio: 506, 508, 499, 503, 504, 510, 497, 512, 514, 505, 493, 496, 506, 502, 509, 496. Si el peso medio de cada caja es una variable aleatoria normal con una desviación standard de 5 gramos.
Obtener el intervalo de confianza del 90% para la media de llenado de este proceso.

25.- La variable aleatoria "demanda" (miles de pesetas) se distribuye como una normal con desviación típica igual a 10.000.
Se estudia una muestra de 20 elementos y se obtiene una media de 39.000 ptas.

a) Contruir un intervalo de confianza para la media poblacional con un nivel de significación del 5%.
b)) qué tamaño mínimo deberá tener la muestra para que la estimación de la media difiera del valor real en menos de 3.000 ptas?) Y en menos de 1.000 ptas?.

26.- Supongamos que tenemos una población que se comporta como una distribución normal, de la cual no conocemos su media, pero si su desviación típica que es igual a 3. Para

poder estudiar esta población, obtenemos una muestra con los siguientes resultados: 18, 22, 23, 17, 20, 17, 23, 14, 26.
Obtener la estimación puntual y por intervalos para un nivel de confianza del 95%.

27.- Sea una variable X caracterizada por una distribución normal de media desconocida y varianza 4.
Contruir un intervalo de confianza para la media de la población con la siguiente información muestral: 29, 30, 35, 36, 40, 42, 45, 47, 48.

28.- Un fabricante de fibras sintéticas desea estimar la tensión de ruptura media de un tipo de fibra. Para ello, diseña un experimento en el que se observan las tensiones de ruptura, en libras, de 16 hilos seleccionados aleatoriamente. Las tensiones obtenidas son 20.8, 20.6, 21.0, 20.9, 19.9, 20.2, 19.8, 19.6, 20.9, 21.1, 20.4, 20.6, 19.7, 19.6, 20.3, 20.7
Supóngase que la tensión de ruptura de una fibra se encuentra modelada por una distribución normal con desviación típica de 0.45 libras.

Construir un intervalo de confianza estimado del 98% para el valor real de la tensión de ruptura promedio de la fibra.

29.- Se desea obtener un intervalo de confianza de la media de una población normal de parámetros desconocidos a partir de una muestra dada por (5, 7, 5, 3, 2, 8, 11, 14, 16, 18), con un nivel de significación del 10%.

30.- Una muestra de 10 medidas de diámetro de una esfera dió una media de 4.38 pulgadas y una desviación típica de

0.06 pulgadas. Hallar los límites de confianza para el diámetro verdadero del 95% de nivel de confianza.

31.- Se está estudiando el absentismo laboral femenino en una empresa de determinado sector. Se eligieron, para ello, 10 trabajadoras al azar de una determinada sección, anotándose el número de días que faltan al trabajo por diferentes motivos, en los últimos 4 meses:

EMPL. N1	1 2 3 4 5 6 7 8 9 10
FALTAS (días)	5 4 1 6 8 7 4 2 7 6

Determinar un intervalo de confianza del 99% para el n1 medio de días de absentismo laboral femenino en la empresa en el último cuatrimestre.

32.- Una muestra de 150 bombillas del fabricante A dieron una vida media de 1400 horas teniendo esta variable una desviación poblacional de 120 horas y otra muestra de 100 bombillas del fabricante B dio una vida media de 1200 horas con una desviación poblacional de 80 horas. Hallar los límites de confianza del 99% para la diferencia de las vidas medias de las poblaciones A y B.

33.- Dos universidades públicas tienen métodos distintos para matricular a sus alumnos. Los dos desean comparar el tiempo que tardan los alumnos en completar el trámite de matriculación. Para ello, se anotaron los tiempos de matriculación en cada universidad para grupos de 100 alumnos seleccionados al azar. Las medias y desviaciones típicas muestrales son las siguientes:

media 1 = 50.2 media 2 = 52.9
desviación 1 = 4.8 desviación 2 = 5.4

Si se supone que el muestreo se lleva a cabo sobre dos distribuciones normales e independientes, obtener los intervalos de confianza para el 90%, 95% y el 99%, para la diferencia de medias del tiempo de matriculación para las dos universidades.

34.- Sean X e Y dos poblaciones normales de medias desconocidas y de varianzas comunes. A partir de muestras de tamaño n1=10 y n2=12, construir un intervalo de confianza para la diferencia de medias poblacionales, si las medias muestrales son: 15.7 para la población 1 y 14.3 para la población 2, y las desviaciones muestrales son 10.2 y 9.6 respectivamente, para niveles de confianza 90, 95 y 99%.

35.- Si una variable aleatoria tiene por función de densidad

$$f(x;a)=\frac{2a}{1-a}_x^{\left(\frac{3a-1}{1-a}\right)} \quad si \;\; 0 \le x \le 1 \; y \; a > 0$$

Hallar el estimador máximo verosimil del parámetro a, para muestras aleatorias simples de tamaño n.

36.- Dada la población $N(m,s)$, calcular el intervalo de confianza para la varianza sabiendo que en una muestra de tamaño 8, hemos obtenido una varianza muestral igual a 9. Para un nivel de significación del 5%, del 10% y del 1%.

37.- Un proceso produce cierta clase de cojinetes de bola cuyo diámetro interior es de 3 cm. Se seleccionan de forma aleatoria 12 de estos cojinetes y se miden sus diámetros

internos, resultando una varianza muestral de 0.0005455. Suponiendo que el diámetro es una variable aleatoria de normalmente distribuida, determinar el intervalo de confianza del 99% para la varianza.

38.- Una muestra aleatoria de 20 cojinetes fabricados por una máquina durante una jornada dio una media de 20.92 mm. y una desviación de 1.067 mm.
Hallar los intervalos de confianza del 95%, 98%, 99% y 90% de los diámetros medios de los cojinetes.

39.- Una muestra tomada de una población normal dio los siguientes valores: 521, 509, 539, 516, 508, 517, 529, 501, 521, 532, 547, 504, 529.
Determinar con una confianza del 95%, si es admisible el valor 10 como posible desviación típica poblacional.

40.- Para comparar las puntuaciones medias de 2 asignaturas se han considerado las calificaciones obtenidas en cada una de ellas, sobre una muestra de 50 alumnos, obteniéndose los siguientes resultados:

$\overline{x_a} = 5.1 \quad S_a = 2.5 \quad \overline{x_b} = 6.2 \quad S_b = 3.1$

Obtener un intervalo de confianza para la diferencia de medias a un nivel del 95%.

41.- De dos grupos análogos de enfermos A y B formados por 50 y 100 individuos, respectivamente, al primero le fue dado un nuevo tipo de somnífero y al segundo u tipo convencional.
Para los pacientes del primer grupo, el número medio de horas de sueño fue 7.82 con una desviación de 0.24 horas. Para los del grupo B fueron 6.75 y 0.30 horas respectivamente.

Hallar los límites de confianza del 95% y 99% para la diferencia del número medio de horas de sueño inducidas por los dos tipos de somníferos.

42.- Sabiendo que el tanto por uno de los clientes no morosos que corresponde a cada uno de los representantes de una empresa, se distribuye según densidad de probabilidad dada por la curva de la familia:

$f(x) = (k+1+)_ x^k \;\; si \;\; 0 < x < 1$

donde k es un parámetro que puede tomar valores mayores que -1 y que depende del representante en cuestión.
Se pide obtener un estimador del parámetro k, basado en una muestra de n elementos, a través del método de máxima verosimilitud.

43.- Una cuenta contable formada por un gran número de asientos presenta una varianza entre las cifras registradas igual a 2.224. ¿ cuál deberá ser el tamaño de la muestra para estimar el importe medio de todos los apuntes contabilizados con un error de 12 pesetas, si suponemos una población básica normal y un 95% de confianza?.

44.- El tiempo que tarda en cubrirse cierto trayecto de una línea de guagua es aleatorio con distribución normal de parámetros desconocidos. Teniendo en cuenta que una muestra de 20 autobuses daba para este mismo recorrido una media de 32 minutos y una varianza de 25, determinar:
a) un intervalo de confianza para el tiempo promedio en el que se cubre dicho recorrido, con un 95% de confianza.
b) el nivel de confianza que permitiría trabajar con una precisión de (+/-) 2,4 minutos.

45.- Sean dos muestras distintas que han sido diseñadas

para estimar el gasto medio anual en talleres mecánicos de automóvil.

	coches < 25000 k. de rodaje	coches > 60000 k. de rodaje
Tamaño muestral	137	104
Gasto medio	9642.7	18366.5
Desviación del gasto	1421.1	3298.1

Obtener con un nivel de confianza del 80%:

a) un intervalo para el gasto medio anual de cada grupo de automóviles.
b) un intervalo de confianza para la diferencia de gasto medio entre uno y otro grupo.