INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

Si realizamos una recapitulación de lo estudiado hasta el momento, encontramos tres partes bien diferenciadas:

1) ESTADÍSTICA DESCRIPTIVA
2) CÁLCULO DE PROBABILIDAD
3) DISTRIBUCIONES DE PROBABILIDAD

ESTADISTICA DESCRIPTIVA

En ella se aprende una serie de técnicas para organizar, presentar y analizar un conjunto finito de observaciones, que según el objetivo del estudio, constituyen una población o una muestra.

CÁLCULO DE PROBABILIDAD

En esta parte se define la probabilidad como una medida de la posibilidad de ocurrencia de cada resultado de una experiencia aleatoria, extendiendo la noción de frecuencia relativa a las poblaciones infinitas.

DISTRIBUCIONES DE PROBABILIDAD

A través de ellas se presentan modelos matemáticos del comportamiento (en términos probabilísticos) de las poblaciones.
Cada distribución surge como consecuencia de hipótesis establecidas sobre el comportamiento del fenómeno aleatorio analizado.
Tales hipótesis son las que permiten identificar una población con la correspondiente distribución. A su vez, cada distribución depende de parámetros matemáticos cuyo valor hemos supuesto conocido.

En la cuarta y última parte de este curso se estudian métodos que nos permiten obtener los valores de tales parámetros poblacionales basándonos en los resultados muestrales. En estos métodos se encuentra una integración de las tres partes anteriores, ya que usan a la probabilidad como una medida de la confianza de nuestras conclusiones.

MUESTREO

Sabemos que una muestra es un subconjunto finito de una población.

Nada hemos dicho, hasta ahora, de cómo obtener la misma, es decir, de cómo se realiza la selección de las *unidades elementales*, sobre las cuales se observa o mide una característica de interés (*variable*) y cuyos valores constituyen la muestra.

En el párrafo anterior aparecen dos conceptos claves en todo problema de muestreo. Ellos son: unidades elementales y variable.

Ambos deben ser definidos previo a la selección de la muestra.

Un planteo correcto del objetivo del muestreo, lleva implícito una definición precisa de la población a analizar y, en consecuencia, una correcta identificación de las unidades elementales y la variable que se hayan asociadas a tal población.

Consideremos por ejemplo un lote de 100 artículos enviados por un fabricante a un cliente. Supongamos que el cliente está interesado en analizar la *calidad de los artículos*. Así planteado el problema indica que las observaciones se realizarán sobre los artículos, siendo por lo tanto *cada artículo una unidad elemental*. La observación de la calidad obliga al cliente a definir *qué es la calidad*, es decir, qué observará en cada artículo (unidad elemental) para decidir sobre la misma.

Si sólo le interesa clasificar los artículos en buenos o defectuosos, o si le interesa determinar un intervalo de valores para la característica en observación (longitud, diámetro, duración, etc.) En el primer caso la variable en estudio es la calidad del artículo, en el segundo la característica elegida. Si la variable es la calidad del artículo, ésta toma dos valores: bueno o defectuoso.

El planteo ambiguo del problema con respecto al objetivo del análisis nos lleva a considerar dos opciones:

1º) Si el cliente desea sólo concluir con respecto a la calidad de los artículos que componen el lote, la población estará constituida por todos los valores (buenos o defectuosos) correspondientes a los 100 artículos. Estamos ante una *población finita*.

2ª) Si el cliente desea concluir con respecto a la calidad del proceso de producción del fabricante, la población estará formada por los infinitos valores (buenos o defectuosos) correspondientes a los infinitos artículos que se producirán bajo este proceso si éste continuara operando indefinidamente. Evidentemente la *población es infinita* y en este caso los valores de la variable de los 100 artículos que constituyen el lote, son una muestra de tal población.

La diferencia crucial que determina si el lote debe ser considerado una población o una muestra dependerá del tipo de decisión a tomarse: si va a evaluarse la calidad de este lote en particular o la calidad del proceso de manufactura del proveedor.

Una vez que el objetivo del estudio se ha especificado, la población queda identificada, y en consecuencia el conjunto de las unidades elementales. Ahora la muestra ya puede ser seleccionada.

Existen dos métodos de selección de muestras:

1º) MÉTODOS NO PROBABILÍSTICOS

En estos métodos la selección de la muestra se realiza de una manera subjetiva, decidiendo el observador las unidades elementales a analizarse.

2º) MÉTODOS PROBABILÍSTICOS

Con ellos las unidades elementales se seleccionan a través de métodos aleatorios. La ventaja de estos métodos con respecto al primero es que permite proporcionar una medida, expresada en probabilidad, de extraer conclusiones erróneas acerca de la población. Es decir permite controlar los llamados *errores de muestreo*, que son los que se producen al inferir de la muestra a la población, por el hecho de no trabajar con la población completa sino con un subconjunto de la misma.

Existen otro tipo de errores, no asignables al muestreo en sí, sino al plan de muestreo, y a los que el muestreo probabilística no controla. Es muy frecuente que un plan de muestreo mal diseñado nos lleve a muestrear una población que no es la del objeto de estudio.

Así por ejemplo si se quiere analizar cierta característica de los habitantes de la ciudad de Rosario y la muestra se elige seleccionando nombres al azar de la guía telefónica, la población física muestreada resulta ser la formada por los habitantes de la ciudad de Rosario que poseen teléfono y todas las conclusiones que se extraigan a partir de esta muestra serán válidas para tal población pero no para *todos* los habitantes de Rosario.

MUESTRAS ALEATORIAS SIMPLES

Sea **X** la variable aleatoria que representa la población en estudio y **f(x)** su función de densidad.

Diremos que una muestra extraída de esta población es de extensión **n** si consta de **n** observaciones. Este conjunto de n observaciones puede ser representado como un vector numérico n dimensional $(x_1, x_2, \ldots, x_n)$.

Supongamos que extraemos sucesivas muestras aleatorias de extensión n de la mencionada población. Los vectores que representan a las distintas muestras son:

$(x_1^1, x_2^1, \ldots, x_n^1)$ 1ª muestra
$(x_1^2, x_2^2, \ldots, x_n^2)$ 2ª muestra
⋮
$(x_1^r, x_2^r, \ldots, x_n^r)$ r-ma muestra
siendo
x_i^j el valor de la i-ma observación de la j-ma muestra.

Evidentemente no tenemos por qué pensar que el valor de la primera observación, para cada una de las muestras, va a ser el mismo. Por el contrario, es lógico suponer que existe variabilidad. El mismo razonamiento podemos hacer para las iésimas observaciones de las r muestras.
Esto quiere decir que *antes* de la extracción de la muestra, cada una de las observaciones puede ser pensada como una variable aleatoria, en consecuencia una muestra aleatoria puede ser representada como un vector aleatorio n dimensional y la notaremos
$\rightarrow$
$M = (X1, X_2, \ldots, X_n)$
Siendo
$\rightarrow$
$M_o = (x_1, x_2, \ldots, x_n)$
un valor observado de la muestra aleatoria.
En particular llamaremos *MUESTRA ALEATORIA SIMPLE* (M.A.S.) a una muestra aleatoria que verifica:
1º) Cada una de las variables aleatorias X_i tiene la misma función de densidad f(x) que la variable X en estudio y por lo tanto se verifica:
$E(X_i) = E(X)$
$V(X_i) = V(X)$
2º) Las variables aleatorias X_i son independientes entre sí.
Observemos que el primer supuesto nos indica que para cada observación a realizar la población debe permanecer inalterada e igual a la original.
El segundo supuesto pide que la aparición de una observación no aumente o disminuya la probabilidad de aparición de otras observaciones.

En caso de población finita estos supuestos exigen que el muestreo se realice *con reposición.*

Si la población es infinita el muestreo puede ser con o sin reposición

INFERENCIA ESTADÍSTICA PARAMÉTRICA

Una vez obtenidos los valores de una muestra, ellos serán usados con el objeto de obtener información con respecto a la población de la cual la muestra fue extraída.

Recordemos nuevamente que una población queda identificada al dar: la variable aleatoria, su distribución de probabilidad y sus parámetros matemáticos; es decir al dar X y f(x, θ), función de densidad de X con parámetro matemático θ.

Supongamos que la ley f resulta conocida ya sea por experiencias pasadas o por hipótesis sobre el fenómeno en estudio pero desconocemos el valor del parámetro.

Así por ejemplo en un proceso de producción se conoce que la introducción de una modificación en el mismo produce un desplazamiento de la distribución, es decir la ley de distribución es la misma pero se corre la esperanza matemática, siendo este nuevo valor desconocido.

Otro ejemplo es el caso de una población que surge por la variabilidad de las mediciones de una magnitud δ con un determinado proceso de medición. Podemos suponer que las mediciones tienen distribución normal por el teorema central del límite, y además podemos conocer la precisión del instrumento, es decir σ^2. Luego nos interesará estimar el parámetro δ que coincide con la esperanza matemática de la distribución.

Son dos los tipos de problemas a los que nos podemos enfrentar cuando necesitamos información acerca del valor de un parámetro:

- La necesidad de darle un valor numérico al parámetro que servirá como aproximación del valor exacto, pero desconocido del mismo, por ejemplo para cálculos posteriores de probabilidades.
- Nos interesa conocer no un valor particular del parámetro sino un rango de valores posibles, es decir si excede un número dado, si es menor que éste o dentro de qué intervalo tiene su posible valor.

El primer caso es un problema de *estimación puntual* mientras que el segundo es de *estimación por intervalos de confianza*, aunque la separación entre ambas formas de estimación no es tan neta sino que se encuentran íntimamente relacionadas como veremos más adelante.

ALGUNOS ESTADÍSTICOS Y SUS DISTRIBUCIONES

Sea X una variable aleatoria con esperanza matemática μ y variancia σ^2 y ($X_1, X_2, \ldots, X_n$) una M.A.S. de tamaño n.

Si Y = H ($X_1, X_2, \ldots, X_n$) es una variable aleatoria que surge como función del vector aleatorio muestral, Y es llamado un *estadístico.*

Los estadísticos que analizaremos en particular son:

$\overline{X} = \frac{1}{n}\sum X_i$ media muestral

$S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2$ variancia muestral

Tanto $\overline{X}$ como S^2 son variables aleatorias (los valores que asumen pueden variar de una muestra a otra)

LA VARIABLE ALEATORIA MEDIA MUESTRAL ($\overline{X}$)

Si $\overline{X} = \frac{1}{n}\sum X_i$, bajo el supuesto de que ($X_1, X_2, \ldots, X_n$) es una M.A.S. de X (cada variable aleatoria X_i tiene la misma distribución y los mismos parámetros que la variable aleatoria X de la cual la muestra fue extraída, es decir $E(X_i) = \mu$), entonces la esperanza matemática de $\overline{X}$ es:

$E(\overline{X}) = E(\frac{1}{n}\sum X_i) = \frac{1}{n}\sum E(X_i) = \frac{1}{n}\, n\, \mu = \mu.$

Por otra parte la variancia de $\overline{X}$ es: $V(\overline{X}) = V(\frac{1}{n}\sum X_i) = \frac{1}{n^2}\sum V(X_i)$

Como las X_i son independientes entre sí, y además $V(X_i) = \sigma^2 \ \forall\ i$ resulta:

$V(\overline{X}) = \frac{1}{n^2}\sum V(X_i) = \frac{\sigma^2}{n}$

Por lo tanto la variancia de la variable aleatoria $\overline{X}$ es la variancia de la variable X dividido el tamaño de la muestra.

Estas dos propiedades de los parámetros de $\overline{X}$ nos indican que cualquiera sea la distribución de la misma, a medida que aumenta el tamaño de la muestra, la $V(\overline{X})$ tiende a cero y en consecuencia las medias muestrales tienden a concentrarse alrededor del parámetro μ.

Con respecto a la distribución de $\overline{X}$ podemos decir que:

a) si la variable $X \sim N(\mu, \sigma^2)$ entonces por la propiedad reproductiva de la distribución normal, $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$,

b) si la variable X tiene cualquier distribución, pero $n \geq 30$, por el Teorema Central del Límite, la distribución de $\overline{X}$ tiende a $N(\mu, \frac{\sigma^2}{n})$.

LA VARIABLE ALEATORIA VARIANZA MUESTRAL (S^2)

Presentamos la distribución de la variable aleatoria S^2 sólo en el caso en que la variable en estudio $X \sim N(\mu, \sigma^2)$.

Bajo este supuesto la variable aleatoria $\frac{(n-1)S^2}{\sigma^2}$ tiene una distribución chi-cuadrada con n -1 grados de libertad (notamos $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$).

Además $E(S^2) = \sigma^2$ y $V(S^2) = \frac{2\sigma^4}{n-1}$

Observamos que la media poblacional de S^2 coincide con la variancia de X y la variancia de S^2 tiende a cero cuando n crece. Al crecer el número de observaciones la distribución de S^2 se concentra cada vez más alrededor del valor σ^2.

ESTIMACION PUNTUAL - ERROR DE ESTIMACION – ESTIMACION POR INTERVALOS DE CONFIANZA

Cuando un estadístico es usado para obtener información con respecto al valor de un parámetro poblacional se lo llama *estimador.*

Si θ es un parámetro desconocido, al estimador de θ lo notamos $\hat{\theta}$.

De las propiedades analizadas en las distribuciones de $\overline{X}$ y S^2, surge que estos estadísticos son buenos estimadores de la esperanza poblacional μ y de la varianza poblacional σ^2 respectivamente, en el sentido de que las distribuciones de probabilidad de los mismos las podemos concentrar tanto como queramos alrededor de los parámetros desconocidos (μ o σ^2 respectivamente) aumentando el tamaño de la muestra.

Luego

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

$\hat{\sigma}^2 = S^2$

Dijimos que realizar una estimación puntual es asignarle al parámetro desconocido un valor, o sea un número.

Este valor se obtiene partiendo de los resultados muestrales $(x_1, x_2, \ldots, x_n)$. Se calcula el valor del estimador elegido, el que se le dará al parámetro desconocido.

O sea:

A μ se asigna $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$

A σ^2 se asigna $s^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2$

Dado que el valor de estos estimadores está dependiendo de la muestra obtenida, no tenemos porque pensar que el mismo coincidirá con el valor del parámetro a estimar. Sabemos que los valores posibles de cada estimador presentan variabilidad dentro de un determinado rango.

Esto nos lleva a tratar de medir el error que cometemos cuando a un parámetro le asignamos el valor del estimador, es decir, el *error de estimación.*

Trataremos cada caso por separado:

a) Estimación de μ con σ^2 conocido

b) Estimación de μ con σ^2 desconocido

c) Estimación de la proporción poblacional (p)

d) Estimación de la varianza poblacional (σ^2)

a) Estimación de μ con σ^2 conocido

Sea X una variable aleatoria con distribución normal, E(X) = μ desconocida y varianza σ^2 conocida.

Con la finalidad de estimar μ se extrae una muestra de tamaño n que asume los valores (x_1, x_2,. . ., x_n). En la misma se calcula $\bar{x}$. Este es el valor que se toma como estimación puntual de μ.

¿Qué error se comete al asignarle a μ el valor de $\bar{x}$?

El error de estimación se mide por $|\bar{x} - \mu|$.

Para poder conocer con exactitud cuánto vale $|\bar{x} - \mu|$ deberíamos conocer el valor exacto de μ; no es esta nuestra situación, por lo tanto debemos contentarnos con dar una cota, ε, del error de estimación, a través de analizar los valores posibles de $\overline{X}$ cuando la muestra es de tamaño n.

La situación ideal sería poder obtener el valor de ε con certeza, sin embargo sabemos que a partir de una muestra no podemos obtener conclusiones acerca de la población con

seguridad total, así es que debemos ser menos ambiciosos y aceptar trabajar con una probabilidad 1 - α cercana a 1, llamada *coeficiente de confianza*. Luego la pregunta anterior debe ser formulada de la siguiente manera:

¿Cuál es el máximo error de estimación que podemos cometer con probabilidad 1 - α, al asignarle a μ el valor de $\bar{x}$?

Es decir debemos encontrar ε tal que se verifique

$$P(|\bar{X} - \mu| < \varepsilon) = 1 - \alpha$$

Esto es equivalente a:

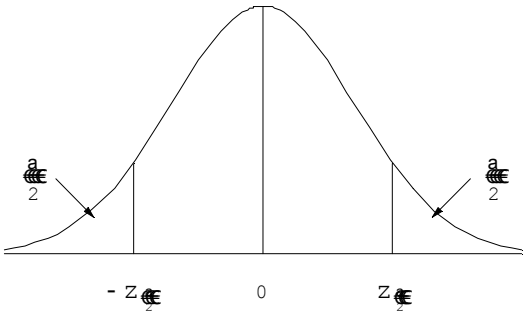
$$P(\mu - \varepsilon < \bar{X} < \mu + \varepsilon) = 1 - \alpha$$

Estandarizando obtenemos:

$$P(-\frac{\varepsilon}{\sigma}\sqrt{n} < Z < \frac{\varepsilon}{\sigma}\sqrt{n}) = 1 - \alpha$$

El valor $\frac{\varepsilon}{\sigma}\sqrt{n}$ debe ser igualado a un valor $z_{\frac{\alpha}{2}}$ que es el valor de la variable normal estándar Z que verifica

$$P(Z \leq -z_{\frac{\alpha}{2}}) = P(Z \geq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$



Luego

$$\varepsilon = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \qquad (1)$$

Dado que σ es un valor supuesto conocido, n está dado y el valor de z también es fijo ya que depende de la confianza fijada (1 - α), luego el valor de ε puede ser calculado.

Observemos que ε se encuentra en relación inversa al tamaño de la muestra (a mayor tamaño de muestra, menor error de estimación), y en relación directa a la confianza (a mayor confianza, mayor error de estimación).

Supongamos que el error calculado no resulta satisfactorio (demasiado grande), para disminuirlo debemos disminuir la confianza o aumentar el tamaño de la muestra. Si la confianza no se quiere modificar, nos queda como opción modificar n.

¿Cuántas observaciones son necesarias para que al estimar μ con $\bar{x}$,el error máximo de estimación sea ε (fijado) con una confianza (1 - α) (fijada)?

De la expresión (1) obtenemos:

$$n = z^2_{\frac{\alpha}{2}} \cdot \frac{\sigma^2}{\varepsilon^2}$$

El valor del error obtenido en (1), indica que:

$$P(\left|\bar{X} - \mu\right| < z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

trabajando algebraicamente obtenemos:

$$P(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

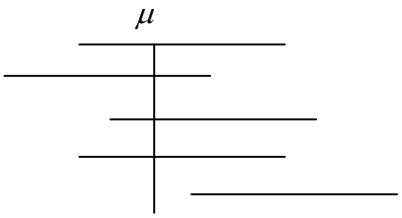
$(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}})$ es un INTERVALO ALEATORIO para el parámetro μ.

Una vez que la muestra ha sido extraída y $\bar{x}$ calculada, reemplazando en la expresión anterior del intervalo aleatorio, obtenemos el INTERVALO DE CONFIANZA para μ, que es un intervalo numérico.

$$(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}})$$

En un intervalos aleatorio, la parte aleatoria son los extremos del mismo, mientras que el parámetro es un valor fijo. Por lo tanto la probabilidad (1 - α) debe ser interpretada como la probabilidad de que un intervalo aleatorio cubra el verdadero valor del parámetro. Pensada la probabilidad como una frecuencia relativa nos indica que si se extraen un número suficientemente grande de muestras de extensión n y con cada

una de ellas se construye un intervalo de confianza para μ, aproximadamente (1 - α)% de tales intervalos cubrirán en verdadero valor de μ.



Cuando el intervalo de confianza ha sido calculado, éste cubre o no el verdadero valor del parámetro, por lo tanto pierde sentido hablar de la probabilidad (1 - α), este valor debe ser interpretado como una medida de la *confianza* del experimentador de obtener el cubrimiento de μ con el intervalo calculado.

Un ejemplo:
Un fabricante produce anillos para los pistones de un motor de automóvil. El diámetro de un anillo es una variable aleatoria X con distribución normal y desviación estándar σ = 0.001 mm. Para una muestra aleatoria de 15 anillos se observó un diámetro promedio $\bar{x}$ = 74.036 mm. Obtenga un intervalo de confianza del 95 % y 99 % para el diámetro promedio, es decir E(X).

Si $\bar{x}$ es la media muestral observada en una muestra aleatoria de tamaño n, de una variable aleatoria X con distribución normal y variancia σ^2 conocida, entonces un intervalo de confianza para μ = E(X) del 100 (1 - α) % está dado por:

$$(\bar{x} - z_{\frac{\alpha}{2}} . \frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + z_{\frac{\alpha}{2}} . \frac{\sigma}{\sqrt{n}})$$

Para α = 0.05 se obtiene:

$$(74.036 - 1.96 . \frac{0.001}{\sqrt{15}} \ , \ 74.036 + 1.96 . \frac{0.001}{\sqrt{15}}) = (74.0355, \ 74.0365)$$

Para α = 0.01 se obtiene:

$$(74.036 - 2.58 . \frac{0.001}{\sqrt{15}} \ , \ 74.036 + 2.58 . \frac{0.001}{\sqrt{15}}) = (74.0353, \ 74.0367)$$

Observamos que para un tamaño de muestra fijo, a mayor confiabilidad se corresponde menor precisión ¿Es esto razonable?

b) Estimación de μ con σ^2 desconocido

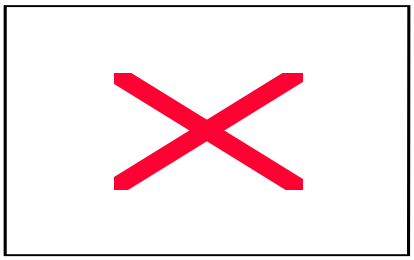
Dado que la distribución de $\overline{X}$ depende de la varianza poblacional σ^2, cuando esta es desconocida debe ser estimada a través de S^2.

El estadístico

$$\frac{\overline{X} - \mu}{S}\sqrt{n}$$

deja de tener una distribución normal estandarizada y se le conoce su distribución sólo en el caso en que la variable en estudio X esté distribuida NORMALMENTE. En tal situación la distribución del estadístico mencionado es la DISTRIBUCION t STUDENT con N -1 GRADOS DE LIBERTAD.

Esta distribución t es de forma campanular y simétrica con eje de simetría en x = 0, siendo su parámetro matemático un número natural n llamado GRADOS DE LIBERTAD.



Cuando el número de grados de libertad tiende a infinito, la distribución t-student se aproxima a una distribución normal estandarizada.

Para estimar la esperanza matemática de una variable aleatoria $X \sim N(\mu, \sigma)$ – ambos parámetros desconocidos), extraemos una M.A.S. de tamaño n y sobre ella calculamos $\overline{x}$, que tomaremos como valor del parámetro μ.

Realizando el mismo razonamiento que en el caso a), el análisis del error de estimación se efectúa a través de la distribución de $\overline{X}$. Es decir que fijado el tamaño de la muestra y la confianza deseada, queremos calcular la cota de error partiendo de:

$$P(|\overline{X} - \mu| < \varepsilon) = 1 - \alpha \qquad (2)$$

Donde ε es desconocido.

Recordemos que:

$$\frac{\overline{X} - \mu}{S}\sqrt{n} \sim t_{n-1}$$

La expresión (2) puede transformarse en

$$P\left(\left|\frac{\overline{X} - \mu}{S}\sqrt{n}\right| < \frac{\varepsilon}{S}\sqrt{n}\right) = 1 - \alpha$$

$\frac{\varepsilon}{S}\sqrt{n}$ debe ser igualado a $t_{n-1,\frac{\alpha}{2}}$, donde $t_{n-1,\frac{\alpha}{2}}$ es el valor de una variable aleatoria t-student con n-1 grados de libertad que verifica:

$$P\left(\frac{\overline{X} - \mu}{S}\sqrt{n} > t_{n-1,\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \quad \text{y} \quad P\left(\frac{\overline{X} - \mu}{S}\sqrt{n} < -t_{n-1,\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$

Luego $P\left(-t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \le \overline{X} - \mu \le t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$ o equivalentemente

$$P\left(\overline{X} - t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

$\left(\overline{X} - t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)$ es un intervalo aleatorio de μ, mientras que

($\bar{x}$ - $t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu <$ $\bar{x}$ + $t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$) es un intervalo de confianza (sus extremos son valores numéricos)

Observemos que $\varepsilon = t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ depende al igual que en el caso a) de la confianza fijada y del tamaño de muestra elegido, pero se diferencia de aquel en que depende del valor que asume la variable aleatoria S. Por lo tanto la cota del error resulta ser aleatoria.

Una vez que la muestra fue extraída, si el valor de ε resulta inapropiado, podemos disminuirlo reduciendo la confianza o aumentando el tamaño de la muestra. Señalemos que en este caso, el valor de n necesario para obtener la cota del error deseada, no puede ser determinado algebraicamente, en razón de que el valor de t también depende del tamaño de la muestra. Lo único que podemos concluir es que el tamaño de muestra debe ser aumentado, pero no sabemos cuánto.

Un ejemplo

Se seleccionaron al azar 15 resistores de la producción de un proceso. La resistencia media observada en la muestra fue de 9.8 ohms, mientras que la desviación estándar muestral fue de 0.5 ohms. Determine un intervalo de confianza del 95% para la resistencia media poblacional. Se supone que la variable en estudio tiene distribución normal.

Si $\bar{x}$ y s son la media aritmética y la desviación estándar observada en una muestra de tamaño n, de una variable X con distribución normal y variancia σ^2 desconocida, entonces un intervalo de confianza para μ_X = E(X) del 100(1-α)% está dado por:

$$(\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}})$$

Para α=0.05 se obtiene de la tabla el valor t=2.145 resultando el intervalo de confianza para μ_X : $(9.8 - \frac{2.145x0.5}{\sqrt{15}} ; 9.8 + \frac{2.145x0.5}{\sqrt{15}}) = (9.523 ; 10.077)$

c) Estimación de la proporción poblacional (p)

En ocasiones nos interesa conocer la proporción p o frecuencia relativa de veces que se presenta cierto suceso A en una población, o lo que es equivalente, conocer la probabilidad de que ocurra el suceso A.

Sea por ejemplo el suceso A: una unidad producida por un proceso es defectuosa. Supongamos que P(A)=p es desconocida.

Para estimar p vamos a considerar una variable aleatoria X a la que le asignamos el valor 1 cuando ocurre el suceso A (una unidad es defectuosa) y el valor 0 cuando ocurre el suceso $\bar{A}$ (una unidad es buena).

La variable aleatoria X que asume los valores 0 y 1 con probabilidades 1-p y p respectivamente, se denomina variable aleatoria con distribución de Bernoulli, de parámetro p. Para tal variable verifique que: E(X)=p y V(X)=p(1-p).

Si se inspeccionan en forma independiente n unidades del proceso de producción y se anotan los valores para $X_1, X_2, ...,X_n$ donde X_i =1 si la iésima unidad inspeccionada tiene defectos y X_i =0 si no es así, entonces una variable de interés es:

$Y = X_1 + X_2 + ...+X_n$ que representa el número total de unidades defectuosas en la muestra de tamaño n. ($X_1, X_2, ...,X_n$ constituye una M.A.S de X). La variable aleatoria $\frac{Y}{n} = \frac{X_1 + X_2 + ...X_n}{n}$ denota la frecuencia relativa de unidades defectuosas en una muestra de tamaño n y verifica:

$E(\frac{Y}{n}) = E(\frac{X1 + X2 + ... + Xn}{n}) = \frac{1}{n}[E(X_1) + E(X_2) + ...+E(X_n)] = \frac{1}{n}np = p$

$V(\frac{Y}{n}) = V(\frac{X1 + X2 + ... + Xn}{n}) = \frac{1}{n^2}[V(X_1) + V(X_2) + ...+V(X_n)] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$

Por el teorema del límite central $\frac{Y}{n}$ tiende a distribuirse normalmente con parámetros p y $\frac{p(1-p)}{n}$.

Usaremos $\frac{Y}{n}$ como estimador de p por cuanto para n convenientemente grande la variable aleatoria $\frac{Y}{n}$ asume valores que se concentran alrededor de p.

Si planteamos $P(\left|\frac{Y}{n} - p\right| < \varepsilon) = 1 - \alpha$ y operamos del mismo que en a) resulta:

P($\frac{Y}{n} - z\sqrt{\frac{p(1-p)}{n}} < p < \frac{Y}{n} + z\sqrt{\frac{p(1-p)}{n}}) = 1-\alpha$, donde z es un valor que se obtiene de la tabla normal estándar o reducida, que verifica: P($Z \leq z$)=1-$\frac{\alpha}{2}$ o equivalentemente P($Z \geq z$)= $\frac{\alpha}{2}$.

Observamos la existencia de un problema que no había aparecido antes. Los límites del intervalo aleatorio que hemos obtenidos están dependiendo del parámetro que se desea estimar.

El problema puede superarse si sustituimos el valor de p por el valor de la frecuencia relativa observada en la muestra, es decir el valor que asume $\frac{Y}{n}$ en la muestra y que notamos con f_A (frecuencia relativa del suceso A en la muestra)

De este modo $(f_A - z\sqrt{\frac{f_A(1-f_A)}{n}}, f_A + z\sqrt{\frac{f_A(1-f_A)}{n}})$ constituye un intervalo de confianza para p.

Observación:

Podemos obtener una cota del error $\varepsilon = z\sqrt{\frac{p(1-p)}{n}}$ si tenemos en cuenta que la función cuadrática g(p) = p(1-p) para $0 \leq p \leq 1$ asume su valor máximo cuando p=$\frac{1}{2}$. Para p=$\frac{1}{2}$, g($\frac{1}{2}$) = $\frac{1}{4}$, luego $\varepsilon \leq z\sqrt{\frac{1}{4n}} = z\frac{1}{2\sqrt{n}}$

Un ejemplo

Una inspección cuidadosa de 70 soportes de concreto precolado reveló que 28 estaban fisurados. Construya un intervalo de confianza del 95% de la verdadera proporción de soportes con fisura.

Sea A :un soporte de concreto precolado está fisurado.

De acuerdo a los datos $f_A = \frac{28}{70}$

De la tabla de la normal estándar o reducida se obtiene para un nivel de confianza del 95% el valor z=1.96 (P($Z \leq 1.96$) =0.975)

Luego un intervalo aproximado del 95% de confianza para p es:

$$(\frac{28}{70}-1.96\sqrt{\frac{\frac{28}{70}(1-\frac{28}{70})}{70}}\ ;\ \frac{28}{70}+1.96\sqrt{\frac{\frac{28}{70}(1-\frac{28}{70})}{70}}) = (0.285\ ;\ 0.515)$$

d) Estimación de la variancia en una población con distribución normal

Ya hemos visto que la variable aleatoria S^2 es un buen estimador de la variancia σ^2 en razón de que $E(S^2) = \sigma^2$ y $V(S^2) = \frac{2}{n-1}\sigma^4$.

En la unidad anterior se vio que si X_1, X_2, ...X_n son n variables aleatorias independientes, donde cada una tiene distribución N(0 , 1) entonces la variable aleatoria $T= X_1^2+ X_2^2+ ....+X_n^2$ tiene una distribución chi –cuadrada con n grados de libertad y notamos: $T\sim \chi^2_n$

Si $X \sim N(\mu , \sigma)$ y X_1, X_2, ...,X_n es una M.A.S de X entonces $\sum_i^n (\frac{Xi-\mu}{\sigma})^2 \sim \chi^2_n$.

Cuando se sustituye la media poblacional μ por la media muestral $\overline{X}$, la variable aleatoria resultante tiene una distribución chi cuadrada con n-1 grados de libertad.

Se nota: $\sum_1^n (\frac{Xi-\overline{X}}{\sigma})^2 \sim \chi^2_{n-1}$

Siendo $S^2=\frac{1}{n-1}\sum_1^n (Xi-\overline{X})^2$ podemos concluir que $\frac{(n-1)S^2}{\sigma^2} = \sum_1^n (\frac{Xi-\overline{X}}{\sigma})^2 \sim \chi^2_{n-1}$

cuando X_1, X_2, ...,X_n es una M.A.S de una variable aleatoria X, normalmente distribuida con media μ y desviación estándar σ.

Si planteamos $P(c_1 \le \frac{(n-1)S^2}{\sigma^2} \le c_2) = 1-\alpha$ y operamos algebraicamente obtenemos

que: $P(\frac{(n-1)S^2}{c_2} < \sigma^2 < \frac{(n-1)S^2}{c_1}) = 1-\alpha$, donde c_1 y c_2 son valores que se obtienen de la tabla chi cuadrada y verifican:

$P(\frac{(n-1)S^2}{\sigma^2} \ge c_1) = 1-\frac{\alpha}{2}$ y $P(\frac{(n-1)S^2}{\sigma^2} \ge c_2) = \frac{\alpha}{2}$.

En síntesis:

- $(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1})$ es un intervalo aleatorio que contiene con probabilidad 1-α a σ^2 , siendo P($\frac{(n-1)S^2}{\sigma^2} \geq c_1$) = 1-$\frac{\alpha}{2}$ y P($\frac{(n-1)S^2}{\sigma^2} \geq c_2$) = $\frac{\alpha}{2}$.
- $(\frac{(n-1)s^2}{c_2}, \frac{(n-1)s^2}{c_1})$ es un intervalo con (1-α)100% de confianza para σ^2 .

Un ejemplo

En la producción de resistores, la variancia de las resistencias refleja la estabilidad del proceso de manufactura. Se desea estimar con un nivel de confianza igual a 0.90, la variancia poblacional de las resistencias, sabiendo que en una muestra de 15 resistores se observó una desviación estándar igual a 0.5 ohms.

De la tabla chi cuadrado, para 14 grados de libertad, se obtienen los valores:

c_2 = 23.68 y c_1 =6.57 (la probabilidad de que una variable aleatoria con distribución chi cuadrada y 14 grados de libertad supere los valores 23.68 y 6.57 es 0.05 y 0.95 respectivamente)

A partir de los datos de la muestra, el intervalo con 90% de confianza para σ^2 es:

$$\left(\frac{14(0.5)^2}{23.68};\frac{14(0.5)^2}{6.57}\right) = (0.148 ; 0.533)$$

BIBLIOGRAFÍA

- Canavos, George. " Probabilidad y Estadística. Aplicaciones y Métodos". México. McGraw Hill
  1988.
- Devore Jay L. "Probabilidad y Estadística para Ingeniería y Ciencias". México. Thomson Editores 2001.
- Meyer Paul L. " Probabilidad y Aplicaciones Estadísticas". México. Addison Wesley Iberoamericana 1993.
- Miller Irwin y Freund J. "Probabilidad y Estadística para Ingenieros". Prentice Hall 1993.
- Milton Susan, Arnold Jesse. "Probabilidad y Estadística con aplicaciones para ingeniería y ciencias computacionales". México. McGrawHill 2004.
- Montgomery Douglas, Runger George. " Probabilidad y Estadística Aplicadas a la Ingeniería". México. McGraw Hill 1996.
- Navidi,William. "Estadística para ingenieros y científicos". McGrawHill 2006.
- Scheaffer Richard, McClave James. " Probabilidad y Estadística para Ingeniería" . México. Grupo Editorial Iberoamericana1993.
- Walpole Ronald, Myers Raymond. " Probabilidad y Estadística" . México. McGrawHill 2001.

En estos textos podrá ahondar en el tema como asimismo encontrar más ejemplos y problemas para resolver.