Tema 7
Intervalos de confianza

Hugo S. Salinas

Introducción

Hemos definido la **inferencia estadística** como un proceso que usa información proveniente de la muestra para generalizar y tomar decisiones acerca de toda la población en estudio. Sin embargo, hasta el momento hemos trabajado la muestra y la población por separado.

Luego, trabajamos herramientas útiles en el análisis exploratorio de los datos provenientes de una muestra, tanto gráficos como resúmenes numéricos para extraer información de interés para la inferencia. Hablamos de distribuciones de frecuencias y estadísticos.

En el tema 6, a través del lenguaje de la probabilidad, tratamos los modelos para las poblaciones que pueden ser de interés, sobre las cuales nos interesa sacar conclusiones, o tomar una decisión. Definimos las variables aleatorias, sus distribuciones de probabilidad, parámetros y algunos modelos frecuentes.
Podemos hacer un cuadro comparativo entre características del análisis exploratorio de datos y de la inferencia estadística:

Introducción cont.

Análisis exploratorio de datos	*Inferencia estadística*
Su objetivo es la exploración de los datos muestrales, en busca de regularidades interesantes.	Su objetivo es responder preguntas concretas sobre la población, planteadas antes de la obtención de los datos.
Las conclusiones sólo se aplican a las unidades de análisis y a las circunstancias para las cuales se obtuvieron los datos.	Las conclusiones se extienden a toda la población en estudio.
Las conclusiones se basan en lo que "vemos" en los datos.	Las conclusiones se explicitan con un grado de confianza.

Muchas de las técnicas utilizadas en inferencia exigen, también, que la distribución de los datos tenga determinadas características. El análisis de datos es de gran ayuda en este aspecto, para descubrir observaciones atípicas y otras desviaciones que puedan perturbar una correcta inferencia. Por lo tanto, en la práctica podemos observar como el análisis exploratorio de los datos y la inferencia estadística se complementan.

Introducción cont.

Como se sabe, muy frecuentemente es necesario seleccionar una **muestra** de unidades de la población, para extraer conclusiones respecto de la misma, en base a las observaciones muestrales .

Resumiendo:

Cuando el interés reside en generalizar las conclusiones de los resultados observados a la población en estudio o queremos tomar una decisión sobre la población en base a una muestra, estamos frente a un problema de inferencia estadística. Para que este proceso sea adecuado, debemos tener en cuenta:

- ❖ Plantear claramente el problema.
- ❖ Delimitar la población en estudio.
- ❖ Definir si el objetivo reside en estimar el valor de un parámetro desconocido de la población (por ej. **μ, σ, p**) a partir de un ***estadístico*** calculado con los datos de una muestra o decidir sobre valores hipotéticos que asignamos a dichos parámetros.
- ❖ Hacer un correcto diseño para la obtención de los datos muestrales. Los resultados de las técnicas para la inferencia que se utilizarán sólo serán válidos si la muestra es obtenida por métodos aleatorios, que son los métodos que dan *confianza* de seleccionar **muestras representativas** de la población. Un buen diseño para la obtención de los datos, es la mejor garantía de que la inferencia tenga valor.
- ❖Tener en cuenta y verificar los requerimientos de las técnicas a aplicar

Parámetros y estadísticos

Un **parámetro** es un número que describe algún aspecto de la población en estudio. En la práctica, en la mayoría de los casos (población infinita, pruebas destructivas, etc) el valor del parámetro es desconocido.

Un **estadístico** es un número que se calcula a partir de los datos muestrales. Si se utiliza para estimar un parámetro desconocido, se le conoce con el nombre de **estimador.**

Tengamos en cuenta que el valor del parámetro es fijo, mientras que el valor de un estadístico está en función de la muestra seleccionada y por lo tanto podrá variar de una muestra a otra.

Si de alguna manera, pudiéramos medir la precisión de este proceso, es decir, si pudiéramos evaluar si el valor del estadístico va a estar cerca del valor del parámetro correspondiente, para cualquier muestra extraída de la población, entonces estaríamos en condiciones de hacer ***buenas inferencias***. Es aquí donde la técnica de muestreo y el tamaño de la muestra juegan un papel fundamental.

Trabajaremos con muestras aleatoria simples, en donde cada elemento de una muestra de tamaño n es una variable aleatoria, siendo X_1, X_2,…, X_n, variables independientes entre sí.

Sólo cuando se utiliza el azar para escoger los elementos que conforman una muestra, podemos describir cómo varía el estadístico. Al obtener de forma repetida de una población distintas muestras del mismo tamaño, podemos encontrar la ***distribución muestral del estadístico,*** como veremos ahora.

Distribución de la media muestral

Distribución muestral de la media muestral $\overline{X}$

Si las muestras aleatorias simples de tamaño n son tomadas de una población con media poblacional μ y desvío estándar poblacional σ, la distribución muestral de $\overline{X}$ tiene las siguientes propiedades:

⇨ 1) $\mu_{\overline{x}} = E(\overline{X}) = \mu$

Es decir, el promedio de todos los posibles valores de $\overline{X}$ es igual al parámetro μ

⇨ 2) $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

Cuando el tamaño de la muestra aumenta, la medida de dispersión disminuye. Es decir, a medida que el número de observaciones obtenidas aumenta, el promedio de los valores observados se acerca más y más a μ (Ley de los grandes números)

⇨ 3) Si la población de la cual se extraen las muestras es normal, la distribución de $\overline{X}$ es también normal con media y desvío como los dados en los puntos anteriores, para cualquier tamaño muestral n.

Distribución de la media muestral cont.

⇨ 4) Si la población de la cual se extraen las muestras no es normal, pero el tamaño muestral es "suficientemente" grande, la distribución de $\overline{X}$ es aproximadamente normal con media y desvío como los dados en los puntos anteriores. Suficientemente grande en la práctica significa un tamaño de muestra n ≥ 30 (Teorema Central del Límite).

El tamaño n de la muestra, necesario para que $\overline{X}$ se aproxime a una distribución normal depende de la distribución de la población. En el caso de que las muestras se extraigan de una población uniforme son suficiente 6 observaciones para que la distribución del promedio muestral sea aproximadamente normal.

⇨ 5) Si la población de la cual se extraen las muestras es normal, con media poblacional μ y desvío estándar poblacional σ, pero ésta es desconocida, se reemplaza σ por **S** (desvío estándar muestral) y la estadística $\frac{(\overline{x}-\mu)}{S/\sqrt{n}}$ deja de tener distribución normal estandarizada y tiene una distribución *t Student con n-1 grados de libertad* [(a)] :

$$\frac{(\overline{X}-\mu)}{S/\sqrt{n}} \sim t_{n-1;\alpha}$$

Distribución de la media muestral cont.

(a) La apariencia general de la **distribución t** es similar a la de la distribución normal estándar: ambas son simétricas y unimodales y el valor máximo de la ordenada se alcanza en la media $\mu = 0$. Sin embargo esta distribución tiene colas más amplias (o más pesadas) que la normal. Existe una distribución *t* distinta para cada tamaño de muestra. Una ***distribución t*** viene determinada por un parámetro llamado **grados de libertad**. A medida que aumentan los grados de libertad, la curva de densidad *t* se parece más a la curva de la *N(0,1),* ya que la estimación de σ por S (desviación estándar muestral) se va haciendo más precisa.

La propiedad 1 indica que el estimador **X es *insesgado,*** ya que el centro de su distribución muestral es igual al valor del parámetro poblacional correspondiente.

La propiedad 2 hace a la **variabilidad o precisión** del estimador y vemos que a medida que el tamaño muestral crece la precisión del estimador es mayor, ya que la variación alrededor del parámetro desconocido disminuye (propiedad de ***convergencia).*** Si la distribución de un estadístico muestra valores muy alejados, se dice que carece de precisión.

Idealmente buscamos un estimador que cumpla estas dos propiedades: que sea insesgado y convergente:

Distribución de la media muestral cont.

Un estadístico es **insesgado** si el centro de su distribución muestral es igual al valor del parámetro poblacional correspondiente.

Un estadístico es **convergente** si su desviación estándar disminuye a medida que el tamaño de muestra crece.

El estadístico $\overline{X}$, por poseer estas propiedades, es un buen estimador de μ.

Estas propiedades también se cumplen para la proporción muestral o frecuencia relativa (f_r) y la varianza muestral (S^2_{n-1}), siendo por lo tanto respectivamente, buenos estimadores de la proporción poblacional (p) y varianza poblacional (σ^2).

En general, la notación que utilizaremos para los estimadores es la siguiente:

Parámetro	Estimador
μ	$\hat{\mu} = \overline{X}$
p	$\hat{p} = f_r$
σ^2	$\hat{\sigma}^2 = S^2_{n-1}$

Distribución de la frecuencia relativa o proporción muestral

El estadístico $\hat{p} = f_r$ es un buen estimador del parámetro **p** (proporción poblacional o probabilidad).

Si simuláramos tomar muchas muestras de igual tamaño y en cada una de ellas calculáramos la proporción de veces que ocurre un suceso A, hallaríamos:

- ⇒ La distribución de la proporción muestral es aproximadamente normal
- ⇒ Su media se encuentra cerca de la proporción poblacional p
- ⇒ Su desviación estándar se hace menor a medida que el tamaño de la muestra se hace mayor.

Distribución muestral de la $\hat{p} = f_r$ (proporción muestral)

Si de una población donde p representa la proporción de elementos que tienen cierta característica A, se toman muestras aleatorias simples de tamaño n, la distribución muestral de la proporción muestral o frecuencia relativa ($\hat{p} = f_r$) de las veces que ocurre A en n, tiene las siguientes propiedades:

⇨ 1) **$E(f_r) = p$**

Es decir, el promedio de todos los posibles valores de f_r es igual al parámetro **p**.

⇨ 2) $$\sigma_{\hat{p}} = \sqrt{Var(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$$

Distribución de la frecuencia relativa o proporción muestral cont.

Cuando el tamaño de la muestra aumenta, la medida de dispersión disminuye. Es decir, a medida que el número de observaciones obtenidas aumenta, el promedio de los valores observados se acerca más y más a p (Ley de los grandes números).

Observe que para un tamaño de muestra fijo, la máxima desviación estándar se encuentra en p = 0,5

⇨ 3) Si n es "suficientemente" grande , la distribución de $\hat{p} = f_r$ se comporta aproximadamente como una distribución normal con media y desviación estándar como las dadas en los puntos 1 y 2.

$$\hat{p} \text{ es aproximadamente } N\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$$

Distribución de la varianza muestral

El estadístico S^2_{n-1} es un buen estimador del parámetro σ^2 (varianza poblacional).

Si simuláramos tomar muchas muestras de igual tamaño y en cada una de ellas calculáramos la varianza muestral, hallaríamos:

- ⇒ La media de la varianza muestral se encuentra cerca de la varianza poblacional σ^2
- ⇒ Su desviación estándar se hace menor a medida que el tamaño de la muestra se hace mayor.

Distribución muestral de la S^2 (varianza muestral)

Si de una población se toman muestras aleatorias simples de tamaño n, la distribución muestral de la varianza muestral S^2_{n-1}, tiene las siguientes propiedades:

⇨ 1) $E(S^2) = \sigma^2$

Es decir, el promedio de todos los posibles valores de S^2_{n-1} es igual al parámetro σ^2

⇨ 2) $V(S^2) = \sigma^2_{S^2} = \dfrac{2\sigma^4}{n-1}$

Cuando el tamaño de la muestra aumenta, la medida de dispersión disminuye. Es decir, a medida que el número de observaciones obtenidas aumenta, el promedio de los valores observados de S^2 se acerca más y más a σ^2 (Ley de los grandes números).

Distribución de la varianza muestral cont.

⇨ 3) Si la población de la cual se extraen las muestras es normal, la variable $\frac{(n-1)\,S^2}{\sigma^2}$ tiene una distribución ji cuadrado (χ^2) con n-1 grados de libertad [(b)]:

$$\frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1}$$

⇨ 4) Si n es "suficientemente" grande, la distribución de la variable χ^2 se ve como una distribución normal con media y desviación estándar como las dadas en los puntos 1 y 2.

(b) Las distribuciones **ji cuadrado** (o chi cuadrado) son una familia de distribuciones que sólo toman valores positivos y que son asimétricas hacia la derecha. Una distribución ji cuadrado viene determinada por un parámetro llamado **grados de libertad**. A medida que aumentan los grados de libertad, las curvas de densidad son menos asimétricas y por lo tanto, los valores mayores son más probables.

Distribución de la varianza muestral cont.

Resumiendo, hemos tratado el comportamiento de las distribuciones muestrales de algunos estimadores cuando se toman muestras aleatorias simples.
Se analizó que si el tamaño de muestra es más grande, la distribución de estos estimadores tiende a centrarse más y más alrededor del valor del parámetro que se quiere estimar.

En la práctica no se conocerá el **verdadero parámetro** poblacional (**por eso la estimación**) y se tomará una sola muestra (no muchas como cuando se simuló la distribución del promedio muestral), pero son las propiedades (**insesgado y convergencia**) las que garantizan que cuando la muestra que se toma sea grande habrá una alta probabilidad de que el valor que toma el estimador (**estimación**) esté cerca del verdadero valor del parámetro que se quiere estimar.

Intervalos de confianza (IC)

Inferir significa sacar conclusiones. La inferencia estadística nos proporciona métodos para sacar conclusiones sobre una población a partir de los datos que surjan de una muestra de dicha población, utilizando la probabilidad para expresar la fuerza de nuestras conclusiones.

Los dos procedimientos más ampliamente utilizados de inferencia estadística son: la construcción de un **intervalo de confianza** cuando el objetivo sea estimar un parámetro poblacional y la **prueba de hipótesis,** cuando el objetivo sea tomar una decisión respecto de una hipótesis que se formula sobre el valor de un parámetro poblacional.

Sólo cuando se utiliza el azar para escoger los elementos que conforman una muestra, podemos describir cómo varía el estadístico. Pudimos contestar preguntas como ¿qué tan cercana queda la media de la muestra **X , de la media de la población μ?.**

En este tema y en el próximo vamos a invertir el argumento. A partir de una muestra conocida que se ha extraído de una población ¿qué se puede concluir acerca de los parámetros desconocidos de la misma? Este proceso involucra una inducción, o inferencia estadística: ir de lo particular (muestra) a lo general (población). Siempre nos basaremos en datos que proceden de una muestra aleatoria simple de una población. Seleccionaremos, para inferir **buenos estimadores:** estimadores insesgados del parámetro poblacional desconocido y convergentes al mismo.

Intervalos de confianza cont.

Si no se conoce el valor de un parámetro poblacional, el mismo se puede estimar a partir de un **intervalo de confianza** para dicho parámetro.
A todo intervalo de confianza, calculado a partir de los datos de una muestra aleatoria, se le fija un nivel de confianza que mide la probabilidad de que el intervalo contenga el verdadero valor del parámetro.

Por ejemplo: un intervalo para un parámetro poblacional, calculado con un 95% de confianza, es un intervalo que tiene una **probabilidad de 95%** de contener el verdadero valor del parámetro.

El objetivo de este tema es describir los razonamientos utilizados en la construcción de un **intervalo de confianza.** Podremos estar interesados en estimar **μ, σ^2** o **p,** obteniendo una medida de la **precisión** de la estimación y otra sobre cuál es nuestra **confianza** de que el resultado sea correcto, como veremos a continuación.

Nos apoyaremos en un ejemplo de estimación del parámetro desconocido **μ,** cuando los datos son una muestra aleatoria simple de tamaño n. El intervalo se basa en el hecho de que la distribución de la media muestral es normal o aproximadamente normal.

IC para la media poblacional con varianza conocida

Anteriormente, suponíamos conocida la media poblacional μ y estudiamos para muestras de distintos tamaños, qué tan cerca o lejos podía esperarse encontrar el valor de la media muestral.

Por ejemplo, si se considera una población normal donde μ = 4.5 y la desviación poblacional σ=1, y se extraen muestras de tamaño 100, la variable promedio muestral se distribuye normalmente con esperanza 4.5 y desviación estándar 1/10. En símbolos:

$$\bar{X} \sim N\,(4.5\ ;\ 1/10)$$

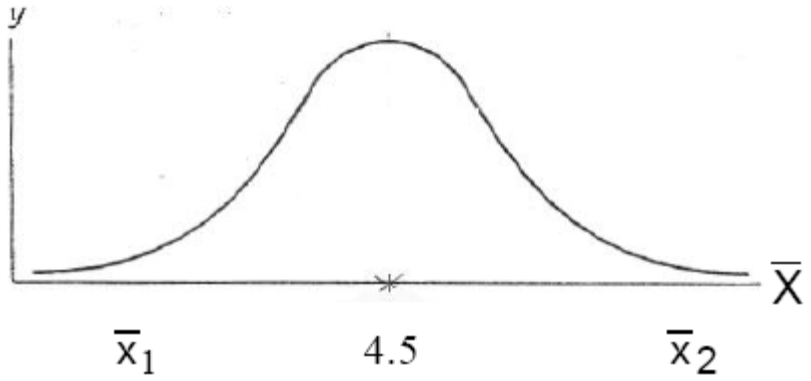
¿Entre qué valores se encuentran el 95 % de los promedios muestrales centrales? Es decir buscamos los valores $\bar{x}_1$ y $\bar{x}_2$ que satisfagan:

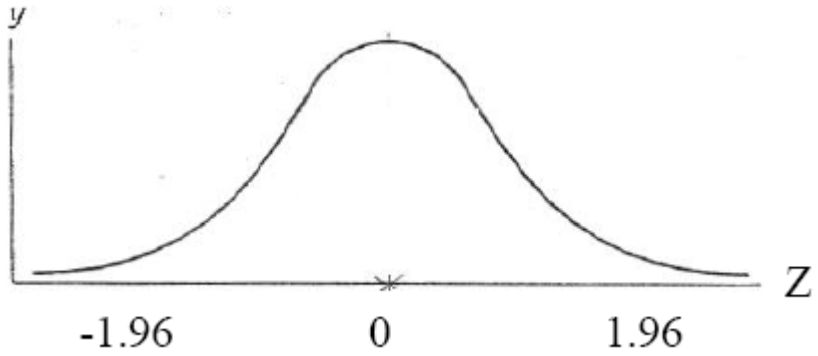
$$P\,(\bar{x}_1 \leq \bar{X} \leq \bar{x}_2) = 0.95$$

$$P\left(\frac{\bar{x}_1 - \mu}{\sigma/\sqrt{n}} \leq Z \leq \frac{\bar{x}_2 - \mu}{\sigma/\sqrt{n}}\right) = P\left(\frac{\bar{x}_1 - 4.5}{1/\sqrt{100}} \leq Z \leq \frac{\bar{x}_2 - 4.5}{1/\sqrt{100}}\right) =$$

$$P\,(-1.96 \leq Z \leq 1.96) = 0.95$$

IC para la media poblacional con varianza conocida cont.





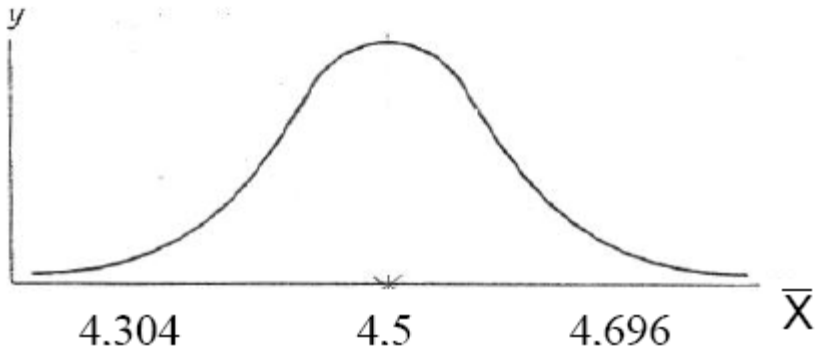
$$\frac{\bar{x}_1 - 4.5}{1/10} = -1.96 \Rightarrow \bar{x}_1 = 4.304$$

$$\frac{\bar{x}_2 - 4.5}{1/10} = 1.96 \Rightarrow \bar{x}_2 = 4.696$$

Es decir $P(4.304 \leq \bar{X} \leq 4.696) = 0.95$

IC para la media poblacional con varianza conocida cont.

Por lo tanto, el 95% de los promedios muestrales estarán entre 4.304 y 4.696:



En general : $$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Supongamos ahora que la media poblacional μ es desconocida y se conoce la desviación estándar poblacional σ. Entonces, invirtiendo μ por la media muestral $\overline{X}$ resulta:

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \qquad ①$$

IC para la media poblacional con varianza conocida cont.

Esta expresión es el **intervalo aleatorio**, a partir del cual se estima la μ desconocida.

Se debe tener cuidado en la interpretación de ①. El promedio poblacional no se ha convertido en una variable, sigue siendo una constante de la población. Los *límites del intervalo sí son aleatorios* ya que dependen de la variable aleatoria promedio muestral.

Cuando se selecciona una muestra, los límites dejan de ser aleatorios dado que obtenemos un valor del promedio de la muestra seleccionada y en consecuencia hablamos de un **intervalo de confianza de 95% para el promedio poblacional.**

$$IC_{\mu} = \left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

Aunque en la práctica sólo se selecciona una **sola muestra** de tamaño n y se calcula el promedio muestral $\bar{X}$ para dicha muestra, es necesario pensar en el conjunto hipotético de todas las muestras posibles, cada una del mismo tamaño n, a fin de entender el significado del intervalo de confianza.

Supongamos las siguientes situaciones para una muestra extraída de tamaño 100:

IC para la media poblacional con varianza conocida cont.

❖ La media muestral resulta igual a 4.35. Luego el intervalo de confianza para μ es:

$$\bar{X} \pm 1.96 \, \sigma/\sqrt{n} = 4.35 \pm 1.96 \, 1/\sqrt{100} = 4.35 \pm 0.196$$

Este intervalo (4.154 ; 4.546) contiene a la media poblacional μ = 4.5. Esta muestra nos llevaría a decir que 4.5 es un valor posible de μ.

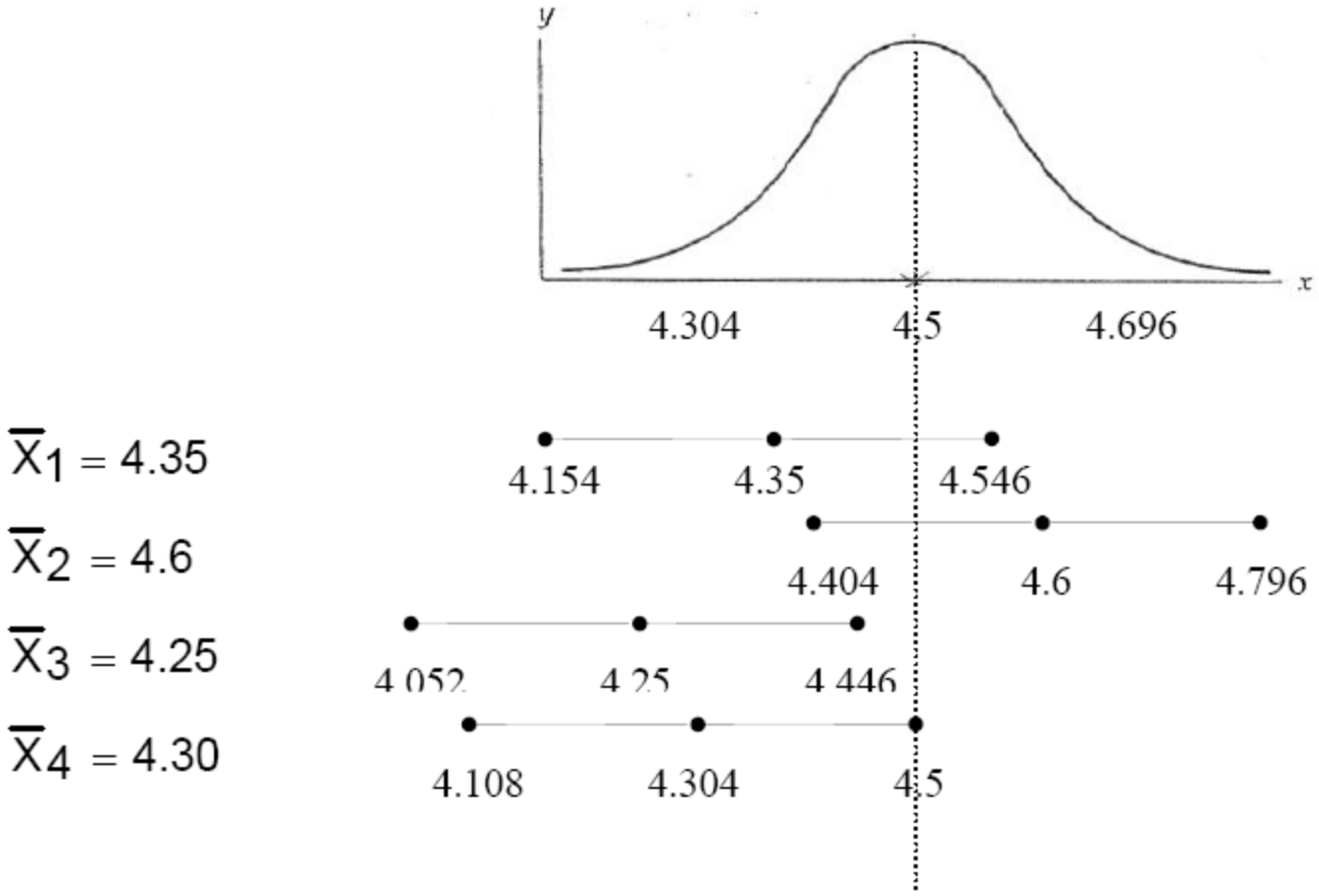
❖ La media muestral resulta igual a 4.6. El intervalo de confianza obtenido a partir de este valor :(4,404 ; 4.796), también nos llevaría a decir que 4.5 es un valor posible de μ.

❖ La media muestral resulta igual a 4.25. El intervalo de confianza obtenido a partir de este valor: (4.052 ; 4.446) no contiene al parámetro; nos llevaría a decir que 4.5 no es un valor posible de μ.

❖ La media muestral resulta igual a 4.304. El intervalo de confianza obtenido a partir de este valor: (4.108 ; 4.5) nos llevaría a decir que 4.5 es un valor posible de μ.

❖ Todas estas situaciones se pueden visualizar en la figura siguiente:

IC para la media poblacional con varianza conocida cont.



O sea que para algunas muestras, el intervalo de confianza contiene al verdadero valor de μ, mientras que para otras no.

IC para la media poblacional con varianza conocida cont.

En este ejemplo, siempre que la media (o promedio) esté situada a una distancia de a lo sumo 0.196 de μ, el intervalo cubrirá al verdadero valor del promedio poblacional y esto sucederá en un 95 % de todas las muestras posibles.

La semi-amplitud del intervalo de confianza se conoce como **error de estimación** y es una medida de la precisión de la estimación. En el ejemplo se trabaja con un error de estimación para un intervalo de 95 % de confianza igual a 0.196.

$$\varepsilon = 1.96 \, \sigma/\sqrt{n} = 1.96 \, 1/\sqrt{100} = 0.196$$

En la práctica sólo se selecciona una muestra y se desconoce μ. Nunca se sabe con seguridad si el intervalo obtenido incluye la media poblacional. Por ejemplo, si se extrae una muestra y su media resulta igual a 4.6 decimos que tenemos una confianza de 95 % de que la media poblacional desconocida se encuentre en el intervalo (4.404 ; 4.796). Este intervalo es el que varía en función de la muestra que sale seleccionada. El valor del parámetro es único.

IC para la media poblacional con varianza conocida cont.

Si a partir del mismo ejemplo se hubiera trabajado con una confianza de 99%, el error de estimación resultaría:

$$\varepsilon = 2.58 \; \sigma/\sqrt{n} = 2{,}58 \; 1/\sqrt{100} = 0.258$$

En general, para una confianza de 100 (1-α) % el error de estimación resulta:

$$\varepsilon = z_\alpha \; \sigma/\sqrt{n}$$

Y el intervalo de **confianza para la media poblacional con varianza conocida** es:

$$IC_\mu = \left(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}} \right)$$

Si la población es Normal, no interesa el tamaño de la muestra aleatoria que se selecciona para estimar μ. Si la población no es Normal, se necesita un tamaño de muestra de por lo menos 30 observaciones (Teorema Central del Límite) para usar la expresión anterior del intervalo de confianza.

Tamaño de la muestra para estimar μ

Siempre es necesario planificar la inferencia conjuntamente con la obtención de los datos.

Si el error de estimación para la media es: $\varepsilon = z_\alpha \frac{\sigma}{\sqrt{n}}$

El tamaño de muestra para un error de estimación ε y un nivel de confianza determinado se deduce de la ecuación anterior, resultando:

$$n = \left(z_\alpha \frac{\sigma}{\varepsilon} \right)^2$$

Es importante tener claro que lo que determina el tamaño de la muestra es el error de estimación y la confianza que se pretende para realizar la estimación y no el tamaño de la población, ya que éste no influye sobre el tamaño de muestra que se necesita para la inferencia.

Esta fórmula (redondear n siempre hacia arriba) no se puede utilizar ligeramente. En la práctica la obtención de observaciones cuesta tiempo y dinero. Puede ocurrir que el tamaño de la muestra ideal sea inviable por razones económicas y/o de otro tipo.

IC para la media con varianza desconocida

En el punto anterior estimamos el promedio poblacional suponiendo la desviación estándar poblacional conocida. En la práctica, es poco probable conocer el valor de σ.

Antes vimos que si la población de la cual se extraen las muestras es normal, con media poblacional **μ** y desviación estándar poblacional **σ** desconocido, se reemplaza σ por S (desvío muestral) y la estadística $\frac{(\overline{X} - \mu)}{S/\sqrt{n}}$

deja de tener distribución normal estandarizada y tiene una distribución **t Student** con n-1 grados de libertad, es decir:

$$\frac{(\overline{X} - \mu)}{S/\sqrt{n}} \sim t_{n-1;\alpha}$$

Para obtener un intervalo de confianza para el promedio poblacional cuando σ era conocida trabajamos con la variable **normal estandarizada** (z_α). Ahora trabajaremos con la variable e **t de Student** ($t_{n-1;\alpha}$).

En consecuencia, para una confianza de 100 (1-α) % el error de estimación resulta:

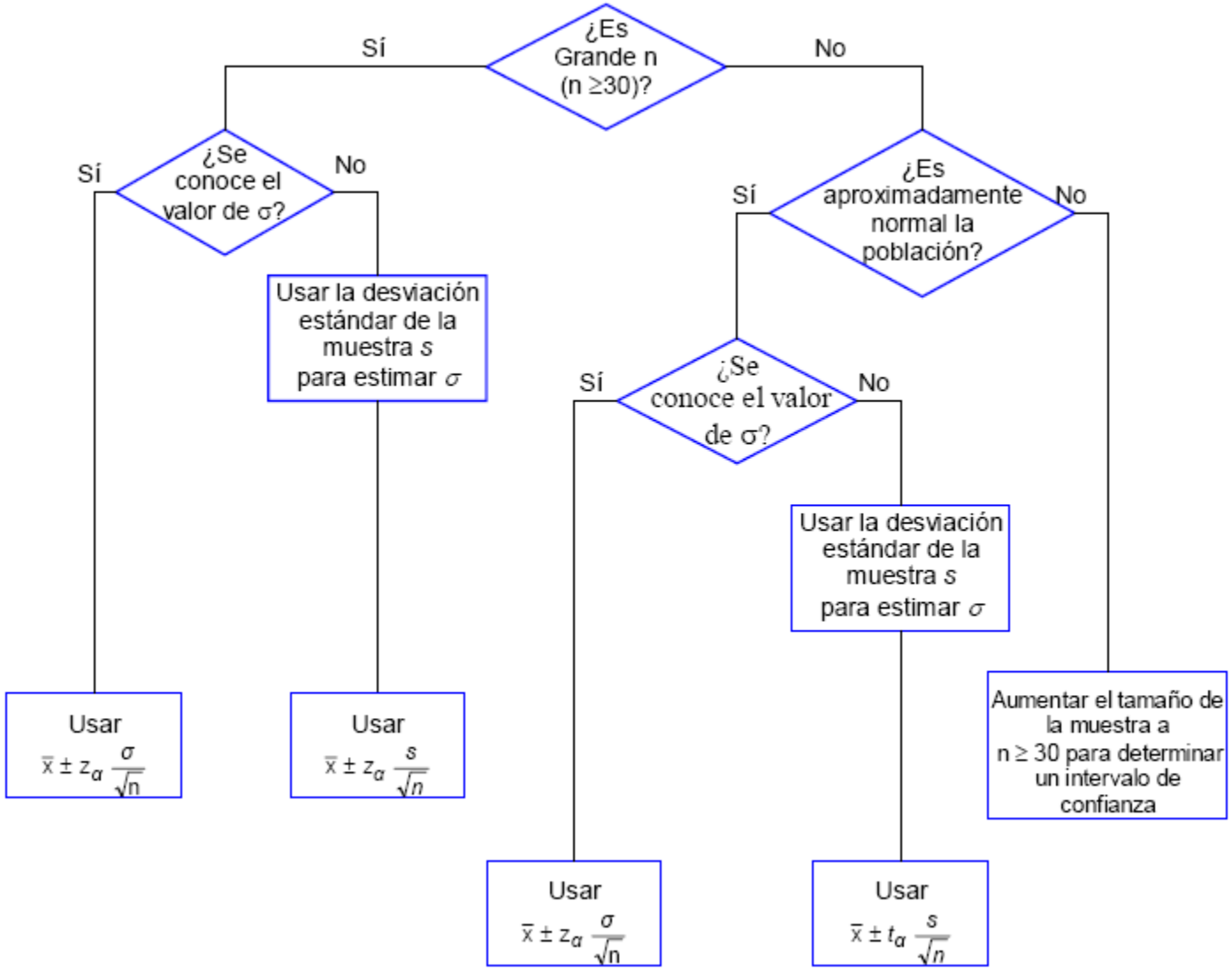
$$\varepsilon = t_{n-1;\alpha} \; S/\sqrt{n}$$

IC para la media con varianza desconocida

Y el intervalo de confianza para la media poblacional con varianza desconocida es:

$$IC_{\mu} = \left(\bar{x} - t_{n-1;\alpha} \frac{s}{\sqrt{n}} \; ; \; \bar{x} + t_{n-1;\alpha} \frac{s}{\sqrt{n}} \right)$$

Resumen IC para la media poblacional



IC para la proporción poblacional

Se sabe que la distribución del estadístico frecuencia relativa o proporción muestral. Si n es *suficientemente* grande, la distribución de **fr** se comporta aproximadamente como una distribución normal con media p y desviación estándar $\sqrt{\frac{p(1-p)}{n}}$

Es decir:

$$\hat{p} \text{ es aproximadamente } N\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$$

Con el mismo razonamiento que empleamos en la estimación de la media poblacional, el planteo inicial para estimar la proporción poblacional o probabilidad de un suceso cualquiera A, es encontrar dos valores $\hat{p}_1$ y $\hat{p}_2$ que verifiquen:

$$P(\hat{p}_1 \leq \hat{p} \leq \hat{p}_2) = 0.95$$

A esta expresión la podemos escribir de la siguiente forma:

$$P(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}) = 0.95$$

IC para la proporción poblacional cont.

Y en consecuencia el Intervalo de confianza para la proporción poblacional para un nivel de confianza de 100 (1- α) % es:

$$IC_{p,(1-p)} = \left(\hat{p} - z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

A partir de esta expresión el error de estimación resulta:

$$\varepsilon = z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Tamaño de la muestra para estimar p

El tamaño de muestra **n** para un error de estimación **ε** y un nivel de confianza de **100 (1-α)%**, se deduce de la ecuación anterior, resultando:

$$n = \left(\frac{z_{\alpha}}{\varepsilon}\right)^2 p(1-p)$$

IC para la proporción poblacional cont.

Para utilizar la fórmula anterior se necesita reemplazar a p por una estimación de la misma.

Esta se puede obtener:

- de la estimación de la proporción muestral en una muestra anterior
- calculando la **estimación de p** en una muestra preliminar (o piloto)

Si estas alternativas no son posibles, otra forma para calcular el tamaño de la muestra requerida, es considerar que siempre p(1 – p) es máximo para p = 0.5. Es decir, que una cota superior para n (para una confianza de 100 (1- α)% y un error ε está dada por:

$$n = \left(\frac{z_\alpha}{\varepsilon}\right)^2 (0.25)$$

Ejemplo :

Una empresa de cable desea conocer qué proporción de sus clientes se informan de las noticias a través de los noticiarios que difunden. Para ello seleccionó una muestra aleatoria de 200 clientes. De las 200 personas, 110 respondieron que se informan a través de los noticieros televisivos. El intervalo obtenido para una confianza de 95% resultó:

$$IC_{p,95\%} = \left(0.55 - 1.96\sqrt{\frac{0.55 * 0.45}{200}} \le p \le 0.55 + 1.96\sqrt{\frac{0.55 * 0.45}{200}}\right) =$$

$$= 0.55 - 0.07 \le p \le 0.55 + 0.07 = (0.48\,;\,0.62)$$

IC para la proporción poblacional cont.

Es decir que con una confianza de 95 % se puede inferir que la proporción de clientes que se informan a través de los noticieros se encuentra entre el 48% y el 62%.
La empresa considera que el error de estimación es alto y por lo tanto, este intervalo no le brinda demasiada información.
A tal fin decide consultar a más clientes. El tamaño de muestra que lo llevaría a cometer un error de 4%, con la misma confianza, y utilizando la proporción muestral ya obtenida, resulta:

$$n = \left(\frac{z_\alpha}{\varepsilon}\right)^2 \hat{p}(1-\hat{p}) = \left(\frac{1.96}{0.04}\right)^2 0.55 * 0.45 = 594.25$$

Es decir que se necesita un tamaño de muestra mayor o igual a 595 clientes.

IC para la varianza poblacional de una distribución normal

Se sabe que la distribución de la varianza muestral S^2. Si la población de la cual se extraen las muestras es normal, la variable $\frac{(n-1)\,S^2}{\sigma^2}$ tiene una distribución ji cuadrado ($\chi 2$) con (n -1) grados de libertad

$$\frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Nuestro planteo inicial es ahora encontrar dos valores de la distribución χ^2 $(\chi^2_a \; y \; \chi^2_b)$ que verifiquen:

$$P\left(\chi^2_a < \frac{(n-1)\,S^2}{\sigma^2} < \chi^2_b\right) = 1 - \alpha$$

Trabajando algebraicamente la expresión anterior obtenemos:

$$P\left(\frac{(n-1)\,S^2}{\chi^2_b} \leq \sigma^2 \leq \frac{(n-1)\,S^2}{\chi^2_a}\right) = 1 - \alpha$$

Es decir que el intervalo de confianza para la varianza poblacional de una población normal (σ^2) con una confianza de 100 (1 - α)% resulta :

$$I.C._{\sigma^2,(1-\alpha)} = \left(\frac{(n-1)\,S^2}{\chi^2_b} \; ; \; \frac{(n-1)\,S^2}{\chi^2_a}\right)$$

IC para la varianza poblacional de una distribución normal cont.

Ejemplo:

En un criadero de peces se crían truchas para aprovisionar ríos y lagos. El peso del pez en el momento que es liberado se puede controlar variando la alimentación. El criadero espera una desviación estándar de 21.5 gramos en el peso de los peces. Para evaluar si el plan de alimentación que se aplica cumple lo deseado, se toma una muestra de 25 peces obteniéndose una desviación para el peso de 28.9 gramos. El intervalo de 95% de confianza para la varianza poblacional resulta:

$$I.C._{\sigma^2,\,0.95} = \left(\frac{(25-1)\,28.9}{39.36} \;;\; \frac{(25-1)\,28.9}{12.4} \right) = (509.27\,;1616.54)$$

Es decir, que con un 95% de confianza, el desvío estándar poblacional se encuentra en el intervalo (22.57 ; 40.21). Por lo tanto se concluye, con un 95 % de confianza, que el desvío estándar del peso de los peces es superior al deseado por el criadero.