Conceptos generales de inferencia estadística.
Estimación de parámetros. Intervalos de confianza.

Inferencia estadística: Dijimos en la primera clase que inferencia estadística es el estudio de técnicas que permiten, a partir de los datos de una muestra, obtener alguna información sobre la población (de la que se extrajo la muestra).

Veamos algunos ejemplos de problemas en los que se pueden aplicar técnicas de inferencia estadística:

Ejemplo 1: Quiero conocer la media de las estaturas de los niños de 8 años varones que viven en la ciudad de Buenos Aires. Para ello elijo 200 niños al azar y les mido su estatura.

Ejemplo 2: Quiero conocer el contenido de hierro de una muestra de cierto mineral. Hago 10 determinaciones.

Ejemplo 3: Quiero conocer la proporción de personas que están de acuerdo con cierta medida del gobierno entre las personas mayores de 18 años que viven en la ciudad de Buenos Aires. Para ello elijo 400 personas al azar y le pregunto a cada una si está o no de acuerdo.

Modelos probabilísticos que puedo usar para estos problemas:

Para el ejemplo 1: Llamo X_i a la estatura del i-ésimo niño seleccionado al azar, entonces supongo:

$$X_1, X_2, ..., X_{200} \text{ vs. as. i.i.d } N(\mu,\sigma^2)$$

Donde no conozco ni μ ni σ^2
μ representa la media poblacional de las estaturas y σ^2 la varianza poblacional.

Para el ejemplo 2: se suele usar un modelo similar: $X_1, X_2, ..., X_{10}$ vs. as. i.i.d $N(\mu,\sigma^2)$
¿Que representan X_i , μ y σ^2 en este ejemplo?

Para el ejemplo 3:

1er modelo: Si llamo X=nro de personas que responden que están a favor, entonces

$$X \sim Bin(n=400, p)$$ donde p es desconocido

¿Que representan p en este ejemplo?

Otro modelo: Si llamo X_i a la variable que vale 1 si la i-ésima persona encuestada responde que está a favor y 0 en caso contrario, entonces

$$X_1, X_2, ..., X_{400} \text{ vs. as. i.i.d } Bin(1,p)$$ (o Bernoulli(p) que es lo mismo).

Como puede apreciarse en estos ejemplos en inferencia estadística planteamos primero modelos probabilísticos que consideramos adecuados para el problema que queremos resolver. En estos tres ejemplos y en muchos otros los modelos son de este tipo:

$X_1, X_2, ..., X_n$ vs. as. i.i.d., cada una con una distribución de la forma $F(x,\theta,\lambda,...)$ donde la función F es conocida, pero los parámetros θ, λ, etc. son desconocidos.

Estimación de parámetros

En el ejemplo 1 queremos estimar (conocer aproximadamente) la media μ.
Un estimador intuitivamente razonable es la media muestral

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

En el ejemplo 3, un estimador intuitivamente razonable de la proporción poblacional p de personas que están de acuerdo con el gobierno es la proporción muestral $\hat{p} = \frac{X}{n}$ si planteamos el primer modelo o la proporción muestra $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$ si planteamos el segundo.

En general si $X_1, X_2, ..., X_n$ son vs. as. cuya distribución depende de un parámetro desconocido θ un estimador de θ es una función de $X_1, X_2, ..., X_n$ "convenientemente" elegida, de modo de que el error de estimación sea "pequeño".

Un criterio para medir la bondad de un estimador: Error cuadrático Medio.
Sean $X_1, X_2, ..., X_n$ vs. as. cuya distribución depende de un parámetro desconocido θ y sea $\hat{\theta} = g(X_1, X_2, ..., X_n)$ un estimador θ. Se define error cuadrático medio del estimador $\hat{\theta}$

$$\text{ECM de } \hat{\theta} = E(\hat{\theta} - \theta)^2$$

Buscaremos estimadores que tengan ECM "pequeño".

Ejemplo: Sean $X_1, X_2, ..., X_n$ vs.as. i.i.d. con cualquier distribución con $E(X_i)=\mu$ y $Var(X_i)= \sigma^2$. Hallar el ECM de $\overline{X}$ como estimador de μ .

Una propiedad de un estimador: estimadores insesgados.
Definición: Sean $X_1, X_2, ..., X_n$ vs. as. cuya distribución depende de un parámetro desconocido θ y sea $\hat{\theta} = g(X_1, X_2, ..., X_n)$ un estimador θ. Se dice que $\hat{\theta}$ es un estimador insesgado de θ si

$$E(\hat{\theta}) = \theta$$

Ejemplo:Sean $X_1, X_2, ..., X_n$ vs.as. i.i.d. con cualquier distribución con $E(Xi)=\mu$ y $Var(Xi) = \sigma^2$.
a) Es $\overline{X}$ un estimador insesgado de μ ?
b) Encontrar un estimador insesgado de σ^2

Sesgo de un estimador
Sean $\hat{\theta}$ un estimador θ. Se define sesgo de $\hat{\theta}$:

$$\text{sesgo de } \hat{\theta} = E(\hat{\theta}) - \theta$$

Relación entre ECM, varianza y sesgo:
Se puede demostrar facilmente que

$$\text{ECM de } \hat{\theta} = \text{Var}(\hat{\theta}) + (\text{sesgo de } \hat{\theta})^2$$

por lo que buscamos estimadores que tengan "poca" varianza y "poco" sesgo. Muchas veces se buscan estimadores insesgados (o sea con sesgo cero) pero esta propiedad no es imprescindible.

Una propiedad de un estimador para muestras grandes: estimadores consistentes.
Definición: Sean $X_1, X_2, ..., X_n$ vs. as. cuya distribución depende de un parámetro desconocido θ y sea $\hat{\theta}_n = g_n(X_1, X_2, ..., X_n)$ una sucesión de estimadores de θ. Se dice que $\hat{\theta}_n$ es consistente si

$$\lim_{n\to\infty} \hat{\theta}_n = \theta$$

Intervalo de confianza para un parámetro

Antes de dar la definición de intervalo de confianza, vamos a encontrarlo para un ejemplo.
Supongamos que se conoce por la experiencia previa que un método de medición del porcentaje de hierro en un mineral tiene una desviación standard de 0.12. También se sabe por experiencia que si hacemos muchas determinaciones de la misma muestra la distribución de las mediciones es aproximadamente normal (el histograma de las mediciones hechas en una misma muestra se parece a la curva de Gauss con σ=0.12). Hacemos 4 determinaciones del contenido de hierro en un trozo de mineral y obtenemos:

15.18% 15.32% 15.46% 15.25%

Teniendo en cuenta la experiencia previa, podemos considerar el siguiente modelo:

Modelo: $X_1, X_2, ..., X_n$ vs. as. i.i.d $N(\mu,\sigma^2)$ con n=4 y σ=0.12

El unico parametro desconocido de este modelo es la media μ. El estimador de μ es la media muestral $\overline{X}$ que en este ejemplo da 15.3025 o, redondeando, 15.30.
Pero cuanto vale μ? Obviamente μ no tiene por que valer 15.30. Pero vamos a encontrar un intervalo que contiene al verdadero valor de μ con probabilidad alta.

Intervalo de confianza para la media de un población, basado en una muestra normal con varianza conocida.
Consideramos el siguiente modelo:

$X_1, X_2, ..., X_n$ vs. as. i.i.d $N(\mu,\sigma^2)$ con σ conocido. (1)

El estimador de μ

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

Sabemos que, bajo el modelo (1)

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Estandarizando:

$$\frac{\overline{X} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0,1) \qquad (2)$$

Como en la curva de Gauss el area entre –1.96 y 1.96 es 0.95 resulta que:

$$P\left(-1.96 \leq \frac{\overline{X} - \mu}{\sqrt{\sigma^2 / n}} \leq 1.96\right) = 0.95$$

Pasando términos de miembro con el objetivo de dejar μ en el centro obtenemos:

$$P\left(\overline{X} - 1.96 * \sigma / \sqrt{n} \leq \mu \leq \overline{X} + 1.96 * \sigma / \sqrt{n}\right) = 0.95 \qquad (3)$$

El intervalo

$$\left[\overline{X} - 1.96 * \sigma / \sqrt{n} \;,\; \overline{X} + 1.96 * \sigma / \sqrt{n}\right] \qquad (4)$$

se llama **intervalo de confianza para μ con nivel de confianza del 0.95 (o 95%).**

Ejemplo: en el ejemplo de las 4 determinaciones de hierro obtenemos:

$$[\,15.3025 - 0.1176 \;,\; 15.3025 + 0.1176\,]$$
$$[\,15.1849 \;,\; 15.4201\,]$$

o rendondeando, el intervalo de confianza al 95% para la media μ es [15.18, 15.42].

Que significa que el intervalo de confianza tenga nivel 95%?

Significa que si extrajésemos muchas (muchísimas!) muestras y para cada muestra calculásemos el intervalo de confianza, entonces el 95% de los intervalos contendrían al verdadero valor del parámetro de la población (en nuestro ejemplo el parámetro es la media μ).

Comentarios:

1) Volvamos al ejemplo de las determinaciones de hierro: Habíamos obtenido que el intervalo de confianza para μ al 95% es [15.11 , 15.49]
Aunque sea intuitivamente razonable no es correcto escribir

$$P(15.11 \leq \mu \leq 15.49) = 0.95$$

por que?

2) Qué ventajas y que desventajas tiene calcular un IC al 99% en vez del 95%.

3) Qué es lo que hay que cambiar en (4) para obtener un IC al 99%?

Continuando con el modelo (1), o sea con el modelo de una muestra normal con varianza conocida, la expresión para el IC para con cualquier nivel es similar a (4) pero cambiando el valor 1.96. Llamemos $z_{\alpha/2}$ al valor que deja un area de $\alpha/2$ en la cola de la curva normal estándar (y por lo tanto un area 1-α entre -$z_{\alpha/2}$ y $z_{\alpha/2}$). El intervalo

$$[\ \overline{X} - z_{\alpha/2} * \sigma\ /\ \sqrt{n}\ \ ,\ \ \ \overline{X} + z_{\alpha/2} * \sigma\ /\ \sqrt{n}\ ] \qquad (5)$$

es un **intervalo de confianza para μ con nivel de confianza 1-α.**

Definición general de intervalo de confianza:

Sean $X_1, X_2, ..., X_n$ vs. as. cuya distribución depende de un parámetro desconocido θ.
Un **intervalo de confianza para θ con nivel de confianza 1-α** es un intervalo de extremos aleatorios (que dependen de la muestra $X_1, X_2, ..., X_n$)

$$[\ a(X_1, X_2, ..., X_n)\ \ ,\ \ b(X_1, X_2, ..., X_n)\ ]$$

tal que

$$P\ (\ a(X_1, X_2, ..., X_n) \leq\ \theta\ \leq\ b(X_1, X_2, ..., X_n)\ )\ = 1 - \alpha$$

Comentario: La expresión (5) nos dá un intervalo de confianza para μ para variables con distribución normal y varianza conocida. También vamos a poder calcular IC para μ sin conocer la varianza, o sin conocer la distribución de la variable o intervalos de confianza para otros parámetros. Las expresiones para calcularlos van a ser diferentes en cada caso, pero el objetivo y las ideas generales son siempre los mismos.

Cálculo del tamaño de la muestra.

Con la muestra de las cuatro determinaciones de hierro obtuvimos el siguiente IC al 95% para la media:

$$[\ 15.3025 - 0.1176\ \ ,\ 15.3025\ + 0.1176\ ]$$
$$[\ 15.1849\ \ ,\ 15.4201\ ]$$

El intervalo obtenido tiene longitud = 2*0.1176 = 0.2352

Que se puede hacer si se desea un IC al 95% con menor longitud?

Llamando L a la longitud del IC, se puede obtener de (4) que

$$L\ = 2 * 1.96 * \sigma\ /\ \sqrt{n}$$

y entonces, puede observarse que aumentando n disminuye la longitud del IC y que esta longitud puede hacerse tan pequeña como se desee tomando n suficientemente grande.

Si por ejemplo deseamos calcular un IC de longitud $L \leq 0.10$ deberá ser

$$L = 2 * 1.96 * 0.12 / \sqrt{n} \leq 0.10$$

De esta inecuación se despeja

$$(2 * 1.96 * 0.12) / 0.10 \leq \sqrt{n}$$

$$[(2 * 1.96 * 0.12) / 0.10]^2 \leq n$$

$$22.13 \leq n$$

o sea que haciendo 23 (o más) determinaciones se obtiene un IC de longitud ≤ que 0.10.

Comentarios:
1) Cuando se duplica el tamaño de la muestra (n), que ocurre con la longitud del IC?
2) Cuando se multiplica por 4 el tamaño de la muestra?

Intervalo de confianza para la media de un población, basado en una muestra normal con varianza desconocida.

Consideramos ahora el siguiente modelo:

$$X_1, X_2, ..., X_n \text{ vs. as. i.i.d } N(\mu,\sigma^2) \text{ con } \sigma \text{ conocido.} \quad (6)$$

El estimador de μ sigue siendo $\overline{X}$ Sigue valiendo que, bajo el modelo (6)

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Estandarizando:

$$\frac{\overline{X} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0,1) \quad (7)$$

Si seguimos los mismos pasos que para el caso σ conocido, volveríamos a obtener las expresion (4) o la expresión más general (5). Pero (4) o (5) no son intervalos de confianza para μ, porque los extremos del intervalo dependen de σ y por lo tanto no cumplen la definición de IC.
Además es intuitivo que (4) o (5) no son de utilidad ya que no pueden calcularse, verdad?
Para solucionar este problema reemplacemos en (7) σ^2 por la varianza muestral s^2 y obtenemos

$$\frac{\overline{X} - \mu}{\sqrt{s^2 / n}} \quad (8)$$

La distribución de (7) es N(0,1) pero la de (8) no. W. S. Gosset, que escribía bajo el seudónimo de Student, a principios del siglo XX encontró la distribución de (8) (su función de densidad ya que es una variable aleatoria continua). La función de densidad de (8) depende del tamaño de la muestra (n), La función de densidad de (8) tiene una forma de campana similar a la curva normal, pero tiene mayor varianza que la distribución N(0,1) (esto último es intuitivamente razonable, por qué). Cuando n es grande se parece a la curva N(0,1) (esto también es intuitivamente esperable). En honor a Student, la distribución de (8) se llama distribución "t de Student con n-1 grados de libertad", lo que notaremos:

$$\frac{\overline{X} - \mu}{\sqrt{s^2 / n}} \sim t_{n-1} \qquad (9)$$

Si llamamos $t_{n-1;\, \alpha/2}$ al valor que deja un área de $\alpha/2$ en la cola de la función de densidad t_{n-1}, siguiendo una deducción análoga al caso σ conocido obtenemos:

$$P\left(\overline{X} - t_{n-1;\, \alpha/2} * s / \sqrt{n} \leq \mu \leq \overline{X} + t_{n-1;\, \alpha/2} * s / \sqrt{n} \right) = 1-\alpha$$

y por lo tanto el intervalo

$$\left[\overline{X} - t_{n-1;\, \alpha/2} * s / \sqrt{n} \leq \mu \leq \overline{X} + t_{n-1;\, \alpha/2} * s / \sqrt{n} \right] \qquad (10)$$

es un **intervalo de confianza para μ con nivel de confianza $1-\alpha$.**

Ejemplo: en el ejemplo de las 4 determinaciones de hierro calculamos $\overline{X}$=15.3025, s=0.1195. Buscamos en un tabla o en Statistix el valor de la curva de t con n-1=3 grados de libertad que deje un area de 0.025 en la cola que es 3.18. Reemplazamos en (10)

$$[\ 15.3025 - 3.18*0.1195 / \sqrt{4}\ ,\ 15.3025 + 3.18*0.1195 / \sqrt{4}\ ]$$

$$[\ 15.3025 - 0.1900\ ,\ 15.3025 + 0.1900\ ]$$

$$[\ 15.11125\ ,\ 15.4925\ ]$$

o rendondeando, el intervalo de confianza al 95% para la media μ es [15.11, 15.49].

Este intervalo puede calcularse sin hacer ninguna cuenta con Statistix.
Para ello hay que ingresar las cuatro determinaciones hechas y luego marcar "Summary Statistics", "Descriptive Statistics" y poner un tilde en el casillero Conf Int. Se obtiene:

```
DESCRIPTIVE STATISTICS

VARIABLE          N   LO 95% CI        MEAN   UP 95% CI          SD
HIERRO            4      15.112      15.303      15.493      0.1195
```

Comentario:
Los valores de la tabla de t que dejan un area $1-\alpha$ en el centro y $\alpha/2$ en cada cola disminuyen **cuando n aumenta y tienden al valor $z_{\alpha/2}$ cuando $n \to \infty$.**
Por ejemplo para $1-\alpha = 0.95$, el valor $t_{n-1;\alpha/2}$ es 3.18 para n=4, 2.26 para n=10, 1.98 para n=100 y su limite para $n \to \infty$ es 1.96.

Intervalo de confianza con nivel asintótico para la media de una población para cualquier distribución (válidos para muestras "grandes").

La limitación que tienen los modelos anteriores es que suponen distribución normal. Gracias al Teorema Central del Límite vamos a poder calcular intervalos de confianza para μ aunque sepamos que la variable no tiene distribución normal o no tengamos ninguna información previa sobre la distribución de la variable. Pero estos intervalos sólo van a valer para n "grande".

Consideramos el siguiente modelo:

$$X_1, X_2, ..., X_n \text{ vs. as. i.i.d con cualquier distribución} \qquad (11)$$

Llamemos $\mu=E(X_i)$ y $\sigma^2= Var(X_i)$. Vamos a suponer (es el caso más usado) que no conocemos σ.

El estimador de μ sigue siendo $\overline{X}$ Sigue valiendo que, bajo el modelo (11), gracias al TCL cumple

$$\overline{X} \quad \tilde{a} \quad N(\mu, \frac{\sigma^2}{n}) \qquad \text{para n grande}$$

donde hemos usado la notación $\tilde{a}$ para indicar "tiene distribución aproximadamente"

Estandarizando:

$$\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}} \quad \tilde{a} \quad N(0,1) \quad \text{para n grande}$$

Si n es grande, la DS muestral es un estimador consistente de la DS poblacional, así que ambas van a ser "parecidas"

$$s \cong \sigma ,$$

por lo tanto

$$\frac{\overline{X}-\mu}{\sqrt{s^2/n}} \quad \tilde{a} \quad N(0,1) \quad \text{para n grande}$$

Luego:

$$P(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sqrt{s^2/n}} \le z_{\alpha/2}) \cong 1-\alpha \text{ para n grande}$$

o, más claramente escrito:

$$\lim_{n\to\infty} P(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sqrt{s^2/n}} \le z_{\alpha/2}) = 1-\alpha$$

Pasando de miembro se llega a que

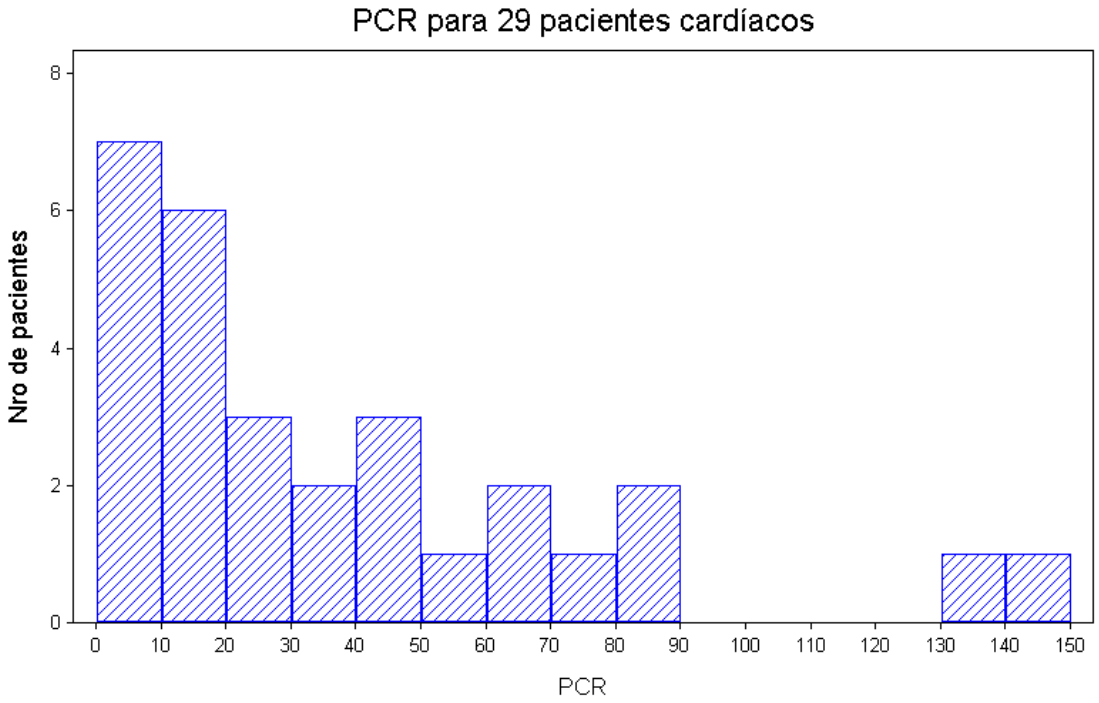
$$\lim_{n\to\infty} P(\overline{X} - z_{\alpha/2} * s/\sqrt{n} \le \mu \le \overline{X} + z_{\alpha/2} * s/\sqrt{n}) = 1-\alpha$$

y por esto el intervalo

$$[\overline{X} - z_{\alpha/2} * s/\sqrt{n} \;,\; \overline{X} + z_{\alpha/2} * s/\sqrt{n}] \qquad (12)$$

se llama **intervalo de confianza para μ con nivel de confianza asintótico 1-α.**

Ejemplo: En la clase de estadística descriptiva mostramos el siguiente histograma de PCR (proteína C reactica) en pacientes cardíacos.



Se puede ver que la distribución no es gaussiana. Estamos en el límite de poder aplicar el TCL porque habíamos dicho como "receta" que en general daba una aproximación aceptable para n≥30. Acá n=29. Apliquémoslo igual, sabiendo que el nivel de confianza no es exacto sino una aproximación.
Calculamos primero la media y la DS con el Statistix:

```
DESCRIPTIVE STATISTICS

                     PCR
N                     29
MEAN              38.665
SD                38.537
1ST QUARTI        9.5150
MEDIAN            25.440
3RD QUARTI        57.970
```

Si queremos un IC al 95%, entonces $z_{\alpha/2} = 1.96$. Reemplazamos en (12):

$$[38.665 - 1.96 * 38.537 / \sqrt{29} \ , \ 38.665 - 1.96 * 38.537 / \sqrt{29}]$$
$$[38.665 - 14.026 \ , \ 38.665 - 14.026]$$

$$[24.639 \ , \ 52.691]$$

o rendondeando "el intervalo de confianza al 95% para la media de PCR en la población de enfermos cardíacos con tales características.... es [24.6, 52.7]" o, como se suele escribir resumiendo: "La media de PCR los 29 pacientes cardíacos estudiados es38.7 (IC al 95%: 24.6 a 52.7)."

Si en Statistix, en Descriptive Statistics tildamos la casilla que dice "Conf. Int." obtenemos:

```
DESCRIPTIVE STATISTICS

VARIABLE          N   LO 95% CI        MEAN   UP 95% CI          SD
PCR              29      24.007      38.665      53.324      38.537
```

Se observa que en el Statistix el IC al 95% resulta ser [24.0, 53.3] y nosotros obtuvimos [24.6, 52.7]. ¿Hay alguno erróneo? ¿Cuál de ellos? ¿Por qué esta discrepancia?

Tamaño de la muestra cuando no se conoce σ.

Con la muestra de 29 pacientes cardíacos el IC al 95% para la media (poblacional) de PCR resultó:

$$[38.665 - 14.026 \ , \ 38.665 - 14.026]$$

$$[24.639 \ , \ 52.691]$$

El intervalo obtenido tiene longitud = $2*14.026 = 28.05$

Que se puede hacer si se desea un IC al 95% con menor longitud?

Llamando L a la longitud del IC, se puede obtener de (12) que

$$L = 2 * z_{\alpha/2} * s \ / \sqrt{n}$$

Si deseamos obtener un IC al 95% para la media de PCR de longitud $L \leq 10$ deberá ser

$$L = 2 * 1.96 * 38.537 \ / \sqrt{n} \ \leq \ 10$$

Despejando n:

$$(2 * 1.96 * 38.537 \ ) / 10 \ \leq \ \sqrt{n}$$

$$[(2 * 1.96 * 38.537 \ ) / 10]^2 \leq \ n$$

$$228.2 \leq \ n$$

o sea que habría que tomar una muestra de 229 pacientes (o más) para obtener un IC de longitud menor o igual que 10.
Pero este procedimiento es aproximado, porque no podemos asegurar que tomando 229 pacientes la longitud del IC sea ≤ 10. Esto se debe que en el razonamiento anterior hay una pequeña trampita: cuando extraemos la nueva muestra de 229 pacientes, s ya no va a ser igual a 38.537 que es el valor que obtuvimos con 29 pacientes.

Lo que hemos aplicado es un procedimiento aproximado que podemos resumir en los siguiente pasos:

Procedimiento aproximado para obtener un IC de una longitud deseada:

1er paso: se obtiene una estimación de la desviación standard de algún estudio previo, de una muestra piloto, de la literatura o de la experiencia del investigador. Llamemos s* a esa estimación.

2do paso: usando s* se calcula el tamaño de la muestra:

$$L = 2 * 1.96 * s / \sqrt{n} \cong 2 * 1.96 * s^* / \sqrt{n} \leq \text{Longitud Deseada}$$

y de la última inecuación se despeja n.

3er. paso: Se extrae la muestra del tamaño calculado en el 2do paso y se calcula el IC basándose en esta muestra.

Con este procedimiento:
- No se asegura que la longitud del IC sea la deseada. Es un procedimiento aproximado.
- Si la estimación previa de la DS no es buena, esto puede hacer que el IC calculado no sea correcto y no tenga el nivel de confianza deseado del 95%?

Comentarios:
1) Hay otra aproximación en este procedimiento que en el ejemplo anterior no importó. En el ejemplo anterior n=229, pero si hubiésemos obtenido un n chico (digamos n<30), para calcular el IC en el 3er. paso, no habría que usar (12) sino (10) (siempre que la variable tenga distribución aproximadamente gaussiana) y esto no lo hemos tenido en cuenta en el primer paso. Pero esto es lo más fácil de arreglar (pensar o consultar). Lo más difícil de arreglar es lo que dijimos antes: el desconocimiento de s.
2) Existe un procedimiento exacto para calcular el n que asegure un IC de longitud ≤ long deseada. No lo explicaremos en este curso. Se usa poco, ya que en general el investigador quiere tener una idea aproximada de la longitud del IC, pero no tiene establecida la obligación de obtener un intervalo de longitud prefijada.

Intervalo de confianza con nivel asintótico para una proporción poblacional (o para el parámetro "p" de la distribución binomial).

Pensemos en un problema como el del ejemplo 3 en el que deseamos estimar una probabilidad o una proporción poblacional.
Podemos considerar el modelo:

$$X \sim \text{Bin}(n, p)$$

o el modelo

$$X_1, X_2, ..., X_n \text{ vs. as. i.i.d } \text{Bin}(1,p)$$

El estimador de p es la proporción muestral

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

Sabemos que si n es grande, la distribución binomial se puede aproximar por la normal.

$$X \;\tilde{a}\; N(np, np(1-p)) \quad \text{para n grande}$$

$$\hat{p} \;\tilde{a}\; N(p, \frac{p(1-p)}{n}) \quad \text{para n grande}$$

Estandarizando

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \;\tilde{a}\; N(0,1) \quad \text{para n grande}$$

Pero $\hat{p}$ es un estimador consistente de p, luego si n es grande ,

$$\hat{p} \cong p ,$$

por lo tanto

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \;\tilde{a}\; N(0,1) \quad \text{para n grande}$$

Luego:

$$\lim_{n\to\infty} P\left(-z_{\alpha/2} \le \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \le z_{\alpha/2}\right) = 1-\alpha$$

y continuando como en el caso del IC para μ se deduce que el intervalo

$$\left[\hat{p} + z_{\alpha/2} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \;,\; \hat{p} + z_{\alpha/2} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \qquad (13)$$

es un **intervalo de confianza para p con nivel de confianza asintótico 1-α.**

Ejemplo: Supongamos que en el ejemplo 3, de las 400 personas encuestadas 80 (20%) responden que están a favor de la medida de gobierno. Entonces

$$\hat{p} = \frac{80}{400} = 0.20$$

Si queremos un IC al 95% para p, entonces $z_{\alpha/2} = 1.96$ y reemplazando en (13) obtenemos:

$$\left[0.20 - 1.96 * \sqrt{\frac{0.20(1-0.20)}{400}} \;,\; 0.20 + 1.96 * \sqrt{\frac{0.20(1-0.20)}{400}}\right]$$

$$[0.20 - 0.0392 \;,\; 0.20 + 0.0392]$$

$$[0.1608 \;,\; 0.2392]$$

Redondeando el IC al 95% para la proporción de personas que están de acuerdo con la medida del gobierno es **[0.16 , 0.24].**