Tamaño de muestra e intervalos de confianza

26 de noviembre de 2024

1. $X \equiv N(\mu, \sigma)$

Sea X una variable aleatoria de la que se quiere extraer una muestra aleatoria simple para calcular un intervalo de confianza de μ con coeficiente de confianza $1 - \alpha = 0'95$ y error máximo de estimación de $e = 0'5$. Calcula el tamaño de muestra que se necesita en las siguientes condiciones:

1. $X \equiv N(\mu, \sigma)$ y la desviación típica es $\sigma = 2$.
2. $X \equiv N(\mu, \sigma)$ y la desviación típica está acotada superiormente por 6.
3. $X \equiv N(\mu, \sigma)$, no se conoce la desviación típica pero por un estudio previo se dispone de una muestra de tamaño 10 con los siguientes resultados:
    a) $\vec{x}_1 =$ `c(7.37, 2.88, 4.39, 5.19, 6.91, 6.42, 1.85, 2.41, 7.44, 4.09)`
    b) $\vec{x}_2 =$ `c(3.56, 4.28, 6.07, 5.27, 4.40, 4.10, 5.82, 5.32, 2.80, 7.73)`
    c) $\vec{x}_3 =$ `c(8.23, 6.34, 8.85, 2.54, 6.15, 3.30, 6.48, 4.76, 0.88, 6.59)`

Solución.

1. $X \equiv N(\mu, \sigma)$ y la desviación típica es $\sigma = 2$.

    El intervalo de confianza de μ es de la forma:

    $$\left(\bar{X} - z\frac{\sigma}{\sqrt{n}},\ \bar{X} + z\frac{\sigma}{\sqrt{n}}\right)$$

    Por tanto el error de estimación debe verificar que $e \leq z\frac{\sigma}{\sqrt{n}}$ y el valor de n viene dado por:

    $$n \geq \left(\frac{z\,\sigma}{e}\right)^2$$

    En las condiciones del problema se obtiene $n \geq 61'46334$, es decir, se necesitan al menos 62 observaciones en la muestra.

```
## Mínimo tamaño de muestra
alfa <- 0.05
e <- 0.5
## X = N(mu, sigma=2)
sigma <- 2
z <- qnorm(1-alfa/2, 0, 1)
n_minimo <- ceiling((z*sigma/e)^2)
```

2. $X \equiv N(\mu, \sigma)$ y la desviación típica está acotada superiormente por 6.

   Para garantizar las condiciones del problema se toma n que verifique :

$$n \geq \left(\frac{z\ \text{cota}(\sigma)}{e}\right)^2 \geq \left(\frac{z\ \sigma}{e}\right)^2$$

   La solución es 553'1701 y por tanto $n = 554$, un incremento muy importante debido a una aproximación poco precisa de σ.

```
## X = N(mu, sigma) ; sigma <= 6
cota_sigma <- 6
z <- qnorm(1-alfa/2, 0, 1)
n_minimo <- ceiling((z*cota_sigma/e)^2)
```

3. $X \equiv N(\mu, \sigma)$, no se conoce la desviación típica pero por un estudio previo se dispone de una muestra de tamaño 10 con los siguientes resultados.

   - $\vec{x}_1 = (7'37, 2'88, 4'39, 5'19, 6'91, 6'42, 1'85, 2'41, 7'44, 4'09)$
     La estimación de σ es $S = 2'0946$, muy parecida al valor real y por tanto el tamaño estimado de la muestra sale 68, muy parecido al real.
   - $\vec{x}_2 = (3'56, 4'28, 6'07, 5'27, 4'40, 4'10, 5'82, 5'32, 2'80, 7'73)$
     El valor de $S = 1'4144$ infraestima σ y el valor de n= 31 es la mitad del que debería ser.
   - $\vec{x}_3 = (8'23, 6'34, 8'85, 2'54, 6'15, 3'30, 6'48, 4'76, 0'88, 6'59)$
     El valor de $S = 2'524726$ sobrestima la dispersión y el tamaño estimado de muestra, $n = 98$, es mayor al necesario.

```
## X=N(mu, sigma) ; muestra piloto de tamaño 10
x <- round(rnorm(10, 5, 2), 2)
S <- sd(x)
z <- qnorm(1-alfa/2, 0, 1)
n_minimo <- ceiling((z*S/e)^2)
```

2. $X \not\equiv N(\mu, \sigma)$

Se desconoce la distribución de X y por tanto no se puede usar directamente la distribución normal para calcular los límites de la región que permite obtener el intervalo.

En este caso se podría usar la aproximación por el teorema central del límite, o utilizar la cota de Chebishof.

$$P\left(|\bar{X}-\mu| \leq e\right) = P\left(\bar{X}-e \leq \mu \leq \bar{X}-e\right) \geq 1 - \frac{\sigma^2/n}{e^2}$$

Para que el intervalo tenga el coeficiente de confianza $1-\alpha$ se debe cumplir:

$$1 - \frac{\sigma^2/n}{e^2} \geq 1-\alpha \Longleftrightarrow \frac{\sigma^2/n}{e^2} \leq \alpha$$

y por tanto la cota inferior para el tamaño de la muestra vale:

$$n \geq \frac{\sigma^2}{\alpha\, e^2}$$

Aplicando este criterio a los apartados anteriores se obtienen los siguientes tamaños de muestra:

	$\sigma = 2$	$\sigma \leq 6$	$S = 2'0873$
n	320	2880	349

```
## Chebishof ; n >= sigma^2/(alfa * e^2)
n3 <- ceiling(S^2/(alfa*e^2))
```

3. Teorema central del límite

Sea X una variable aleatoria con distribución de Bernoulli(p) de la que se quiere extraer una muestra aleatoria simple para calcular un intervalo de confianza de p con coeficiente de confianza $1-\alpha = 0'95$ y error máximo de estimación de $e = 0'10$.

Teniendo en cuenta que el estimador máximo-verosímil de p es $\bar{X}$ y el teorema central del límite, se puede aproximar el intervalo de confianza mediante la expresión:

$$(\bar{X} - z\frac{\sqrt{p\,(1-p)}}{n}, \bar{X} + z\frac{\sqrt{p\,(1-p)}}{n})$$

y la cota inferior para el tamaño de la muestra vale:

$$n \geq \frac{z^2\, p\,(1-p)}{e^2}$$

Dado que p es desconocido, el valor de n se puede aproximar usando una cota para $p\,(1-p)$ o estimándolo mediante una muestra.

1. Cota para $\text{Var}(X) = p\,(1-p)$. La $\text{Var}(X) \leq 0\text{'}25$ ya que $p\,(1-p)$ es una parábola con máximo en $p = 0\text{'}5$, es decir, cuando la incertidumbre sobre el resultado que va a ocurrir es máxima.

   En las condiciones del enunciado de este apartado se estima el mínimo tamaño de muestra en tres casos diferentes:

   a) No se sabe nada sobre p

   b) $p \geq 0\text{'}70$

   c) $p \leq 0\text{'}20$

	$\text{Var}(X) \leq 0\text{'}25$	$\text{Var}(X) \geq 0\text{'}70$	$\text{Var}(X) \leq 0\text{'}20$
n	97	81	62

2. Se toma una muestra para tener una idea aproximada del valor de p y sustituir la varianza por $\hat{p}\,(1-\hat{p})$.

   $\vec{x} = (1, 0, 0, 1, 1, 0, 0, 0, 0, 1)$; $\hat{p} = 0\text{'}4$

   El tamaño estimado de la muestra vale: $n \geq 93$

```
## Generación de la muestra
p <- runif(1, 0, 1)
x <- rbinom(10, 1, p)
p_est <- mean(x)
var <- p_est*(1-p_est)
n_minimo <- ceiling((z*sqrt(var)/e)^2)
```