Ejercicios Resueltos TP8

Índice

1. Ejercicio 3

Enunciado

Una máquina llenadora de latas de café dosifica cantidades variables con distribución normal de desvío estándar de 15 gramos. A intervalos regulares se toman muestras de 10 envases con el fin de estimar la dosificación media. Una de estas muestras arrojó una media de 246 gramos.

a) Obtener un intervalo de confianza del 90 % para la dosificación media.

b) ¿Cuántos envases más habría que pesar para poder obtener una estimación cuyo error de muestreo fuera 5 gramos?

Resolución

En esta parte de la materia, es de muchísima importancia diferenciar en lo que pasa antes y después de la muestra. Haremos esta distinción utilizando los términos "pre-muestra" y "post-muestra". Las variables aleatorias consideradas "pre-muestra" son denotadas con letras mayúsculas (por ejemplo, X_i) y los valores que toman "post-muestra" serán nombradas con letras minúsculas (por ejemplo, x_i).

Además, como esta parte de la materia trata de sacar conclusiones cuando hay parámetros poblacionales desconocidos, también podemos hacer un apartado con dichos parámetros. Por otro lado, en esta guía, querremos estimar el valor de uno de estos parámetros desconocidos, por lo que además habrá que hacer alguna mención sobre el parámetro que se busca estimar.

Por otro lado, haremos la distinción entre parámetros poblacionales (características de la población) y muestrales (características de la muestra). Generalmente, salvo indicación contraria, los parámetros poblacionales son denotados con letras griegas (μ, σ) y los parámetros muestrales son escritos mediante letras latinas $(\overline{x}, s)$.

Observación: Los parámetros poblacionales NO son variables aleatorias. Es un valor desconocido pero **fijo**, ya que se trata de una característica de la población, y la población se mantiene constante (o al menos eso debe ser posible asumirlo antes de arribar a cualquier conclusión). Justamente, al ser **fijo**, no puede ser considerado una variable. Por lo tanto, no diremos que los parámetros tienen una distribución de probabilidad.

Lo aleatorio proviene de la muestra elegida, ya que de toda la población, podemos elegir distintas muestras a partir de ella. En este ejercicio, elegimos $n = 10$ latas de café, y todos los valores obtenidos describirán esa muestra. Sin embargo, distintas muestras de 10 elementos tendrán aparejadas distintas medias. De ahí que todo aquello determinado por la muestra, antes de la extracción, será aleatorio.

Variables aleatorias

Definamos la siguiente variable aleatoria:

$$X_i = \text{peso de la } i\text{-ésima lata de café}$$

Esta definición es pre-muestra. Además, i toma valores entre 1 y n =10.

Parámetros y estimadores

El único parámetro desconocido es $\mu = E[X_i]$ y es el parámetro que se busca estimar, la media **poblacional**. Por otro lado, el parámetro **poblacional** $\sigma = 15$ es conocido.

El estimador para μ es la media **muestral** $\overline{X}_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$.

Datos

Tenemos los siguientes datos:

- $X_i \sim N(\mu, 15)$ (pre-muestra)
- Nivel de Confianza: 90 % (pre-muestra)
- Media **muestral**: $\overline{x}_{10} = \dfrac{\sum_{i=1}^{10} x_i}{10} = 246$ (post-muestra)

Además, debemos suponer que las variables aleatorias X_i son independientes e idénticamente distribuidas.

1.1. Item a

Pre-muestra

Debemos construir un intervalo de confianza del 90 % para μ. Es decir, valores a y b de forma que:

$$P(a \leq \mu \leq b) = 0{,}9$$

Sin embargo, como μ es desconocido, estos valores no pueden depender de μ. Pueden depender de lo conocido hasta el momento (σ) y lo que conoceremos después de tomar la muestra: $X_1, \cdots, X_{10}$.

Es decir, buscamos $a(X_1, \cdots, X_{10}, \sigma)$ y $b(X_1, \cdots, X_{10}, \sigma)$, de forma que encierre a la media poblacional μ con probabilidad 0.9.

Como fue mencionado, los parámetros μ (desconocido) y σ (conocido) no tienen distribución de probabilidad. Sin embargo, previo a tomar la muestra, influyen sobre la distribución de cada variable aleatoria X_i ya que sabemos que $E[X_i] = \mu$ y $\sigma[X_i] = \sigma$ para todo $1 \leq i \leq 10$.

Al ser la muestra de $n = 10$ elementos, no podemos hacer uso del Teorema Central del Límite. Sin embargo, al ser las X_i normales e independientes entre sí, al sumarlas y multiplicarlas por constantes, también son normales. Es decir, su promedio dado por:

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

también es normal. Para terminar de conocer su distribución, necesitamos su media y desvío:

$$E[\overline{X}_n] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{\sum_{i=1}^{n} \overbrace{E[X_i]}^{\mu}}{n} = \frac{\mu \cdot \sum_{i=1}^{n} 1}{n} = \mu$$

$$V[\overline{X}_n] = V\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n^2} \cdot V\left[\sum_{i=1}^{n} X_i\right] \overset{IND}{=} \frac{1}{n^2} \cdot \sum_{i=1}^{n} V\left[X_i\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \Rightarrow \sigma[\overline{X}_n] = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Por lo tanto,

$$\overline{X}_n \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N\left(0; 1\right)$$

Es decir, podemos realizar todos nuestros cálculos probabilísticos respecto de este último estadístico, ya que nos resulta de mayor simpleza encontrar dos valores $k_1, k_2 \in \mathbb{R}$ que cumplan:

$$P\left(k_1 \leq \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq k_2\right) = 0{,}9$$

Cualquier combinación de valores de k_1 y k_2 que cumplan esto permiten construir un intervalo de confianza adecuado. Sin embargo, recordemos que en un intervalo de confianza siempre conviene que la longitud del intervalo

sea lo menor posible, ya que la estimación en ese caso resulta más precisa. Se puede demostrar que para distribuciones simétricas (respecto del cero) el intervalo más estrecho viene dado por el caso en que estos límites también son simétricos, es decir, $k_1 = -k_2$. Para simplificar la notación, tomaremos $k_2 = k$ y $k_1 = -k$:

$$P\left(-k \leq \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq k\right) = 0{,}9 \Rightarrow \Phi(k) - \Phi(-k) = 0{,}9 \Rightarrow \Phi(k) - (1 - \Phi(k)) = 0{,}9 \Rightarrow 2\Phi(k) - 1 = 0{,}9 \Rightarrow$$

$$\Rightarrow \Phi(k) = \frac{0{,}9 + 1}{2} \Rightarrow k = z_{0{,}95}$$

Es decir, podemos encontrar el intervalo de confianza a partir de lo obtenido:

$$P\left(-z_{0{,}95} \leq \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{0{,}95}\right) = 0{,}9 \Rightarrow P\left(-z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}} \leq \overline{X}_n - \mu \leq z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}9 \Rightarrow$$

Multiplicando por -1 todos los miembros de la inecuación, se invierten los signos:

$$P\left(z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu - \overline{X}_n \geq -z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}9 \Rightarrow P\left(\overline{X}_n + z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu \geq \overline{X}_n - z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}9$$

Es decir, cuando se obtenga la muestra y su media muestral, el intervalo de confianza de 90 % será:

$$IC_{90\,\%}(\mu) = [a(X_1, \cdots, X_{10}, \sigma); b(X_1, \cdots, X_{10}, \sigma)] = \left[\overline{X}_n - z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X}_n + z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Comentarios:

- Notemos que este cálculo nunca requirió saber los valores de σ ni de n. Es decir, el cálculo es extrapolable a cualquier valor que puedan tomar estos parámetros. Lo que hemos utilizado es el nivel de confianza (90 %), aunque se pueden repetir los mismos pasos para cualquier nivel de confianza.
- Un error común es reemplazar la media muestral en el último cálculo probabilístico. Es decir, como la media muestral ($\overline{x}_{10}$) toma el valor 246, la expresión errónea es la siguiente:

$$\underbrace{P\left(\overline{X}_n + z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu \geq \overline{X}_n - z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}9}_{\textbf{Bien}} \Rightarrow \underbrace{P\left(246 + z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu \geq 246 - z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0{,}9}_{\textbf{Mal}}$$

  Esta expresión exhibe errores conceptuales y matemáticos. Un error conceptual es utilizar el valor de la media muestral (claramente, obtenida "post-muestra") en el cálculo probabilístico, en el que justamente se intenta predecir lo que sucederá una vez obtenida la muestra. Por otro lado, confunde $\overline{X}_n$ (una variable aleatoria) con la media muestral $\overline{x}_n$ (un número).

  Por último, recordando que μ es un parámetro poblacional, todos los valores de la expresión errónea son **fijos**, es decir, no se pueden cumplir las inecuaciones con una cierta probabilidad. Esto describe un error matemático. Al ser valores fijos, las inecuaciones se cumplen o no se cumplen. Dicho de otro modo, la "probabilidad"de que se cumplan sólo pueden ser 0 o 1.

Post-muestra

Después de tomar la muestra, sabemos que la media muestral es $\overline{x}_{10} = 246$. Además, recordemos que $\sigma = 15$, $n = 10$ y $z_{0{,}95} \approx 1{,}644854$. Es decir, el intervalo de confianza está dado por:

$$IC_{90\,\%}(\mu) = \left[\overline{x}_{10} - z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X}_n + z_{0{,}95} \cdot \frac{\sigma}{\sqrt{n}}\right] \approx \left[246 - 1{,}644854 \cdot \frac{15}{\sqrt{10}}; 246 + 1{,}644854 \cdot \frac{15}{\sqrt{10}}\right] = [238{,}1978; 253{,}8022]$$

Comentario: Otro error común es decir que este intervalo tiene un 90 % de probabilidad de contener a la media poblacional, y como hemos dicho antes, todos los valores son fijos, por lo que no puede contener la media con una probabilidad que no sea 0 ni 1.

Hay dos formas de interpretar un intervalo de confianza, y ambas son previas a tomar la muestra:

- Un intervalo construido mediante la fórmula

$$IC_{90\,\%}(\mu) = \left[\overline{X}_n - z_{0,95} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X}_n + z_{0,95} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

  tiene un 90 % de probabilidad de incluir a la media poblacional μ.

- Si se toman M muestras de tamaño n, aproximadamente el 90 % de los intervalos construidos de esta manera contendrán a la media poblacional μ. Por ejemplo, si se toman $M = 1000$ muestras, aproximadamente 900 de los intervalos construidos a partir de ellas contendrán a la media.

1.2. Item b

Nos piden un error de muestreo de 5 gramos. El error de muestreo es el error máximo de la distancia de cualquiera de los puntos del intervalo al límite más cercano. Es decir, esta distancia máxima se da en el centro del intervalo, ya que tiene la máxima distancia a ambos límites.

Por lo tanto, el error de muestreo es la semiamplitud del intervalo. Y como dicho intervalo está dado por:

$$IC_{90\,\%}(\mu) = \left[\overline{X}_n - z_{0,95} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X}_n + z_{0,95} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

el error de muestreo E se obtiene del siguiente modo:

$$E = z_{0,95} \cdot \frac{\sigma}{\sqrt{n}}$$

Para obtener este nivel de error deseado, debemos encontrar un valor de n que satisfaga dicha condición, ya que es justamente el tamaño de muestra que debemos determinar:

$$E = z_{0,95} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} = z_{0,95} \cdot \frac{\sigma}{E} \Rightarrow n = z_{0,95}^2 \cdot \frac{\sigma^2}{E^2} = 1{,}644854^2 \frac{15^2}{5^2} \approx 24{,}3499$$

Obviamente, no se puede tomar un número fraccionario de envases, por lo que habrá que determinar un número entero a través de alguna desigualdad. En el caso de tener que elegir, nos gustaría que el error sea el menor valor posible, por lo que si no podemos obtener $E = 5$, trataremos de que $E \leq 5$:

$$5 \geq E = z_{0,95} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} \geq z_{0,95} \cdot \frac{\sigma}{5} \Rightarrow n = z_{0,95}^2 \cdot \frac{\sigma^2}{5^2} = 1{,}644854^2 \frac{15^2}{5^2} \approx 24{,}3499 \Rightarrow n \geq 25$$

Cuidado: hay que prestar atención al enunciado. Si bien hemos obtenido un número de muestra de 25 latas, nos piden cuántos envases **más** debemos seleccionar, a partir del tamaño de muestra inicial ($n = 10$). Es decir, la respuesta correcta es:

Deben pesarse 15 envases más.

2. Ejercicio 8

Enunciado

En la siguiente tabla se presentan los datos del contenido de silicio en una muestra de 150 coladas de hierro :

Contenido de sílice	Cantidad de coladas
0,333 – 0,433	4
0,433 – 0,533	12
0,533 – 0,633	19
0,633 – 0,733	28
0,733 – 0,833	48
0,833 – 0,933	25
0,933 – 1,033	14

Estimar con una confianza del 95 % el contenido medio de sílice por colada.

Resolución

Variables aleatorias

Consideramos las siguientes variables aleatorias:

$$X_i = \text{contenido de silicio en la } i\text{-ésima colada de hierro.}$$

Parámetros y estimadores

En este caso, tanto la media **poblacional** $E[X_i] = \mu$ como el desvío **poblacional** $\sigma(X_i) = \sigma$ son desconocidos.
Para estimar μ, usamos la media **muestral** $\overline{X}_n$.
Para estimar σ, utilizamos el desvío **muestral** s.

Datos

Nos piden un nivel de confianza de 95 %. Deberíamos asumir que las variables son independientes e idénticamente distribuidas.

Pre-muestra

No tenemos certeza sobre la normalidad de los datos, por lo tanto, no es adecuado utilizar la distribución t de Student. Sin embargo, como $n = 150$ es suficientemente grande, podemos usar el teorema central del límite para hipotetizar que la media **muestral** tenga distribución aproximadamente normal.

Por lo tanto, habrá que usar los cuantiles dados por $z_{\frac{1+0,95}{2}} = z_{0,975}$ y el intervalo de confianza vendrá dado por:

$$IC_{95\,\%}(\mu) = \left[\overline{X}_n - z_{0,975} \cdot \frac{s}{\sqrt{n}}, \overline{X}_n + z_{0,975} \cdot \frac{s}{\sqrt{n}}\right]$$

El mayor problema es que los datos vienen agrupados, y no podemos calcular los parámetros muestrales con las sumatorias que empleamos usualmente. Debemos calcularlos con las fórmulas de datos agrupados:

$$\overline{X}_n^{Ag} = \frac{\sum_{j=1}^{L} X_j \cdot f_j}{n}$$

$$s^{Ag} = \sqrt{\frac{\sum_{j=1}^{L} (X_j - \overline{X}_n^{Ag})^2 \cdot f_j}{n-1}}$$

donde L es la cantidad de intervalos, X_j es la marca de clase del j-ésimo intervalo y f_j es la frecuencia absoluta del j-ésimo intervalo.

Es decir, de forma más precisa, el intervalo de confianza vendrá dado por:

$$IC_{95\,\%}(\mu) = \left[\overline{X}_n^{Ag} - z_{0,975} \cdot \frac{s^{Ag}}{\sqrt{n}}, \overline{X}_n^{Ag} + z_{0,975} \cdot \frac{s^{Ag}}{\sqrt{n}}\right]$$

Post-muestra

Luego de la muestra tenemos la siguiente tabla de datos agrupados:

L_{inf}	L_{sup}	f_j
0,333	0,433	4
0,433	0,533	12
0,533	0,633	19
0,633	0,733	28
0,733	0,833	48
0,833	0,933	25
0,933	1,033	14

Agregando las columnas necesarias:

L_{inf}	L_{sup}	f_j	x_j	$x_j \cdot f_j$	$x_j - \overline{x}_{150}^{Ag}$	$(x_j - \overline{x}_{150}^{Ag})^2 \cdot f_j$
0,333	0,433	4	0,383	1,532	−0,356	0,5088
0,433	0,533	12	0,483	5,796	−0,256	0,7905
0,533	0,633	19	0,583	11,077	−0,156	0,4663
0,633	0,733	28	0,683	19,124	−0,0566	0,0899
0,733	0,833	48	0,783	37,584	0,0433	0,0901
0,833	0,933	25	0,883	22,075	0,143	0,5136
0,933	1,033	14	0,983	13,762	0,243	0,8289
	N	150	$\sum_{j=1}^{L} x_j \cdot f_j$	110,95	$\sum_{j=1}^{L} (x_j - \overline{x}_{150}^{Ag})^2 \cdot f_j$	3,288
			$\overline{x}_{150}^{Ag}$	0,7396	s^{Ag}	0,1485

Por lo tanto, como $z_{0,975} = 1{,}9599$:

$$IC_{90\,\%}(\mu) = \left[\overline{x}_{150}^{Ag} - z_{0,975} \cdot \frac{s^{Ag}}{\sqrt{n}}; \overline{x}_{150}^{Ag} + z_{0,975} \cdot \frac{s^{Ag}}{\sqrt{n}}\right] = \left[0{,}7396 - 1{,}9599 \cdot \frac{0{,}1485}{\sqrt{150}}; 0{,}7396 + 1{,}9599 \cdot \frac{0{,}1485}{\sqrt{150}}\right]$$

$$IC_{90\,\%}(\mu) = [0{,}7158; 0{,}7633]$$

3. Ejercicio 17

Enunciado

El rating de un programa de televisión se mide como el porcentaje de hogares que está viendo el programa en un momento dado. Una compañía medidora de rating cuenta con un panel de 600 hogares colaboradores, en los cuales ha instalado un people meter (dispositivo que registra cada minuto si el televisor está encendido y en qué canal, y envía telefónicamente la información a la base de datos durante la noche). Se ha registrado el rating del programa La noche del 10 en 25 puntos. Es decir que el 25 % de los hogares del panel vió todo el programa ese día.

a) Calcular un intervalo de confianza del 90 % para el rating de La noche del 10.

b) Sabiendo que cada people meter cuesta $\$C$, calcular la inversión adicional en dispositivos necesaria para medir el rating con un error de muestreo de $\pm 1\,\%$

Resolución

Variables aleatorias

Podemos considerar la siguiente variable aleatoria:

$$X_n = \text{cantidad de televisores que sintonizan el programa entre el panel de } n \text{ hogares.}$$

Por otro lado, podríamos considerar:

$$\widehat{p}_n = \frac{X_n}{n} = \textbf{proporción} \text{ de televisores que sintonizan el programa entre el panel de } n \text{ hogares.}$$

Parámetros y estimadores

El parámetro **poblacional** es p. Es decir, la proporción de hogares que sintonizan el programa en el total de la población. El estimador de p que utilizamos es $\widehat{p}_n$.

Notar que a diferencia de lo que ocurre para la media, el parámetro poblacional no está representado con una letra griega. Sin embargo, es otra notación usual utilizar un acento circunflejo ($\widehat{\ }$) para un estimador. Es decir, aquí para diferenciar lo muestral de lo poblacional, se utiliza dicho acento para estimadores muestrales y se omite para parámetros poblacionales.

Datos

El panel de $n = 600$ hogares de la muestra son seleccionados a partir de la población. Por lo tanto, asumiendo que la probabilidad de que los hogares sintonicen el programa son independientes entre sí, y que tienen la misma probabilidad p de mirar el programa que el resto de la **población**, obtenemos que X_n tiene una distribución binomial:

$$X_n \sim \text{Bi}(n, p)$$

donde p es la proporción **poblacional**.

3.1. Item a

Pre-muestra

Ahora debemos encontrar dos valores (a y b) a partir de los valores conocidos (n) y los que serán conocidos a partir de la muestra ($\widehat{p}_n$), de forma que contengan con alta probabilidad a la proporción poblacional p:

$$P\left(a(\widehat{p}_n, n) \leq p \leq b(\widehat{p}_n, n)\right) = 0{,}9$$

Para determinar los extremos del intervalo $a(\widehat{p}_n, n)$ y $b(\widehat{p}_n, n)$, debemos partir de alguna distribución conocida. Como $X_n \sim \text{Bi}(n, p)$ y $n = 600$, podemos utilizar el Teorema Central del Límite para decir que X_n tiene distribución aproximadamente normal:

$$X_n \overset{(a)}{\sim} N(np; \sqrt{np(1-p)}) \Rightarrow \widehat{p}_n = \frac{X_n}{n} \overset{(a)}{\sim} N\left(p; \sqrt{\frac{p(1-p)}{n}}\right) \Rightarrow \frac{\widehat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0; 1)$$

Por lo tanto,

$$P\left(-z_{0,95} \leq \frac{\widehat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{0,95}\right) = 0{,}9 \Rightarrow P\left(-z_{0,95}\sqrt{\frac{p(1-p)}{n}} \leq \widehat{p}_n - p \leq z_{0,95}\sqrt{\frac{p(1-p)}{n}}\right) = 0{,}9$$

$$P\left(z_{0,95}\sqrt{\frac{p(1-p)}{n}} \geq p - \widehat{p}_n \geq -z_{0,95}\sqrt{\frac{p(1-p)}{n}}\right) = 0{,}9$$

$$P\left(\widehat{p}_n + z_{0,95}\sqrt{\frac{p(1-p)}{n}} \geq p \geq \widehat{p}_n - z_{0,95}\sqrt{\frac{p(1-p)}{n}}\right) = 0{,}9$$

Es decir, hemos establecido dos valores a y b que contienen a la proporción poblacional con alta probabilidad. Sin embargo, notar que estos límites dependen del parámetro p que no se puede conocer aunque tomemos la muestra. Por otro lado, por la ley de los grandes números,

$$\lim_{n \to +\infty} P(|\widehat{p}_n - p| \geq \varepsilon) = 0, \ \forall \varepsilon > 0$$

Es decir, para valores grandes de n, es muy baja la probabilidad de que $\widehat{p}_n$ y p tomen valores muy diferentes. Por lo tanto, como $n = 600$, podemos asumir que $\widehat{p}_n \approx p$ y que por lo tanto:

$$0{,}9 = P\left(\widehat{p}_n + z_{0,95}\sqrt{\frac{p(1-p)}{n}} \geq p \geq \widehat{p}_n - z_{0,95}\sqrt{\frac{p(1-p)}{n}}\right) \approx P\left(\underbrace{\widehat{p}_n + z_{0,95}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}}_{b(\widehat{p}_n, n)} \geq p \geq \underbrace{\widehat{p}_n - z_{0,95}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}}_{a(\widehat{p}_n, n)}\right)$$

Obviamente, el valor p del miembro medio de la inecuación se mantiene fijo ya que si también es reemplazado por $\widehat{p}_n$ dejamos de tener un intervalo de confianza.

De este modo, tenemos que, una vez tomada la muestra, el intervalo de confianza vendrá dado por:

$$IC_{90\,\%}(p) = \left[\widehat{p}_n - z_{0,95}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}; \widehat{p}_n + z_{0,95}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right]$$

Comentario: Notar que esto no depende del valor de n y se usa el mismo cálculo para otros tamaños muestrales.

Post-muestra

Luego de tomar la muestra, se obtiene $\widehat{p}_{600} = 0{,}25$ y como $z_{0{,}95} = 1{,}6448$, el intervalo de confianza viene dado por:

$$IC_{90\,\%}(p) = \left[0{,}25 - 1{,}6448 \cdot \sqrt{\frac{0{,}25 \cdot 0{,}75}{600}}; 0{,}25 + 1{,}6448 \cdot \sqrt{\frac{0{,}25 \cdot 0{,}75}{600}}\right] = [0{,}2209; 0{,}2791] = [22{,}09\,\%; 27{,}91\,\%]$$

3.2. Item b

Ahora nos piden calcular la inversión extra si se requiere un margen de error de 1 %. El costo extra estará vinculado con un nuevo tamaño muestral n.

Para que el margen de error se reduzca a los niveles deseados debe ocurrir lo siguiente:

$$z_{0{,}95} \cdot \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}} \leq 0{,}01 \Rightarrow \sqrt{n} \leq z_{0{,}95} \cdot \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{0{,}01} \Rightarrow n \geq \frac{z^2_{0{,}95} \cdot \widehat{p}_n(1-\widehat{p}_n)}{0{,}01^2}$$

El problema es que al tomar mayor muestra $\widehat{p}_n$ ya no es necesariamente igual a $\widehat{p}_{600} = 0{,}25$. Sin embargo, sabemos que, dado que $\widehat{p}_n \in (0, 1)$, $\widehat{p}_n(1-\widehat{p}_n) \leq 0{,}25$. Por lo tanto, si se cumple:

$$n \geq \frac{z^2_{0{,}95} \cdot 0{,}25}{0{,}01^2} \geq \frac{z^2_{0{,}95} \cdot \widehat{p}_n(1-\widehat{p}_n)}{0{,}01^2}$$

entonces se cumple lo pedido. Por lo tanto, determinamos el tamaño muestral según:

$$n \geq \frac{z^2_{0{,}95} \cdot 0{,}25}{0{,}01^2} \Rightarrow n \geq 6763{,}859 \Rightarrow n = 6764$$

Es decir, como cada people-meter cuesta \$C, entonces la inversión **extra** es \$6164 C (recordar que ya habían instalados 600 people-meters).

Dependiendo del valor de C, este costo extra puede ser muy considerable y se puede intentar buscar una alternativa de menor costo. Por eso, haremos el cálculo de otra manera.

En el cálculo utilizamos una cota, basados en que desconocíamos el valor de $\widehat{p}_n$. Sin embargo, siguiendo el mismo razonamiento en la aproximación de p por $\widehat{p}_{600} = 0{,}25$, podemos asumir también que $p \approx \widehat{p}_n$ y por lo tanto $\widehat{p}_{600} \approx \widehat{p}_n$:

$$n \geq \frac{z^2_{0{,}95} \cdot \widehat{p}_n(1-\widehat{p}_n)}{0{,}01^2} \approx \frac{z^2_{0{,}95} \cdot 0{,}25(1-0{,}25)}{0{,}01^2} = 5072{,}894 \Rightarrow n = 5073$$

Por lo tanto, la inversión **extra** sería de \$4473 C. Por lo tanto, esta aproximación más precisa disminuyó considerablemente la inversión extra respecto a la anterior estrategia más conservadora.

Sobre las aproximaciones

En este último ejercicio hubo algunas aproximaciones que pueden generar alguna inquietud sobre su implementación.

Primero, en el cálculo del intervalo de confianza, se aproxima el desvío real de $\widehat{p}_n$ por una aproximación que no involucra al parámetro p.

Para analizar el impacto de dicha aproximación, se hizo una simulación en la que se toman distintos valores de p y n, y se construyeron los siguientes intervalos de confianza, utilizando el desvío real y el desvío aproximado:

$$IC^R_{95\,\%}(p) = \left[\widehat{p}_n - z_{0{,}975}\sqrt{\frac{p(1-p)}{n}}; \widehat{p}_n + z_{0{,}975}\sqrt{\frac{p(1-p)}{n}}\right]$$

$$IC^E_{95\,\%}(p) = \left[\widehat{p}_n - z_{0{,}975}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}; \widehat{p}_n + z_{0{,}975}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right]$$

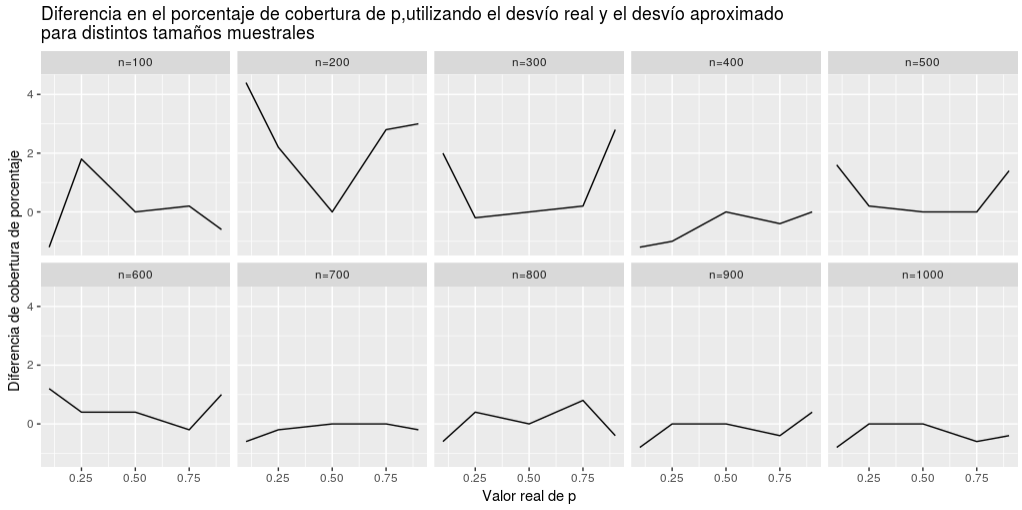
como es una simulación, el valor de p es conocido y podemos calcular el intervalo IC^R que en la práctica es imposible o demasiado costoso.

Luego, para cada combinación de n y p, se tomaron $M = 500$ muestras binomiales con dichos parámetros, se calcularon los intervalos de confianza con ambas fórmulas, y se determinó el porcentaje de cobertura de p, basado en las $M = 500$ iteraciones.

Es decir, para cada combinación de n y p, se calculó:

- $PC^R = 100 \cdot \dfrac{\#\text{intervalos } (IC^R) \text{ que contienen a } p}{M}\,\%$
- $PC^E = 100 \cdot \dfrac{\#\text{intervalos } (IC^E) \text{ que contienen a } p}{M}\,\%$

El próximo gráfico muestra $PC^R - PC^E$ para cada combinación de p y n:

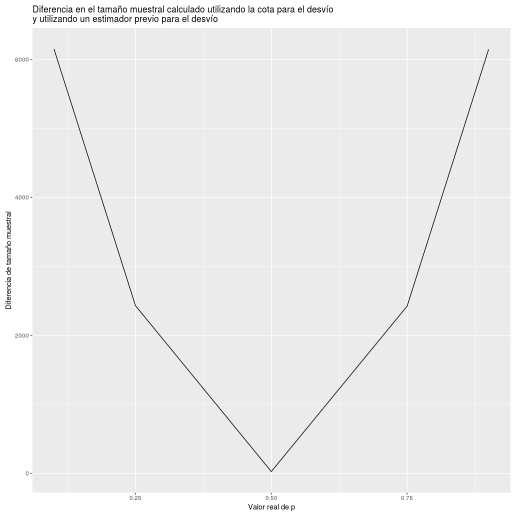


Notar que si la diferencia es negativa, es porque el intervalo con el desvío estimado tuvo un mayor porcentaje de cobertura.

Hay dos cuestiones que se observan:

- A medida que aumenta el tamaño muestral, la diferencia de porcentaje cubierto es menor. Es decir, más cercano al cero. Tiene sentido dado que a mayor tamaño muestral, mejor será la estimación de p a través de $\widehat{p}_n$.
- Cuando $p = 0{,}5$, la diferencia siempre es muy cercana a cero. Es decir, la diferencia se agudiza en valores lejanos a $p = 0{,}5$. Esto tiene su explicación dado que a medida que p se aleja de 0,5, el desvío se achica, por lo que la longitud de los intervalos se hace menor. De este modo, al ser más angosto el intervalo, menos chances tiene de contener a p.

Respecto a la aproximación del tamaño muestral, a continuación mostramos la diferencia con el n calculado a partir de la cota $p(1-p) \leq 0{,}25$ y la aproximación utilizando la estimación previa de p, considerando distintos valores para n



Vemos que como la cota asume $p = 0{,}5$, la diferencia se hace cero para este valor. Sin embargo, notemos que la diferencia puede ser realmente considerable a medida que p se aleja de 0,5, con diferencias de hasta 6000 muestras.