Probabilidad y Estadística (93.24)
Estimación de parámetros

Índice

1. Repaso de algunos conceptos

Generalidades de Estimación

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias cuya distribución está relacionada con un *parámetro* θ. Si θ es desconocido, se necesita un algoritmo para calcular una aproximación al mismo. A este algoritmo o regla de cálculo se lo denomina *estimador* de θ. Es común escribir al estimador como $\hat{\theta}$. Dado que el estimador es una función de las variables aleatorias, $\hat{\theta}(X_1, X_2, \cdots, X_n)$ es aleatorio. Si se realiza uno o varios experimentos de los cuales se determinan valores $x_1, x_2, \cdots, x_n$, a la aproximación resultante $\hat{\theta}(x_1, x_2, \cdots, x_n)$ se la denomina una *estimación* del valor del parámetro θ.
Por ejemplo, si las variables aleatorias son i.i.d., θ puede ser $\mu = \mathrm{E}[X_1]$. Un estimador común de μ es el promedio:

$$\hat{\mu} = \hat{\mu}(X_1, \cdots, X_n) = \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Sabemos que $\hat{\mu}$ es una variable aleatoria. Si se usa $n = 2$ y se realizan experimentos en los que se obtienen $x_1 = 2$, $x_2 = 4$, una estimación del valor esperado μ será

$$\hat{\mu}(2, 4) = \frac{1}{2}(2 + 4) = 3.$$

Debemos destacar que $\bar{X}$ es lo que se conoce como un *estimador puntual*, ya que devuelve un solo valor como aproximación del parámetro.

Una forma habitual de medir la bondad de un estimador puntual es mediante el *error cuadrático medio*:

$$\mathrm{ECM}(\hat{\theta}) = \mathrm{E}\left[\left(\hat{\theta} - \theta\right)^2\right].$$

En líneas generales, se busca que los estimadores tengan el menor error cuadrático medio. No es muy difícil demostrar que

$$\mathrm{ECM}(\hat{\theta}) = \mathrm{sesgo}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta}),$$

donde

$$\mathrm{sesgo}(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta, \qquad \mathrm{Var}(\hat{\theta}) = E\left[\left(\hat{\theta} - \mathrm{E}[\hat{\theta}]\right)^2\right].$$

Cuando $\mathrm{E}[\hat{\theta}] = \theta$, se dice que el estimador es *insesgado*. En ciertos casos, se puede encontrar el estimador "ideal", esto es, el *estimador insesgado de mínima varianza*.

En ocasiones, el estimador puede ser insesgado en general, pero sí cuando se toma un n grande. En decir, se dice que $\hat{\theta}$ es asintóticamente insesgado sii

$$\lim_{n\to\infty} \mathrm{E}[\hat{\theta}(X_1, \cdots, X_n)] - \theta = 0.$$

En relación con este concepto asintótico, se define un estimador *consistente* como aquel para el cual

$$\lim_{n\to\infty} \hat{\theta}(X_1, \cdots, X_n) = \theta.$$

Estimador de máxima verosimilitud

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias continuas tales que la función de densidad de probabilidad conjunta se puede escribir como $\mathrm{f}(x_1, \cdots, x_n; \theta)$, siendo θ un parámetro a a estimar. A la función $\mathrm{f}(\cdots; \theta)$ se la denomina verosimilitud. Dada una muestra, el estimador de máxima verosimilitud está dado por

$$\hat{\Theta} = \arg\max_{\theta} \mathrm{f}(X_1, \cdots, X_n; \theta), \tag{1}$$

es decir, el valor de θ que maximiza la "probabilidad" para los valores de la muestra obtenida. En el caso de variables aleatorias discretas, hay que cambiar la función de densidad por una función de masa de probabilidad.
Se puede demostrar que los estimadores de máxima verosimilitud son asintóticamente consistentes. Otras propiedades están fuera del alcance de este curso.
Un ejemplo: $X_i \sim \mathcal{N}(\mu, 1)$ i.i.d., con μ desconocido. Luego,

$$\mathrm{f}(X_1, \cdots, X_n; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2}}. \tag{2}$$

Dado que el logaritmo es una función creciente, podemos trabajar con

$$\mathcal{L}(\mu) = \ln \mathrm{f}(X_1, \cdots, X_n; \mu) = -\frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2}. \tag{3}$$

Derivando respecto de μ e igualando a cero, podemos encontrar el máximo:

$$\left.\frac{\partial \mathcal{L}}{\partial \mu}\right|_{\mu=\hat{\mu}} = \sum_{i=1}^{n}(X_i - \hat{\mu}) = \sum_{i=1}^{n} X_i - n\hat{\mu} = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{4}$$

Método de los momentos

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias i.i.d.. Sea w_k una función tal que el momento k-ésimo ($k \in \mathbb{N}$) puede escribirse como

$$\mu_k = \mathrm{E}\left[X_i^k\right] = w_k(\theta), \tag{5}$$

donde θ es un parámetro desconocido de la distribución de X_i. Supongamos que existe un estimador para dicho momento, esto es, $\hat{\mu}_k(X_1, X_2, \cdots, X_n)$. Luego, se puede estimar el parámetro como

$$\hat{\Theta} = w_k^{-1}(\hat{\mu}_k). \tag{6}$$

Como ejemplo concreto, sean $X_i \sim \mathrm{Exp}(\theta)$. Un estimador para el primer momento (ver más abajo) es:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{7}$$

es decir, la media poblacional se puede estimar por la media muestral. Dado que $\mu = 1/\theta$, tenemos:

$$\frac{1}{\hat{\Theta}} = \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow \hat{\Theta} = \frac{n}{\sum_{i=1}^{n} X_i}. \tag{8}$$

En el caso de que existan varios parámetros a estimar, se pueden utilizar varios momentos. Además, también pueden usarse momentos centrales (como la varianza).

Máximo *a posteriori* (MAP)

Esta forma de encontrar estimadores permite incorporar información sobre el parámetro que se tenga **antes** de tomar la muestra. Para ello, el primer paso es **considerar al parámetro desconocido Θ como una variable aleatoria**. La información sobre él es incorporada a través de la distribución del mismo $\mathrm{f}_\Theta(\theta)$. Esta distribución se conoce como *a priori*, porque condensa información que se tiene **antes** de tomar la muestra.
Sean $X_1, X_2, \cdots, X_n$ variables aleatorias continuas tales que la función de densidad de probabilidad conjunta **dado el valor de** θ se puede escribir como $\mathrm{f}(x_1, \cdots, x_n|\theta)$.
Usando el Teorema de Bayes (por eso, este método se denomina *Bayesiano*), podemos escribir la densidad de probabilidad del parámetro Θ dada la muestra como:

$$\mathrm{f}(\theta|X_1, \cdots, X_n) = \frac{\mathrm{f}(X_1, \cdots, X_n|\theta)\mathrm{f}_\Theta(\theta)}{\int \mathrm{f}(X_1, \cdots, X_n|\zeta)\mathrm{f}_\Theta(\zeta)d\zeta}. \tag{9}$$

Esta densidad se llama *a posteriori*, porque se puede evaluar **después** de conocer la muestra
El estimador del parámetro es

$$\hat{\theta} = \arg\max_\theta \mathrm{f}(\theta|X_1, \cdots, X_n). \tag{10}$$

Como ejemplo concreto, sean $X_i \sim \mathcal{N}(M, 1)$ i.i.d., con media M desconocida. Asumamos que se sabe que M está cerca de cierto valor μ_p y esa información *a priori* se puede escribir suponiendo que $M \sim \mathcal{N}(\mu_p, \sigma_p)$. Luego,

$$\mathrm{f}(X_1, \cdots, X_n|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2}}, \tag{11}$$

$$\mathrm{f}_M(\mu) = \frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-\frac{(\mu-\mu_p)^2}{2\sigma_p^2}}. \tag{12}$$

La densidad *a posteriori* es

$$\mathrm{f}(\mu|X_1, \cdots, X_n) = \frac{\frac{1}{\sqrt{2\pi\sigma_p^2}} e^{-\frac{(\mu-\mu_p)^2}{2}} \times \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2\sigma_p^2}}}{\Lambda}, \tag{13}$$

donde Λ es un número que **no depende** de μ. Dado que el logaritmo es una función creciente, podemos trabajar con

$$\mathcal{Q}(\mu) = -\frac{n+1}{2}\ln(2\pi) - \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2} - \ln(\sigma_p) - \frac{(\mu-\mu_p)^2}{2\sigma_p^2} - \ln(\Lambda). \tag{14}$$

Derivando respecto de μ e igualando a cero, podemos encontrar el máximo:

$$\left.\frac{\partial \mathcal{Q}}{\partial \mu}\right|_{\mu=\hat{\mu}} = \sum_{i=1}^{n}(X_i - \hat{\mu}) - \frac{(\hat{\mu}-\mu_p)}{\sigma_p^2} = 0 \Rightarrow \hat{\mu} = \frac{1}{1+n\sigma_p^2}\mu_p + \frac{n\sigma_p^2}{1+n\sigma_p^2}\frac{1}{n}\sum_{i=1}^{n} X_i. \tag{15}$$

Obsérvese que, a medida que la muestra crece, la información *a priori* pierde importancia (ver el límite cuando $n \to \infty$).

Estimación puntual de la media

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias i.i.d. con media $\mu = \mathrm{E}[X_1]$. Un estimador de la media es el promedio:

$$\hat{\mu} = \hat{\mu}(X_1, \cdots, X_n) = \bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Algunos hechos:

1. Es un estimador insesgado.
2. Por la ley de los grandes números, es un estimador consistente.
3. Si $\sigma^2 = \mathrm{Var}\,[X_1]$,
$$\mathrm{ECM}(\hat{\mu}) = \mathrm{Var}\left[\bar{X}\right] = \frac{\sigma^2}{n}.$$
4. Si $X_1 \sim \mathcal{N}(\mu, \sigma)$, entonces $\hat{\mu} \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$.
5. Si $\sigma^2 = \mathrm{Var}\,[X_1] < \infty$, por el Teorema Central del Límite, $\hat{\mu}$ es asintóticamente normal. Es decir que, para n grande, la distribución del estimador $\hat{\mu}$ se puede aproximar por la de una variable aleatoria $\sim \mathcal{N}(\mu, \sigma/\sqrt{n})$.

Estimación puntual de la varianza

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias i.i.d. con media $\mu = \mathrm{E}[X_1]$ y varianza $\sigma^2 = \mathrm{Var}\,[X_1]$. Un estimador de la varianza es la varianza muestral:

$$\hat{\sigma}^2 = \hat{\sigma}^2(X_1, \cdots, X_n) = S^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2 = \frac{1}{n-1}\sum_{k=1}^{n} X_k^2 - \frac{n}{n-1}\bar{X}^2.$$

Algunos hechos:

1. Es un estimador insesgado (¡gracias al $(n-1)$ en el denominador!).
2. La varianza del estimador es
$$\mathrm{ECM}(\hat{\sigma}^2) = \mathrm{Var}\left[S^2\right] = \frac{\sigma^4}{n}\left(\kappa + 2 + \frac{2}{n-1}\right),$$
donde
$$\kappa = \text{curtosis} = \frac{\mathrm{E}\left[(X_1 - \mathrm{E}[X_1])^4\right]}{\sigma^4} - 3.$$
Si X_1 es normal, se reduce a $\mathrm{Var}\left[S^2\right] = 2\sigma^4/(n-1)$.
3. Si $X_1 \sim \mathcal{N}(\mu, \sigma)$, entonces $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.
4. Si $\kappa < \infty$, Chebychev nos permite probar una versión débil de los grandes números y que se trata de un estimador consistente.
5. Si $\kappa < \infty$, por una versión del Teorema Central del Límite, S^2 es asintóticamente normal. Es decir que, para n grande, la distribución de estimador S^2 se puede aproximar por la de una variable aleatoria $\sim \mathcal{N}(\sigma^2, \sqrt{\mathrm{Var}\,[S^2]})$.

Estimación puntual de una proporción

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias i.i.d. con $X_1 \sim \text{Bernoulli}(p)$. Es decir, cada X_i marca la ocurrencia o no de un dado evento en una serie de experimentos independientes. Un estimador de p es la frecuencia relativa de ocurrencia:

$$\hat{p} = \hat{p}(X_1, \cdots, X_n) = F = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Algunos hechos:

1. Es un estimador insesgado.
2. Por la ley de los grandes números, es un estimador consistente.
3. El error cuadrático medio es
$$\text{ECM}(\hat{p}) = \text{Var}\left[\hat{p}\right] = \frac{p(1-p)}{n}.$$
4. $n\hat{p} \sim \text{Binomial}(n, p)$.
5. Por el Teorema Central del Límite, $\hat{p}$ es asintóticamente normal. Es decir que, para n grande, la distribución del estimador $\hat{p}$ se puede aproximar por la de una variable aleatoria $\sim \mathcal{N}(p, \sqrt{p(1-p)/n})$.

Estimación de intervalos

En ciertas ocasiones, en vez de dar una estimación puntual de un parámetro θ, se realiza la estimación de un *intervalo de confianza* $(\hat{\theta}_l^\alpha, \hat{\theta}_u^\alpha)$ tal que

$$\text{P}\left(\hat{\theta}_l^\alpha(X_1, \cdots, X_n) < \theta < \hat{\theta}_l^\alpha(X_1, \cdots, X_n)\right) = 1 - \alpha,$$

donde $1-\alpha$ (cercano a 1) es el *nivel de confianza* del estimador.

La interpretación frecuentista de un intervalo de confianza es la siguiente. Supongamos que extraemos un número grande N de muestras de tamaño n, es decir, tenemos $M_{i\,i=1}^{n}$, donde $M_i = \{x_1^{(i)}, x_2^{(i)}, \cdots x_n^{(i)}\}$ es la i-ésima muestra. Con cada muestra, se generará un nuevo intervalo $I^{(i)} = (\hat{\theta}_l^{\alpha,(i)}, \hat{\theta}_u^{\alpha,(i)})$. Aproximadamente $(1-\alpha)N$ intervalos contendrá al verdadero valor del parámetro θ.

Si se fija $\hat{\theta}_l^\alpha = -\infty$, se dice que se trata de un *intervalo unilateral a derecha*. Si, por el contrario, se hace $\hat{\theta}_u^\alpha = +\infty$, se trata de un *intervalo unilateral a izquierda*. Si ambos $|\hat{\theta}_l^\alpha|, |\hat{\theta}_l^\alpha| < \infty$, entonces es un *intervalo bilateral*.

Intervalo para la media con varianza conocida

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias i.i.d. normales con media $\mu = \mathrm{E}[X_1]$ **desconocida** y varianza $\sigma^2 = \mathrm{Var}[X_1]$ **conocida**. Luego, se tienen los siguientes intervalos con un nivel de confianza de $100\gamma\,\%$:

1. Unilateral a derecha:
$$I_r^\gamma = \left(-\infty, \bar{X} + z_\gamma \frac{\sigma}{\sqrt{n}}\right),$$
donde $z_p = \Phi^{-1}(p)$.

2. Unilateral a izquierda:
$$I_l^\gamma = \left(\bar{X} - z_\gamma \frac{\sigma}{\sqrt{n}}, +\infty\right).$$
Recordemos que $z_\alpha = -z_{1-\alpha}$.

3. Bilateral:
$$I_{lr}^\gamma = \left(\bar{X} - z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}\right).$$

Si las variables aleatorias no son normales, pero n es grande, el Teorema Central del Límite nos permite utilizar estos mismos intervalos de confianza.

Sean o no normales las variables aleatorias, si σ^2 es desconocido, para n muy grande, se pueden utilizar estos mismos intervalos de confianza reemplazando σ desconocido por el estimador S.

Intervalo para la proporción con muestras grandes

Sean p una probabilidad y $\hat{p}$ un estimador resultante de n experimentos independientes. Si n es grande, se tienen los siguientes intervalos con un nivel de confianza de $100\gamma\,\%$:

1. Unilateral a derecha:
$$I_r^\gamma = \left[0, \hat{p} + z_\gamma \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$
2. Unilateral a izquierda:
$$I_l^\gamma = \left(\hat{p} - z_\gamma \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, 1\right].$$
3. Bilateral:
$$I_{lr}^\gamma = \left(\hat{p} - z_{\frac{1+\gamma}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{1+\gamma}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

Estos intervalos de confianza son muy parecidos a los de la media con varianza conocida. Esto es debido a que p es el valor medio de variables aleatorias i.i.d. con distribución Bernoulli, $\hat{p}$ es la media muestral y al hecho que podemos aplicar el Teorema Central del Límite. Sin embargo, hay algunas diferencias:

- Se sabe que $p \in [0, 1]$.
- Se desconoce la varianza real y se la reemplaza por el estimador $\hat{p}(1-\hat{p})$. Se puede mostrar que esta aproximación es buena para n grande.

Intervalo para la media de variables aleatorias normales

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias i.i.d. normales con media $\mu = \mathrm{E}[X_1]$ y varianza $\sigma^2 = \mathrm{Var}[X_1]$ **desconocidas**. Luego, se tienen los siguientes intervalos con un nivel de confianza de $100\gamma\,\%$:

1. Unilateral a derecha:
$$I_r^\gamma = \left(-\infty, \bar{X} + t_{n-1,\gamma} \frac{S}{\sqrt{n}}\right),$$
donde $t_{k,p}$ es el $100p$ percentil de una variable aleatoria con distribución t-Student con k grados de libertad.
2. Unilateral a izquierda:
$$I_l^\gamma = \left(\bar{X} - t_{n-1,\gamma} \frac{S}{\sqrt{n}}, +\infty\right).$$
Recordemos que $t_{n-1,\alpha} = -t_{n-1,1-\alpha}$.
3. Bilateral:
$$I_{lr}^\gamma = \left(\bar{X} - t_{n-1,\frac{1+\gamma}{2}} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\frac{1+\gamma}{2}} \frac{S}{\sqrt{n}}\right).$$

Dado que para k muy grande la distribución t de Student es muy parecida a la normal, para $n > 200$ no hace mucha diferencia si se utilizan los percentiles de una distribución Gaussiana.

2. Función de distribución de la variable aleatoria con distribución normal estándar

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

3. Fractiles de la distribución normal estándar

α	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.50	0.0000	0.0025	0.0050	0.0075	0.0100	0.0125	0.0150	0.0175	0.0201	0.0226
0.51	0.0251	0.0276	0.0301	0.0326	0.0351	0.0376	0.0401	0.0426	0.0451	0.0476
0.52	0.0502	0.0527	0.0552	0.0577	0.0602	0.0627	0.0652	0.0677	0.0702	0.0728
0.53	0.0753	0.0778	0.0803	0.0828	0.0853	0.0878	0.0904	0.0929	0.0954	0.0979
0.54	0.1004	0.1030	0.1055	0.1080	0.1105	0.1130	0.1156	0.1181	0.1206	0.1231
0.55	0.1257	0.1282	0.1307	0.1332	0.1358	0.1383	0.1408	0.1434	0.1459	0.1484
0.56	0.1510	0.1535	0.1560	0.1586	0.1611	0.1637	0.1662	0.1687	0.1713	0.1738
0.57	0.1764	0.1789	0.1815	0.1840	0.1866	0.1891	0.1917	0.1942	0.1968	0.1993
0.58	0.2019	0.2045	0.2070	0.2096	0.2121	0.2147	0.2173	0.2198	0.2224	0.2250
0.59	0.2275	0.2301	0.2327	0.2353	0.2378	0.2404	0.2430	0.2456	0.2482	0.2508
0.60	0.2533	0.2559	0.2585	0.2611	0.2637	0.2663	0.2689	0.2715	0.2741	0.2767
0.61	0.2793	0.2819	0.2845	0.2871	0.2898	0.2924	0.2950	0.2976	0.3002	0.3029
0.62	0.3055	0.3081	0.3107	0.3134	0.3160	0.3186	0.3213	0.3239	0.3266	0.3292
0.63	0.3319	0.3345	0.3372	0.3398	0.3425	0.3451	0.3478	0.3505	0.3531	0.3558
0.64	0.3585	0.3611	0.3638	0.3665	0.3692	0.3719	0.3745	0.3772	0.3799	0.3826
0.65	0.3853	0.3880	0.3907	0.3934	0.3961	0.3989	0.4016	0.4043	0.4070	0.4097
0.66	0.4125	0.4152	0.4179	0.4207	0.4234	0.4261	0.4289	0.4316	0.4344	0.4372
0.67	0.4399	0.4427	0.4454	0.4482	0.4510	0.4538	0.4565	0.4593	0.4621	0.4649
0.68	0.4677	0.4705	0.4733	0.4761	0.4789	0.4817	0.4845	0.4874	0.4902	0.4930
0.69	0.4958	0.4987	0.5015	0.5044	0.5072	0.5101	0.5129	0.5158	0.5187	0.5215
0.70	0.5244	0.5273	0.5302	0.5330	0.5359	0.5388	0.5417	0.5446	0.5476	0.5505
0.71	0.5534	0.5563	0.5592	0.5622	0.5651	0.5681	0.5710	0.5740	0.5769	0.5799
0.72	0.5828	0.5858	0.5888	0.5918	0.5948	0.5978	0.6008	0.6038	0.6068	0.6098
0.73	0.6128	0.6158	0.6189	0.6219	0.6250	0.6280	0.6311	0.6341	0.6372	0.6403
0.74	0.6433	0.6464	0.6495	0.6526	0.6557	0.6588	0.6620	0.6651	0.6682	0.6713
0.75	0.6745	0.6776	0.6808	0.6840	0.6871	0.6903	0.6935	0.6967	0.6999	0.7031
0.76	0.7063	0.7095	0.7128	0.7160	0.7192	0.7225	0.7257	0.7290	0.7323	0.7356
0.77	0.7388	0.7421	0.7454	0.7488	0.7521	0.7554	0.7588	0.7621	0.7655	0.7688
0.78	0.7722	0.7756	0.7790	0.7824	0.7858	0.7892	0.7926	0.7961	0.7995	0.8030
0.79	0.8064	0.8099	0.8134	0.8169	0.8204	0.8239	0.8274	0.8310	0.8345	0.8381
0.80	0.8416	0.8452	0.8488	0.8524	0.8560	0.8596	0.8632	0.8669	0.8706	0.8742
0.81	0.8779	0.8816	0.8853	0.8890	0.8927	0.8965	0.9002	0.9040	0.9078	0.9116
0.82	0.9154	0.9192	0.9230	0.9269	0.9307	0.9346	0.9385	0.9424	0.9463	0.9502
0.83	0.9542	0.9581	0.9621	0.9661	0.9701	0.9741	0.9782	0.9822	0.9863	0.9904
0.84	0.9945	0.9986	1.0027	1.0069	1.0110	1.0152	1.0194	1.0237	1.0279	1.0322
0.85	1.0364	1.0407	1.0451	1.0494	1.0537	1.0581	1.0625	1.0669	1.0714	1.0758
0.86	1.0803	1.0848	1.0893	1.0939	1.0985	1.1031	1.1077	1.1123	1.1170	1.1217
0.87	1.1264	1.1311	1.1359	1.1407	1.1455	1.1503	1.1552	1.1601	1.1650	1.1700
0.88	1.1750	1.1800	1.1850	1.1901	1.1952	1.2004	1.2055	1.2107	1.2160	1.2212
0.89	1.2265	1.2319	1.2372	1.2426	1.2481	1.2536	1.2591	1.2646	1.2702	1.2759
0.90	1.2816	1.2873	1.2930	1.2988	1.3047	1.3106	1.3165	1.3225	1.3285	1.3346
0.91	1.3408	1.3469	1.3532	1.3595	1.3658	1.3722	1.3787	1.3852	1.3917	1.3984
0.92	1.4051	1.4118	1.4187	1.4255	1.4325	1.4395	1.4466	1.4538	1.4611	1.4684
0.93	1.4758	1.4833	1.4909	1.4985	1.5063	1.5141	1.5220	1.5301	1.5382	1.5464
0.94	1.5548	1.5632	1.5718	1.5805	1.5893	1.5982	1.6072	1.6164	1.6258	1.6352
0.95	1.6449	1.6546	1.6646	1.6747	1.6849	1.6954	1.7060	1.7169	1.7279	1.7392
0.96	1.7507	1.7624	1.7744	1.7866	1.7991	1.8119	1.8250	1.8384	1.8522	1.8663
0.97	1.8808	1.8957	1.9110	1.9268	1.9431	1.9600	1.9774	1.9954	2.0141	2.0335
0.98	2.0537	2.0748	2.0969	2.1201	2.1444	2.1701	2.1973	2.2262	2.2571	2.2904
0.99	2.3263	2.3656	2.4089	2.4573	2.5121	2.5758	2.6521	2.7478	2.8782	3.0902

4. Fractiles de la distribución t-Student

GDL	0.800	0.900	0.950	0.975	0.990	0.995
1	1.3764	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467
16	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.8591	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.8583	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.8575	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.8569	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.8562	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.8557	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.8551	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.8546	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.8542	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.8538	1.3104	1.6973	2.0423	2.4573	2.7500
31	0.8534	1.3095	1.6955	2.0395	2.4528	2.7440
32	0.8530	1.3086	1.6939	2.0369	2.4487	2.7385
33	0.8526	1.3077	1.6924	2.0345	2.4448	2.7333
34	0.8523	1.3070	1.6909	2.0322	2.4411	2.7284
35	0.8520	1.3062	1.6896	2.0301	2.4377	2.7238
36	0.8517	1.3055	1.6883	2.0281	2.4345	2.7195
37	0.8514	1.3049	1.6871	2.0262	2.4314	2.7154
38	0.8512	1.3042	1.6860	2.0244	2.4286	2.7116
39	0.8509	1.3036	1.6849	2.0227	2.4258	2.7079
40	0.8507	1.3031	1.6839	2.0211	2.4233	2.7045
41	0.8505	1.3025	1.6829	2.0195	2.4208	2.7012
42	0.8503	1.3020	1.6820	2.0181	2.4185	2.6981
43	0.8501	1.3016	1.6811	2.0167	2.4163	2.6951
44	0.8499	1.3011	1.6802	2.0154	2.4141	2.6923
45	0.8497	1.3006	1.6794	2.0141	2.4121	2.6896
46	0.8495	1.3002	1.6787	2.0129	2.4102	2.6870
47	0.8493	1.2998	1.6779	2.0117	2.4083	2.6846
48	0.8492	1.2994	1.6772	2.0106	2.4066	2.6822
49	0.8490	1.2991	1.6766	2.0096	2.4049	2.6800
50	0.8489	1.2987	1.6759	2.0086	2.4033	2.6778

GDL	0.800	0.900	0.950	0.975	0.990	0.995
51	0.8487	1.2984	1.6753	2.0076	2.4017	2.6757
52	0.8486	1.2980	1.6747	2.0066	2.4002	2.6737
53	0.8485	1.2977	1.6741	2.0057	2.3988	2.6718
54	0.8483	1.2974	1.6736	2.0049	2.3974	2.6700
55	0.8482	1.2971	1.6730	2.0040	2.3961	2.6682
56	0.8481	1.2969	1.6725	2.0032	2.3948	2.6665
57	0.8480	1.2966	1.6720	2.0025	2.3936	2.6649
58	0.8479	1.2963	1.6716	2.0017	2.3924	2.6633
59	0.8478	1.2961	1.6711	2.0010	2.3912	2.6618
60	0.8477	1.2958	1.6706	2.0003	2.3901	2.6603
61	0.8476	1.2956	1.6702	1.9996	2.3890	2.6589
62	0.8475	1.2954	1.6698	1.9990	2.3880	2.6575
63	0.8474	1.2951	1.6694	1.9983	2.3870	2.6561
64	0.8473	1.2949	1.6690	1.9977	2.3860	2.6549
65	0.8472	1.2947	1.6686	1.9971	2.3851	2.6536
66	0.8471	1.2945	1.6683	1.9966	2.3842	2.6524
67	0.8470	1.2943	1.6679	1.9960	2.3833	2.6512
68	0.8469	1.2941	1.6676	1.9955	2.3824	2.6501
69	0.8469	1.2939	1.6672	1.9949	2.3816	2.6490
70	0.8468	1.2938	1.6669	1.9944	2.3808	2.6479
71	0.8467	1.2936	1.6666	1.9939	2.3800	2.6469
72	0.8466	1.2934	1.6663	1.9935	2.3793	2.6459
73	0.8466	1.2933	1.6660	1.9930	2.3785	2.6449
74	0.8465	1.2931	1.6657	1.9925	2.3778	2.6439
75	0.8464	1.2929	1.6654	1.9921	2.3771	2.6430
76	0.8464	1.2928	1.6652	1.9917	2.3764	2.6421
77	0.8463	1.2926	1.6649	1.9913	2.3758	2.6412
78	0.8463	1.2925	1.6646	1.9908	2.3751	2.6403
79	0.8462	1.2924	1.6644	1.9905	2.3745	2.6395
80	0.8461	1.2922	1.6641	1.9901	2.3739	2.6387
81	0.8461	1.2921	1.6639	1.9897	2.3733	2.6379
82	0.8460	1.2920	1.6636	1.9893	2.3727	2.6371
83	0.8460	1.2918	1.6634	1.9890	2.3721	2.6364
84	0.8459	1.2917	1.6632	1.9886	2.3716	2.6356
85	0.8459	1.2916	1.6630	1.9883	2.3710	2.6349
86	0.8458	1.2915	1.6628	1.9879	2.3705	2.6342
87	0.8458	1.2914	1.6626	1.9876	2.3700	2.6335
88	0.8457	1.2912	1.6624	1.9873	2.3695	2.6329
89	0.8457	1.2911	1.6622	1.9870	2.3690	2.6322
90	0.8456	1.2910	1.6620	1.9867	2.3685	2.6316
91	0.8456	1.2909	1.6618	1.9864	2.3680	2.6309
92	0.8455	1.2908	1.6616	1.9861	2.3676	2.6303
93	0.8455	1.2907	1.6614	1.9858	2.3671	2.6297
94	0.8455	1.2906	1.6612	1.9855	2.3667	2.6291
95	0.8454	1.2905	1.6611	1.9853	2.3662	2.6286
96	0.8454	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.8453	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.8453	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.8453	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.8452	1.2901	1.6602	1.9840	2.3642	2.6259

5. Guía de ejercicios

1. Sean $X_1, X_2, \cdots, X_n$ variables aleatorias i.i.d. con distribución conocida. Encuentre un estimador del parámetro α desconocidos usando el criterio de máxima verosimilitud cuando:

   a) $X \sim \text{Laplace}(0, \alpha)$, es decir:
   $$\mathrm{f}_X(x) = \frac{1}{2\alpha} e^{-\frac{|x|}{\alpha}}. \tag{16}$$

   b) $X \sim \text{Uniforme}(3, \alpha)$.

   c) $X \sim \text{Uniforme}(\alpha, 2)$.

2. Repita el ejercicio anterior, pero usando el método de los momentos.

3. Siempre que sea posible, encuentre un intervalo de confianza con un nivel de confianza del 95 % para los estimadores del ejercicio anterior. Para ello, haga uso del Teorema de Central del Límite, asumiendo que n es suficientemente grande.

4. Determine si los estimadores de los primeros dos ejercicios son insesgados.

5. Sean $X_i \sim \text{Bernoulli}(P)$, $i = 1, 2, \cdots, n$, variables aleatorias independientes. La información *a priori* sobre el valor de P puede representarse asumiendo que $P \sim \text{Beta}(\alpha, \beta)$, es decir,
   $$f_P(p) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{(\beta-1)} & p \in (0,1), \\ 0 & p \notin (0,1), \end{cases} \tag{17}$$
   con $\alpha, \beta > 0$. Determine un estimador de P usando el máximo *a posteriori* (MAP). *Ayuda: aplique logaritmos a la función de densidad a posteriori.*

6. Sea X una variable aleatoria con media μ y varianza σ^2. Considere dos muestras aleatorias independientes de tamaños n_1 y n_2, con medias muestrales $\overline{X}_1$ y $\overline{X}_2$. Se define el estadístico
   $$Y = a\,\overline{X}_1 + b\,\overline{X}_2$$
   donde $0 < a < 1$.

   a) Determine b en función de a para que Y sea un estimador insesgado de μ.

   b) Utilizando la relación entre b y a hallada en el ítem anterior, encuentre el valor de a que minimiza la varianza de Y si $n_2 = 2\,n_1$.

   c) Suponga que $n_2 = n_1 = n$ y $b = 1 - a$. Demuestre que $\text{Var}(Y) < \text{Var}(\overline{X}_1) = \text{Var}(\overline{X}_2)$ si $a \in (0, 1)$ y que el mínimo de $\text{Var}(Y)$ es $\sigma^2/2\,n$ para $a = \frac{1}{2}$.

7. El tiempo de duración de una pieza de un equipo puede considerarse una variable aleatoria normal con desvío estándar de 4 horas y una media μ que se desea estimar. Una muestra aleatoria de 100 piezas que fueron probadas produjo una media muestral de 501.2 horas de duración. Obtener un intervalo de confianza para la media μ con un nivel de confianza de: a) 0.95 (95 %), b) 0.99 (99 %).

8. En una fábrica de materiales eléctricos se desea estimar el peso promedio del último lote de rollos de alambre de cobre salido de producción. Para ello se eligió al azar una muestra de 20 rollos que arrojó un promedio de 38 kg por rollo. Se conoce además, de registros históricos, que el desvío estándar del peso de un rollo es de 4.2 kg.

a) Estimar el peso medio de los rollos con un intervalo de confinza del 95 %.

b) ¿Cuántos rollos más habría que pesar para poder obtener una estimación del peso medio de un rollo con un error de muestreo de 1 kg?

9. 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 aquí. Video con el ejercicio explicado.

10. Se conectan 25 lámparas de luz infrarroja en un invernadero, de tal manera que si falla una lámpara, otra se enciende inmediatamente (se enciende solamente una lámpara a la vez). Las lámparas funcionan en forma independiente y el tiempo de duración puede suponerse una variable aleatoria normal con una vida media de 50 horas y una desviación estándar de 4 horas. Si T es el tiempo total de operación de las 25 lámparas de luz infrarroja,

    a) hallar la probabilidad de que dicho tiempo T exceda 1300 horas. Explicar el procedimiento utilizado detallando la variable aleatoria con la que se calcula la probabilidad solicitada.

    b) La variable aleatoria $T/25$ puede considerarse como una estimación puntual de la media de la duración de una lámpara en base a una muestra de tamaño 25. Obtener un intervalo de confianza del 95 % para la vida media de una lámpara si el tiempo total de operación fue 1251 horas.

11. Se sabe que determinaciones hechas sobre la densidad de cierto producto químico se distribuyen normalmente alrededor de la media poblacional desconocida. Esa densidad puede considerarse una variable aleatoria con distribución normal con un desvío estándar de 0.005 g/cm^3. Se desea estimar la densidad media con un intervalo de confianza del 95 % y con un error menor que 0.002 g/cm^3. ¿Cuál deberá ser el tamaño mínimo de la muestra?

12. La media muestral y la dispersión muestral de la carga máxima soportada por un cable en una muestra de tamaño 60 son 11.09 y 0.73 Tn respectivamente. Hallar los límites de confianza del 95 % para la carga máxima media.

13. 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 aquí. Video con el ejercicio explicado.

14. Los contenidos de 7 recipientes similares para ácido sulfúrico son: 9.8, 10.2, 10.4, 9.8, 10.0, 10.2 y 9.6 litros. Obtener intervalos de confianza del 95 % para la media del contenido de los recipientes de esa clase asumiendo que el contenido de ácido en los recipientes es una variable aleatoria normal.

    Resolución aquí.

15. En una industria textil hay un lote de tambores de 100 litros de capacidad que contienen un suavizante textil, que se han usado parcialmente, por lo que se desea estimar el contenido medio de los mismos. A tal efecto se tomó una muestra aleatoria de 15 tambores, se midieron sus contenidos y se obtuvo una media muestral de 63 litros con un desvío estándar muestral de 12.5 litros. Determinar los límites de confianza del 90 % para el contenido medio de los tambores de la población. Suponga que el contenido de suavizante en los recipientes tiene distribución normal.

16. Se hicieron pruebas de inmersión en solución corrosiva de una aleación determinándose el porcentaje de pérdida en resistencia a la tracción (que se supone una variable aleatoria con distribución normal). Los resultados obtenidos fueron: 6.4, 4.6, 4.6, 6.4, 3.2, 5.2, 6.5, 4.9, 4.3, 5.6, 3.7 y 4.6. Encontrar un intervalo de confianza del 95 % para el porcentaje esperado de pérdida en resistencia a la tracción.

17. De lo producido en cierto proceso se tomó una muestra aleatoria de 6 artículos y se midió una longitud característica (que se supone una variable aleatoria con distribución normal), obteniéndose los siguientes valores (en cm): 2.9, 2.92, 2.9, 2.84, 2.79 y 2.8. Determinar un intervalo de confianza del 95 % para la longitud característica media.

18. Las ventas de una revista semanal han sido las siguientes (en miles de ejemplares) en las últimas 4 semanas: 15.4, 18.5, 16.3, y 19.2. Calcular los límites de confianza del 95 % para el promedio de las ventas semanales (se supone que el volumen semanal de ventas es una variable aleatoria con distribución normal).

19. Una máquina produce varillas de metal usadas en el sistema de suspensión de un auto. Se toma una muestra aleatoria de 15 varillas y se miden los diámetros. Los datos obtenidos (en mm) se detallan más abajo. Asumiendo que el diámetro de las varillas está distribuido normalmente obtener un intervalo de confianza del 95 % para el diámetro medio de la varilla.

Diam. (mm)	8.24	8.23	8.20	8.21	8.20	8.28	8.23	8.26
	8.24	8.24	8.25	8.19	8.25	8.26	8.23	

20. De un proceso productivo de una pieza seriada se tomó una muestra de 300 unidades en la que se encontraron 18 defectuosas.

    a) Calcular los límites de confianza del 90 % para el porcentaje defectuoso del proceso.

    b) Calcular el tamaño de muestra adicional para tener un intervalo del mismo nivel de confianza pero de semiamplitud 0.01 (o sea del 1 % de semiamplitud).

    c) Con la muestra dada de 300 unidades calcular el porcentaje defectuoso máximo del proceso con 90 % de confianza (o sea un porcentaje tal que la probabilidad de que el verdadero porcentaje defectuoso lo exceda sea 0.1).

    Resolución aquí.

21. Una muestra de 100 votantes elegidos al azar indicó que 55 % de ellos estaba a favor de un candidato. Obtener un intervalo de confianza del 95 % para la proporción de votantes que en el total de la población votan por este candidato.

22. 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). Se ha registrado el rating del programa *Bailando por bailar* 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 del programa.

    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 ± 1 %.

    Resolución aquí. Video con el ejercicio explicado.

23. En un diario de gran circulación de esta ciudad se publicó el resultado de una encuesta en la que se indicaban, para una muestra de tamaño n, varias *proporciones* de preferencia por varios equipos de fútbol. En ese reporte figuraba una ficha técnica en la que se indicaba lo siguiente:

    - Tamaño de la muestra $n = 9300$
    - Nivel de confianza 95 %
    - Error máximo: 1 % (semiamplitud máxima de un intervalo de confianza).

    Explicite claramente un planteo en el cuál dadas dos de las tres cantidades indicadas se obtenga la tercera. Verificar, por ejemplo, que para ese nivel de confianza y tamaño muestral el error máximo es aproximadamente 0.01 (1 %).
    En esa misma encuesta se indicaba que *los hinchas de Boca no son la mitad mas uno* dado que el 41 % (0.403) de la muestra correspondió a hinchas de Boca. Explique el significado de esta afirmación a partir del intervalo de confianza hallado.

24. Considere una cierta población de tamaño finito N en la que se encuentra definido un atributo A en sus integrantes. Un parámetro poblacional de interés es la proporción p de integrantes de la población que tienen el atributo. Un estimador insesgado de p es $\hat{p} = \frac{X}{n}$, donde X es el número de miembros de una muestra de tamaño n, extraída de esta población de tamaño N, que tienen el atributo bajo análisis. Si la muestra se extrae sin reposición y vale la aproximación normal de la distribución hipergeométrica, entonces se puede demostrar que la semiamplitud de un intervalo de confianza del $100\,(1-\alpha)$ % para p, viene dada por

$$E = z_{1-\frac{\alpha}{2}} \sqrt{\hat{p}\,(1-\hat{p})} \sqrt{\frac{N-n}{N\,n}}.$$

    a) Verificar que si $N \gg n$ entonces se obtiene el resultado para poblaciones infinitas. Analizar el caso en que $N = n$ (la muestra en este caso es una sola posible).

    b) Sea $B = \frac{E}{z_{1-\frac{\alpha}{2}} \sqrt{\hat{p}\,(1-\hat{p})}}$. Probar que el tamaño de la muestra necesario para estimar p, con un nivel de confianza $100\,(1-\alpha)$ %, viene dado por

$$n = \frac{1}{B^2 + \frac{1}{N}}.$$

    En este caso el valor de $\hat{p}$ es el que por ejemplo corresponda a una estimación *a priori* de p (por ejemplo el resultado de una muestra piloto).

    c) Analizar n como función de N si por ejemplo $E = 0.01$ (1 % es la semiamplitud del intervalo de confianza para p), $z = 2$ (o $(1-\alpha) \approx 0.955$) y $\hat{p} = 0.5$. Usar para N valores del la forma 10^k con k tomando los valores $2, 3, \ldots, 7$. Extraiga alguna conclusión.

25. Usted ha sido contratado en un criadero de chanchos donde se ha implementado una nueva dieta para los animales. El dueño le informa que, al final del proceso de crianza, el peso de un chancho adulto tomado al azar puede considerarse una variable aleatoria normal con desvío $\sigma = 5$ kg.

    a) El dueño le pide que determine el peso medio de los chanchos adultos al final de la crianza con un nivel de confianza del 99 % y un margen de error de 1 kg. ¿Cuál es el número mínimo de chanchos que debería pesar? Justifique su respuesta.

    b) Al final, usted pesa 25 chanchos tomados al azar y el peso promedio le da $\overline{x} = 200$ kg. Determine el intervalo de confianza con un nivel de 99 % para el peso medio.

    c) Usted no confía en el dueño del criadero y descree que el desvío poblacional sea igual a 5 kg. Por ello, calcula el desvío muestral a partir de los datos de los 25 chanchos, obteniendo $s = 4.4692$. Determine el intervalo de confianza con un nivel de 95 % para el peso medio.

Apéndice: Una simulación en Octave.
Optativo y recomendable para fijar conceptos.

Se sugiere ejecutar los códigos de *Octave* donde se repite N veces la operación de extraer una muestra de tamaño n y en esa muestra se evalúa un estimador. De esta manera se tienen N realizaciones del estadístico y se puede realizar una estadística de ese conjunto de valores.

a) La secuencia que sigue a continuación se refiere al problema de la estimación de una proporción poblacional a partir de una muestra de tamaño n:

```
p = rand; n = 100; N = 1000; X = (rand(n,N)<p);
M = mean(X);
hist(M)
[ mean(M) p std(M) sqrt(p*(1-p)/n)]
```

Breve explicación: Se genera el valor de `p` con un número aleatorio, luego se obtiene la matriz `X` de `n` filas y `N` columnas con 1 o 0 dependiendo que la condición entre paréntesis sea verdadera o falsa, respectivamente. Entonces la proporción de 1 (en el largo plazo o para n grande) es `p`. El vector `M` almacena `N` valores de la media (la proporción muestral $\hat{p}$) de una muestra de tamaño `n` de cada columna de `X`. El comando `hist` genera el histograma de los N valores del vector `M`. Luego se imprimen cuatro números: el promedio de los `n` valores de `M`, el valor de `p`, el desvío estándar del vector `M`, y el desvío estándar del estimador $\hat{p}$). Una ejecución de este código devolvió estos cuatro números 0.15451, 0.15377, 0.036667 y 0.036073.

b) El código que sigue se refiere al problema de estimar la media de una variable aleatoria normal a partir de una muestra de tamaño n:

```
mu = 100*rand;sigma = 0.1*mu;
n = 100; N = 1000; X=mu+sigma*randn(n,N);
M = mean(X);
hist(M)
mean(M)
[ mean(M) mu std(M) sigma/sqrt(n)]
```

Breve explicación: Se generan los valores de μ y σ, el primero como 100 veces un número aleatorio, y σ como el 10 % de μ. Luego se obtiene la matriz `X` de `n` filas y `N` columnas con números aleatorios provenientes de una variable aleatoria normal con la media y el desvío generados. El vector `M` almacena `N` valores de la media (la media muestral $\hat{\mu} = \overline{x}$) de una muestra de tamaño `n` de cada columna de `X`. El comando `hist` genera el histograma de los N valores del vector `M`. Luego se imprimen cuatro números: el promedio de los `n` valores de `M`, el valor de μ, el desvío estándar del vector `M`, y el desvío estándar del estimador $\hat{\mu}$. Una ejecución de este código devolvió estos cuatro números 71.611, 71.579, 0.71086 y 0.71579.

c) A continuación se generan N intervalos de confianza para la estimación de la media μ de una variable aleatoria normal a partir de una muestra de tamaño n, y se mide la frecuencia de intervalos que tienen la propiedad de contener el valor de μ:

```
mu = 100*rand; sigma = 0.1*mu;
n = 100; N = 1000; X=mu + sigma*randn(n,N);
M = mean(X);
A = M - 1.6449*sigma/sqrt(n); B = M + 1.6449*sigma/sqrt(n);
p = mean((A < mu).*(mu < B))
```

Breve explicación: Se generan los valores de μ y σ, el primero como 100 veces un número aleatorio, y σ como el 10 % de μ. Luego se obtiene la matriz `X` de `n` filas y `N` columnas con números aleatorios provenientes de una variable aleatoria normal con la media y el desvío generados. El

vector `M` almacena `N` valores de la media (la media muestral $\hat{\mu} = \overline{x}$) de una muestra de tamaño `n` de cada columna de `X`. Los vectores `A` y `B` contienen los `N` extremos izquierdo y derecho de intervalos de confianza del 90 % con centro en cada uno de los valores generados de la media muestral obtenida en base a muestras de tamaño `n`. El valor de `p` es la proporción de esos intervalos que contienen el valor de μ. Una ejecución de este código devolvió `p = 0.90900`.

6. Respuestas

7. a) (500.42, 501.98) b) (500.17, 502.23).

8. a) 38 ± 1.84 b) 48 rollos más.

9. a) 246 ± 7.8 b) 15 envases más.

10. a) 0.6 % (0.0062) b) (48.47, 51.61).

11. $n \geq 25$.

12. (10.905, 11.275).

13. (0.716, 0.763)

14. (9.74, 10.26)

15. (57.3, 68.7).

16. (4.323, 5.677).

17. Para la media: (2.8, 2.92)

18. (14.5, 20.2) y (1.03, 44.76)

19. (8.22, 8.25).

20. a) 0.060 ± 0.023 b) 1226 c) 0.0776 (7.76 %).

21. Para el 95 % el intervalo es (0.452, 0.648).

22. a) (0.22, 0.28) b) $4473\,C$.

24. a) 166 b) $(197.4242, 202.5758)$ c) $(198.1552, 201.8448)$.

7. Ejercicios resueltos

Ejercicio 9

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:

$$X_i = \text{peso de la } i\text{-ésima lata de café} \tag{18}$$

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

Parámetros y estimadores

El único parámetro desconocido es $\mu = \mathrm{E}[X_i]$ 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.

Ítem a

Pre-muestra

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

$$\mathrm{P}\left(a \leq \mu \leq b\right) = 0.9 \tag{19}$$

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 $\mathrm{E}[X_i] = \mu$ 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:

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n} \tag{20}$$

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

$$\mathrm{E}[\overline{X}_n] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{\sum_{i=1}^{n} \overbrace{\mathrm{E}[X_i]}^{\mu}}{n} = \frac{\mu \cdot \sum_{i=1}^{n} 1}{n} = \mu \tag{21}$$

$$\mathrm{Var}\left[\overline{X}_n\right] = \mathrm{Var}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n^2} \cdot \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \overset{IND}{=} \frac{1}{n^2} \cdot \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \tag{22}$$

$$\Rightarrow \sigma[\overline{X}_n] = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \tag{23}$$

Por lo tanto,

$$\overline{X}_n \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N\left(0; 1\right) \tag{24}$$

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:

$$\mathrm{P}\left(k_1 \leq \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq k_2\right) = 0.9 \tag{25}$$

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 $k_1 = -k$:

$$\begin{aligned}\mathrm{P}\left(-k \leq \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq k\right) = 0.9 &\Rightarrow \Phi(k) - \Phi(-k) = 0.9 \\ &\Rightarrow \Phi(k) - (1 - \Phi(k)) = 0.9 \\ &\Rightarrow 2\Phi(k) - 1 = 0.9 \\ &\Rightarrow \Phi(k) = \frac{0.9+1}{2} \Rightarrow k = z_{0.95}\end{aligned} \tag{26}$$

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

$$\mathrm{P}\left(-z_{0.95} \leq \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{0.95}\right) = 0.9 \Rightarrow \mathrm{P}\left(-z_{0.95} \cdot \frac{\sigma}{\sqrt{n}} \leq \overline{X}_n - \mu \leq z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.9 \tag{27}$$

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

$$\begin{aligned}\mathrm{P}\left(z_{0.95} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu - \overline{X}_n \geq -z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.9 \Rightarrow \\ \Rightarrow \mathrm{P}\left(\overline{X}_n + z_{0.95} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu \geq \overline{X}_n - z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.9\end{aligned} \tag{28}$$

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

$$IC_{90\,\%}(\mu) = [a(X_1, \cdots, X_{10}, \sigma); b(X_1, \cdots, X_{10}, \sigma)] = \left[\overline{X}_n - z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X}_n + z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{29}$$

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:

$$\underbrace{\mathrm{P}\left(\overline{X}_n + z_{0.95} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu \geq \overline{X}_n - z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.9 \Rightarrow}_{\textbf{Bien}} \tag{30}$$

$$\underbrace{\mathrm{P}\left(246 + z_{0.95} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu \geq 246 - z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.9}_{\textbf{Mal}} \tag{31}$$

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:

$$\begin{aligned} IC_{90\,\%}(\mu) &= \left[\overline{x}_{10} - z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X}_n + z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right] \\ &\approx \left[246 - 1.644854 \cdot \frac{15}{\sqrt{10}}; 246 + 1.644854 \cdot \frac{15}{\sqrt{10}}\right] = [238.1978; 253.8022] \end{aligned} \tag{32}$$

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

$$IC_{90\,\%}(\mu) = \left[\overline{X}_n - z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X}_n + z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{33}$$

  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.

Ítem 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:

$$IC_{90\,\%}(\mu) = \left[\overline{X}_n - z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}; \overline{X}_n + z_{0.95} \cdot \frac{\sigma}{\sqrt{n}}\right] \tag{34}$$

el error de muestreo E se obtiene del siguiente modo:

$$E = z_{0.95} \cdot \frac{\sigma}{\sqrt{n}} \tag{35}$$

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:

$$E = z_{0.95} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} = z_{0.95} \cdot \frac{\sigma}{E} \Rightarrow n = z_{0.95}^2 \cdot \frac{\sigma^2}{E^2} = 1.644854^2 \frac{15^2}{5^2} \approx 24.3499 \tag{36}$$

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 $E \leq 5$:

$$5 \geq E = z_{0.95} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} \geq z_{0.95} \cdot \frac{\sigma}{5} \Rightarrow n = z_{0.95}^2 \cdot \frac{\sigma^2}{5^2} = 1.644854^2 \frac{15^2}{5^2} \approx 24.3499 \Rightarrow n \geq 25 \tag{37}$$

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. (38)

Ejercicio 13

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

(39)

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

Resolución

Variables aleatorias

Consideramos las siguientes variables aleatorias:

$$X_i = \text{contenido de silicio en la } i\text{-ésima colada de hierro.} \tag{40}$$

Parámetros y estimadores

En este caso, tanto la media **poblacional** $\mathrm{E}[X_i] = \mu$ 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:

$$IC_{95\,\%}(\mu) = \left[\overline{X}_n - z_{0.975} \cdot \frac{s}{\sqrt{n}}, \overline{X}_n + z_{0.975} \cdot \frac{s}{\sqrt{n}}\right] \tag{41}$$

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:

$$\overline{X}_n^{Ag} = \frac{\sum_{j=1}^{L} X_j \cdot f_j}{n} \tag{42}$$

$$s^{Ag} = \sqrt{\frac{\sum_{j=1}^{L}(X_j - \overline{X}_n^{Ag})^2 \cdot f_j}{n-1}} \tag{43}$$

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:

$$IC_{95\,\%}(\mu) = \left[\overline{X}_n^{Ag} - z_{0.975} \cdot \frac{s^{Ag}}{\sqrt{n}}, \overline{X}_n^{Ag} + z_{0.975} \cdot \frac{s^{Ag}}{\sqrt{n}}\right] \tag{44}$$

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

(45)

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

(46)

Por lo tanto, como $z_{0.975} = 1.9599$:

$$\begin{aligned} IC_{90\,\%}(\mu) &= \left[\overline{x}_{150}^{Ag} - z_{0.975} \cdot \frac{s^{Ag}}{\sqrt{n}}; \overline{x}_{150}^{Ag} + z_{0.975} \cdot \frac{s^{Ag}}{\sqrt{n}}\right] \\ &= \left[0.7396 - 1.9599 \cdot \frac{0.1485}{\sqrt{150}}; 0.7396 + 1.9599 \cdot \frac{0.1485}{\sqrt{150}}\right] \\ &= [0.7158; 0.7633] \end{aligned} \tag{47}$$

Ejercicio 14

Los contenidos de 7 recipientes similares para ácido sulfúrico son: 9.8, 10.2, 10.4, 9.8, 10.0, 10.2 y 9.6 litros. Obtener intervalos de confianza del 95 % para la media del contenido de los recipientes de esa clase asumiendo que el contenido de ácido en los recipientes es una variable aleatoria normal.

Resolución

Podemos tabular los datos como sigue:

k	x_k	x_k^2
1	9.8	96.04
2	10.2	104.04
3	10.4	108.16
4	9.8	96.04
5	10.0	100.00
6	10.2	104.04
7	9.6	92.16
Suma	70.0	700.48

A partir de esta tabla, podemos calcular

$$\bar{x} = \frac{1}{7}\sum_{k=1}^{7} x_k = 10.0, \tag{48}$$

$$s^2 = \frac{1}{7-1}\sum_{k=1}^{7} x_k - \frac{7}{7-1}\bar{x}^2 = 0.08 \Rightarrow s \approx 0.2828. \tag{49}$$

Para encontrar el intervalo de confianza para la media, notamos que se trata de una variable aleatoria normal de la cual se desconocen tanto su valor medio como la varianza. Por lo tanto, necesitamos encontrar $t_{6,0.975}$. En la tabla de la fractiles de la t de Student encontramos $t_{6,0.975} = 2.4469$. Por lo tanto, la semi-amplitud del intervalo es

$$t_{6,0.975}\frac{s}{\sqrt{7}} \approx 0.2616, \tag{50}$$

y el intervalo queda

$$10 \pm 0.2616 = (9.7384, 10.2616). \tag{51}$$

Veamos cómo podemos resolver este problema usando, por ejemplo, Octave.

```
>> x = [9.8 10.2 10.4 9.8 10 10.2 9.6];
>> n = length(x);
>> xbar = mean(x);
>> s = std(x);
>> alpha = 0.05;
>> t = tinv(1-alpha/2,n-1);
>> intervalo = [xbar - t*s/sqrt(n), xbar + t*s/sqrt(n)]
intervalo =

    9.7384   10.2616
```

Una forma de hacerlo en Python es la siguiente:

```
In [46]: import numpy as np
In [47]: from scipy import stats as st
In [48]: x = [9.8,10.2,10.4,9.8,10,10.2,9.6]
In [49]: n = len(x)
In [50]: xbar = np.mean(x)
In [51]: sn = st.sem(x)
In [52]: alpha = 0.05
In [53]: t = st.t.ppf(1-alpha/2,n-1)
In [54]: [xbar - t*sn, xbar + t*sn]
Out[54]: [9.7384141201766834, 10.261585879823317]
```

Obsérvese que `scipy.stats.sem()` devuelve $s/\sqrt{n}$. Otra forma de hacerlo en Python:

```
In [55]: st.t.interval(0.95, len(x)-1, loc=np.mean(x), scale=st.sem(x))
Out[55]: (9.7384141201766834, 10.261585879823317)
```

Y una última forma en Python:

```
In [56]: import statsmodels.stats.api as sm
In [57]: sm.DescrStatsW(x).tconfint_mean(alpha=0.05)
Out[57]: (9.7384141201766834, 10.261585879823317)
```

Ejercicio 21

De un proceso productivo de una pieza seriada se tomó una muestra de 300 unidades en la que se encontraron 18 defectuosas.

1. Calcular los límites de confianza del 90 % para el porcentaje defectuoso del proceso.
2. Calcular el tamaño de muestra adicional para tener un intervalo del mismo nivel de confianza pero de semi-amplitud 0.01 (o sea del 1 % de semi-amplitud).
3. Con la muestra dada de 300 unidades calcular el porcentaje defectuoso máximo del proceso con 90 % de confianza (o sea un porcentaje tal que la probabilidad de que el verdadero porcentaje defectuoso lo exceda sea 0.1).

Resolución

Llamemos p a la proporción de defectuosos. Dado que n es grande (> 100), no hay dificultad en aproximar la distribución del estimador $\hat{p}$ por una normal. La estimación en el caso de la parte 1 del problema, está dada por

$$\hat{p} \pm z_{0.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{18}{300} \pm 1.6449\sqrt{\frac{\frac{18}{300}(1-\frac{18}{300})}{300}} = 0.06 \pm 0.0226 = (0.0374, 0.0826), \tag{52}$$

donde $z_{0.95}$ se obtuvo de la tabla correspondiente.

La parte 3 del ejercicio pide un intervalo de confianza unilateral a derecha (por eso habla del porcentaje defectuoso *máximo*). El lado derecho de este intervalo es

$$\hat{p} + z_{0.90}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{18}{300} + 1.2816\sqrt{\frac{\frac{18}{300}(1-\frac{18}{300})}{300}} = 0.06 + 0.0176 = 0.0776. \tag{53}$$

La parte 2 del ejercicio pide calcular el tamaño de muestra adicional para tener una semi-amplitud del intervalo de confianza bilateral menor o igual a 0.01. Es decir, hay que buscar el m tal que

$$z_{0.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{m}} \leq 0.01. \tag{54}$$

Si no supiéramos nada acerca de p y, por tanto, de los valores aproximados que tomará $\hat{p}$, deberíamos procurar satisfacer esta desigualdad para cualquier valor posible $\hat{p}$. Ya hemos visto que el máximo de la función $g(p) = p(1-p)$ se obtiene cuando $p = 0.5$, por lo que deberíamos requerir que

$$z_{0.95}\sqrt{\frac{\frac{1}{4}}{m}} \leq 0.01 \Rightarrow m \geq \left(\frac{z_{0.95}}{2 \times 0.01}\right)^2 \approx 6763.9. \tag{55}$$

Es decir, se necesitan al menos 6464 muestras adicionales. Este resultado *no está mal*, pero es muy conservador. De hecho, tenemos una estimación puntual para p a partir de una gran cantidad de muestras. Si bien cambiará al variar el número de muestras, es poco probable que se aleje mucho del valor ya obtenido. Por lo tanto, es más razonable pedir que

$$z_{0.95}\sqrt{\frac{0.06 \times 0.94}{m}} \leq 0.01 \Rightarrow m \geq \left(\frac{z_{0.95}}{0.01}\right)^2 \times 0.06 \times 0.94 \approx 1525.9. \tag{56}$$

Por lo tanto, necesitamos al menos 1226 muestras adicionales.

Ejercicio 17

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:

$$X_n = \text{cantidad de televisores que sintonizan el programa entre el panel de } n \text{ hogares.} \tag{57}$$

Por otro lado, podríamos considerar:

$$\widehat{p}_n = \frac{X_n}{n} = \textbf{proporción} \text{ de televisores que sintonizan el programa entre el panel de } n \text{ hogares.} \tag{58}$$

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:

$$X_n \sim \text{Bi}(n, p) \tag{59}$$

donde p es la proporción **poblacional**.

Ítem 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:

$$\text{P}\left(a(\widehat{p}_n, n) \leq p \leq b(\widehat{p}_n, n)\right) = 0.9 \tag{60}$$

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:

$$X_n \overset{(a)}{\sim} N(np; \sqrt{np(1-p)}) \Rightarrow \widehat{p}_n = \frac{X_n}{n} \overset{(a)}{\sim} N\left(p; \sqrt{\frac{p(1-p)}{n}}\right) \Rightarrow \frac{\widehat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0;1) \tag{61}$$

Por lo tanto,

$$\text{P}\left(-z_{0.95} \leq \frac{\widehat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{0.95}\right) = 0.9 \Rightarrow \text{P}\left(-z_{0.95}\sqrt{\frac{p(1-p)}{n}} \leq \widehat{p}_n - p \leq z_{0.95}\sqrt{\frac{p(1-p)}{n}}\right) = 0.9 \tag{62}$$

$$\text{P}\left(z_{0.95}\sqrt{\frac{p(1-p)}{n}} \geq p - \widehat{p}_n \geq -z_{0.95}\sqrt{\frac{p(1-p)}{n}}\right) = 0.9 \tag{63}$$

$$\text{P}\left(\widehat{p}_n + z_{0.95}\sqrt{\frac{p(1-p)}{n}} \geq p \geq \widehat{p}_n - z_{0.95}\sqrt{\frac{p(1-p)}{n}}\right) = 0.9 \tag{64}$$

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,

$$\lim_{n \to +\infty} \text{P}(|\widehat{p}_n - p| \geq \varepsilon) = 0, \ \forall \varepsilon > 0 \tag{65}$$

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:

$$\begin{aligned} 0.9 &= \text{P}\left(\widehat{p}_n + z_{0.95}\sqrt{\frac{p(1-p)}{n}} \geq p \geq \widehat{p}_n - z_{0.95}\sqrt{\frac{p(1-p)}{n}}\right) \\ &\approx \text{P}\left(\underbrace{\widehat{p}_n + z_{0.95}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}}_{b(\widehat{p}_n, n)} \geq p \geq \underbrace{\widehat{p}_n - z_{0.95}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}}_{a(\widehat{p}_n, n)}\right) \end{aligned} \tag{66}$$

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:

$$IC_{90\,\%}(p) = \left[\widehat{p}_n - z_{0.95}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}; \widehat{p}_n + z_{0.95}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right] \tag{67}$$

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:

$$\begin{aligned} IC_{90\,\%}(p) &= \left[0.25 - 1.6448 \cdot \sqrt{\frac{0.25 \cdot 0.75}{600}}; 0.25 + 1.6448 \cdot \sqrt{\frac{0.25 \cdot 0.75}{600}}\right] \\ &= [0.2209; 0.2791] = [22.09\,\%; 27.91\,\%] \end{aligned} \tag{68}$$

Ítem 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:

$$z_{0.95} \cdot \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}} \leq 0.01 \Rightarrow \sqrt{n} \leq z_{0.95} \cdot \frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{0.01} \Rightarrow n \geq \frac{z_{0.95}^2 \cdot \widehat{p}_n(1-\widehat{p}_n)}{0.01^2} \tag{69}$$

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:

$$n \geq \frac{z_{0.95}^2 \cdot 0.25}{0.01^2} \geq \frac{z_{0.95}^2 \cdot \widehat{p}_n(1-\widehat{p}_n)}{0.01^2} \tag{70}$$

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

$$n \geq \frac{z_{0.95}^2 \cdot 0.25}{0.01^2} \Rightarrow n \geq 6763.859 \Rightarrow n = 6764 \tag{71}$$

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 $\widehat{p}_{600} \approx \widehat{p}_n$:

$$n \geq \frac{z_{0.95}^2 \cdot \widehat{p}_n(1-\widehat{p}_n)}{0.01^2} \approx \frac{z_{0.95}^2 \cdot 0.25(1-0.25)}{0.01^2} = 5072.894 \Rightarrow n = 5073 \tag{72}$$

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:

$$IC^R_{95\,\%}(p) = \left[\widehat{p}_n - z_{0.975}\sqrt{\frac{p(1-p)}{n}}; \widehat{p}_n + z_{0.975}\sqrt{\frac{p(1-p)}{n}}\right] \tag{73}$$

$$IC^E_{95\,\%}(p) = \left[\widehat{p}_n - z_{0.975}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}; \widehat{p}_n + z_{0.975}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right] \tag{74}$$

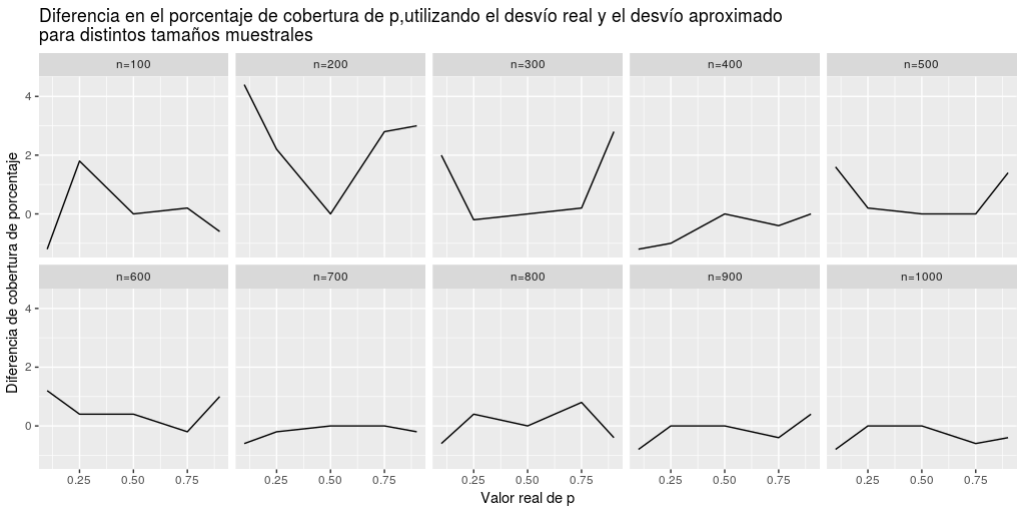
como es una simulación, el valor de p es conocido y podemos calcular el intervalo IC^R que en la práctica es imposible o demasiado costoso.

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:

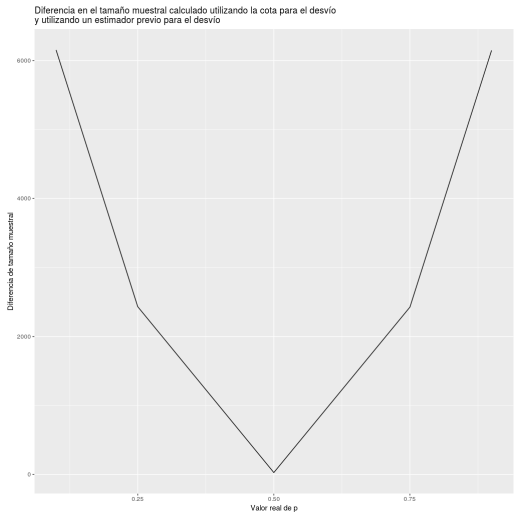


Notar que si la diferencia es negativa, es porque el intervalo con el desvío estimado tuvo un mayor porcentaje de cobertura.

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.