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 $\mu = E[X_1]$. 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*:

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

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

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

donde

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

Cuando $E[\hat{\theta}] = \theta$, 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

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

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

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

Estimación puntual de la media

Sean $X_1, X_2, \cdots, X_n$ variables aleatorias i.i.d. con media $\mu = E[X_1]$. 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 $\mu = E[X_1]$ 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
$$\mathrm{ECM}(\hat{\sigma}^2) = \mathrm{Var}\left[S^2\right] = \frac{\sigma^4}{n}\left(\kappa - 1 + \frac{2}{n-1}\right),$$
donde
$$\kappa = \text{curtosis} = \frac{E\left[(X_1 - E[X_1])^4\right]}{\sigma^4}.$$
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

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

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 $\mu = E[X_1]$ **desconocida** y varianza $\sigma^2 = \text{Var}[X_1]$ **conocida**. Luego, se tienen los siguientes intervalos de confianza:

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

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

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

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 de confianza:

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

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 $\mu = E[X_1]$ y varianza $\sigma^2 = \mathrm{Var}[X_1]$ **desconocidas**. Luego, se tienen los siguientes intervalos de confianza:

1. Unilateral a derecha:
$$I_r^\alpha = \left(-\infty, \bar{X} + t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}\right),$$
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:
$$I_l^\alpha = \left(\bar{X} - t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}, +\infty\right).$$
Recordemos que $t_{n-1,\alpha} = -t_{n-1,1-\alpha}$.
3. Bilateral:
$$I_{lr}^\alpha = \left(\bar{X} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right).$$

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.

Intervalo para la varianza de variables aleatorias normales

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

1. Unilateral a derecha:
$$I_r^\alpha = \left(-\infty, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha}} \right),$$
donde $\chi^2_{k,p}$ es el $100p$ percentil de una variable aleatoria con distribución χ^2_k.

2. Unilateral a izquierda:
$$I_l^\alpha = \left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha}}, +\infty \right).$$
Recordemos que $t_{n-1,\alpha} = -t_{n-1,1-\alpha}$.

3. Bilateral:
$$I_{lr}^\alpha = \left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \right).$$

Si n es muy grande, se pueden utilizar los intervalos de confianza para la media con varianza conocida, donde en lugar de $\bar{X}$ se coloca S^2 y en lugar de σ^2 escribimos $2S^4/(n-1)$ (estimador de la varianza de S^2).

Ejercicio 9 de la guía 8

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 y la dispersión del contenido de los recipientes de esa clase asumiendo que el contenido de ácido en los recipientes es una variable aleatoria normal.

Respuesta:
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

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

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

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

$$t_{6,0.975}\frac{s}{\sqrt{7}} \approx 0.2616, \tag{3}$$

y el intervalo queda

$$10 \pm 0.2616 = (9.7384, 10.2616). \tag{4}$$

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 15 de la guía 8

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).

Respuesta:

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

$$\hat{p} \pm z_{0.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{18}{300} \pm 1.6449\sqrt{\frac{\frac{18}{300}(1-\frac{18}{300})}{300}} = 0.06 \pm 0.0226 = (0.0374, 0.0826), \tag{5}$$

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

$$\hat{p} + z_{0.90}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{18}{300} + 1.2816\sqrt{\frac{\frac{18}{300}(1-\frac{18}{300})}{300}} = 0.06 + 0.0176 = 0.0776. \tag{6}$$

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

$$z_{0.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{m}} \leq 0.01. \tag{7}$$

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

$$z_{0.95}\sqrt{\frac{\frac{1}{4}}{m}} \leq 0.01 \Rightarrow m \geq \left(\frac{z_{0.95}}{2 \times 0.01}\right)^2 \approx 6763.9. \tag{8}$$

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

$$z_{0.95}\sqrt{\frac{0.06 \times 0.94}{m}} \leq 0.01 \Rightarrow m \geq \left(\frac{z_{0.95}}{0.01}\right)^2 \times 0.06 \times 0.94 \approx 1525.9. \tag{9}$$

Por lo tanto, necesitamos al menos 1226 muestras adicionales.