Práctica 1

Ejercicio 1. Sean $X_1, \ldots, X_n$ una muestra aleatoria con distribución $N(\mu, \sigma^2)$. Supongamos que σ es conocido. Calcular el estimador de máxima verosimilitud de $\theta = \mu^2$, $\widehat{\theta}_{MV}$. Calcular la distribución asíntotica de $\sqrt{n}(\widehat{\theta}_{MV} - \theta)$.

Ejercicio 2. Sea $X_1, \ldots, X_n$ una muestra aleatoria con distribución Binomial de parametros k y p. Si k es un número conocido

1. Encontrar el estimador de momentos y el de máxima verosimilitud de p.
2. Encontrar el estimador de máxima verosimilitud de las odds
$$O = \frac{p}{1-p}.$$
¿Es un estimador consistente?
3. Utilizando el método delta obtener la distribución aproximada de $\widehat{O}_{mv}$.
4. Estimar el valor de las odds y del error de estimación para una muestra con $k = 3$ que se observaron un total de 16 ceros, 10 unos, 3 dos y 1 tres.

Solucion

Sabemos que un estimador $\widehat{\theta}_n$ es consistente si $\widehat{\theta}_n \xrightarrow{P} \theta = O$

Sabemos por la ley de grandes números que $\overline{X}_n \xrightarrow{P} p$ por lo que también vale que $1 - \overline{X}_n \xrightarrow{P} 1 - p$. Por lo tanto vale que

$$\widehat{\theta}_n = \frac{\overline{X}_n}{1 - \overline{X}_n} \xrightarrow{P} \frac{p}{1-p} = O$$

como queríamos ver.

■

Ejercicio 3. Una empresa que fabrica autopartes necesita comprar láminas de acero. Para que una lámina de acero sea útil para el fabricante debe pasar un control de calidad que realiza la misma empresa. Dos proveedores de láminas de acero, llamados A y B, le ofrecen al fabricante de autopartes las láminas al mismo precio. Para decidir cuál proveedor contratar, un directivo de la empresa autopartista sugiere el siguiente plan. Llamemos p a la probabilidad de que una lámina de acero producida por el proveedor A, elegida al azar, pase el control de calidad del fabricante de autopartes y q a la probabilidad de que una lámina de acero producida por el proveedor B, elegida al azar, pase el control de calidad del fabricante de autopartes. El directivo de la empresa autopartista sugiere contratar al proveedor A si $p-q \geq 0$ y contratar al proveedor B si $p-q<0$. Como p y q son desconocidos, el directivo le pide a cada una de los proveedores una muestra de n láminas de acero y verifica si cada lámina pasa el control de calidad.

Llamemos X_i, $i=1,\ldots,n$ a la variable aleatoria que vale 1 si la i-ésima lámina de acero del proveedor A pasa el control de calidad y que vale 0 si la i-ésima lámina de acero del proveedor A no pasa el control de calidad. Llamemos Y_i, $i=1,\ldots,n$ a la variable aleatoria que vale 1 si la i-ésima lámina de acero del proveedor B pasa el control de calidad y que vale 0 si la i-ésima lámina de acero del proveedor B no pasa el control de calidad. Supongamos que $X_1,\ldots,X_n$ forman una muestra aleatoria, que $Y_1,\ldots,Y_n$ forman otra muestra aleatoria y que $X_1,\ldots,X_n,Y_1,\ldots,Y_n$ son independientes entre si. Finalmente, llamemos

$$\widehat{d}_n = \frac{\sum\limits_{i=1}^{n}(X_i-Y_i)}{n}.$$

1. Mostrar que $\widehat{d}_n$ es un estimador consistente de $p-q$.

2. Calcular el desvío estándar de $\widehat{d}_n$. Mostrar que el desvío estándar de $\widehat{d}_n$ es menor o igual que
$$\sqrt{\frac{1}{2n}}.$$

3. ¿Qué tamaño de muestra se necesita para garantizar que, con probabilidad al menos 0.95, el error de estimar $p-q$ usando $\widehat{d}_n$ sea menor a 0,1? Resolver este inciso de dos maneras: usando la desigualdad de Tchebyshev primero y el TCL despues (obteniendo en este último caso una respuesta aproximada).

4. Mostrar que
$$\widehat{v}_n = \sqrt{\frac{\sum_{i=1}^{n}(X_i-Y_i)^2}{n} - (\widehat{d}_n)^2}.$$
es un estimador consistente de el desvío standard de X_1-Y_1.

5. Calcular el límite en distribución de
$$\sqrt{n}\frac{\left(\widehat{d}_n-(p-q)\right)}{\widehat{v}_n}.$$

Solución Tenemos que

$$X_i \overset{\text{iid}}{\sim} Ber(p)$$
$$Y_i \overset{\text{iid}}{\sim} Ber(q)$$

Luego, $E(X_i)=p, E(Y_i)=q, Var(X_i)=p(1-p), Var(Y_i)=q(1-q)$. Además, notar que

$$\widehat{d}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i-Y_i) = \frac{1}{n}\sum_{i=1}^{n}X_i - \frac{1}{n}\sum_{i=1}^{n}Y_i = \overline{X}_n - \overline{Y}_n$$

1. $Var(X_i) = p(1-p) < +\infty$ y $Var(Y_i) = q(1-q) < +\infty$. Como X_i e Y_i son iid, podemos aplicar LGN:

$$\overline{X}_n \xrightarrow{P} E(X_1) = p$$
$$\overline{Y}_n \xrightarrow{P} E(Y_1) = q$$

Por propiedad de la convergencia en probabilidad:

$$\widehat{d}_n = \overline{X}_n - \overline{Y}_n \xrightarrow{P} p - q$$

Entonces, $\widehat{d}_n$ estima consistentemente $p-q$.

2. Por independencia **entre muestras**, $\overline{X}_n$ es independiente de $\overline{Y}_n$:

$$Var\left(\widehat{d}_n\right) = Var\left(\overline{X}_n - \overline{Y}_n\right) = Var\left(\overline{X}_n\right) + Var\left(\overline{Y}_n\right)$$

Por independencia **entre las observaciones de cada muestra**, la varianza de la media muestral es

$$Var\left(\overline{X}_n\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}$$
$$Var\left(\overline{Y}_n\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} q(1-q) = \frac{q(1-q)}{n}$$

Luego

$$Var\left(\widehat{d}_n\right) = Var\left(\overline{X}_n\right) + Var\left(\overline{Y}_n\right) = \frac{p(1-p)}{n} + \frac{q(1-q)}{n} = \frac{p(1-p)+q(1-q)}{n}$$

Por lo que el error estándar de $\widehat{d}_n$ es

$$\sigma_{\widehat{d}_n} = \sqrt{Var\left(\widehat{d}_n\right)} = \sqrt{\frac{p(1-p)+q(1-q)}{n}}$$

$g(t) = t(1-t)$ es una función que alcanza su máximo en $t = 1/2$. El valor en el máximo es $g(1/2) = 1/4$. Entonces,

$$p(1-p) \leq \frac{1}{4}$$
$$q(1-q) \leq \frac{1}{4}$$
$$\sigma_{\widehat{d}_n} = \sqrt{\frac{p(1-p)+q(1-q)}{n}} \leq \sqrt{\frac{1/4+1/4}{n}} = \sqrt{\frac{1}{2n}}$$

3. Utilizemos en primer lugar la desigualdad de Tchebyshev. Buscamos n tal que

$$\begin{aligned} P\left(\left|\widehat{d}_n-(p-q)\right|<0{,}1\right) &\geq 0{,}95 \\ 1-P\left(\left|\widehat{d}_n-(p-q)\right|\geq 0{,}1\right) &\geq 0{,}95 \\ -P\left(\left|\widehat{d}_n-(p-q)\right|\geq 0{,}1\right) &\geq -0{,}05 \\ P\left(\left|\widehat{d}_n-(p-q)\right|\geq 0{,}1\right) &\leq 0{,}05 \end{aligned}$$

Como $E\left(\widehat{d}_n\right)=p-q$ y $Var\left(\widehat{d}_n\right)<+\infty$, por Tchebycheff

$$P\left(\left|\widehat{d}_n-(p-q)\right|\geq 0{,}1\right)\leq \frac{Var\left(\widehat{d}_n\right)}{0{,}1^2}\leq \frac{\frac{1}{2n}}{0{,}01}=\frac{50}{n}$$

Entonces, para asegurarnos de tener la precisión buscada, n debe ser tal que

$$\begin{aligned} \frac{50}{n} &\leq 0{,}05 \\ \frac{50}{0{,}05} &\leq n \\ n &\geq 1000 \end{aligned}$$

Luego, veamos que resultado obtenemos de manera aproximada. Sea $D_i=X_i-Y_i$. Los D_i son iid porque cada muestra aleatoria es iid, con $E(D_i)=E(X_i)-E(Y_i)=p-q$ y $V(D_i)=V(X_i)+V(Y_i)=p(1-p)+q(1-q)<+\infty$. Veamos que

$$\widehat{d}_n=\frac{1}{n}\sum_{i=1}^{n}(X_i-Y_i)=\frac{1}{n}\sum_{i=1}^{n}D_i$$

Por lo que $\widehat{d}_n$ es la media muestral de D_i. Como $E(D_i)=p-q$ y $\sqrt{Var(D_i)}=\sqrt{v^2}<+\infty$, por TCL tenemos que

$$\sqrt{n}\left(\frac{\widehat{d}_n-(p-q)}{\sqrt{v^2}}\right)\xrightarrow{D} Z\sim N(0,1)$$

Entonces, veamos que

$$\begin{aligned} P\left(\left|\widehat{d}_n-(p-q)\right|<0{,}1\right) &= P\left(-0{,}1<\widehat{d}_n-(p-q)<0{,}1\right) \\ &= P\left(-0{,}1\sqrt{\frac{n}{v^2}}<\sqrt{\frac{n}{v^2}}(\widehat{d}_n-(p-q))<0{,}1\sqrt{\frac{n}{v^2}}\right) \\ &\approx P\left(-0{,}1\sqrt{\frac{n}{v^2}}<Z<0{,}1\sqrt{\frac{n}{v^2}}\right) \end{aligned}$$

Luego, como Z sigue una distribución normal estándar que es continua y simétrica, obtenemos lo siguiente:

$$P\left(-0{,}1\sqrt{\frac{n}{v^2}} < Z < 0{,}1\sqrt{\frac{n}{v^2}}\right) = 2P\left(Z < 0{,}1\sqrt{\frac{n}{v^2}}\right) - 1$$

Finalmente, lo que buscamos es encontrar cuál es el tamaño de muestra n que verifica que

$$\begin{aligned} P\left(\left|\widehat{d}_n - (p-q)\right| < 0{,}1\right) &\approx 2P\left(Z < 0{,}1\sqrt{\frac{n}{v^2}}\right) - 1 \geq 0{,}95 \\ &\iff P\left(Z < 0{,}1\sqrt{\frac{n}{v^2}}\right) \geq 1{,}975 \\ &\iff 0{,}1\sqrt{\frac{n}{v^2}} \geq 1{,}96 \\ &\iff \sqrt{n} \geq 19{,}6\sqrt{v^2} \end{aligned}$$

Recordando que $\sqrt{v^2} \leq \sqrt{\frac{1}{4} + \frac{1}{4}}$, la desigualdad se transforma en $n \geq \frac{19{,}6}{\sqrt{2}} = 192{,}08$. Lo que quiere decir que necesitamos $n \geq 193$.

4. Es importante notar que D_i es acotado: solamente puede tomar valores en $\{-1, 0, 1\}$ por lo que $E(D_i^4) < +\infty$. Esto implica $Var(D_i^2) < +\infty$. Luego, por LGN

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 = \frac{1}{n}\sum_{i=1}^{n} D_i^2 \xrightarrow{P} E(D_1^2)$$

Ya habíamos mostrado que

$$\widehat{d}_n \xrightarrow{P} p - q = E(X_1) - E(Y_1) = E(D_1)$$

Como el cuadrado es una función continua:

$$\left(\widehat{d}_n\right)^2 \xrightarrow{P} E(D_1)^2$$

Por propiedad de la convergencia en probabilidad:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 - \left(\widehat{d}_n\right)^2 \xrightarrow{P} E\left(D_1^2\right) - E(D_1)^2 = Var(D_1)$$

Como la raíz cuadrada es una función continua:

$$\widehat{v}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 - \left(\widehat{d}_n\right)^2} \xrightarrow{P} \sqrt{Var(D_1)} = v$$

Donde $v = \sqrt{Var(D_1)} = \sqrt{Var(X_1 - Y_1)}$ es el desvío estándar de $X_1 - Y_1$, por lo que $\widehat{v}_n$ es consistente.

5. Llamando a $\sqrt{Var(D_i)} = v < +\infty$, por TCL tenemos

$$\sqrt{n}\left(\frac{\widehat{d}_n - (p-q)}{v}\right) \xrightarrow{D} Z \sim N(0,1)$$

Como $\widehat{v}_n$ es consistente:

$$\frac{v}{\widehat{v}_n} \xrightarrow{P} 1$$

Suponiendo que $v > 0$, es decir que D_1 no es una constante.
Finalmente, por Slutsky:

$$\sqrt{n}\left(\frac{\widehat{d}_n - (p-q)}{\widehat{v}_n}\right) = \sqrt{n}\left(\frac{\widehat{d}_n - (p-q)}{v}\right)\frac{v}{\widehat{v}_n} \xrightarrow{D} Z \sim N(0,1)$$

Ejercicio 4. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución con parámetro θ. Sea $\widehat{\theta}_n$ un estimador de θ con varianza finita e insesgado. Demuestre que

$$ECM\left(\widehat{\theta}_n\right) = Var\left(\widehat{\theta}_n\right)$$

SoluciónRecordemos la definición de error cuadrático medio.

$$ECM\left(\widehat{\theta}_n\right) = E_\theta\left(\left(\widehat{\theta}_n - \theta\right)^2\right)$$

Llamando $e = \widehat{\theta}_n - \theta$ vale que $ECM(\widehat{\theta}_n) = E(e^2) = Var(e) + (E(e))^2$
Por lo que

$$ECM(\widehat{\theta}_n) = Var\left(\widehat{\theta}_n - \theta\right) + \underbrace{\left(E\left(\widehat{\theta}_n - \theta\right)\right)^2}_{sesgo\,=\,0 \text{ en este ejercicio}}$$

Entonces, tenemos que $ECM(\widehat{\theta}_n) = Var\left(\widehat{\theta}_n - \theta\right) = Var\left(\widehat{\theta}_n\right)$ porque θ es una constante, por lo que probamos los que queríamos ver.

■

Ejercicio 5. (Opcional) Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución con un parámetro de interés asociado θ. Mostrar que si

$$ECM\left(\widehat{\theta}_n\right) \to 0$$

entonces $\widehat{\theta}_n$ es consistente para θ.
Sugerencia: Escribir la definición de convergencia en probabilidad y considerar usar la desigualdad de Markov de manera conveniente.

Solucion
Sabemos que $ECM\left(\widehat{\theta}_n\right) \to 0$, que es lo mismo que decir que

$$E\left(\left(\widehat{\theta}_n - \theta\right)^2\right) \to 0.$$

Queremos ver que para cualquier $\varepsilon > 0$ vale que

$$\lim_{n\to\infty} P\left(|\widehat{\theta}_n - \theta| \geq \varepsilon\right) = 0$$

Como $P\left(|\widehat{\theta}_n - \theta| \geq \varepsilon\right)$ es una probabilidad, vale que

$$0 \leq P\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) = P\left(\left(\widehat{\theta}_n - \theta\right)^2 \geq \varepsilon^2\right) \underbrace{\leq}_{\text{Markov}} \frac{1}{\varepsilon^2} \cdot E\left[\left(\widehat{\theta}_n - \theta\right)^2\right]$$

Por lo tanto se tiene que:

$$0 \leq P\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} \cdot E\left[\left(\widehat{\theta}_n - \theta\right)^2\right]$$

Tomando $\lim\limits_{n\to\infty}$ se tiene que

$$0 \leq \lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \lim_{n\to\infty} \frac{1}{\varepsilon^2} \cdot E\left[\left(\widehat{\theta}_n - \theta\right)^2\right] \underbrace{=}_{ECM(\widehat{\theta}_n) \to 0} 0$$

Por teorema del sandwich vemos que

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) = 0$$

para cualquier valor de $\varepsilon > 0$ por lo tanto $\widehat{\theta}_n \xrightarrow{P} \theta$. Entonces $\widehat{\theta}$ es un estimador consistente para θ.

Ejercicio 6. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución con esperanza finita igual a μ y varianza finita, $Var(X)$.

1. Es $(\bar{X}_n)^2$ un estimador insesgado de $E(X)^2$?
2. Es $(\bar{X}_n)^2$ un estimador consistente de $E(X)^2$?
3. Para qué valores de k es $(\bar{X}_n)^2 - kS^2$ un estimador insesgado y consistente de $E(X)^2$?

Solución
Antes de comenzar, recordamos los siguientes resultados:
$E(X_i^2) = \mu^2 + Var(X_i)$

$E(\bar{X}_n^2) = \mu^2 + \dfrac{1}{n} Var(X_1)$

- a

$$
\begin{aligned}
E\left(\bar{X_n}^2\right) &= E\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right) \\
&= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \left(E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\right)^2 \\
&= \frac{1}{n} Var(X) + \mu^2 \neq E(X)^2 = \mu^2
\end{aligned}
$$

Por lo tanto, no es insesgado.

- b

Por la ley de grandes números sabemos que $\bar{X_n} \xrightarrow{P} \mu$ y como $g(x) = x^2$ es una función continua entonces, por propiedades de la convergencia en probabilidad, vale que $\left(\bar{X_n}\right)^2 \xrightarrow{P} \mu^2 = [E(X)]^2$

Por lo tanto, este estimador es consistente.

- c

$$
\begin{aligned}
E\left(\left(\bar{X_n}\right)^2 - kS^2\right) &= E\left(\bar{X_n}^2\right) - kE\left(S^2\right) \\
&= \frac{1}{n} Var\left(X\right) + \mu^2 - kVar\left(X\right) \\
&= \left(\frac{1}{n} - k\right) Var\left(X\right) + \mu^2
\end{aligned}
$$

Entonces si $k = \dfrac{1}{n}$ se cumple que el estimador es insesgado

$$
E\left(\left(\bar{X_n}\right)^2 - kS^2\right) = \mu^2
$$

Para verificar consistencia de $\bar{X_n}^2 + \dfrac{1}{n}S^2$, notamos que $\bar{X_n}^2 \xrightarrow{P} \mu^2$ y que $\dfrac{1}{n}S^2 \xrightarrow{P} 0$ pues $S^2 \xrightarrow{P} Var(X_1)$ y $\dfrac{1}{n} \to 0$

■

Ejercicio 7. Una investigadora quiere estimar la fracción de estudiantes en cierta universidad que ha violado el código ético de la institución. Para esto obtiene una muestra aleatoria de n estudiantes. La investigadora se da cuenta que preguntarles directamente a los estudiantes si han violado el código ético probablemente de lugar a algunas respuestas deshonestas. Considera entonces la siguiente técnica, llamada respuesta aleatorizada.

La investigadora arma un mazo de 100 cartas, donde 50 cartas tienen escrita la pregunta 'Ha usted violado el código ético de la universidad?' y las otras 50 cartas tienen escrita la pregunta 'Es el último dígito de tu número de teléfono 0, 1 o 2?'

A cada estudiante en la muestra aleatoria se le pide que mezcle el mazo de cartas, saqué una carta y responda a la pregunta en la carta elegida. Cómo la mitad de las cartas tienen escrita una pregunta irrelevante, responder que sí ya no estigmatiza a los alumnos.

Sea p la proporción de alumnos de la universidad que han violado el código ético y llamemos q a la probabilidad de que un alumno responda Si en la técnica de respuesta aleatorizada.

1. Usando la Ley de Probabilidad Total, mostrar que $q = 1/2 \times p + 1/2 \times 3/10$.

2. Sea Y la variable aleatoria que cuenta el número de respuestas positivas. Mostrar que Y/n es un estimador insesgado de q. Usando esto, proponer un estimador insesgado de p basado en Y. Si $n = 80$ e $y = 20$, ¿cúanto vale el estimador propuesto?

Solución

Sea $X_i = 1$ si el alumno responde que sí y 0 en caso contrario. Sea $A_i = 1$ si el alumno sacó una carta con la pregunta respecto a la violación del código ético y 0 en caso contrario.

1. Asumimos que el sistema funciona (es decir, que los alumnos responden honestamente porque no se sienten estigmatizados). Es razonable pensar que el último dígito del número de teléfono, la carta sacada del mazo y el hecho de haber violado o no el código ético son independientes. Además, suponemos que todos los dígitos tienen igual probabilidad de aparecer al final del número telefónico de cualquier estudiante. Luego, por la Ley de Probabilidad Total:

$$
\begin{aligned}
q &= P(X_i = 1) \\
&= P(X_i = 1 \wedge A_i = 1) + P(X_i = 1 \wedge A_i = 0) \\
&= P(A_i = 1)P(X_i = 1|A_i = 1) + P(A_i = 0)P(X_i = 1|A_i = 0) \\
&= \frac{1}{2} \times p + \frac{1}{2} \times \frac{3}{10} \\
&= \frac{10p + 3}{20}
\end{aligned}
$$

   Donde usamos

$$
\begin{aligned}
P(A_i = 1) = P(A_i = 0) &= \frac{1}{2} \\
P(X_i = 1|A_i = 1) &= p \\
P(X_i = 1|A_i = 0) &= \frac{3}{10}
\end{aligned}
$$

2. Tenemos

$$
E\left(\frac{Y}{n}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} q = q
$$

   Considerar el siguiente estimador de p:

$$
\widehat{p} = \frac{2Y}{n} - \frac{3}{10}
$$

Se trata de un estimador insesgado:

$$E(\widehat{p}) = 2E\left(\frac{Y}{n}\right) - \frac{3}{10} = 2q - \frac{3}{10} = 2\left(\frac{10p+3}{20}\right) - \frac{3}{10} = p + \frac{3}{10} - \frac{3}{10} = p$$

Si $n = 80$ e $y = 20$ tenemos

$$\widehat{p} = \frac{2 \times 20}{80} - \frac{3}{10} = \frac{1}{5}$$

■

Ejercicio 8. Sean $X_1, \ldots, X_n$ una muestra aleatoria de variables aleatorias que se distribuyen $N(\theta, \theta^2)$ con $\theta > 0$. Encuentre los estimadores del método de momentos que podrían obtenerse a partir del primer y segundo momento.