Universidad de la República | **Probabilidad y Estadística**
Facultad de Ingeniería - IMERL | **Primer Semestre 2017**

SEGUNDO PARCIAL
SÁBADO 24 DE JUNIO 2017.

Número de Parcial	Cédula	Nombre y Apellido

PARA USO DOCENTE		
Ej. 1	Ej. 2	TOTAL

Ejercicio 1. [**27 puntos**] La Tabla 1 muestra la frecuencia de partidos según la cantidad de goles en el mundial de fútbol de Francia 98. En total se jugaron 64 partidos y se hicieron 170 goles.

Table 1: Goles en el mundial de Francia 98

Número total de goles por partido	0	1	2	3	4	5	6	7	Total
Frecuencia de partidos con esa cantidad de goles	5	11	12	18	11	6	0	1	64

Por ejemplo, hubo 5 partidos en los que no hubo goles, y hubo un solo partido en los que se convirtieron 7 goles. Sea X la variable aleatoria que cuenta la cantidad de goles en un partido de fútbol de 90 minutos de duración.

1. En esta parte asumiremos que X tiene distribución de Poisson de parámetro λ.
   (a) Probar que $\overline{X}_n$ es un estimador insesgado de λ. Calcular el error cuadrático medio del estimador.
   (b) Usando los datos de la Tabla 1, determinar un intervalo de confianza para λ al nivel de confianza 0.9.
2. Dada una muestra $X_1, \dots, X_n$ de variables aleatorias independientes e idénticamente distribuidas, consideramos F_n la función de distribución empírica de los datos, esto es $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le x\}}$.
   (a) Para la muestra $X_1, \dots, X_{64}$ definida por la Tabla 1, graficar $F_n(x)$.
   (b) Indicar moda, mediana, primer y tercer cuartil de la muestra. Realizar el boxplot correspondiente.
3. Sea p la probabilidad de que un partido termine con 1 solo gol.
   (a) A partir de los datos de la Tabla 1, determinar un intervalo de confianza asintótico (aproximado) para p al nivel de confianza 0.9.
   (b) Un amiga/o quiere apostar en una penca para el mundial de Rusia 2018 (también son 64 partidos) y en su predicción resulta que hay 17 partidos terminados con un solo gol. ¿Cuál es tu opinión sobre su predicción? Justifica tu respuesta.
4. Sea Y_i una variable aleatoria definida por:

$$Y_i = \begin{cases} 1 & \text{si hay un gol en el } i\text{-ésimo minuto,} \\ 0 & \text{si no.} \end{cases}$$

   Asumiendo que las variables $Y_1, \dots, Y_{90}$ son independientes e idénticamente distribuidas, y que la probabilidad de que se haga un gol en un determinado minuto es pequeña, justificar que la distribución de X se puede aproximar por una Poisson.

Ejercicio 1. Solución

1. (a) Tenemos $X_1, \ldots, X_n$ es una muestra iid con distribución Poisson(λ). Sabemos por tanto que $\mathrm{E}(X) = \lambda$ y por la LGN se puede afirmar que $\bar{X}_n \xrightarrow[n]{\mathbb{P}} \mathrm{E}(X) = \lambda$ y por lo tanto $\bar{X}_n$ es un estimador del parámetro λ.
Además $\mathrm{E}\left(\bar{X}_n\right) = \mathrm{E}(X_1) = \lambda$, lo que muestra que $\bar{X}_n$ es un estimador insesgado. Como el estimador es insesgado, se tiene que $\mathtt{ECM}(\bar{X}_n) = \mathrm{Var}\left(\bar{X}_n\right) = \frac{\mathrm{Var}(X_1)}{n} = \frac{\lambda}{n}$.

   (b) En esta parte tenemos la opción de construir intervalos de confianza exacto o aproximados:

   Exacto: Un intervalo de confianza exacto a nivel $1-\alpha$ para $\mathrm{E}(X_1) = \lambda$, es:

$$I_{\alpha,n}(\lambda) = \left[\frac{\chi_{1-\alpha/2}(2S_n)}{2n}, \frac{1}{2n}\chi_{\alpha/2}(2S_n+2)\right],$$

donde $\chi_\alpha(N)$ es el punto que deja área α a la derecha en una distribución chi-cuadrado con N grados de libertad. En este caso tenemos que la suma total de goles es $S_n = X_1 + \ldots, +X_{64} = 170$. Usando una tabla de chi- cuadrado, para $\alpha = 0.9$ se tiene que $\chi_{0.95}(340) = 298.27$ y $\chi_{0.05}(342) = 386.12$. El intervalo de confianza resulta entonces $I_{\alpha,n} = [2.33, 3.02]$.

Apoximado: Utilizando la aproximación por TCL tenemos que un intervalo de confianza aproximado a nivel $1-\alpha$ para $\mathrm{E}(X_1) = \lambda$ es:

$$I_{\alpha,n}(\lambda) = \left[\bar{X}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2}\frac{s_n}{\sqrt{n}}\right]$$

De la tabla, podemos calcular

$$\bar{X}_{64} = \frac{S_{64}}{64} = \frac{170}{64} = \frac{1}{64}(0\times 5 + 1\times 11 + 2\times 12 + 3\times 18 + 4\times 11 + 5\times 6 + 6\times 0 + 7\times 1)$$

De la misma manera se puede calcular s_n desvío estándar empírico de la muestra:

$$s_n^2 = \frac{n}{n-1}\sigma_n^2 \quad \text{donde} \quad \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2$$

A partir de la tabla se obtiene que:

$$\sum_{i=1}^{64} X_i^2 = (0^2\times 5 + 1^2\times 11 + 2^2\times 12 + 3^2\times 18 + 4^2\times 11 + 5^2\times 6 + 6^2\times 0 + 7^2\times 1) = 596$$

Y por lo tanto $\sigma_n^2 = 2.24$ y $s_n^2 = 2.27$ (observar que la diferencia entre los dos valores es muy pequeña por lo que cualquiera de los dos podría ser utilizado como estimador del desvío estándar).

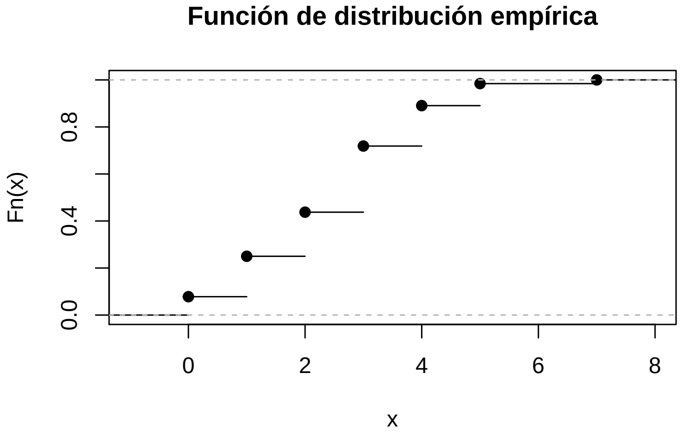
Finalmente para $\alpha = 0.9$, se tiene que $z_{\alpha/2} = z_{0.05} = 1.65$, y el intervalo de confianza resulta entonces: $I_{\alpha,n} = [2.66 - 0.31, 2.66 + 0.31] = [2.35, 2.97]$.

Otra alternativa posible es utilizar que $\mathrm{Var}(X_1) = \lambda$ y por lo tanto $\sqrt{\bar{X}_n}$ es también un estimador del desvío estándar, de donde el intervalo

$$I_{\alpha,n}(\lambda) = \left[\bar{X}_n - z_{\alpha/2}\frac{\sqrt{\bar{X}_n}}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2}\frac{\sqrt{\bar{X}_n}}{\sqrt{n}}\right],$$

es también un intervalo de confianza aproximado para λ.

Sustituyendo los valores (ya calculados) tenemos que $I_{\alpha,n} = [2.66 - 0.34, 2.66 + 0.34] = [2.32, 3]$.



2. La moda es el valor con mayor frecuencia de aparición en la muestra, en este caso, la moda es 3.

   El primer cuartil es $\hat{x}_{1/4} = 1$, la mediana es $\hat{x}_{1/2} = 3$ y el tercer cuartil es $\hat{x}_{3/4} = 4$.

   Para realizar el boxplot debemos calcular los límites inferior y superior para calcular los bigotes. En este caso el RIC vale $\hat{x}_{3/4} - \hat{x}_{1/4} = 3$, de donde Linf$= \hat{x}_{1/4} - 1.5 \times \text{RIC} = -3.5$, y el mínimo valor de la muestra que supera este límite es 0. De la misma manera, Lsup$= \hat{x}_{3/4} + 1.5 \times \text{RIC} = 8.5$, y el máximo valor de la muestra que es menor que este límite es 7. Por lo tanto los bigotes son 0 y 7 y no hay datos atípicos.

3. (a) Sea Z_i una variable aleatoria definida por:

$$Z_i = \begin{cases} 1 & \text{si el } i\text{-ésimo partido terminó con un solo gol,} \\ 0 & \text{si no.} \end{cases}$$

   Por lo tanto tenemos que $Z_1, \ldots, Z_{64}$ es una muestra de variables aleatorias iid con distribución Ber(p) con p la probabilidad de que un partido termine con un sol gol. Un intervalo de confianza para p es entonces:

$$I_{\alpha,n} = \left[\bar{Z}_n - z_{\alpha/2} \frac{\sqrt{\bar{Z}_n(1-\bar{Z}_n)}}{\sqrt{n}}, \bar{Z}_n + z_{\alpha/2} \frac{\sqrt{\bar{Z}_n(1-\bar{Z}_n)}}{\sqrt{n}} \right].$$

   Calculamos a partir de los datos $\hat{p}$ la proporción empírica de partidos que finalizaron con un solo gol, esto es $\hat{p} = \bar{Z}_n = 11/64 = 0.17$, de donde $I_{\alpha,n} = [0.17 - 0.077, 0.17 + 0.077] = [0.093, 0.247]$.

   (b) En este caso, nuestro amigo/a estaría apostando a una proporción de partidos culminados en un solo gol de $q = 17/64 = 0.265$ que no pertence al intervalo de confianza para dicha proporción a nivel de confianza de 0.9. Por lo tanto, le diríamos que tiene muy pocas chances de ganar la penca.

4. Por como se definieron las variables Y_i, tenemos que $X = \sum_{i=1}^{90} Y_i$ donde para todo i tenemos que $Y_i \sim \text{Ber}(p)$ siendo p la probabilidad de que se haga un gol en el i-ésimo minuto. Las variables pueden asumirse independientes e idénticamente distribuidas. Resulta entonces que X tiene distribución Binomial de parámetros $n = 90$ y p. Considerando que n es grande y p es pequeña es que podemos utilizar la aproximación Poisson a la Binomial. Es claro que esto es una aproximación pues no todos los minutos son igualmente probables y la independencia es también un supuesto fuerte. Sin embargo, si comparamos las probabilidades empíricas con las probabilidades teóricas de una distribución Poisson de parámetro $\lambda = 2.66$ tendríamos un muy buen ajuste de las mismas.

Ejercicio 2. [**33 puntos**] La distribución de riqueza entre las personas de un país suele modelarse mediante una distribución Pareto. Se dice que una variable aleatoria X tiene distribución Pareto de parámetros $a > 0$ y $\gamma > 1$ si es absolutamente continua con densidad dada por:

$$f_X(x) = \begin{cases} \gamma \frac{a^\gamma}{x^{\gamma+1}} & \text{si } x \geq a, \\ 0 & \text{en otro caso.} \end{cases}$$

Además la función de distribución de X está dada por:

$$F_X(x) = \begin{cases} 1 - \left(\frac{a}{x}\right)^\gamma & \text{si } x \geq a, \\ 0 & \text{en otro caso.} \end{cases}$$

Consideremos entonces X la variable aleatoria que indica los ingresos de una persona en un determinado país, que tiene distribución Pareto de parámetros a y γ.

1. Probar que $\mathrm{E}(X) = a\frac{\gamma}{\gamma-1}$. Hallar la mediana de X.
2. Calcular la probabilidad $\mathbf{P}(X > \mathbf{E}(X))$. ¿Qué ocurre con esta probabilidad cuando γ decrece a 1? La yapa[1]: ¿Le parece que el promedio es un indicador representativo del ingreso de una persona en este caso?
3. Sea $X_1, X_2, \dots, X_n$ una muestra de variables aleatorias i.i.d con distribución Pareto de parámetros a y γ. Asumiendo que a es conocido:
   (a) Hallar el estimador por momentos y por máxima verosimilitud de γ. ¿Se le ocurre algún otro estimador de γ diferente de los dos anteriores?
   (b) Asumimos ahora que $\gamma > 2$. Hallar σ^2 la varianza de X y probar que $g(\overline{X}_n)$ es un estimador de σ^2 siendo
   $$g(x) = \frac{(x-a)^2 x}{2a - x}$$
   definida para $x < 2a$.
4. Se tienen datos de ingresos en el Uruguay según la encuesta de hogares del 2014. De estos, se tomaron aquellos salarios mayores a 12.000 pesos (salario mínimo nacional) y menores que 89.000, que resulta en una muestra de $n = 16662$ personas. En el cuadro que sigue se muestra un resumen estadístico de dichos datos (los datos se expresan en miles de pesos):

min	q_1	m_X	$\overline{x}$	q_3	max	s_n
12	15	20	23.71	28	89	12.26

   donde min es el mínimo de los datos, max es el máximo, m_X es la mediana empírica, $\overline{x}$ es el promedio, q_1 y q_3 son el primer y tercer cuartil de la muestra respectivamente y s_n es el desvío estándar. Para las siguientes preguntas, se asume que $a = 12$.
   (a) Estimar γ e indicar un intervalo de confianza asintótico (aproximado) a nivel 0.95 para γ.
   (b) Estimar la probabilidad de que el promedio de ingresos de la muestra sea menor o igual a 25.000 pesos uruguayos.

Ejercicio 2. Solución

1.

$$\mathrm{E}(X) = \int_{\mathbb{R}} 1 - F_X(x)dx = \int_0^a 1dx + \int_a^\infty \left(\frac{a}{x}\right)^\gamma dx = a + a^\gamma \frac{x^{-\gamma+1}}{-\gamma+1}\Big|_a^\infty = a + \frac{a}{\gamma-1} = \frac{\gamma}{\gamma-1}a$$

Observar que para resolver el límite en $+\infty$ hemos usado que $\gamma > 1$ y por lo tanto $x^{-\gamma+1}$ tiende a cero. El mismo resultado se obtiene obviamente utilizando la fórmula de valor esperado en términos de la densidad. Observar que si $\gamma < 1$ entonces el valor esperado es infinito.

Para calcular la mediana m_X basta observar que F_X es continua, entonces $m_X = F_X^{-1}(1/2)$. Por lo tanto igualando $F_X(m_X) = 1/2$ resulta que $\left(\frac{a}{m_X}\right)^\gamma = \frac{1}{2}$, de donde $m_X = a2^{1/\gamma}$.

[1]Esta última pregunta es por dos puntos extras.

2.
$$\mathbb{P}(X > \mathrm{E}(X)) = \mathbb{P}\left(X > a\frac{\gamma}{\gamma-1}\right) = 1 - F_X\left(a\frac{\gamma}{\gamma-1}\right) = \left(\frac{\gamma-1}{\gamma}\right)^{\gamma} \underset{\gamma\to 1}{\longrightarrow} 0$$

La yapa: la densidad es decreciente con un máximo en $x = a$, por lo tanto en términos de riquezas, diríamos que hay muchas personas con poca riqueza (con un mínimo que es a) y pocas personas con mucha riqueza. Por lo tanto el valor esperado no es muy representativo, ya que "promedia a ambas poblaciones". Observar además que si $\gamma \to 1$, el valor esperado tiende a infinito y la probabilidad de encontrar a una persona con riqueza mayor al valor esperado es cero. Observar que esto no sucede por ejemplo con la mediana: $m_X \to 2a$ cuando $\gamma \to 1$.

3. (a) El estimador por momentos se obtiene directamente de aplicar la LGN:
$$\bar{X}_n \to \mathrm{E}(X) = a\frac{\gamma}{\gamma-1}$$

Asumiendo a conocido basta definir $\hat{\gamma}$ tal que $\bar{X}_n = a\frac{\hat{\gamma}}{\hat{\gamma}-1}$. Despejando, obtenemos que $\hat{\gamma} = \frac{\bar{X}_n}{\bar{X}_n - a}$ es el estimador por momentos de γ. Observar que $X_i \geq a \,\forall\, i = 1, \cdots, n$ y por lo tanto $\bar{X}_n > a$.

La función de verosimilitud está definida por:
$$L(\gamma) = \prod_{i=1}^{n} f_X(X_i) = \prod_{i=1}^{n} \gamma\frac{a^{\gamma}}{X_i^{\gamma+1}} \quad \text{siempre que} \quad X_i \geq a \,\forall\, i = 1, \cdots, n$$

Al ser logaritmo una función monónota creciente, es equivalente maximizar $L(\gamma)$ que $\log(L(\gamma))$. Tomando logaritmo resulta entonces que:
$$\log L(\gamma) = n\log(\gamma) + n\gamma\log(a) - (\gamma+1)\sum_{i=1}^{n}\log(X_i)$$

Derivando respecto a γ e igualando a cero, tenemos que:
$$\log L(\gamma)' = \frac{n}{\gamma} + n\log(a) - \sum_{i=1}^{n}\log(X_i) = 0 \text{ si y solo si}$$
$$\hat{\gamma}_n = \frac{n}{\sum_{i=1}^{n}\log(X_i) - n\log(a)} = \frac{n}{\sum_{i=1}^{n}\log(X_i/a)} = \frac{1}{\bar{Y}_i} \quad \text{donde} \quad Y_i = \log(X_i/a).$$

Estudiando el signo de la función se puede ver que $\hat{\gamma}$ es efectivamente un máximo de la verosimilitud. Además con un poco más de trabajo se pude verificar que $\hat{\gamma}_n$ converge en probabilidad a γ.

Es posible obtener otro estimador de γ a partir de la mediana empírica: $\hat{\gamma} = \frac{\log(2)}{\log(m_X/a)}$.

(b) $\mathrm{E}\left(X^2\right) = \int_{\mathbb{R}} x^2 f_X(x)dx = \int_a^{\infty} x^2\gamma\frac{a^{\gamma}}{x^{\gamma+1}}dx = \frac{\gamma}{\gamma-2}a^2$ donde usamos que $\gamma > 2$ para resolver los límites en infinito. Observar que si $\gamma < 2$ entonces la varianza es infinita.

La varianza resulta entonces que $\sigma^2 = \mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 = \frac{\gamma a^2}{(\gamma-2)(\gamma-1)^2}$.

Para probar que $g(\bar{X}_n)$ es un estimador de σ^2 basta ver que $\sigma^2 = g(\mathrm{E}(X))$ siendo g una función continua. Sea $\mu = \mathrm{E}(X) = a\frac{\gamma}{\gamma-1}$, entonces:
$$\sigma^2 = \frac{\mu^2}{\gamma(\gamma-2)} \quad \text{y} \quad \gamma(\gamma-2) = \frac{\mu(2a-\mu)}{\mu-a}$$

Juntando ambas igualdades llegamos a que:

$$\sigma^2 = \frac{\mu(\mu - a)^2}{(2a - \mu)} = g(\mu).$$

4. (a) Ya vimos que asumiendo a conocido (en este caso $a = 12$) un estimador de γ es $\hat{\gamma} = \frac{\bar{X}_n}{\bar{X}_n - a} = \frac{23.71}{23.71-12} = 2.025$.

Un intervalo de confianza aproximado a nivel $1 - \alpha$ para el valor esperado de X es:

$$I_{\alpha,n} = \left[\bar{X}_n - z_{\alpha/2} \frac{\sqrt{g(\bar{X}_n)}}{\sqrt{n}}, \bar{X}_n - z_{\alpha/2} \frac{\sqrt{g(\bar{X}_n)}}{\sqrt{n}} \right]$$

En este caso, $1 - \alpha = 0.95$, de donde $z_{0.025} = 1.96$. Además de los datos tenemos que $\bar{X}_n = 23.71$ y $g(\bar{X}_n) = \frac{a^2\hat{\gamma}}{(\gamma-1)^2(\hat{\gamma}-2)} = 11215,4$, de donde $\sqrt{g(\bar{X}_n)} = 105.9$.

Por lo tanto, $I_{\alpha,n}(\mu) = [23.71 - 1.61, 23.71 + 1.61] = [22.1, 25.32] = [x_i, x_d]$.

Sin embargo, este es un intervalo para $\mathrm{E}\,(X) = a\frac{\gamma}{\gamma-1}$ y no para γ. Despejando γ en función del valor esperado tenemos que:

$$I_{\alpha,n}(\gamma) = \left[\frac{x_d}{x_d - a}, \frac{x_i}{x_i - a} \right] = [1.90, 2.188]$$

Observar que el estimador $\bar{\gamma} = 2.025$ pertence al intervalo.

Si utilizamos s_n como estimador de la varianza se obtiene que $I_{\alpha,n}(\mu) = [23.71 - 0.186, 23.71 + 0.186] = [23.524, 23.896] = [x_i, x_d]$, y despejando al igual que antes:

$$I_{\alpha,n}(\gamma) = [2.008, 2.041].$$

Nuevamente, observar que el estimador $\bar{\gamma}$ pertenece al intervalor:

(b) Utilizando la aproximacin por TCL, tenemos que:

$$\begin{aligned} \mathbb{P}(\bar{X}_{16662} \leq 25) &= \mathbb{P}\left(\sqrt{16.662} \frac{\bar{X}_{16662} - \bar{x}}{\sqrt{g(\bar{X}_n)}} \leq \sqrt{16.662} \frac{25 - \bar{x}}{\sqrt{g(\bar{X}_n)}} \right) \\ &\approx \Phi\left(\sqrt{16.662} \frac{25 - 23.71}{105.9} \right) = \Phi(1.57) = 0.94 \end{aligned}$$

Usando $s_n = 12.26$, resulta que $\mathbb{P}(\bar{X}_{16662} \leq 25) \approx \Phi(13.58) = 1$.

Nota: Se puede observar que $\sqrt{g(\bar{X}_n)}$ y s_n dan estimaciones muy diferentes para el desvío estándar. Esto se debió a un error al indicar s_n que debió ser 12.26^2. Sin embargo, esto no afecta la realización del ejercicio y por supuesto todas las respuestas a la parte 4 donde se utilice $s_n = 12.26$ como estimador del desvío serán consideradas correctas.