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

Examen Julio
Sábado 22 de julio 2017.

| Número de Examen | Cédula | Nombre y Apellido |
|---|---|---|
| | | |

| PARA USO DOCENTE | | | | |
|---|---|---|---|---|
| Ej. 1 | Ej. 2 | Ej. 3 | Ej. 4 | TOTAL |
| | | | | |

**Ejercicio 1.** [**15 puntos**]

De los correos que llegan a una cuenta de mail, se sabe que el 20 % son spam (correos no deseados). Un detector automático sencillo de spam busca frases que se repitan en dichos correos, en particular se encontró que la frase"dinero gratis" aparece en el 10 % de los correos no deseados, mientras que aparece en apenas el 1 % de los correos normales (es decir que no son considerados spam).

1. Calcular la probabilidad de que un correo entrante contenga la frase "dinero gratis".
2. Un correo entrante contiene la frase "dinero gratis¿'cuál es la probabilidad de que sea spam? ¿y de que no lo sea?

## Ejercicio 1. Solución

Se consideran los siguientes sucesos: $F = \{$el correo entrante continene la frase "dinero gratis"$\}$ y $S = \{$el correo entrante es spam$\}$. De la letra tenemos que $\mathbf{P}(S) = 0{,}2$, $\mathbf{P}(F|S) = 0{,}1$ y $\mathbf{P}(F|S^c) = 0{,}01$.

1. Usando la fórmula de la probabilidad total resulta que:

$$\mathbf{P}(F) = \mathbf{P}(F|S)\mathbf{P}(S) + \mathbf{P}(F|S^c)\mathbf{P}(S^c) = 0{,}1 \times 0{,}2 + 0{,}01 \times 0{,}8 = 0{,}028.$$

2. Usando la fórmula de Bayes tenemos que:

$$\mathbf{P}(S|F) = \frac{\mathbf{P}(F|S)\mathbf{P}(S)}{\mathbf{P}(F)} = \frac{0{,}1 \times 0{,}2}{0{,}028} = 0{,}71.$$

Además $\mathbf{P}(S^c|F) = 1 - \mathbf{P}(S|F) = 1 - 0{,}71 = 0{,}29$ .

**Ejercicio 2.** [**30 puntos**] Los siguientes datos representan las alturas de una muestra de 10 hombres y 10 mujeres elegidos al azar del curso de PyE 2017 (primer semestre).

| Mujeres (cm) | 168 | 170 | 167 | 168 | 159 | 170 | 173 | 167 | 160 | 165 |
|---|---|---|---|---|---|---|---|---|---|---|
| Hombres (cm) | 188 | 165 | 193 | 173 | 179 | 173 | 168 | 174 | 179 | 183 |

Denotamos por $X$ la variable aleatoria que mide la altura de las mujeres e $Y$ la de los hombres. Suponemos que $X$ e $Y$ tienen distribución normal.

1. *a*) Hallar un intervalo de confianza exacto, al nivel de confianza 0,9, para la esperanza de $X$.
   *b*) Hallar un intervalo de confianza exacto, al nivel de confianza 0,9, para la varianza de $Y$.

2. Se sabe que en dicho curso la proporción de mujeres es de 0,2. Si se sabe que la altura de un estudiante elegido al azar del curso está entre 168 cm y 169 cm. ¿Es más probable que sea un hombre o una mujer? Asumir que el valor esperado y la varianza de $X$ e $Y$ son los estimados a partir de la muestra.

## Ejercicio 2. Solución

1. *a*) La variable $X$ tiene distribución normal de parámetros $\mu_X$ y $\sigma_X^2$. Un estimador de $\mu_X$ es el promedio muestral $\overline{X}_n = 166{,}7$. El estimador de $\sigma_X$ es $s_n = 4{,}37$. El intervalo de confianza exacto para $\mu_X$ al nivel de confianza $1-\alpha$ es

$$I = \left[\overline{X}_n \pm \frac{t_{\alpha/2}(n-1)s_n}{\sqrt{n}}\right]$$

en donde $t_{\alpha/2}(n-1)$ es el cuantil de la distribución de Student con $n-1$ grados de libertad. Para $\alpha = 0{,}1$ y $n = 10$ el cuantil vale $t_{\alpha/2}(n-1) = 1{,}83$. Así, el intervalo de confianza queda

$$I = [166{,}7 \pm 2{,}53] = [164{,}2; 169{,}2]\,.$$

*b*) El intervalo de confianza exacto para $\sigma_X^2$ al nivel de confianza $1-\alpha$ es

$$I = \left[\frac{n-1}{\chi_{\alpha/2}(n-1)}s_n^2, \frac{n-1}{\chi_{1-\alpha/2}(n-1)}s_n^2\right]$$

donde $s_n = \sigma_Y = 8{,}72$. $\chi_{\alpha/2}(n-1)$ y $\chi_{1-\alpha/2}(n-1)$ son los cuantiles de la distribución $\chi^2$ con $n-1$ grados de libertad. Si $\alpha = 0{,}1$ estos valen

$$\chi_{\alpha/2}(n-1) = 16{,}9 \text{ y } \chi_{1-\alpha/2}(n-1) = 3{,}3.$$

Así, el intervalo de confianza para $\sigma_Y^2$ queda

$$I = [40{,}5; 205{,}9]\,.$$

2. Llamemos $Z$ a la variable que indica la altura de un estudiante elegido al azar. Asumimos que $X \sim N(166{,}7; 19{,}1)$ y que $Y \sim N(177{,}5; 76{,}1)$.

Por el Teorema de Bayes la probabilidad de que sea mujer ($M$) está dada por

$$\mathbf{P}\left(M|Z \in [168,169]\right) = \frac{\mathbf{P}\left(Z \in [168,169]|M\right)\mathbf{P}(M)}{\mathbf{P}\left(Z \in [168,169]\right)}$$

Además $\mathbf{P}\left(Z \in [168,169]|M\right) = \mathbf{P}(X \in [168,169])$, y

$$\mathbf{P}(X \in [168,169]) = \mathbf{P}\left(0{,}3 \leq \frac{X-\mu_X}{\sigma_X} \leq 0{,}53\right) = \Phi(0{,}53) - \Phi(0{,}3) = 0{,}084.$$

Entonces

$$\boxed{\mathbf{P}\left(M|Z \in [168,169]\right) = \frac{0{,}084 \times 0{,}2}{\mathbf{P}\left(Z \in [168,169]\right)}}$$

Del mismo modo podemos calcular la probabilidad de que sea hombre ($H$)

$$\mathbf{P}\left(H|Z \in [168,169]\right) = \frac{\mathbf{P}\left(Z \in [168,169]|H\right)\mathbf{P}(H)}{\mathbf{P}\left(Z \in [168,169]\right)}$$

En este caso $\mathbf{P}\left(Z \in [168,169]|H\right) = \mathbf{P}(Y \in [168,169])$, y

$$\mathbf{P}(Y \in [168,169]) = \mathbf{P}\left(-1{,}09 \leq \frac{Y-\mu_Y}{\sigma_Y} \leq -0{,}97\right) = \Phi(0{,}53) - \Phi(0{,}3) = 0{,}028.$$

Entonces

$$\boxed{\mathbf{P}\left(H|Z \in [168,169]\right) = \frac{0{,}028 \times 0{,}8}{\mathbf{P}\left(Z \in [168,169]\right)}}$$

Juntando los dos resultados recuadrados obtenemos que

$$\frac{\mathbf{P}\left(H|Z \in [168,169]\right)}{\mathbf{P}\left(M|Z \in [168,169]\right)} = \frac{0{,}028 \times 0{,}8}{0{,}084 \times 0{,}2} = 1{,}33.$$

Es decir, es más probable que sea hombre ($H$).

**Ejercicio 3.** [**30 puntos**] Una moneda tiene probabilidad $p$ de salir cara. Sea $X$ la variable aleatoria que cuenta el número de lanzamientos de la moneda hasta que salga cara por segunda vez (si sale cara en el primer y en el segundo lanzamiento $X$ vale 2).

1. Sea $X_1, \ldots, X_n$ una muestra i.i.d. de $X$. Hallar el estimador de máxima verosimilitud de $p$.
2. La tabla siguiente resume $n = 500$ valores observados de $X$.

| Valor observado de $X$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frecuencia | 74 | 98 | 96 | 66 | 47 | 42 | 26 | 20 | 11 | 7 | 2 | 6 | 0 | 3 | 1 | 1 |

   Estimar $p$ a partir de estos datos.
3. ¿Se le ocurre otro estimador de $p$? En caso afirmativo calcularlo.

## Ejercicio 3. Solución

1. La función de probabilidad puntual de $X$ es $\mathbf{P}(X = k) = (k-1)p^2(1-p)^{k-2}$. Es posible reconocer a $X$ como una v.a. con distribución Binomial Negativa con parámetros $k = 2$ y $p$. Entonces, la función de verosimilitud es

$$L_n(p) = \prod_{i=1}^{n} (X_i - 1)p^2(1-p)^{X_i - 2}.$$

   Tomando logaritmos a ambos lados

$$\log(L_n(p)) = \sum_{i=1}^{n} \log(X_i - 1) + 2\log(p) + (X_i - 2)\log(1-p).$$

   Derivando respecto de $p$ obtenemos

$$\frac{d}{dp}\left[\log(L_n(p))\right] = \sum_{i=1}^{n} \frac{2}{p} - \frac{(X_i - 2)}{1-p} = \frac{2n}{p} - \frac{\sum_{i=1}^{n}(X_i - 2)}{1-p}$$

   Igualando a cero y despejando $p$ obtenemos que

$$\boxed{\hat{p} = \frac{2}{\overline{X}_n}}$$

2. Usando el estimador de máxima verosimilitud hallado en la parte anterior tenemos que

$$\boxed{\hat{p} = \frac{2}{4{,}974} = 0{,}402}$$

3. Otra forma de estimar $p$ es usando la probabilidad de que $X$ valga 2. De la tabla podemos aproximar esta probabilidad por $74/500 = 0{,}148$. La probabilidad teórica se escribe en función de $p$ como $\mathbf{P}(X = 2) = p^2$, de donde

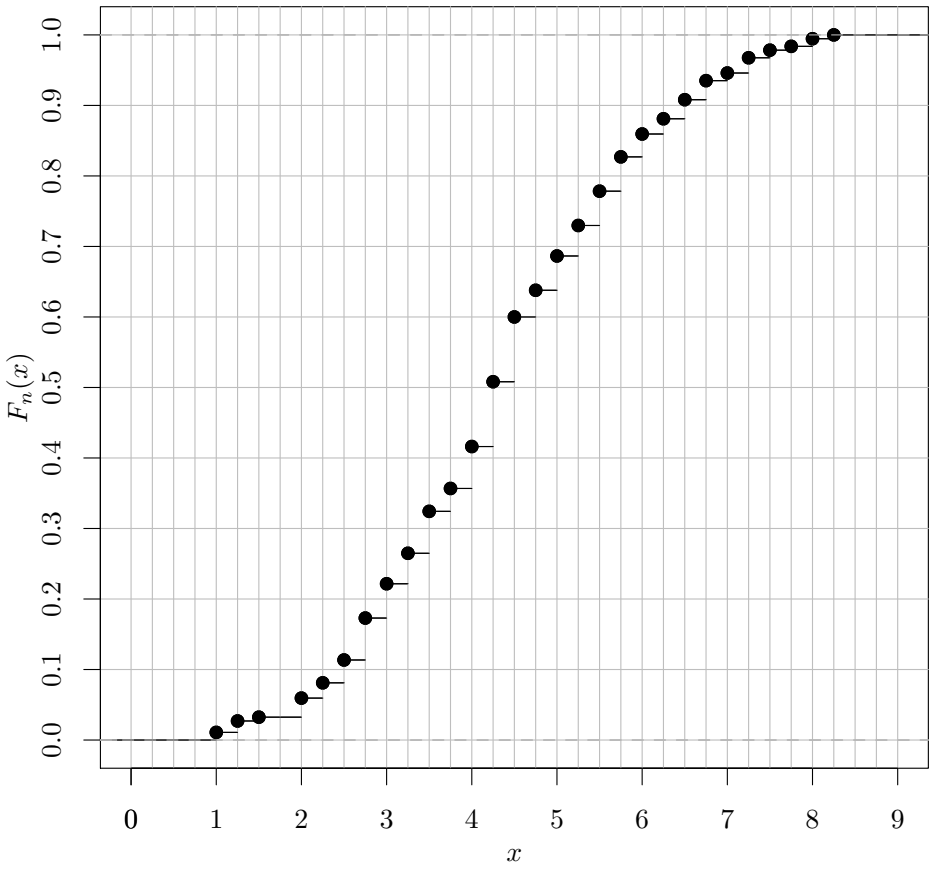
$$\boxed{\hat{p} = \sqrt{0{,}148} = 0{,}385}$$

   es otra estimación para $p$.

**Ejercicio 4.** [**25 puntos**]

Se consideran las 4 últimas cifras de un número de teléfono celular. Se asume que cada una de estas cifras puede ser modelada por una variable aleatoria uniforme en el conjunto $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Se asume además que las cifras son independientes entre sí. Para $i = 1, 2, 3, 4$ se define $X_i$ como la $i$-ésima cifra de las últimas cuatro: si el número es 099123456, entonces $X_1 = 3$, $X_2 = 4$, $X_3 = 5$ y $X_4 = 6$.

1. Se define $\bar{X}_4 = \frac{1}{4}\sum_{i=1}^{4} X_i$. Si utilizamos el teorema central del límite (TCL) para aproximar la distribución de $\bar{X}_4$ ¿qué distribución obtendríamos? ¿cuáles son sus parámetros?
2. Sea $X$ una variable aleatoria con la distribución hallada en la parte anterior. Hallar los cuantiles $x_p$ de $X$ para $p \in \{0{,}2; 0{,}4; 0{,}5; 0{,}6; 0{,}8; 0{,}9\}$.
3. A continuación se muestra la función de distibución empírica $F_n(x)$ de los promedios de las últimas cuatro cifras de $n = 185$ números de teléfono recolectados durante el último curso de PyE. Hallar los cuantiles empíricos $\hat{x}_p$ de dicha muestra para los mismos valores de $p$ que antes.




4. Graficar $(x_p, \hat{x}_p)$ para los valores de $p$ anteriores (graficar $x_p$ en el eje de las $x$ y $\hat{x}_p$ en el eje de las $y$). Si los puntos están cerca de la recta $y = x$ entonces podemos decir que hay un buen ajuste entre la distribución del promedio y su aproximación. ¿Cuál es su conclusión sobre el ajuste en este caso?

## Ejercicio 4. Solución

1. Dado que $X_1, X_2, X_3, X_4$ son variables aleatorias independientes e idénticamente distribuidas con valor esperado finito y varianza finita, el TCL permite aproximar $\bar{X}_4$ por una v.a. $X$ con distribución Normal de parámetros $\mu = \mathbf{E}\left(\bar{X}_4\right) = \mathbf{E}\left(X_1\right)$ y $\sigma^2 = \mathrm{Var}\left(\bar{X}_4\right) = \frac{\mathrm{Var}(X_1)}{4}$ donde,

- $\mathbf{E}(X_1) = \sum_{k=0}^{9} k\mathbf{P}(X_1 = k) = \sum_{k=0}^{9} k\frac{1}{10} = \frac{1}{10}\frac{9\times 10}{2} = 4{,}5$.
- $\mathbf{E}\left(X_1^2\right) = \sum_{k=0}^{9} k^2\mathbf{P}(X_1 = k) = \sum_{k=0}^{9} k^2\frac{1}{10} = \frac{285}{10} = 28{,}5$ y,
- $\mathrm{Var}(X_1) = \mathbf{E}\left(X_1^2\right) - \mathbf{E}(X_1)^2 = 28{,}5 - (4{,}5)^2 = 8{,}25$.

2. El objetivo de esta parte y la siguiente es ver si la aproximación por TCL es razonable aún en este caso en que el valor de $n = 4$ es pequeño. Sea $X$ es una variable aleatoria con distribución Normal de parámetros $\mu = 4{,}5$ y $\sigma^2 = \frac{8{,}25}{4} = 2{,}06$. Por ser $X$ absolutamente continua, tenemos que:

$$p = \mathbf{P}(X \leq x_p) = \mathbf{P}\left(\frac{X-\mu}{\sigma} \leq \frac{x_p - \mu}{\sigma}\right) = \Phi\left(\frac{x_p - \mu}{\sigma}\right)$$

donde en la última igualdad usamos que $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

Por lo tanto, resulta que $\frac{x_p-\mu}{\sigma} = z_p$ siendo $z_p$ el cuantil $p$ de una normal estándar, de dónde $x_p = \sigma z_p + \mu = 1{,}44 z_p + 4{,}5$. El cuantil $z_p$ se obtiene de la tabla de la normal estándar.

| $p$ | $z_p$ | $x_p$ |
|---|---|---|
| 0.2 | -0.84 | 3.28 |
| 0.4 | -0.26 | 4.12 |
| 0.5 | 0 | 4.5 |
| 0.6 | 0.26 | 4.87 |
| 0.8 | 0.84 | 5.71 |
| 0.9 | 1.29 | 6.38 |

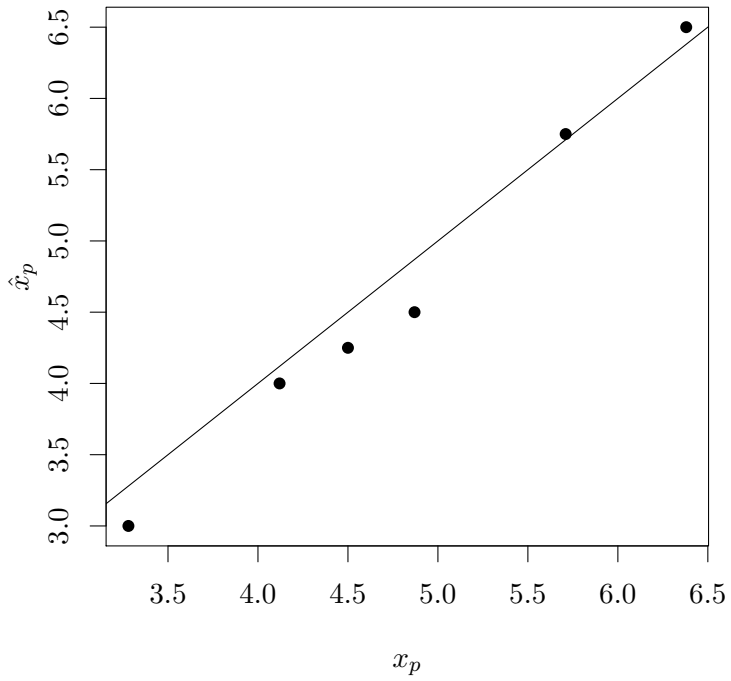
Obervar que $z_{1-p} = -z_p$ y que $z_{0,5} = 0$.

3. El cuantil empírico es $\hat{x}_p = \text{mín}\{x : F_n(x) \geq p\}$ y se obtiene de la gráfica dada:

| $p$ | $\hat{x}_p$ |
|---|---|
| 0.2 | 3 |
| 0.4 | 4 |
| 0.5 | 4.25 |
| 0.6 | 4.5 |
| 0.8 | 5.75 |
| 0.9 | 6.5 |

4. Vamos a graficar ahora los valores de $(x_p, \hat{x}_p)$:

| $p$ | $\hat{x}_p$ | $x_p$ |
|---|---|---|
| 0.2 | 3 | 3.28 |
| 0.4 | 4 | 4.12 |
| 0.5 | 4.25 | 4.5 |
| 0.6 | 4.5 | 4.87 |
| 0.8 | 5.75 | 5.71 |
| 0.9 | 6.5 | 6.38 |




Por lo tanto, vemos que los puntos están bastante cerca de la recta $y = x$ lo que indica un ajuste bueno de la distibución de $\bar{X}_4$ a la distribución dada por el TCL.