Examen de Introducción a la Econometría (LECO).

Departamento de Economía Cuantitativa. Universidad Complutense de Madrid.
16 de junio de 2008. Duración: **2 horas y 30 minutos**.

Apellidos:	Nombre:	DNI:
Nombre del profesor:		Grupo:

No desgrape las hojas de este cuadernillo. El examen está compuesto por tres problemas y 10 cuestiones cortas. Elija SÓLO DOS DE LOS TRES problemas propuestos y respóndalos en hojas aparte. Las cuestiones cortas deben ser respondidas en el espacio debajo del enunciado en este mismo cuardernillo. Junto a cada pregunta y apartado se indica la puntuación.

Problemas

Problema 1. (2.5 pt.) Se dispone de datos sobre la velocidad en km/h y la distancia recorrida por un automóvil hasta que se detiene (en metros) para cinco vehículos.

i	Vel. (x_i)	Dist. (y_i)
1	18	5
2	24	16
3	11	7
4	19	7
5	40	26
Suma	112	61

Con los datos de la tabla se han calculado los siguientes estadísticos: $nS_x^2 = 473{,}2$, $nS_y^2 = 310{,}8$ y $nS_{xy} = 357{,}6$. Se pide (realice los cálculos con dos decimales de precisión):

1. (0.25 pt) Formule un modelo teórico de relación lineal entre distancia de frenado y velocidad del coche.

   Suponemos que es la distancia hasta la detención la que depende de la velocidad de vehículo. Así, tenemos dos versiones de la relación:

   - $E(Y/X = x_i) = a + bx_i$
   - $y_i = a + bx_i + \varepsilon_i$, donde $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma_\varepsilon^2$, $\forall i$.

2. (0.25 pt) Estime el modelo por mínimos cuadrados. El estimador por el método de los momentos o de mínimos cuadrados del modelo anterior es:

   - $\hat{b} = \frac{S_{xy}}{S_x^2} = \frac{357{,}6/n}{473{,}2/n} = 0{,}7557.$
   - $\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{61}{5} - 0{,}7557\frac{112}{5} = -4{,}7278.$

El modelo estimado se puede expresar: $\widehat{E(Y/X = x_i)} = \hat{y}_i = -4{,}73 + 0{,}7557x_i$ o también $y_i = -4{,}73 + 0{,}7557x_i + \hat{\varepsilon}_i$.

3. (0.25 pt) Interprete los parámetros estimados.

   La pendiente de la recta estimada indica que por cada km/h más de velocidad, la distancia de frenado aumenta, por término medio, 0.76 metros. Esta relación se mantiene para velocidades entre 11 y 40 km/h. En este ejercicio, la constante no tiene una interpretación física, y sólo contribuye a un ajuste adecuado de la recta a los datos.

4. (0.25 pt) Calcule los residuos. ¿Qué vehículo frena peor?

   Los residuos se calculan: $\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (-4{,}73 + 0{,}7557x_i)$, $i = 1, \ldots, 5$. En la siguiente tabla aprecen los residuos y sus cuadrados (que se necesitan más adelante).

i	$\hat{\varepsilon}_i$	$\hat{\varepsilon}_i^2$
1	-3.87	15.01
2	2.59	6.71
3	3.42	11.66
4	-2.63	6.92
5	0.50	0.25
Suma	0	40.56

   El vehículo que "peor frena" es aquel que más supera la distancia de frenado esperada dada su velocidad o, de otra forma, el que presenta un residuo más grande y positivo, que resulta ser el coche $i = 3$.

5. (0.75 pt) Contraste la hipótesis nula de que la velocidad influye en la distancia de frenado. Exprese claramente la hipótesis nula, la alternativa y la región crítica del contraste.

   La hipótesis nula es que la velocidad no influye en la distancia, $H_0 : b = 0$, frente a la alternativa de que el efecto es positivo $H_1 : b > 0$. No tiene sentido considerar que la velocidad pudiera influir negativamente en la distancia de parada.

   El estadístico de contraste, ya particularizado en la nula, es $\frac{\hat{b}}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum(x_i-\bar{x})^2}}} \sim t_3$. Para calcular su valor es necesario obtener $\hat{\sigma}_\varepsilon^2$, ya se dispone del resto de componentes. El estimador de la varianza de las perturbaciones es $\hat{\sigma}_\varepsilon^2 = \frac{1}{3}\sum_{i=1}^5 \hat{\varepsilon}_i^2 = \frac{40{,}56}{3} = 13{,}52$.

   El valor del estadístico es $z = \frac{0{,}7557}{\sqrt{\frac{13{,}52}{473{,}2}}} = 4{,}47$, y la región crítica $\{z \geq c^*\}$. El valor crítico se obtiene de las tablas buscando $P[t_3 \geq c^*] = 0{,}05$, y resulta $c^* = 2{,}35$. Así, no se puede aceptar la nula, lo que significa que la velocidad influye positivamente en la distancia de frenado.

   El contraste también se podría haber realizado calculando el p-valor en vez del valor crítico, que en este caso es $\alpha^* = P[t_3 \geq 4{,}47] = 0{,}0104$.

6. (0.75 pt) Obtenga un intervalo de confianza al 95 % para el efecto de la velocidad sobre la distancia de frenado. ¿Qué interpretación tiene el intervalo calculado?

   Para obtener el intervalo de confianza usamos el procedimiento del estadístico pivote. El cálculo del intervalo de puede plantear:

$$P\left[-\lambda \leq \frac{\hat{b} - b}{\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum(x_i-\bar{x})^2}}} \leq \lambda\right] = 0{,}95.$$

El valor de λ se puede encontrar en las tablas de la distribución t_3 y resulta $\lambda = 3{,}18$. Obsérvese que este valor es diferente del valor crítico del apartado anterior porque ahora la signficación queda repartida en las dos colas.

Despejando, el intervalo de confianza resulta:

$$\left(\hat{b} - \lambda\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum(x_i - \bar{x})^2}}\ , \qquad \hat{b} + \lambda\sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum(x_i - \bar{x})^2}}\right),$$

que tiene un resultado numérico: (0,2178, 1,294). Esto es, el efecto de un km/h más de velocidad alarga la distancia de frentado entre 0,22 y 1,29 metros, con una confianza del 95 %. A la vista del intervalo, se habría aceptado la hipótesis nula de que $b = 1$.

Problema 2. (2.5 pt) En la centralita de una empresa hay un empleado que atiende las llamadas entrantes y que últimamente se siente saturado de trabajo. La empresa supone que las llamadas que entran por minuto siguen una distribución de Poisson. Se ha recogido, aleatoriamente, el número de llamadas que han entrado en la centralita por minuto en cien minutos diferentes. Los resultados aparecen en la siguiente tabla:

Llamadas	0	1	2	3	4	5	6	7	8
Veces	1	2	11	28	20	23	12	3	1

1. (1 pt) Contraste si **los datos se ajustan** a los supuestos de la empresa (las llamadas son una $\mathcal{P}(1)$). Formule claramente la hipótesis nula, la alternativa y halle el p-valor.

   Dado que la variable aleatoria cuya distribución se pretende contrastar es discreta (una Poisson), el contraste apropiado es el chi-cuadrado. El contraste de Kolmogorov-Smirnov, en este caso, podría ser aceptable puesto que la variable toma "bastantes" valores.

   Llamando X al número de llamadas que entran en la centralita por minuto, la hipótesis a contrastar es $H_0 : X \sim \mathcal{P}(1)$ frente a la alternativa de que X no sigue la distribución de la nula.

   Para llevar a cabo el contraste se necesitan las frecuencias esperadas (o teóricas), que se calculan $E_j = nP_j$, donde P_j es la probabilidad del intervalo (en este caso, la probabilidad de cada valor) suponiendo que la hipótesis nula es cierta. Utilizando la tabla de la $\mathcal{P}(1)$ se obtiene:

	0	1	2	3	4	5	6	7	8
P_j	0.37	0.37	0.18	0.06	0.02	0.00	0.00	0.00	0.00
E_j	36.79	36.79	18.39	6.13	1.53	0.31	0.05	0.01	0.00

   Como se puede apreciar, en varios intervalos la frecuencia esperada es inferior a cinco, por lo que es necesario agrupar varios (los valores 3 a 8). La tabla con las frecuencias observadas y esperadas para aplicar el contraste aparece a continuación:

Llamadas	0	1	2	3–8
O_j	1	2	11	87
E_j	36.79	36.79	18.39	8.53

El estadístico de contraste es:

$$Q = \sum_{j=1}^{4} \frac{(O_j - E_j)^2}{E_j} = \frac{(1 - 36{,}79)^2}{36{,}79} + \frac{(2 - 36{,}79)^2}{36{,}79} + \frac{(11 - 18{,}39)^2}{18{,}39} + \frac{(87 - 8{,}53)^2}{8{,}53} = 792{,}55,$$

que es necesario comparar con una χ^2_3, aunque obviamente se rechaza la hipótesis nula. El valor crítico del contraste para una significación del 5 % es $c^* = 7{,}81$ y el p-valor es $\alpha^* = 0$.

2. (1.5 pt) Realice un **contraste paramétrico** para la hipótesis nula de que el parámetro que caracteriza la entrada de llamadas es $\lambda = 2$ frente a la alternativa de que es mayor. Para ello:

   a) (0.5 pt) Encuentre la **mejor región crítica**.

   La mejor región crítica (la uniformemente más potente) se obtiene de la aplicación del teorema de Neyman-Pearson. En este ejercicio, la mejor región crítica viene dada por:

$$\frac{\pounds(\lambda_0)}{\pounds(\lambda_1)} = \frac{e^{-n\lambda_0}\lambda_0^{\sum x_i} / \prod x_i!}{e^{-n\lambda_1}\lambda_1^{\sum x_i} / \prod x_i!} = e^{-n(\lambda_0 - \lambda_1)} \left(\frac{\lambda_0}{\lambda_1}\right)^{\sum x_i}.$$

   Ahora es necesario transformar la región crítica en un suceso equivalente en función de un estadístico cuya distribución sea conocida bajo la nula:

$$\begin{aligned} e^{-n(\lambda_0-\lambda_1)} \left(\frac{\lambda_0}{\lambda_1}\right)^{\sum x_i} &\leq k, \\ -n(\lambda_0 - \lambda_1) + \log\left(\frac{\lambda_0}{\lambda_1}\right) \sum x_i &\leq k', \\ \log\left(\frac{\lambda_0}{\lambda_1}\right) \sum x_i &\leq k'', \\ \sum_{i=1}^{n} x_i &\geq k'''. \end{aligned}$$

   En el último paso se ha tenido en cuenta que $\log(\lambda_0/\lambda_1) < 0$ puesto que $0 < \lambda_0 < \lambda_1$. Como podemos encontrar la distribución del estadístico resultante ($\sum_{i=1}^n x_i \sim \mathcal{P}(n\lambda_0)$), el último paso ya contiene el estadístico y la región crítica.

   b) (0.5 pt) Halle el valor crítico a un 5 % de significación.

   Para hallar el valor crítico utilizamos la probabilidad de error de tipo I:

$$P\left[\sum_{i=1}^{n} x_i \geq c^* \Big/ \sum_{i=1}^{n} x_i \sim \mathcal{P}(200)\right] = 0{,}05.$$

   El problema ahora es que no se dispone de tablas de la distribución $\mathcal{P}(200)$. Dado que se trata de la suma de un número grande de variables aleatorias (que además son independientes e idénticamente distribuidas), se puede utilizar el teorema central del límite:

$$\frac{\sum_{i=1}^n x_i - E(\sum_{i=1}^n x_i)}{\sqrt{V(\sum_{i=1}^n x_i)}} = \frac{\sum_{i=1}^n x_i - 200}{\sqrt{200}} \to N(0,1),$$

   y la probabilidad de error tipo I resulta:

$$P\left[\frac{\sum_{i=1}^n x_i - 200}{\sqrt{200}} \geq \frac{c^* - 200}{\sqrt{200}} \Big/ \frac{\sum_{i=1}^n x_i - 200}{\sqrt{200}} \to N(0,1)\right] = 0{,}05.$$

   El valor de la normal estándar que deja a su derecha un 5 % de probabilidad es 1.6448, por lo que se puede despejar $c^* = 200 + 1{,}6448\sqrt{200} = 223{,}26$. Así, la región crítica pedida es $\{\sum_{i=1}^n x_i \geq 223{,}26\}$ o usando la aproximación normal $\left\{\frac{\sum_{i=1}^n x_i - 200}{\sqrt{200}} \geq 1{,}64\right\}$.

c) (0.5 pt) Aplique el contraste obtenido a los datos de la tabla.

Para aplicar el contraste es necesario calcular el estadístico para la muestra disponible:

$$\sum_{i=1}^{n} x_i = \sum_{j=0}^{8} j \cdot n_j = 0 \cdot 1 + 1 \cdot 2 + 2 \cdot 11 + \ldots + 7 \cdot 3 + 8 \cdot 1 = 404.$$

Puesto que el valor del estadístico es muy superior al valor crítico, no se puede aceptar la hipótesis nula. El empleado tiene razones objetivas para sentirse saturado de llamadas, puesto que entran, al menos, dos por minuto. De hecho, el estimador máximo verosímil del número esperado de llamadas por minuto es $\hat{\lambda} = \bar{x} = 4{,}04$.

Nota: La función de cuantía de una distribución $\mathcal{P}(\lambda)$ es $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots$

Problema 3. (2.5 pt) Sea una población X que se distribuye $N(\mu, 2)$. Se quiere llevar a cabo el contraste $H_0 : \mu = 0$ frente a $H_1 : \mu = 1$ con una significación $\alpha = 0{,}05$. Para ello se dispone de una muestra aleatoria simple de tamaño $n = 10$. Se consideran dos regiones críticas: $RC_1 = \{\bar{x} > k_1\}$ y $RC_2 = \{\frac{x_1+x_n}{2} > k_2\}$, donde $\bar{x}$ es la media muestral y x_1 y x_n son, respectivamente, el primer y el último elemento de la muestra.

1. (0.5 pt) Halle los valores de k_1 y k_2 para que ambas regiones críticas posean el mismo nivel de significación.

   Antes de calcular los valores críticos es necesario obtener la distribución de los estadísticos de contraste. Suponiendo una población general $X \sim N(\mu, \sigma^2)$, y una muestra aleatoria simple tamaño n, se obtiene que $\bar{x} \sim N(\mu, \sigma^2/n)$. Denotando por $\tilde{x} = \frac{x_1+x_n}{2}$ al segundo estadístico, resulta $E(\tilde{x}) = \mu$ y $V(\tilde{x}) = \sigma^2/2$, por lo que $\tilde{x} \sim N(\mu, \sigma^2/2)$.

   Para la primera región crítica, el valor crítico se obtiene de:

   $$\alpha = 0{,}05 = P\left[\bar{x} \geq k_1/\bar{x} \sim N(0, 2/10)\right] = P\left[\frac{\bar{x}}{\sqrt{2/10}} \geq \frac{k_1}{\sqrt{2/10}} \Big/ \frac{\bar{x}}{\sqrt{2/10}} \sim N(0,1)\right]$$

   y resulta $\frac{k_1}{\sqrt{2/10}} = 1{,}64$, por lo que $k_1 = 1{,}64\sqrt{2/10} = 0{,}74$. Siguiendo los mismos pasos para la segunda región crítica se tiene:

   $$\alpha = 0{,}05 = P\left[\tilde{x} \geq k_2/\tilde{x} \sim N(0, 2/2)\right] = P\left[\frac{\tilde{x}}{\sqrt{2/2}} \geq \frac{k_2}{\sqrt{2/2}} \Big/ \frac{\tilde{x}}{\sqrt{2/2}} \sim N(0,1)\right]$$

   y $k_2 = 1{,}64$.

   Las regiones críticas resultan $RC_1 = \{\bar{x} > 0{,}74\}$ y $RC_2 = \{\frac{x_1+x_n}{2} = \tilde{x} > 1{,}64\}$.

2. (0.5 pt) Dado que ambos contrastes tienen el mismo nivel de significación, ¿esto implica que ambos son igualmente adecuados para realizar el contraste?

   No. Los valores críticos se han elegido de forma que ambas regiones críticas tengan la misma significación, por lo que no se puede usar para compararlas. Para decidir qué región crítica es mejor será necesario calcular la potencia.

3. (1 pt) Calcule la potencia de ambos contrastes ¿Qué se puede decir acerca de la idoneidad de las regiones críticas?

La potencia de un contraste es $1-P(\text{error tipo II})$, y $P(\text{error tipo II}) = P(\text{Aceptar } H_0/H_0 \text{ falsa})$. Así, la potencia se puede calcular como $P(\text{Rechazar } H_0/H_0 \text{ falsa})$. Para la primera región crítica resulta:

$$\begin{aligned} w_1 &= P\left[\bar{x} \geq 0{,}74/\bar{x} \sim N(1, 2/10)\right] = P\left[\frac{\bar{x}-1}{\sqrt{2/10}} \geq \frac{0{,}74-1}{\sqrt{2/10}} \Big/ \frac{\bar{x}-1}{\sqrt{2/10}} \sim N(0,1)\right] \\ &= 1 - \Phi(-0{,}5912) = \Phi(0{,}5912) = 0{,}7228, \end{aligned}$$

donde $\Phi()$ es la función de distribución de la normal estánadar.

Siguiendo los mismos pasos para la segunda región crítica:

$$\begin{aligned} w_2 &= P\left[\tilde{x} \geq 1{,}64/\tilde{x} \sim N(1, 2/2)\right] = P\left[\frac{\tilde{x}-1}{\sqrt{2/2}} \geq \frac{1{,}64-1}{\sqrt{2/2}} \Big/ \frac{\tilde{x}-1}{\sqrt{2/2}} \sim N(0,1)\right] \\ &= 1 - \Phi(0{,}64) = 0{,}2611. \end{aligned}$$

La primera región crítica es claramente mejor que la segunda.

4. (0.5 pt) ¿Cómo cambiaría su respuesta a los apartados anteriores si aumentara el tamaño muestral a $n = 20$?

   En el primer apartado del ejercicio se mostró que la distribución del estadístico $\tilde{x}$ no dependende del tamaño de la muestra, por lo que tanto el valor crítico para una significación dada como la potencia no cambian al aumentar el tamaño muestral.

   Al contrario ocurre para la primera de las regiones críticas consideradas. El valor crítico para un tamaño muestral genérico n se obtiene de:

$$\alpha = 0{,}05 = P\left[\bar{x} \geq k_1/\bar{x} \sim N(0, 2/n)\right] = P\left[\frac{\bar{x}}{\sqrt{2/n}} \geq \frac{k_1}{\sqrt{2/n}} \Big/ \frac{\bar{x}}{\sqrt{2/n}} \sim N(0,1)\right],$$

   y resulta $\frac{k_1}{\sqrt{2/n}} = 1{,}64$, por lo que $k_1 = 1{,}64\sqrt{2/n}$, que tiende hacia cero a medida que aumenta el tamaño muestral. Si $n = 20$, el valor crítico es $k_1 = 1{,}64\sqrt{2/20} = 0{,}52$.

   La potencia en función de n es:

$$\begin{aligned} w_1(n) &= P\left[\bar{x} \geq k_1/\bar{x} \sim N(1, 2/n)\right] = P\left[\frac{\bar{x}-1}{\sqrt{2/n}} < \frac{k_1-1}{\sqrt{2/n}} \Big/ \frac{\bar{x}-1}{\sqrt{2/n}} \sim N(0,1)\right] \\ &= 1 - \Phi\left(\frac{1{,}64\sqrt{2/n}-1}{\sqrt{2/n}}\right) = 1 - \Phi\left(1{,}64 - \sqrt{n/2}\right). \end{aligned}$$

   Es evidente que la potencia aumenta con el tamaño muestral. En el caso concreto de $n = 20$, resulta $w_1(20) = 1 - \Phi(1{,}64 - \sqrt{20/2}) = 0{,}9360$, mayor que la que se tenía con $n = 10$.

   Como comentario general, el estadístico $\tilde{x}$ no sólo es peor porque tiene menor potencia, sino también porque no mejora (no aumenta su potencia) cuando se dispone de más información (mayor tamaño muestral), cosa que cabe esperar de cualquier buen estadístico de contraste.

Cuestiones cortas

Debe obtener más de 2 puntos en estas preguntas para que se califiquen los problemas.

Enunciado para las tres cuestiones siguientes. Sea el par de variables aleatorias (X, Y), cuya distribución de probabilidad conjunta viene dada por la siguiente tabla:

	$X = 1$	$X = 2$	$X = 3$
$Y = 0$	$1/9$	$1/3$	$2/9$
$Y = 1$	$1/18$	$1/6$	$1/9$

Cuestión 1. (0.5 pt) Halle el coeficiente de correlación lineal ρ_{XY}.

El coeficiente de correlación lineal simple se define $\rho_{XY} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}}$, por lo que necesitamos los momentos de segundo orden, y para calcularlos, también los de primer orden. Las distribuciones marginales de X e Y se han añadido a la tabla de la función de cuantía conjunta. Sin embargo, una mínima precaución sugiere comprobar primero la posible independencia, porque si X e Y son independientes, entonces $\rho_{XY} = 0$.

	$X = 1$	$X = 2$	$X = 3$	$P(Y = y)$
$Y = 0$	$2/18$	$6/18$	$4/18$	$12/18$
$Y = 1$	$1/18$	$3/18$	$2/18$	$6/18$
$P(X = x)$	$3/18$	$9/18$	$6/18$	

Se puede comprobar en la tabla que $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$, $\forall x, y$, por lo tanto, $\rho_{XY} = 0$.

Cuestión 2. (0.5 pt) Calcule $E(Y/X = 2)$.

Puesto que las variables aleatorias son independientes, sabemos que $E(Y/X = 2) = E(Y)$, que resulta $E(Y) = \sum_{y=0}^{1} yP(Y = y) = 0\frac{12}{18} + 1\frac{6}{18} = \frac{1}{3}$.

Cuestión 3. (0.5 pt) Formule la función generatriz de momentos conjunta $M_{XY}(t, s)$.

Aplicando la definición:

$$\begin{aligned} M_{XY}(t,s) &= E(e^{tX+sY}) = \sum_{x=1}^{3}\sum_{y=0}^{1} e^{tx+sy}P(X = x, Y = y) \\ &= e^{1\cdot t+0\cdot s}\frac{2}{18} + e^{1\cdot t+1\cdot s}\frac{1}{18} + e^{2\cdot t+0\cdot s}\frac{6}{18} + e^{2\cdot t+1\cdot s}\frac{3}{18} + e^{3\cdot t+0\cdot s}\frac{4}{18} + e^{3\cdot t+1\cdot s}\frac{2}{18} \\ &= e^{t}\frac{2}{18} + e^{t+s}\frac{1}{18} + e^{2t}\frac{6}{18} + e^{2t+s}\frac{3}{18} + e^{3t}\frac{4}{18} + e^{3t+s}\frac{2}{18}. \end{aligned}$$

Cuestión 4. (0.5 pt) Demuestre que la suma de los datos de una muestra tamaño n en desviaciones a su media es cero; esto es, que $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0,$$

puesto que $n\bar{x} = \sum_{i=1}^{n} x_i$.

Cuestión 5. (0.5 pt) "El p-valor de un contraste de normalidad es 0,085. Esto quiere decir que para un nivel de significación del 0,05 no se puede aceptar la hipótesis nula de que la distribución es normal" ¿Es cierta la afirmación? Razone la respuesta.

No, la afirmación es falsa. Si el p-valor es mayor que la significación seleccionada esto implica que el estadístico de contraste es menor que el valor crítico, por lo que se acepta la hipótesis nula con la significación dada.

Para un p-valor de 0,085 se rechazaría la nula de normalidad con una significación del 0,10, pero se aceptaría con un 0,05.

Cuestión 6. (0.5 pt) De 2.000 encuestados, 1.420 son favorables a la retirada de los parquímetros y 580 no lo son. El Ayuntamiento manifiesta que "la mayoría de la población está a favor de los parquímetros". Contraste la veracidad de la afirmación del Ayuntamiento con un nivel de significación del $\alpha = 0{,}05$ **utilizando un contraste paramétrico**. Indique claramente hipótesis nula y alternativa.

Se supone que la opinión de un ciudadano sobre mantener los parquímetros es $X \sim B(1,p)$. En este caso, la hipótesis nula es $H_0 : p = 0{,}5$ frente a la alternativa que representa la opinión del Ayuntamiento $H_1 : p > 0{,}5$ (si la mayoría está a favor, la probabilidad de que un ciudadano quiera mantener los parquímetros debe ser mayor que 0,5).

Puesto que se dispone de muchos datos, se puede utilizar el estadístico $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1)$. La región crítica es $\{\frac{\hat{p}-0{,}5}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \geq c^*\}$.

A partir de los datos muestrales, $\hat{p} = \frac{580}{2000} = 0{,}29$ y el valor del estadístico es:

$$\frac{0{,}29 - 0{,}5}{\sqrt{\frac{0{,}29(1-0{,}29)}{2000}}} = -20{,}70.$$

Alternativamente, se podría haber utilizado $\frac{0{,}29-0{,}5}{\sqrt{\frac{0{,}5(1-0{,}5)}{2000}}} = -18{,}78$.

En ambos casos, con un nivel de significación del 5 %, el valor crítico es 1,64, por lo que claramente no se puede rechazar la hipótesis nula. De hecho, se aceptaría que a favor sólo se encuentra el 30 %.

Cuestión 7. (0.5 pt) Ilustre gráficamente y obtenga numéricamente el p-valor del contraste de la cuestión anterior.

El p-valor es el area bajo la curva normal estándar a la derecha de $z = -20{,}70$, que es la unidad.

Cuestión 8. (0.5 pt) Discuta la veracidad o falsedad de la siguiente afirmación: "Si se rechaza H_0 con un nivel de significación $\alpha = 0{,}05$ también se rechaza H_0 con un nivel de significación $\alpha = 0{,}10$".

La afirmación es cierta. Para rechazar al 0,05 el p-valor debe ser inferior a la significación ($\alpha^* < 0{,}05$), por lo que también es inferior a 0,10.

Cuestión 9. (0.5 pt) Suponga la siguiente función de densidad marginal de una variable aleatoria X: $f_1(x) = \frac{1}{x^2}, x \geq 1$. Compruebe si es función de densidad y calcule $E(X)$.

Es claro que la función propuesta toma valores no negativos en el soporte. La segunda condición para que sea función de densidad es que la integral en el soporte sea unitaria:

$$\int_1^\infty \frac{1}{x^2}dx = \left[-x^{-1}\right|_1^\infty = \left(\lim_{x\to\infty} -\frac{1}{x}\right) - (-1) = 1.$$

La esperanza es:

$$E(X) = \int_1^\infty x\frac{1}{x^2}dx = \int_1^\infty \frac{1}{x}dx = [\ln x|_1^\infty = \left(\lim_{x\to\infty} \ln x\right) - (-1) = \infty.$$

La variable aleatoria X no tiene esperanza finita, cosa que no debería sorprendernos porque no hay garantía de que existan todos los momentos de una variable aleatoria.

Cuestión 10. (0.5 pt) Sea $(X,Y)'$ una normal bivariante con vector de esperanzas $\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ y matriz de varianzas-covarianzas $\Sigma = \begin{pmatrix} 1 & \sigma_{XY} \\ \sigma_{XY} & 1 \end{pmatrix}$. ¿Cuánto es la $P(X-Y>0)$ si $\sigma_{XY}=0$? ¿Y si $\sigma_{XY}=0{,}5$?

Llamando $Z = X - Y$, Z es normal porque es combinación lineal de normales, con momentos $E(Z) = E(X) - E(Y) = -1$ y $V(Z) = V(X) + V(Y) - 2cov(X,Y) = 2 - 2\sigma_{XY}$; esto es, $Z \sim N(-1, 2(1-\sigma_{XY}))$. Las probabilidades pedidas son:

$$\begin{aligned} \text{Si } \sigma_{XY} = 0 \quad &\Rightarrow \quad P(Z>0) = P\left(\frac{Z-(-1)}{\sqrt{2}} > \frac{-(-1)}{\sqrt{2}}\right) = 1 - \Phi\left(\frac{1}{\sqrt{2}}\right) = 0{,}2398. \\ \text{Si } \sigma_{XY} = 0{,}5 \quad &\Rightarrow \quad P(Z>0) = P\left(\frac{Z-(-1)}{\sqrt{2\cdot 0{,}5}} > \frac{-(-1)}{\sqrt{2\cdot 0{,}5}}\right) = 1 - \Phi\left(1\right) = 0{,}1587, \end{aligned}$$

donde $\Phi()$ es la función de distribución de la normal estándar.

Fórmulas de posible utilidad

Transformación de variables. Sea $X \sim f_X(x)$ y se define $Y = h(X)$. Entoces $f_Y(y) = f_X\left(h^{-1}(y)\right)\left|\frac{dh^{-1}(y)}{dy}\right|$ donde $h^{-1}(\cdot)$ es la *función inversa* de $h(\cdot)$.

Aproximación lineal a la esperanza condicional.

$$E^*(Y/X=x) = E(Y) - E(X)\frac{Cov(X,Y)}{V(X)} + E(X)\frac{Cov(X,Y)}{V(X)}\cdot x$$

Varianza condicional de la normal bivariante. $V(Y/X=x) = V(Y)(1-\rho^2_{XY})$.

Modelo de regresión lineal. Sea $E(Y_i/X_i = x_i) = a + bx_i$ (o también $Y_i = a + bx_i + \varepsilon_i$, $\varepsilon_i \underset{iid}{\sim} N(0,\sigma^2_\varepsilon)$). Si $\hat{a}$ y $\hat{b}$ son los estimadores por el método de los momentos (o de mínimos cuadrados) de a y b, entonces:

$$\frac{\hat{a}-a}{\sqrt{\frac{\hat{\sigma}^2_\varepsilon \sum x_i^2}{T\sum(x_i-\bar{x})^2}}} \sim t_{n-2}; \qquad \frac{\hat{b}-b}{\sqrt{\frac{\hat{\sigma}^2_\varepsilon}{\sum(x_i-\bar{x})^2}}} \sim t_{n-2}; \qquad \hat{\sigma}^2_\varepsilon = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Distribuciones de funciones de variables aleatorias. Si $X \sim N(\mu,\sigma^2)$, $X_1 \sim N(\mu_1,\sigma_1^2)$ y $X_2 \sim N(\mu_2,\sigma_2^2)$ e independientes y se dispone de muestras de tamaños n, n_1 y n_2 respectivamente:

$$\frac{\bar{x}-\mu}{\sqrt{\sigma^2/n}} \sim N(0,1); \qquad \frac{\bar{x}-\mu}{\sqrt{s^2/n}} \sim t_{n-1}; \qquad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}; \qquad \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_s^2} \sim F_{n_1-1,n_2-1}$$

$$\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1); \qquad \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{n_1s_1^2+n_2s_2^2}{n_1+n_2-2}}} \sim t_{n+m-2}$$

donde s^2 denota la *cuasivarianza* muestral.

Proporciones. $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \to N(0,1)$. Con dos poblaciones y muestras de tamaños n y m:

$$\frac{(\hat{p_1}-\hat{p_2})-(p_1-p_2)}{\sqrt{\left(\frac{n+m}{n\cdot m}\right)\hat{p_T}(1-\hat{p_T})}} \to N(0,1),$$

donde $\hat{p_T} = \frac{n\hat{p_1}+m\hat{p_2}}{n+m}$.

Contraste de Jarque-Bera. $JB = n\left[\frac{AS^2}{6}+\frac{(K-3)^2}{24}\right] \to \chi^2_2$.

Contraste Chi cuadrado. $\sum_{i=1}^{k}\frac{(T_i-O_i)^2}{T_i} \sim \chi^2$ donde T_i y O_i son, respectivamente las i-ésimas frecuencias absolutas esperadas y observadas.

Contrastes de Kolmogorov-Smirnov. Para una muestra $D_n = \sup|F_n^*(x)-F(x)|$. Para dos muestras $D_{n,m} = \sup|F_n^*(x)-G_m^*(x)|$. $F_n^*(x)$ y $G_m^*(x)$ son funciones de distribución empíricas (o muestrales) y $F(x)$ es una función de distribución teórica.

Contraste de Wilcoxon. El estadístico $T = T^+ - T^-$, bajo H_0 cumple $E(T)=0$ y $V(T) = \frac{n(n+1)(2n+1)}{6}$.

Contraste de Mann-Whitney. $U = \text{mín}(U_1,U_2)$, donde $U_1 = n_1n_2+\frac{n_1(n_1+1)}{2}-R_1$ y $U_2 = n_1n_2+\frac{n_2(n_2+1)}{2}-R_2$. Bajo H_0 se cumple $E(U)=\frac{n_1n_2}{2}$ y $V(U)=\frac{n_1n_2(n_1+n_2+1)}{12}$.

Aproximación a los valores críticos en los contrastes de Kolmogorov-Smirnov. Para el contraste de una muestra, el valor crítico c^* con un nivel de significación α se aproxima mediante $c^*_\alpha = k_\alpha\sqrt{1/n}$, donde k_α es 1.07, 1.22, 1.36, 1.52 y 1.63 para niveles de significación del 20 %, 10 %, 5 %, 2 % y 1 %, respectivamente.

Para el contraste de dos muestras, el valor crítico aproximado se calcula:

$$c^*_\alpha = k_\alpha\sqrt{\frac{n_1+n_2}{n_1\cdot n_2}},$$

donde k_α es 1.07, 1.22 y 1.52 para niveles de significación α del 10 %, 5 % y 1 %, respectivamente.

Tablas estadísticas

	x.xx	x.x1	x.x2	x.x3	x.x4	x.x5	x.x6	x.x7	x.x8	x.x9
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	0.8413	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	0.9772	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	0.9987	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

Cuadro 1: Función de distribución de la $N(0,1)$

r	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.999
2	0.45	0.71	1.02	1.39	1.83	2.41	3.22	4.61	5.99	7.38	9.21	13.82
4	1.65	2.19	2.75	3.36	4.04	4.88	5.99	7.78	9.49	11.14	13.28	18.47
6	3.07	3.83	4.57	5.35	6.21	7.23	8.56	10.64	12.59	14.45	16.81	22.46
8	4.59	5.53	6.42	7.34	8.35	9.52	11.03	13.36	15.51	17.53	20.09	26.12
10	6.18	7.27	8.30	9.34	10.47	11.78	13.44	15.99	18.31	20.48	23.21	29.59
12	7.81	9.03	10.18	11.34	12.58	14.01	15.81	18.55	21.03	23.34	26.22	32.91
14	9.47	10.82	12.08	13.34	14.69	16.22	18.15	21.06	23.68	26.12	29.14	36.12
16	11.15	12.62	13.98	15.34	16.78	18.42	20.47	23.54	26.30	28.85	32.00	39.25
18	12.86	14.44	15.89	17.34	18.87	20.60	22.76	25.99	28.87	31.53	34.81	42.31
20	14.58	16.27	17.81	19.34	20.95	22.77	25.04	28.41	31.41	34.17	37.57	45.31

Cuadro 2: Función de distribución de la χ^2_r.

r	0.6	0.7	0.8	0.9	0.95	0.975	0.99	0.999
1	0.3249	0.7265	1.3764	3.0777	6.3138	12.7062	31.8205	318.3088
2	0.2887	0.6172	1.0607	1.8856	2.9200	4.3027	6.9646	22.3271
3	0.2767	0.5844	0.9785	1.6377	2.3534	3.1824	4.5407	10.2145
4	0.2707	0.5686	0.9410	1.5332	2.1318	2.7764	3.7469	7.1732
5	0.2672	0.5594	0.9195	1.4759	2.0150	2.5706	3.3649	5.8934
6	0.2648	0.5534	0.9057	1.4398	1.9432	2.4469	3.1427	5.2076
7	0.2632	0.5491	0.8960	1.4149	1.8946	2.3646	2.9980	4.7853
8	0.2619	0.5459	0.8889	1.3968	1.8595	2.3060	2.8965	4.5008
9	0.2610	0.5435	0.8834	1.3830	1.8331	2.2622	2.8214	4.2968
10	0.2602	0.5415	0.8791	1.3722	1.8125	2.2281	2.7638	4.1437
11	0.2596	0.5399	0.8755	1.3634	1.7959	2.2010	2.7181	4.0247
12	0.2590	0.5386	0.8726	1.3562	1.7823	2.1788	2.6810	3.9296
13	0.2586	0.5375	0.8702	1.3502	1.7709	2.1604	2.6503	3.8520
14	0.2582	0.5366	0.8681	1.3450	1.7613	2.1448	2.6245	3.7874
15	0.2579	0.5357	0.8662	1.3406	1.7531	2.1314	2.6025	3.7328
16	0.2576	0.5350	0.8647	1.3368	1.7459	2.1199	2.5835	3.6862
17	0.2573	0.5344	0.8633	1.3334	1.7396	2.1098	2.5669	3.6458
18	0.2571	0.5338	0.8620	1.3304	1.7341	2.1009	2.5524	3.6105
19	0.2569	0.5333	0.8610	1.3277	1.7291	2.0930	2.5395	3.5794
20	0.2567	0.5329	0.8600	1.3253	1.7247	2.0860	2.5280	3.5518

Cuadro 3: Función de distribución de la t_r.

							x						
λ	0	1	2	3	4	5	6	7	8	9	10	11	12
1	0.3679	0.7358	0.9197	0.9810	0.9963	0.9994	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
2	0.1353	0.4060	0.6767	0.8571	0.9473	0.9834	0.9955	0.9989	0.9998	1.0000	1.0000	1.0000	1.0000
3	0.0498	0.1991	0.4232	0.6472	0.8153	0.9161	0.9665	0.9881	0.9962	0.9989	0.9997	0.9999	1.0000
4	0.0183	0.0916	0.2381	0.4335	0.6288	0.7851	0.8893	0.9489	0.9786	0.9919	0.9972	0.9991	0.9997
5	0.0067	0.0404	0.1247	0.2650	0.4405	0.6160	0.7622	0.8666	0.9319	0.9682	0.9863	0.9945	0.9980

Cuadro 4: Función de distribución de la $\mathcal{P}(\lambda)$.