Nombre: _____

Examen final de ESTADÍSTICA APLICADA (21 de junio de 2019)
Primer curso del grado en Bioquímica

EXPLICAR DEBIDAMENTE LA RESOLUCIÓN DE CADA PROBLEMA

1. (1.5 puntos) Una ingeniera genética hace cruces de tigre y guepardo. En la primera generación del cruce conjetura los siguientes ratios fenotípicos:

Sólo rayas	4
Sólo puntos	3
Rayas y puntos	9
Total	16

$p_1 = \frac{4}{16} = 0.25$

$p_2 = 3/16 = 0.1875$

$p_3 = 9/16 = 0.5625$

Al realizar los cruces en la práctica observó las siguientes frecuencias en la primera generación:

Sólo rayas	50 $= O_1$
Sólo puntos	41 $= O_2$
Rayas y puntos	85 $= O_3$
Total	176 $= n = O_1 + O_2 + O_3$

A un nivel del 5 %, utiliza un contraste χ^2 para decidir si hay suficiente evidencia muestral en contra de la previsión de la investigadora.

Contraste χ^2 de bondad de ajuste:

H_0: $p_1 = P(\text{sólo rayas}) = \frac{1}{4}$, $p_2 = P(\text{sólo puntos}) = \frac{3}{16}$, $p_3 = P(\text{rayas y puntos}) = 9/16$

$\alpha = 0.05$

H_1: El modelo de H_0 no se ajusta bien a la muestra.

Frecuencias absolutas esperadas bajo H_0:

$e_1 = 176 \cdot \frac{1}{4} = 44 \qquad e_2 = 176 \frac{3}{16} = 33 \qquad e_3 = 99$

Estadístico del contraste:

$$\chi^2 = \frac{O_1^2}{e_1} + \frac{O_2^2}{e_2} + \frac{O_3^2}{e_3} - n = \frac{50^2}{44} + \frac{41^2}{33} + \frac{85^2}{99} - 176 =$$

$$= 180.74 - 176 = 4.74$$

Región de rechazo: $R = \{\chi^2 > \chi^2_{3-1;0.05} = \chi^2_{2;0.05} = 5.99\}$

No hay suficiente evidencia para rechazar H_0. El modelo conjeturado por la investigadora se ajusta bien a los datos.

X = resultado final de la batería de tests $\quad n = 12$

$\bar{x} = 63 \quad s = 17$

Como n es "pequeño" y no podemos aplicar el TCL, en (a), (b) y (c) suponemos que $X \sim N(\mu, \sigma)$ con μ y σ desconocidos.

2. En un estudio acerca de los efectos de la falta de sueño sobre las capacidades cognitivas, 12 estudiantes de universidad privados de sueño durante una noche realizaron al día siguiente una batería de tests con el objetivo de medir su rapidez mental, fluidez, flexibilidad y originalidad. Los resultados finales de estos estudiantes tuvieron una media muestral de 63 y una desviación típica muestral de 17. Una puntuación más baja se interpreta como una capacidad disminuida para pensar creativamente.

a) (0.5 puntos) Construir un intervalo de confianza al 90 % para la puntuación esperada de un sujeto que haya sido privado del sueño en la noche anterior a los tests. Indicar las suposiciones previas necesarias para la resolución del apartado.

b) (0.5 puntos) Construir un intervalo de confianza al 90 % para la varianza de la puntuación de un sujeto que haya sido privado del sueño en la noche anterior a los tests. Indicar las suposiciones previas necesarias para la resolución del apartado.

c) (1 punto) La puntuación esperada de un individuo que haya dormido adecuadamente antes de los tests es 80. A un nivel de significación del 10 %, ¿hay suficiente evidencia muestral para afirmar que dejar de dormir la noche anterior a las pruebas empeora en media la puntuación? Indicar las suposiciones previas necesarias para la resolución del apartado.

d) (0.5 puntos) En el contraste de **c)** di razonadamente todo lo que puedas acerca del p-valor del contraste, por ejemplo, si es mayor o no que 0.05, 0.1, etc. Concluye si es razonable o no rechazar la hipótesis nula.

a) $IC_{90\%}(\mu) = \left(63 \mp 1.796 \frac{17}{\sqrt{12}}\right) = (63 \mp 8.81) = (54.19, 71.81)$

$t_{11;0.05} = 1.796$

b) $IC_{90\%}(\sigma^2) = \left(\frac{11 \cdot 17^2}{19.68}, \frac{11 \cdot 17^2}{4.57}\right) = (161.53, 695.62)$

$\chi^2_{11;0.05} = 19.68 \quad \chi^2_{11;0.95} = 4.57$

c) $H_0: \mu \geq 80$
$H_1: \mu < 80 \quad \alpha = 0.1 \quad R = \{t < t_{11;0.9} = -t_{11;0.1} = -1.363\}$

$t = \frac{63 - 80}{17/\sqrt{12}} = -3.46 \Rightarrow$ Rechazamos H_0 al nivel del 10%

d) $\alpha = 0.05 \longrightarrow t_{11;0.95} = -1.796 > t \longrightarrow$ Rechazamos H_0 al 5%

$\alpha = 0.01 \longrightarrow t_{11;0.99} = -2.718 > t \longrightarrow$ " " " 1%

$\alpha = 0.005 \longrightarrow t_{11;0.995} = -3.106 > t \longrightarrow$ " " " 0.5%

$\alpha = 0.0005 \longrightarrow t_{11;0.9995} = -4.437 > t \longrightarrow$ No " " " 0.05%

$0.0005 < p\text{-valor} < 0.005$

Como el p-valor es menor que los niveles de significación habituales, lo razonable es rechazar H_0.

3. Se mide el grado X de expresión de un gen en el tejido ovárico de 23 mujeres sanas[1], obteniéndose los siguientes datos (ordenados de menor a mayor):

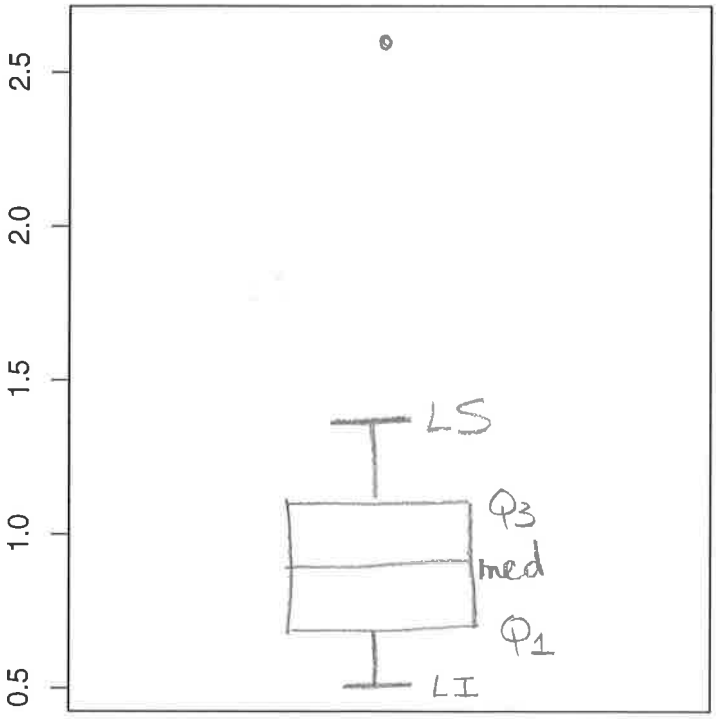
$x_{(1)}$ 0.51	0.52	0.62	0.67	0.67	0.70	0.76	0.76	0.79	0.81	0.81	0.84
0.89	0.94	1.01	1.09	1.15	1.15	1.16	1.27	1.35	1.37	2.63	

$x_{(13)}$ $x_{(14)}$ $x_{(15)}$ $x_{(16)}$ $x_{(17)}$ $x_{(18)}$ $x_{(19)}$ $x_{(20)}$ $x_{(21)}$ $x_{(22)}$ $x_{(23)}$

Para estos datos se obtiene $\sum_{i=1}^{23} x_i = 22.4736$ y $\sum_{i=1}^{23} x_i^2 = 26.2135$.

a) (1.2 puntos) Especificando claramente las fórmulas empleadas, calcular la media, la varianza, la desviación típica, el coeficiente de variación, la mediana, los cuartiles, el rango y el rango intercuartílico de los datos.

b) (0.8 puntos) En el siguiente gráfico, realiza un diagrama de caja de los datos, especificando claramente cómo se determina cada parte del diagrama.



c) (1.25 puntos) Se mide el grado de expresión del mismo gen en el tejido ovárico de 30 mujeres con cáncer de ovario, obteniéndose la siguiente muestra:

0.81	0.70	0.64	0.67	0.60	0.42	0.70	0.55	0.98	1.10
0.69	0.34	0.60	0.49	1.19	0.87	2.33	1.16	0.50	0.95
0.81	2.78	1.25	0.69	1.03	0.69	0.57	0.72	0.72	0.94

cuya media y desviación típica son 0.88 y 0.51 respectivamente. A nivel $\alpha = 0.05$, ¿hay suficiente evidencia muestral de que el nivel esperado de expresión de ese gen es diferente en mujeres sanas y en pacientes con cáncer de ovario?

[1]Fuente de los datos: Pepe *et al.* (2003). Selecting Differentially Expressed Genes from Microarray Experiments. *Biometrics*, 59,133–142.

$Z \sim N(0,1)$

X

4. Diversos estudios han establecido que un nivel de hemoglobina glicosilada por debajo de 7 reduce notablemente el riesgo de sufrir enfermedades cardiovasculares. Se supone que en cierta población este nivel sigue una distribución normal de media 6 y desviación típica 2.

a) (0.5 puntos) Calcular el porcentaje aproximado de la población cuyo nivel de hemoglobina glicosilada está por encima del valor de riesgo 7.

b) (1 punto) Si se seleccionan aleatoria e independientemente 100 personas en esta población, calcular la probabilidad aproximada de que más de 35 tengan un nivel de hemoglobina glicosilada por encima de 7.

c) (1 punto) Si se seleccionan aleatoria e independientemente 10 personas en esta población, calcular la probabilidad de que como mucho 2 tengan un nivel de hemoglobina glicosilada por encima de 7.

d) (0.25 puntos) Si se seleccionan aleatoria e independientemente dos personas en esta población, determinar la distribución de probabilidad de la suma de sus dos niveles de hemoglobina glicosilada.

a) $P\{X>7\} = P\left\{Z > \frac{7-6}{2}\right\} = P\{Z>0.5\} = 0.3085$

b) $Y = $ nº de personas de esas 100 que tienen un nivel de hemoglobina glicosilada por encima de 7

$Y \sim B(100, 0.3085) \underset{TCL}{\approx} N(30.85, \sqrt{100 \cdot 0.3085(1-0.3085)} \approx 4.62)$

$P\{Y>35\} = P\{Y>35.5\} \approx P\left\{Z > \frac{35.5-30.85}{4.62}\right\} =$

$= P\{Z>1.01\} = 0.1562$

c) $Y = $ nº de personas de esas 10 que tienen un nivel de hemoglobina glicosilada por encima de 7

$P\{Y \leq 2\} = (1-0.3085)^{10} + 10\, 0.3085\, (1-0.3085)^9 +$

$+ \underbrace{\binom{10}{2}}_{45} 0.3085^2 (1-0.3085)^8 =$

$= 0.3604$

d) $X_1 \sim N(6,2)$ $\quad X_2 \sim N(6,2)$ independientes

$\underbrace{X_1 + X_2}_{\text{suma de 2 normales independientes}} \sim N(12, \sqrt{2^2+2^2} = 2\sqrt{2})$

(3) a) $\bar{x} = \frac{1}{23}\sum_{i=1}^{23} x_i = \frac{22.4736}{23} = 0.98$

s_x^2 = varianza muestral $= \frac{1}{22}\sum_{i=1}^{23} x_i^2 - \frac{23}{22}\bar{x}^2 = \frac{1}{22}\left(26.2135 - 23 \cdot 0.98^2\right)$

$= 0.19$

$s_x = \sqrt{0.19} = 0.43$

$cv = \frac{s_x}{|\bar{x}|} = \frac{0.43}{0.98} = 0.44$

$med = x_{(12)} = 0,84$

$Q_1 = 0.5\, x_{(6)} + 0.5\, x_{(7)} = 0.5\ 0.70 + 0.5\ 0.76 = 0.73$

$0.25 \cdot 22 + 1 = 6.5$

$Q_3 = 0.5\, x_{(17)} + 0.5\, x_{(18)} = 0.5\ 1.15 + 0.5\ 1.15 = 1.15$

$0.75 \cdot 22 + 1 = 17.5$

rango $= x_{(23)} - x_{(1)} = 2.63 - 0.51 = 2.12$

$RI = Q_3 - Q_1 = 1.15 - 0.73 = 0.42$

b) $Q_1 - 1.5\, RI = 0.1$

$Q_3 + 1.5\, RI = 1.78$

El único dato que se sale del intervalo $[0.1, 1.78]$ es $x_{(23)} = 2.63$, que es un dato atípico, por tanto, entonces $LI = x_{(1)} = 0.51$ y $LS = x_{(22)} = 1.37$ son la menor y mayor observaciones respectivamente que caen dentro de $[0.1, 1.78]$

c) Denotamos por Y la expresión del gen en el tejido ovárico de una mujer con cáncer de ovario. Si denotamos

$\mu_1 = E(X)$ $\mu_2 = E(Y)$ $V(X) = \sigma^2 = V(Y)$

queremos contrastar $H_0: \mu_1 = \mu_2$ frente a $H_1: \mu_1 \neq \mu_2$.

Suponiendo que $X \sim N(\mu_1, \sigma)$ e $Y \sim N(\mu_2, \sigma)$ son independientes la región de rechazo del contraste es $R = \{|t| > t_{51;0.025}\}$ con

$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.98 - 0.88}{\sqrt{0.23}\sqrt{\frac{1}{23} + \frac{1}{30}}} = 0.75$

$t_{51;0.025} \simeq \frac{t_{40;0.025} + t_{60;0.025}}{2}$

$= \frac{2.021 + 2.000}{2}$

$= 2.01$

$s_p^2 = \frac{22\ 0.19 + 29\ 0.51^2}{51} = 0.23$

No hay evidencia para rechazar H_0.