1	2	3	4	Calificación

Probabilidad y Estadística (C)

Recuperatorio del segundo parcial - 2C 2018 - 19/12/2018

Complete esta hoja y entréguela con el resto del examen. ***Realizar cada ejercicio en hoja separada****. Escribir el nombre en cada una. Al retirarse debe firmar una hoja de asistencia.*

Apellido y nombres:.................................... Nº de libreta: ....

mail:............................@.................... Firma:......................

Turno: | Tarde: 14 a 17 hs | Noche: 19 a 22 hs | Nº de hojas entregadas:

Criterio de aprobación: Para aprobar este examen es necesario sumar al menos 60 puntos.

Recuerde definir con palabras los eventos y/o las variables aleatorias involucradas, nombres y parámetros de las distribuciones. ***Justifique claramente sus afirmaciones****.*

1. Un juego de preguntas y respuestas consiste en responder 50 cuestionarios que contienen dos preguntas cada uno. Cada cuestionario es independiente de los otros. Sean X e Y el resultado de acertar la primera y segunda pregunta del cuestionario respectivamente. Es decir X (resp. Y) vale 1 si el jugador acierta la primera (resp. la segunda) pregunta y 0 en caso contrario. Sea $Z = X + Y$ el resultado final del cuestionario. Se sabe que la probabilidad conjunta de X e Y esta dada por:

$p_{X,Y}(x,y)$	$Y = 0$	$Y = 1$
$X = 0$	0.1	0.4
$X = 1$	0.4	0.1

a) (7 pts.) Hallar la covarianza y la correlación de X e Y. ¿Es consistente esto con la afirmación de que una de las dos preguntas es fácil y la otra es dificil?

b) (7 pts.) Calcular la esperanza y varianza de Z.

c) (7 pts.) Dar una cota inferior para la probabilidad de obtener entre 40 y 60 puntos en el juego.

d) (4 pts.) Aproximar la probabilidad de obtener más de 55 puntos.

2. (25 p) Sea $X_1, \ldots, X_n \overset{iid}{\sim} LN(\mu, \sigma^2)$ una muestra aleatoria log-normal, es decir $Y_i = \ln(X_i) \sim N(\mu, \sigma^2)$.

a) (5 p) Hallar los estimadores de μ y de σ^2 basados en los dos primeros momentos de X. Sugerencia: usar la función generadora de momentos de una normal.

b) (4 p) Hallar la densidad de $X \sim LN(\mu, \sigma^2)$.

c) (4 p) Calcular $\ell_{LN}(\mu, \sigma^2)$ la función de log-verosimilitud basado en $X_1, ..., X_n$.

d) (3 p) Calcular $\ell_N(\mu, \sigma^2)$ la función de log-verosimilitud basado en $Y_1, ..., Y_n$.

e) (5 p) Deducir que los estimadores de máxima verosimilitud de una muestra de μ y σ^2 para una muestra log-normal cuando ambos parámetros son desconocidos son:

$$\widehat{\mu}_{ML} = \frac{\sum_{i=1}^n \ln(X_i)}{n}, \qquad \widehat{\sigma^2}_{ML} = \frac{\sum_{i=1}^n (\ln(X_i) - \widehat{\mu}_{ML})^2}{n}$$

Aclaración: puede usar sin demostrar cuales son los estimadores de máxima verosimilitud de μ y σ^2 de una muestra normal.

f) (4 p) La distribución log-normal se usa (entre otras cosas) como un modelo paramétrico de la distribución del ingreso. Se sabe que el coeficiente Gini se puede calcular en términos de los parámetros de acuerdo a la fórmula $G = P(|Z| \leq \frac{\sigma}{\sqrt{2}})$ donde $Z \sim N(0,1)$. Hallar el estimador de máxima verosimilitud de G. Si se obtiene en la muestra una estimación de $\hat{\sigma}^2(\underline{x}) = 0.75$, calcular la estimación del Gini de máxima verosimilitud.

3. El ministerio de medio ambiente desea estimar la contaminación radioactiva de una empresa que trabaja con materiales radioactivos. Para esto se mide con un contador Geiger que cuenta una cantidad de partículas radioactivas por minuto. La cantidad de partículas que cuenta por minuto tiene distribución Poisson, Se realizan 40 mediciones independientes en la planta y alrededores y se obtiene una media muestral de 5.8 partículas por minuto.

a) (10 pts.) Construir un intervalo de confianza de nivel aproximado $1 - \alpha = 0.95$, y calcular el intervalo obtenido en base a la muestra.

b) (7 pts.) Si en condiciones normales la cantidad de particulas por minuto es 5, ¿Hay evidencia para afirmar que la empresa está contaminando el medio ambiente?

c) (5 pts.) Construir otro intervalo distinto al del item a) si conocemos el desvío muestral y es $s = 2.3$.

d) (3 pts.) ¿Es correcto decir que la probabilidad de que el intervalo obtenido en base a la muestra contenga a λ es aproximadamente 0.95? Justifique.

4. En el pasado, una planta química ha producido un promedio de 1100 kilogramos de productos químicos al día. Los registros de este año, con base en 260 días de operación, muestran lo siguiente:

$$\overline{y} = 1060 \text{ kilogramos/día} \qquad s = 340 \text{ kilogramos/día}$$

Deseamos probar si el promedio de producción diaria bajó sensiblemente el año pasado.

a) (4 pts.) Formule las hipótesis nula y alternativa apropiadas.

b) (8 pts.) Dar un test apropiado con nivel de significación aproximado 0.05. Dar el estadístico del test, su distribución bajo H_0 y determine la región de rechazo correspondiente.

c) (3 pts.) ¿Los datos aportan suficiente evidencia para indicar una caída en el promedio de producción diaria con un nivel de significación aproximado 0.05?

d) (4 pts.) Calcule el p-valor obtenido.

e) (6 pts.) Si ahora el gerente de la empresa decide que la produción bajó sensiblemente cuando la producción es menor a 1070 kilogramos/día, y además se sabe que $\sigma = 300$ kilogramos/día, plantear el nuevo test que propone gerente y calcular su nivel de significación.

Solución:

1. a) $\text{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0.1 - (0.5)^2 = -0.15$, $\rho = \frac{-0.15}{0.25} = -0.6$. Sí, es consistente pues la correlación negativa me dice que si respondo correctamente una (X sube o viceversa), es muy poco probable que conteste bien la otra (Y baja o viceversa).

   b) $E(Z) = E(X+Y) = E(X) + E(Y) = \frac{1}{2} + \frac{1}{2} = 1$. $V(Z) = V(X) + V(Y) + 2\text{Cov}(X,Y) = 0.5 - 0.3 = 0.2$.

   c) Sea $S_{50} = \sum_{i=1}^{50} Z_i$, entonces

$$P(40 \leq S_{50} \leq 60) = P(|S_{50} - 50| \leq 10) = 1 - P(|S_{50} - 50| \geq 11) \geq 1 - \frac{50.0.2}{11^2} \approx 1 - 0.082 = 0.918$$

.

d) (Con corrección por continuidad) $P(S_{50} \geq 55) = P(S_{50} \geq 54.5) = P(\frac{S_{50}-50}{\sqrt{50.0.2}} \geq \frac{54.5-50}{50.0.2}) \approx 1 - 0.922 = 0.078$. (Sin corrección da ≈ 0.057).

2. a) $E(X) = E(e^{\ln(X)}) = e^{\mu+\frac{\sigma^2}{2}}$, $E(X^2) = E(e^{2\ln(X)}) = e^{2\mu+\frac{\sigma^2}{2}2^2} = e^{2\mu+2\sigma^2}$, por el método de momentos reemplazo los momentos muestrales,

$$\overline{X} = e^{\hat{\mu}+\frac{\hat{\sigma}^2}{2}}, \quad \overline{X^2} = e^{2\hat{\mu}+2\hat{\sigma}^2}$$

$$\ln(\overline{X}) = \hat{\mu} + \frac{\hat{\sigma}^2}{2}, \quad \ln(\overline{X^2}) = 2\hat{\mu} + 2\hat{\sigma}^2$$

$$\hat{\mu} = 2\ln(\overline{X}) - \frac{1}{2}\ln(\overline{X^2}), \quad \hat{\sigma}^2 = \ln(\overline{X^2}) - 2\ln(\overline{X})$$

b) Si $x > 0$, $F_X(x) = P(X \leq x) = P(\ln(X) \leq \ln(x)) = F_{\ln(X)}(\ln(x))$, luego derivando

$$f_X(x) = f_{\ln(X)}(\ln(x)).\frac{1}{x}\mathbb{1}_{(0,+\infty)}(x) = \frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}\mathbb{1}_{(0,+\infty)}(x)$$

c) $\mathcal{L}_{LN}(\mu, \sigma^2, \underline{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma x_i}e^{-\frac{(\ln(x_i)-\mu)^2}{2\sigma^2}}\mathbb{1}_{(0,+\infty)}(x_i)$, luego

$$\ell_{LN}(\mu, \sigma^2, \underline{x}) = -\sum_{i=1}^{n}\ln(x_i) - n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln(x_i)-\mu)^2 + \sum_{i=1}^{n}\ln(\mathbb{1}_{(0,+\infty)}(x_i))$$

d) $\mathcal{L}_N(\mu, \sigma^2, \underline{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}$, luego

$$\ell_N(\mu, \sigma^2, \underline{y}) = -n\ln(\sigma) - n\ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

e) Observar que dados $\underline{x} = x_1, ..., x_n$, sean $\underline{y} = y_1, ..., y_n$ sus logaritmos, entonces

$$\arg\max_{\mu,\sigma^2} \ell_{LN}(\mu, \sigma^2, \underline{x}) = \arg\max_{\mu,\sigma^2} \ell_N(\mu, \sigma^2, \underline{y})$$

ya que estas dos funciones solo difieren en una constante de la muestra. Como conocemos los estimadores de máxima verosimilitud de una normal, se tiene que

$$\widehat{\mu}_{ML} = \frac{\sum_{i=1}^{n}\ln(X_i)}{n}, \qquad \widehat{\sigma^2}_{ML} = \frac{\sum_{i=1}^{n}(\ln(X_i) - \widehat{\mu}_{ML})^2}{n}$$

f) Por la propiedad de invariancia del MLE, $\widehat{G}_{ML} = P(|Z| \leq \frac{\widehat{\sigma}_{ML}}{\sqrt{2}}) = 2\phi(\frac{\widehat{\sigma}_{ML}}{\sqrt{2}}) - 1$, luego

$$\widehat{G}(\underline{x}) = 2\phi(\frac{\sqrt{0.75}}{\sqrt{2}}) - 1 \approx 2\phi(0.61) - 1 \approx 0.4582$$

3. a) Sean $X_1, ..., X_n \sim \mathcal{P}(\lambda)$, como $E(X) = \lambda, V(X) = \lambda$, entonces $\sqrt{n}\frac{\overline{X}-\lambda}{\sqrt{\lambda}} \xrightarrow{\mathcal{D}} N(0,1)$, por el teorema de Slutsky se tiene que $T = \sqrt{n}\frac{\overline{X}-\lambda}{\sqrt{\overline{X}}} \xrightarrow{\mathcal{D}} N(0,1)$ es un pivote asintótico, luego

$$0.95 \approx P(-1.96 \leq \sqrt{n}\frac{\overline{X}-\lambda}{\sqrt{\overline{X}}} \leq 1.96) = P(\overline{X} - 1.96\sqrt{\frac{\overline{X}}{n}} \leq \lambda \leq \overline{X} + 1.96\sqrt{\frac{\overline{X}}{n}})$$

luego $IC(0.95) = [\overline{X} - 1.96\sqrt{\frac{\overline{X}}{n}}, \overline{X} + 1.96\sqrt{\frac{\overline{X}}{n}}]$. Se obtiene en la muestra $IC_{obs}(0.95) = [5.053, 6.546]$.

b) Sí, hay evidencia estadística para rechazar que $\lambda = 5$, ya que este no pertenece al intervalo obtenido en base a la muestra, y tenemos confianza 0.95 de que el verdadero λ esté contenido en este intervalo. De hecho nos da la misma evidencia que un test bilateral. Alternativamente se puede plantear un test, que rechazara $H_0 : \lambda = 5, \quad \text{vs.} \quad H_1 : \lambda > 5$.

c) Similarmente al item a) se tiene que $T = \sqrt{n}\frac{\overline{X}-\lambda}{S} \xrightarrow{\mathcal{D}} N(0,1)$ luego

$$0.95 \approx P(-1.96 \leq \sqrt{n}\frac{\overline{X}-\lambda}{S} \leq 1.96) = P(\overline{X} - 1.96\frac{S}{\sqrt{n}} \leq \lambda \leq \overline{X} + 1.96\frac{S}{\sqrt{n}})$$

Se obtiene en la muestra $IC_{obs}(0.95) = [5.087, 6.513]$.

d) No, no es correcto ya que el intervalo obtenido con la muestra es determinístico y λ también es desconocido pero fijo, por lo tanto lo contiene o no, la probabilidad es 0 o 1, hablamos de confianza 0.95.

4. a) $H_0 : \mu = \mu_0 = 1100 \quad \text{vs.} \quad H_1 : \mu < 1100$.

b) $T = \sqrt{n}\frac{\overline{Y}-\mu_0}{S} \approx N(0,1)$ bajo H_0 por Slutsky. La región de rechazo es de la forma $\mathcal{R} = \{t \,|\, t < k\}$, pero necesito que

$$0.05 = P_{H_0}(T \in \mathcal{R}) = P_{H_0}(T < k)$$

como T es aproximadamente normal, $k = -z_{0.05} = -1.645$.

c) $T_{obs} = \sqrt{260}\frac{1060-1100}{340} \approx -1.897$, luego se rechaza H_0.

d) p-valor $= P_{H_0}(T < T_{obs}) = P_{H_0}(T < -1.897) \approx 0.0294$.

e) Ahora el estadístico es $\overline{Y}$ y rechazamos cuando $\overline{Y} < 1070$. Luego

$$P_{H_0}(\overline{Y} < 1070) = P_{H_0}(\sqrt{260}\frac{\overline{Y}-1100}{300} < \sqrt{260}\frac{1070-1100}{300}) \approx P(Z < -1.61) \approx 0.0537$$