Tarea 1

IN3401 - Estadística para la Economía y la Gestión
Departamento de Ingeniería Industrial, Universidad de Chile
Primavera 2019
Prof. Raimundo Undurraga
Fecha entrega: Martes 10 de Septiembre, hasta las 00:00 horas.

** Preguntas puden ser respondidas en inglés o español. Todas las preguntas valen lo mismo.

I. Teoría de Distribuciones

1. Defina qué es un estimador sesgado. Defina qué es un estimador inconsistente. Explique cuidadosamente cuál es la diferencia entre ambos. Todos los estimadores sesgados son inconsistentes? Todos los estimadores inconsistentes son sesgados? Ilustre con un ejemplo en cada caso.

2. Sean M_i con $i = 1, ..., N$ un conjunto de variables aleatorias independientes e idénticamente distribuidas, de media μ y varianza σ^2. Se define el siguiente estimador de la varianza:

$S^2_\star = \frac{1}{N}\sum_{i=1}^{N}(M_i - \bar{M})^2$ donde $\bar{M} = \frac{1}{N}\sum_{i=1}^{N} M_i$ corresponde a la media muestral.

a. Demuestre que $S^2_\star$ es un estimador sesgado de la varianza poblacional.

b. Proponga un estimador insesgado de la varianza poblacional.

3. Un dado de 9 caras es lanzado 100 veces. Usando la aproximación normal, encuentre la probabilidad de obtener un 2 entre 3 y 7 veces (estríctamente). Encuentre la probabilidad de que la suma de los 100 resultados sea menor a 300. (Hint: Para aproximar una variable continua a una discreta se utiliza una "corrección de continuidad": restar 0.5 al límite inferior y sumar 0.5 al límite superior. Por ejemplo, 4 a 6 se analiza como 3.5 a 6.5.)

4. La variable aleatoria X tiene distribución uniforme continua con $0 < X < 7$.

(i) Cuál es la función de densidad de X?
(ii) Cuál es la función de distribución acumulada de X?
(iii) Encuentre la media y la varianza de X.
(iv) Cuál es la densidad condicional de $X|X > 4$?
(v) Cuál es la esperanza condicional y la varianza condicional de esta variable?

5. Considere la distribución conjunta de dos variables aleatorias, y, que representa el número de delitos al día en la Región Metropolitana, y x, la densidad promedio de carabineros en la Región Metropolitana. Note que x es continua e y discreta. Suponga que la densidad condicional de y es $f(y|x) = \frac{e^{-\beta x}(\beta x)^y}{y!}$, con $y = 0, 1, ..., x \geq 0$, y $\beta > 0$, mientras que la distribución marginal de x es $f(x) = \theta e^{-\theta x}$, con $\theta > 0$. Por lo tanto, condicionada en x, y distribuye Poisson de parámetro βx, mientras que x, no condicionada, tiene distribución exponencial.

a. Cuál es la distribución conjunta, $f(x, y)$ de estas dos variables aleatorias, x e y?

b. Muestre que el inverso de la densidad marginal de y es $\frac{1}{\delta(1-\delta)^y}$, donde el inverso de δ es $\frac{\beta+\theta}{\theta}$.

c. Muestre que $\frac{1}{E[x]} = \theta$ y que $(\frac{1}{Var[x]})^{1/2} = \theta$

d. Si $E[y|x] = \beta x$, obtenga $E[y]$, $Var[y]$, y $Cov[x, y]$. Exprese en términos de β y θ.

6. Las variables aleatorias y y x distribuyen como una normal con medias 2 y 1, varianzas 5 y 2 y covarianza 6.

(i) Calcule la pendiente y el intercepto en la función de esperanza condicional $E[y|x]$.
(ii) Calcule la correlación cuadrada entre y y x.
(iii) Calcule la correlación cuadrada entre y y $E[y|x]$.

7. Suponga que (w, z) distribuyen como una normal bivariada con $E[w] = 0$, $E[z] = 0$, $Var[w] = 1$, $Var[z] = 1$, $Cov[w, z] = \epsilon$. Tenemos una muestra aleatoria (w_i, z_i), con $i = 1, , n$. Nos interesa estimar ϵ, el parámetro desconocido de la distribución.

(i) Usando la ley de los grandes números la covarianza muestral puede ser utilizada para estimar ϵ consistentemente. Explique.
(ii) Una forma distinta de verlo: Sabemos que $E[w|z] = \alpha + \beta z$ donde $\beta = \frac{Cov[z,w]}{Var[z]} = \epsilon$ y $\alpha = E[w] - \beta E[z] = 0$. Luego, $w = \epsilon z + \rho$. Entonces, la regresión lineal de w sobre z nos permite hacer una estimación consistente de ϵ. Es esto cierto? Explique.
   (a) Cuál es la distribución asintótica de este estimador?
   (b) Cómo se diferencia este estimador con el estimador de la parte (i)?
(iii) Un observador alerta de la parte b) nota que la ecuación implica que $z = \frac{1}{\epsilon}w - \frac{1}{\epsilon}\rho = \delta w + \delta\rho$. Dado esto, él sugiere que también podemos hacer una regresión lineal de z sobre w para estimar δ. Luego, gracias al Teorema de Slutsky, se puede obtener un estimador consistente de ϵ tomando el recíproco del estimador de δ. Es esto verdadero o falso? Explique.

8. Sea W una variablea aleatoria que distribuye normal de parámetros $(\eta, 9m^2)$. Encuentre $\Pr(W > \eta + m)$.

9. Mostrar que si una variable W sigue una $Normal(0, 1)$, W^2 sigue una distribución Chi-cuadrado con 1 grado de libertad. Mostrar además que esta distribución es un caso particular de distribuciones gamma.

II. Test de Hipótesis

10. Sea $\overline{X}_{45}$ la media muestral de una muestra de tamaño 45 de una distribución Gamma$(8, \lambda)$. Use el Teorema del Límite Central (TLC) para encontrar un aproximación para el intervalo de confianza de λ al 90% de confianza. (Hint : Use el hecho de que cuando $X \sim$ Gamma$(8, \lambda)$, $E(X) = \frac{8}{\lambda}$ y $Var(X) = \frac{8}{\lambda^2}$ y aplique el TLC a la variable aleatoria $\left(\overline{X}_{45} - \frac{8}{\lambda}\right) \Big/ \sqrt{\frac{8}{45\lambda^2}}$.)

11. Se observa la muestra 1.6, 2.9, 1.3, 0.4, 2.4, 1.8, 1.2, 2.8, 2.5, 3.5 de tamaño 10 de una distribución $N(2, \sigma^2)$, i.e., tiene media conocida y equivale a 2. Construya un intervalo al 99% de confianza para σ^2. (Hint: use el hecho de que si $X_i \sim N(\mu, \sigma^2)$, luego $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2$ distribuye Chi-cuadrado con n grados de libertad.)

12. De acuerdo con la Asociación Nacional de Atletas Universitarios, "los atletas de las principales universidades se graduaron, en general, al mismo ritmo que otros estudiantes". La encuesta incluía a estudiantes que habían ingresado a la universidad justo antes que se pusiera en práctica la proposición 48 de la Asociación (la proposición 48 establece que los atletas deben mantener un promedio de calificaciones igual o mayor a 4 en una escala de 0 a 7). Suponga que una nueva encuesta aplicada a 800 atletas indica que se gradúan 568.

a. Determine un intervalo de confianza al 90% para p, la proporción de atletas universitarios que se gradúan.

b. El intervalo incluye el valor $p = 0.51$, porcentaje de estudiantes que se graduaron antes que entrara en vigencia la proposicin 48? Qué puede concluir de esto?

III. Ejercicios Empíricos

(Necesita utilizar un software para esta parte de la tarea. Puede utilizar cualquier programa que le acomode: Stata, R, Excel, etc. Los cálculos son directos.)

13. El archivo P13.csv es un archivo Excel (portable) CSV que contiene los datos de la cuenta de la luz y el número de piezas para 144 hogares.

(i) Haga un simple scatter (X-Y) plot con la variable N PIEZAS en el eje horizontal y CUENTA LUZ en el eje vertical. Qué conclusiones puede obtener de la relación entre N PIEZAS y CUENTA LUZ?

(ii) Note que N PIEZAS solo toma unos pocos valores, $3, 4, 5, ..., 11$. Compute la media de CUENTA LUZ para cada valor de N PIEZAS. Que puede concluir sobre la media condicional? Grafique la media contra el número de piezas. Qué observa? Explique.

14. Se realiza un experimento que varía la cantidad de fertilizante que se usa para hacer crecer manzanos. Se testean 4 concentraciones, utilizando 3 grupos de 12 árboles para cada una. Más adelante en el año, se mide el peso promedio de las manzanas para cada grupo de árboles. Si todo hubiese salido bien, tendríamos 3 observaciones del peso promedio para cada una de las 4 concentraciones. Sin embargo, 2 grupos fueron destruidos accidentalmente por una persona en un *bulldozer*. Dado esto, sólo se obtienen los siguientes 10 datos:

Table 1: Tratamiento de Manzanas

Observación	Tipo de Tratamiento	Peso Manzanas
1	1	127.5
2	1	123.8
3	1	114.4
4	2	58.9
5	2	60.4
6	2	68.9
7	3	80.4
8	3	96.9
9	4	97.7
10	4	77.3

Cómo construiría un estadístico F para testear si existen diferencias significativas entre las 4 concentraciones en términos del peso de las manzanas?

15. Use Excel para generar 500 muestras aleatorias de tamaño $n = (10; 30; 100)$ a partir de una distribución exponencial con $\lambda = 0.5$ (y por tanto una media poblacional $\mu_n = \frac{1}{0.5} = 2$ y una varianza poblacional $\sigma_n^2 = \frac{1}{(0.5)^2} = 4$). Para cada tamaño muestral, crear una base de datos diferente en una hoja particular del documento Excel (hoja $1 : n = 10$; hoja $2 : n = 30$; hoja $3 : n = 100$). Sugerencia: cada fila es una muestra, i.e., cada una de las 3 hojas debería tener 500 filas correspondiente a 500 muestras diferentes. Sin embargo, para $n = 10$ tendremos 10 columnas (10 observaciones cada muestra), para $n = 30$ tendremos 30 columnas (30 observaciones cada muestra), y para $n = 100$ tendremos 100 columnas (100 observaciones cada muestra)

1. En cada una de las tres hojas, calcule el promedio $\overline{y}_r$ para cada muestra r;
2. Plotee un histograma de $\overline{y}_r$ imponiendo una densidad normal con media μ_n y varianza σ_n^2. Deberías obtener 3 histogramas, una para cada tamaño muestral. Pegar cada histograma en una hoja separada, con nombre $hist10$, $hist30$, e $hist100$. (Nota: Para la inversa de la distribucion acumulada de una exponencial, use $-ln(1-p)/\lambda$).
3. Crear “qqplots” para cada caso, i.e., cuando $n = (10; 30; 100)$. Pegar cada qqplot en una hoja separada, con nombre $qqplot10$, $qqplot30$, e $qqplot100$. Qué se puede concluir respecto al supuesto de normalidad en cada caso?
4. Para cada muestra $r = 1 \dots 500$, en cada caso $n = (10; 30; 100)$, crear un intervalo de confianza asintótico del 95% para la media poblacional y crear un indicador igual a 1 si el intervalo de confianza contiene la media poblacional, y 0 si no. Expresar los resultados en una hoja separada llamada ”IC”. Para qué tamaños muestrales (n=10,30, o 100) el teorema del límite central predice correctamente el área de los intervalos de confianza? Explique.