Estudio de caso # 4

Universidad Externado de Colombia

Departamento de Matemáticas

Estadística 2

Juan Sosa, Ph. D.

October 10, 2018

Instrucciones generales

- **Puede hacer el examen solo o puede asociarse con otra persona,** entendiendo que la calificación del examen será la misma para ambas personas.
- El reporte final se debe enviar a más tardar el **miércoles 31 de octubre de 2018** a las **11:59 p.m.** a la cuenta de correo: `juan.sosa@uexternado.edu.co`.
- Reportar las cifras utilizando la **cantidad adecuada de decimales,** dependiendo de lo que se quiera mostrar y las necesidades del problema.
- Numerar figuras y tablas (`http://unilearning.uow.edu.au/report/1fi.html`) y proporcionarles un **tamaño adecuado** que no distorsione la información que estas contienen.
- El archivo del reporte final debe ser un **archivo** `pdf` con el siguiente formato: Letra Calibri, tamaño 12, interlineado sencillo con espacio entre párrafos y texto justificado. Márgenes: Normal. Tamaño: Carta. Orientación: Vertical.

- Especificar el software donde se llevó a cabo el computo e **incluir el código** correspondiente como un anexo al final del reporte con el siguiente formato: Letra Courier New, tamaño 10, interlineado sencillo.

- El objetivo principal de este trabajo es la claridad lógica y la interpretación de LOS resultados. **El informe no necesita ser extenso.** Recuerde ser minimalista escribiendo el reporte. Se deben incluir solo aquelLOS gráficos y tablas (¡y valores en la tabla!) que son relevantes para la discusión.

- Hacer el informe ya sea en inglés o español. No ambos!

- **Cualquier evidencia de plagio o copia se castigará severamente** tal y como el reglamento de la Universidad Externado de Colombia lo estipula.

  Si está claro que (por ejemplo) dos grupos han trabajado juntos en una parte de un problema que vale 20 puntos, y cada respuesta habría ganado 16 puntos (si no hubiera surgido de una colaboración ilegal), entonces cada grupo recibirá 8 de los 16 puntos obtenidos colectivamente (para una puntuación total de 8 de 20), **y me reservo el derecho de imponer penalidades adicionales a mi discreción**.

  Si un grupo resuelve un problema por su cuenta y luego comparte su solución con cualquier otro grupo (porque rutinariamente Usted hace esto, o por lástima, o bondad, o por cualquier motivo que pueda creer tener; no importa!), Usted es tan culpable de colaboración ilegal como la persona que tomó su solución, y ambos recibirán la misma penalidad. Este tipo de cosas es necesario hacerlas ya que muchas personas no hacen trampa, y debo asegurarme de que sus puntajes son obtenidos de manera genuina. En otras clases, personas perdieron la clase debido a una colaboración ilegal; **no deje que le suceda a Usted!**

Encuesta de Transición de la Escuela al Trabajo

Considere la Encuesta de Transición de la Escuela al Trabajo que realizó el DANE en 2013 y 2015. La descripción completa de la operación estadística y la descripción de las variables se encuentra en los siguientes enlaces:
`http://microdatos.dane.gov.co/index.php/catalog/517/get_microdata`
`http://microdatos.dane.gov.co/index.php/catalog/518/get_microdata`.
Esta encuesta contiene ocho módulos, a saber:

- Características generales.
- Educación.
- Jóvenes ocupados.
- Jóvenes desocupados.
- Inicio del historial de actividades.
- Perspectivas laborales y fuerza de trabajo.
- Jóvenes inactivos.
- Historial de actividades.

Considere el módulo de JOVENES OCUPADOS (107 variables). Los datos correspondientes a los años 2013 y 2015 se encuentran disponibles en los archivos `JO13.csv` y `JO15.csv`, respectivamente. Este caso de estudio es de alguna manera *longitudinal* porque se quiere investigar si hubo cambios considerables de 2013 a 2015 respecto a las características más relevantes de los jóvenes ocupados. Entre otras, las preguntas EP689 y EP6500 son de especial interés porque se relacionan con los ingresos de los jóvenes ocupados:

EP689 ¿Existe un nivel mínimo de ingreso mensual por debajo del cual no aceptaría un trabajo? (1 = Sí, 2 = No).

EP6500 Antes de descuentos ¿cuánto ganó el mes pasado en este empleo? (Incluya propinas y comisiones, y excluya viáticos y pagos en especie). Valor mensual en pesos ($).

Sea π_1 y π_2 la proporción poblacional de jóvenes para los cuales sí existe un nivel mínimo de ingreso mensual por debajo del cual no aceptaría un trabajo en 2013 y 2015, respectivamente. Sea σ_1^2 y σ_2^2 (μ_1 y μ_2) la varianza (media) poblacional de los ingresos de los jóvenes en 2013 y 2015, respectivamente.

1. De acuerdo con el diseño estadístico de la encuesta (ver `Metodologia ETET 2015.pdf`, por ejemplo), ¿cuál es el universo, la población objetivo, la fuente de datos, y la cobertura y desagregación geográfica?

2. Importar las bases de datos y remover todas las variables exceptuando aquellas correspondientes a las preguntas EP689 y EP6500. Luego, remover todos los registros que tengan algún dato faltante. En seguida, remover todos aquellos individuos que en la pregunta EP6500 tengan registrados los valores 0, 98, o 99, o valores superiores a $10,000,000. Después de este filtro, la cantidad de registros completos teniendo en cuenta únicamente las preguntas EP689 y EP6500 es 2,288 en 2013 y 2,228 en 2015.

3. Clasificar las variables según su naturaleza y su escala de medición (ver Sosa et al. 2012, Cap. 1, por ejemplo).

4. Describir numérica y gráficamente las variables en ambos años. ¿Parece haber cambios de un año a otro en alguna de las variables? Comentar los resultados obtenidos.

5. Usando un nivel de significancia del 5%, ¿existen diferencias significativas entre π_1 y π_2? Responder esta pregunta usando tanto intervalos de confianza (reportar el error estándar, el margen de error, el coeficiente de variación y el intervalo) como pruebas de hipótesis (mostrar explícitamente todos los pasos de la prueba incluyendo la región de rechazo y el valor p).

6. Para la prueba del numeral 5., graficar la potencia de la prueba para valores de $\pi_1 - \pi_2$ entre -1 y 1. Comentar los resultados obtenidos acerca de la potencia y el error tipo II de la prueba.

7. ¿Existen diferencias significativas entre σ_1^2 y σ_2^2? Proceder del mismo modo que en el numeral 5.

8. ¿Existen diferencias significativas entre μ_1 y μ_2? Proceder del mismo modo que en el numeral 5.

9. Para la prueba del numeral 8., graficar la potencia de la prueba para valores de $\mu_1 - \mu_2$ en un rango apropiado. Comentar los resultados obtenidos acerca de la potencia y el error tipo II de la prueba.

10. Graficar un histograma (y sobre este la curva normal correspondiente) y un gráfico cuantil-cuantil normal (con las bandas de confianza) para los ingresos de los jóvenes en el 2013. Probar el sistema de hipótesis correspondiente usando la prueba de Shapiro-Wilk y la prueba de Kolmogorov-Smirnov. ¿Los ingresos de los jóvenes en 2013 parecen provenir de una distribución normal?

    Nota: Para llevar a cabo la prueba de Kolmogorov-Smirnov, use la función `ks.test` de `R`. Más información acerca de esta función la puede encontrar en el siguiente enlace: `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ks.test.html`.

11. Repetir el numeral anterior para los ingresos de los jóvenes en 2015.

12. ¿El supuesto de normalidad de las poblaciones se satisface? ¿Es absolutamente necesario que se satisfaga dado que los tamaños muestra son grandes? ¿Por qué?

13. Al igual que en el caso de estudio # 3, ajustar el modelo log-normal usando el método de máxima verosimilitud para los ingresos de los jóvenes en el 2013. Graficar el histograma de los ingresos y sobre este la curva del modelo log-normal ajustado.

    Nota: No es necesario hacer las demostraciones relacionadas con el modelo log-normal. Utilice las fórmulas directamente.

14. Repetir el numeral anterior para los ingresos de los jóvenes en el 2015.

15. Sea θ_1 y θ_2 la media poblacional de los ingresos de los jóvenes en 2013 y 2015, respectivamente, usando el modelo log-normal. Usando el mismo procedimiento presentado en el casos de estudio #3, calcular un intervalo de confianza para $\theta_1 - \theta_2$ usando una confiabilidad del 95% (reportar el error estándar, el margen de error, el coeficiente de variación y el intervalo).

16. Es posible probar el sistema de hipótesis $H_0 : \theta_1 - \theta_2 = 0$ frente a $H_1 : \theta_1 - \theta_2 \neq 0$ usando el modelo log-normal. Para ello, el estadístico

de prueba es

$$\text{Estadístico de prueba} = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\mathbb{V}\text{ar}\left[\hat{\theta}_1 - \hat{\theta}_2\right]}},$$

donde $\hat{\theta}_1$ y $\hat{\theta}_2$ son los estimadores de máxima verosimilitud de θ_1 y θ_2, respectivamente. Recuerde que bajo el supuesto de independencia de las observaciones en ambos años, se tiene que

$$\mathbb{V}\text{ar}\left[\hat{\theta}_1 - \hat{\theta}_2\right] = \mathbb{V}\text{ar}\left[\hat{\theta}_1\right] + \mathbb{V}\text{ar}\left[\hat{\theta}_2\right].$$

En este caso, dado que los tamaños de muestra son grandes, la distribución de referencia para el estadístico de prueba es una distribución z (normal estándar).

Usando un nivel de confianza del 95%, probar el sistema de hipótesis anterior (mostrar explícitamente todos los pasos de la prueba incluyendo la región de rechazo y el valor p).

17. Resumir los resultados de los numerales 8., 15., y 16. en una tabla y comprar los resultados. ¿Las conclusiones son equivalentes? ¿Por qué?

Ahora considere el módulo de PERSPECTIVAS LABORALES Y FUERZA DE TRABAJO (25 variables). Los datos correspondientes a los años 2013 y 2015 se encuentran disponibles en los archivos `PLFT13.csv` y `PLFT15.csv`, respectivamente. Se quiere (a) identificar si en 2015 el optimismo sobre las expectativas laborales futuras es independiente del sector donde los jóvenes quieren conseguir trabajo, y (b) probar si existió un cambio significativo del 2013 al 2015 en términos del sector donde los jóvenes quieren conseguir trabajo. Para ello se tienen en cuenta las preguntas EP732 y EP741 como sigue:

EP732 En términos generales, ¿se siente optimista sobre sus expectativas laborales futuras? (1 = Sí, 2 = No).

EP741 ¿Con quién le gustaría trabajar principalmente? (1 = Solo (negocio propio/granja), 2 = Trabajar para el gobierno/sector público, 3 = Trabajar para una empresa privada, 4 = Trabajar sin remuneración en un negocio familiar/granja, 5 = Otra).

1. Importar las bases de datos y remover todas las variables exceptuando aquellas correspondientes a las preguntas EP732 y EP741. Remover todos los registros que tengan algún dato faltante.

   La cantidad de registros completos teniendo en cuenta únicamente las preguntas EP732 y EP741 es 6,416 en 2013 y 6,524 en 2015.

2. Clasificar las variables según su naturaleza y su escala de medición.

3. Describir numérica y gráficamente las variables en 2015. ¿Parece haber una relación entre las variables? Comentar los resultados obtenidos.

4. Probar si las variables correspondientes a las preguntas EP732 y EP741 son dependientes en 2015. Responder esta pregunta usando pruebas de hipótesis chi cuadrado de independencia (mostrar explícitamente todos los pasos de la prueba incluyendo el valor p). Usar un nivel de significancia del 5%.

5. Calcular las frecuencias relativas de la pregunta EP741 en el año 2013. Considerando estas frecuencias como referente histórico (valores hipotéticos), probar si hubo cambios significativos de 2013 a 2015. Proceder del mismo modo que en el numeral anterior, pero esta vez usando pruebas de hipótesis chi cuadrado de bondad de ajuste.

Finalmente, teniendo en cuenta los resultados obtenidos en los dos módulos anteriores, ¿cuáles serían sus recomendaciones al Ministerio de Trabajo con miras al diseño de una política pública? Explicar brevemente.