Proyecto 2020

ASIGNATURA LIC MI CARRERA

Orientaciones metodológicas:

Este es el tema de mi clase

Estudiante Pérez Pérez *

Ejercicios

Ejercicio 1 *Genere una Población normal de tamaño 500, seleccione 8 muestras de tamaños varios(Muy mayor que 30, mayor que 30, 30, 20), 4 muestras con remplazo y 4 sin remplazo.*

a. *Calcule para cada una de las muestras los Estadísticos Descriptivos, de la Conferencia 1.*

b. *Calcúlelos en la población inicial. Analice las diferencias.*

c. *Grafique los resultados.*

d. *Para cada muestra calcule los intervalos de confianza para la media y la varianza.*

e. *Analice las diferencias en los resultados de las muestras de tamaños similares.*

Propuesta de distintos ejercicios de la clase, para desarrollar las habilidades a crear durante la clase.

Ejercicio 2 *De acuerdo a su set de datos(Equipo 10):*

a. *Utilice los Estadísticos Descriptivos estudiados en la Conferencia 1. Para describir el comportamiento de tres de sus variables. Seleccione las que sean más importantes y explique porque seleccionó estas.*

b. *Grafique los resultados.*

c. *Interprete los resultados en términos del problema.*

Ejercicio 3 *Analizando los datos del archivo "adult.data.csv"(Equipo 10), ¿Hay diferencias significativas entre el promedio de años dedicados a la educación y la cantidad de ingreso de los censados?*

Objetivos

a. Esta sección va dedicada a los objetivos de la clase, las metas para el encuentro y ciertas especificidades que considere de importancia resaltar durante el trancurso de la clase.

b. Según la temática se pueden hacer alusión a los medios de enseñanza utilizados convenientemente.

Introducción

(Xmin')

(Como introducir mi clase?)

- Recursos para motivar la clase.
- Recuento por los antecedentes de los resultados o investigadores.
- Esta no tiene que venir acompañada por plecas, solo es un ejemplo.

Ejercicio 1

(código en doucmento adjunto)

Se genera una poblacion inicial con 500 valores y una distribución normal con media 0 y varianza 1. Luego

*Soy estudiante de X año de mi carrera.

de esta se extraen 4 muestras sin remplazo, cada una de tamaño 200, 60, 30 y 20 respectivamente. Luego se extraen otras 4 de igual tamaño a las anteriores y con remplazo.

La exactitud de los estimadores puntuales fluctua en cada prueba realizada, la fluctuacion era mayor o menor dependiendo del tamaño de las muestras. Las muestras de mayor tamaño presentan sus estimadores son generalmente más exactos, en cambio, las de menor tamaño suelen estar más alejadas del valor real. Esto se debe a que en las muestras más grandes cuando tienen un caso extremo no representativo de la población su impacto queda disminuido por el resto de los datos no extremos, mientras que en las muestras pequeñas la existencia de uno de estos altera considerablemente la información extraída. Esta razón refleja la dependencia directa que existe entre los resultados de una muestra y la calidad de sus datos. Podemos deducir entonces la importancia de que las muestras esten compuestas por datos fiables, e incluso más cuando la muestra es pequña.

La extracción de muestras con remplazo en lo que se refiere a la media no tiene muchos cambios.

Ejercicio 2

El dataset contaba con valores faltantes. Para lidiar con este problema se realizó un preprocesamiento de los datos. Se tenían dos opciones: descartar dichos valores faltantes o sustituirlos por la media de los conocidos. Se optó por la última opción. De esta forma se obtuvo un set de datos más limpio y completo que facilitó el trabajo en el posterior anális.

Como parte del análisis se tomaron tres variables que se consideraron las más importantes. Primeramente tomamos 'education', pues consideramos que el nivel de educación influye en gran medida en el salario de una persona. No suelen ganar lo mismo una persona con nivel universitario queuno que solamente tiene un nivel medio-superior. La otra variable escogida para analizar fue 'occupation'. El salario de una persona está determinado principalmente por su profesión o el tipo de trabajo que realiza. Como tercera y última variable a analizar se tomó 'sex'. Se decidió tomar esta variable porque se sabe que hoy en la actualidad todavía existen diferencias en cuanto al salario de dos personas que se realizan el mismo trabajo en dependencia del sexo, por lo que resulto de interés trabajar con esta variable.

Para garantizar la veracidad de los planteamientos anteriores se hicieron pruebas de hipótesis para cada uno de estos atributos. El resultado coincidió. Tanto la educación como la ocupación y el sexo influyen significativamente en los ingresos anuales de una persona. Sin embargo, se decidió analizar el resto de variables y se pudo observar un hecho interesante: la gran mayoría de los atributos del dataset influyen significativamente en los ingresos anuales de una persona. Por lo que podemos decir que se hizo una muy buena seleccion de variables a la hora de recopilar información referente al problema.

Con respecto a los resultados obtenidos a través del análisis descriptivo de las tres variables escogidas, la tabla 4 muestra los resultados para la educación, ocupación y el sexo.En las figuras de la 1 a la 6 se muestran algunos resultados gráficos. En este caso, todas las variables escogidas eran categóricas por lo que se codificaron en valores enteros. En las tablas 1,2 y 3 se puede apreciar que valor representa cada número.

Interpretación de los resultados

Table 1: Valores numérico de las categorias de educación

Categoría	**Valor**
Bachelors	1
Some-college	2
11th	3
HS-grad	4
Prof-school	5
Assoc-acdm	6
Assoc-voc	7
9th	8
7th-8th	9
12th	10
Masters	11
1st-4th	12
10th	13
Doctorate	14
5th-6th	15
Preschool	16

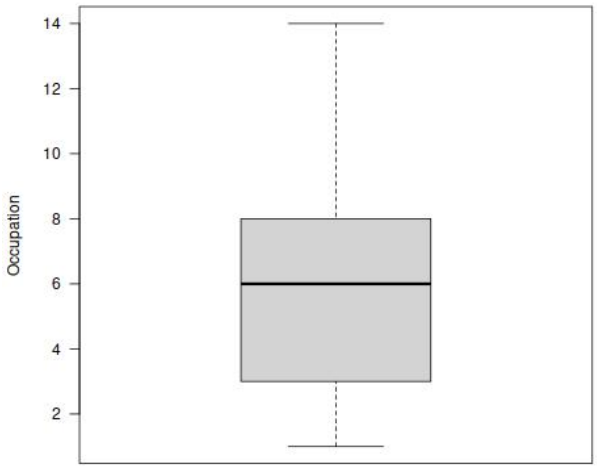

Figure 1: Boxplot para ocupacion

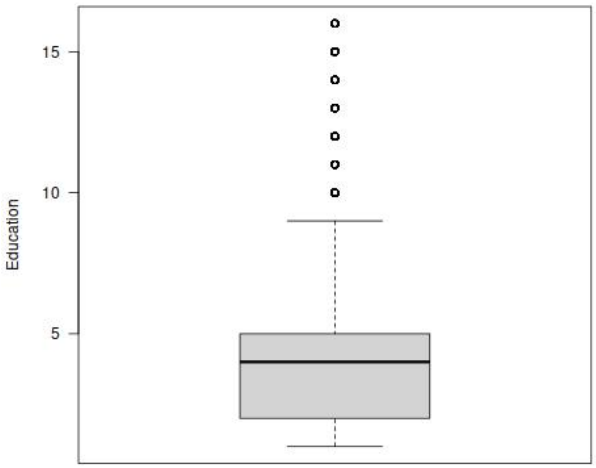

Figure 3: Boxplot para educacion

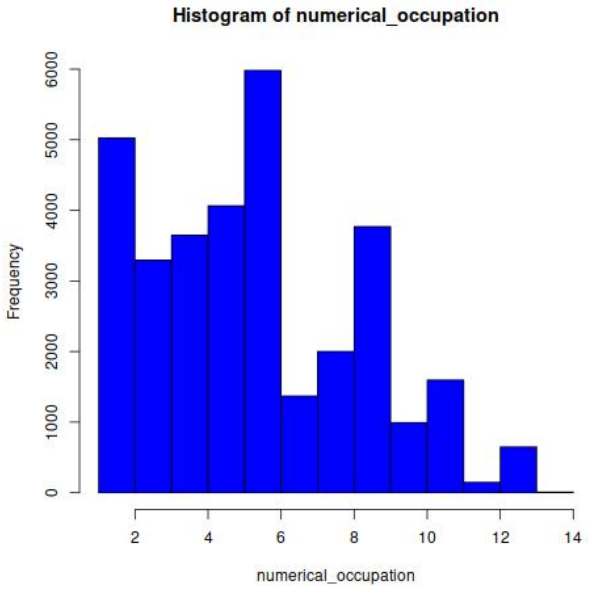

Figure 2: Histograma de ocupacion

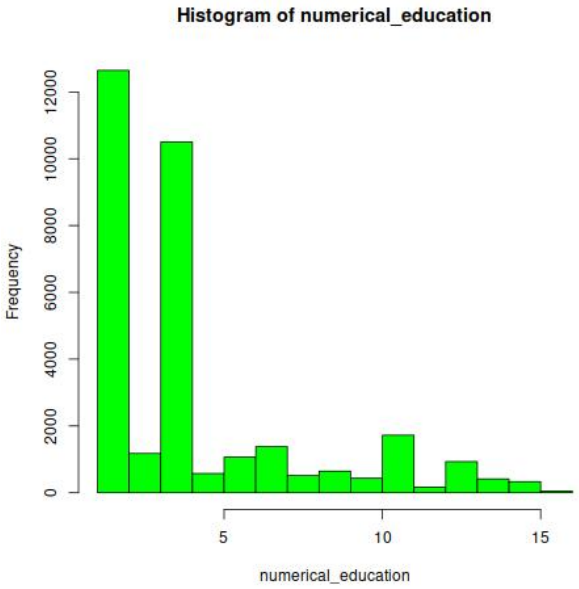

Figure 4: Histograma de educacion

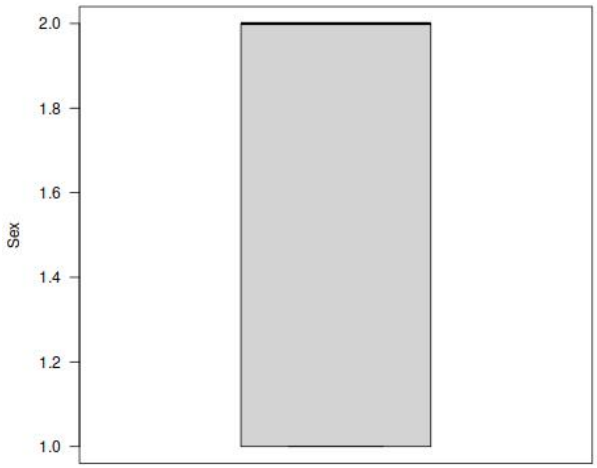

Figure 5: Boxplot para sexo

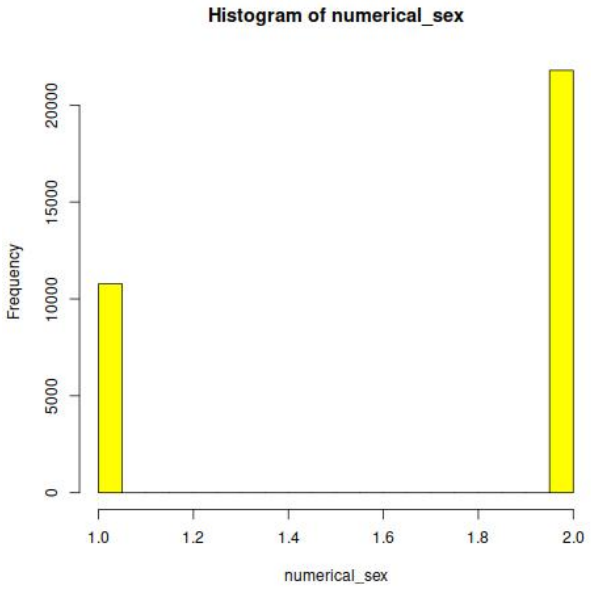

Figure 6: Histograma de sexo

Table 2: Valores numéricos de las categorias de ocupación

Categoría	**Valor**
Tech-support	1
Craft-repair	2
Other-service	3
Sales	4
Exec-managerial	5
Prof-specialty	6
Handlers-cleaners	7
Machine-op-inspct	8
Adm-clerical	9
Farming-fishing	10
Transport-moving	11
Priv-house-serv	12
Protective-serv	13
Armed-Forces	14

Table 3: Valores numéricos de las categorias de sexo

Categoría	**Valor**
Femenino	1
Masculino	2

Table 4:

Estadígrafo	**Ocupación**	**Educación**	**Sexo**
Media	6	4	2
Mediana	6	4	2
Varianza	8.35	11.98	0.22
DS	2.89	3.46	0.47
Moda	6	4	2
Min	1	1	1
Lower-Hinge	3	2	1
Median	6	4	2
Upper-Hinge	8	5	2
Max	14	16	2

Ejercicio 3

El problema trata sobre comparar la medias de años de educacion de los grupos de ingresos $\leq 50K$ y $> 50K$, y ver si hay alguna diferencia significativa entre esas medias.
El código referente a este ejercicio esta en el archivo exercise_3.R

Para lograr lo anterior se dividió todos los datos del archivo csv en dos grupos: un grupo tiene ingresos $\leq 50K$ y otro grupo tiene ingresos de $> 50K$.
Luego de realizar la división se escogieron dos muestras sin reemplazo de tamaño $N = 60$ cada una.
Por lo que el problema se reduce a una prueba de hipótesis para la comparación de las medias de dos poblaciones Normales.
Las hipótesis para la prueba que se escogieron fueron:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Donde μ_1 es la media de años de educación de la población de personas que poseen ingresos por debajo de los $50K$(inclusivo) y μ_2 es la media de años de educación de la población de personas que poseen ingresos por encima de los $50K$.
Se asumió que no se conocían las varianzas de dichos grupos(o que era muy costoso calcularlas). Por lo que para hacer la prueba de hipótesis de la media se hace primero una prueba de hipótesis para la igualdad de las varianzas.

Parte del codigo para la hipotesis de varianza

```
result <- var.test(sample1,
sample2, alternative = "two.sided")
```

Luego de establecer la igualdad o desigualdad de varianza se procede a realizar la prueba de la media:

Parte del codigo para la hipotesis de la media

```
result <-t.test(sample1,sample2,
alternative = alt, var.equal = varequal)
```

Finalmente se compara el resultado del *p-value* de *result* con un valor α preestablecido y si es menor se rechaza la hipótesis nula.

Analíticamente

Se hace primero la prueba de varianza.
Los datos se tomaron del script referente al ejercicio.
$H_0 : \sigma_1^2 = \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{S_1^2}{S_2^2} = \frac{6.37}{5.13} = 1,24$$

$$F_{1-\alpha/2}(n_1 - 1, n_2 - 1) = F_{0.975}(59, 59) = 1.67$$

$$F_{\alpha/2}(n_1 - 1, n_2 - 1) = F_{0.025}(59, 59) = 0.597$$

Como $F > F_{\alpha/2}(n_1 - 1, n_2 - 1)$ y $F < F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ no se puede descartar H_0 por lo que se asume que $\sigma_1^2 = \sigma_2^2$

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} = 3.7$$

$$t_{1-\alpha/2}(n_1 + n_2 - 2) = t_{0.975}(116) = 1.98$$

Como $|T_{\bar{X}-\bar{Y}}| > t_{1-\alpha/2}(n_1+n_2-2)$ se cumple la región crítica y se descarta H_0 pudiendo afirmarse que se aprecian diferencias significativas.

Conclusiones

(Cierta cantidad de minutos')

Se resumirán los resultados más destacados ejercitados en la actividad.

Se puede hacer mención de aplicaciones del método estudiado, posibles investigaciones o repercusiones en la cotidianidad. Así como los elementos de mayor significación.

Estudio Independiente

(Algun tiempo')

Orientar y comentar los ejercicios siguientes:

Ejercicio 4 *De creerlo conveniente, la asignación de tareas para el estudio independiente, o la asignación de evaluaciones.*

Ejercicio 5 *La cantidad de los mismos es a conveniencia aunque podría ser de ayuda su justificación.*

Ejercicio 3

Para concluir, la solución de los ejercicios propuestos.

Ejercicio 4

El esquema de clase es variable y queda sujeto a la voluntad del participante, lo que si deberá ajustarse a los requisitos de la convocatoria oficial.