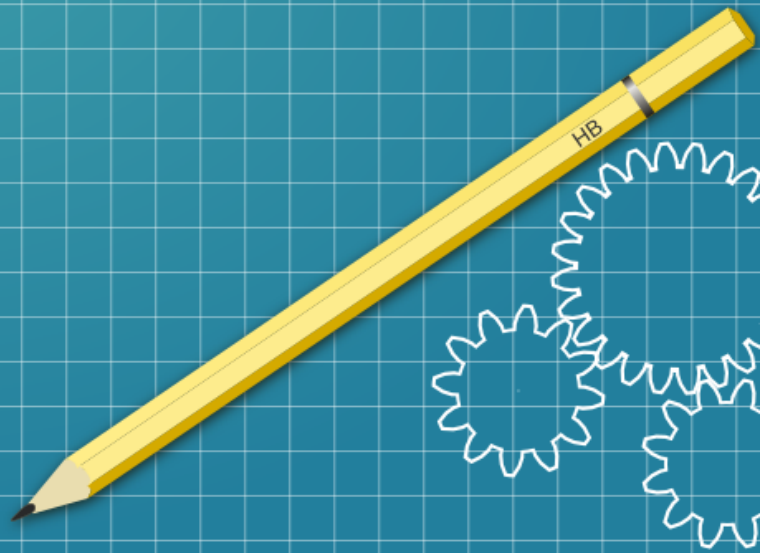
Trabajo Integrador PyE 2023

Alumno: Jofré Santiago

Registro: 21260



Análisis del dataset

- El dataset datos relacionados al consumo de cigarrillos de una determinada marca. Las variables relevantes son:
- Consumo
- Precios
- Ingresos

Estadística Descriptiva

- Variables: cualitativas, cuantitativas
- Medidas de tendencia central.
- Medidas de dispersión.

Medidas de tendencia central

- Media: $\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$
- Mediana
- Moda: valor/es del conjunto de datos que tienen mayor frecuencia

Medidas de dispersión

- Varianza: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- Desviación Típica: $\sqrt{}$ Var(x)

- Coeficiente de variación: $CV(x) = \frac{s(x)}{\bar{x}}$

Estadística descriptiva – Medidas de tendencia Central: Salidas de Python

- Variable: consumo

-------Variable: consumo-------

Tipo de variable: cuantitativa continua

----Medidas de tendencia central----

Media muestral:

4.86

Mediana:

4.84

Moda(s):

0 4.96

Name: consumo, dtype: float64

Estadística descriptiva – Medidas de tendencia Central: Salidas de Python

-------Variable: precio-------

Tipo de variable: cuantitativa continua

----Medidas de tendencia central----

Media muestral:

0.2

Mediana:

0.2

Moda(s):

0 0.12

Name: precio, dtype: float64

Estadística descriptiva – Medidas de tendencia Central: Salidas de Python

- Variable: Ingresos

-------Variable: ingresos-------

Tipo de variable: cuantitativa continua

----Medidas de tendencia central----

Media muestral:

4.75

Mediana:

4.73

Moda(s):

0 4.59

Name: ingreso, dtype: float64

Estadística descriptiva – Medidas de dispersión: Salidas de Python

- Variable: consumo

----Medidas de dispersion----

Varianza:

0.04

Desviacion tipica:

0.19

Coeficiente de variacion:

0.04

De la varianza y desviación típica se concluye que no hay una dispersión significativa entre los datos de la variable

Del coeficiente de variación se tiene que los datos se alejan de la media un 4% y que, la media del conjunto de datos es representativa y por lo tanto, los datos son ‘homogéneos’.

Estadística descriptiva – Medidas de dispersión: Salidas de Python

- Variable: precio

----Medidas de dispersion----

Varianza:

0.01

Desviacion tipica:

0.08

Coeficiente de variacion:

0.41

De la varianza y desviación típica se observa que no hay gran dispersión entre los datos del conjunto

El coeficiente de variación, indica que los datos se alejan de la media un 41%, por lo que el conjunto de datos tiene una forma más ‘heterogénea’.

Estadística descriptiva – Medidas de dispersión: Salidas de Python

- Variable: ingresos

----Medidas de dispersion----

Varianza:

0.02

Desviacion tipica:

0.15

Coeficiente de variacion:

0.03

De la varianza y la desviación típica también se observa que los datos no están tan dispersos entre sí ya que lo valores obtenidos son pequeños

Del coeficiente de variación se concluye que el conjunto de datos difiere de la media solo un 3% por lo que el

Otras conclusiones de los coeficientes de variación obtenidos

- El CV al ser un número adimensional permite comparar distintos conjunto de datos, incluso si son de naturaleza distinta.
- En el caso de la variables analizadas se concluye que el CV más pequeño lo tiene la variable ingresos (0.03), mientras que el CV más grande lo tiene la variable precio (0.41) ambas variables con muestra de igual tamaño (n=46)

Cuartiles – Salidas de Python

- Variable: Consumo

-------Variable: consumo-------

Q1=4.72

Q2=4.84

Q3=4.98

De los cuartiles observados se puede concluir que el 25% de los datos son menores o iguales que 4.72; el 50% de ellos menor o igual a 4.84 y el 75% de ellos menor o igual a 4.98

Cuartiles – Salidas de Python

- Variable: Precio

-------Variable: precio-------

Q1=0.15

Q2=0.2

Q3=0.24

De los cuartiles obtenidos se observa que el 25% de los datos son menores o igual a 0.15; el 50% son menores o iguales a 0.20 y el 75% de ellos son menores o iguales a 0.24

Cuartiles – Salidas de Python

- Variable: Ingresos

-------Variable: ingresos-------

Q1=4.62

Q2=4.73

Q3=4.84

De los cuartiles obtenidos se concluye que el 25% de los datos son menores o iguales a 4.62; el 50% de los datos son menores o iguales a 4.73 y el 75% menores o iguales a 4.84

Coeficiente de asimetría y Curtosis

- El coeficiente de asimetría de un conjunto de datos es el numero:

$$\mathrm{sk} = \frac{1}{s^3}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3\right)$$

- La curtosis es un número que denotaremos por la letra k, y se define de la siguiente manera:

$$\mathrm{k} = \frac{1}{s^4}\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4\right)$$

Coeficiente de asimetría y Curtosis – Salidas de Python

Coeficiente de asimetria de las variables:

consumo: -0.01

precio : -0.16

ingreso : 0.65

Para el caso de la variables consumo y precio se obtuvieron números negativos, por lo que en las gráficas de las frecuencias se verán colas más alargada a la izquierda de las medias (asimetría negativa)

Para el caso de la variable ingreso se observa un número positivo, esto indica que en la gráfica habrá un cola mas alargada a la derecha de la media (asimetría positiva)

Coeficiente de asimetría y Curtosis – Salidas de Python

Curtosis:

consumo: 0.18

precio : 0.13

ingreso: -0.13

En el caso de la variable consumo y precio observamos un valor de curtosis positivo, por lo que las variables tienen una distribución leptocúrtica, que indica la presencia de colas cortas y un decrecimiento rápido.

Para las variable ingresos se obsera un valor de curtosis negativo, lo que indica que la variable tiende a una distribución platicúrtica, existiendo colas anchas y un decrecimiento lento

Histograma: interpretación de las gráficas

- En el histograma de la variable consumo se puede por un lado observar la presencia de valores atípicos hacia la derecha e izquierda, y por otro lado, se observa un asimetría negativa teniendo colas mas alargadas hacia la derecha de la media.
- En el caso de la variable precio, se observa un valor atípico negativo hacia la izquierda, presentándose además una asimetría negativa, teniéndose colas más alargadas a la izquierda de la media.
- En el caso de la variable ingreso se observa una asimetría positiva, es decir, colas más alargadas hacia la derecha de la media.

Estadística Inferecial: Intervalos de confianza

Definición: Un intervalo de confianza para un parámetro desconocido θ de una distribución de probabilidad es un intervalo aleatorio de la forma $(\hat{\theta}_1, \hat{\theta}_2)$, en donde $\hat{\theta}_1$ y $\hat{\theta}_2$ son dos estadísticos que satisfacen:

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1 - \alpha$$

Intervalo de confianza para la media y varianza de una distribución normal

- La función empleada para el cálculo del intervalo de confianza para la media es t.interval(1-alpha,df,loc,scale)
- Para el calculo del intervalo de confianza de la varianza se empleó la formula vista en la teoría:

$$\frac{(n-1)}{\sigma^2}S^2 \sim \chi^2(n-1)$$

$$P(\chi^2_{\frac{\alpha}{2}} < \frac{S^2(n-1)}{\sigma^2} < \chi^2_{1-\frac{\alpha}{2}}) = 1-\alpha$$

$$\frac{(n-1).S^2}{\chi^2_{(1-\frac{\alpha}{2},n-1)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(\frac{\alpha}{2},n-1)}}$$

- Donde los cuantiles se obtuvieron mediante la funcion: chi2.ppf(p,df,loc=0,scale=1)

Salidas de Python: estimación de la media poblacional de las variables

INTERVALOS DE CONFIANZA PARA LA MEDIA POBLACIONAL

La media de consumo pertenece al intervalo: (4.475562186238289, 5.23922042245736 5) con una confianza del 95%

La media de precio pertenece al intervalo: (0.0332699165023 5647, 0.364990953062 8609) con una confianza del 95%

La media de ingresos pertenece al intervalo: (4.453794953149256, 5.044465916415962) con una confianza del 95%

Salidas de Python: estimacion de la varianza para las variables

INTERVALOS DE CONFIANZA PARA LA Varianza POBLACIONAL

La varianza de consumo pertenece al intervalo (0.02472531761120617 6,0.05701467509168827 6) con una confianza del 95%

La varianza de precio pertenece al intervalo (0.00466541011381202 1,0.01075807583106281) con una confianza del 95%

La varianza de ingresos pertenece al intervalo (0.01479227740821404 5,0.03410985062182593) con una confianza del 95%

Test de hipótesis: Dócimas para la media y varinza

- H0: $\mu=\mu 1$ vs H1: $\mu \neq \mu 1$ o H1: $\mu < \mu 1$ o H2: $\mu > \mu 1$
- Para el calculo se emplea la función ttest_1samp(variable,valor,alternative)
- Retorna el p-valor y el estadístico pivotal
- Para el caso de la varianza, se aplica la variable pivotal vista en la teoria

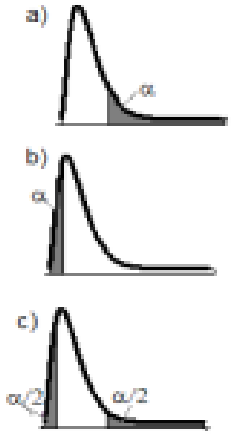
$$\chi^2_{H_0} = \frac{(n-1)s^2}{\sigma_0^2}$$

Reglas de decisión

$H_1: \sigma^2 > \sigma_0^2$ SE RECHAZA H_0 SI $\chi^2_{H_0} > \chi^2_{(n-1),(1-\alpha)}$

$H_1: \sigma^2 < \sigma_0^2$ SE RECHAZA H_0 SI $\chi^2_{H_0} < \chi^2_{(n-1),\alpha}$

$H_1: \sigma^2 \neq \sigma_0^2$ SE RECHAZA H_0 SI $\chi^2_{H_0} < \chi^2_{(n-1),(\alpha/2)}$ o $\chi^2_{H_0} > \chi^2_{(n-1),(1-\alpha/2)}$



Contrastes no paramétricos: Prueba de bondad de ajustes

- H0: “La variable sigue una distribución normal” vs H1: “La variable NO sigue una distribución normal”
- Para esta prueba de bondad de ajuste se emplea la funcion normaltest(dato)
- Que retorna el p-valor y el estadistico $s^2 + k^2$
- Esta prueba se basa en la prueba D’agostino k^2 cuya prueba se basa a su vez en las transformaciones de la prueba de curtosis y asimetría

Salidas de python: Prueba de bondad de ajuste- H0: “X tiene distribución normal” vs H1: “X NO tiene distribución normal”

-------Variable: consumo-------

NormaltestResult(statistic=0.24113113431143535, pvalue=0.8864189654696161)

Como el p-valor registrado es mayor que 0.05 podemos afirmar con una confianza del 95% que la variable paquetes sigue

una distribucion normal

-------Variable: precio-------

NormaltestResult(statistic=0.3990247564076488, pvalue=0.8191300813910641)

Como el p-valor registrado es mayor que 0.05 podemos afirmar con una confianza del 95% que la variable precio sigue una distribucion normal

-------Variable: ingresos-------

NormaltestResult(statistic=3.4159266455457726, pvalue=0.18123453323345426)

Como el p-valor registrado es mayor que 0.05 podemos afirmar con una confianza del 95% que la variable ingreso sigue una distribucion normal

Regresión lineal

- La regresión lineal es un modelo matemática que describe la relación de dependencia entre dos o más variables. Donde se tiene una variable Y dependiente, y m variables independientes o regresorias.
- El modelo lineal responde a la siguiente ecuación:
- $Y = \beta_0 + \beta_1 X + ... + \beta m\ Xm + \varepsilon$
- Donde: Y es la variable dependiente; Xi la variable dependiente para i=1..m ; Bi los parámetros o estimadores asociados a cada variable dependiente para i=1.m y ε (epsilon) la función de pérdida.

Regresión lineal simple: cálculo de la pendiente y ordenada al origen

- Se realiza la relación lineal simple entre dos variables: X variable independiente e Y variable dependiente, obteniendo los parámetros w y b del modelo lineal y=wx+b
- Para el cálculo se emplea la función linealregress(x,y) donde x e y son las variables a analizar
- Devuelve, entre otros valores el valor de w (pendiente) y b (ordenada al origen), stderr e intercept_stderr.
- Para la construcción del intérvalo de confianza para la pendiente se suma y resta al valor de la pendiente el error estándar de la misma.

Regresión lineal: salidas de python

Iniciando calculo de los parametros w y b

Modelo lineal simple: y=wx+b

LinregressResult(slope=0.2703334232048171, intercept=-1.0847182528983557, rvalue=0.4813626086039644, pvalue=0.0007073914185734368, stderr=0.07421020903021644, intercept_stderr=0.35259826348534895)

Los parametos finales son: w=0.27,b=-1.08

Intervalo de confianza del 95% para la pendiente: 0.2703334232048171 +/- 0.0504708482896445

Fin