Actividad por pares

Jairo Buitrago
jaabuitragov@correo.udistrital.edu.co

Septiembre 2020

Consideraciones generales

- La mayor parte del análisis fue desarrollado utilizando el lenguaje de programación Python, por tanto aunque se mostraran resultados buena parte de la evidencia entregada será el código implementado para el calculo de valores.
- Las gráficas también fueron generadas haciendo uso del mismo lenguaje, no se mostrará ninguna gráfica generada en CODAP, para estas no se añadirá el código como evidencia ya que lo necesario es la gráfica.
- la tabla entregada en el documento será identificada mediante la palabra "datos" en las secciones de código del documento.

1. Ejercicio

1.1. Descripción de la variable carne

Para empezar la descripción de esta variable hallaremos las distintos medidas de tendencia central y de variabilidad para darnos una idea inicial de como se comporta la variable. Para ello se utilizo el siguiente fragmento de código:

```
print("Sumatoria de la variable carne:",np.sum(datos.Carne))
print("Media de la variable carne:",datos.Carne.mean())
print("Mediana de la variable carne:",datos.Carne.median())
print("Desviación estandar de la variable carne:",np.std(datos.Carne))
```

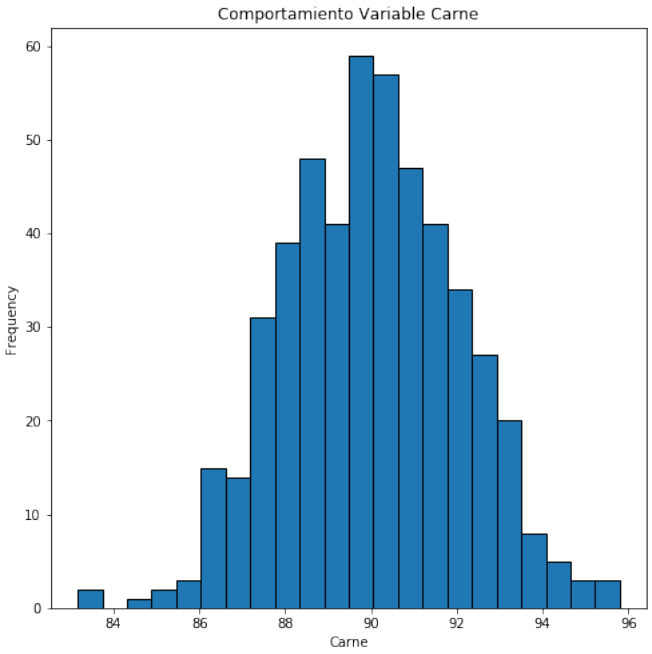
los resultados obtenidos fueron:

$$\sum x_i = 45010{,}9035\ gramos$$
$$\bar{x} = 90{,}022\ gramos$$
$$\tilde{x} = 90{,}026\ gramos$$
$$\sigma = 2{,}04\ gramos$$

Primero observamos que el valor de la media se encuentra en 90,022 gramos con una variación o desviación estándar de 2,04 gramos, por ultimo el valor más central o la mediana es de 90,026 gramos, vamos a explorar un poco la distribución de la variable realizando un histograma. Para ello calcularemos algunos datos básicos para saber la construcción del diagrama. (Solo calcularemos el número de intervalos, la amplitud de estos y el rango ya que son los necesarios para que el lenguaje utilizado genere el gráfico.)

$$Rango = 95{,}799 - 83{,}162 = 12{,}637 \approx 13$$
$$N^o\ intervalos = \sqrt{500} = 22{,}36 \approx 22$$
$$Ancho\ de\ intervalo = \frac{12{,}637}{22{,}36} = 0{,}57$$

Con estos datos podemos dibujar el histograma para mirar de manera más detallada el comportamiento de la variable.



Como se puede notar en el histograma hay 22 barras, una por cada intervalo, a primera vista la distribución se ve simétrica. Sin embargo se nota un intervalo alejado del resto de la distribución, vamos a calcular los cuartiles para construir un gráfico de caja , para ello se utilizó el siguiente código:

```
datos.Carne.quantile([0.25,0.50,0.75])
```

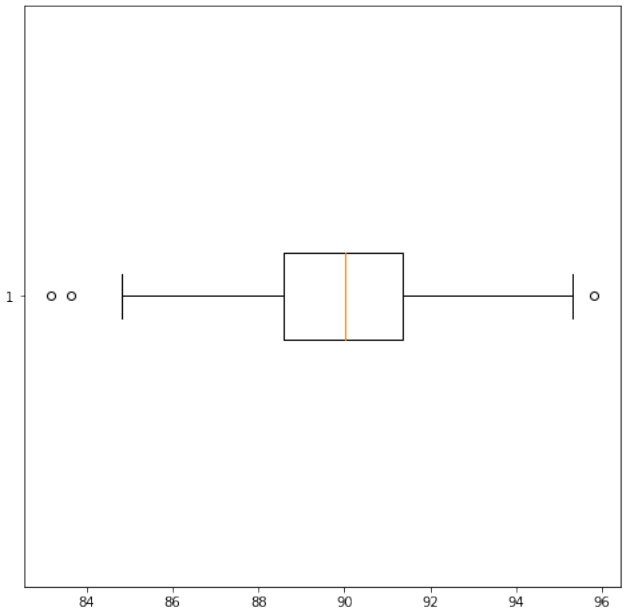
El resultado obtenido fue el siguiente:

$$0{,}25\ o\ Q_1 = 88{,}596$$

$$0{,}50\ o\ Q_2 = 90{,}026$$

$$0{,}75\ o\ Q_3 = 91{,}369$$

Al realizar el gráfico de caja obtenemos:



En el gráfico se notan algunos datos atípicos por tanto podemos deducir que que la mediana puede ser una tendencia a ser tomada más en cuenta que la media ya que esta se puede ver sesgada por la presencia

de estos datos. Como otra posible interpretación del gráfico dada la posición de la caja es posible afirmar que los datos tienen una asimetría hacia la izquierda, esto quiere decir que los datos tienden a estar concentrados más a la derecha de la media que a la izquierda.

1.2. Descripción de la variable Salsa

Para el análisis de la variable salsa haremos algo parecido a lo que se hizo con la variable carne, primero medidas de tendencia central y variabilidad, para esto se implemento el siguiente código:

```
print("Sumatoria de la variable salsa:",np.sum(datos.Salsa))
print("Media de la variable salsa:",datos.Salsa.mean())
print("Mediana de la variable salsa:",datos.Salsa.median())
print("Desviación estandar de la variable salsa:",np.std(datos.Salsa))
```

Los resultados obtenido para este caso fueron los siguientes:

$$\sum x_i = 2804{,}055$$

$$\bar{x} = 5{,}608$$

$$\tilde{x} = 5{,}6$$

$$\sigma = 0{,}234$$

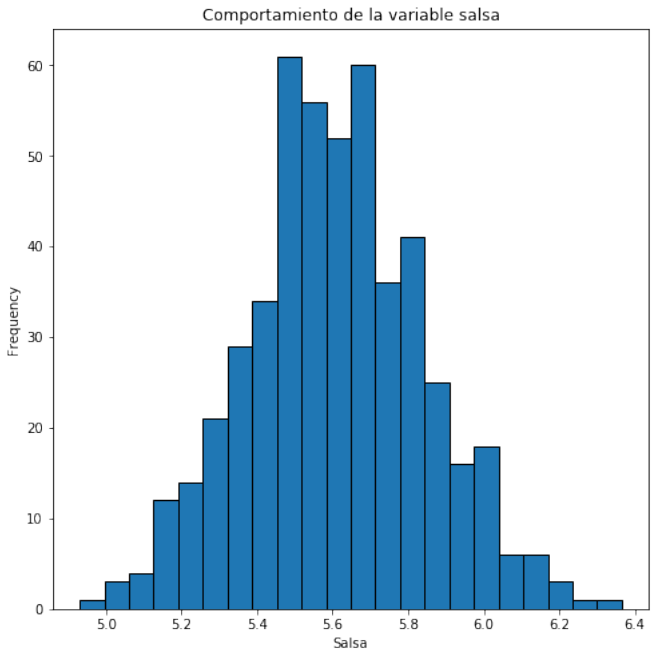
Después de haber obtenido los valores de la media, la mediana y la desviación estándar nos damos una idea de en que datos, aproximadamente, se encuentran los valores de la variable. Ahora al igual que con la variable carne calcularemos los valores necesarios para hacer un histograma:

$$Rango = 6{,}366 - 4{,}93 = 1{,}436$$

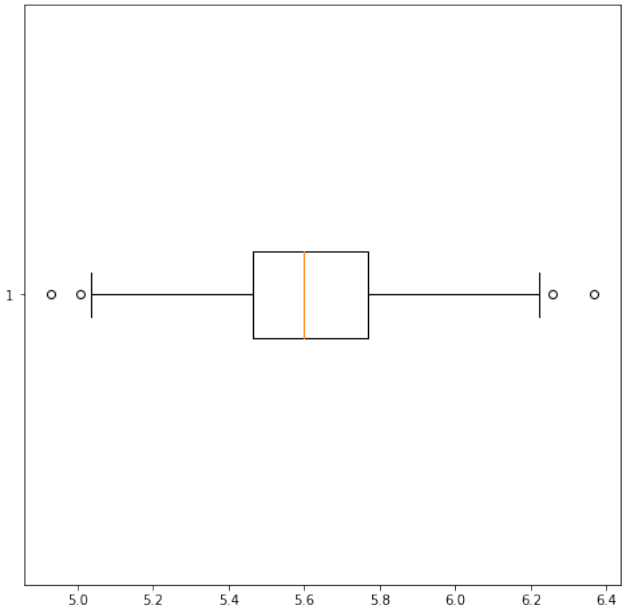
$$N^o\ intervalos = \sqrt{500} = 22{,}36 \approx 22$$

$$Ancho\ de\ intervalo = \frac{1{,}436}{22{,}36} = 0{,}064$$

Con estos valores generaremos el histograma para ver la distribución de esta variable:



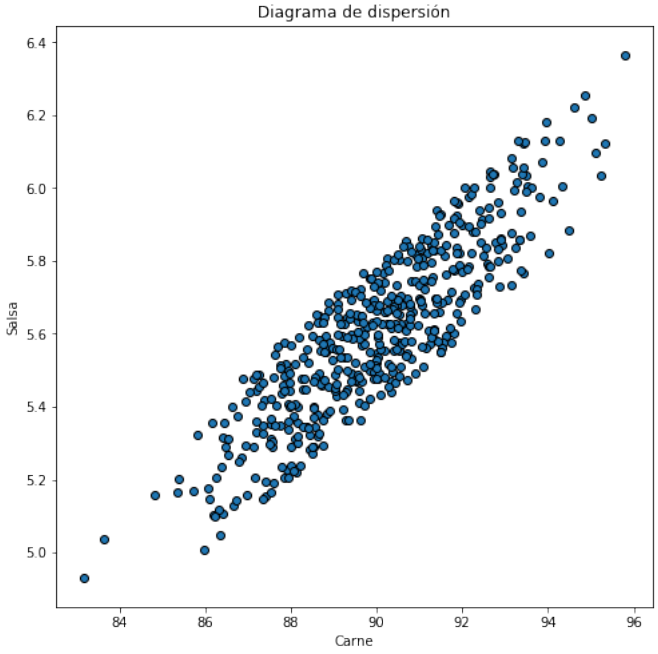
A primera vista este gráfico también se ve simétrico, sin embargo algunos intervalos centrales tienen una frecuencia menor comparativamente con sus intervalos contiguos. Nuevamente utilizaremos el gráfico de caja para ver que si existen datos atípicos y notar hacia donde tienden a estar un poco más los valores de la variable.



Podemos notar que esta variable tiene algunos datos atípicos asi com también que si distribución es simétrica con una pequeña tendencia de los datos hacia la izquierda.

1.3. Análisis conjunto de las variables Salsa y Carne

Para empezar el análisis conjunto haremos un diagrama de dispersión que nos permita ver si las variables se pueden ajustar a algún tipo de función.



Se puede ver que los datos se pueden ajustar a una linea recta, por tanto construimos una recta de ajuste para indicar la relación de estas variables, estableceremos $x = Carne$ y $y = Salsa$:

$$\sum x_i = 45010{,}9035 \quad \sum y_i = 2804{,}0550$$

$$\sum x_i y_i = 252636{,}5895 \quad \sum x_i^2 = 4054044{,}0967 \quad \sum y_i^2 = 15752{,}8347$$

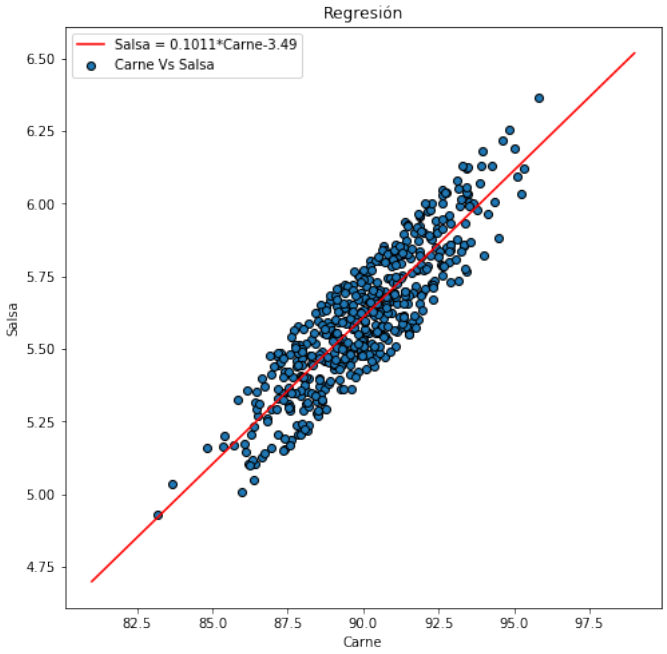
Formamos un sistema de ecuaciones para hallar los valores de a y b que nos permitan describir la ecuación de la recta de ajuste:

$$\begin{aligned} 2804{,}0550 = &\quad 500a + 45010{,}9035b \\ 252636{,}5895 = &\quad 45010{,}9035a + 4054044{,}0967b \end{aligned}$$

Al resolver este sistema obtenemos que $a = -3{,}49$ y $b = 0{,}1011$. Por tanto el sistema queda remplazando x y y por las variables Carne y Salsa:

$$Salsa = 0{,}1011 * Carne - 3{,}49$$

Al gráficar esta recta de ajuste con el diagrama de de dispersión obtenemos:



Si calculamos el coeficiente de correlación obtenemos un valor de $r = 0{,}8815$ por tanto podemos decir que la recta se ajusta de forma aceptable a los datos.

1.4. Descripción del comportamiento de las variables Papas y Refresco

Para el análisis de esto haremos una tabla de doble entrada, para ello lo primero que hallaremos será el total de las Papas y los refrescos de cada categoría, para ello se implemento el siguiente código:

```
datos.Papas.value_counts()
datos.Refresco.value_counts()
```

El resultado obtenido fue el siguiente:

Papas		Refresco	
Chicas	87	Chico	167
Medianas	247	Mediano	250
Grandes	166	Grande	83

utilizaremos el siguiente código para calcular cuantos Refrescos existen asociados a un tipo de Papas:

```
datos.groupby("Papas").Refresco.value_counts()
```

El resultado Obtenido fue:

Papas	Refresco	Conteo
Chicas	Mediano	36
	Chico	35
	Grande	16
Grandes	Mediano	80
	Chico	54
	Grande	32
Medianas	Mediano	134
	Chico	78
	Grande	35

Con estos datos construimos la tabla de doble entrada que contemplara todo lo calculado, además nos dejara hallar la matriz de probabilidad conjunta:

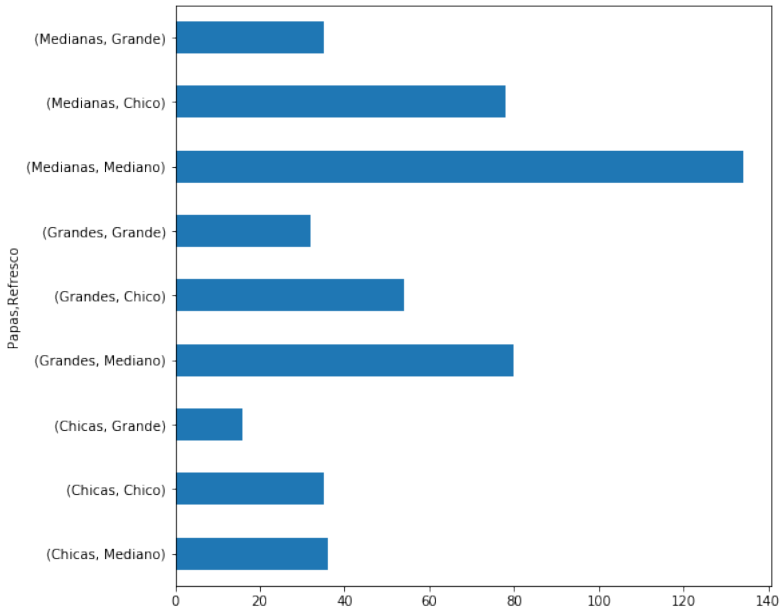
		Papas			
		Grandes	Medianas	Chicas	Total
Refresco	Grande	32	35	16	83
	Mediano	80	134	36	250
	Chico	54	78	35	167
	Total	166	247	87	500

Finalmente con esta tabla de doble entrada podemos construir la matriz de probabilidad conjunta:

		Papas			
		Grandes	Medianas	Chicas	Total
Refresco	Grande	0.064	0.07	0.032	0.166
	Mediano	0.16	0.268	0.072	0.5
	Chico	0.108	0.156	0.07	0.334
	Total	0.332	0.494	0.174	500

En esta ultima matriz se puede evidenciar la probabilidad de ocurrencia de dos eventos al mismo tiempo así como también las probabilidades marginales de cada cada evento.

Finalmente para facilitar la visualización de esta tabla vamos a observar este comportamiento mediante un gráfico:



Por ultimo se puede concluir que la combinación de papas y refresco con mayor frecuencia es la de "Medianas, Mediano".

2. Ejercicio 2

2.1. ¿Cuánto vale el coeficiente de correlación entre las variables Carne y Salsa?

El coeficiente de correlación será igual a:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n * (\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2\right]\left[n * (\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} y_i)^2\right]}}$$

$$r = \frac{500 * 252636{,}59 - 45010{,}90 * 2804{,}06}{\sqrt{[500 * 4054044{,}1 - 45010{,}90^2][500 * 15752{,}83 - 2804{,}06^2]}} = \frac{105030{,}746}{119254{,}93} = 0{,}88$$

Este coeficiente nos indica que la recta tiene un buen ajuste a los datos.

2.2. ¿Qué cantidad de salsa, en gramos, se esperaría que un cliente le ponga a su hamburguesa si ésta tiene 89 gramos de carne? Redondea a dos decimales.

Teniendo en cuenta la ecuación de la recta de ajuste:

$$Salsa = 0{,}1011 * Carne - 3{,}49$$

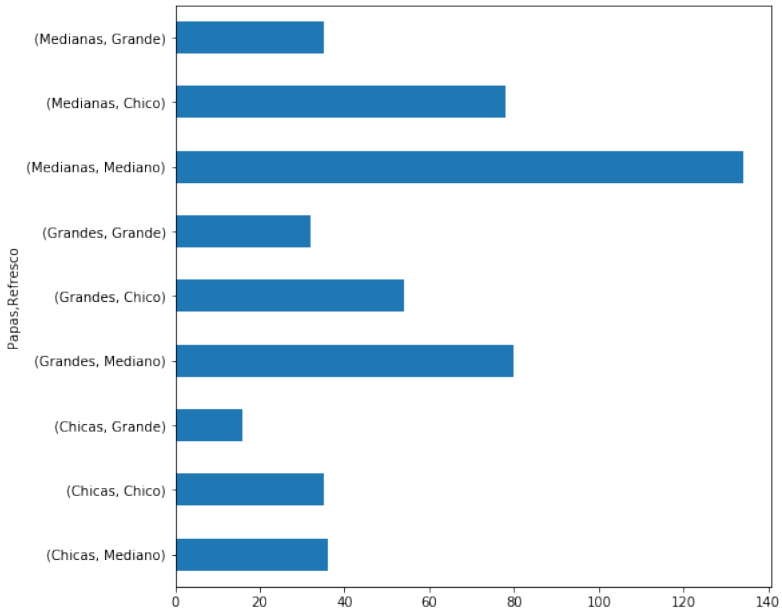
Reemplazando los valores obtenemos

$$Salsa = 0{,}1011 * 89 - 3{,}49 = 5{,}5\ gramos$$

Por tanto la cantidad de salsa que se esperaría pida un cliente si su hamburguesa es de 89 gramos es de 5.5 gramos.

2.3. ¿Qué combinación de papas y refresco es la más frecuente?

Para esta pregunta y la siguiente se anexa el gráfico generado a partir de la tabla de doble entrada



Con la tabla de doble entrada y el gráfico hechos en el ejercicio 1 podemos decir que la combinación de papas y Refresco más frecuente es la de "Papas medianas y Refresco mediano".

2.4. ¿Qué combinación de papas y refresco es la menos frecuente?

De acuerdo al gráfico y la tabla de doble entrada es evidente que la combinación de menos frecuencia es la de "papas chicas y refresco grande".

2.5. Calcula la probabilidad que hay de que un cliente seleccionado al azar haya pedido...

Para esta sección del ejercicio será añadida la matriz de probabilidad conjunta hallada en el ejercicio 1

		Papas			
		Grandes	Medianas	Chicas	Total
Refresco	Grande	0.064	0.07	0.032	0.166
	Mediano	0.16	0.268	0.072	0.5
	Chico	0.108	0.156	0.07	0.334
	Total	0.332	0.494	0.174	500

a. Papas medianas.

$$P(papas\ medianas) = \frac{247}{500} = 0{,}494$$

La probabilidad de que un cliente pida papas medianas es de 49.4 %

b. Papas medianas o refresco chico.

$$P(papas\ medianas\ o\ refresco\ chico) = P(papas\ medianas) + P(refresco\ chico) - P(papas\ medianas\ y\ refresco\ chico)$$

Reemplazando

$$P(papas\ medianas\ o\ refresco\ chico) = 0{,}49 + 0{,}33 - 0{,}156 = 0{,}664$$

La probabilidad de que un cliente pida papas medianas o refresco chico es de 66.4 %

c. Papas grandes y refresco chico.

$$P(papas\ grandes\ y\ refresco\ chico) = \frac{54}{500} = 0{,}108 \approx 0{,}11$$

La probabilidad de que un cliente pida papas grandes y refresco chico es de 11 %

d. Refresco chico si ya pidió papas grandes

$$P(refresco\ chico|papas\ grandes) = \frac{P(papas\ grandes\ y\ refresco\ chico)}{P(papas\ grandes)}$$

Reemplazando

$$P(refresco\ chico|papas\ grandes) = \frac{0{,}11}{0{,}332} = 0{,}331$$

La probabilidad de que un cliente pida un refresco chico dado que ya pidio papas grandes es de 33.1 %

2.6. ¿Los eventos papas grandes y refresco grande son independientes? Sí, No y Por qué.

Sabemos que $P(papas\ grandes\ y\ refresco\ grande) = 0{,}064$ si los eventos son independientes debe cumplirse la igualdad:

$$P(papas\ grandes\ y\ refresco\ grande) = P(papas\ grandes)P(refresco\ grande) = 0{,}064$$

$$P(papas\ grandes)P(refresco\ grande) = 0{,}332 * 0{,}166 = 0{,}055$$

Como no se cumple la igualdad podemos concluir que los eventos no son independientes, por tanto podemos decir que la ocurrencia de un evento puede afectar al otro.

Por ultimo dejo el enlace por si desea mirar toda la implementación del código para el análisis:
`https://github.com/jairosam/Smulacion/blob/master/An%C3%A1lisis%20DataSet%20Hamburguesas.ipynb`