Banjercito

Diplomado en Ciencia de Datos

Notas del Curso

Módulo III — Semana 1

Instructor: Alan Badillo Salas

Agosto 2024

1 Introducción

En el Módulo I hemos profundizado los conceptos más importantes del uso de Excel y Power BI, como el tablas dinámicas en Excel y la adquisición de datos mediante Power Query y generación de informes en PoweBI.

En el Módulo II hemos aprendido los concetos fundamentales de *Python* como el manejo de listas, series y tablas de pandas. Así como una introducción a la probabilidad y estadística de las que partiremos en este curso.

En este módulo III *"Ciencia de Datos"* comenzaremos por revisar los conceptos más importantes de la estadística como los tipos de análisis cuantitativos y cualitativos, las medidas de tendencia central, las medidas de dispersión, las distribuciones de probabilidad, la inferencia estadística, la regresión y correlación, el análisis de la varianza (ANOVA) y la estadística descriptiva, exploratoria y muestreos.

Adicionalmente, profundarizaremos en el uso de Excel, PowerBI y Excel según sea conveniente en cada estudio o ambos en caso de que se quieran dominar ambas herramientas al mismo tiempo.

1.1 Contenido de la Semana 1

Esta semana revisaremos los siguientes temas:

1. Tipos de análisis estadístico
   (a) Análisis cuantitativo
   (b) Análisis cualitativo
2. Medidas de tendencia central
   (a) Media
   (b) Mediana
   (c) Moda
   (d) Cuartiles
   (e) Percentiles

2 Tipos de análisis estadístico

El análisis estadístico es una técnica que permite obtener información de los datos relativa a los mismos datos, por ejemplo, si poseemos una lista de edades, podríamos describir la edad mínima, la edad máxima, la edad promedio, la edad más repetida e incluso que tanto se dispersa una edad respecto a otra. A estos valores los llamaremos estadísticos y su base de fondo será la probabilidad de que un evento pueda ocurrir.

Existen idealmente dos tipos de datos que pueden ser analizados mediante la estadística y son los datos cuantitativos o que se pueden cuantificar, por ejemplo, sumar, promediar, escalar, y los datos cualitativos o que no se pueden dimensionar y se refieren a características de relaciones entre los datos, por ejemplo, a qué categoría o grupo pertenecen, la valoración que da un usuario a un producto o el nivel de satisfacción en una compra.

En general haremos un *Análisis Cuantitativo* si los datos están relacionados a valores numéricos o un *Análisis Cualitativo* si los datos son categóricos o relacionales.

2.1 Análisis cuantitativo

El análisis cuantitativo se refiere a un análisis sobre datos numéricos que pueden ser continuos como el precio de la gasolina, la estatura de una persona, el peso de un producto o la temperatura del ambiente. También sobre datos numéricos que pueden ser discretos como la edad de una persona, el número de hijos de un trabajador o el número de horas extras trabajadas.

En en análisis cuantitativo los valores continuos pueden tomar valores intermedios unos de otros, por ejemplo, podríamos tener que una persona pesa 64.5kg y otra 72.8kg, pero en cualquier momento podríamos recibir un dato intermedio de una persona que pesa 67.9kg. Esto implica que no siempre tendremos valores repetidos y la probabilidad de que una persona pese 65kg o 70kg sea más difícil de calcular.

En el análisis cuantitativo los valores discretos no pueden tomar valores intermedios, generalmente son números enteros que toman valores finitos o infinitos, pero no intermedios, por ejemplo, la edad de una persona teoricamente es finita, pero también se podría ver como posiblemente infinita, en fines prácticos nadie podría tener 23.7 años ya que o tiene 23 años o 24 años cumplidos y en algún momento podría tomar un valor de 140 años si se prolongara la vida de una persona. Casi siempre ocurren estos tipos de casos, por ejemplo, el número de hijos que tiene un trabajador podría ser útil para la empresa para otorgar un crédito o aumentar su salario. Y aunque es más probable que tenga 0, 1, 2, 3 o incluso 4, podría a llegar a tener 8, 10 o 12 hijos. Este tipo de complejidad debe ser considerada al analizar los datos discretos y aunque se asume que alguien no podría tener 1.4 hijos, tampoco se conoce el número máximo de hijos.

2.2 Análisis cualitativo

El análisis cualitativo consiste en establecer información sobre propiedades relacionales de los datos, por ejemplo, para un carro algunos valores cualitativos serían: marca, modelo, color, valoración del cliente, nivel de satisfacción de la compra etc. Generalmente se asocian los análisis cualitativos con categorías o valoraciones (rating o rankeo). Se podría pensar que la valoración de un cliente o el nivel de satisfacción es un valor cuantitativo, por ejemplo, si la valoración de un cliente es del 1 al 5, podríamos pensar que es algo cuantitativo y discreto o continuo, pero en realidad, la valoración de un cliente es algo subjetivo que no se midió directamente del carro y no es una propiedad natural de carro, sino que es un valor relacionado o asociado al carro de manera inderecta.

Los análisis cualitativos sirven generalmente para dividir los datos o agruparlos en categorías o grupos de categorías, por ejemplo, todos los carros color rojo o todos los carros con una valoración del cliente mayor a 3. Esto permitirá hacer análisis cuantitativo en subespacios enfocados y ajustados a cualidades del objeto de análisis, por ejemplo, cuál es el promedio del precio (algo cuantitativo) sobre los carros con un nivel de satisfacción menor a 2 (algo cualitativo).

3 Medidas de Tendencia Central

Los datos cuantitativos (numéricos) poseen características matemáticas que podemos calcular para poder entender mejor la naturaleza del dato.

Las principales medidas de tendencia central son:

- **Media** — Represental el valor promedio de un conjunto de valores y es el elemento con la menor distancia a cada uno de los datos, por ejemplo, si tenemos un conjunto de edades, la edad promedio será la que represente mejor al grupo, ya que estará a la menor distancia posible de cada una de las edades.

- **Mediana** — Representa el valor medio de un conjunto de valores ordenados, y es el valor que queda en medio o a la mitad de todos los valores, si el número de valores es par se tomará el promedio de los dos valores centrales. La mediana podría ser diferente a la media y su sentido es entender qué tan sesgados están los datos. Si la mediana es menor a la media significará que hay muchos valores pequeños en el conjunto de valores y por lo tanto un sesgo a la izquierda y si la mediana es mayor a la media significará que los valores grandes dominan más de la mitad del conjunto de valores y el sesgo estará a la derecha. Esto nos sirve para determinar si los datos estarán centrados o desplazados.

- **Moda** — Representa el valor más repetido o de mayor frecuencia de un conjunto de valores ordenados por frecuencia de repetición. Este generalmente se usa en datos discretos e indica cuál es el elemento que más se repite. Con esto podríamos entender mejor los datos, por ejemplo, ordenarlos por frecuencia de aparición y obtener diferentes modas (la primera, segunda, tercera y así sucesivamente).

- **Cuartiles** — Representa el valor promedio a un 25% (Primer Cuartil — Q_1), 50% (Segundo Cuartil — Q_2) o 75% (Tercer Cuartil — Q_3). Estos son importantes para construir la caja estadística que representa el espacio donde vive la mayoría de los datos. Así el 25% de los datos nos dará un límite inferior que nos dirá que abajo de ese valor promedio vive solo el 25% de los datos y el 75% nos dará de manera similar el límite superior que nos dirá cuál es el 25% de los datos que viven arriba de ese límite. De esta manera nos quedamos con el 50% de los datos y podemos ver si el sesgo está cargado hacia el 75% de los datos (sesgo a la derecha) o cargado hacia el 25% de los datos (sesgo a la izquierda).

- **Percentiles** — Representa la generalización del promedio de datos a cierto porcentaje de la población, por ejemplo, al 30% representará el valor promedio al 30% de la población. Para obtenerlo debemos ordenar el conjunto de valores del menor al mayor y ver qué porcentaje de progreso tiene, luego calcular el valor promedio al x%.

3.1 Media

Para calcular la media basta con sumar todos los valores del conjunto y dividir dicha suma entre el número total de elementos en el conjunto. Esto representará un valor único para todo el conjunto que aproxime el valor ideal de lo que debería

valer cada dato del conjunto si todos representaran una misma idea o naturaleza.

La ecuación para calcular el valor medio es:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

La ecuación (1) representa el valor promedio (valor medio) de los datos y geométricamente es el que queda al centro de todos los datos (en un espacio euclidiano). Este valor sirve para conocer dónde se encuentra el valor central que podría representar a todo un conjunto de valores, si es que su dispresión es baja.

3.1.1 Ejemplo para calcular la media

Supongamos que tenemos la siguiente lista de valores:

Edad
23
24
18
20
27
35

Table 1: Conjunto de edades

La edad promedio se calcula como:

$$\begin{aligned} \overline{edad} &= \frac{1}{6} \sum_{i=1}^{6} Edad_i \\ &= \frac{23 + 24 + 18 + 20 + 27 + 35}{6} \\ &= \frac{147}{6} \\ &= 24.5 \end{aligned} \tag{2}$$

Entonces la edad promedio $\overline{edad} = 24.5$ y representa la edad más cercana a todas las edades. También nos indica que las edades están cercanas a los 24.5 años y aunque podríamos pensar que nadie puede tener estrictamente 24.5 años, la edad media si puede y significa que entre 24 y 25 años tendría una persona en promedio.

3.2 Mediana

Para calcular la mediana primero debemos ordenar los datos y encontrar el valor del centro, si el número de valores es par habrán dos centros y entonces

la mediana será el promedio de estos dos valores centrales, sino, si es impar habrá un único centro y será la mediana directamente. El valor de la mediana representa el valor central o el valor que tomaría aproximadamente el 50% de la población, por ejemplo, si ordenamos a todos los alumnos de un curso de secundaria por estatura de la menor a la mayor, la estatura central nos indicaría que tan alto es el alumno que se encuentra en medio de todos los alumnos, si este alumno es más alto que el promedio, entonces los alumnos serían altos en general, pero si es más bajo que el promedio, esto indicaría que muchos de los alumnos son más bajos que el promedio ideal.

La ecuación de la mediana se calcula como:

$$\widehat{x} = \begin{cases} x_p & N \equiv 1 \pmod 2, p = \frac{N+1}{2} \\ \frac{x_p + x_{p+1}}{2} & N \equiv 1 \pmod 2, p = \frac{N}{2} \end{cases} \tag{3}$$

La ecuación (3) representa el valor de en medio de los datos ordenados e indica qué tan bajo o alto es respecto al promedio para entender si los datos se cargan a la izqueirda o a la derecha.

3.2.1 Ejemplo para calcular la mediana

Supongamos que tenemos la siguiente lista de valores:

Edad
23
24
18
20
27
35

Table 2: Conjunto de edades

Ordenando los valores de menor a mayor tenemos:

Edad	
1	18
2	20
3	23
4	24
5	27
6	35

Table 3: Conjunto de edades ordenadas de menor a mayor

La edad mediana se calcula como (caso par para 6 valores):

$$\begin{aligned} p &= \frac{6}{2} = 3 \\ \widehat{edad} &= \frac{edad_3 + edad_4}{2} \\ &= \frac{23 + 24}{2} \\ &= \frac{47}{2} \\ &= 23.5 \end{aligned} \tag{4}$$

Entonces la edad mediana $\widehat{edad} = 23.5$ y representa el centro de las edades. Vemos que es ligeramente menor al promedio por lo que los datos están ligeramente sesgados a la izqueirda, esto significa que hay más edades menores al promedio que edades mayores al promedio. Por lo que podemos asumir que hay mayor población más joven que el promedio. Es decir, 4 personas son menores al promedio mientras solo 2 son mayores al promedio.

3.3 Moda

Para calcular la moda hay que ordenar todos los datos por frecuencia en que se repiten, esto suele funcionar únicamente en valores discretos, ya que dos valores continuos podrían no repetirse, por ejemplo, si tenemos el precio del gas de \$17.36 y otro de \$17.37 serán distintos idealmente. Sin embargo, podríamos crear intervalos o *bins* como veremos más adelante en la construcción de un histograma.

La ecuación de la moda se calcula como:

$$\dot{x} = x_m, max_{m=1..N} freq(x_m) \tag{5}$$

La ecuación (5) representa el valor que más se repite, es decir de mayor frecuencia de repetición ($freq(x_m)$). Este valor de moda indica cuál es el valor que más se repite en el conjunto de datos, se usa mucho para saber cuál es el producto más vendido, la edad más común o el número de hijos con mayor repetición en el conjunto de valores.

3.3.1 Ejemplo para calcular la moda

Supongamos que tenemos la siguiente lista de valores:

Índice	**Edad**	**Índice**	**Edad**
1	**18**	11	**18**
2	**18**	12	**20**
3	**19**	13	**28**
4	**18**	14	**20**
5	**18**	15	**19**
6	**19**	16	**18**
7	**18**	17	**24**
8	**20**	18	**18**
9	**18**	19	**20**
10	**18**	20	**18**

Table 4: Conjunto de las edades del grupo de primer año de licenciatura

Podemos observar que muchas de las edades se repiten, sobre todo las de 18 años.

Si ordenamos los datos por frecuencia de mayor a menor tenemos:

Edad	**Frecuencia**
18	**11**
20	**4**
19	**3**
24	**1**
28	**1**

Table 5: Conjunto de las edades del grupo de primer año de licenciatura, ordenados por frecuencia descendente

Podemos observar que la moda es $\dot{edad} = 18$, sin embargo podemos ver que el siguiente valor más repetido es 20 años y luego 19 años con $11, 4, 3$ repticiones respectivamente. Esto nos daría las primeras 3 modas y sabríamos que 90% de los datos son $18, 20, 19$.

3.4 Cuariles

Para calcular los tres cuartiles Q_1, Q_2, Q_3 necesitamos ordenar los datos usaremos una idea similar a la mediana para ubicar el dato al $Q_1 \to 25\%$, $Q_2 \to 50\%$ (corresponde a la mediana) y $Q_3 \to 75\%$.

La ecuación del primer cuartil Q_1 se calcula como:

$$Q_1 = x_a, \quad a = \left\lceil \frac{N}{4} \right\rceil \tag{6}$$

La ecuación del segundo cuartil Q_2 se calcula como:

$$Q_2 = x_b, \quad b = \left\lceil \frac{N}{2} \right\rceil \tag{7}$$

La ecuación del tercer cuartil Q_3 se calcula como:

$$Q_3 = x_c, \quad c = \left\lceil \frac{3 \cdot N}{4} \right\rceil \tag{8}$$

Para ser más precisos si hay dos centros cercanos al pivote p debemos promediar dichos valores.

3.4.1 Ejemplo para calcular los cuartiles

Supongamos que tenemos la siguiente lista de valores:

Índice	**Edad**	**Índice**	**Edad**
1	**18**	11	**18**
2	**18**	12	**20**
3	**19**	13	**28**
4	**18**	14	**20**
5	**18**	15	**19**
6	**19**	16	**18**
7	**18**	17	**24**
8	**20**	18	**18**
9	**18**	19	**20**
10	**18**	20	**18**

Table 6: Conjunto de las edades del grupo de primer año de licenciatura

Si ordenamos los datos de menor a mayor (de forma ascendente) tenemos:

Índice	**Edad**	**Índice**	**Edad**
1	**18**	11	**18**
2	**18**	12	**19**
3	**18**	13	**19**
4	**18**	14	**19**
5	**18**	15	**20**
6	**18**	16	**20**
7	**18**	17	**20**
8	**18**	18	**20**
9	**18**	19	**24**
10	**18**	20	**28**

Table 7: Conjunto de las edades del grupo de primer año de licenciatura ordenados

Entonces el valor de los cuartiles serán:

$$\begin{aligned} Q_1 = x_a = x_5 = \mathbf{18}, \quad a = \left\lceil \frac{20}{4} \right\rceil = 5 \\ Q_2 = x_b = x_{10} = \mathbf{18}, \quad b = \left\lceil \frac{20}{2} \right\rceil = 10 \\ Q_3 = x_c = x_{15} = \mathbf{20}, \quad c = \left\lceil \frac{3 \cdot 20}{4} \right\rceil = 15 \end{aligned} \tag{9}$$

Así, podemos observar que la edad para el 25% y 50% de la población sigue siendo 18 años, esto significa que las edades están cargadas hacia los 18 años en su 50% y apenas alcanza los 20 años al 75% de la población.

3.5 Percentiles

Extendiendo la idea los cuartiles que indican el valor medio en los porcentajes 25%, 50% y 75% podemos calcular el valor central para cada porcentaje desde el 0% hasta el 100%. Esto significa que cada valor tiene asociado un porcentaje de progreso y podemos calcular cada promedio relacionado a un porcentaje. Aquí los datos deben estar ordenados de menor a mayor.

La ecuación del percentil P_j con $j = 0, 1, 2, ..., 100$ se calcula como:

$$P_j = \overline{x_i}, \forall \left\lceil \frac{100 \cdot i}{N} \right\rceil = j \tag{10}$$

Esto significa que P_j será el promedio de todos los valores x_i tales que su progreso o porcentaje de avance del índice $i = 1...N$ sea igual a j en valor su redondeado.

3.5.1 Ejemplo para calcular el percentil

Supongamos que tenemos la siguiente lista de valores:

Índice	**Edad**	**Índice**	**Edad**	**Índice**	**Edad**	**Índice**	**Edad**
1	**18**	26	**19**	51	**19**	76	**18**
2	**18**	27	**19**	52	**19**	77	**20**
3	**19**	28	**20**	53	**20**	78	**22**
4	**18**	29	**20**	54	**20**	79	**21**
5	**18**	30	**19**	55	**19**	80	**19**
6	**19**	31	**18**	56	**18**	81	**22**
7	**18**	32	**21**	57	**21**	82	**20**
8	**20**	33	**19**	58	**19**	83	**23**
9	**18**	34	**22**	59	**22**	84	**19**
10	**18**	35	**20**	60	**20**	85	**22**
11	**18**	36	**19**	61	**19**	86	**21**
12	**20**	37	**20**	62	**20**	87	**20**
13	**28**	38	**18**	63	**18**	88	**19**
14	**20**	39	**22**	64	**22**	89	**20**
15	**19**	40	**21**	65	**21**	90	**22**
16	**18**	41	**20**	66	**20**	91	**21**
17	**24**	42	**19**	67	**19**	92	**23**
18	**18**	43	**22**	68	**22**	93	**20**
19	**20**	44	**18**	69	**18**	94	**22**
20	**18**	45	**23**	70	**23**	95	**21**
21	**18**	46	**20**	71	**20**	96	**22**
22	**19**	47	**21**	72	**21**	97	**19**
23	**18**	48	**19**	73	**19**	98	**21**
24	**22**	49	**20**	74	**20**	99	**20**
25	**20**	50	**21**	75	**21**	100	**23**

Table 8: Conjunto de las edades de los grupos primer año de licenciatura

Esta tabla contiene 100 edades, sin embargo, la mayoría de las veces tendremos más valores por lo que al ordenar más de un valor caerá en el mismo percentil.

Si ordenamos los valores de menor a mayor:

Índice	**Edad**	**Índice**	**Edad**	**Índice**	**Edad**	**Índice**	**Edad**
1	**18**	26	**19**	51	**20**	76	**21**
2	**18**	27	**19**	52	**20**	77	**21**
3	**18**	28	**19**	53	**20**	78	**21**
4	**18**	29	**19**	54	**20**	79	**21**
5	**18**	30	**19**	55	**20**	80	**21**
6	**18**	31	**19**	56	**20**	81	**22**
7	**18**	32	**19**	57	**20**	82	**22**
8	**18**	33	**19**	58	**20**	83	**22**
9	**18**	34	**19**	59	**20**	84	**22**
10	**18**	35	**19**	60	**20**	85	**22**
11	**18**	36	**19**	61	**20**	86	**22**
12	**18**	37	**19**	62	**20**	87	**22**
13	**18**	38	**19**	63	**20**	88	**22**
14	**18**	39	**19**	64	**20**	89	**22**
15	**18**	40	**19**	65	**20**	90	**22**
16	**18**	41	**19**	66	**20**	91	**22**
17	**18**	42	**19**	67	**20**	92	**22**
18	**18**	43	**20**	68	**21**	93	**22**
19	**18**	44	**20**	69	**21**	94	**23**
20	**18**	45	**20**	70	**21**	95	**23**
21	**19**	46	**20**	71	**21**	96	**23**
22	**19**	47	**20**	72	**21**	97	**23**
23	**19**	48	**20**	73	**21**	98	**23**
24	**19**	49	**20**	74	**21**	99	**24**
25	**19**	50	**20**	75	**21**	100	**28**

Table 9: Conjunto de las edades de los grupos primer año de licenciatura

Como la tabla coíncide en 100 valores y 100 porcentajes, es fácil calcular cada percentil, el cuál será el de su índice. Por ejemplo, el percentil 17 es $P_{17} = 18$ y el percentil 65 es $P_{65} = 20$ y el perccentil 95 es $P_{95} = 23$.

Si la tabla tuviera 500 valores, entonces el índice 17 ya no correspondería al percentil 17, sino que sería $\lceil \frac{17}{500} \rceil = 3$, que corresponde al percentil al 3% de progreso. Pero entonces los índices $i = 15, 16, 17, 18, 19$ también corresponden al percentil 3%. Por lo que el percentil 500 sería $P_{500} = \frac{x_{15}+x_{16}+x_{17}+x_{18}+x_{19}}{5}$, esto significa el promedio de todos los valores cuyo índice tienen una progresión al 3%.

4 Problemas

Los siguientes problemas desarrollan los conceptos expuestos. Resuelve al menos un problema y presenta los resultados de la solución indicando el proceso completo para resolverlo.

4.1 Problema 1 — Ejemplos de tipos de análisis

Muestra un ejemplo de al menos 10 valores para los datos cuantitativos o cualitativos listados, sin usar los ejemplos siguientes:

- **Datos cuantitativos y continuos** — Por ejemplo, el precio del gas en 10 días distintos.
- **Datos cuantitativos y discretos** — Por ejemplo, la edad de 20 personas.
- **Datos cualitativos de tipo categoría** — Por ejemplo, el departamento al que pertenecen 30 productos.
- **Datos cualitativos de tipo etiqueta** — Por ejemplo, el color de 14 automóviles.
- **Datos cualitativos de tipo valoración** — Por ejemplo, la valoración del 1 al 5 de 50 películas diferentes.
- **Datos cualitativos de tipo satisfacción** — Por ejemplo, la satisfacción *"Muy Satisfecho"*, *"Satisfecho"*, *"Poco Satisfecho"* o *"Insatisfecho"* para 17 clientes que reciben un servicio.

Construye una hoja de excel para cada ejemplo propuesto y sus valores, por ejemplo:

Día	**Precio del Gas**
1	17.45
2	17.95
3	17.23
4	16.13
5	19.21
6	19.32
7	18.29
8	17.67
9	15.12
10	11.98

Table 10: heading

4.2 Problema 2 — Determinar la media

Para el siguiente conjunto de valores, determina su media (valor promedio):

Índice	Edad	Índice	Edad	Índice	Edad	Índice	Edad
1	**18**	26	**19**	51	**19**	76	**18**
2	**18**	27	**19**	52	**19**	77	**20**
3	**19**	28	**20**	53	**20**	78	**22**
4	**18**	29	**20**	54	**20**	79	**21**
5	**18**	30	**19**	55	**19**	80	**19**
6	**19**	31	**18**	56	**18**	81	**22**
7	**18**	32	**21**	57	**21**	82	**20**
8	**20**	33	**19**	58	**19**	83	**23**
9	**18**	34	**22**	59	**22**	84	**19**
10	**18**	35	**20**	60	**20**	85	**22**
11	**18**	36	**19**	61	**19**	86	**21**
12	**20**	37	**20**	62	**20**	87	**20**
13	**28**	38	**18**	63	**18**	88	**19**
14	**20**	39	**22**	64	**22**	89	**20**
15	**19**	40	**21**	65	**21**	90	**22**
16	**18**	41	**20**	66	**20**	91	**21**
17	**24**	42	**19**	67	**19**	92	**23**
18	**18**	43	**22**	68	**22**	93	**20**
19	**20**	44	**18**	69	**18**	94	**22**
20	**18**	45	**23**	70	**23**	95	**21**
21	**18**	46	**20**	71	**20**	96	**22**
22	**19**	47	**21**	72	**21**	97	**19**
23	**18**	48	**19**	73	**19**	98	**21**
24	**22**	49	**20**	74	**20**	99	**20**
25	**20**	50	**21**	75	**21**	100	**23**

Table 11: Conjunto de las edades de los grupos primer año de licenciatura

Contesta las siguientes preguntas:

- ¿Cuántos valores están por debajo de la media?
- ¿Cómo es el sesgo de los datos, hacia la izquierda o a la derecha?
- ¿Si los datos estuvieran ordenados la media sería distinta?
- ¿La media de los primeros 50 datos más la media de los 50 restantes equivale a la media de los 100 datos totales?

4.3 Problema 3 — Determinar la mediana

Determina el valor de la mediana para los siguientes valores no numéricos:

	Género		**Género**		**Género**		**Género**
1	HOMBRE	21	HOMBRE	41	MUJER	61	HOMBRE
2	MUJER	22	HOMBRE	42	MUJER	62	MUJER
3	MUJER	23	MUJER	43	HOMBRE	63	HOMBRE
4	MUJER	24	HOMBRE	44	MUJER	64	HOMBRE
5	HOMBRE	25	MUJER	45	MUJER	65	HOMBRE
6	HOMBRE	26	MUJER	46	HOMBRE	66	HOMBRE
7	HOMBRE	27	HOMBRE	47	HOMBRE	67	MUJER
8	HOMBRE	28	HOMBRE	48	MUJER	68	HOMBRE
9	MUJER	29	MUJER	49	HOMBRE	69	MUJER
10	MUJER	30	HOMBRE	50	MUJER	70	HOMBRE
11	MUJER	31	HOMBRE	51	HOMBRE	71	HOMBRE
12	MUJER	32	MUJER	52	HOMBRE	72	MUJER
13	HOMBRE	33	MUJER	53	MUJER	73	HOMBRE
14	HOMBRE	34	HOMBRE	54	MUJER	74	HOMBRE
15	MUJER	35	HOMBRE	55	HOMBRE	75	HOMBRE
16	MUJER	36	HOMBRE	56	HOMBRE	76	HOMBRE
17	HOMBRE	37	HOMBRE	57	MUJER	77	HOMBRE
18	HOMBRE	38	HOMBRE	58	HOMBRE	78	HOMBRE
19	MUJER	39	MUJER	59	MUJER	79	HOMBRE
20	MUJER	40	MUJER	60	HOMBRE	80	MUJER

Contesta las siguientes preguntas:

- ¿Se puede calcular la mediana de la lista con valores no numéricos?
- ¿Qué criterio usaste para ordenar los datos, primero las mujeres o primero los hombres?
- ¿Qué valor queda en el centro con el índice 40 cuándo se ordenan los datos, un hombre o una mujer?
- ¿Qué sentido tiene calcular la mediana en valores no numéricos?
- ¿Qué pasaría si los valores fueran impares y la mediana estuviera entre hombre y mujer?
- ¿Si hubieran más categorías sería posible ordenar los datos y calcular la mediana?

4.4 Problema 4 — Determinar la moda

Determina el valor de la moda para los siguientes valores:

	Alerta		**Alerta**		**Alerta**		**Alerta**
1	3	21	3	41	2	61	3
2	2	22	3	42	2	62	2
3	4	23	4	43	2	63	2
4	4	24	2	44	4	64	2
5	2	25	4	45	4	65	2
6	2	26	4	46	2	66	2
7	2	27	2	47	0	67	5
8	0	28	0	48	5	68	0
9	5	29	5	49	0	69	5
10	5	30	0	50	5	70	0
11	5	31	0	51	0	71	0
12	5	32	5	52	0	72	5
13	3	33	5	53	5	73	3
14	3	34	3	54	2	74	3
15	2	35	3	55	3	75	4
16	2	36	4	56	4	76	4
17	1	37	1	57	2	77	1
18	1	38	1	58	0	78	0
19	3	39	3	59	3	79	0
20	2	40	2	60	5	80	2

Contesta las siguientes preguntas:

- ¿Se puede calcular la mediana de la lista con valores no numéricos?
- ¿Qué criterio usaste para ordenar los datos, primero las mujeres o primero los hombres?
- ¿Qué valor queda en el centro con el índice 40 cuándo se ordenan los datos, un hombre o una mujer?
- ¿Qué sentido tiene calcular la mediana en valores no numéricos?
- ¿Qué pasaría si los valores fueran impares y la mediana estuviera entre hombre y mujer?
- ¿Si hubieran más categorías sería posible ordenar los datos y calcular la mediana?

4.5 Problema 5 — Determinar los Cuartiles

Determina el valor de los tres cuartiles Q_1, Q_2, Q_3 para los siguientes valores:

	Alerta		**Alerta**		**Alerta**		**Alerta**
1	3	21	3	41	2	61	3
2	2	22	3	42	2	62	2
3	4	23	4	43	2	63	2
4	4	24	2	44	4	64	2
5	2	25	4	45	4	65	2
6	2	26	4	46	2	66	2
7	2	27	2	47	0	67	5
8	0	28	0	48	5	68	0
9	5	29	5	49	0	69	5
10	5	30	0	50	5	70	0
11	5	31	0	51	0	71	0
12	5	32	5	52	0	72	5
13	3	33	5	53	5	73	3
14	3	34	3	54	2	74	3
15	2	35	3	55	3	75	4
16	2	36	4	56	4	76	4
17	1	37	1	57	2	77	1
18	1	38	1	58	0	78	0
19	3	39	3	59	3	79	0
20	2	40	2	60	5	80	2

Contesta las siguientes preguntas:

- ¿Qué significa el valor de Q_1?
- ¿Qué significa el valor de Q_2?
- ¿Qué significa el valor de Q_3?
- ¿Si tuvieramos que quedarnos con el 50% de los datos, cuáles serían los que deberíamos escojer y bajo qué criterio?
- ¿Puedes dibujar una caja estadística que muestre dónde se ubican los valores de Q_1, Q_2, Q_3?

4.6 Problema 6 — Determinar los Percentiles

Determina el valor de los 100 percentiles $P_1, P_2, ..., P_{99}, P_{100}$ para los siguientes valores:

	Alerta		**Alerta**		**Alerta**		**Alerta**
1	3	21	3	41	2	61	3
2	2	22	3	42	2	62	2
3	4	23	4	43	2	63	2
4	4	24	2	44	4	64	2
5	2	25	4	45	4	65	2
6	2	26	4	46	2	66	2
7	2	27	2	47	0	67	5
8	0	28	0	48	5	68	0
9	5	29	5	49	0	69	5
10	5	30	0	50	5	70	0
11	5	31	0	51	0	71	0
12	5	32	5	52	0	72	5
13	3	33	5	53	5	73	3
14	3	34	3	54	2	74	3
15	2	35	3	55	3	75	4
16	2	36	4	56	4	76	4
17	1	37	1	57	2	77	1
18	1	38	1	58	0	78	0
19	3	39	3	59	3	79	0
20	2	40	2	60	5	80	2

Contesta las siguientes preguntas:

- ¿Fue posible calcular los 100 percentiles?
- ¿Qué pasa cuando tenemos menos de 100 valores?
- ¿Qué pasaría si tuvieramos más de 100 valores?
- ¿Qué significa el percentil al 5% (P_5)?
- ¿Qué significa el percentil al 95% (P_{95})?
- ¿Si tuvieramos que quedarnos con el 90% de los valores más significativos, cuáles serían?