

Nivelatorio: Estadística para la Ciencia de Datos

Presentación 1: Utilizar e interpretar medidas de tendencia central, de posición y de dispersión para describir datos

Diego Antonio Bohórquez Ordóñez
dabohorquez@lces.edu.co

Maestría en Ciencia de Datos

Favor ingresar a:

<https://forms.gle/nat3cDMhka9YaM5r5>

Formación académica:

Economista y Negociador Internacional
Pregrado – Universidad Icesi

Docencia universitaria
Diplomado – Universidad Icesi

Magíster en Economía
Posgrado – Universidad Icesi

Magíster en Ciencia de Datos
Posgrado – Universidad Icesi

Master in Applied Sport Performance Analysis
Posgrado – Loughborough University, UK
(actualmente)

Experiencia laboral:

Asistente de investigación
CIENFI – Universidad Icesi (jul 2016 – feb 2017)

Joven investigador
Alianza CAOBA (feb 2017 – ago 2018)

Investigador asociado
CIENFI – Universidad Icesi (feb 2017 – act.)

Coordinador permanencia y graduación estudiantil
Universidad Icesi (ago 2018- act.)

Profesor HC del departamento TIC
Universidad Icesi (ago 2019 – act.)

Data scientist
Sunderland AFC (ago 2022 – act.)

Agenda

1. Estructura del nivelatorio y material de apoyo
2. ¿Qué es la estadística y por qué es útil en la Ciencia de Datos?
3. Tipos de variables
4. Descripción de los datos para variables cualitativas
5. Descripción de los datos para variables cuantitativas
6. Práctica en Python

Estructura del nivelatorio y material de apoyo

Sesión	Título de la clase	Cap libro guía
1	Utilizar e interpretar medidas de tendencia central, de posición y de dispersión para describir datos	1,2,3,4
2	Utilizar e interpretar medidas de tendencia central, de posición y de dispersión para describir datos	1,2,3,4
3	Examinar las principales distribuciones de probabilidad que se emplean para describir datos	5,6,7
4	Describir una población a partir de una muestra, mediante el uso de estimadores puntuales, intervalos de confianza y pruebas de hipótesis	8,9,10
5	Realizar comparaciones entre dos o más poblaciones a partir de una muestra, empleando pruebas de hipótesis paramétricas y no paramétricas	11,12,17,18

- Para este nivelatorio, vamos a seguir el texto:

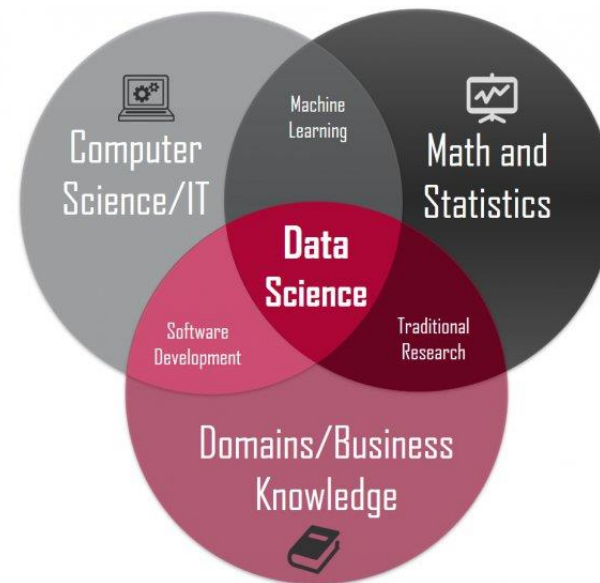
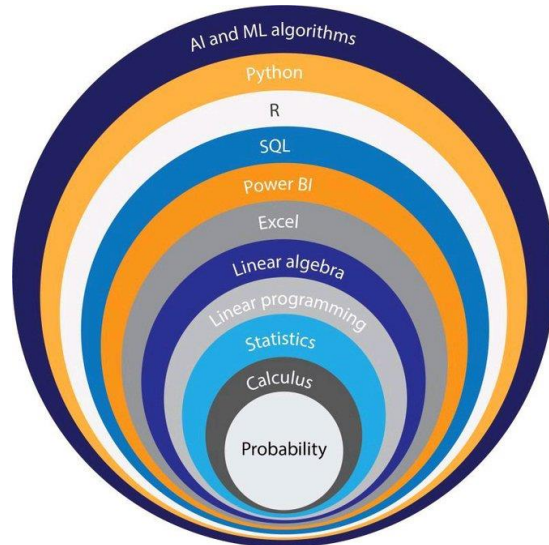
Lind, D. A., Marchal, W. G., & Wathen, S. A. (2015). Estadística aplicada a los negocios y la economía. McGraw-Hill, Ed 16.

¿Cómo se aprueba este nivelatorio?:

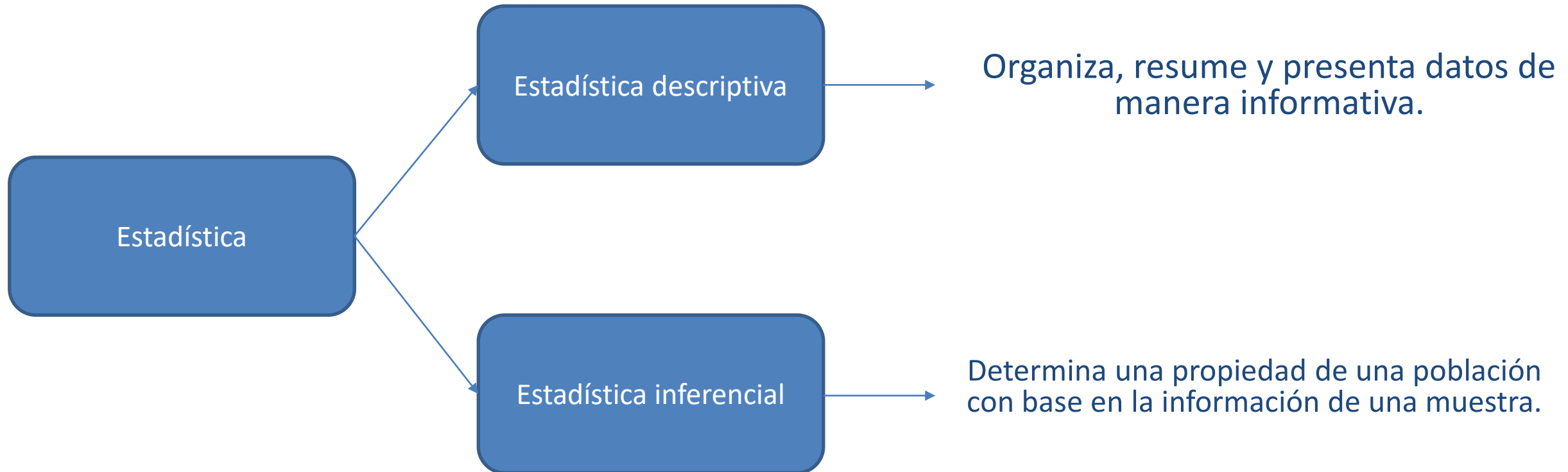
- Entregar completo cada una de las actividades propuestas (40%)
- Entregar completo el trabajo final del curso (60%)

¿Qué es la estadística y por qué es útil en la Ciencia de datos?

- ¿Qué es?: “Ciencia que recoge, organiza, presenta, analiza e interpreta datos con el fin de propiciar la toma de decisiones más eficaz” (Lind, Marchal & Wathen, 2015)
- ¿Por qué es útil en la Ciencia de datos?:



Existen dos tipos de estadística:

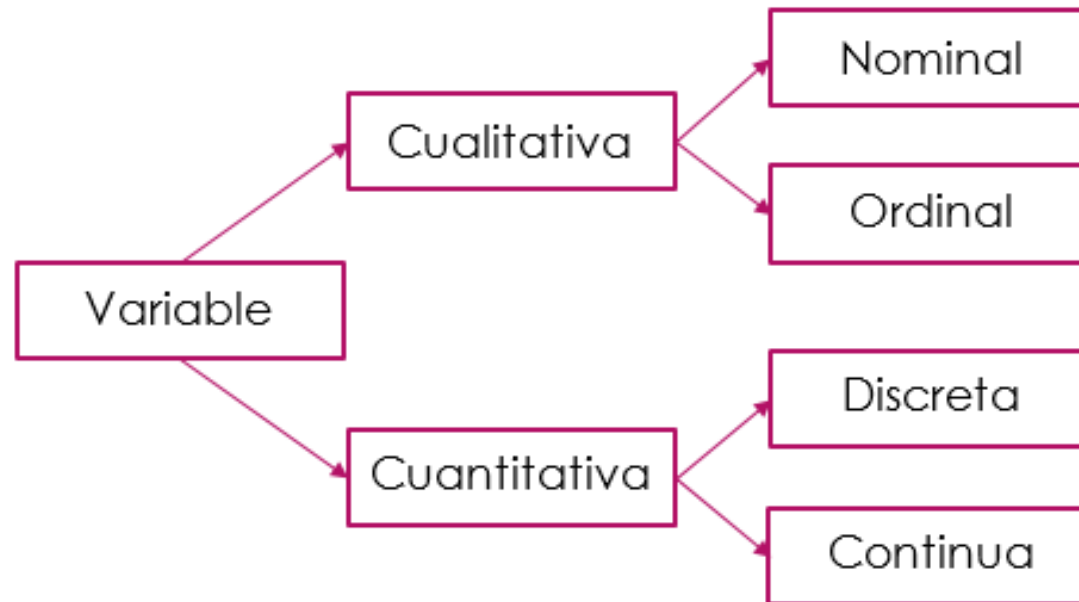


Tipos de variables

- ¿Qué es una variable?

Una variable es una característica o cualidad

- ¿Cómo se clasifican?



Variables cualitativas:

Es una característica o cualidad de naturaleza no numérica

1. Nominal: Característica que se define en categorías mutuamente excluyentes, sin ordenación.

Ejemplo: Género, departamento en el que trabaja, ciudad de procedencia.

2. Ordinal: Característica que se define en categorías mutuamente excluyentes, donde sí existe ordenación.

Ejemplo: Escala de satisfacción del cliente, perfiles de riesgo, estrato, niveles de desempeño.

Variables cuantitativas:

Es una característica o cualidad de naturaleza numérica

1. Discreta: Característica que asume un número contable de valores entre dos valores (conteos).

Ejemplo: Número de clientes, número de goles, cantidad de visitas la página web.

2. Continua: Característica que asume un número infinito de valores entre dos valores.

Ejemplo: Utilidades del mes, probabilidad de que un estudiante se gradúe, tiempo de espera en la parada del bus.

Ejercicio: ¿Cuál es la variable y qué tipo de variable es?

1. Una base de datos contiene la ciudad de origen de cada uno de los estudiantes.
2. Una base de datos contiene la distancia que existe entre cada casa de un estudiante y la Universidad Icesi.
3. En la encuesta de fin de carrera se encontró que la mayoría de los graduandos está muy satisfechos con la experiencia educativa de Icesi.
4. En una base de datos se tiene el mes de nacimiento de cada uno de los estudiantes, para felicitarlos al inicio de cada mes vía correo electrónico.
5. Durante un mes se recolectó la cantidad de vehículos que ingresan a la Universidad Icesi en un día.
6. Gestión Humana desea premiar a los colaboradores más leales, por lo que consultó la base que contiene el número de años en servicio de cada uno de los colaboradores.

Sobre la transformación de tipos de variables:

- Podemos transformar variables cuantitativas a cualitativas.

Ejemplo: Para premiar la lealtad de los colaboradores, Gestión Humana construyó una base de datos que contiene el número de años que un colaborador lleva de servicio.

- Es una variable cuantitativa continua, pero la podemos transformar en variable cualitativa ordinal.
- Generando categorías: menos de 5 años, entre 5 y 9,99 años, 10 años o más.

Descripción de los datos con variables cualitativas

De manera numérica:

Se utilizan tablas de frecuencias, presentando la frecuencia relativa de cada categoría.

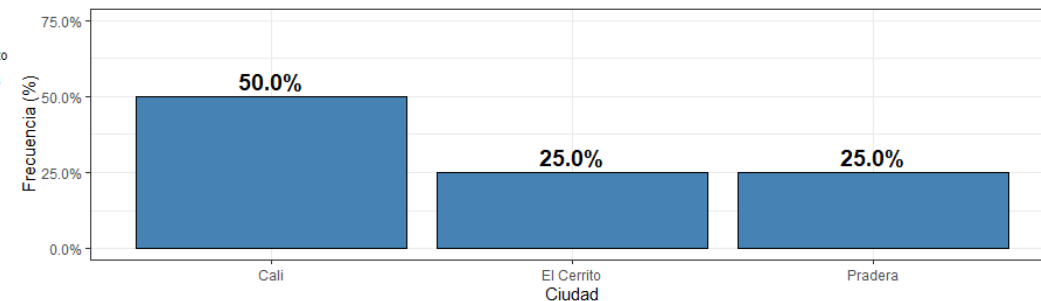
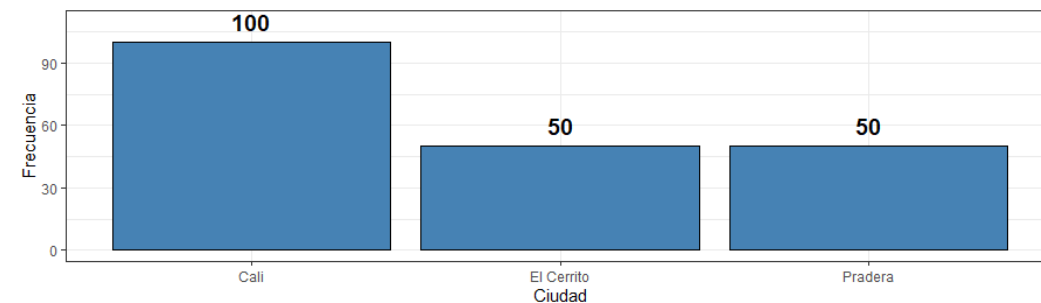
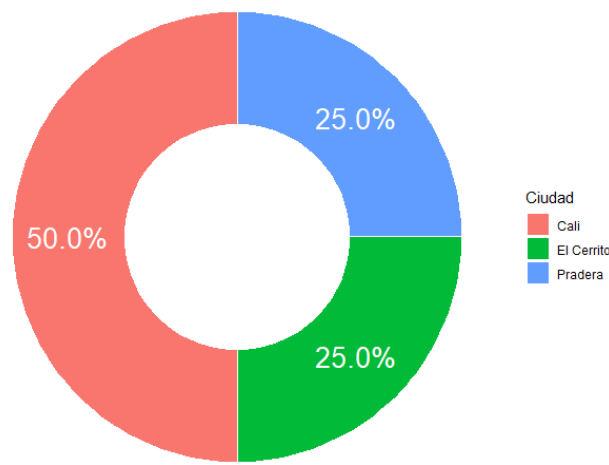
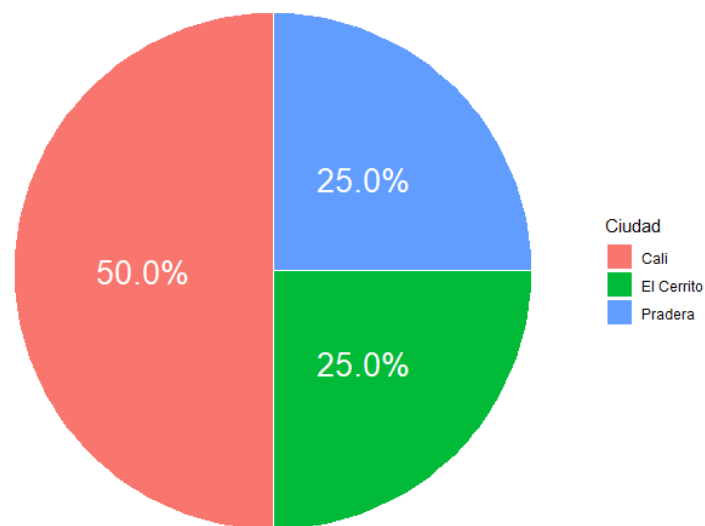
Ejemplo: En una encuesta de 200 propietarios de vehículos, 100 eran de Cali, 50 eran de Pradera y 50 eran de El Cerrito.

Ciudad	Frecuencia	Frecuencia relativa (%)
Cali	100	50,00%
Pradera	50	25,00%
El Cerrito	50	25,00%
Total	200	100,00%

De manera gráfica:

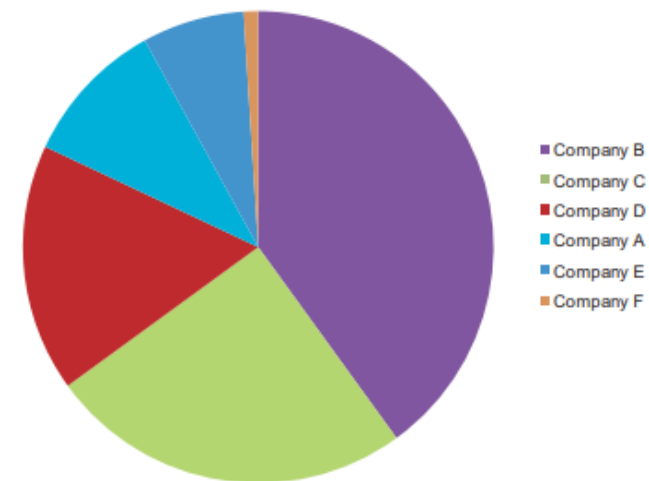
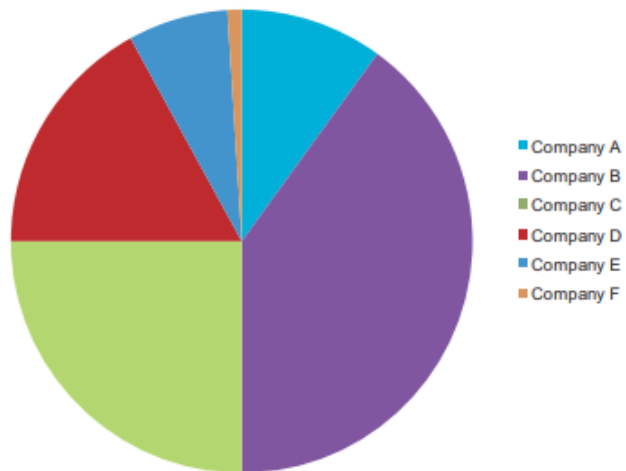
Se utiliza el gráfico circular (torta) o de barras simples.

Ejemplo: En una encuesta de 200 propietarios de vehículos, 100 eran de Cali, 50 eran de Pradera y 50 eran de El Cerrito.



Ejemplo torta:

¿En cuál figura es más grande el área verde? ¿y la morada?

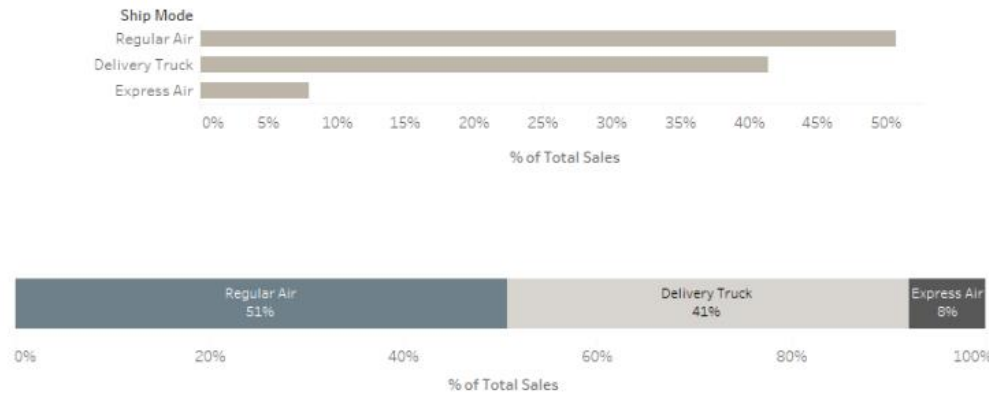


Revisar: http://www.perceptualedge.com/articles/visual_business_intelligence/save_the_pies_for_dessert.pdf?_ga=2.131808975.173300834.1611611179-1532428875.1611611179

What other chart types we can use to show the proportion.

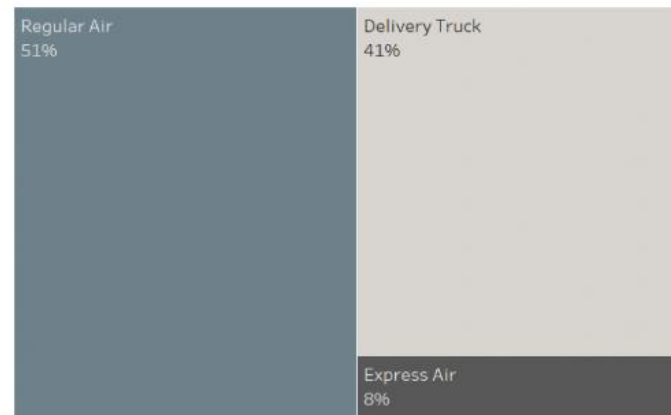
SIMPLE BAR CHART OR STACKED BAR CHART

Definitely, the best alternative for a pie chart/ donut chart is a simple bar graph because in that case we only have to compare one dimension, length with more clarity and less clutter.



Treemap

Ben Shneiderman, the founder of the treemap, which shows the hierarchical data in areas of rectangles.



Descripción de los datos con variables cuantitativas

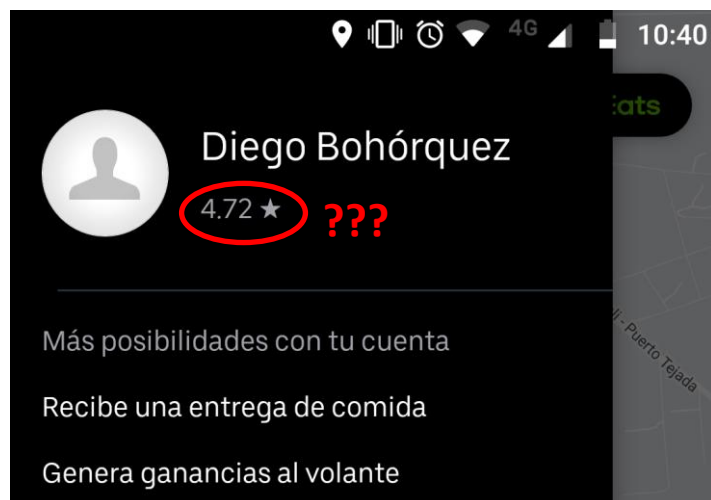
De manera numérica:

Se utilizan medidas de tendencia central, de posición y de dispersión, para describir la distribución de la variable.

- Medidas de tendencia central: El objetivo es encontrar el “centro” de los datos o alrededor de qué punto se agrupan los datos.
 1. Media: Valor promedio de los datos.
 2. Mediana: Es el valor que está por encima del 50% de los datos, al ordenarlos de menor a mayor.
 3. Moda: Valor que aparece con mayor frecuencia.

Medida	Cualitativa nominal	Cualitativa ordinal	Cuantitativa discreta	Cuantitativa continua
Media	NO	NO	SÍ (para aquellas que presentan conteos)	SÍ
Mediana	NO	SÍ (no muy usado)	SÍ	SÍ
Moda	SÍ	SÍ	SÍ (no muy usado)	SÍ (no muy usado)

Ejemplos clásicos del mal uso de las medidas de tendencia central:



“La satisfacción del cliente es en promedio 8.5, por tanto, estamos bien”

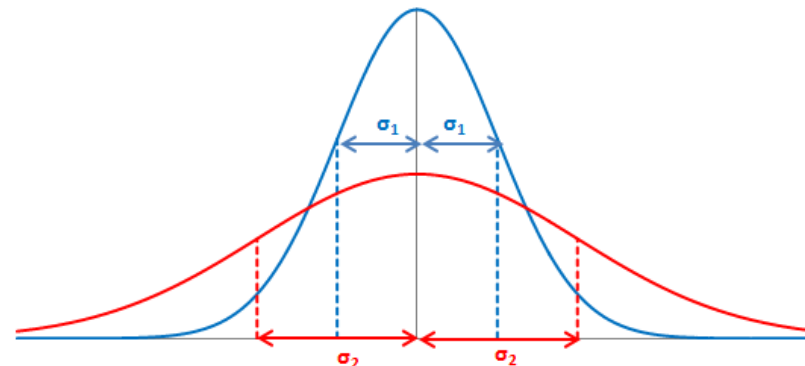
Problemas de la media:

- Se ve afectada por valores extremos (deja de ser útil)

Datos	Media	Mediana
1,2,2,3,4	2,4	2
1,2,2,3,8	3,2	2
1,2,2,3,16	4,8	2
1,2,2,3,32	8	2

- Saber el “centro” no nos dice todo sobre la distribución de una variable (mismo problema con la mediana)

Dos distribuciones normales con diferentes desviaciones típicas σ



- **Medidas de dispersión:** El objetivo es saber qué tan agrupados están los datos alrededor del centro. En otras palabras, qué tan lejos o qué tan cerca, en promedio, están del centro.
 1. **Desviación estándar:** Promedio de la desviación con respecto a la media. La varianza es la desviación estándar al cuadrado (mide el promedio de las desviaciones respecto a la media al cuadrado).
 2. **Coeficiente de variación:** Mide qué tan grande es la dispersión, como proporción de la media.

Medida	Cualitativa nominal	Cualitativa ordinal	Cuantitativa discreta	Cuantitativa continua
Desviación estándar	NO	NO	SÍ (para aquellas que presentan conteos)	SÍ
Coeficiente de variación	NO	NO	SÍ (para aquellas que presentan conteos)	SÍ

Comentarios sobre el coeficiente de variación:

- Su utilidad radica en proveer la posibilidad de comparar la dispersión de dos variables.
- Sin embargo, también existen algunas reglas empíricas sobre cuál debería ser la relación adecuada entre desviación estándar y media.
- Para algunos investigadores, un coeficiente de variación superior al 20% debe generar preocupación por la “alta” dispersión de los datos. (Para el DANE, en sus mediciones, más del 15% es preocupante, siendo “aceptable” entre 5% y 10%)

Ejemplo: Se recolectó la frecuencia de visitas a una de las sucursales más concurridas de un banco, por 18 días hábiles seguidos.

Datos		
30	23	34
28	36	30
29	30	27
27	35	31
25	32	29
26	32	33

Media: 29,83

Mediana: 30

Desviación estándar: 3,5

Coeficiente de variación: 11,7%

Problema de estas medidas:

- Al utilizar la media como referencia de centro, entonces valores extremos generan problemas en el cálculo. (si la media no “está bien”, estas medidas tampoco)

- Medidas de posición (cuantiles): El objetivo es describir la posición que tiene un valor específico en relación con el resto de datos. Se dividen los datos en partes iguales, luego de haberlos ordenado de menor a mayor.
 1. Cuartiles: Se dividen los datos en 4 grupos iguales (25% de los datos en cada grupo)
 2. Deciles: Se dividen los datos en 10 grupos iguales (10% de los datos en cada grupo)
 3. Percentiles: Se dividen los datos en 100 grupos iguales (1% de los datos en cada grupo)

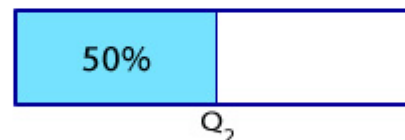
Medida	Cualitativa nominal	Cualitativa ordinal	Cuantitativa discreta	Cuantitativa continua
Cuartiles	NO	NO	SÍ (para aquellas que presentan conteos)	SÍ
Deciles	NO	NO	SÍ (para aquellas que presentan conteos)	SÍ
Percentiles	NO	NO	SÍ (para aquellas que presentan conteos)	SÍ

Los más usados son los cuartiles:

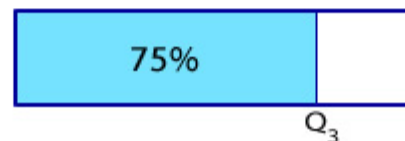
El **primer cuartil** (Q_1) es el valor de la variable que supera a lo más el 25 % de los datos y es superado por a lo más el 75 % de ellos en la distribución ordenada de menor a mayor.






El **segundo cuartil** (Q_2) es un valor que supera a lo más el 50 % de los datos y es superado por a lo más el 50 % de ellos, es decir, Q_2 coincide con la mediana.



El **tercer cuartil** (Q_3) es un valor que supera a lo más al 75 % de los datos y es superado por a lo más el 25 % de ellos.



Miremos un ejemplo con percentiles (Saber 11 y Saber Pro):




REPORTE DE RESULTADOS
ESTUDIANTE
•SABER 11.

PUNTAJE GLOBAL

De 500 puntos posibles, su puntaje global es **283**

¿EN QUÉ PERCENTIL ME ENCUENTRO?

▲ Con respecto a los estudiantes del país, usted está aquí.


Reporte de resultados de estudiantes Saber Pro

Información Básica

Fecha de aplicación: 29 de octubre de 2017
Fecha de publicación de resultados: 24 de febrero de 2018
Número de registro:
Identificación:
Nombres y apellidos:
Institución: UNIVERSIDAD FRANCISCO DE PAULA SANTANDER-OCAÑA
Código SNIES:
Programa: INGENIERIA MECANICA

Puntaje Global

De 300 puntos posibles, su puntaje global es **185**

Grupo Referencia

INGENIERÍA

¿En qué percentil me encuentro?

▲ Respecto a los estudiantes a nivel nacional, usted está aquí.
▼ Respecto a los estudiantes de su grupo de referencia, usted está aquí.


Módulos Competencias Genéricas		
Módulos	De 300 puntos posibles, su puntaje es	¿En qué percentil me encuentro?
Comunicación escrita	216	
Razonamiento cuantitativo	201	
Lectura crítica	188	
Competencias ciudadanas	143	
Inglés	178	

Volviendo al Ejemplo: Se recolectó la frecuencia de visitas a una de las sucursales más concurridas, por 18 días hábiles seguidos.

Datos		
30	23	34
28	36	30
29	30	27
27	35	31
25	32	29
26	32	33

Media: 29,83

Mediana: 30

Desviación estándar: 3,5

Coeficiente de variación: 11,7%

Mínimo: 23

1er Cuartil: 27,25

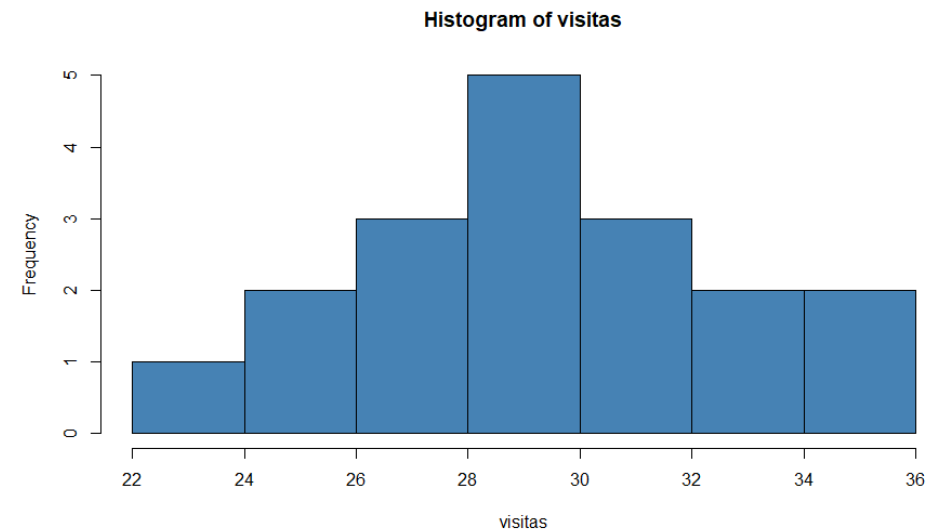
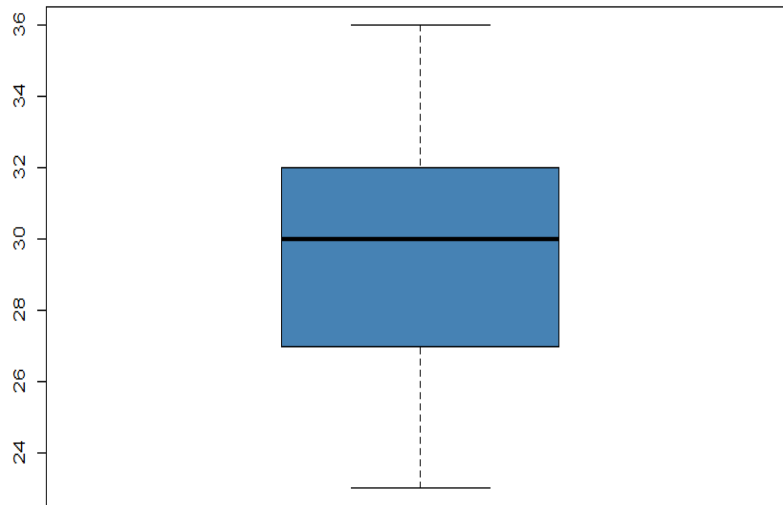
3er Cuartil: 32

Máximo: 36

De manera gráfica:

- Se utiliza el histograma o el diagrama de cajas.

Ejemplo: Se recolectó la frecuencia de visitas a una de las sucursales más concurridas de un banco, por 18 días hábiles seguidos.

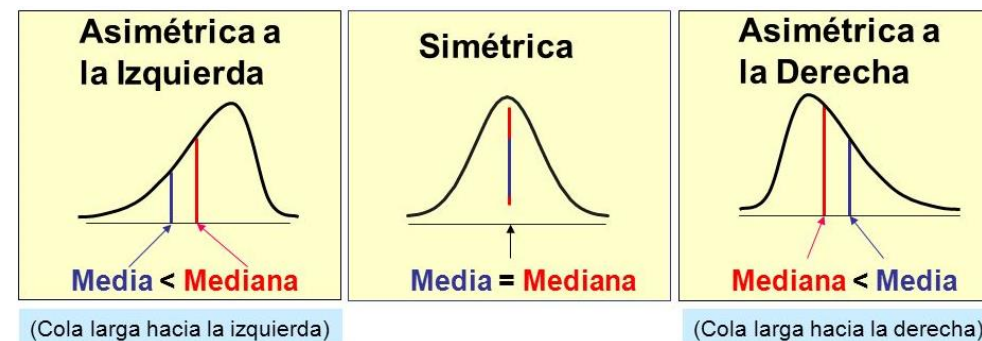


Finalmente, hablemos un poco sobre la forma de la distribución:

1. Sobre la simetría (de manera intuitiva):

- Datos simétricos: Media es igual a la mediana y los valores se dispersan uniformemente alrededor de éstos.
- Sesgo a derecha o sesgo positivo: Media es mayor que la mediana y los valores se dispersan más hacia la derecha que hacia la izquierda.
- Sesgo a izquierda o sesgo negativo: Media es menor que la mediana y los valores se dispersan más hacia la izquierda que hacia la derecha.

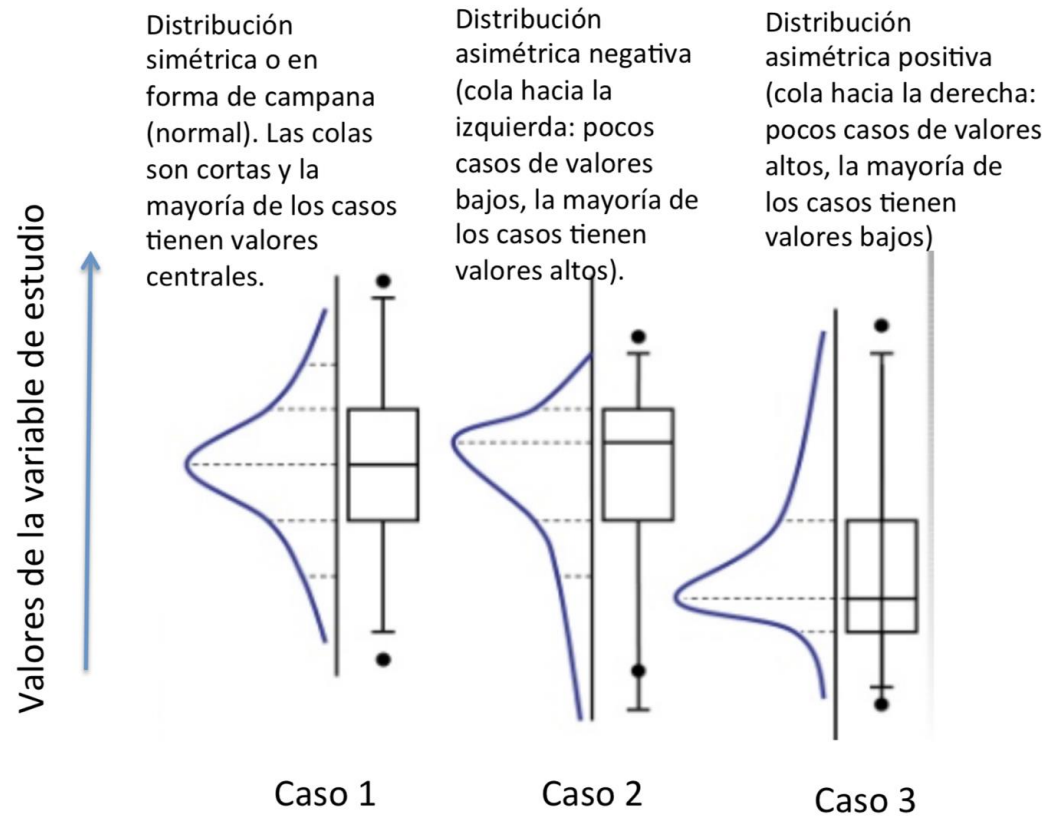
De manera gráfica:



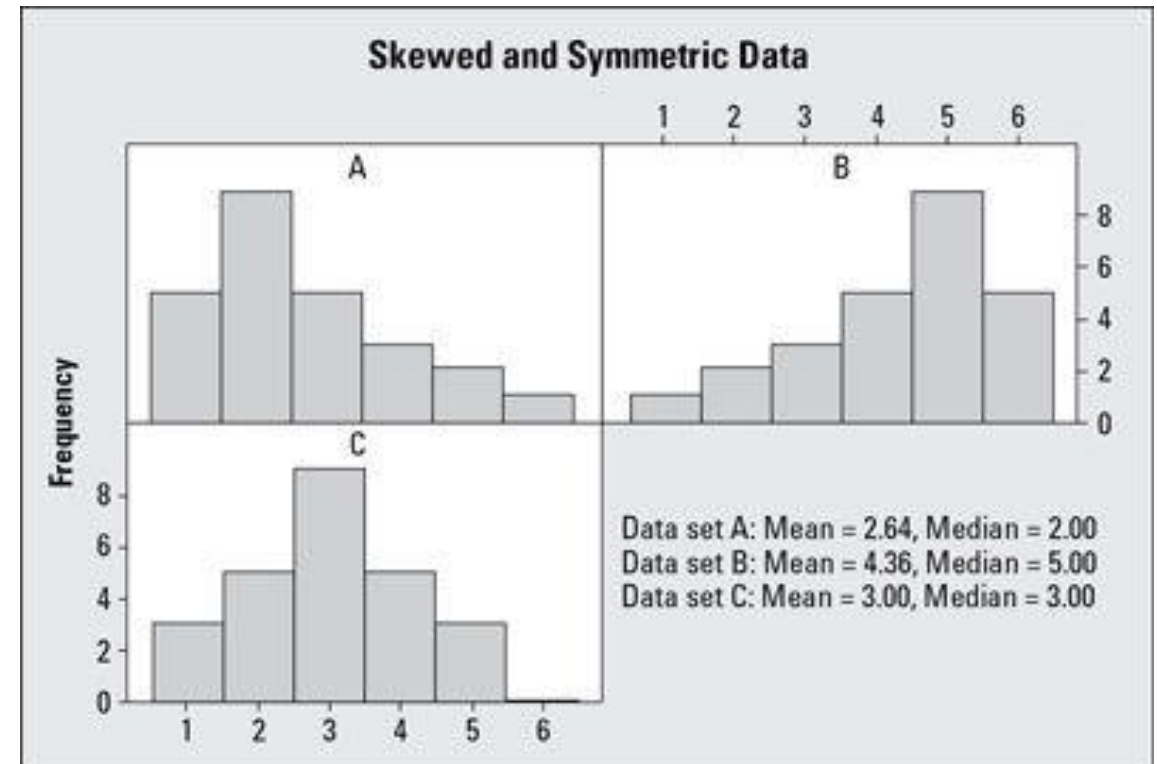
De manera matemática: Las fórmulas parten de la diferencia entre cada observación y la media, siendo importante el signo de la diferencia.

- Coef. Asim. igual o "muy cercano" a 0 (simetría)
- Coef. Asim. mayor a 0 (sesgo positivo)
- Coef. Asim. menor a 0 (sesgo negativo)

Sesgo en diagrama de cajas:



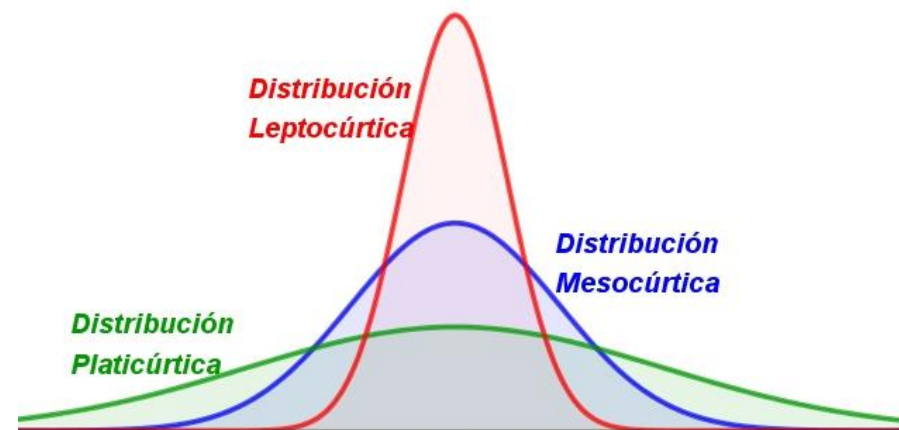
Sesgo en histograma:



Finalmente, hablemos un poco sobre la forma de la distribución:

2. Sobre la curtosis: Se refiere a la forma de “punta” del centro de la distribución.
- Mesocúrtica: Tiene forma igual a la de una distribución normal.
- Leptocúrtica: Forma “puntiaguda” en el centro. (Muy concentrada en el centro).
- Platicúrtica: Forma “achatada” en el centro. (Poco concentrada en el centro).

De manera gráfica:



De manera matemática: Las fórmulas parten de la diferencia entre cada observación y la media.

- Coef. curtosis igual o “muy cercano” a 0 (mesocúrtica)
- Coef. curtosis mayor a 0 (leptocúrtica)
- Coef. curtosis menor a 0 (platicúrtica)

En resumen...

Tutorial en Python con base gapminder

Ejercicio práctico en Python

Suponga que nos contrataron para trabajar en un Instituto donde enseñan inglés. Se nos ha entregado una base de datos que contiene información del último periodo, en la que se cuenta con 1779 estudiantes distribuidos en los 6 niveles que se ofertan. La base de datos cuenta con las siguientes columnas en la pestaña "Notas":

1. Nivel: Identificador del nivel de inglés al que pertenece el estudiante.
2. Código: Identificador del código del estudiante en las Institución.
3. 1P: Nota del primer parcial.
4. 2P: Nota del segundo parcial.
5. 3P: Nota del tercer parcial.
6. NF: Nota definitiva.
7. Género: Identificador del Género del estudiante.

Por su parte, en la pestaña "INFO_S11":

1. Código: Identificador del código del estudiante en las Institución.
2. PG: Puntaje global en las pruebas Saber 11.
3. PING: Puntaje de inglés en las pruebas Saber 11.
4. NIVING: Nivel de desempeño en inglés en las pruebas Saber 11.

Primera parte: Antes que nada, debemos preparar los datos.

1. Cargar las dos bases de datos. (read_excel)
2. Unir las dos bases de datos. (merge)
3. Definir qué tipo de variable es cada una de las columnas (intuitivamente)
4. Crear una nueva columna que identifique si el estudiante aprobó o no el curso. (np.where)
5. Crear una nueva columna que identifique cuantos parciales aprobó cada estudiante (np.where).

Segunda parte: Nuestro jefe le ha solicitado:

1. ¿Cuántos estudiantes hay en cada nivel?
2. ¿Cuál es la distribución por género en el Instituto?
3. Explore la distribución del puntaje de inglés en Saber 11
4. Explore la distribución de la nota definitiva
5. Explore la tasa de aprobación por nivel. (pd.crosstab)
6. Filtre por el nivel con la tasa de aprobación más baja y muestre con gráficas la respuesta a esta pregunta: ¿Existe diferencia en la distribución de las notas de los tres parciales?

Actividad en casa

Actividad en casa: Seleccione 2 variables (una cualitativa y una cuantitativa) de una base de datos que use en su organización y descríbalas:

1. ¿Cuál es el contexto de la base de datos? (cómo y por qué se recoge, para qué se usa, qué variables están disponibles)
2. Describa cada variable de manera numérica y gráfica, según su tipo. Es decir, debe presentar para cada variable una gráfica y un párrafo que la describe.

Entregar un Jupyter Notebook (Python), a través de Intu.