DD MM AA

Estadística Descriptiva

Permite
- Conocer reglas y métodos usados en el tratamiento de datos
- Evaluar y cuantificar la importancia de los resultados estadísticos
- Entender fenomenos de la realidad (sociales, biologicos....)
- Dar vision clara de la información proveniente de distintas fuentes

Estadística
↳ Manejo de Información Cuantificable
- Descripción de datos
- Inferencia = concluir

↳ Estudia fenomenos aleatorios y su comportamiento (variabilidad)

↳ Funciones
- Resumir = Gráficamente o Numericamente
- Simplificar=
- Comparar
- Relacionar
- Proyectar

↳ Tareas
1. Delimitar estudio
   - Establecer Unidades
   - Definir Variables
2. Observar
   - Censo
   - Muestreo
   - Diseño experimental
3. Recolección y registro de datos
4. Depurar información
5. Resumir = Graf. Numeric.
6. Interpretar

Topicos
- Diseño experimental
- Estadística Descriptiva (Tareas)
- Inferencia
- Estadística no paramétrica (modelos sin supuestos previos)
- Elementos de regresión (Variables respuesta y explicativas)

IPOS DE VARIABLES
↳ Características que varían
- Estatura
- Presión Sanguínea

↳
1. Continuas = Mediciones, Estatura de un grupo
2. Discretas = Conteos, # autos en un semaforo, horas
3. Categoricas = clasificación, Sexo, estrato

NIVELES DE MEDICION

1. Nominal = representan categorias, Estado civil = C, S, M, no cumple 1≤2≤3
   No medidas basicos de resumen
2. Ordinal = representan Jerarquias, Cuadrados = 1, 2, 3 se cumple 1≤2≤3
   No medidas basicos de resumen
3. Intervalo = cuantitativas
4. Razon = Cuantitativa con escalas, estatura
   Presion sanguinea

DESCRIPCION DE DATOS

Distribuciones de frecuencia

- Tabla de frecuencia
  - Encuentre Minimo y Maximo de los valores registrados
  - Defina Intervalos de clase
    - Sturges ✗
      $K = 1 + 3{,}33 \log_{10}(n)$
    - Velleman n Pequeño
      $K = 2\sqrt{n}$
    - Dixon y Kronmal n Grande
      $K = 10 \cdot \log_{10}(n)$

K = numero de clases
n = numero de datos

Notas:
* Entre 5 y 25 clases es lo adecuado
* Pocas clases ⇒ Gran perdida de informacion
* Pocos datos, Muchos clases ⇒ No se evidencian comportamientos
* Dentro de los clases es importante Organizar (menor a mayor)

- Cuente las observaciones por cada subintervalo (Frecuencia de clase)
- Calcule la frecuencia relativa

$$FR = \frac{\text{Frecuencia de clase}}{\#\text{ total de observaciones}}$$

Ejemplo: tabla dada:

Masa	Estatura	Edad	Genero	Estrato	Gasto	Longitud Abdomen	Tipo colegio	Horas a estudio
Kg	CM	Dias	H, M	1 a 5	$	CM	Pu Opri	hora

Tiempo en llegar a la U	Longitud pies Prom	Longitud manos Prom
Min	CM	CM

- Tabla de frecuencia para genero

	Frecuencia	Porcentaje
Hombre	46	68,66
Mujer	21	31,34

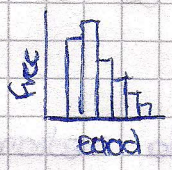
67 DATOS

- En la Edad el grafico de barras no es diciente (Se hizo intervalos de clase) (8 clases)

Categoria	Frecuencia	Frecuencia Relativa
≥ 6300	20	0,299
(6300, 6900)	24	0,358
(6900, 7500)	11	0,164
(7500, 8100)	4	0,060
(8100, 8700)	3	0,045
(8700, 9300)	2	0,030
(9300, 10000)	2	0,030
> 10000	1	0,015

Se aplico la formula

* Nuevo Grafico frecuencia Vs edad



- Para la estatura definieron clases

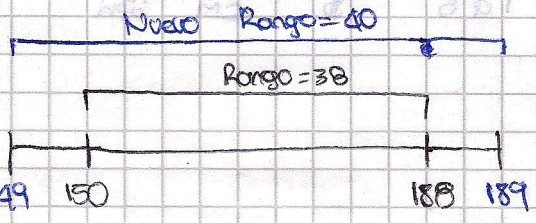
Storges → $K = 1 + 3{,}33 \log_{10}(67) = 7{,}08 \approx \boxed{8}$

* El valor maximo es 188 cm y el minimo el rango es igual a 38

La Amplitud de cada intervalo estara dada por

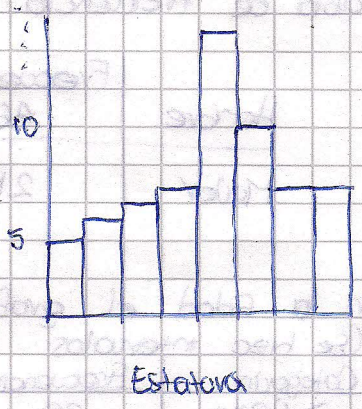
$$A = \frac{\text{Rango}}{K} = \frac{38}{8} = 4{,}75 \approx \boxed{5}$$

Como hay un excedente de dos se reparte equitativamente



La tabla queda dada

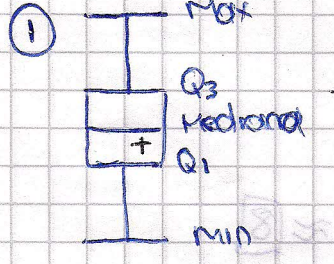
Intervalo	Frecuencia	FR	Marca
149, 154	5	0,075	151,5
154, 159	6	0,090	156,5
159, 164	7	0,104	161,5
164, 169	8	0,119	166,5
169, 174	17	0,254	171,5
174, 179	10	0,149	176,5
176, 184	7	0,104	181,5
184, 189	7	0,104	186,5



NOTA: Las tablas de frecuencia deben contar con
- Uniformidad: Clases de igual amplitud o variable que dependan del tipo de datos
- Unicidad: Clases no traslapados
- Completez: Cada dato pertenece a una y sólo una clase

Box-Plot (Cajas y Bigotes)

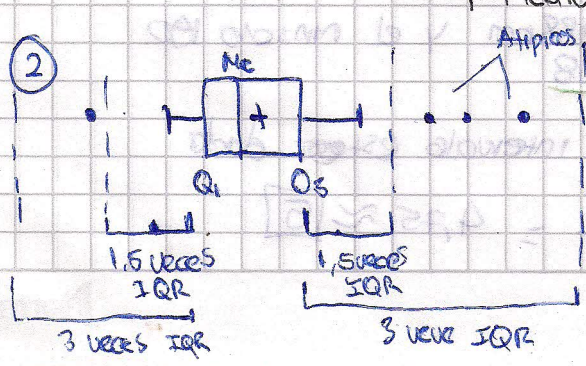
Es señala las características importantes de un conjunto de datos



Q = Cuartil

50% de los datos mas centrales

Nota = la mediana y la moda, no siempre estan en el mismo lugar



Medidas de localizacion y de tendencia central
Medidas de dispersion (variabilidad

DD MM AA

ANALISIS DESCRIPTIVO DE UN CONJUNTO DE n DATOS

- Datos agrupados = Crear clases y hacer frecuencia
- Datos sin agrupar = Manipulacion de la informacion como se tomo

Datos Agrupados

Media Muestral

$$\bar{X} = \frac{\sum_{i=1}^{\#celdas} \text{Marca de clase} \cdot \text{Frecuencia de clase}}{\text{Total de frecuencia}}$$

Ejemplo

$\bar{X} = 170,38$

Moda

↳ Marca de clase con la mayor frecuencia

Ejemplo

Moda = 171,5

Percentiles (X_p)

↳ Valores abajo y arriba de los cuales se encuentra una cierta proporcion de datos del conjunto

$P_{50} \Rightarrow p = 0,5$
$P_{25} \Rightarrow p = 0,25$

$$X_p = L + \frac{(P-a) \times h}{f}$$

Donde:

F = FR de la clase que contiene el percentil

L = Limite inferior de la clase que contiene el percentil

P = Percentil a calcular ($P_{50} \Rightarrow p = 0,5$)

Datos Sin agrupar

Media Muestral

Sensible a valores extremos

$\bar{x}$ = Promedio

$$= \frac{\Sigma \text{ de datos}}{\#\text{ de datos}}$$

Moda

↳ Dato con mayor frecuencia (Dato que mas se repite)

Ordenar datos

Percentiles

↳ Homologo de los cuartiles (%)

a= FR Acumulada del intervalo anterior al del percentil

h= Longitud de clase del Percentil

Ejemplo P50

$$P50 = 169 + \frac{(0,5 - 0,38) \times 5}{0,254}$$

$$P50 = 171,2$$

Mediana

↳ Es igual al P50

Varianza

$$S_a^2 = \sum_{i=1}^{k} \frac{(m_i - \bar{x}_a)^2 \times f_i}{n-1}$$

Rango intercuartil

↳ diferencia entre el percentil 75 y el 25 (Datos mas centrales)

$$Q_{RANGE} = Q_3 - Q_1 = P_{75} - P_{25}$$

Ejemplo

$S_a^2 = 101,4$
$S_a = 10,07$
$P_{25} = 163,1$
$P_{75} = 177,6$
$Q_{RANGE} = 14,5$

Mediana

↳ Dato central

Impar = 1

Datos ordenados — Par = el prom de los dos centrales

$$\tilde{x} = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2} \quad \text{Par}$$

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} \quad \text{impar}$$

- 5 [7] 9
- 5 8 3 7
  3 5 7 8

↳ $\frac{5+7}{2} = 6$

Varianza

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

A mayor varianza mayor dispersion

- **Desviacion Estandar**

↳ Raiz de la varianza

Rango intercuartil

$$Q_{RANGE} = Q_3 - Q_1 = P_{75} - P_{25}$$

Coeficiente de Variacion

$$C.V. \frac{\sigma}{\bar{x}} = \frac{\text{Desviacion estandar}}{\text{media}}$$

a mayor valor mayor dispersion

DD MM AA

1. Importar Datos

%>% = Clic

Datos %>% Importar datos %>% Desde archivo Excel
Desde archivo de texto
...

2. Resumen y tipo de datos

Summary (df)
Str (df)

3. Modificar Datos

3.1 Conversion a factores

Datos %>% Modificar Variable del %>% Recodificar variables
conjunto de datos
activos

% >% Escoger Variables

1 = = "verdad"
2 = = "falso"

Para Proporciones

3.2 Nuevas Variables

Datos %>% Modificar variable del %>% Calcular nueva variable
conjunto de datos
activo

- factor (Ifelse (Condicion, "verdadero", "falso")) Prop
- Formula : Variable 1 * variable 2 =

3.3 Eliminar factores no necesarios

Datos %>% Modificar variable del %>% Descartar niveles sin
conjunto de datos uso
activo

☑ Todos los factores

4. Agrupar Datos (filtrar)

Datos %>% Conjunto de datos activos %>% Filtrar el conjunto
de datos activos

☑ Incluir todos las variables

expresion de seleccion

- Para un factor : df == "factor 1"
- Para 2 o mas factores: df == "factor 1" | df == "factor 2"

DD MM AA

5. Estadísticos de Interes (Pruebas No Normales o σ conocidas)

Estadísticos %>% Resumenes %>% Resumenes numericos %>%

Resumir por grupo %>% Seleccionar variables %>% Estadísticos de interes

- ☑ Media
- ☑ Desviación típica

6. Juego de hipotesis Para medias

H_0 = Complemento
H_a = Lo que queremos : cola derecha / cola izquierda / Bilateral ?

6.1 Test de Normalidad

- Prueba de hipotesis

Estadísticos %>% Resumenes %>% Test de Normalidad

{P-Value > α Son normales} %>% Test por grupos

- Q-Q plot

Graficos %>% Graficas de comparacion de Cuantiles

- Histograma

Graficos %>% Histograma

6.2 Homocedasticidad (Razon de varianzas)

Cuidado: "factor en orden alfabetico"

{Teniendo los dos factores agrupados}

Estadísticos %>% Varianzas %>% Test F para dos varianzas %>% Bilateral

Convertir var # en factor

{P-Value > α Son iguales $\sigma_1^2 = \sigma_2^2$}

6.3 Ph (T-Student) Para 1 Media

Estadísticos %>% Medias %>% Test t para una muestra

- Ingresar Ha
  MU= X

- Revisar α
- Analizar P-Value
- Concluir

6.4 Ph (T-Student) Para dos Medias

Estadisticos %>% Medias %>% Test t para muestras independientes %>% Seleccionar Variable

Revisar factores
Voltear Ha en
caso de estar
truncados

- Seleccionar Ha
- Revisar α
- Varianzas iguales
  - Si
  - No

- Verificacion Ph (Graficamente)

Graficos %>% Graficos de las Medias

⊙ Intervalos de confianza

6.5 Ph para una proporcion

Estadisticos %>% Proporciones %>% Test de proporciones para una muestra

Ho = P0

(Si los factores estan truncados
Ho = 1 - P0

6.6. Ph (Multinomial) (Chi-Cuadrado)

Ho = Todas las probabilidades son iguales a las propuestas

Ha = Al menos una de las probabilidades es diferente a la propuesta

Estadisticos > Resumenes > Distribucion de frecuencias

- Ingresar valores

☑ Test Chi-Cuadrado de bondad de ajuste

DD MM AA

- No hay evidencia en los datos con una significancia