Conceptos básicos

miércoles, 23 de abril de 2025 11:13

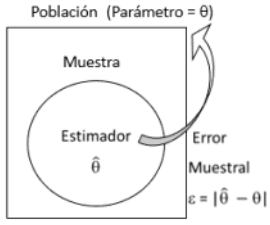
Estadística: Ciencia matemática que se encarga del análisis de datos y las inferencias que podemos realizar a partir de ellos.

- **Estadística descriptiva:** Técnicas para describir y representar información
- **Probabilidades:** Modelos matemáticos para representar el comportamiento de variables aleatorias
- **Estadística inferencial:** Técnicas para realizar inferencias a partir de estimadores muestrales

PROCESO DE INFERENCIA ESTADÍSTICA

- **Población:** Estudiantes usuarios del Banner en el período 2025 de la ESPE Matriz en modalidad presencial
- **Variable:** Accesos exitosos al sistema/semana, Total de requerimientos de acceso, Tiempos de espera para el acceso, Rechazos de acceso
- **Dato:**
- Parámetro: Son desconocidos!! (se pueden calcular únicamente a través de un censo!!)
- Muestra: Representatividad = Tamaño, Selección
- Estimador
- Error muestral

EL PROCESO DE INFERENCIA ESTADÍSTICA



Población: Conjunto total de observaciones, debidamente delimitado.

Variables: Característica de interés de un elemento de la población.

Dato: Información o respuesta obtenida sobre la variable

Parámetro: Medida de resumen poblacional de la variable

Muestra: Subconjunto representativo de la población.

Estimador: Medida resumen de la variable obtenida a partir de la muestra

Error muestral: Diferencia que existe entre el estimador y el parámetro

N = tamaño poblacional = 5.000 estudiantes
n = tamaño muestral = 500 estudiantes (fórmula estadística)

TIPOS DE DATOS: Cuantitativos

Cuantitativos o Numéricos:	**Discretos:** Se expresan por número enteros	Procesos de conteo: Número de defectos por botella defectuosa
Se expresa por medio de números	**Continuos:** Se expresa por números reales	Procesos de medición: Contenido neto

Cualitativos o Categóricos:	**Ordinales:** Las categorías tienen un orden implícito.	Calidad del etiquetado: *Bueno, Regular, Malo.*
Se expresa por medio de categorías	**Nominales**: No tienen orden, sino indican pertenencia	Tipo de botella: Vidrio, PET

OJO: Las técnicas estadísticas son diferentes para cada tipo de dato!!

Codificación: Escala de Likert — Valores arbitrarios

1	2	3	4	5
Totalmente en Desacuerdo	PD	N	PA	TA

Variable = accesos exitosos
¿cuántas veces accediste exitosamente al Banner esta semana?
= 0, 1, 2, 3…. (discreta)

- 0 veces
- 1 a 3 veces
- 4 a 5 veces
- 6 o más veces

} categorías

PARAMETRO VS ESTIMADOR

PARAMETRO	ESTIMADOR
Resumen poblacional	Resumen muestral
Desconocido	Conocido
Constante	Aleatorio (al azar)

Estadística descriptiva

viernes, 25 de abril de 2025 11:34

Objetivo: Proporcionar técnicas para describir, sistematizar, representar gráficamente y organizar información (datos).

OJO: Las técnicas descriptivas dependen del tipo de datos (Numérico o Categóricos)

Tipos de técnicas descriptivas

- Medidas descriptivas: Calcular resúmenes numéricos de los datos que permitan entender el comportamiento de los datos
- Gráficas descriptivas: Representaciones gráficas del comportamiento de los datos

TECNICAS	DATO CUANTITATIVO	DATO CUALITATIVO
MEDIDAS DESCRIPTIVAS	- Tendencia central - Dispersión - Posición - Forma	- Frecuencias o Conteos - Proporciones o Porcentajes
GRAFICAS BÁSICAS	- Diagramas de puntos - Histogramas, Polígonos de frecuencia y Ojivas de frecuencia - Diagrama de caja o Boxplot	- Gráfico de barras - Gráficos de sectores

Análisis descriptivo para datos numéricos

viernes, 25 de abril de 2025 11:39

- Medidas descriptivas:

Tipo de medidas	**Para que sirven**	**Técnicas**
Tendencia central	Identificar el valor central de la concentración de los datos	Promedio, Mediana, Moda
Dispersión	Medir el grado de dispersión (concentración) de los datos	Rango, Varianza, Desviación estándar
Posición	Identificar puntos de corte a partir de la distribución de los datos	Percentiles, Deciles, Cuartiles
Forma	Identificar si la distribución de los datos se ajustan a un modelo determinado (campana de Gauss = distribución normal)	Coeficiente de Asimetría Coeficiente de Curtosis

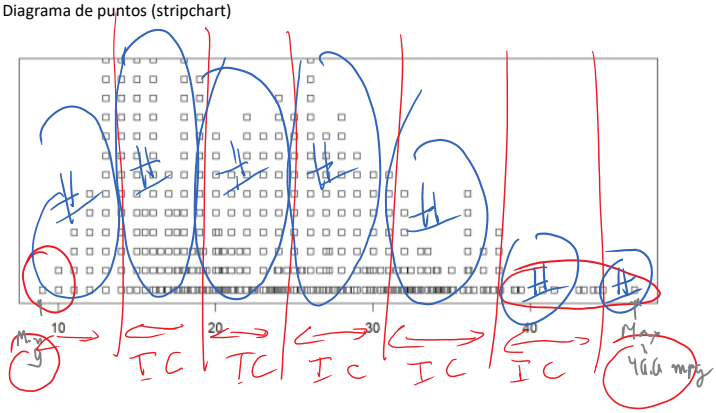
- Gráficos descriptivos:

Tipos de gráficos	**Para que sirven**
Diagrama de puntos	Representa la distribución de los datos, desde el mínimo al máximo
Histograma	Representa la distribución de los datos, a partir de las frecuencias observadas en distintos intervalos de clase de amplitud fija
Ojiva de frecuencias	Representa la frecuencia acumulada de los datos a partir de intervalos de clase
Boxplot	Identificar la presencia de outliers

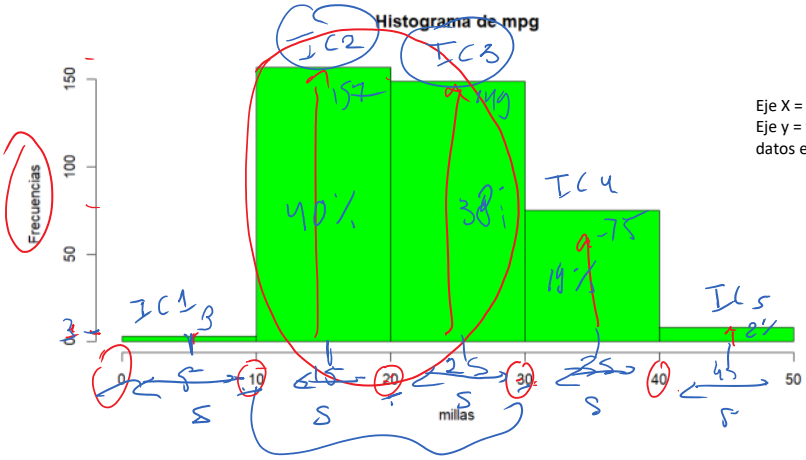
Práctica de R: análisis descriptivo datos numéricos

viernes, 25 de abril de 2025 11:51

Diagrama de puntos (stripchart)



Histograma



Eje X = clases
Eje y = frecuencias = conteos de datos en cada clase

```
> histo$breaks #limites de clase
[1]  0 10 20 30 40 50
> histo$counts #frecuencias absolutas
[1]  3 157 149  75   8
> histo$mids #marcas de clase
[1]  5 15 25 35 45
```

Mayor concentración de datos: 10 y 30 mpg (aprox el 78% de los autos)

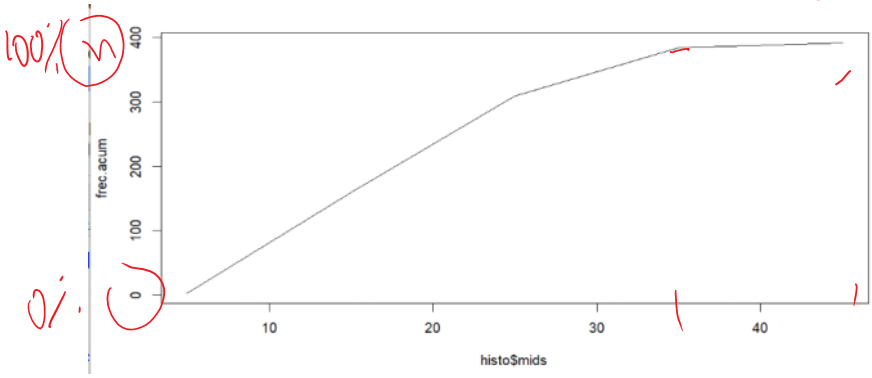
```
> sum(histo$counts)
[1] 392
> histo$counts / n # frecuencias relativas
[1] 0.007653061 0.400510204 0.380102041 0.191326531
[5] 0.020408163
```

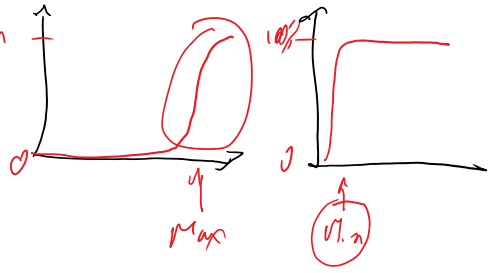
```
> frec.acum
[1]   3 160 309 384 392
```

TABLA DE DISTRIBUCION DE FRECUENCIAS

Clase	Lim Inf	Lim Sup	Frecuencia abs	Frec relat	Frec Abs acum
1	0	10	3	0.7%	3
2	10	20	157	40%	160
3	20	30	149	38%	309
4	30	40	75	19%	384
5	40	50	8	2%	392

OJIVA DE FRECUENCIAS



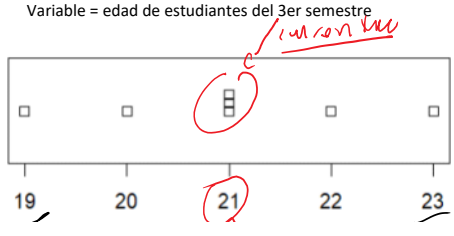


Analizar en que región de los datos se presenta mayor acumulación de información

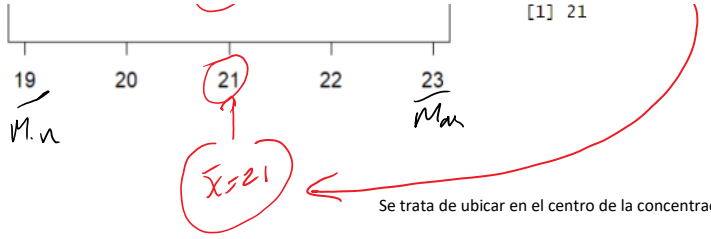
PROMEDIO MUESTRAL

$$\bar{x} = \frac{\sum x}{n}$$

Variable = edad de estudiantes del 3er semestre

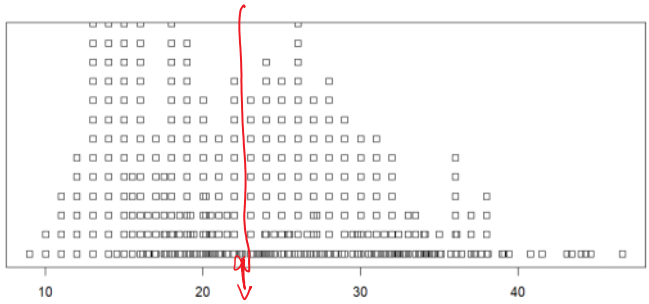



```
> (21+21+21+22+20+19+23)/7
[1] 21
> sum(ejm)/7
[1] 21
> mean(ejm)
[1] 21
```



Se trata de ubicar en el centro de la concentración de los datos

```
> mean(millas)
[1] 23.44592
```



Mediana: Ubicar mediante posición el valor que divida al conjunto de datos en dos subconjuntos de tamaños iguales (n/2 datos = 50% c/u)

- Primero ordenar los datos de menor a mayor
- Ubicar el dato que se encuentre en la posición más cercana a la posición n/2
- 392/2 = 196

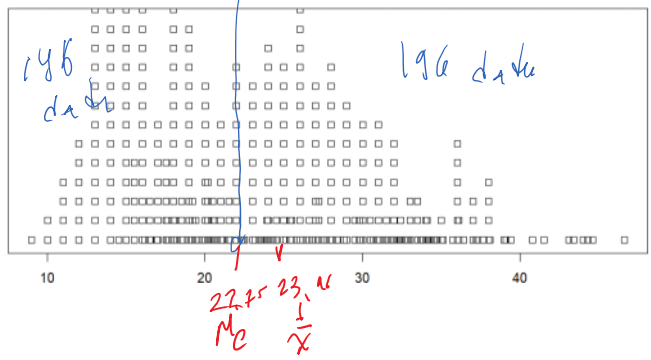
```
> sort(ejm)
[1] 19 20 21 21 21 22 23
```

```
> median(ejm)
[1] 21
```

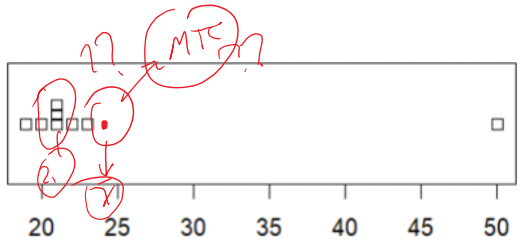
Ojo: si los datos son simétricos (se concentran en el punto medio de la escala), el promedio y la mediana será muy parecidas (o iguales)

```
> median(millas)
[1] 22.75
```

```
> mean(millas)
[1] 23.44592
```



DATOS ATÍPICOS O OUTLIERS:
Datos que se encuentran muy alejados de la concentración de los datos

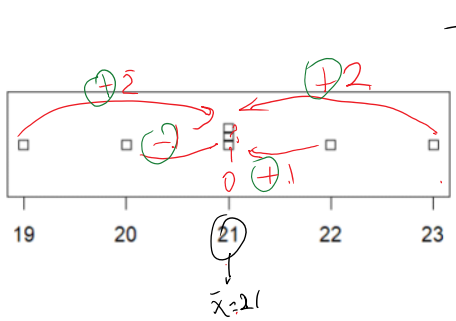



```
> ejm2 = c(21, 21, 21, 22, 20, 19, 23, 50)
> stripchart(ejm2, method = "stack")
> mean(ejm2)
[1] 24.625
> median(ejm2)
[1] 21
> sort(ejm2)
[1] 19 20 21 21 21 22 23 50
```

OJO: El promedio muestral es una buena MTC siempre y cuando no existan datos atípicos. Caso contrario, es suele utilizar a la mediana como MTC más apropiada.

Propiedad del promedio muestral: La suma de desviaciones de cada datos respecto al promedio es nula

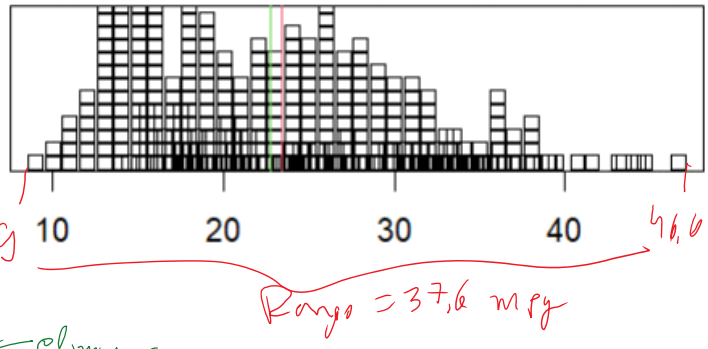
$$\sum (x - \bar{x}) = 0$$




```
> sum(millas - mean(millas))
[1] 4.760636e-13
```

El promedio muestral se puede entender como el centro de gravedad del conjunto de datos

Medidas de dispersión: Medir el grado de dispersión de los datos!



Rango: Max - Min

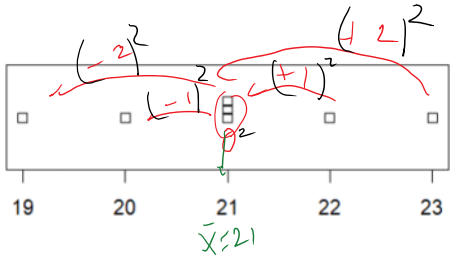
```
> rango = max - min
> rango
[1] 37.6
```

Varianza muestral:

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

Cuasi promedio de las desviaciones cuadradas respecto al promedio muestral

GRADOS DE LIBERTAD = número de datos que puedes variar libremente en un conjunto de datos, prefijando previamente cierta información (en la varianza, primero debemos fijar el promedio)



$$\sum (x - \bar{x}) = 0$$

$$\frac{\sum d^2(+)}{n} = \bar{d}^2$$

```
> sum((ejm - 21)^2)/ (7-1)
[1] 1.666667
> var(ejm) # 1.66 a?os cuadrados ???
[1] 1.666667
```

Ojo: la varianza está elevado al cuadrado, por tanto sus unidades también

```
> var(millas)
[1] 60.91814
```

Mpg al cuadrado????? - varianza no es interpretable!!

Desviación estándar:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

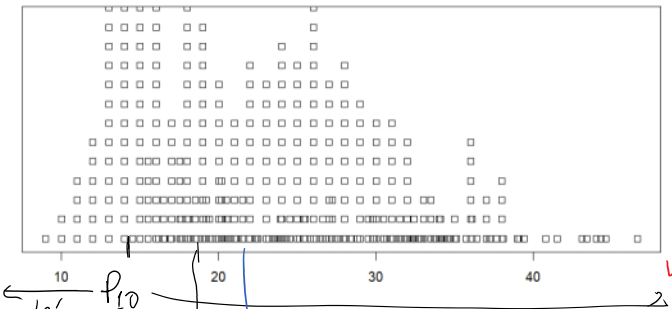
Raiz cuadrado del cuasi promedio de desviaciones cuadradas
Ventaja = tiene las mismas unidades de los datos originales

```
> sqrt(var(ejm))
[1] 1.290994
> sd(ejm)
[1] 1.290994
```

```
> sd(millas)
[1] 7.805007
```

Análisis de las medidas de dispersión: A mayor valor de la varianza o desv. Estándar, los datos son más dispersos.!!

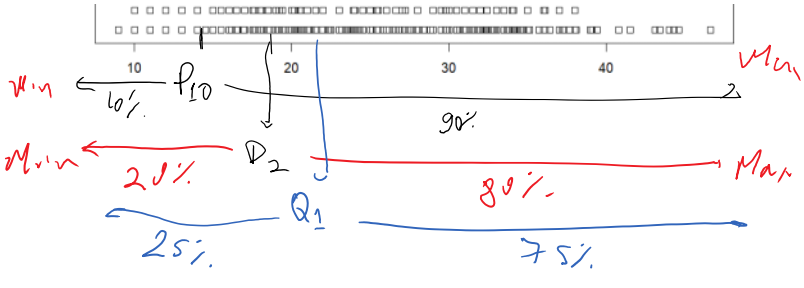
MEDIDAS DE POSICIÓN



PERCENTILES: Puntos de corte al 1% de datos: P10
DECILES: Puntos de corte al 10% de los datos: D2 = decil 2 = P20
CUARTILES: Puntos de corte al 25% de los datos: Q1 = P25

OJO: Mediana = Q2 = D5 = P50

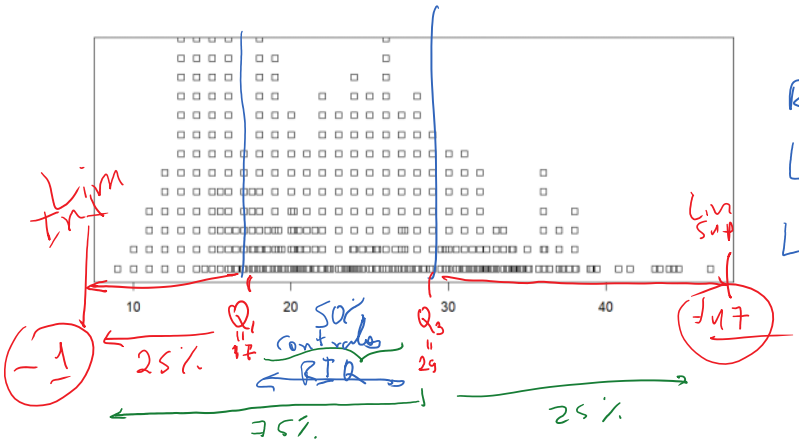
```
> quantile(millas, probs = 0.05) #P5
5%
13
> quantile(millas, probs = 0.50) #P50 = mediana = segundo cuartil
50%
22.75
```




```
> quantile(millas, probs = 0.50) #P50 = mediana = segundo cuartil
50%
22.75

> median(millas)
[1] 22.75
> quantile(millas, probs = 0.25) #P25 = Q1 =primer cuartil
25%
17
> quantile(millas, probs = 0.75) #P75 = Q3 = tercer cuartil
75%
29
```

DIAGRAMA DE CAJAS - BOXPLOT: Gráfico basado en las Q1, Q2 y Q3 que se utiliza para identificar datos outliers

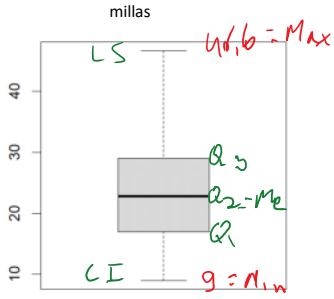


$RIQ = Q_3 - Q_1 = 29 - 17$

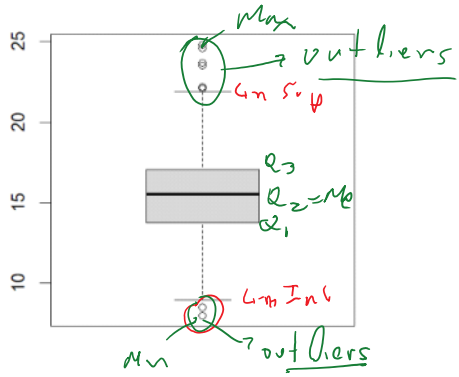
$Lim\ Inf = Q_1 - 1.5\ RIQ$

$Lim\ Sup = Q_3 + 1.5\ RIQ$

RIQ = 29 - 17 = 12 pmg
Lim Inf = 17 - 1,5*12 = -1
Lim Sup = 29 + 1,5*12 = 47



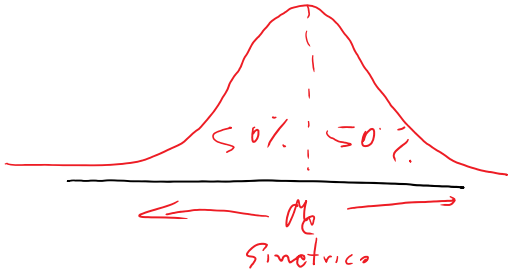
Para la variable millas, no existen datos atípicos!!

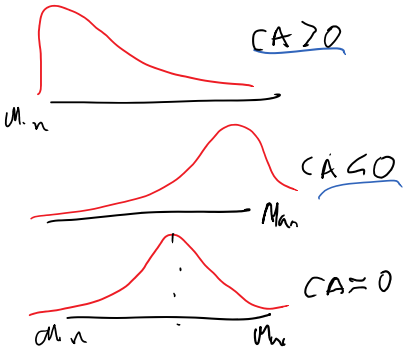


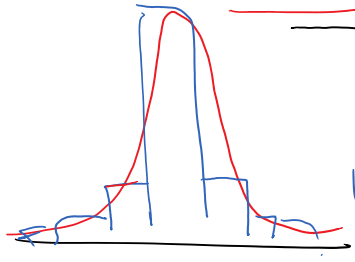
FUNCION DESCRIBE - PAQUETE PSYCH

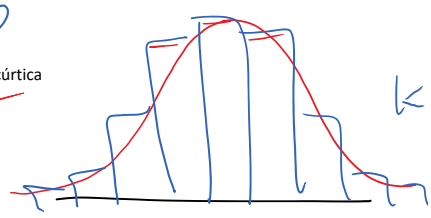
```
> library(psych)
> ?describe
> describe(millas)
   vars   n  mean   sd median trimmed mad min  max range
X1    1 392 23.45 7.81  22.75   22.99 8.6   9 46.6  37.6
   skew kurtosis   se
X1 0.45    -0.54 0.39
```

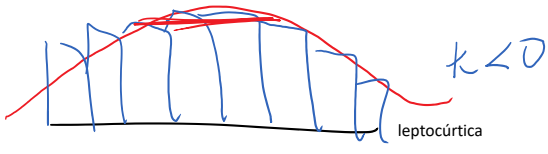
CA = ASIMETRIA K = KURTOSIS

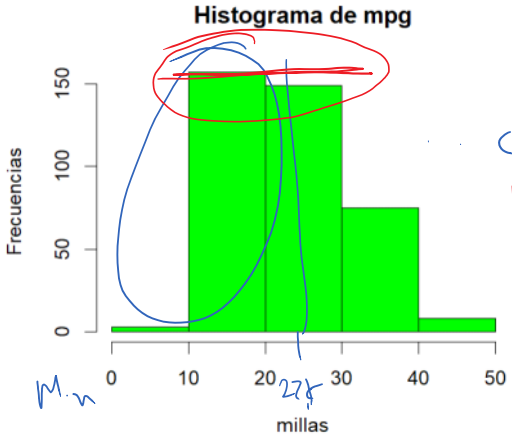












Datos no son simétricos (tienen una tendencia hacia el mínimo)
Datos presentan un pequeño grado de aplanamiento

Análisis descriptivo de datos categóricos

lunes, 28 de abril de 2025 11:50

Medidas descriptivas	Frecuencias o conteos: # de datos que corresponden a cada categoría (nivel) Proporciones: % de datos que corresponden a cada categoría
Gráficos	Barras: sirve para comparar las frecuencias entre todas las categorías Circular: sirve para representar la composición porcentual de la muestra

VARIABLE CODIFICADA!!

origin

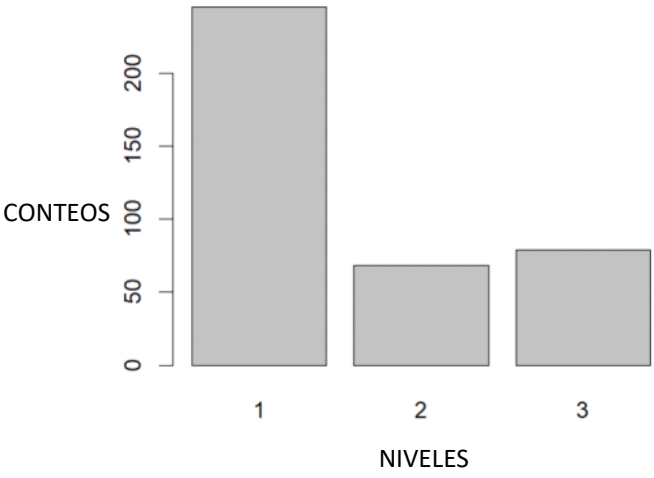
Origin of car (1. American, 2. European, 3. Japanese)

```
> conteos = table(origen) # table = conteos de cada nivel
> addmargins(conteos) # frecuencias absolutas
origen
  1   2   3 Sum
245  68  79 392
```

```
> proporciones = prop.table(conteos) # proporciones
> addmargins(proporciones) # frecuencias relativas
origen
        1         2         3       Sum
0.6250000 0.1734694 0.2015306 1.0000000
```

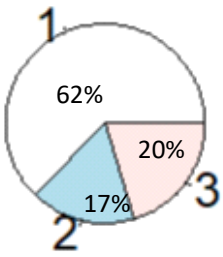
245/392 = 0,625 79/392 = 0,2015

GRAFICO DE BARRAS



Ojo: Barra más alta = clase modal (Moda)

GRAFICO CIRCULAR



Ojo: el grafico circular no es adecuado cuando existen demasiadas categorías

ESTE GRAFICO NO SIRVE!!

ANALISIS CONJUNTO = BIVARIANTE

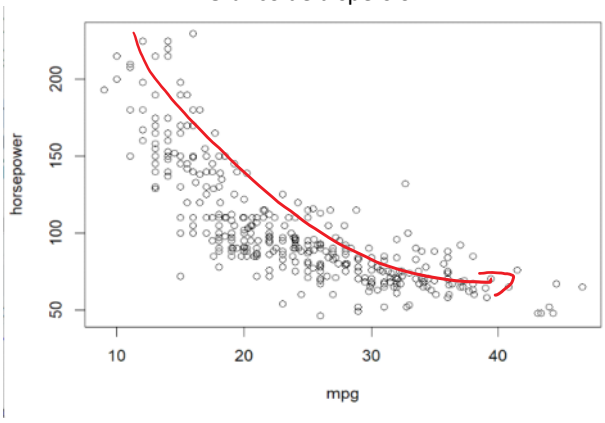
lunes, 28 de abril de 2025 12:11

Objetivo: describir el comportamiento conjunto de dos variables X y Y

TIPOS DE VARIABLES	TECNICAS DESCRIPTIVAS
X y Y son numéricas	- Correlación lineal: R - Gráfico de dispersión
X y Y son categóricas	- Tabla de contingencia (cruzada) - Gráficos de barras apiladas
X es numérica y Y categóricas (o al revés)	- Análisis descriptivo por niveles (Y) - Boxplot para cada nivel (Y)

Caso 1: Ambas numéricas
X = mpg, Y = horsepower

Gráfico de dispersión

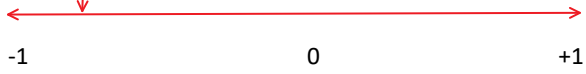


Se evidencia una relación INVERSAMENTE PROPORCIONAL entre mpg y horsepower!!

```
> plot(mpg, horsepower) # diagrama de dispersion
> cor(mpg, horsepower) # correlacion
[1] -0.7784268
```

Correlación lineal de Pearson = R = -0.77

R = -0.77



-1 0 +1

INVERSAMENTE PROPORCIONAL | INCORRELADAS | DIRECTAMENTE PROPORCIONAL

OJO: No confundir correlación con causalidad!!
Correlación = asociación (comportamiento conjunto)
Causalidad = causa - efecto (explicativo)

CASO 2: Ambas variables categóricas
X = year (modelo del auto), Y = origen

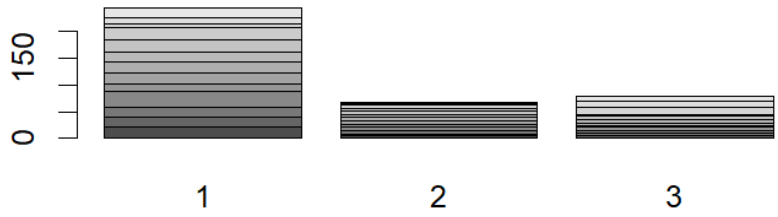
TABLA CRUZADA / CONTINGENCIA

```
> tabla = table(year, origen) # tabla de contingencia
> tabla
      origen
year   1  2  3
  70  22  5  2
  71  19  4  4
  72  18  5  5
  73  29  7  4
  74  14  6  6
  75  20  6  4
  76  22  8  4
  77  18  4  6
  78  22  6  8
  79  23  4  2
  80   6  8 13
  81  13  3 12
  82  19  2  9
```

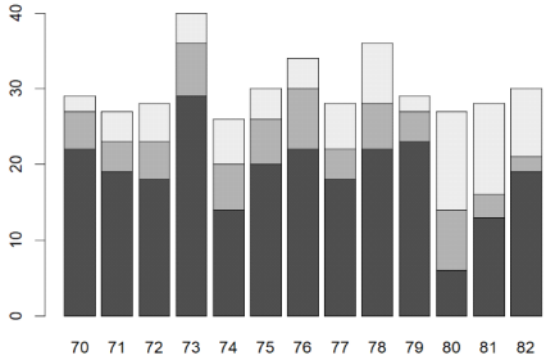
TABLA DE PROBABILIDADES

```
> addmargins(prop.table(tabla)) # tabla de probabilidad
     origen
year           1           2           3         Sum
  70  0.056122449 0.012755102 0.005102041 0.073979592
  71  0.048469388 0.010204082 0.010204082 0.068877551
  72  0.045918367 0.012755102 0.012755102 0.071428571
  73  0.073979592 0.017857143 0.010204082 0.102040816
  74  0.035714286 0.015306122 0.015306122 0.066326531
  75  0.051020408 0.015306122 0.010204082 0.076530612
  76  0.056122449 0.020408163 0.010204082 0.086734694
  77  0.045918367 0.010204082 0.015306122 0.071428571
  78  0.056122449 0.015306122 0.020408163 0.091836735
  79  0.058673469 0.010204082 0.005102041 0.073979592
  80  0.015306122 0.020408163 0.033163265 0.068877551
  81  0.033163265 0.007653061 0.030612245 0.071428571
  82  0.048469388 0.005102041 0.022959184 0.076530612
  Sum 0.625000000 0.173469388 0.201530612 1.000000000
```

GRAFICO DE BARRAS APILADAS

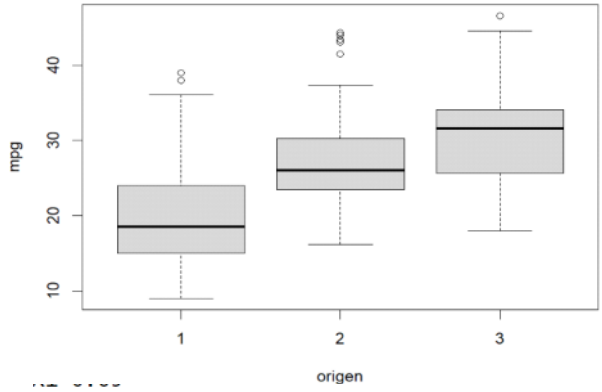



```
> tabla2 = table(origen, year) # tabla de contingencia
> barplot(tabla2) # grafico de barras apiladas
```



CASO 3: Una numérica y otra categórica
OJO: Repetir el análisis descriptivo de la variable numérica para cada nivel de la variable categórica

X = mpg (numérica), Y = origen (categórica)



El comportamiento estadístico de los subgrupos es distinto al comportamiento de la muestra en general!!

- mpg de los americanos es muy inferior al de los japoneses!!

```
> describeBy(mpg, origen) # analisis descriptivo por subgrupo

group: 1
   vars   n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 245 20.03 6.44   18.5   19.37 6.67   9  39    30 0.83     0.03 0.41
------------------------------------------------------------
group: 2
   vars  n mean   sd median trimmed  mad  min  max range skew kurtosis  se
X1    1 68 27.6 6.58     26    27.1 5.78 16.2 44.3  28.1 0.73     0.31 0.8
------------------------------------------------------------
group: 3
   vars  n  mean   sd median trimmed  mad min  max range skew kurtosis   se
X1    1 79 30.45 6.09   31.6   30.47 6.52  18 46.6  28.6 0.01    -0.39 0.69
```