Sesión 2: Estadística Descriptiva

EUSEBIO ANGULO SÁNCHEZ-HERRERA

LABORATORIO DE ESTADÍSTICA

ESCUELA SUPERIOR DE INFORMÁTICA CIUDAD REAL

Estadística Descriptiva

- **Estadística Descriptiva:** Describe, analiza y representa un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan la información contenida en ellos.

- **Tipos de variables:**
  - **Variables cuantitativas (discretas y continuas)**
    - Gráficos: Histogramas y *Boxplot*
    - Medidas de tendencia central, dispersión, posición y forma.
  - **Variables cualitativas (nominales y ordinales)**
    - Gráficos: G. de barras y G. Circular.
    - Medidas de tendencia central, dispersión y posición.

Estadística Descriptiva

- **Objetivo:** recoger, organizar y analizar datos.
- **Mediante:** descripciones gráficas y numéricas.
- Y solamente sacando conclusiones sobre **datos procesados.**
- Primero determinar tipo de variable, por ejemplo debido a su naturaleza tenemos:

A Continua

B Discreta

☐ Número de botellas manufacturadas

2 3 12 23 0 4 12 4

13 2 0 2 1 4 0 2

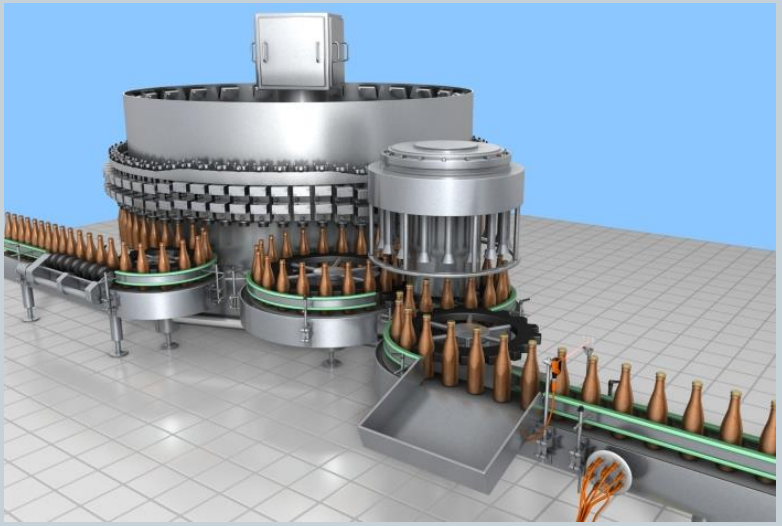
☐ Contenido neto de las botellas

215 195 190 192 192 206

205 199 190 215 207 192

Ejemplo

- El fichero “espera.txt” contiene el tiempo entre paradas en proceso de llenado.

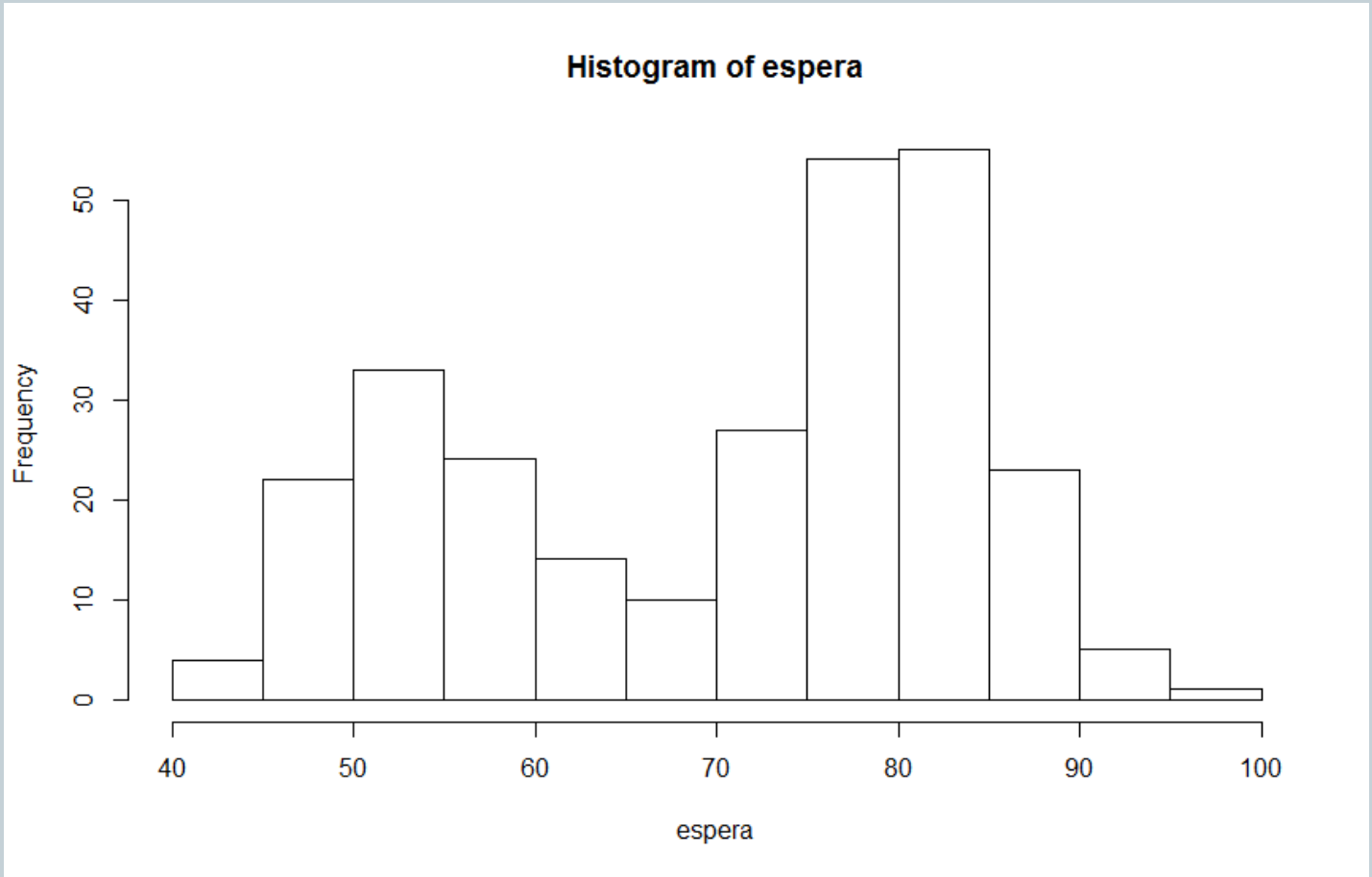


1. **Modificar directorio de trabajo:** >setwd(“…”) o …(menú)…
2. **Cargar datos:** >espera<-scan(file = “espera.txt”)
3. **Summary:** >summary(espera)

```
> summary(espera)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   43.0    58.0    76.0    70.9    82.0    96.0
```

Histograma

- **>hist(espera)**



Histograma

> hist(espera)$counts #obtener un valor de un objeto

[1] 4 22 33 24 14 10 27 54 55 23 5 1

- **Se pueden indicar el número de barras**

>hist(espera,breaks=3)

>hist(espera,breaks=3)$breaks area en funcion de los intervalos que cogemos

>hist(espera,breaks=30)$breaks #no hace las 30

>hist(espera,probability=TRUE, main="Proceso de llenado", xlab="tiempo")

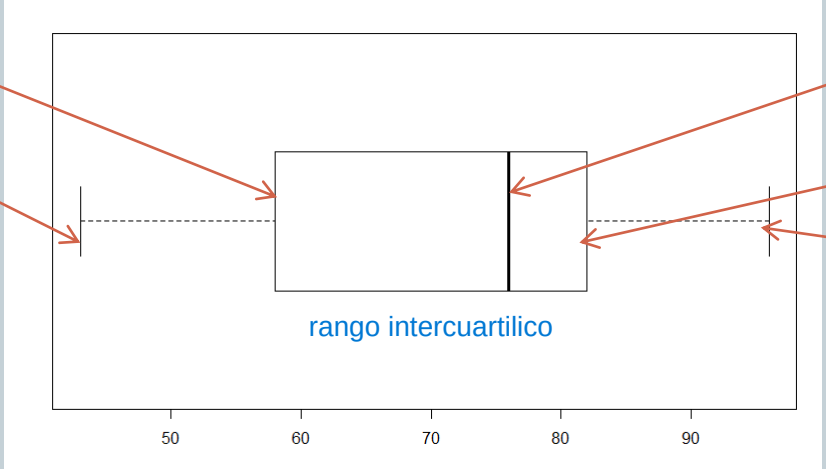
>?hist

Boxplot

- **>boxplot(espera,horizontal=TRUE)**

25%

min



mediana

75%

max

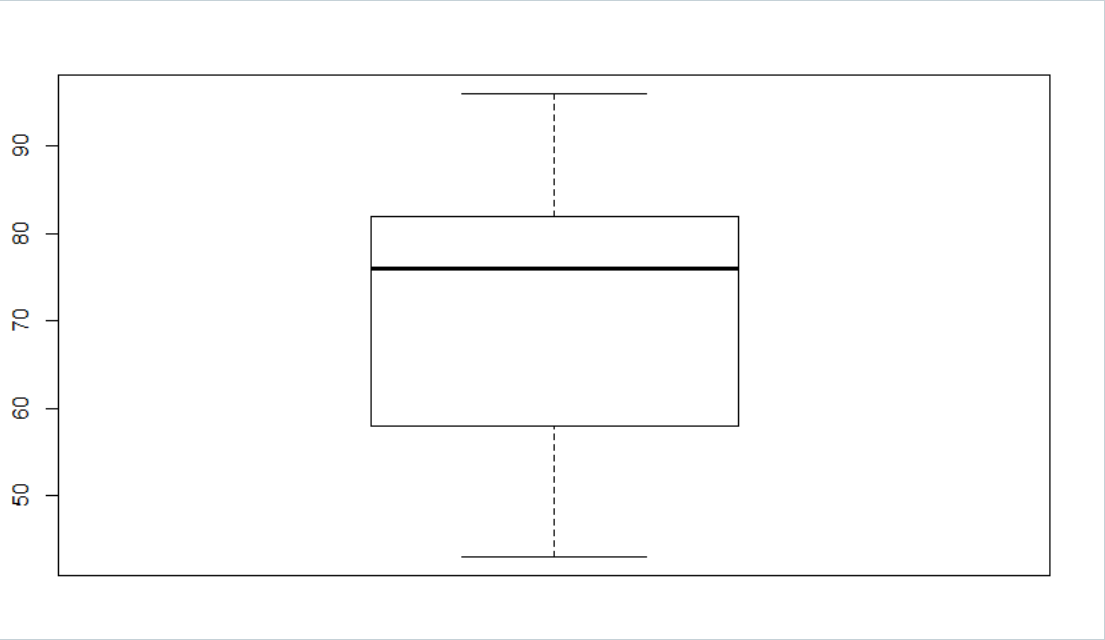
- Cuartiles, min y max (salvo que haya *outliers*)
- I=Mínimo de los datos ϵ [Q1-1.5*IQR,Q3+1.5*IQR]
- Q_1, Q_2 =Me (50%), Q_3
- S= Máximo de los datos ϵ [Q1-1.5*IQR,Q3+1.5*IQR]

valor minimo bigote

Boxplot

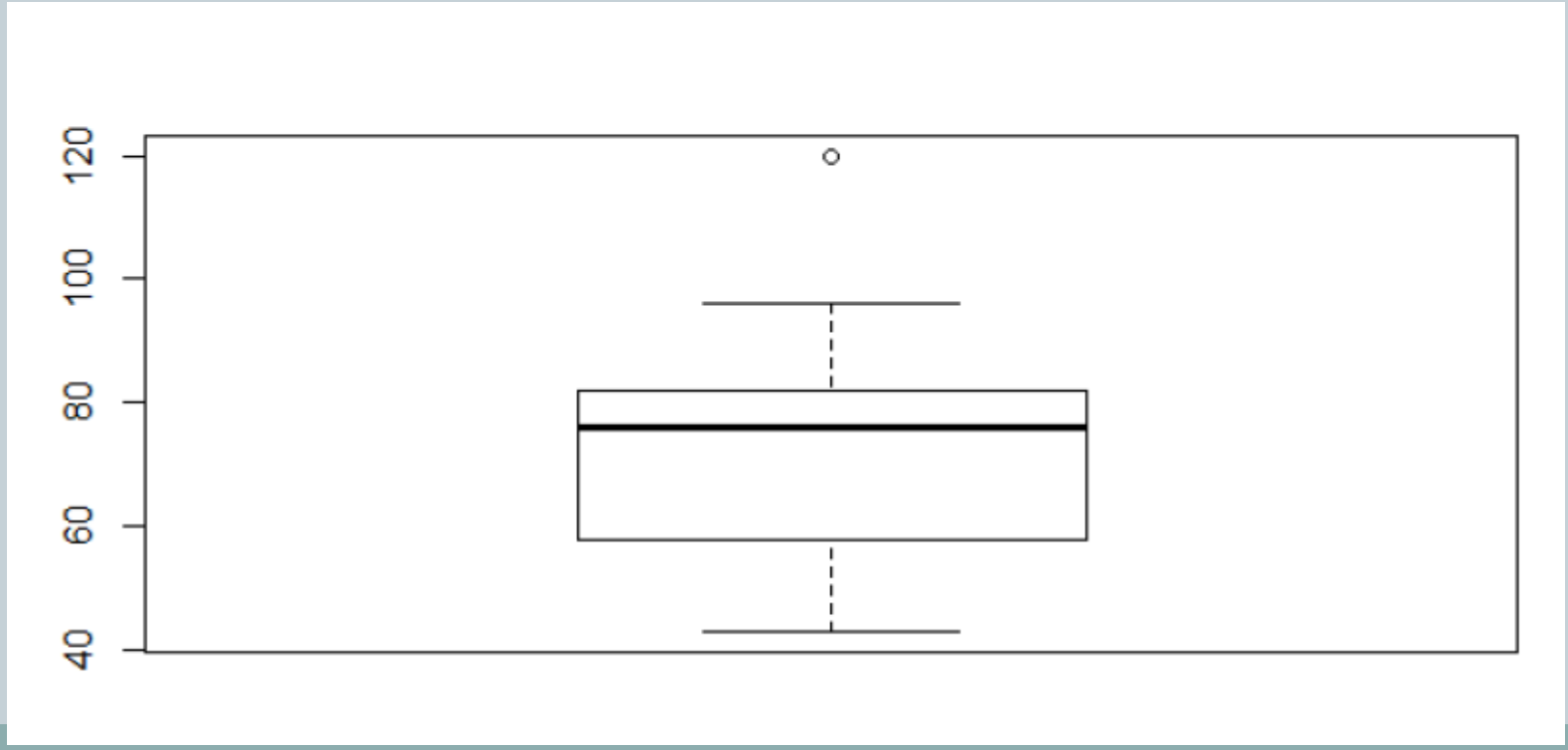
- >boxplot(espera)$stats devuelve bigote mas pequeño 25% mediana 75% y bigote mas grande no se ve afectado por aurier
- >fivenum(espera) #En este caso son ambos igual pero no siempre es así

te devuelve el valor maximo real si se tienen aurier



Boxplot: outliers

- > esperaE=c(espera,120) #en este caso stats y fivenum no coinciden
- > box_var=boxplot(esperaE)
- > box_var$out te devuelve un vector con los aurier que tiene



Medidas para la tendencia central

- **Media, mediana y moda**
  - >mean(espera)
    [1] 70.89706
  - >median(espera)
    [1] 76

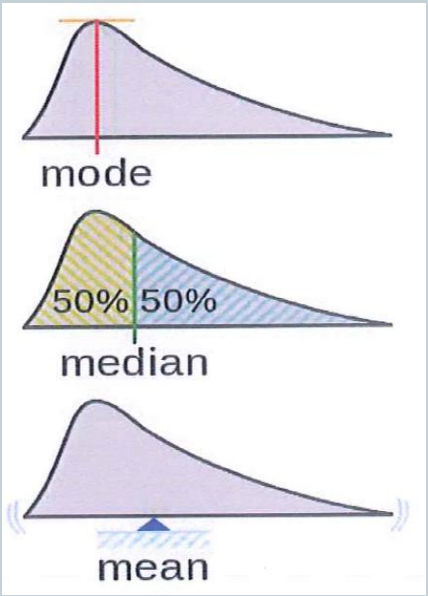


table agrupa los datos por frecuencia

  - > which.max(table(espera)) #genera tabla f.
    78 # valor más repetido
    34 # posición
  - Comprobar como queda la media, mediana y moda (*outilier*)
  - >esperaE<-c(espera,150)

Medidas de dispersión

- >var(espera) # Varianza

  [1] 184.8233
- >sd(espera) # Desviación típica

  [1] 13.59497
- >sd(espera)/abs(mean(espera)) # Coeficiente de variación

  [1] 0.1917565 CV = desviacion tipica X / media X coeficiente de pirson para elegir variable dependiente o independiente
- >IQR(espera) # Rango Intercuartil

  [1] 24
- >range(espera)

  [1] 43 96 vector con el valor mas pequeño y mayor d elos datos

Medidas de posición

- **Cuartiles y Percentiles**
  - >quantile(espera) # Cuartiles

    0% 25% 50% 75% 100%

    43 58 76 82 96
  - >quantile(espera, 0.1)

    10%

    51
  - >quantile(espera, c(.21,.15,.9))

    21% 15% 90%

    55 53 86
  - ?quantile #type(tiene 9 algoritmos implementados)

Medidas

si tenemos aurier no es simetrico

- **Coeficiente de asimetría de Fisher (unimodal)**

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{S_x^3}$$

  - En función del signo del estadístico diremos que la asimetría es positiva o negativa. Si distribución es simétrica, asimetría es cero.
  - formula skewness
  - sum((espera-mean(espera))^3)/(length(espera)*sd(espera)^3)

    [1] -0.414025
  - Descargar paquete fBasics (comandos o directorio paquetes)
  - > install.packages ("fBasics"); library("fBasics"); skewness(espera)

    [1] -0.414025 #otra forma sencilla: fBasics::skewness(espera)

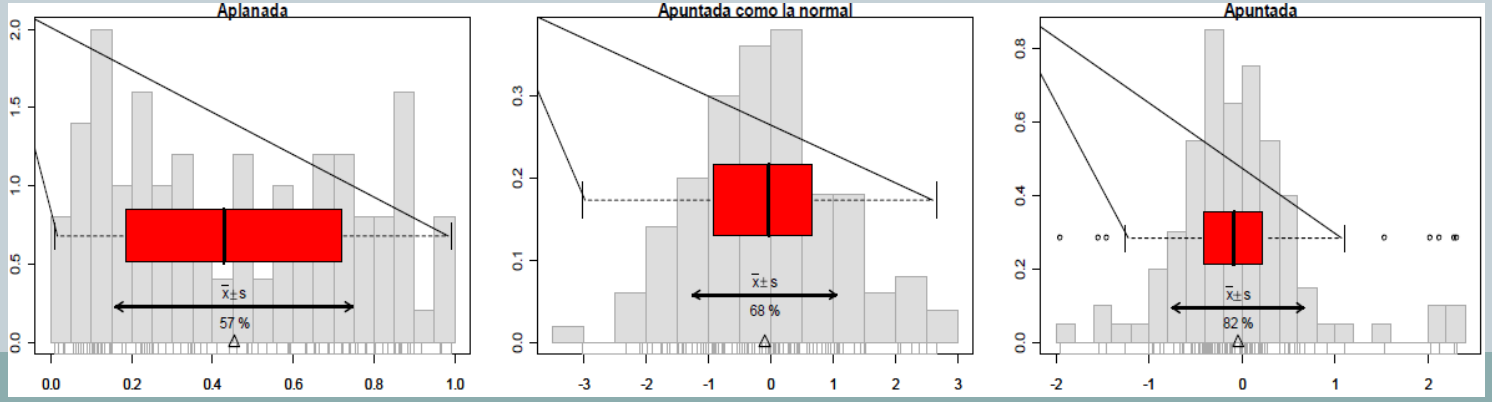
Medidas

- **Coeficiente de Fisher's Kurtosis (unimodal)**

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{S_x^4} - 3$$

La curtosis nos indica el grado de apuntamientos (aplastamiento) de una distribución con respecto a la distribución normal o gaussiana

- Platicúrtica (aplanada): curtosis < 0
- Mesocúrtica (como la normal): curtosis =0
- Leptocúrtica (apuntada): curtosis > 0



Medida

- > sum((espera-mean(espera))^4)/ (length(espera)*sd(espera)^4)-3

  [1] -1.156263

- > kurtosis(espera)

  [1] -1.156263

- A tener en cuenta: el paquete **moments** incluye
  - skewness: pero utiliza otro coeficiente distinto a Fisher
  - kurtosis: pero no resta el 3 de la fórmula
  - moments::skewness(), moments::kurtosis()

Variables cualitativas

- Las variables cualitativas pueden ser por su naturaleza Nominales o Ordinales. no se puede hace varianza media
- Ejemplos:
- Resultados de una encuesta (si, no, no contesta):

  >encuesta=

  c(“si”,”no”,”no”,”si”,”nc”,”si”,”no”,”si”,”nc”,”no”)

- Clientes grado de satisfacción (1-3): (1=Poor, 2=Ok, 3=Good)

  >satisfaccion=c(1,3,2,1,1,2,1,2,3,3,1,1,2,3,2,1,2,3,3,2)

Tabla de frecuencias

- Frecuencias absolutas:

  >table(encuesta)

```
nc no si
 2  4  4
```

  >table(satisfaccion)

- Frecuencias relativas:

  >table(encuesta)/length(encuesta)

```
  nc   no   si
0.2  0.4  0.4
```

  >table(satisfaccion)/length(satisfaccion)

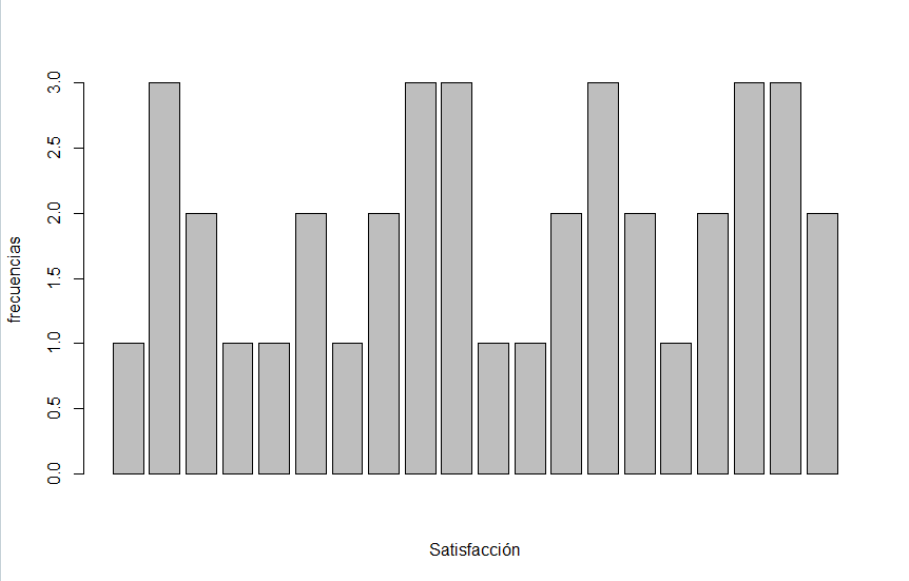
- Frecuencias acumuladas:

  >cumsum(table(encuesta))

Diagramas de barras

cualitativas rep grafica

- >barplot(satisfaccion,xlab=“Satisfacción”,ylab=“frecuencias”)
- > barplot(table(satisfaccion),xlab="Satisfacción", ylab="frecuencias")



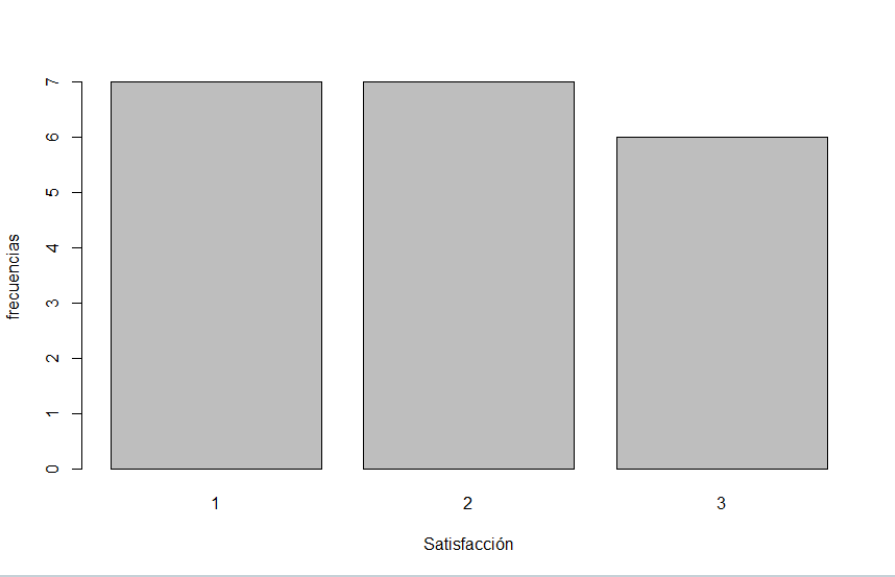
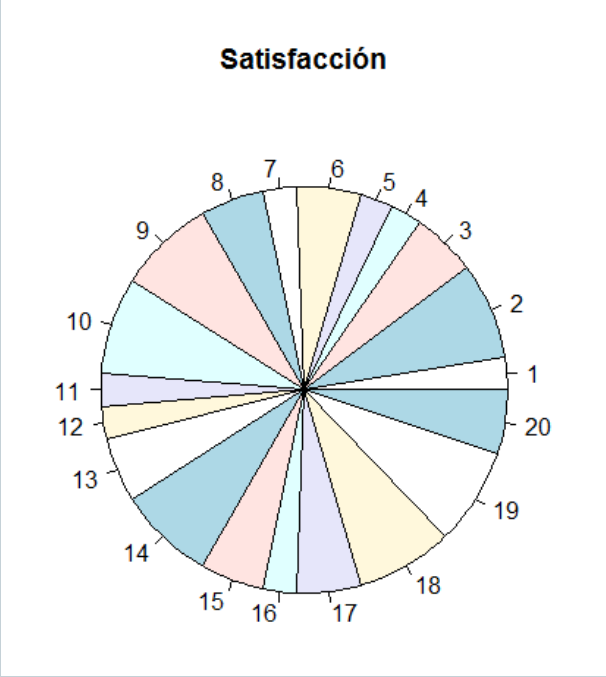


Gráfico circular

- >pie(satisfaccion,main="Satisfacción") # de todas las variables
- > pie(table(satisfaccion),main="Satisfacción") # tabla frecuenc



Medidas de tendencia central

- **Caso encuesta**:

```
>which.max(table(encuesta)) # Significativa cualitativas
no                          # Medidas no significativas para Variables
2                           # Cualitativas  Media y mediana?
```

- **Caso sastisfacción**:

```
>which.max(table(satisfaccion))
1
1
>mean(satisfaccion); median(satisfaccion) # No tienen sentido
[1] 1.95
[1] 2
```

Medidas de dispersión y posición

- Interpretar los resultados obtenidos en teoría con los resultados obtenidos con estas instrucciones para la **variable satisfacción**

  >var(satisfaccion)

  >sd(satisfaccion)

  >sd(satisfaccion)/abs(mean(satisfaccion))

  >IQR(satisfaccion)

  >quantile(satisfaccion)

  ??? #no tienen sentido ni para satisfacción ni para encuesta

- Repetir para **variable encuesta**.

Estructuras de control en R
Condicionales

- **if(){}else{}**

```
x <- 3
if(x>2){
    y <- 3
}else{
    y <- 5
}
y
```

- Puede asignarse el resultado a una variable:

```
    y <- if(x > 3) { 10} else { 0}
```

- Pueden anidarse

Estructuras de control en R
Bucles

- **for(){}** Repite una acción un nº determinado de veces

```
x<-c("a","b","c","d")
for(i in 1:length(x)) {
   print(x[i])
}
```

- **while(){}** Repite una acción mientras se cumpla la condición

```
count <- 0
while(count < 10) {
   print(count)
   count <- count + 1
}
```

- Pueden anidarse

Estructuras de control en R Ejemplo

comparar dos valores para ver cual tiene mas valor

- **Tipificación** $$Z_x = \frac{X - \overline{X}}{S_x}$$

```
# ¿En que conjunto de calificaciones tiene más mérito un 8?
calificaciones=replicate(10, runif(20)*10) #matriz 10x20 aleatoria
tip=NULL #inicializamos la variable
nota=8
for(i in 1:ncol(calificaciones)){ #recorremos por columnas
   aux=calificaciones[,i] #guardamos la columna
   tip=c(tip,(nota-mean(aux))/sd(aux)) #añadimos la tip al vector
}
which(max(tip)==tip) #preguntamos cual es la mayor
```

Estructuras de control en R
Ejemplo

- **Tipificación** $$Z_x = \frac{X - \bar{X}}{S_x}$$

Repetir un ejemplo de tipificación generando solo dos grupos de notas aleatoriamente.

Ejercicios

- Los datos del fichero Session3_var.Rdata representan las notas obtenidas por 20 estudiantes en las 10 asignaturas que tuvieron en un curso completo.
  - 1) Obtener *summary* con los principales valores de la estadística descriptiva de todas las asignaturas contenidas en el *dataset*.
  - 2) Obtener *summary* con los principales valores de la estadística descriptiva de la primera asignatura.
  - 3) ¿Cuál es la asignatura con mayor nota media?
  - 4) Realizar un *histograma* completo del *dataset*.
  - 5) ¿Hay algún *outliers* en alguna asignatura? Piensa que podría causar.

Ejercicios

- Los datos del fichero Session3_var.Rdata representan las notas obtenidas por 20 estudiantes en las 10 asignaturas.
  - 6) ¿Cuántas asignaturas tienen más del 40% de estudiantes con una nota mayor que 8?
  - 7) Comparando la séptima y la novena asignatura: ¿Cual de ellas es más similar a una distribución normal? ¿Como son llamadas dependiendo del valor del parámetro utilizado?
  - 8) Comparando la quinta, octava y décima asignatura: ¿Cual de ellas podemos considerar más simétrica?¿Qué podemos decir de las otras asignaturas?
  - 9) ¿Qué percentil representa una nota de 8 para la primera asignatura?
  - 10) ¿En qué asignatura tiene más mérito obtener un 9?

Sumario

- Repasar conceptos vistos en Tema 2 de Estadística Descriptiva. Se utilizan diferentes funciones con Variables Cualitativas y Cuantitativas
- Manejar entrada y salida de datos en R
- Representación de gráficas e imágenes
- Paquetes
- Analizar e interpretar los resultados obtenidos
- Estructuras de control en R: condicionales y bucles