Analisi Univariata
&
Esercizi

Metodi Quantitativi per Economia, Finanza e Management

Esercitazione n°3

Riepilogo lezioni precedenti…

LEZIONE 1: Introduzione a R

LEZIONE 2: Il questionario

Nota:

- Rispettare l'ordine delle sezioni del questionario:

  - domande comportamentali (inerenti all'obiettivo)

    NB.: inserire almeno 15 domande con scale di punteggio

  - domande anagrafiche/socio-demo

  - domande attitudinali

Lavoro di Gruppo

- Se qualche studente fosse intenzionato a svolgere l'esame da frequentante ma non ha ancora formato un gruppo da 3-4 persone, venga a riferircelo a fine lezione in modo tale da poter formare noi i gruppi

- Inviare entro il **19/10/2017** via e-mail il questionario da validare

- Attendere la validazione con eventuali correzioni via e-mail prima di iniziare la somministrazione

Prima di iniziare..

- Controllare se sul pc su cui state lavorando esiste già una cartella C:\corso. In tal caso eliminare tutto il contenuto. In caso contrario creare la cartella **corso** all'interno del disco C

- Andare sul disco condiviso F nel percorso ***F:\corsi\Metodi_Quantitativi_EFM_1819\esercitazione3*** e copiare il contenuto nella cartella C:\corso

- Aprire il programma R (Start → All Programs → R)

- Cambiare la directory di lavoro puntando il percorso fisico C:\corso, utilizzando l'istruzione

```
setwd('C:/Corso')
```

- Importare il file CSV telefonia.csv nell'oggetto R *telefonia* con il comando

```
telefonia=read.csv('telefonia.csv', header=TRUE)
```

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:

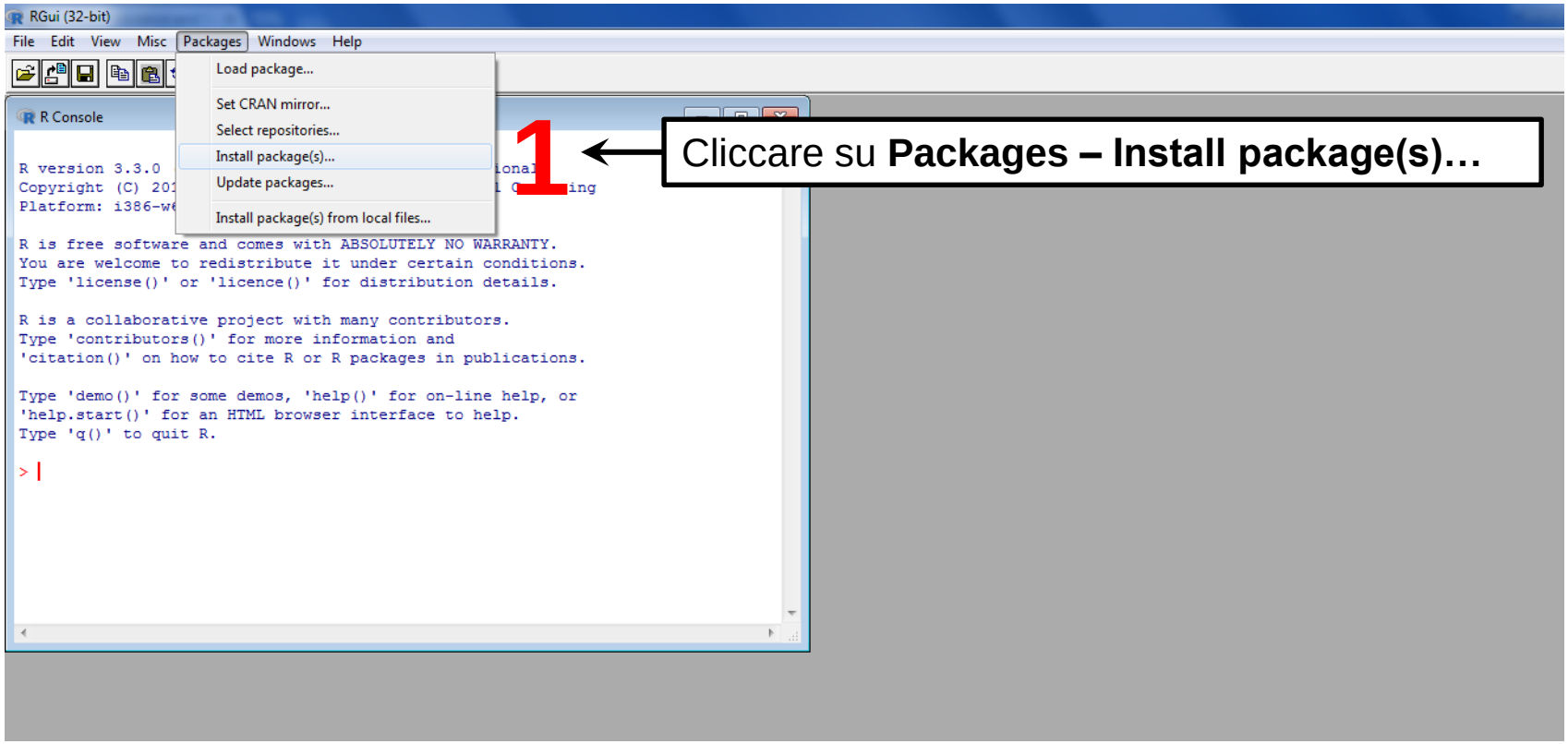
Installazione dei pacchetti

Con l'installazione del software R vengono scaricati numerosi pacchetti di base, ma molte altre funzioni possono essere aggiunte grazie a pacchetti e plugins aggiuntivi, disponibili in un apposito sito (repository): il **CRAN** (Comprehensive R Archive Network)

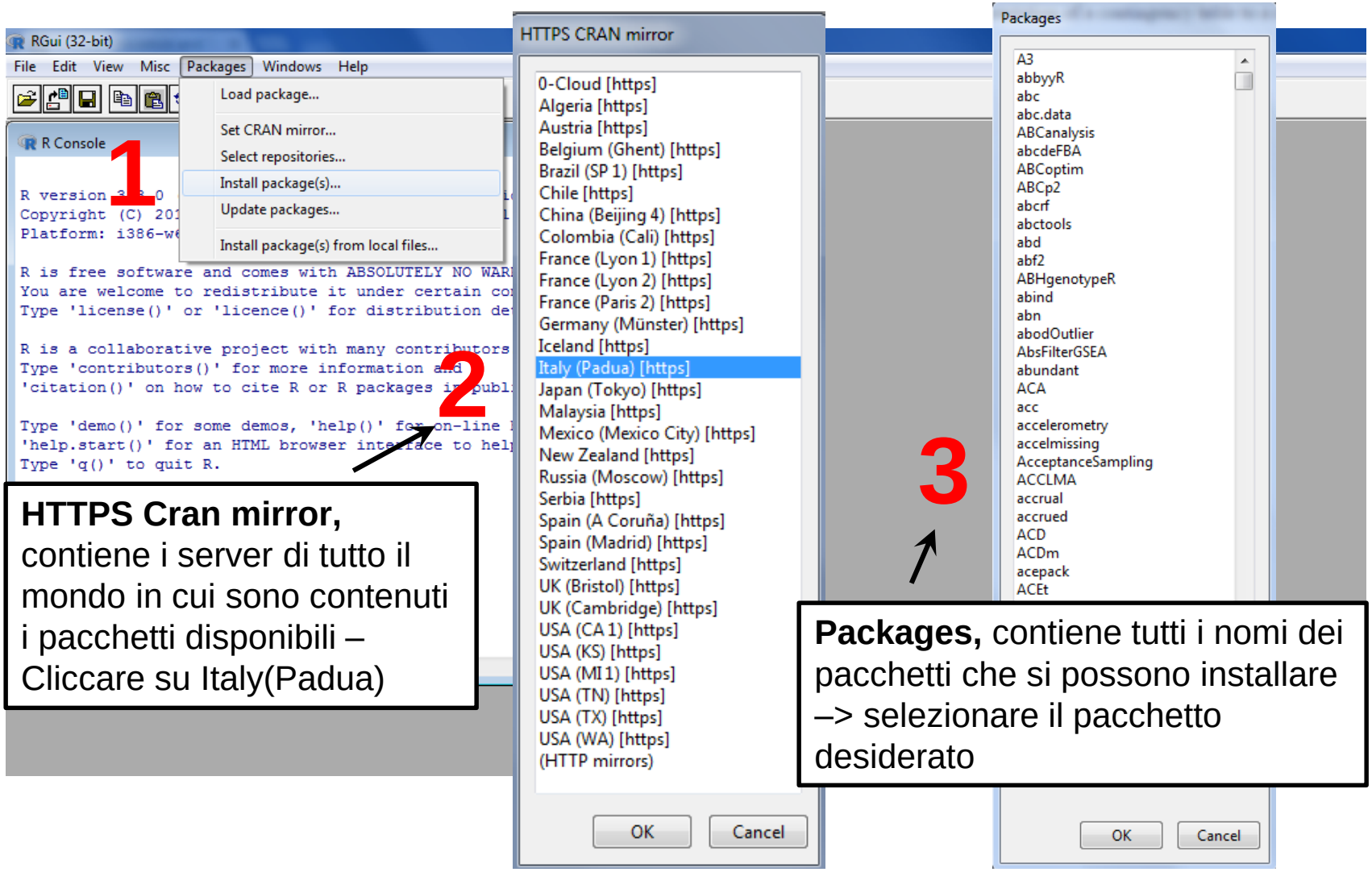
Installazione dei pacchetti

N.B.: L'installazione dei pacchetti deve essere fatta solo una volta dopo l'installazione di R e non ad ogni sua successiva apertura.

Come installare un pacchetto:



Installazione dei pacchetti



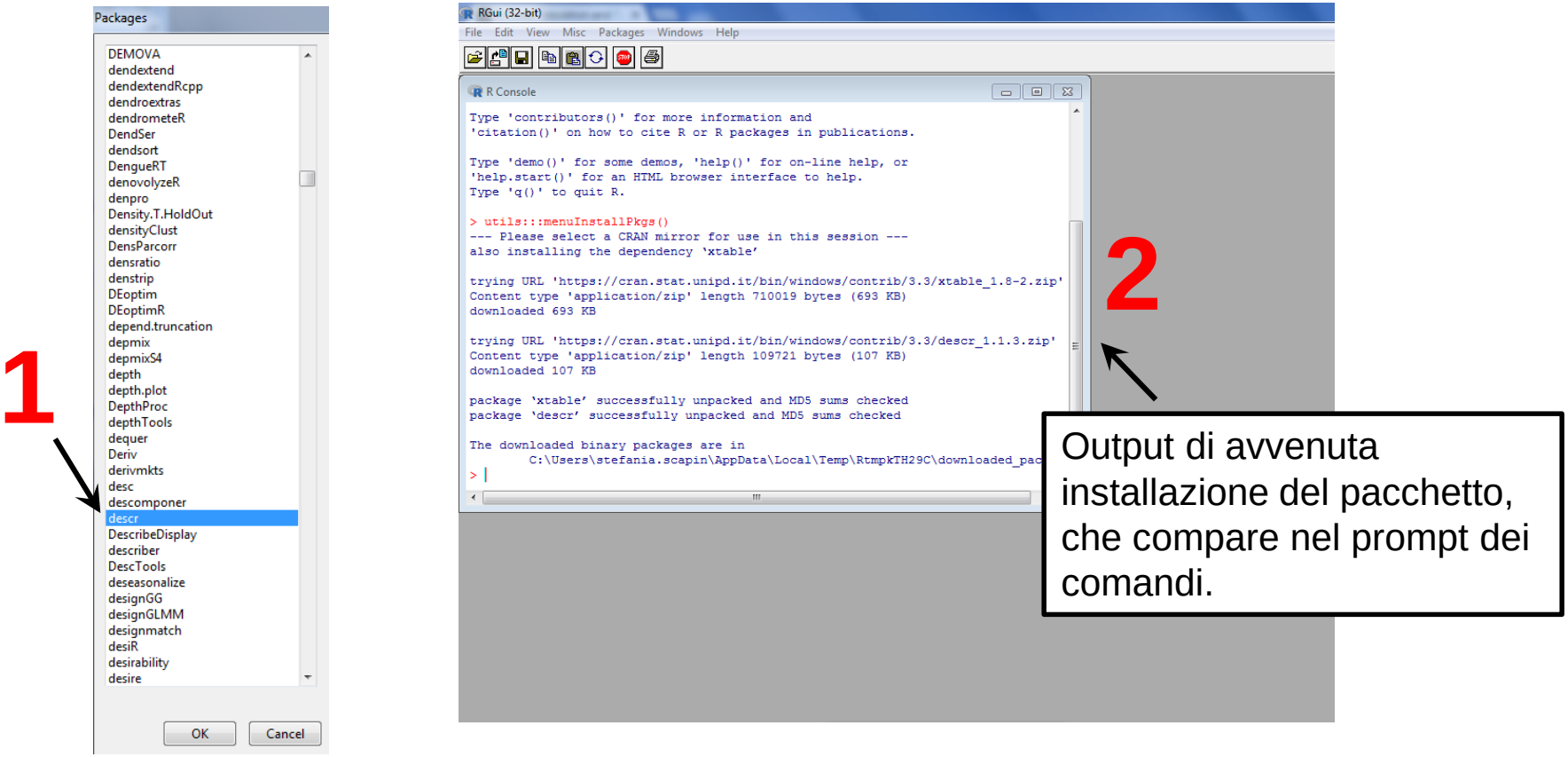
HTTPS Cran mirror, contiene i server di tutto il mondo in cui sono contenuti i pacchetti disponibili – Cliccare su Italy(Padua)

Packages, contiene tutti i nomi dei pacchetti che si possono installare –> selezionare il pacchetto desiderato

Installazione dei pacchetti

Per questa esercitazione, serve installare il pacchetto DESCR.

Seguendo il procedimento della slide precedente, trovare il pacchetto di riferimento e installarlo.



Installazione dei pacchetti

In alternativa si può utilizzare il seguente comando:

```
install.packages("nome_pacchetto")
```

Per esempio per installare il pacchetto *descr:*

```
> install.packages("descr")
Installing package into ‘C:/Users/anna.pozzebon/Documents/R/win-library/3.5’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/descr_1.1.4.zip
'
Content type 'application/zip' length 181251 bytes (177 KB)
downloaded 177 KB

package ‘descr’ successfully unpacked and MD5 sums checked
```

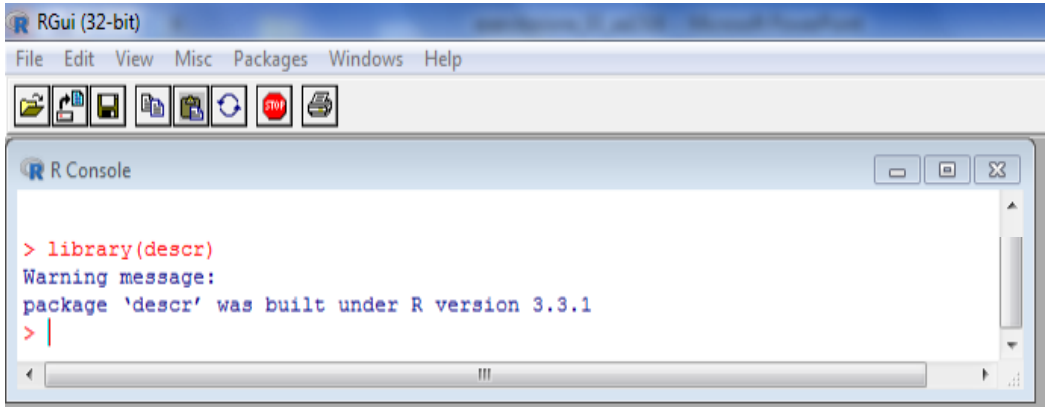
Installazione dei pacchetti

N.B.: Ogni volta che apriamo R, bisogna **richiamare** i pacchetti installati in modo da poterne utilizzare le funzioni contenute

```
library(descr)
```

Richiamo il pacchetto

Se il pacchetto è stato caricato, troverete:



Altrimenti il risultato sarà:

```
> library(glm)
Error in library(glm) : there is no package called 'glm'
>
```

Pacchetto non ancora installato

Installazione dei pacchetti

Ricapitolando:

Se è necessario utilizzare delle funzioni che non sono incluse nell'installazione base del software R, bisogna:

- **Installare** una sola volta il pacchetto contenente le funzioni desiderate
- **Richiamare**, ad ogni apertura di R, i pacchetti precedentemente installati con il comando:

  `library(nome_pacchetto)`

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:

Analisi Univariata: Procedure R

Studio della distribuzione di ogni variabile, singolarmente considerata, all'interno della popolazione

Funzioni R per l'analisi univariata di una variabile:

R	TIPO VARIABILE	FUNZIONE
freq table frequencyBy	Variabili qualitative o quantitative discrete	Distribuzione di frequenze (frequenze assolute, relative)
summary basicStats IQR CV getmode	Variabili quantitative	Calcolo misure di sintesi di tipo univariato

freq – Sintassi generale

La FREQ permette di calcolare le distribuzioni di frequenza univariate per variabili qualitative e quantitative discrete

```
freq(variabile)
```

table – Sintassi generale

Tramite la funzione table possiamo calcolare le <u>frequenze assolute e relative cumulate</u>.
La sintassi è la seguente:

```
cbind(cumsum(table(variabile)),
  cumsum(table(variabile)/length(variabile)))
```

Legenda funzioni:

cbind = funzione che dispone in Colonna i risultati tra parentesi

table = funzione che calcola le frequenze per ogni categoria

cumsum = operatore che svolge la somma cumulata

length = funzione che indica la lunghezza della variabile specificata (ovvero la sua numerosità totale)

freq: Variabile qualitativa

Frequenze <u>assolute e relative</u>: operatore telefonico

```
freq(telefonia$operatore)
```

Frequenze <u>assolute e relative cumulate</u>: operatore telefonico

Codice relativo alla frequenza assoluta cumulata

```
cbind(cumsum(table(telefonia$operatore))
  ,cumsum(table(telefonia$operatore)/length(telefonia$
 operatore)))
```

Codice relativo alla frequenza relativa cumulata

=

Frequenza assoluta cumulata / TOTALE numerosità (236)

Output freq

***Frequenza assoluta (p)*:** consiste nell'associare a ciascuna categoria, o modalità, il numero di volte in cui compare nei dati

***Frequenza relativa percentuale (p/N*100)*:** rapporto tra la frequenza assoluta ed il numero complessivo delle osservazioni effettuate

```
> library(descr)
> telefonia=read.csv('telefonia.csv',header=TRUE)
> freq(telefonia$operatore)
telefonia$operatore
         Frequency Percent
Tim             55  23.305
Tre             12   5.085
Vodafone       154  65.254
Wind            15   6.356
Total          236 100.000
> cbind(cumsum(table(telefonia$operatore)),cumsum(table(telefonia$operatore)/length(telefonia$operatore)))
         [,1]      [,2]
Tim        55 0.2330508
Tre        67 0.2838983
Vodafone  221 0.9364407
Wind      236 1.0000000
```

Frequenze cumulate

freq: Variabile quantitativa discreta

Frequenze assolute e relative: numero medio di giorni alla settimana di utilizzo del telefono fisso

```
freq(telefonia$fisso_g)
```

Frequenze assolute e relative cumulate

Codice relativo alla frequenza assoluta cumulata

```
cbind(cumsum(table(telefonia$fisso_g))
  ,cumsum(table(telefonia$fisso_g)/length(telefonia$
  fisso_g)))
```

Codice relativo alla frequenza relativa cumulata

=

Frequenza assoluta cumulata / TOTALE numerosità

Output freq

```
freq(telefonia$fisso_g)
```

```
telefonia$fisso_g
        Frequency Percent
0              27  11.441
0.5             9   3.814
1              10   4.237
2              19   8.051
3              21   8.898
4              14   5.932
5              19   8.051
6               9   3.814
7             108  45.763
Total         236 100.000
```

```
cbind(cumsum(table(telefonia$fisso_g))
   ,cumsum(table(telefonia$fisso_g)/le
   ngth(telefonia$fisso_g)))
```

```
     [,1]      [,2]
0      27 0.1144068
0.5    36 0.1525424
1      46 0.1949153
2      65 0.2754237
3      86 0.3644068
4     100 0.4237288
5     119 0.5042373
6     128 0.5423729
7     236 1.0000000
```

Fare attenzione al numero di modalità della variabile

freq: Variabile qualitativa con dati missing

Variabile qualitativa:

secondo motivo di utilizzo mezzi di comunicazione

```
freq(telefonia$motivo_utilizzo_2)
```

N.B.: se la variabile sulla quale vogliamo effetturare una distribuzione di frequenza contiene dei **valori mancanti**, **R li tratta come una modalità**

Output freq

Output

```
> freq(telefonia2$motivo_utilizzo_2)
telefonia2$motivo_utilizzo_2
                     Frequency  Percent Valid Percent
Altro                        2   0.8475       0.9434
Famigliari                  40  16.9492      18.8679
Partner                     22   9.3220      10.3774
Piacere/Tempo libero       128  54.2373      60.3774
Studio                      20   8.4746       9.4340
NA's                        24  10.1695
Total                      236 100.0000     100.0000
```

MISSING, *i valori missing vengono considerati come una categoria della variabile qualitativa*

Frequenze percentuali, *calcolate considerando i valori missing come una categoria*

Frequenze percentuali, *calcolate NON considerando i valori missing come una categoria*

freq: Variabile qualitativa con dati missing

Se la variabile sulla quale vogliamo effetturare una distribuzione di frequenza contiene dei **valori mancanti** e **non** vogliamo che R li tratti come una modalità della variabile qualitativa in analisi, dobbiamo scrivere:

```
freq(na.exclude(telefonia$motivo_
  utilizzo_2))
```

Esclude *i valori missing nel calcolo delle frequenze*

Output freq

Output

```
> freq(na.exclude(telefonia2$motivo_utilizzo_2))
na.exclude(telefonia2$motivo_utilizzo_2)
                      Frequency  Percent
Altro                         2   0.9434
Famigliari                   40  18.8679
Partner                      22  10.3774
Piacere/Tempo libero        128  60.3774
Studio                       20   9.4340
Total                       212 100.0000
>
```

Frequenze entro classe - Sintassi

E' possibile ottenere la distribuzione di frequenza di una variabile rispetto ai valori assunti da un'altra variabile categorica, in modo da osservare se la variabile in analisi ha comportamenti differenti in sottopopolazioni

Attenzione!

Non esiste in R una funzione standard per le frequenze entro classe.

E' possibile, quindi, costruire delle funzioni personalizzate che devono essere richiamate una sola volta all'apertura dell'area di lavoro R (come per il richiamo delle librerie).

```
> frequencyBy<-function(dati,variabile_grouping,variabile_analisi,missing){
+ arguments <- as.list(match.call())
+
+ y = eval(arguments$variabile_grouping, dati)
+ un<-unique(y)
+ print(un)
+ arr<-list()
+
+ for(i in 1:length(un)){
+ vet=un[i]
+ mat<-dati[which(y==vet),]
+
+ x=eval(arguments$variabile_analisi, mat)
+ if(missing=="FALSE"){tabella<-freq(x,plot=FALSE)}
+ else {tabella<-freq(na.exclude(x),plot=FALSE)}
+
+ if (i==1) {
+ arr[[i]]<-tabella
+ }
+
+ else {
+ arr[[i]]<-tabella
+
+ }
+
+ }
+
+ names(arr)<-un
+ print(arr)
+ }
```

Comandi da eseguire (invio) per richiamare la funzione
N.B.: questo codice non va assolutamente modificato!

Frequenze entro classe - Sintassi

```
> frequencyBy(telefonia,sesso,operatore,FALSE)
```

Dopo aver eseguito il comando di cui sopra, per calcolare la frequenza entro classe basta scrivere il nome funzione (in questo caso **frequencyBy**) e la variabile su cui si vuole calcolare l'indice (come per le funzioni R viste fin'ora)

Specificare il nome della tabella su cui stiamo lavorando

```
frequencyBy(nome dataset,var classificazione,
variabile analisi, missing)
```

Variabile per cui si vuole la distribuzione di frequenze

Eliminare o meno i missing dalla variabile di analisi.
Se missing=**TRUE** si ottiene la distribuzione di frequenza con I missing, se presenti.
Se missing=**FALSE** si ottiene la distribuzione di frequenze senza missing

Variabile entro cui calcolare le distribuzioni di frequenze della variabile di analisi

Frequenze entro classe - Output

Ottenere la distribuzione di frequenze della variabile operatore entro le classi della variabile sesso

```
frequencyBy(dataset,sesso, operatore, FALSE)
```

sesso=F

```
[1] F M
Levels: F M
$F
x
          Frequency Percent
Tim              27      27
Tre               7       7
Vodafone         63      63
Wind              3       3
Total           100     100
```

Variabile di analisi

Variabile di classificazione

sesso=M

```
$M
x
          Frequency Percent
Tim              28  20.588
Tre               5   3.676
Vodafone         91  66.912
Wind             12   8.824
Total           136 100.000
```

Analisi Univariata: Procedure R

Studio della distribuzione di ogni variabile, singolarmente considerata, all'interno della popolazione

Procedure SAS per l'analisi univariata di una variabile:

R	TIPO VARIABILE	FUNZIONE
freq table frequencyBY	Variabili qualitative o quantitative discrete	Distribuzione di frequenze (frequenze assolute, relative e cumulate)
summary basicStats IQR CV getmode quantile describeBY	Variabili quantitative	Calcolo misure di sintesi di tipo univariato

Analisi Univariata: Misure di Sintesi

Misure di posizione:

Misure di tendenza centrale:

- Media aritmetica
- Mediana
- Moda

Misure di tendenza non centrale:

- Quantili di ordine p (percentili, quartili)

Misure di variabilità/dispersione:

- Campo di variazione
- Differenza interquartile
- Varianza
- Scarto quadratico medio
- Coefficiente di variazione

Misure di forma della distribuzione:

- Skewness
- Kurtosis

summary – Sintassi

La *summary* permette di calcolare misure di posizione per variabili **<u>quantitative</u>**:

- di tendenza centrale (media, mediana)
- di tendenza non centrale (quartili)

```
summary(nome_dataset$nome_variabile)
```

summary – Esempio

Misure di sintesi della variabile quantitativa discreta: numero medio di messaggi inviati al giorno

```
summary(telefonia$num_sms_e)
```

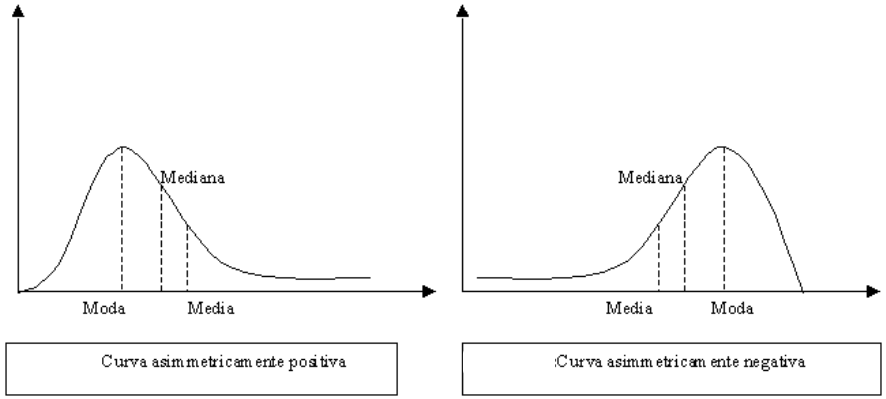
summary - Output

Misure di tendenza centrale

- ***Media aritmetica***: somma dei valori diviso il numero di valori

- ***Mediana***: in una lista ordinata, la mediana è il valore "centrale" (50% sopra, 50% sotto)

```
> summary(telefonia$num_sms_e)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    5.00   10.00   24.31   30.00  100.00
```



summary - Output

Misure di tendenza non centrale

- ***Primo quartile (25%)***: valore per cui ho il 25% dei dati al di sotto e il 75% dei dati sopra questo valore

- ***Terzo quartile (75%)***: valore per cui ho il 75% dei dati al di sotto e il 25% dei dati sopra questo valore

```
> summary(telefonia$num_sms_e)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    5.00   10.00   24.31   30.00  100.00
```

quantile - Sintassi

Misure di tendenza non centrale

- ***Quantili***: il quantile di ordine α è il valore che permette di divider la popolazione in due parti.

Per esmpio il quantile di ordine 25% è il valore per cui il 25% di dati hanno un valore più piccolo del quantile, e il 75% dei dati hanno un valore più grande del quantile.

```
quantile(nome_dataset$nome_variabile,
c(.01,.05,.10,.25,.50,.75,.90,.95,.99) )
```

quantile - Output (1/2)

Quantili della variabile quantitativa discreta: numero medio sms inviati al giorno

```
quantile(telefonia$num_sms_e,
c(.01,.05,.10,.25,.50,.75,.90,.95,.99) )
```

```
> quantile(telefonia$num_sms_e,c(.01,.05,.10,.25,.50,.75,.90,.95,.99))
 1%  5% 10% 25% 50% 75% 90% 95% 99%
  1   2   2   5  10  30  70 100 100
```

1° Quartile (Q1)

Mediana

3° Quartile (Q3)

quantile – Output (2/2)

```
> quantile(telefonia$num_sms_e,c(.01,.05,.10,.25,.50,.75,.90,.95,.99))
 1%  5% 10% 25% 50% 75% 90% 95% 99%
  1   2   2   5  10  30  70 100 100
```

I Quartili dividono la sequenza ordinata dei dati in 4 segmenti contenenti lo stesso numero di valori

- Il primo quartile, Q_1, è il valore per il quale il 25% delle osservazioni sono minori di esso e il 75% sono maggiori
- Q_2 coincide con la mediana (50% sono minori, 50% sono maggiori)
- Il terzo quartile, Q3, è il valore per il quale il 75% delle osservazioni sono minori di esso e il 25% sono maggiori

Interpretazione: in questo esempio, quindi, il 90% della popolazione in analisi ha mandato mediamente al più 70 sms al giorno.

Moda - Sintassi

Attenzione!

Non esiste in R una funzione standard per calcolare la moda.

E' possibile, quindi, costruire delle funzioni personalizzate che devono essere richiamate una sola volta all'apertura dell'area di lavoro R (come per il richiamo delle librerie).

```
>  #####funzione da richiamare per calcolare la moda##########################
>
> getmode <- function(v) {
+ uniqv <- unique(v)
+ tabs<-tabulate(match(v, uniqv))
+ maxtab<-max(tabulate(match(v, uniqv)))
+  uniqv[which(tabs == maxtab)]
+
+ }
>
> #############fine costruzione##############################################
> #calcolo della moda
> getmode(telefonia$num_sms_e)
[1] 10
```

Comandi da eseguire (invio) per richiamare la funzione **N.B.: questo codice non va assolutamente modificato!**

Dopo aver eseguito il comando di cui sopra, per calcolare la moda basta scrivere il nome funzione (in questo caso **getmode**) e la variabile su cui si vuole calcolare l'indice (come per le funzioni R viste fin'ora)

```
getmode(nome_dataset$nome_variabile)
```

Moda – Output (1/2)

Misure di tendenza centrale

- ***Moda***: valore che occorre più frequentemente

Moda della variabile **quantitativa** discreta: numero medio sms inviati al giorno

```
getmode(telefonia$num_sum_e)
```

```
> getmode(telefonia$num_sms_e)
[1] 10
```

N.B.: nel caso in cui una variabile risulti essere bimodale, ovvero ha due modalità con la stessa frequenza massima, vengono riportate entrambe le modalità.

Moda – Output (2/2)

La moda può essere calcolata anche su una **variabile qualitativa**. Restituirà la categoria della variabile con la frequenza assoluta più elevata.

Moda della variabile qualitativa: marca di telefoni più venduta

```
getmode(telefonia$marca)
```

```
> getmode(telefonia$marca)
[1] Nokia
Levels: Altro Lg Motorola Nek Nokia PalmOne Samsung Siemens Sony Ericsson
```

MODA

Tutte le categorie della variabile qualitativa *marca*

basicStats – Sintassi

La ***summary*** è una funzione che permette di calcolare una serie limitata di misure statistiche.

Un'altra funzione più esauriente è, invece, la ***basicStats***
Permette di calcolare indici:

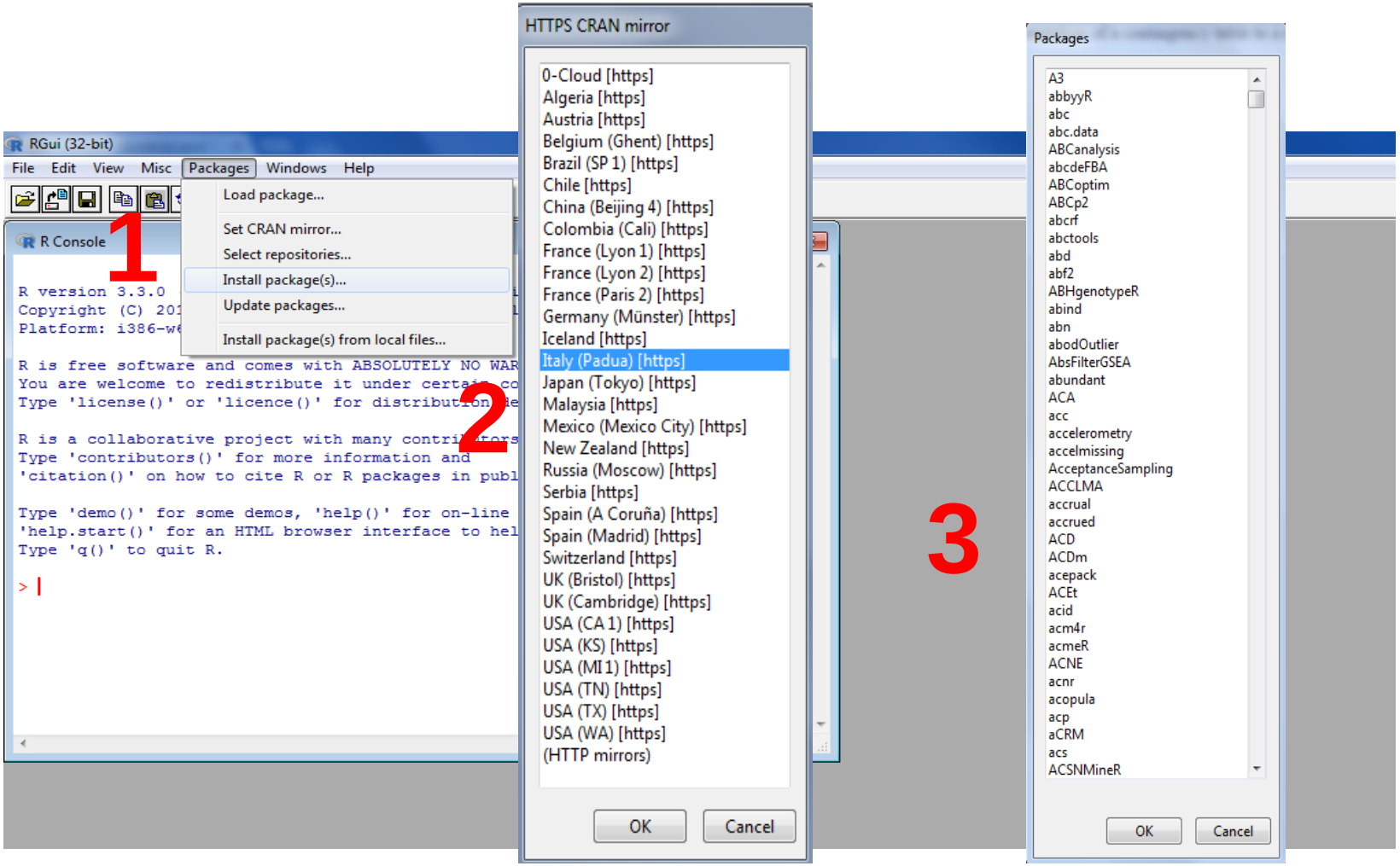
- di posizione
- di variabilità
- di forma della distribuzione

```
basicStats(nome_dataset$nome_variabile)
```

N.B. Per usare questa funzione è necessario scaricare il pacchetto **fBasics**

Installazione pacchetto - fBasics

Riprendiamo la procedura di installazione dei pacchetti:



Installazione pacchetto - fBasics

Riprendiamo la procedura di installazione dei pacchetti:

RGui (32-bit) - [R Console]

File Edit View Misc Packages Windows Help

```
> utils:::menuInstallPkgs()
--- Please select a CRAN mirror for use in this session ---
also installing the dependencies 'timeDate', 'timeSeries', 'gss', 'stabledist'

trying URL 'https://cran.stat.unipd.it/bin/windows/contrib/3.3/timeDate_3012.100.zip'
Content type 'application/zip' length 789672 bytes (771 KB)
downloaded 771 KB

trying URL 'https://cran.stat.unipd.it/bin/windows/contrib/3.3/timeSeries_3022.101.2.zip'
Content type 'application/zip' length 1588518 bytes (1.5 MB)
downloaded 1.5 MB

trying URL 'https://cran.stat.unipd.it/bin/windows/contrib/3.3/gss_2.1-5.zip'
Content type 'application/zip' length 868052 bytes (847 KB)
downloaded 847 KB

trying URL 'https://cran.stat.unipd.it/bin/windows/contrib/3.3/stabledist_0.7-0.zip'
Content type 'application/zip' length 41330 bytes (40 KB)
downloaded 40 KB

trying URL 'https://cran.stat.unipd.it/bin/windows/contrib/3.3/fBasics_3011.87.zip'
Content type 'application/zip' length 1556980 bytes (1.5 MB)
downloaded 1.5 MB

package 'timeDate' successfully unpacked and MD5 sums checked
package 'timeSeries' successfully unpacked and MD5 sums checked
package 'gss' successfully unpacked and MD5 sums checked
package 'stabledist' successfully unpacked and MD5 sums checked
package 'fBasics' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\stefania.scapin\AppData\Local\Temp\RtmpEXEPWc\downloaded_packages
> library(fBasics)
Loading required package: timeDate
Loading required package: timeSeries


Rmetrics Package fBasics
Analysing Markets and calculating Basic Statistics
Copyright (C) 2005-2014 Rmetrics Association Zurich
Educational Software for Financial Engineering and Computational Science
Rmetrics is free software and comes with ABSOLUTELY NO WARRANTY.
https://www.rmetrics.org --- Mail to: info@rmetrics.org
Warning messages:
1: package 'fBasics' was built under R version 3.3.1
2: package 'timeSeries' was built under R version 3.3.1
```

Procedura che indica che il pacchetto fBasics è stato installato

Richiamo il pacchetto nell'area di lavoro

basicStats – Esempio

Misure di sintesi della variabile quantitativa discreta: numero medio sms inviati al giorno

```
basicStats(telefonia$num_sms_e)
```

basicStats – Output

```
> basicStats(telefonia$num_sms_e)
            X..telefonia.num_sms_e
nobs                     236.000000
NAs                        0.000000
Minimum                    0.000000
Maximum                  100.000000
1. Quartile                5.000000
3. Quartile               30.000000
Mean                      24.313559
Median                    10.000000
Sum                     5738.000000
SE Mean                    1.852702
LCL Mean                  20.663532
UCL Mean                  27.963587
Variance                 810.071475
Stdev                     28.461755
Skewness                   1.575958
Kurtosis                   1.349222
```

Misure di posizione

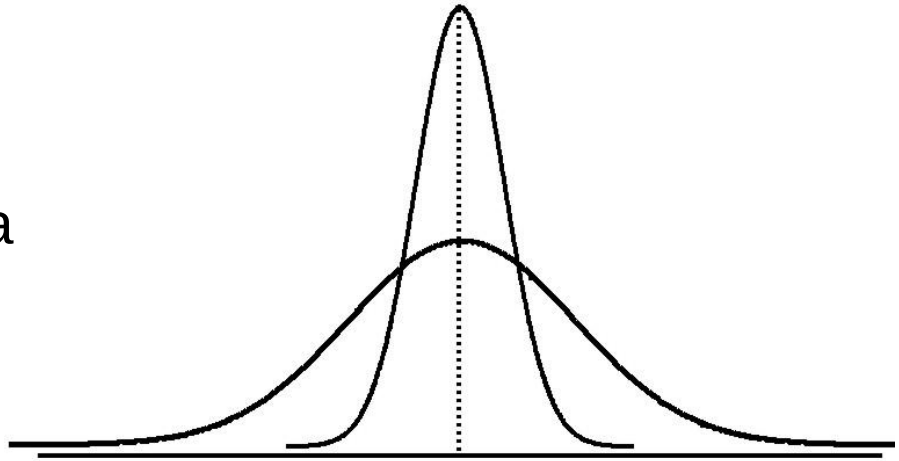
Misure di variabilità e di distribuzione

Misure di Variabilità

Variabilità

- **Campo di Variazione**
- **Differenza Interquartile**
- **Varianza**
- **Scarto Quadratico Medio**
- **Coefficiente di Variazione**

- Le misure di variabilità forniscono informazioni sulla **dispersione** o **variabilità** dei valori.



Stesso centro, diversa variabilità

basicStats – Output

Misure di Variabilità

- ***Varianza*** [Variance]:
  media dei quadrati delle differenze fra ciascuna osservazione e la media

$$\sigma^2 = \frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N}$$

```
> basicStats(telefonia$num_sms_e)
            X..telefonia.num_sms_e
nobs                     236.000000
NAs                        0.000000
Minimum                    0.000000
Maximum                  100.000000
1. Quartile                5.000000
3. Quartile               30.000000
Mean                      24.313559
Median                    10.000000
Sum                     5738.000000
SE Mean                    1.852702
LCL Mean                  20.663532
UCL Mean                  27.963587
Variance                 810.071475
Stdev                     28.461755
Skewness                   1.575958
Kurtosis                   1.349222
```

- ***Scarto Quadratico Medio*** [Std Deviation]:
  mostra la variabilità rispetto alla media (radice quadrata della varianza).
  <u>L'unità di misura è quella delle osservazioni</u>.

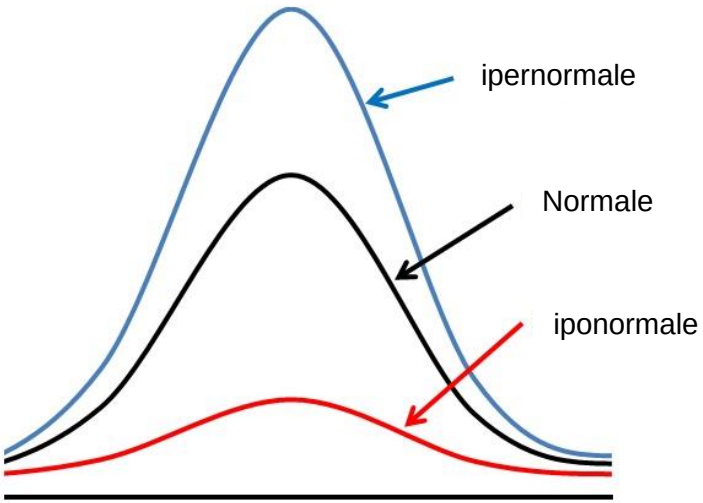
$$\sigma = \sqrt{\sigma^2}$$

basicStats - Output

Misure di Forma della Distribuzione

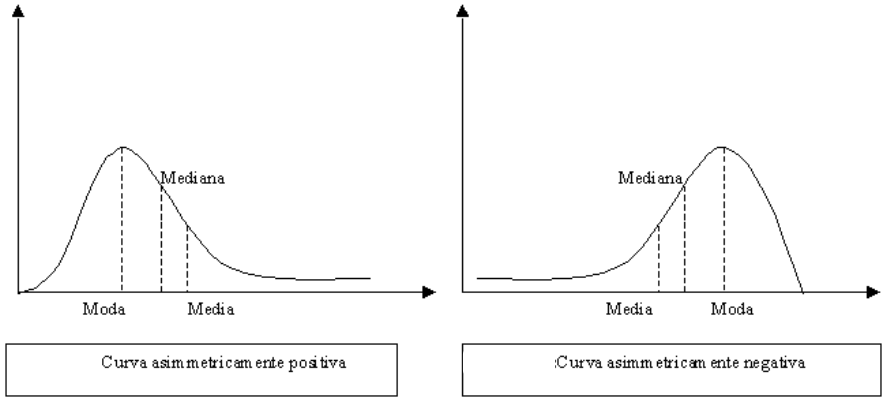
Kurtosis: **indice che permette di verificare se i dati seguono una distribuzione di tipo Normale (simmetrica)**

- β=3 se la distribuzione è "Normale"
- β<3 se la distribuzione è iponormale
- β>3 se la distribuzione è ipernormale



Skewness: **indice che informa circa il grado di simmetria o asimmetria di una distribuzione**

- γ=0 distribuzione simmetrica
- γ<0 asimmetria negativa (mediana>media)
- γ>0 asimmetria positiva (mediana<media)



basicStats - Output

Misure di Forma della Distribuzione

```
> basicStats(telefonia$num_sms_e)
            X..telefonia.num_sms_e
nobs                     236.000000
NAs                        0.000000
Minimum                    0.000000
Maximum                  100.000000
1. Quartile                5.000000
3. Quartile               30.000000
Mean                      24.313559
Median                    10.000000
Sum                     5738.000000
SE Mean                    1.852702
LCL Mean                  20.663532
UCL Mean                  27.963587
Variance                 810.071475
Stdev                     28.461755
Skewness                   1.575958
Kurtosis                   1.349222
```

Kurtosis: **indice che permette di verificare se i dati seguono una distribuzione di tipo Normale (simmetrica)**

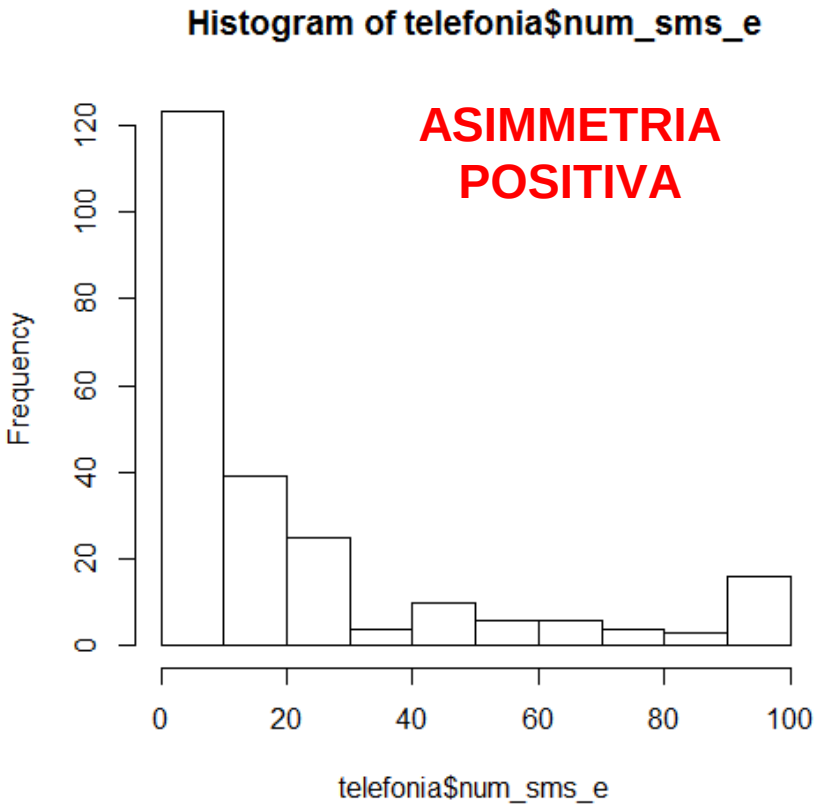
- β=3 se la distribuzione è “Normale”
- β<3 se la distribuzione è iponormale
- β>3 se la distribuzione è ipernormale

Skewness: **indice che informa circa il grado di simmetria o asimmetria di una distribuzione**

- γ=0 distribuzione simmetrica
- γ<0 asimmetria negativa (mediana>media)
- γ>0 asimmetria positiva (mediana<media)

basicStats – Skewness, esempio

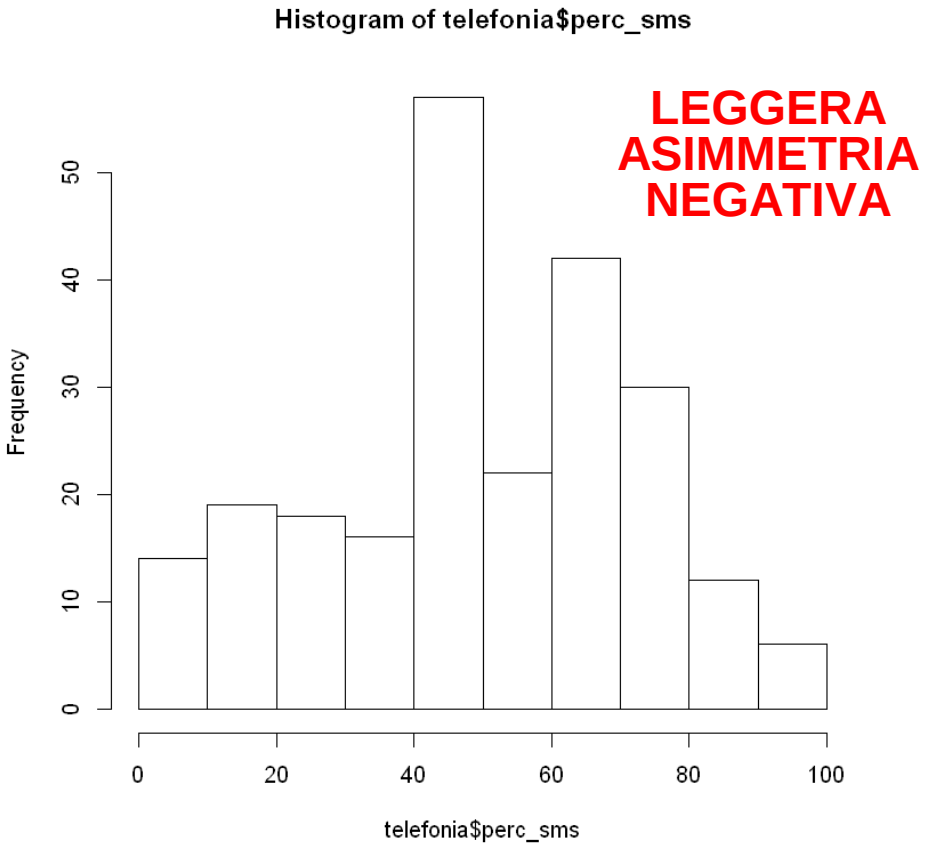
```
> basicStats(telefonia$num_sms_e)
            X..telefonia.num_sms_e
nobs                     236.000000
NAs                        0.000000
Minimum                    0.000000
Maximum                  100.000000
1. Quartile                5.000000
3. Quartile               30.000000
Mean                      24.313559
Median                    10.000000
Sum                     5738.000000
SE Mean                    1.852702
LCL Mean                  20.663532
UCL Mean                  27.963587
Variance                 810.071475
Stdev                     28.461755
Skewness                   1.575958
Kurtosis                   1.349222
```



Skewness: altro esempio

Variabile PERC_SMS del dataset TELEFONIA

```
> basicStats(telefonia$perc_sms)
            X..telefonia.perc_sms
nobs                   236.000000
NAs                      0.000000
Minimum                  0.000000
Maximum                 99.000000
1. Quartile             40.000000
3. Quartile             70.000000
Mean                    53.722458
Median                  50.000000
Sum                  12678.500000
SE Mean                  1.475531
LCL Mean                50.815499
UCL Mean                56.629416
Variance               513.817323
Stdev                   22.667539
Skewness                -0.275620
Kurtosis                -0.686841
```



Skewness più vicina a 0. Distribuzione più simmetrica rispetto all'esempio precedente. Leggera **asimmetria negativa**

Differenza Interquartile (IQR) (1/2)

Le due funzioni *summary* e *basicStats* non restituiscono in output tutte le misure di sintesi di cui necessitiamo. Nelle prossime slides vedremo altre funzioni più specifiche.

Misure di Variabilità

Differenza Interquartile [Interquartile Range]:

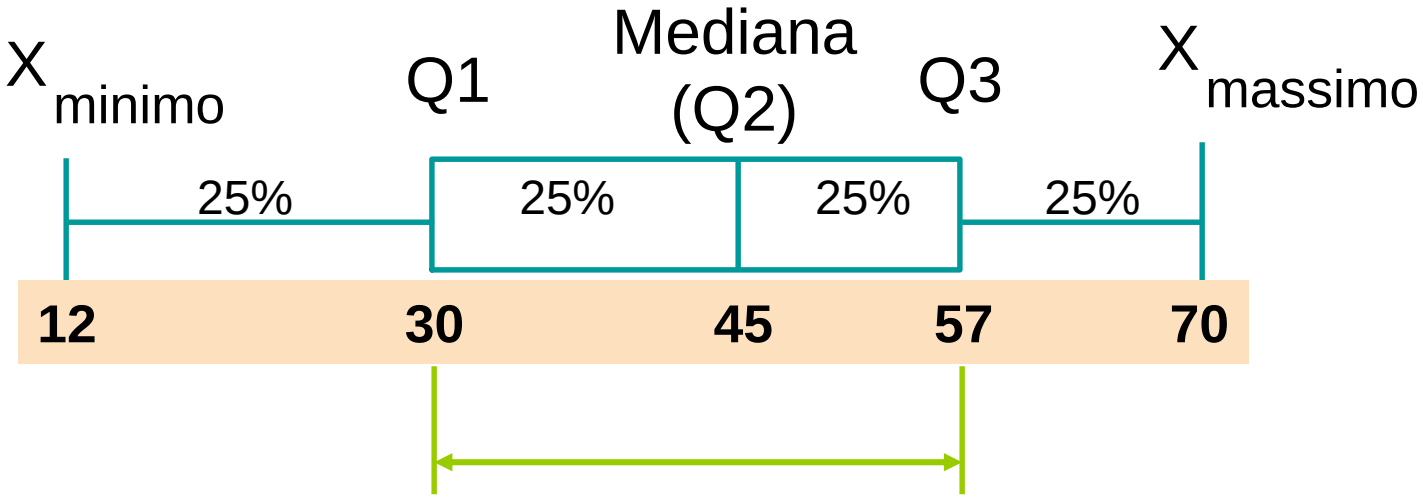
3° quartile – 1° quartile

Lo scarto interquartile è un indice di dispersione, cioè una misura di quanto i valori si allontanino da un valore centrale.

```
IQR(nome_dataset$nome_variabile)
```

Differenza Interquartile (IQR) (2/2)

Misura di Variabilità



Differenza Interquartile

57 – 30 = 27

<u>OUTLIERS</u>:	Q1 - 1,5 * Differenza interquartile
	Q3 + 1,5 * Differenza interquartile

IQR- Output

Scarto interquartile della variabile quantitativa discreta: numero medio sms inviati al giorno

```
IQR(telefonia$num_sms_e)
```

```
> IQR(telefonia$num_sms_e)
[1] 25
```

Campo di Variazione - Sintassi

Misure di Variabilità

- ***Campo di variazione:*** differenza tra il valore massimo e il valore minimo della variabile

```
max(nome_dataset$nome_variabile) -
  min(nome_dataset$nome_variabile)
```

Campo di Variazione - Output

Campo di variazione della variabile quantitativa discreta: numero medio sms inviati al giorno

```
max(telefonia$num_sms_e)-
min(telefonia$num_sms_e)
```

```
> max(telefonia$num_sms_e)-min(telefonia$num_sms_e)
[1] 100
```

Coefficiente di Variazione - Sintassi

Misure di Variabilità

- ***Coeff di variazione*** [Coeff Variation]:
  misura la variabilità relativa
  rispetto alla media (%)

$$CV = \left(\frac{\sigma}{\left| \overline{X} \right|} \right) \cdot 100\%$$

Questo indice si usa per confrontare misure di fenomeni riferite anche ad unità di misura differenti.

```
cv(nome_dataset$nome_variabile)
```

N.B. Per usare questa funzione è necessario installare la libreria **labstatR,** e ricordarsi di richiamare il pacchetto prima di eseguire la funzione

CV- Output

Coefficiente di variazione della variabile quantitativa discreta: numero medio sms inviati al giorno

```
cv(telefonia$num_sms_e)
```

```
> cv(telefonia$num_sms_e)
[1] 1.16813
```

Misure di sintesi (1/2) – Esempio 2

Misure di sintesi della variabile quantitativa continua: numero medio ore utilizzo al giorno del telefono cellulare

```
> basicStats(telefonia$cell_h)
            X..telefonia.cell_h
nobs                  236.000000
NAs                     0.000000
Minimum                 0.250000
Maximum                24.000000
1. Quartile             0.500000
3. Quartile             2.000000
Mean                    2.436441
Median                  1.000000
Sum                   575.000000
SE Mean                 0.253880
LCL Mean                1.936270
UCL Mean                2.936612
Variance               15.211369
Stdev                   3.900175
Skewness                3.308201
Kurtosis               12.576699
```

Misure di sintesi (2/2) – Esempio 2

```
> getmode(telefonia$cell_h)
[1] 0.25
> IQR(telefonia$cell_h)
[1] 1.5
> cv(telefonia$cell_h)
[1] 1.597373
> max(telefonia$cell_h)-min(telefonia$cell_h)
[1] 23.75
> quantile(telefonia$cell_h,c(.01,.05,.10,.25,.50,.75,.90,.95,.99))
   1%    5%   10%   25%   50%   75%   90%   95%   99%
 0.25  0.25  0.25  0.50  1.00  2.00  6.00 10.50 21.20
```

Descrittive entro classe – Sintassi

Statistiche descrittive univariate con variabile di classificazione

```
describeBy(dataset$variabile_quantitativa,
dataset$variabile_classificazione,na.rm=TRUE)
```

TRUE= cancella i valori mancanti dall'analisi
FALSE= non cancella i valori mancanti dall'analisi

N.B. Per usare questa funzione è necessario scaricare e richiamare il pacchetto **psych. Seguire il procedimento illustrato precedentemente**

Descrittive entro classe – Esempi

Misure di sintesi della variabile:
numero medio ore utilizzo al giorno telefono cellulare suddivisa per sesso

```
describeBy(telefonia$cell_h,
telefonia$sesso,na.rm=TRUE)
```

```
> describeBy(telefonia$cell_h, telefonia$sesso,na.rm=TRUE)
group: F
   vars   n mean   sd median trimmed  mad  min max range skew kurtosis   se
X1    1 100 1.45 2.38      1    0.92 1.11 0.25  16 15.75 4.33     21.8 0.24
------------------------------------------------------------
group: M
   vars   n mean   sd median trimmed  mad  min max range skew kurtosis   se
X1    1 136 3.16 4.59      1     2.1 1.11 0.25  24 23.75 2.74     8.22 0.39
>
```

Media oraria dell'utilizzo cellulare per le donne

Massimo numero di ore dell'utilizzo cellulare per gli uomini

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:

Analisi Univariata: GRAFICI

Rappresentazioni grafiche per l'analisi univariata di una variabile:

GRAFICO	TIPO VARIABILE	FUNZIONE
BAR CHART	Variabili qualitative	Bar chart o diagramma a barre (variabili alfanumeriche)
GRAFICO A TORTA	Variabili qualitative	Grafico a torta(variabili alfanumeriche)
HISTOGRAM	Variabili quantitative	Istogramma (variabili numeriche)
BOX PLOT	Variabili quantitative	Rappresentazione grafica di alcune misure di sintesi

BAR CHART – Sintassi (1/2)

Grafico a barre, utilizzato per rappresentare la distribuzione di frequenze di una variabile ordinale.

In questo caso il grafico a barre è uno degli output predefiniti della funzione FREQ vista precedentemente

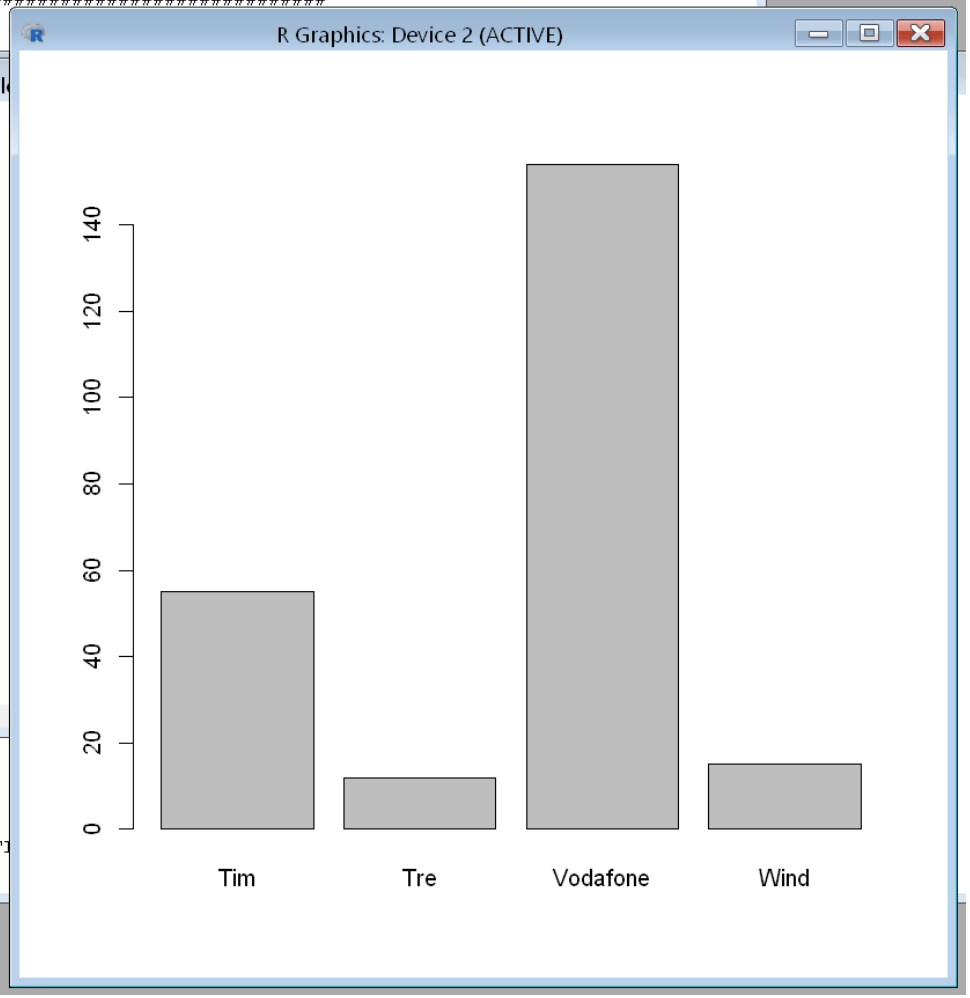
```
freq(nome_dataset$nome_variabile)
```

BAR CHART- Output (2/2)

`freq(telefonia$operatore)`

```
> freq(telefonia$operatore)
telefonia$operatore
         Frequency Percent
Tim             55  23.305
Tre             12   5.085
Vodafone       154  65.254
Wind            15   6.356
Total          236 100.000
>
```

GRAFICO A BARRE



Analisi Univariata: GRAFICI

Rappresentazioni grafiche del modulo SAS INSIGHT per l'analisi univariata di una variabile:

GRAFICO	TIPO VARIABILE	FUNZIONE
BAR CHART	Variabili qualitative	Bar chart o diagramma a barre (variabili alfanumeriche)
GRAFICO A TORTA	Variabili qualitative	Grafico a torta(variabili alfanumeriche)
HISTOGRAM	Variabili quantitative	Istogramma (variabili numeriche)
BOX PLOT	Variabili quantitative	Rappresentazione grafica di alcune misure di sintesi

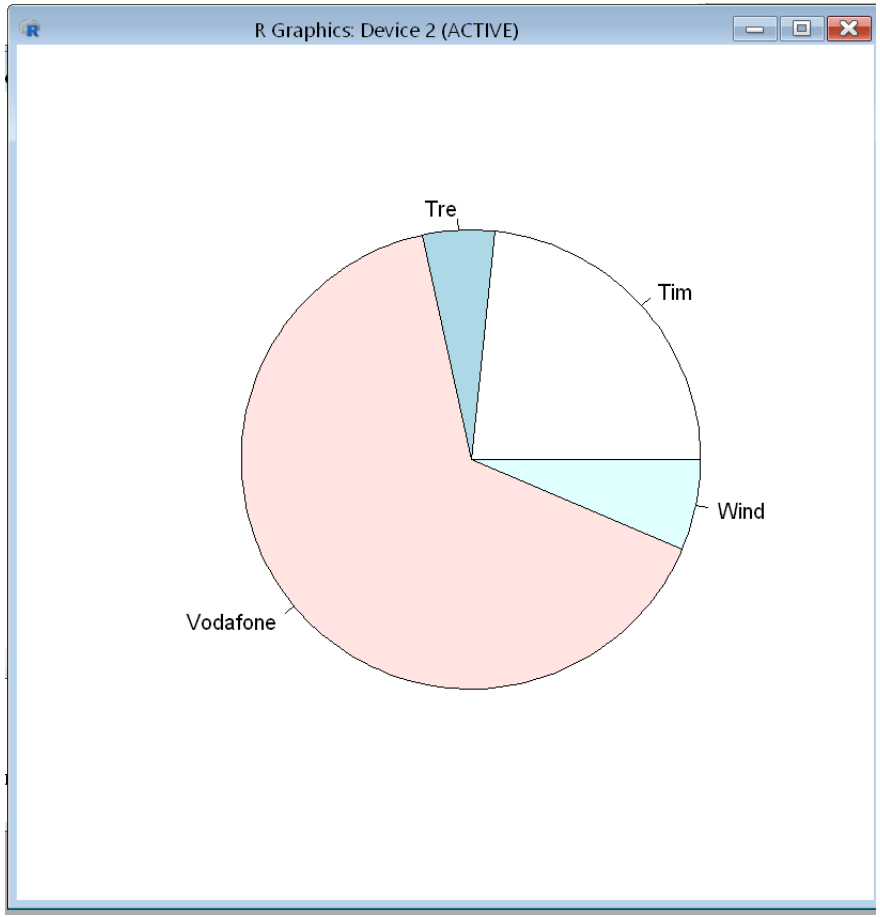
GRAFICO A TORTA – Sintassi (1/2)

Grafico a torta, utilizzato per rappresentare la distribuzione di frequenze di una variabile categorica.

```
pie(table(nome_dataset$nome_variabile))
```

GRAFICO A TORTA - Output (2/2)

```
pie(table(telefonia$operatore))
```



Analisi Univariata: GRAFICI

Rappresentazioni grafiche del modulo SAS INSIGHT per l'analisi univariata di una variabile:

GRAFICO	TIPO VARIABILE	FUNZIONE
BAR CHART	Variabili qualitative	Bar chart o diagramma a barre (variabili alfanumeriche)
GRAFICO A TORTA	Variabili qualitative	Grafico a torta(variabili alfanumeriche)
HISTOGRAM	Variabili quantitative	Istogramma (variabili numeriche)
BOX PLOT	Variabili quantitative	Rappresentazione grafica di alcune misure di sintesi

ISTOGRAMMA – Sintassi (1/2)

L'istogramma permette di visualizzare la forma della distribuzione di una variabile continua.

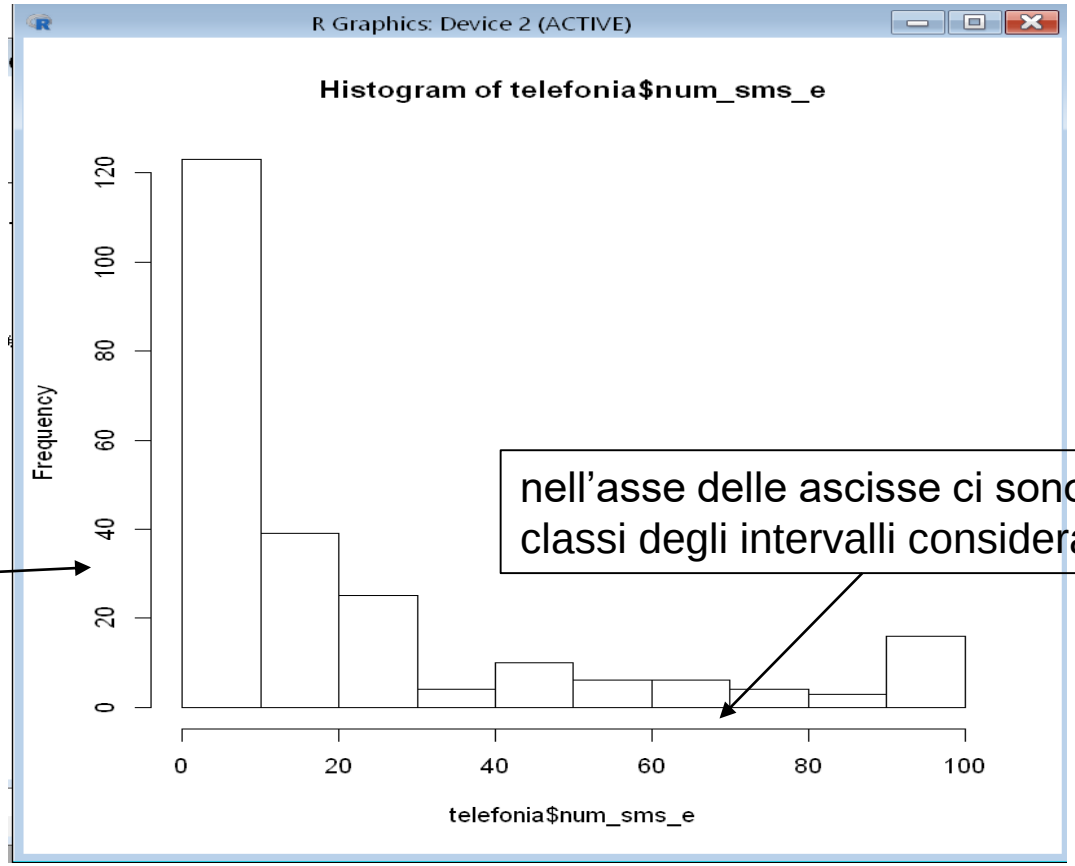
Il comando da eseguire è il seguente

```
hist(nome_dataset$nome_variabile)
```

ISTOGRAMMA – Output (2/2)

```
hist(telefonia$num_sms_e)
```

l'asse delle ordinate rappresenta la densità di frequenza; l'area del rettangolo corrisponde alla frequenza della classe stessa



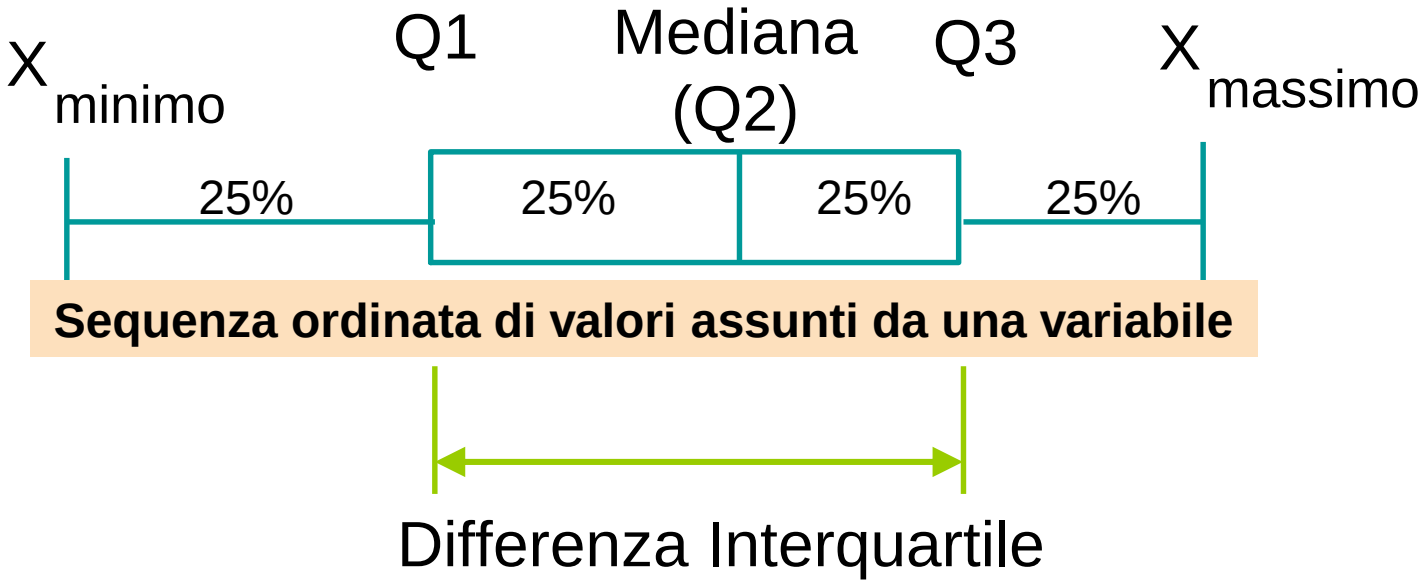
nell'asse delle ascisse ci sono le classi degli intervalli considerati;

Analisi Univariata: GRAFICI

Rappresentazioni grafiche del modulo SAS INSIGHT per l'analisi univariata di una variabile:

GRAFICO	TIPO VARIABILE	FUNZIONE
BAR CHART	Variabili qualitative	Bar chart o diagramma a barre (variabili alfanumeriche)
GRAFICO A TORTA	Variabili qualitative	Grafico a torta(variabili alfanumeriche)
HISTOGRAM	Variabili quantitative	Istogramma (variabili numeriche)
BOX PLOT	Variabili quantitative	Rappresentazione grafica di alcune misure di sintesi

GRAFICI: Box Plot (1/4)



<u>OUTLIERS</u>: Q1 - 1,5 * Differenza interquartile
Q3 + 1,5 * Differenza interquartile

BOXPLOT - Sintassi(2/4)

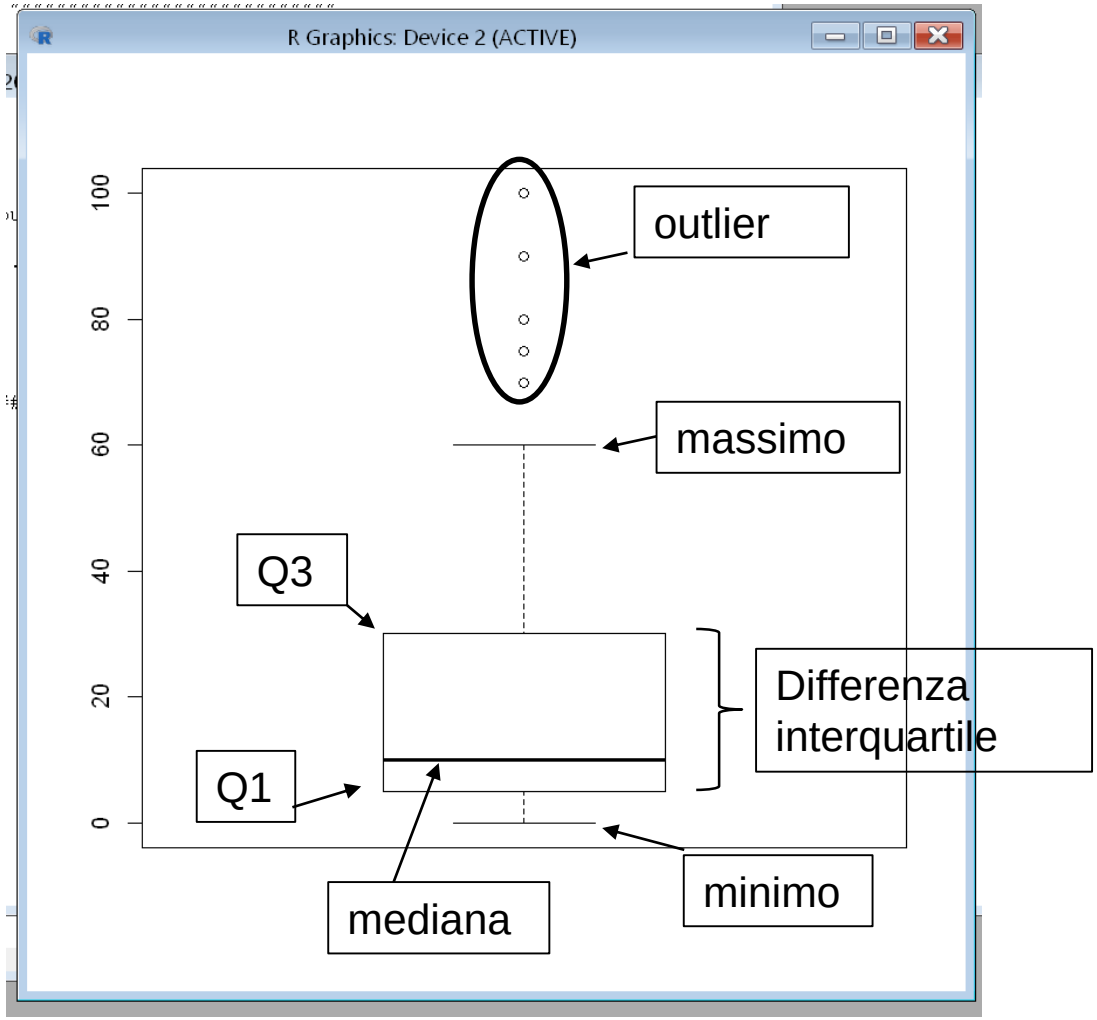
Rappresentazione grafica di alcune misure di sintesi di una variabile quantitativa.
Permette infatti di evidenziare nella distribuzione, i quartili, la media, la differenza interquartile e il campo di variazione

Il comando da eseguire è il seguente

```
boxplot(nome_dataset$nome_variabile)
```

BOXPLOT – Output(3/4)

```
boxplot(telefonia$num_sms_e)
```




```
> basicStats(telefonia$num_sms_e)
            X..telefonia.num_sms_e
nobs                     236.000000
NAs                        0.000000
Minimum                    0.000000
Maximum                  100.000000
1. Quartile                5.000000
3. Quartile               30.000000
Mean                      24.313559
Median                    10.000000
Sum                     5738.000000
SE Mean                    1.852702
LCL Mean                  20.663532
UCL Mean                  27.963587
Variance                 810.071475
Stdev                     28.461755
Skewness                   1.575958
Kurtosis                   1.349222
```

Vengono rappresentati graficamente alcuni indici calcolati precedentemente

BOXPLOT entro classe – Output(4/4)

```
boxplot(dataset$variabile_quantitativa~dataset$
 variabile_categorica)
```

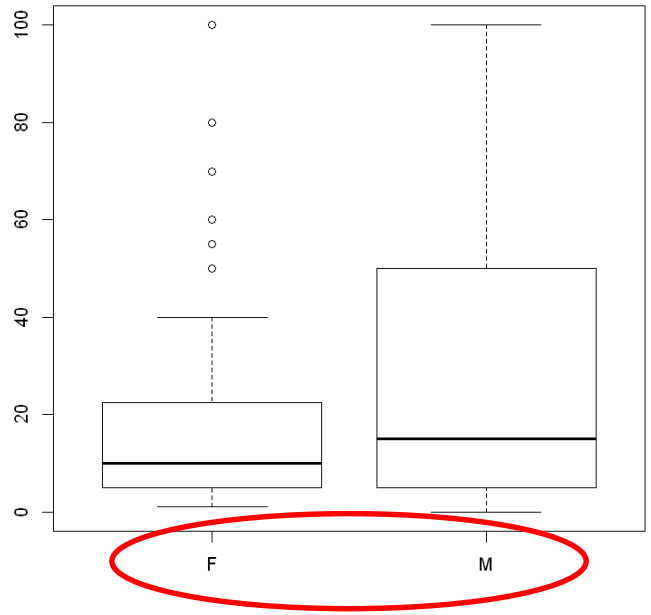
Variabile di classe entro cui rappresentare la distribuzione della variabile quantitativa

Variabile quantitativa da rappresentare

Simbolo tilde, indica una dipendenza tra le due variabili.
Per ottenerlo ALT 126

Distribuzione del numero di sms rispetto al sesso del cliente

```
boxplot(telefonia$num_sms_e~telefonia$sesso)
```



Variabile categorica

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:

Dataset

Il dataset DENTI contiene dati sul consumo di dentifricio (di marca A e di marca B). Le variabili sono:

#	Variable	Type	Label
1	CODCLI	Num	CODICE CLIENTE
2	SESSO	Char	SESSO
3	ETACLASS	Char	CLASSE DI ETA'
4	REGIONE	Char	REGIONE ITALIANA
5	PRESBAMB	Char	PRESENZA BAMBINI (1:SI / 2:NO)
6	TRATTOT	Num	CLIENTE ABITUALE DI DENTIFRICI SI/NO
7	ALTOCON	Num	ALTO CONSUMANTE SI/NO
8	CONSTOT	Num	TOTALE CONSUMO DI DENTIFRICI NEL PERIODO
9	ACQTOT	Num	TOTALE ACQUISTI DI DENTIFRICI NEL PERIODO
10	STOCKTOT	Num	TOTALE ACCUMULO DI DENTIFRICI NEL PERIODO
11	TATTITOT	Num	NUMERO DI CONTATTI PUBBLICITARI TOTALI
12	TRIP	Num	PERIODO OSSERVAZIONE
13	CITYSIZE	Char	DIMENSIONE CITTA' DI RESIDENZA IN CLASSI
14	AREA	Char	AREA GEOGRAFICA
15	ACQ_A	Num	ACQUISTI DI DENTIFRICI DELLA MARCA A NEL PERIODO
16	STOCK_A	Num	ACCUMULO DI DENTIFRICI DELLA MARCA A NEL PERIODO
17	CONS_A	Num	CONSUMO DI DENTIFRICI DELLA MARCA A NEL PERIODO
18	TRAT_A	Num	CLIENTE ABITUALE DI DENTIFRICI DELLA MARCA A SI/NO
19	TATTI_A	Num	NUMERO DI CONTATTI PUBBLICITARI (DENTIFRICI MARCA A)
20	ACQ_B	Num	ACQUISTI DI DENTIFRICI DELLA MARCA B NEL PERIODO
21	STOCK_B	Num	ACCUMULO DI DENTIFRICI DELLA MARCA B NEL PERIODO
22	CONS_B	Num	CONSUMO DI DENTIFRICI DELLA MARCA B NEL PERIODO
23	TRAT_B	Num	CLIENTE ABITUALE DI DENTIFRICI DELLA MARCA B SI/NO
24	TATTI_B	Num	NUMERO DI CONTATTI PUBBLICITARI (DENTIFRICI MARCA B)

Esercizi Analisi univariata

Svolgere i seguenti esercizi utilizzando il dataset DENTI:

1. Allocare la DIRECTORY DI LAVORO (che punta alla cartella che contiene il file DENTI.CSV).

2. Importare in R la tabella DENTI.CSV e salvarla in un oggetto col nome DENTI_NEW.

3. Si può affermare che l'insieme degli intervistati è costituito principalmente da donne?

4. Verificare se i clienti abituali della marca B si distribuiscono in modo differente nelle diverse aree geografiche

5. Verificare se ci sono missing nella variabile ETACLASS

Esercizi Analisi univariata

6. Utilizzare la funzione più opportuna per determinare la modalità con frequenza più alta (moda) delle variabili
   - AREA
   - CONSTOT

7. Determinare l'accumulo medio di dentifrici della marca A

8. Calcolare il quantile al 10% della variabile contatti pubblicitari e interpretarne il valore.

9. Verificare se il consumo medio totale differisce tra uomini e donne

10. Verificare simmetria e normalità della variabile TATTI_A e disegnarne il boxplot