# Analisi Bivariata

*Metodi Quantitativi per Economia, Finanza e Management*

*Esercitazione n°4*

# Lavoro di gruppo

- Attendere la validazione del questionario via mail e procedere alla somministrazione dello stesso

- Argomenti da trattare nel lavoro di gruppo:
  - Analisi Univariata
  - Analisi Bivariata
  - Test Statistici
  - Analisi Fattoriale
  - Regressione Lineare
  - Regressione Logistica

# Lavoro di gruppo – Schema di valutazione

**Topics**

**1. Introduzione**
1.1. Definizione Obiettivi di Ricerca
1.2. Descrizione del Contesto
1.3. Definizione della Popolazione
1.4. Disegno del Campione
1.5. Fieldwork

**2. Analisi Preliminari**
2.1. Controllo Rappresentatività del Campione
2.2. Analisi Univariate
2.3. Analisi Connessione
2.6. Analisi Correlazione
2.7. ANOVA

**3. Analisi Fattoriale**
3.1. Scelta Numero dei fattori
3.2. Interpretazione dei fattori

**4. Regressione Lineare**
4.1. Definizione obiettivo di analisi
4.2. Scelta variabili di input
4.3. Valutazione bontà del modello
4.4. Analisi Multicollinearità
4.5. Interpretazione del modello

**5. Regressione Logistica**
5.1. Definizione obiettivo di analisi
5.2. Scelta variabili di input
5.3. Valutazione bontà del modello
5.4. Analisi Multicollinearità
5.5. Interpretazione del modello

**6. Conclusioni**

**7. Layout**

# 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_1617\esercitazione4*** e copiare il contenuto nella cartella C:\corso

- Aprire il programma R(Start → All Programs → R→ R 3.3.1)

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

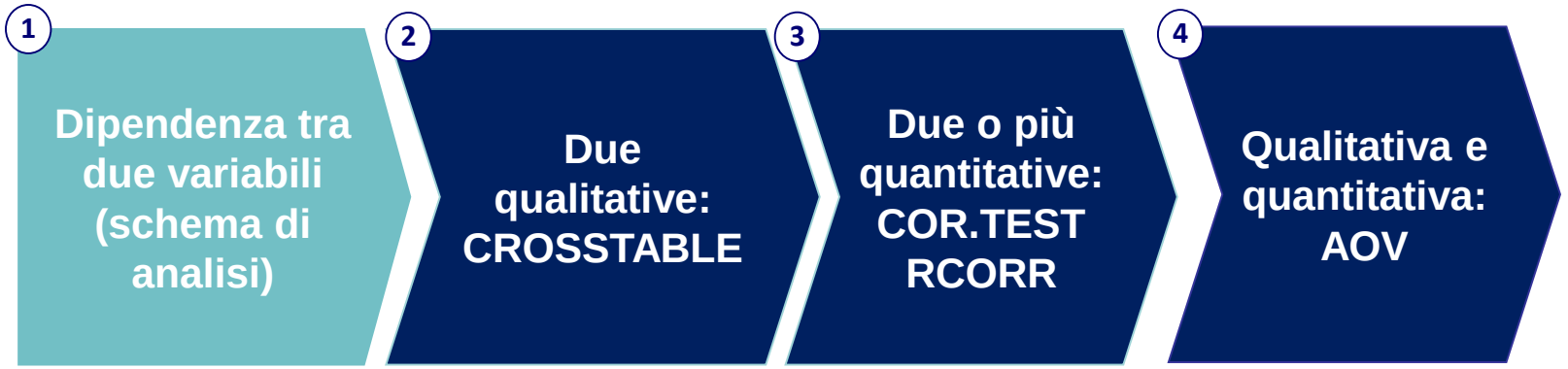
```
setwd('C:/Corso')
```

- 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**:



# Analisi Bivariata

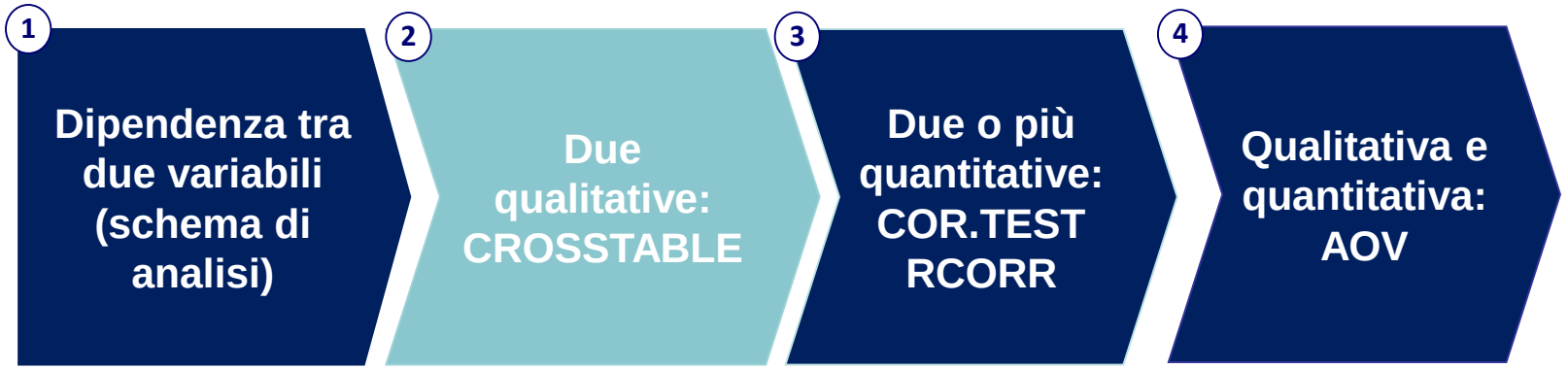
*Studio della distribuzione di due variabili congiuntamente considerate e delle relazioni esistenti tra esse*

**OBIETTIVO:**
studiare la relazione di dipendenza/indipendenza tra due variabili.
L'analisi d'indipendenza dipende dalla natura delle variabili:

| | | |
|---|---|---|
| **Due Variabili Qualitative** | Indipendenza Statistica (indici Chi Quadro, Cramer V) | CROSSTABLE |
| **Due o più Variabili Quantitative** | Indipendenza Lineare (indice: coeff. di correlazione lineare) | COR.TEST RCORR |
| **Una Qualitative e Una Quantitativa continua** | Indipendenza in media (indice: eta-quadro) | AOV |

# Metodi Quantitativi per Economia, Finanza e Management

**Obiettivi di questa esercitazione**:



# Riepilogo teorico (1/2)

**X e Y due variabili qualitative/quantitative discrete**

***Tabelle di Contingenza:***

tabelle a doppia entrata; i valori riportati all'interno della tabella sono le frequenze congiunte assolute (numero di osservazioni per ogni combinazione di modalità di X e Y).

| Colore degli occhi\Colore dei Capelli | $Biondi$ | $NonBiondi$ | $Totale$ |
|---|---|---|---|
| $Chiari$ | 21 | 19 | 40 |
| $NonChiari$ | 9 | 51 | 60 |
| $Totale$ | 30 | 70 | 100 |

**NB**: come vedremo R riporta nell'output anche le distribuzioni marginali (somme per riga e per colonna) e le frequenze relative congiunte (frequenza assoluta congiunta / numero di osservazioni totali)

# Riepilogo teorico (2/2)

***Indipendenza Statistica:***

se al variare di X le distribuzioni subordinate ($Y|X= x_i$) sono tutte uguali tra loro, si può concludere che la distribuzione di Y non dipende da X. Nel caso di indipendenza statistica, la frequenza relativa congiunta è pari al prodotto delle marginali corrispondenti

$$P(x_i,y_j)=P_x(x_i)P_y(y_j)$$

***Indici di connessione:***

- *$\chi^2$ (chi-quadrato)* assume valore nullo se i fenomeni X e Y sono indipendenti. Tende a crescere, al crescere del numero di osservazioni.
- *Cramer V:* basato sul $\chi^2$, è un indice relativo (non risente del numero di osservazioni). Assume valori compresi tra 0 e 1: 0 nel caso di indipendenza statistica, e tende a crescere all'aumentare del grado di dipendenza delle variabili considerate.

# CrossTable - Descrizione

La CROSSTABLE permette di

1. Creare tabelle di contingenza a due o più dimensioni per variabili qualitative e quantitative discrete

2. Calcolare indici di dipendenza relativi a tabelle di contingenza (tra cui chi-quadrato e Cramer V)

# CrossTable – Sintassi generale

Distribuzione di frequenza bivariata (tabelle di contingenza)

```
CrossTable(nome_dataset$nome_variabile1
  , nome_dataset$nome_variabile2,
  prop.chisq=FALSE )
```

È un'opzione che inseriremo sempre

**N.B**. Per usare questa funzione è necessario richiamare la libreria **descr,** scaricata nella lezione 3.

# CrossTable – Esempio 1

**Variabili qualitative**: sesso e operatore telefonico

```
CrossTable(telefonia$sesso,
  telefonia$operatore,
  prop.chisq=FALSE)
```



# Output CrossTable - Esempio 1

```
> CrossTable(telefonia$sesso,telefonia$operatore, prop.chisq=FALSE)
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

=============================================================
                   telefonia$operatore
telefonia$sesso      Tim     Tre   Vodafone    Wind   Total
-------------------------------------------------------------
F                     27       7         63       3     100
                   0.270   0.070      0.630   0.030   0.424
                   0.491   0.583      0.409   0.200
                   0.114   0.030      0.267   0.013
-------------------------------------------------------------
M                     28       5         91      12     136
                   0.206   0.037      0.669   0.088   0.576
                   0.509   0.417      0.591   0.800
                   0.119   0.021      0.386   0.051
-------------------------------------------------------------
Total                 55      12        154      15     236
                   0.233   0.051      0.653   0.064
=============================================================
```

Distribuzioni marginali: frequenze marginali assolute e relative

Frequenze congiunte assolute

Frequenze congiunte relative

Frequenze subordinate di riga e colonna

# Output CrossTable - Esempio 1

```
> CrossTable(telefonia$sesso,telefonia$operatore, prop.chisq=FALSE)
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

=============================================================
                 telefonia$operatore
telefonia$sesso      Tim     Tre   Vodafone    Wind   Total
-------------------------------------------------------------
                      27       7         63       3     100
                   0.270   0.070      0.630   0.030   0.424
                   0.491   0.583      0.409   0.200
                   0.114   0.030      0.267   0.013
-------------------------------------------------------------
M                     28       5         91      12     136
                   0.206   0.037      0.669   0.088   0.576
                   0.509   0.417      0.591   0.800
                   0.119   0.021
-------------------------------------------------------------
Total                 55      12        154      15     236
                   0.233   0.051      0.653   0.064
=============================================================
```

freq. marginale assoluta=28+5+91+12

freq. subordinate:

di riga=27/100

di col=27/55

freq. marginale relativa=(28+5+91+12)/236

freq. congiunta relativa = (28/236)

# Output CrossTable - Esempio 1

```
> CrossTable(telefonia$sesso,telefonia$operatore, prop.chisq=FALSE)
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

***Indipendenza Statistica:***

se al variare di X le distribuzioni subordinate $(Y|X= x_i)$ sono tutte uguali tra loro, si può concludere che la distribuzione di Y non dipende da X. Nel caso di indipendenza statistica, la frequenza relativa congiunta è pari al prodotto delle marginali corrispondenti

$$P(x_i,y_j)=P_x(x_i)P_y(y_j)$$

| | telefonia$operatore | | | | |
|---|---|---|---|---|---|
| telefonia$sesso | Tim | Tre | Vodafone | Wind | Total |
| F | 27 | 7 | 63 | 3 | 100 |
| | 0.270 | 0.070 | 0.630 | 0.030 | 0.424 |
| | 0.491 | 0.583 | 0.409 | 0.200 | |
| | 0.114 | 0.030 | 0.267 | 0.013 | |
| M | 28 | 5 | 91 | 12 | 136 |
| | 0.206 | 0.037 | 0.669 | 0.088 | 0.576 |
| | 0.509 | 0.417 | 0.591 | 0.800 | |
| | 0.119 | 0.021 | 0.386 | 0.051 | |
| Total | 55 | 12 | 154 | 15 | 236 |
| | 0.233 | 0.051 | 0.653 | 0.064 | |

Frequenze subordinate

# CrossTable - Esempio 2

C’è indipendenza statistica tra le variabili sesso del rispondente (SESSO) e possesso del computer (COMPUTER)?

```
CrossTable(telefonia$sesso,
  telefonia$computer,
  prop.chisq=FALSE)
```

# CrossTable– Esempio 2

Cell Contents

| |
|---|
| N |
| N / Row Total |
| N / Col Total |
| N / Table Total |

| telefonia$sesso | telefonia$computer 0 | 1 | Total |
|---|---|---|---|
| F | 16 | 84 | 100 |
| | 0.160 | 0.840 | 0.424 |
| | 0.286 | 0.467 | |
| | 0.068 | 0.356 | |
| M | 40 | 96 | 136 |
| | 0.294 | 0.706 | 0.576 |
| | 0.714 | 0.533 | |
| | 0.169 | 0.407 | |
| Total | 56 | 180 | 236 |
| | 0.237 | 0.763 | |

**Da cosa possiamo dedurre la presenza di dipendenza/ indipendenza tra le due variabili?**

Le variabili sono indipendenti se la distribuzione della variabile “possesso computer” non è influenzata dal sesso…

... Ovvero la distribuzione di chi possiede il computer da chi non lo possiede non varia tra maschi e femmine e corrisponde alla distribuzione marginale della variabile computer

# CrossTable – Esempio 2

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

=========================================
                 telefonia$computer
telefonia$sesso          0         1   Total
-----------------------------------------
F                       16        84     100
                     0.160     0.840   0.424
                     0.286     0.467
                     0.068     0.356
-----------------------------------------
M                       40        96     136
                     0.294     0.706   0.576
                     0.714     0.533
                     0.169     0.407
-----------------------------------------
Total                   56       180     236
                     0.237     0.763
=========================================
```

**Femmine:**

- **16% computer=0**
- **84% computer=1**

**Maschi:**

- **29.4% computer=0**
- **70.6% computer=1**

**Le distribuzioni sono diverse, ci fa pensare alla presenza di dipendenza tra le due variabili!**

# CrossTable – Esempio 2

NB: **la relazione di dipendenza è simmetrica**. Anche analizzando la dipendenza del sesso dalla variabile computer osserviamo un'influenza

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

==========================================
                    telefonia$computer
telefonia$sesso          0          1   Total
------------------------------------------
F                       16         84     100
                     0.160      0.840   0.424
                     0.286      0.467
                     0.068      0.356
------------------------------------------
M                       40         96     136
                     0.294      0.706   0.576
                     0.714      0.533
                     0.169      0.407
------------------------------------------
Total                   56        180     236
                     0.237      0.763
==========================================
```

**Computer=0:**

- **28.6% F**
- **71.4% M**

**Computer=1:**

- **46.7% F**
- **53.3% M**

**Per quantificare il grado di connessione tra le due variabili calcoliamo gli indici di connessione**

# CrossTable - Descrizione

La CROSSTABLE permette di

1. Creare tabelle di contingenza a due o più dimensioni per variabili qualitative e quantitative discrete

2. Calcolare indici di dipendenza relativi a tabelle di contingenza (tra cui Chi-quadrato e Cramer V)

# Chi quadrato – Sintassi generale

Calcolo dell'indice Chi-quadro

```
CrossTable(nome_dataset$variabile1,
  nome_dataset$variabile2,
  prop.chisq=FALSE, options)
```

OPTIONS:

- chisq=TRUE = **calcola l'indice chi-quadro**



# Esempio n°1- Indice Chi-Quadro

C'è indipendenza statistica tra le variabili sesso del rispondente (SESSO) e possesso del computer (COMPUTER)?

```
CrossTable(telefonia$sesso,
  telefonia$computer,
  prop.chisq=FALSE, chisq=TRUE)
```

# Esempio n°1- Indice Chi-Quadro

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

=========================================
                   telefonia$computer
telefonia$sesso         0       1   Total
-----------------------------------------
F                      16      84     100
                    0.160   0.840   0.424
                    0.286   0.467
                    0.068   0.356
-----------------------------------------
M                      40      96     136
                    0.294   0.706   0.576
                    0.714   0.533
                    0.169   0.407
-----------------------------------------
Total                  56     180     236
                    0.237   0.763
=========================================

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 = 5.727462      d.f. = 1      p = 0.0167

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 = 5.010379      d.f. = 1      p = 0.0252
```

Come valutiamo la presenza di indipendenza?

→ **Test d'ipotesi (PROSSIMA LEZIONE)**

# Indice di Cramer V – sintassi generale

Calcolo dell'indice di Cramer V:

```
CramerV(nome_dataset$variabile1,nome_dataset$variabile2)
```

**N.B**. Per calcolare l'indice di Cramer V è necessario scaricare il pacchetto ***DescTools***

e ricordarsi di richiamarlo (*library(DescTools)*)

```
> library(DescTools)
Warning message:
package 'DescTools' was built under R version 3.3.1
```



# Esempio n°1- Indice di Cramer V

C'è indipendenza statistica tra le variabili sesso del rispondente (SESSO) e possesso del computer (COMPUTER)?

```
CramerV(telefonia$sesso,
  telefonia$computer)
```

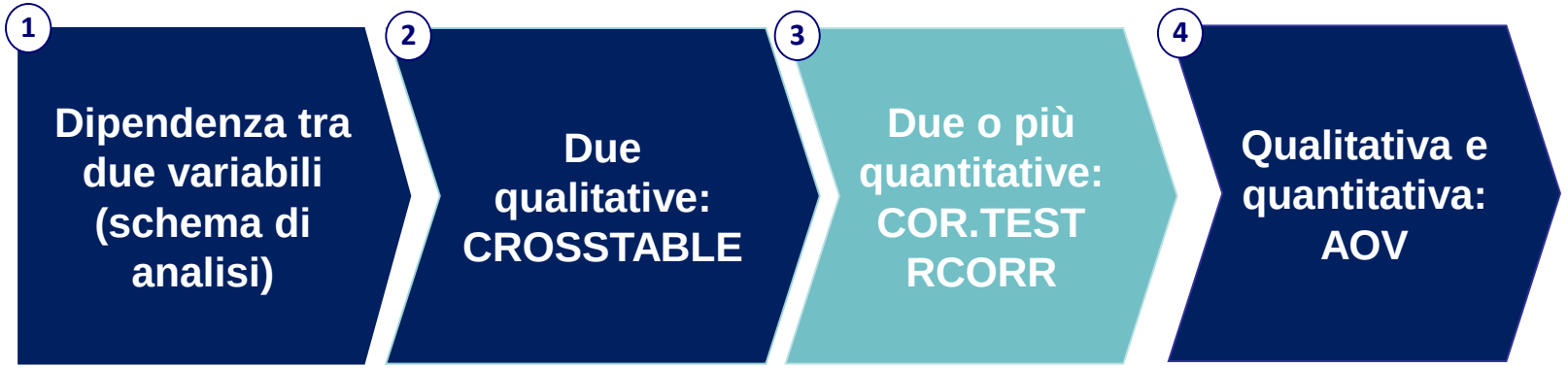
```
> CramerV(telefonia$sesso,telefonia$computer)
[1] 0.1557848
```

Come valutiamo la presenza di indipendenza?

→ **Test d'ipotesi (PROSSIMA LEZIONE)**

# Metodi Quantitativi per Economia, Finanza e Management

**Obiettivi di questa esercitazione**:



# Riepilogo teorico

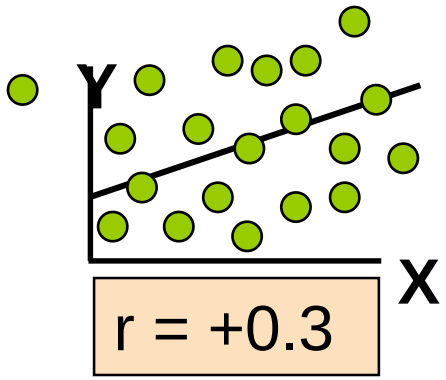
**X e Y due variabili quantitative**

Indaghiamo la presenza di una relazione lineare tra le due variabili

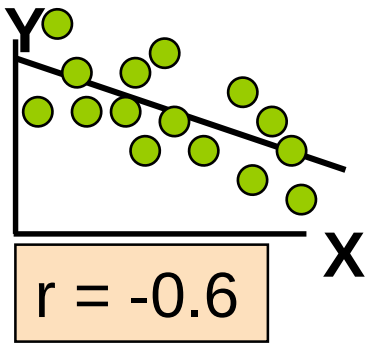
*Coefficiente di correlazione lineare* ***ρ(X,Y)*** : $$\rho = \text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

ρ = 0 → non c'è relazione lineare tra X e Y

ρ > 0 → relazione lineare positiva tra X e Y




ρ < 0 → relazione lineare negativa tra X e Y



# Correlazione tra due variabili
## cor.test - Descrizione

La funzione cor.test permette di

- calcolare la correlazione tra due variabili quantitative

```
cor.test(nome_dataset$variabile1,
  nome_dataset$variabile2)
```

# cor.test - Esempio

Correlazione tra il numero medio di ore di utilizzo del telefono cellulare e del fisso al giorno

```
cor.test(telefonia$cell_h,
  telefonia$fisso_h)
```



# Output cor.test - Esempio

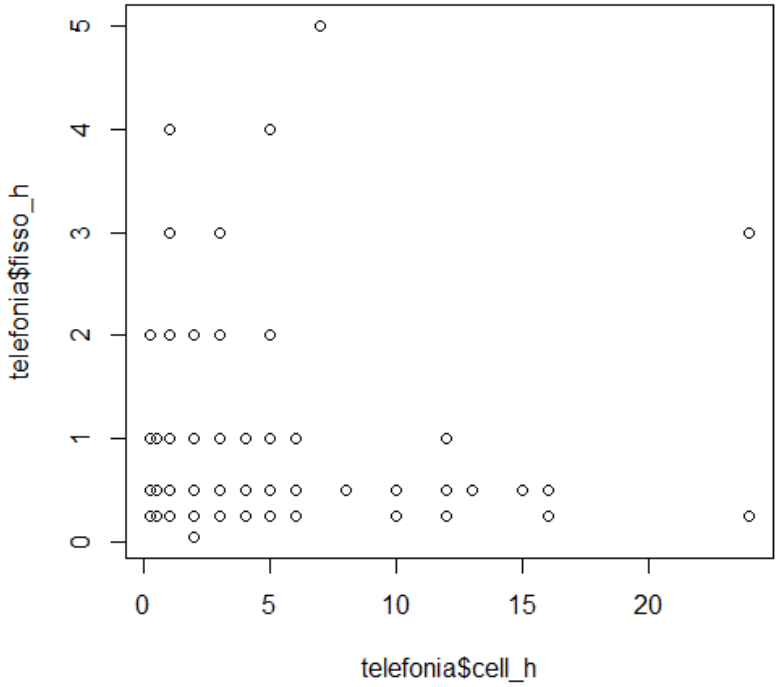
```
> cor.test(telefonia$cell_h, telefonia$fisso_h)

        Pearson's product-moment correlation

data:  telefonia$cell_h and telefonia$fisso_h
t = 3.6117, df = 206, p-value = 0.0003821
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1117014 0.3678627
sample estimates:
      cor
0.2440342
```

*Coefficiente di correlazione lineare ρ(X,Y):*
presenza di relazione lineare positiva

# Output cor.test - Esempio

Se vogliamo visualizzare la correlazione tramite un grafico, possiamo fare un *plot* della distribuzione delle due variabili in essere.

```
plot(telefonia$cell_h,
   telefonia$fisso_h)
```



# Correlazione tra più variabili
## rcorr - Descrizione

La funzione ***rcorr*** permette di calcolare la correlazione tra più di due variabili quantitative, creando così una matrice di correlazione. La diagonale di tale matrice avrà sempre correlazione 1.

```
rcorr(as.matrix(nome_dataset_new))
```

***Per svolgere questa funzione:***

- E' necessario creare un **subset** contenente solo le variabili di interesse su cui applicare l'analisi di correlazione.
- È necessario scaricare il pacchetto **Hmisc**

# Correlazione tra più variabili
## rcorr - Descrizione

L' output della funzione **rcorr** è una lista di elementi di seguito descritti:

- r : è la matrice di correlazione
- n : è la matrice che contiene il numero di osservazioni per ogni coppia di variabile analizzata
- p : p-values corrispondenti al livello di significatività delle osservazioni

# rcorr – Creazione di un subset

## Creazione di un subset

Per creare un nuovo dataset con le solo variabili di interesse, la sintassi è la seguente:

```
Nome_dataset_new=nome_dataset[ ,c("var1",
  "var2", "var3", "varN", …)]
```

# rcorr - Esempio

Vogliamo calcolare la correlazione tra le seguenti variabili:

- durata media delle chiamate effettuate `[durata_chiamate_e]` e:
- durata media delle chiamate ricevute `[durata_chiamate_r]`
- numero medio di ore di utilizzo del telefono cellulare al giorno `[cell_h]`
- numero medio di ore di utilizzo del telefono fisso al giorno `[fisso_h]`

# rcorr - Esempio

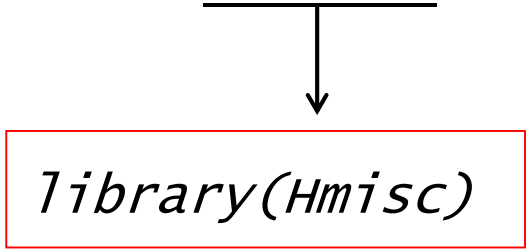
## Creazione di un subset

*tel=telefonia[,c("durata_chiamate_r", "durata_chiamate_e", "fisso_h","cell_h")]*

```
> tel=telefonia[,c("durata_chiamate_r", "durata_chiamate_e", "fisso_h","cell_h")]
> head(tel)
  durata_chiamate_r durata_chiamate_e fisso_h cell_h
1                 4                 1    0.50   1.00
2                 7                 7    0.25   0.25
3                 3                 2    2.00   2.00
4                10                 5    0.50   2.00
5                30                30    2.00   2.00
6                10                10    0.50   3.00
>
```

# rcorr– Installazione pacchetto

Installare il pacchetto Hmisc è richiamarlo.




```
> library(Hmisc)
Loading required package: lattice
Loading required package: survival
Loading required package: Formula
Loading required package: ggplot2

Attaching package: 'Hmisc'

The following objects are masked from 'package:base':

    format.pval, round.POSIXt, trunc.POSIXt, units

Warning messages:
1: package 'Hmisc' was built under R version 3.3.1
2: package 'ggplot2' was built under R version 3.3.1
```

# rcorr - Esempio

## Correlazione tra più variabili

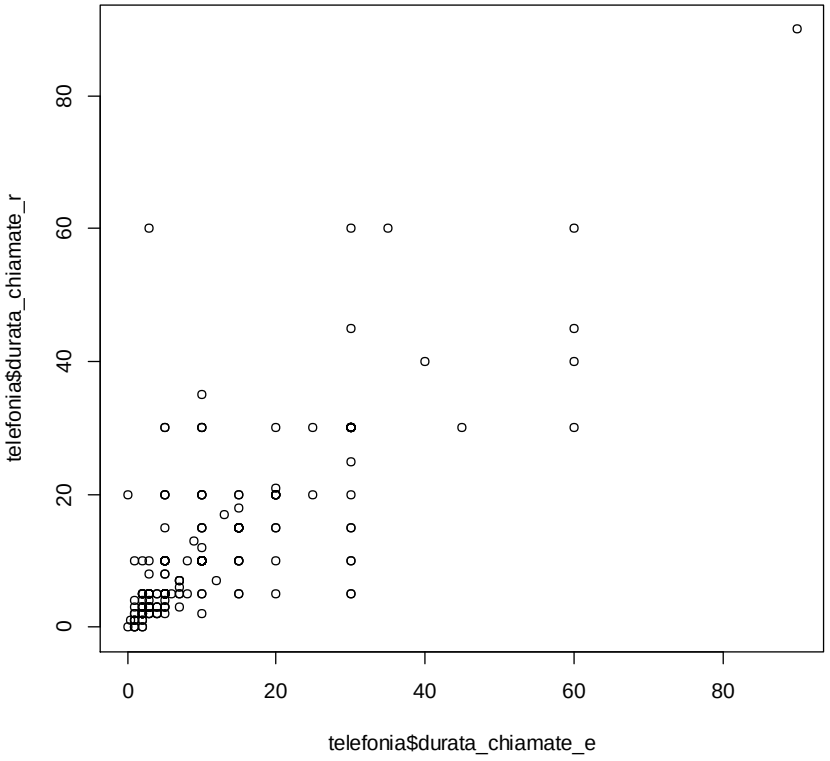
*rcorr(as.matrix(tel))*

```
> rcorr(as.matrix(tel))
                  durata_chiamate_r durata_chiamate_e fisso_h cell_h
durata_chiamate_r              1.00              0.79    0.28   0.32
durata_chiamate_e              0.79              1.00    0.25   0.23
fisso_h                        0.28              0.25    1.00   0.24
cell_h                         0.32              0.23    0.24   1.00

n
                  durata_chiamate_r durata_chiamate_e fisso_h cell_h
durata_chiamate_r               236               236     208    236
durata_chiamate_e               236               236     208    236
fisso_h                         208               208     208    208
cell_h                          236               236     208    236

P
                  durata_chiamate_r durata_chiamate_e fisso_h cell_h
durata_chiamate_r                   0e+00             0e+00   0e+00
durata_chiamate_e 0e+00                               3e-04   3e-04
fisso_h           0e+00             3e-04                     4e-04
cell_h            0e+00             3e-04             4e-04
> |
```



# Output rcorr - Esempio

```
> rcorr(as.matrix(tel))
                 durata_chiamate_r durata_chiamate_e fisso_h cell_h
durata_chiamate_r              1.00                          
durata_chiamate_e              0.79              1.00        
fisso_h                        0.28              0.25    1.00
cell_h                         0.32              0.23    0.24   1.00

n
                 durata_chiamate_r durata_chiam
durata_chiamate_r               236
durata_chiamate_e               236
fisso_h                         208
cell_h                          236

P
                 durata_chiamate_r durata_chiam
durata_chiamate_r                  0e+00
durata_chiamate_e 0e+00
fisso_h           0e+00            3e-04
cell_h            0e+00            3e-04
> 
```

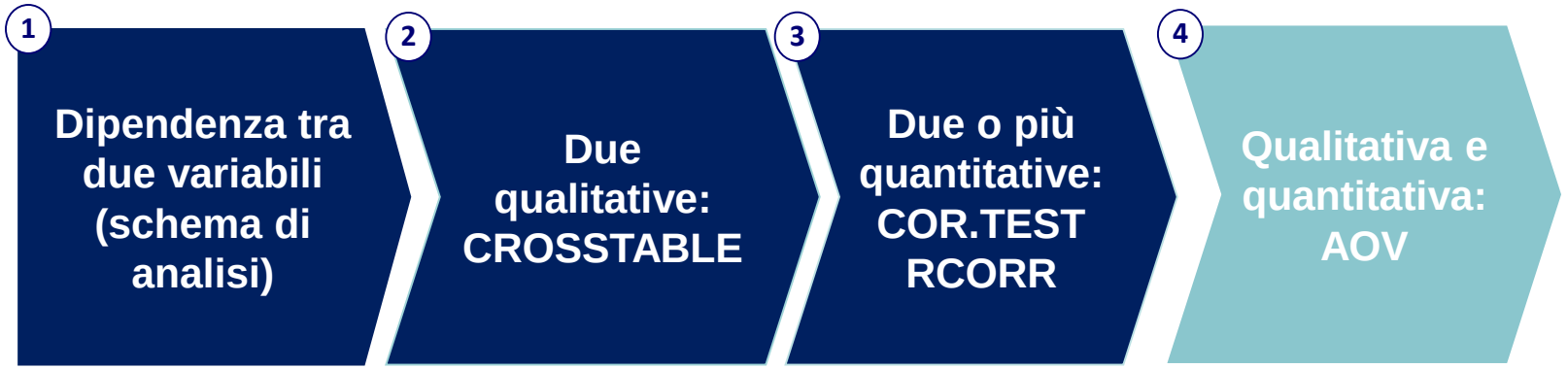


# Correlazione - Game

http://guessthecorrelation.com/

# Metodi Quantitativi per Economia, Finanza e Management

**Obiettivi di questa esercitazione**:



# Riepilogo teorico (1/4)

**X variabile qualitativa e Y variabile quantitativa**

Indaghiamo la relazione esistente confrontando le medie aritmetiche della variabile Y (quantitativa) sui gruppi di osservazioni generati dalle modalità assunte dalla variabile X (qualitativa)

Esempio:

X: sesso

Y: reddito

Le due variabili sono ***indipendenti in media*** se il reddito medio delle donne non è significativamente diverso dal reddito medio degli uomini

# Riepilogo teorico (2/4)

**X variabile qualitativa e Y variabile quantitativa**

$$SQT_y = SQ_{tra} + SQ_{nei}$$

dove

**$SQT_y$** somma dei quadrati degli scarti di ogni valore dalla media generale *(media reddito generale)*

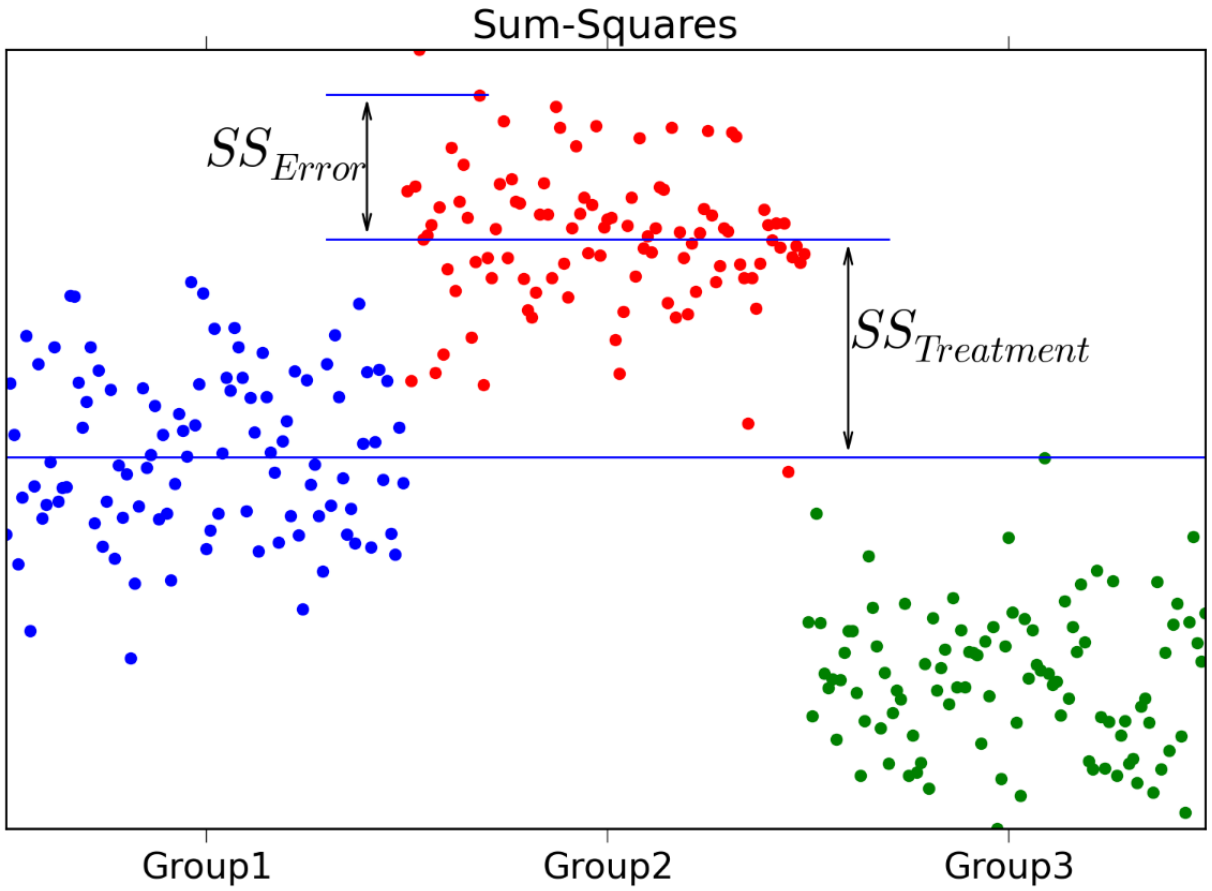
**$SQ_{tra}$** somma dei quadrati degli scarti di ogni media di gruppo *(media reddito donne, media reddito uomini)* dalla media generale *(media reddito generale)*

**$SQ_{nei}$** somma degli scarti al quadrato di ogni valore dalla media del suo gruppo *(media reddito donne o media reddito uomini)*

# Riepilogo teorico (3/4)

**X variabile qualitativa e Y variabile quantitativa**

$$SQT_y = SQ_{tra} + SQ_{nei}$$



# Riepilogo teorico (4/4)

**X variabile qualitativa e Y variabile quantitativa**

Indice relativo per misurare la dipendenza in media:

$$\eta^2 = SQ_{tra} / SQT_y = 1 - (SQ_{nei} / SQT_y)$$

- $\eta^2 = 0 \Rightarrow$ indipendenza in media
- $\eta^2 > 0 \Rightarrow$ dipendenza in media
- $\eta^2 = 1 \Rightarrow$ massima dipendenza in media

$\eta^2$ è sempre compreso tra 0 e 1.

# ANOVA

R prevede diversi modi per condurre l'analisi della varianza (ANOVA), utilizzata per confrontare le medie e le varianze di due o più gruppi di dati, per valutare se tali differenze sono statisticamente significative.

L'Anova si utilizza quindi quando la variabile o le variabili indipendenti sono di tipo categoriale, e la variabile dipendente è numerica.

# aov – Sintassi generale (1/2)

Sia Y una variabile quantitativa e X una variabile qualitativa

```
aov(y~x, data=nome_dataset)
```

~ è il simbolo TILDE, da tastierino numerico tenere premuto ALT e digitare 126 (ALT+126)

# aov – Sintassi generale (2/2)

Sia Y una variabile quantitativa e X una variabile qualitativa

```
anova=aov(y~x, data=nome_dataset)
```

**OUTPUT 1:**

```
model.tables(anova, type="means")
```

**OUTPUT 2:**

```
summary(anova)
```

# Esempio

C’è relazione tra la soddisfazione del cliente (SODDISFAZIONE_GLOBALE) e l’operatore telefonico da lui scelto (OPERATORE)?

```
aov(soddisfazione_globale~operatore,
data=telefonia)
```

# Esempio: Output 1

```
> anova=aov(soddisfazione_globale~operatore, data=telefonia)
> model.tables(anova, type="means")
Tables of means
Grand mean

6.493617

 operatore
      Tim    Tre Vodafone Wind
    6.164  6.417    6.627  6.4
rep 55.000 12.000  153.000 15.0
```

Media totale

Media dei singoli gruppi

La media della soddisfazione globale sembra molto vicina tra i diversi gruppi

# Esempio: Output 2

```
> anova=aov(soddisfazione_globale~operatore, data=telefonia)
> summary(anova)
             Df Sum Sq Mean Sq F value Pr(>F)
operatore     3    8.9   2.977   1.608  0.188
Residuals   231  427.8   1.852
1 observation deleted due to missingness
```

**Interpretazione:**

Nella tabella i valori riportati sono:

- Df = gradi di libertà
- Sum Sq = devianza (alla riga *operatore*, entro gruppi, alla riga *Residuals,* residua)
- Mean Sq = varianza (come sopra)
- F value = test F: *Mean Sq entro gruppi / Mean Sq residua*
- Pr(>F) = p-value

Ai fini dell'interpretazione, si deve ricordare che l'ipotesi nulla è che le varianze siano uguali fra di loro, e che dunque la variabile indipendente non produca effetti sulla variabile dipendente

# Esempio: Output 2

```
> anova=aov(soddisfazione_globale~operatore, data=telefonia)
> summary(anova)
             Df Sum Sq Mean Sq F value Pr(>F)
operatore     3    8.9   2.977   1.608  0.188
Residuals   231  427.8   1.852
1 observation deleted due to missingness
```

**Interpretazione:**

Ai fini dell'interpretazione, si deve ricordare che l'ipotesi nulla è che le varianze siano uguali fra di loro, e che dunque la variabile indipendente non produca effetti sulla variabile dipendente

La probabilità che sia vera l'ipotesi nulla è indicata dal valore Pr (p-value). Nel caso in esempio, la relazione non è significativa (accettiamo H0) ed quindi le due variabili sono quasi perfettamente indipendenti.

# Eta-Quadro

**X variabile qualitativa e Y variabile quantitativa**

Indice relativo per misurare la dipendenza in media:

$$\eta^2 = SQ_{tra} / SQT_y = 1 - (SQ_{nei} / SQT_y)$$

- $\eta^2 = 0 \Rightarrow$ indipendenza in media
- $\eta^2 > 0 \Rightarrow$ dipendenza in media
- $\eta^2 = 1 \Rightarrow$ massima dipendenza in media

$\eta^2$ è sempre compreso tra 0 e 1.

Per calcolare l'indice $\eta^2$ in R, bisogna scaricare il pacchetto *lsr* e richiamarlo.

```
library(lsr)
```

# etaSquared - Sintassi

```
etaSquared(nome_dataset_generato_da_aov)
```

```
> etaSquared( anova)
                eta.sq eta.sq.part
operatore 0.020451   0.020451
```

eta quadro

Anche il valore di eta-quadro è molto vicino a 0 →avvalora l'ipotesi di indipendenza in media

**NB: per una valutazione più oggettiva rimandiamo alla prossima lezione (test d'ipotesi)**

# 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

1. Allocare la DIRECTORY DI LAVORO (che punta alla cartella che contiene  il file DENTI.CSV )
2. Utilizzare la procedura corretta per analizzare la relazione di indipendenza tra area geografica e sesso
3. Utilizzare la procedura corretta per analizzare la relazione di indipendenza tra le variabili consumo di dentifrici della marca A e numero di contatti pubblicitari totali
4. Utilizzare la procedura corretta per analizzare la relazione di indipendenza tra la variabile consumo di dentifrici della marca A e area geografica e confrontarla con quella tra consumo di dentifrici della marca A e dimensione della città di residenza.