

# **Analisi Bivariata:** **Test Statistici**

*Metodi Quantitativi per Economia,  
Finanza e Management*

*Esercitazione n°5*

# Prima di iniziare..

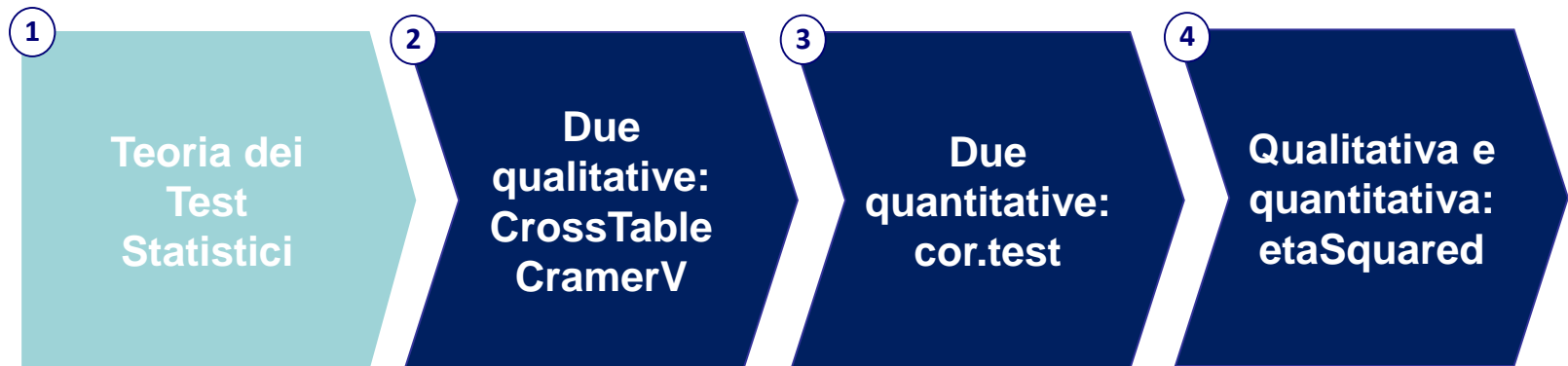
- Controllare se sul pc su cui state lavorando esiste già una cartella C:\corso. In tal caso eliminare tutto il contenuto. In caso contrario creare la cartella **corso** all'interno del disco C
- Andare sul disco condiviso F nel percorso **F:\corsi\Metodi\_Quantitativi\_EFM\_1718\esercitazione5** e copiare 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:**



# Scorsa lezione: Analisi Bivariata

TIPO DI VARIABILI	TIPO DI RELAZIONE INDAGATA	INDICI DI DIPENDENZA	FUNZIONI R
↓ Due Variabili Qualitative	↓ Indipendenza Statistica	↓ Chi Quadro, Cramer V	↓ CrossTable, CramerV
Due Variabili Quantitative	Indipendenza Lineare	coeff. di correlazione lineare	cor.test
Una Qualitative e Una Quantitativa continua	Indipendenza in media	indice eta-quadro	etaSquared



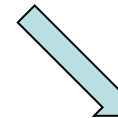
# Teoria dei Test d'Ipotesi (1/6)

## Cos'è un test d'ipotesi?

Il ricercatore fornisce ipotesi riguardo la distribuzione di una o più variabili della popolazione

### Obiettivo del test:

decidere se accettare o rifiutare l'ipotesi statistica alla luce di un risultato campionario



### TEST PARAMETRICI

Il ricercatore conosce la distribuzione delle variabili in analisi a meno di uno o più parametri e formula ipotesi sul valore dei parametri incogniti

### TEST NON PARAMETRICI

Il ricercatore fornisce delle ipotesi sul comportamento delle variabili, indipendentemente dalla conoscenza della loro distribuzione

**TEST per l'INDIPENDENZA DI DUE VARIABILI**



# Teoria dei Test d'Ipotesi (2/6)

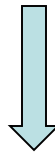
Vengono formulate due ipotesi:

- **H0** IPOTESI NULLA
- **H1** IPOTESI ALTERNATIVA (*rappresenta, di fatto, l'ipotesi che il ricercatore sta cercando di dimostrare*)

## Esempio (test d'indipendenza)

H0: X e Y sono indipendenti

H1: X e Y non sono indipendenti



L'obiettivo è verificare la plausibilità di un'affermazione (**ipotesi statistica**) riguardante la popolazione, ovvero il parametro da cui dipende, sulla base dell'evidenza campionaria



# Teoria dei Test d'Ipotesi (3/6)

Si possono commettere diversi tipi di errore:

Le due variabili sono realmente indipendenti

Esiste in natura una dipendenza tra le variabili

	STATO DI NATURA	
DECISIONE	$H_0$ Vera	$H_0$ Falsa
Accetto $H_0$	No errore	ERRORE SECONDO TIPO ( $\beta$ )
Rifiuto $H_0$	ERRORE PRIMO TIPO ( $\alpha$ )	No errore

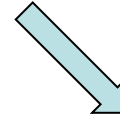
Sulla base del campione decido che c'è indipendenza

Sulla base del campione decido che c'è dipendenza



# Teoria dei Test d'Ipotesi (4/6)

Si possono commettere diversi tipi di errore:



## ERRORE PRIMO TIPO

- Rifiutare un'ipotesi nulla vera
- Considerato un tipo di errore molto serio
- La probabilità dell'errore di primo tipo è  $\alpha$

$\alpha$

**Livello di significatività del test**

## ERRORE SECONDO TIPO

- Non rifiutare un'ipotesi nulla falsa
- La probabilità dell'errore di secondo tipo è  $\beta$
- $(1 - \beta)$  è definito come la **potenza del test** (probabilità che un'ipotesi nulla falsa venga rifiutata)





# Teoria dei Test d'Ipotesi (5/6)

- Il ricercatore fissa a priori il livello di significatività del test (i valori comuni sono 0.01, 0.05, 0.10)
- L'obiettivo è quello di scegliere una delle due ipotesi, in modo che la probabilità di commettere un errore del primo tipo, sulla base dei dati campionari, sia bassa, o meglio inferiore al livello di significatività scelto:

$$P(\text{rifiutare } H_0 \mid H_0 \text{ vera}) < \alpha$$

## **P-value** («livello di significatività osservato»)

- Viene determinato sulla base di una statistica calcolata sui dati campionari (**statistica test**), che dipende dal test che si sta conducendo
- Rappresenta la probabilità di commettere l'errore di primo tipo sulla base del campione
- Deve essere confrontato con il valore di significatività scelto a monte



# Teoria dei Test d'Ipotesi (6/6)

## 1) Sistema di Ipotesi

- Formulazione ipotesi nulla e ipotesi alternativa
- Impostazione a priori del livello di significatività  $\alpha$



## 2) Calcolo Statistica test

- Calcolo del valore della statistica test (specifica del test che si sta conducendo) sulla base dei dati campionari



## 3) Calcolo P-value

- Calcolo del livello di significatività osservato



- Se **p-value**  $< \alpha$  → sulla base dei dati campionari, la probabilità di rifiutare  $H_0$  quando  $H_0$  è vera è inferiore alla soglia scelta → **rifiuto  $H_0$**
- Se **p-value**  $\geq \alpha$  → **accetto  $H_0$**



# Teoria dei Test d'Ipotesi - Esempio

## 1) Sistema di Ipotesi

$\left\{ \begin{array}{l} H_0: X \text{ e } Y \text{ sono indipendenti} \\ H_1: X \text{ e } Y \text{ dipendenti} \end{array} \right.$

- Fissiamo  $\alpha = 0.05$

## 2) Calcolo Statistica test

## 3) Calcolo P-value

- Se **p-value < 0.05** → **rifiuto H0** → *concludo che X e Y sono dipendenti*
- Se **p-value ≥ 0.05** → **accetto H0** → *concludo che X e Y sono indipendenti*



# Test per l'indipendenza tra due variabili

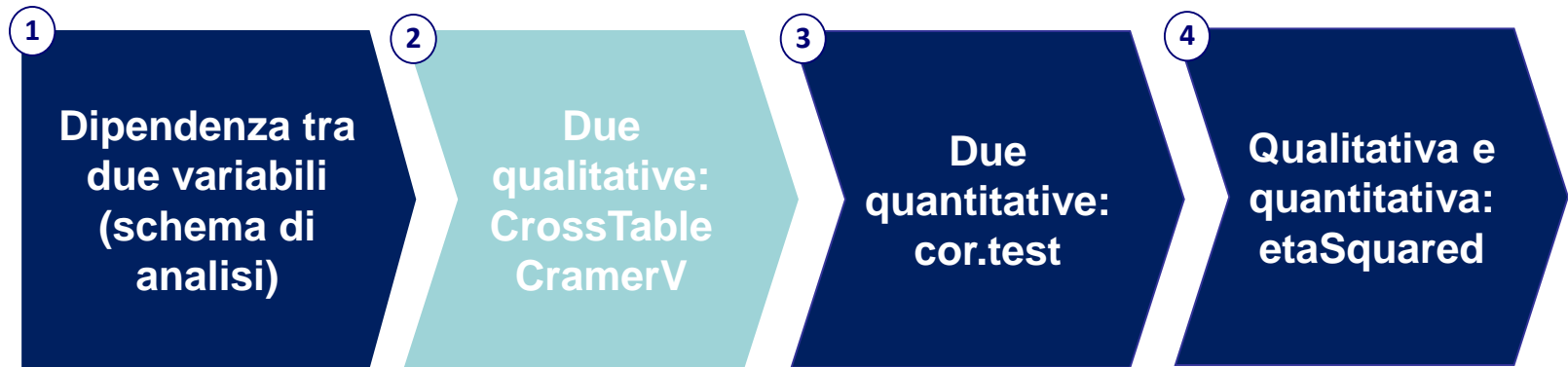
*Il test per la valutazione dell'indipendenza di due variabili dipende dalla natura delle variabili considerate:*

<b>Due Variabili Qualitative</b>	Test per l'Indipendenza Statistica	CrossTable CramerV
<b>Due Variabili Quantitative</b>	Test per l'Indipendenza Lineare	cor.test
<b>Una Qualitative e Una Quantitativa continua</b>	Test per l'Indipendenza in media	etaSquared



# Metodi Quantitativi per Economia, Finanza e Management

**Obiettivi di questa esercitazione:**



# Test per l'indipendenza statistica

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

Ipotesi:

H0: X e Y sono statisticamente indipendenti

H1: X e Y sono statisticamente dipendenti

Statistica test:

Statistica Chi-Quadro

Regola di decisione:

Se  $p\text{-value} < \alpha \rightarrow$  rigetto H0  $\rightarrow$  X e Y sono statisticamente dipendenti

Se  $p\text{-value} \geq \alpha \rightarrow$  accetto H0  $\rightarrow$  X e Y sono statisticamente indipendenti



# CrossTable

Test d'indipendenza statistica tra due variabili qualitative o quantitative discrete

**Variabili qualitative:** sesso e operatore telefonico

```
CrossTable(dataset$variabile1,  
dataset$variabile2,  
prop.chisq=FALSE, chisq=TRUE)
```

Test d'indipendenza statistica tra due variabili qualitative o quantitative discrete

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



# CrossTable – Esempio

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

```
CrossTable(telefonია$sezzo,  
           telefonია$computer,  
           prop.chisq=FALSE, chisq=TRUE)
```

```
CramerV(telefonია$sezzo,  
        telefonია$computer)
```





# Scorsa lezione: tabella di contingenza

Cell Contents

```
-----  
N  
N / Row Total  
N / Col Total  
N / Table Total  
-----
```

```
=====
                telefonia$computer
telefonia$sexo  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:**

- **29.41% computer=0**
- **70.59% computer=1**

Le distribuzioni della variabile computer, condizionate al sesso, sono diverse (viceversa quelle del sesso condizionate al possesso del computer)

→ ci fa pensare alla presenza di dipendenza tra le due variabili!



# Scorsa lezione: Indici di connessione

Statistics for All Table Factors

Pearson's Chi-squared test

-----  
Chi<sup>2</sup> = 5.727462      d.f. = 1      p = 0.0167

Pearson's Chi-squared test with Yates' continuity correction

-----  
Chi<sup>2</sup> = 5.010379      d.f. = 1      p = 0.0252

```
> CramerV(telefoniasesso, telefoniascomputer)
[1] 0.1557848
```

Come valutiamo la presenza di indipendenza a partire dagli indici calcolati? Chi-quadro “vicino” a 0, Cramer V “prossimo” a 0

→ Vediamo cosa dice il **Test d’ipotesi**



# Risultato del Test

Statistics for All Table Factors

Pearson's Chi-squared test

Chi<sup>2</sup> = 5.727462      d.f. = 1      p = 0.0167

Pearson's Chi-squared test with Yates' continuity correction

Chi<sup>2</sup> = 5.010379      d.f. = 1      p = 0.0252

H0: X e Y sono statisticamente indipendenti

H1: X e Y sono statisticamente dipendenti

**P-value=0.0167**

Sia  $\alpha = 0.05$ :

p-value <  $\alpha$  → rigetto H0 →

concludo che X e Y sono statisticamente dipendenti

**Se avessimo scelto un livello di significatività diverso?**

...con  $\alpha = 0.01$ :

p-value  $\geq \alpha$  → accetto H0 → X e Y sono statisticamente indipendenti

**A seconda del livello di significatività fissato possiamo raggiungere conclusioni differenti!**

**NB.** Se considerando i valori più comuni di  $\alpha$  (0.01, 0.05, 0.1), si ottengono conclusioni diverse, si può dire che sulla base del campione la presunta relazione di dipendenza non è così forte.



# CrossTable - Esempio 2

C'è indipendenza statistica tra le variabili SESSO e MARCA?

```
CrossTable(telefonია$sezzo,  
           telefonია$marca, prop.chisq=FALSE,  
           chisq=TRUE)
```



# CrossTable : Esempio 2

telefonია\$secco	telefonია\$marca									Total
	Altro	Lg	Motorola	Nek	Nokia	PalmOne	Samsung	Siemens	Sony Ericsson	
F	2	8	19	2	45	1	15	1	7	100
	0.020	0.080	0.190	0.020	0.450	0.010	0.150	0.010	0.070	0.424
	0.333	0.615	0.365	0.500	0.437	1.000	0.375	0.200	0.583	
	0.008	0.034	0.081	0.008	0.191	0.004	0.064	0.004	0.030	
M	4	5	33	2	58	0	25	4	5	136
	0.029	0.037	0.243	0.015	0.426	0.000	0.184	0.029	0.037	0.576
	0.667	0.385	0.635	0.500	0.563	0.000	0.625	0.800	0.417	
	0.017	0.021	0.140	0.008	0.246	0.000	0.106	0.017	0.021	
Total	6	13	52	4	103	1	40	5	12	236
	0.025	0.055	0.220	0.017	0.436	0.004	0.169	0.021	0.051	

## Attenzione:

molte celle con frequenze congiunte assolute molto basse  
 (<5) → test non affidabile



# CrossTable: Esempio 2

Statistics for All Table Factors

Pearson's Chi-squared test

-----  
Chi^2 = 7.075429      d.f. = 8      p = 0.529

Warning message:

```
In chisq.test(tab, correct = FALSE, ...) :  
  Chi-squared approximation may be incorrect
```

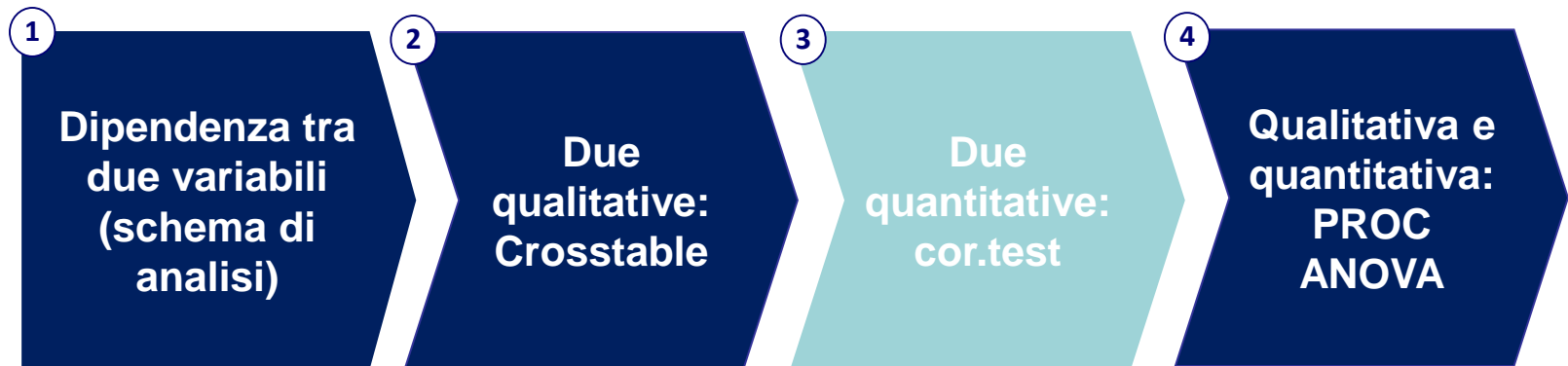
```
> |
```

**Se più del 20% delle celle ha frequenza assoluta < 5, R segnala che il test non è affidabile!**



# Metodi Quantitativi per Economia, Finanza e Management

**Obiettivi di questa esercitazione:**



# Test per l'indipendenza lineare

**X e Y due variabili quantitative**

Ipotesi:

H0: X e Y sono linearmente indipendenti ( $\rho_{\text{popolaz}}=0$ )

H1: X e Y sono linearmente dipendenti ( $\rho_{\text{popolaz}}\neq 0$ )

Statistica test:

Statistica t di Student

Regola di decisione:

Se p-value  $< \alpha \rightarrow$  rigetto H0  $\rightarrow$  X e Y sono linearmente dipendenti

Se p-value  $\geq \alpha \rightarrow$  accetto H0  $\rightarrow$  X e Y sono linearmente indipendenti





# Cor.test – Ripasso sintassi

Test per la correlazione tra due o più variabili quantitative

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



# Cor.test – Esempio1

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)
```



# Scorsa Lezione: Indice di correlazione

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
```

**P-value = 0.00038**

- Sia fissando  $\alpha = 0.05$  che  $\alpha = 0.01$

$p\text{-value} < \alpha \rightarrow$  rigetto  $H_0 \rightarrow X$  e  $Y$  sono linearmente dipendenti

- Conclusione: esiste una relazione lineare tra le due variabili, anche se non molto forte (il coefficiente di correlazione lineare non è nullo, ma ha valore non molto elevato)

*Coefficiente di correlazione lineare  $\rho(X, Y)$ : presenza di relazione lineare positiva*



# Correlazione tra più variabili

## Sintassi

```
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 e richiamare il pacchetto Hmisc



# Correlazione tra più variabili

## Esempio2

Correlazione tra il numero medio di ore di utilizzo del telefono cellulare, numero ore di utilizzo del fisso al giorno, durata media delle chiamate effettuate, durata media delle chiamate ricevute

```
new_telefonia<-telefonía[,c("durata_chiamate_e", "durata_chiamate_r",  
"cell_h", "fisso_h")]
```

```
rcorr(as.matrix(new_telefonia))
```



# Esempio2: Risultati

```
> rcorr(as.matrix(telefonია[,c("durata chiamate e", "durata chiamate r", "cell_h", "fisso_h")])
durata_chiamate_e durata_chiamate_r cell_h fisso_h
durata_chiamate_e 1.00 0.79 0.23 0.25
durata_chiamate_r 0.79 1.00 0.32 0.28
cell_h 0.23 0.32 1.00 0.24
fisso_h 0.25 0.28 0.24 1.00

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

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

Valore della correlazione

La correlazione tra la durata della chiamata in uscita è correlata con la durata delle chiamate in entrata

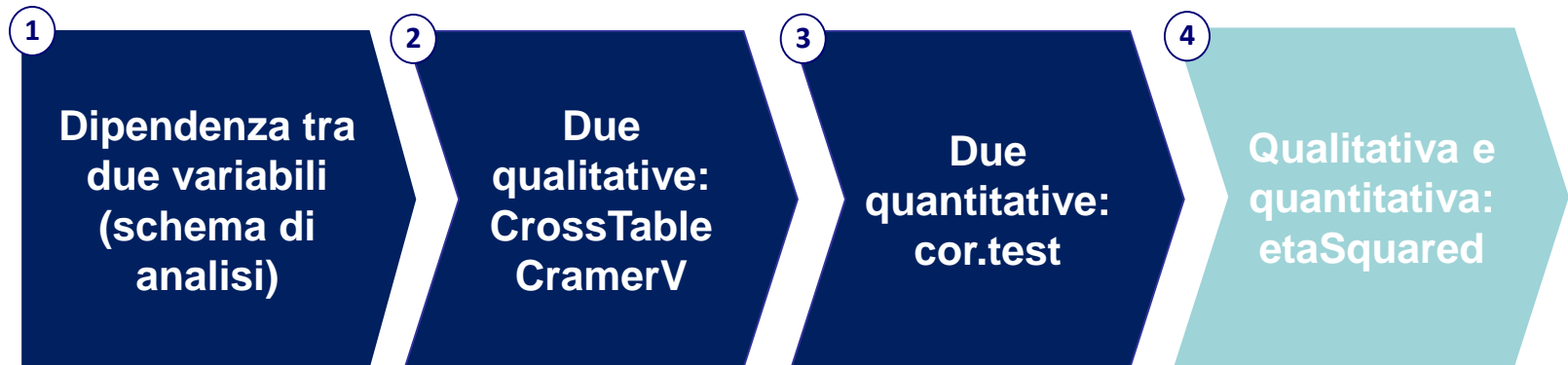
Valore p-value per ogni singola correlazione

La correlazione tra durata chiamata in entrata e durata chiamata in uscita è significativa, poiché p-value pari a zero, quindi rifiutiamo  $H_0$ . Ciò significa che X e Y sono linearmente dipendenti.



# Metodi Quantitativi per Economia, Finanza e Management

**Obiettivi di questa esercitazione:**



# Test per l'indipendenza in media

**X variabile qualitativa, Y variabile quantitativa**

Ipotesi:

$H_0$ : X e Y sono indipendenti in media  $\leftrightarrow$

$\mu_1 = \mu_2 = \dots = \mu_k$  (le medie di Y nei gruppi sono tutte uguali tra loro)

$H_1$ : X e Y sono dipendenti in media  $\leftrightarrow$

le  $\mu_i$  non sono tutte uguali (esistono almeno due medie diverse tra loro)

Statistica test:

Statistica F di Fisher

Regola di decisione:

Se p-value  $< \alpha \rightarrow$  rigetto  $H_0 \rightarrow$  X e Y sono dipendenti in media

Se p-value  $\geq \alpha \rightarrow$  accetto  $H_0 \rightarrow$  X e Y sono indipendenti in media





# 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)?

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

```
model.table(anova, types="means")
```

```
Summary(anova)
```



# Esempio

```
> anova=aov(soddisfazione_globale~operatore, data=telefonია)
> 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  $H_0$ ), poiché p-value > di  $\alpha$ , sia che il livello  $\alpha$  sia 5% che 1%, ed anzi le due variabili sono quasi perfettamente indipendenti.



# etaSquared - Sintassi

```
etaSquared(anova)
```

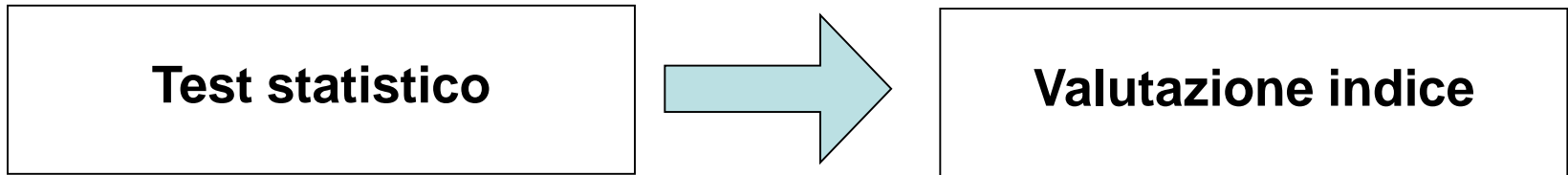
```
> etaSquared( anova)
              eta.sq eta.sq.part
operatore 0.020451 0.020451
```

eta quadro

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



# Approccio di analisi



- 1) Eseguire l'opportuno test statistico in dipendenza dalla tipologia delle variabili poste a confronto;
- 2) Analizzare l'esito del test (p-value confrontato con  $\alpha$ ):
  - a) Indipendenza tra le due variabili → verificare se il valore dell'indice di dipendenza conferma l'esito del test;
  - b) Dipendenza tra le due variabili → valutare il valore dell'indice di dipendenza per indagare la forza della relazione.



# 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	ETACCLASS	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 S/NO
7	ALTOCON	Num	ALTO CONSUMANTE S/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 S/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 S/NO
24	TATTI_B	Num	NUMERO DI CONTATTI PUBBLICITARI (DENTIFRICI MARCA B)

# Esercizi

1. Allocare l'area di lavoro, in modo che punti alla cartella fisica dove è contenuto il file Excel DENTI\_NEW.csv
2. Analizzare la relazione di indipendenza tra area geografica e sex e commentare il test d'ipotesi.
3. Analizzare la relazione di indipendenza tra le variabili consumo di dentifrici della marca A e numero di contatti pubblicitari totali e commentare il test d'ipotesi.
4. 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, commentando i test d'ipotesi.