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_1516\esercitazione5*** e copiare il contenuto nella cartella C:\corso

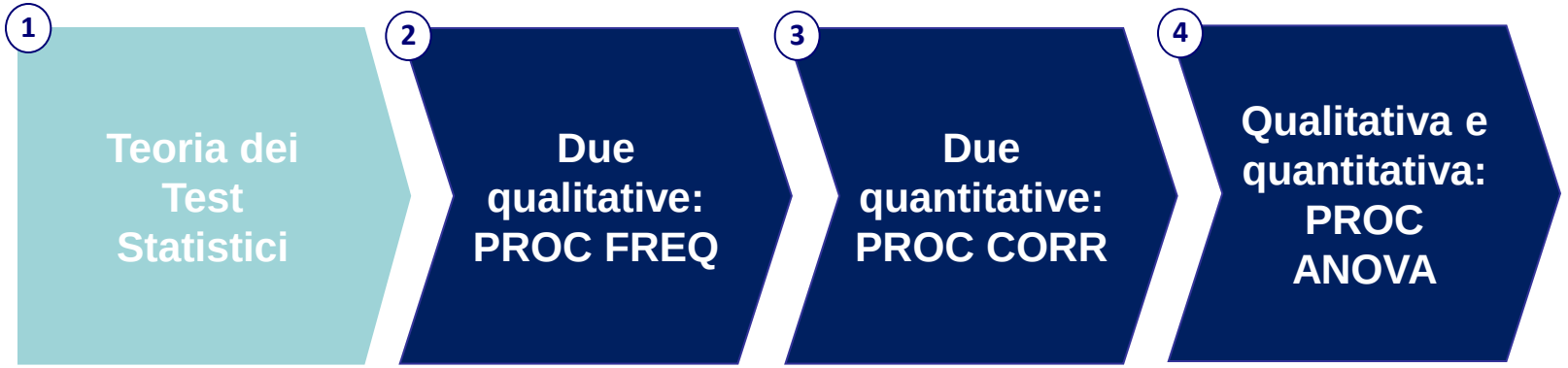
- Aprire il programma SAS (Start → All Programs → SAS → SAS 9.3)

- Allocare la libreria ***corso,*** puntando il percorso fisico C:\corso, utilizzando l'istruzione:

```
libname corso 'C:\corso';
```

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	PROCEDURA SAS
Due Variabili Qualitative	Indipendenza Statistica	Chi Quadro, Cramer V	PROC FREQ
Due Variabili Quantitative	Indipendenza Lineare	coeff. di correlazione lineare	PROC CORR
Una Qualitative e Una Quantitativa continua	Indipendenza in media	indice eta-quadro	PROC ANOVA

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

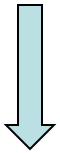
Teoria dei Test d'Ipotesi (2/6)

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 stabilire se, sulla base dei dati campionari osservati, l'ipotesi nulla è «verosimile». Viene rifiutata se il campione osservato è «improbabile» ritenendo vera quell'ipotesi.

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 (β)
Rifiuto H_0	ERRORE PRIMO TIPO (α)	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 α

Livello di significatività del test

ERRORE SECONDO TIPO

- Non rifiutare un'ipotesi nulla falsa
- La probabilità dell'errore di secondo tipo è **β**
- **(1 – β)** 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 α

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 < α** → sulla base dei dati campionari, la probabilità di rifiutare H0 quando H0 è vera è inferiore alla soglia scelta → **rifiuto H0**
- Se **p-value ≥ α** → **accetto H0**

Teoria dei Test d'Ipotesi - Esempio

1) Sistema di Ipotesi

{H0: X e Y sono indipendenti
H1: X e Y dipendenti

- 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*

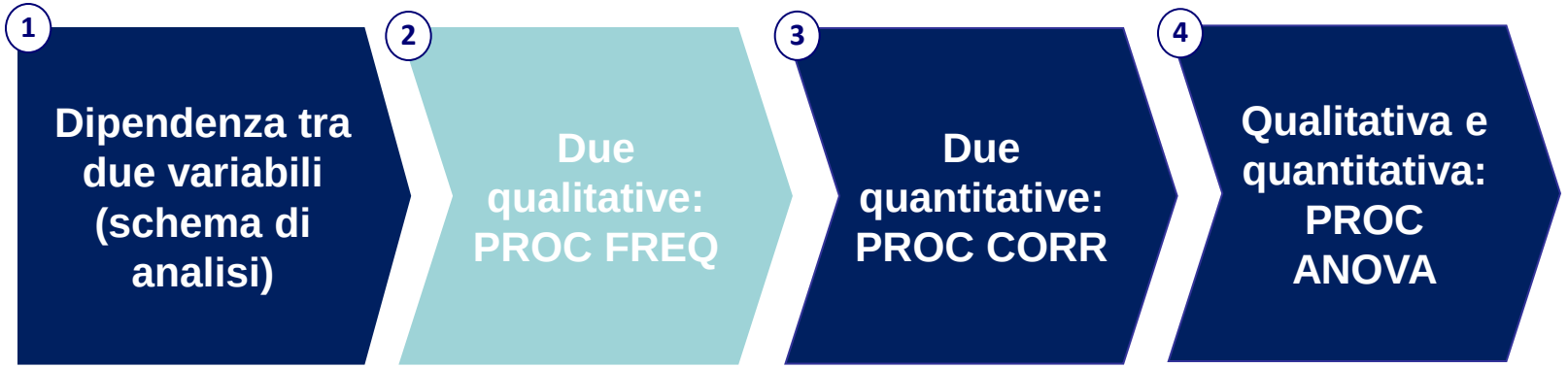
Test per l'indipendenza statistica

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

Due Variabili Qualitative	Test per l'Indipendenza Statistica	PROC FREQ
Due Variabili Quantitative	Test per l'Indipendenza Lineare	PROC CORR
Una Qualitative e Una Quantitativa continua	Test per l'Indipendenza in media	PROC ANOVA

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-value < α → rigetto H0 → X e Y sono statisticamente dipendenti

Se p-value ≥ α → accetto H0 → X e Y sono statisticamente indipendenti

PROC FREQ

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

```
proc freq data= dataset;
    tables variabile1 * variabile2 /chisq;
run;
```

NB: tra le opzioni posso sempre inserire l'opzione missing, per considerare anche i missing nel calcolo delle frequenze:

```
tables variabile1 * variabile2 /missing chisq;
```

PROC FREQ – Esempio

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

```
proc freq data=corso.telefonia;
tables sesso * computer /chisq;
run;
```

Scorsa lezione: tabella di contingenza

Frequency Percent Row Pct Col Pct	Table of sesso by computer			
		computer(computer)		
	sesso(sesso)	0	1	Total
	F	16 6.78 16.00 28.57	84 35.59 84.00 46.67	100 42.37
	M	40 16.95 29.41 71.43	96 40.68 70.59 53.33	136 57.63
	Total	56 23.73	180 76.27	236 100.00

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

Statistic	DF	Value	Prob
Chi-Square	1	5.7275	0.0167
Likelihood Ratio Chi-Square	1	5.9139	0.0150
Continuity Adj. Chi-Square	1	5.0104	0.0252
Mantel-Haenszel Chi-Square	1	5.7032	0.0169
Phi Coefficient		-0.1558	
Contingency Coefficient		0.1539	
Cramer's V		-0.1558	

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

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

Risultato del Test

Statistic	DF	Value	Prob
Chi-Square	1	5.7275	0.0167
Likelihood Ratio Chi-Square	1	5.9139	0.0150
Continuity Adj. Chi-Square	1	5.0104	0.0252
Mantel-Haenszel Chi-Square	1	5.7032	0.0169
Phi Coefficient		-0.1558	
Contingency Coefficient		0.1539	
Cramer's V		-0.1558	

P-value=0.0167

Sia α =0.05:

p-value < α → rigetto H0 →

concludo che X e Y sono statisticamente dipendenti

Se avessimo scelto un livello di significatività diverso?

…con α =0.01:
p-value ≥ α → 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 α (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.

PROC FREQ: Esempio 2

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

```
proc freq data=corso.telefonia;
tables sesso * marca /chisq;
run;
```

PROC FREQ: Esempio 2

Table of sesso by marca										
	marca(marca)									
sesso(sesso)	Altro	Lg	Motorola	Nek	Nokia	PalmOne	Samsung	Siemens	Sony Ericsson	Total
F	2 0.85 2.00 33.33	8 3.39 8.00 61.54	19 8.05 19.00 36.54	2 0.85 2.00 50.00	45 19.07 45.00 43.69	1 0.42 1.00 100.00	15 6.36 15.00 37.50	1 0.42 1.00 20.00	7 2.97 7.00 58.33	100 42.37
M	4 1.69 2.94 66.67	5 2.12 3.68 38.46	33 13.98 24.26 63.46	2 0.85 1.47 50.00	58 24.58 42.65 56.31	0 0.00 0.00 0.00	25 10.59 18.38 62.50	4 1.69 2.94 80.00	5 2.12 3.68 41.67	136 57.63
Total	6 2.54	13 5.51	52 22.03	4 1.69	103 43.64	1 0.42	40 16.95	5 2.12	12 5.08	236 100.00

Attenzione:

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

PROC FREQ: Esempio 2

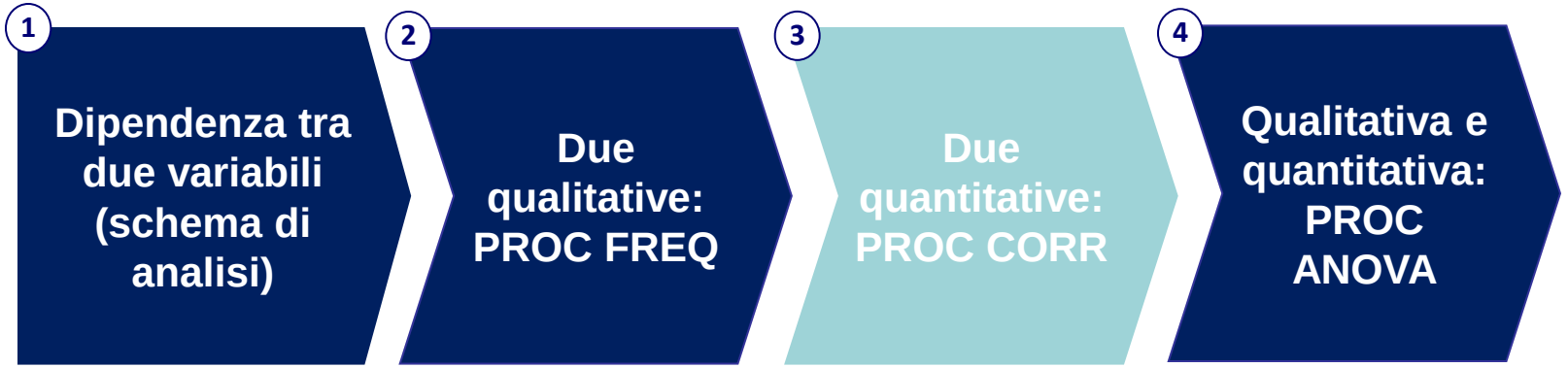
Statistic	DF	Value	Prob
Chi-Square	8	7.0754	0.5285
Likelihood Ratio Chi-Square	8	7.5018	0.4836
Mantel-Haenszel Chi-Square	1	0.0103	0.9191
Phi Coefficient		0.1731	
Contingency Coefficient		0.1706	
Cramer's V		0.1731	

WARNING: 44% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Se più del 20% delle celle ha frequenza assoluta < 5, SAS lo segnala e 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 ($\rho_{popolaz}=0$)
H1: X e Y sono linearmente dipendenti ($\rho_{popolaz}\neq 0$)

Statistica test:

Statistica t di Student

Regola di decisione:

Se p-value < α → rigetto H0 → X e Y sono linearmente dipendenti

Se p-value ≥ α → accetto H0 → X e Y sono linearmente indipendenti

PROC CORR

Test per la correlazione tra due o più variabili quantitative

```
proc corr data= dataset;
   var variabile1 variabile2 … variabilen;
run;
```

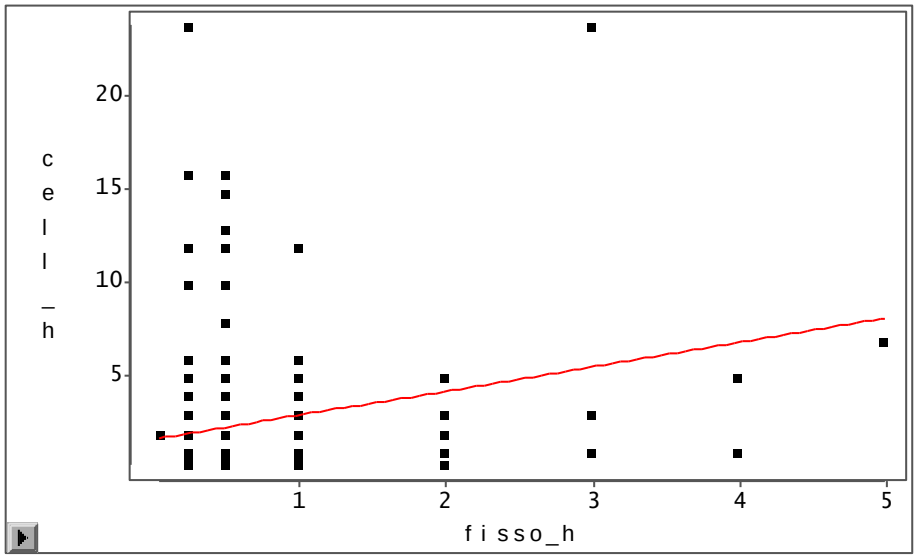
PROC CORR - Esempio

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

```
proc corr data=corso.telefonia;
var cell_h fisso_h;
run;
```

Scorsa Lezione: Indice di correlazione

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	cell_h	fisso_h
cell_h cell_h	1 236	0.24403 0.0004 208
fisso_h fisso_h	0.24403 0.0004 208	1 208



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

Risultato del test

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	cell_h	fisso_h
cell_h cell_h	1 236	0.24403 0.0004 208
fisso_h fisso_h	0.24403 0.0004 208	1 208

P-value = 0.0004

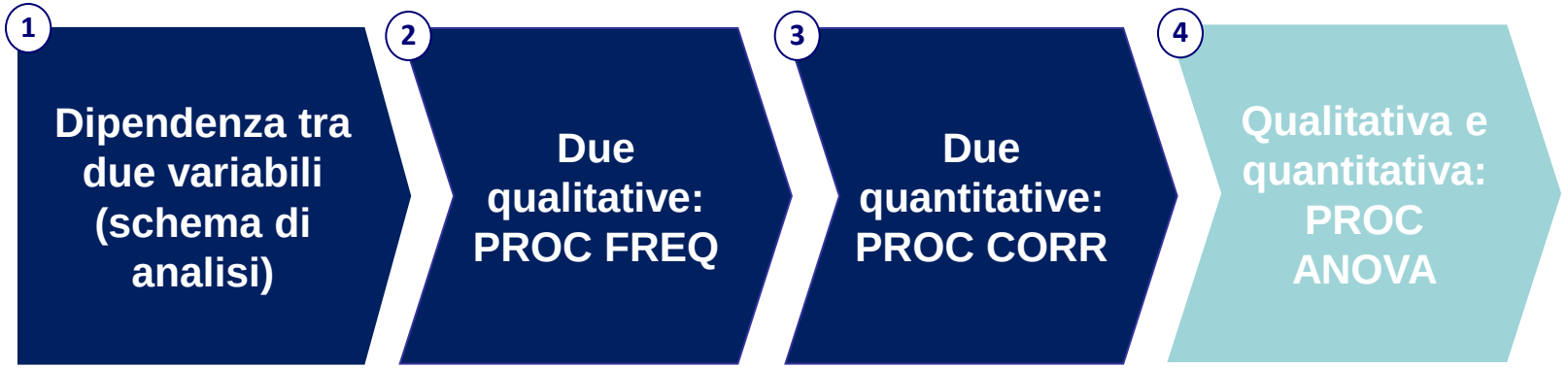
- Sia fissando α =0.05 che α =0.01

p-value < α → rigetto H0 → 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)

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Test per l'indipendenza in media

X variabile qualitativa, Y variabile quantitativa

Ipotesi:

H0: X e Y sono indipendenti in media ↔
μ1 = μ2 = … = μk (le medie di Y nei gruppi sono tutte uguali tra loro)

H1: X e Y sono dipendenti in media ↔
le μi non sono tutte uguali (esistono almeno due medie diverse tra loro)

Statistica test:

Statistica F di Fisher

Regola di decisione:

Se p-value < α → rigetto H0 → X e Y sono dipendenti in media

Se p-value ≥ α → accetto H0 → X e Y sono indipendenti in media

PROC ANOVA

Test d'indipendenza in media tra:

Y variabile quantitativa e X variabile qualitativa

```
PROC ANOVA DATA=dataset;
 CLASS X;
 MODEL Y=X;
 MEANS X;
RUN;
```

PROC ANOVA - Esempio

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

```
PROC ANOVA DATA =corso.telefonia;
 CLASS operatore;
 MODEL soddisfazione_globale=operatore;
 MEANS operatore;
RUN;
```

Scorsa lezione: considerazioni

Level of operatore	N	soddisfazione_globale	
		Mean	Std Dev
Tim	55	6.16363636	1.33004645
Tre	12	6.41666667	1.31137217
Vodafone	153	6.62745098	1.29209313
Wind	15	6.4	2.06328448

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

R-Square	Coeff Var	Root MSE	soddisfazione_globale Mean
0.020451	20.9571	1.360877	6.493617

eta quadro

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

Risultato del Test:

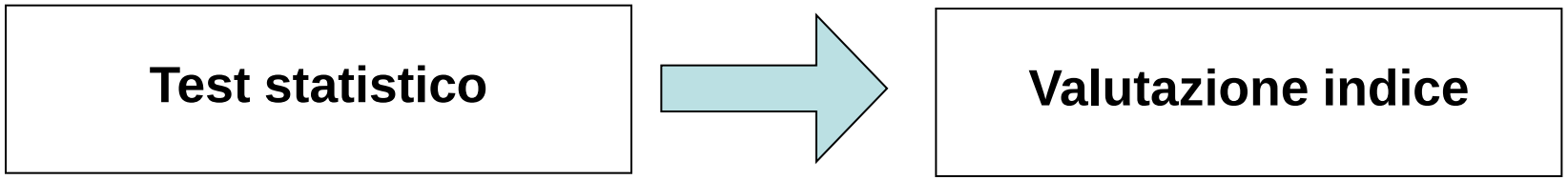
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8.9317803	2.9772601	1.61	0.1884
Error	231	427.8086453	1.8519855		
Corrected Total	234	436.7404255			

P-value = 0.1884

Fissando α =0.05

p-value > α → accetto H0 → X e Y sono indipendenti 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 (pvalue):
   a) Indipendenza tra le due variabili → verificare se il valore dell'indice conferma l'esito del test;
   b) Dipendenza tra le due variabili → valutare il valore dell'indice 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	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 libreria ESER5, in modo che punti alla cartella fisica dove è contenuta la tabella SAS «DENTI_NEW»
2. Analizzare la relazione di indipendenza tra area geografica e sesso
3. Analizzare la relazione di indipendenza tra le variabili consumo di dentifrici della marca A e numero di contatti pubblicitari totali
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.