Analisi Bivariata

Metodi Quantitativi per Economia, Finanza e Management

Esercitazione n°4

Lavoro di gruppo

- Inviare il questionario via mail a gdeppieri@liuc.it e gmagistrelli@liuc.it **entro oggi 30/10/2015**

- 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

  Tre argomenti a scelta tra
  - Analisi fattoriale
  - Regressione lineare
  - Regressione Logistica
  - Serie storiche

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

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

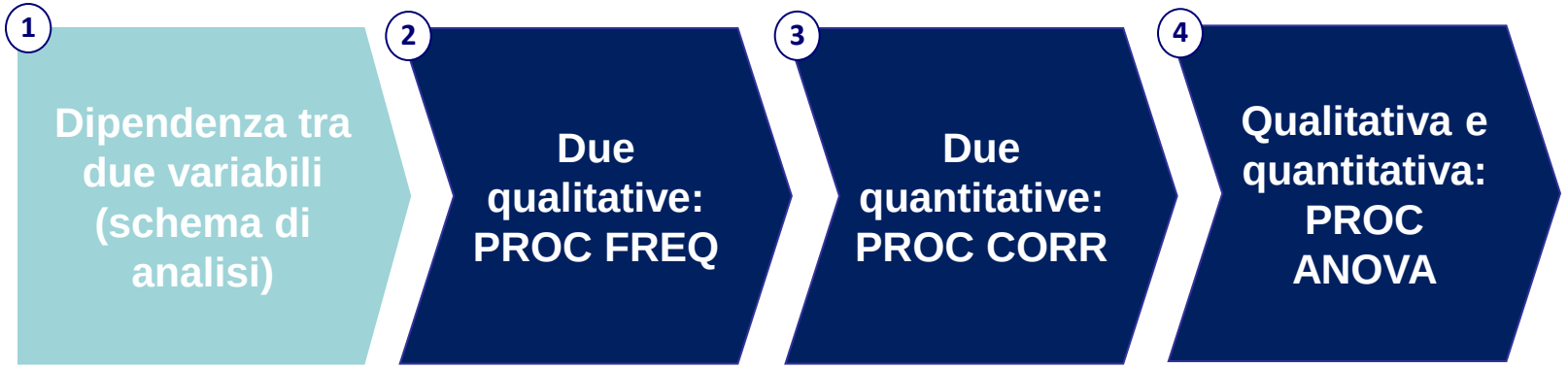
- Allocare la libreria ***corso,*** puntando il percorso fisico C:\corso, utilizzando l'istruzione:

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

- Nella libreria dovreste visualizzare la tabella TELEFONIA

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Analisi Bivariata

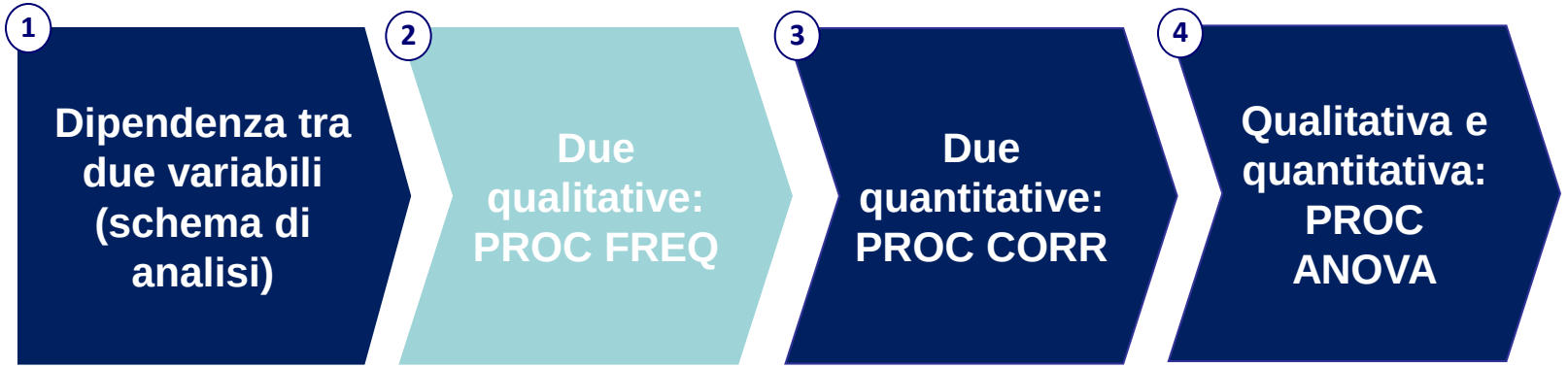
Studio della distribuzione di due variabili congiuntamente considerate e delle relazioni esistenti tra esse

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)	PROC FREQ
Due Variabili Quantitative	Indipendenza Lineare (indice: coeff. di correlazione lineare)	PROC CORR
Una Qualitative e Una Quantitativa continua	Indipendenza in media (indice: eta-quadro)	PROC ANOVA

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 SAS 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:

- *χ^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 χ^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.

PROC FREQ - Descrizione

La PROC FREQ permette di

1. calcolare le distribuzioni di frequenza univariate per variabili qualitative e quantitative discrete

```
proc freq data= dataset;
   tables variabile /option(s);
run;
```

ESERCITAZIONE 3!

2. creare tabelle di contingenza a due o più dimensioni per variabili qualitative e quantitative discrete

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

PROC FREQ – Sintassi generale

Distribuzione di frequenza bivariata (tabelle di contingenza)

```
proc freq data= dataset;
   tables variabile1 * variabile2 /option(s);
run;
```

OPTIONS:

- /missing considera anche i missing nel calcolo delle frequenze

Rispetto alla sintassi della distribuzione di frequenza univariata bisogna aggiungere

**** nome variabile2***

PROC FREQ – Esempio 1

Variabili qualitative: sesso e operatore telefonico

```
proc freq data=corso.telefonia;
tables sesso * operatore;
run;
```

Output PROC FREQ - Esempio 1

Frequenze congiunte assolute e relative

Distribuzioni marginali: frequenze marginali assolute e relative

Frequenze subordinate di riga e colonna

Frequency Percent Row Pct Col Pct	Table of sesso by operatore				
sesso	operatore				Total
	Tre	Tim	Vodafone	Wind	
F	7 2.97 7.00 58.33	27 11.44 27.00 49.09	63 26.69 63.00 40.91	3 1.27 3.00 20.00	100 42.37
M	5 2.12 3.68 41.67	28 11.86 20.59 50.91	91 38.56 66.91 59.09	12 5.08 8.82 80.00	136 57.63
Total	12 5.08	55 23.31	154 65.25	15 6.36	236 100.00

Output PROC FREQ – Esempio 1

freq. congiunta relativa =(7/236)*100

freq. marginale assoluta=7+27+63+3

freq. marginale relativa=(7+27+63+3)/236*100

freq. subordinate:

% di riga=5/136*100

% di col=5/12*100

Frequency Percent Row Pct Col Pct	Table of sesso by operatore				
sesso	operatore				Total
	Tre	Tim	Vodafone	Wind	
F	7 2.97 7.00 58.33	27 11.44 27.00	63 26.69 63.00	3 1.27 3.00	100 42.37
M	5 2.12 3.68 41.67	28 11.86 20.59 50.91	91 38.56 66.91 59.09	12 5.08 8.82 80.00	136 57.63
Total	12 5.08	55 23.31	154 65.25	15 6.36	236 100.00

PROC FREQ – Esempio 2

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

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

Output PROC FREQ – Esempio 2

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

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

Output PROC FREQ – Esempio 2

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

Output PROC FREQ – Esempio 2

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

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

Computer=0:

- **28.57% F**
- **71.43% M**

Computer=1:

- **46.67% F**
- **53.33% M**

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

PROC FREQ - Descrizione

La PROC FREQ permette di

1. calcolare le distribuzioni di frequenza univariate per variabili qualitative e quantitative discrete

```
proc freq data= dataset;
   tables variabile /option(s);
run;
```

ESERCITAZIONE 3!

2. creare tabelle di contingenza a due o più dimensioni per variabili qualitative e quantitative discrete

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

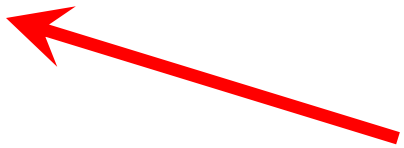
PROC FREQ – Sintassi generale

Calcolo dell'indice chi-quadro e Cramer V

```
proc freq data= dataset;
   tables variabile1 * variabile2 /option(s);
run;
```

OPTIONS:

- /missing considera anche i missing nel calcolo delle frequenze
- **/chisq calcola l'indice chi-quadro e altre misure di associazione basate sul chi-quadro**



Esempio n°1- Indici Chi-Quadro e Cramer V

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

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

Esempio n°1- Indici Chi-Quadro e Cramer V

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	

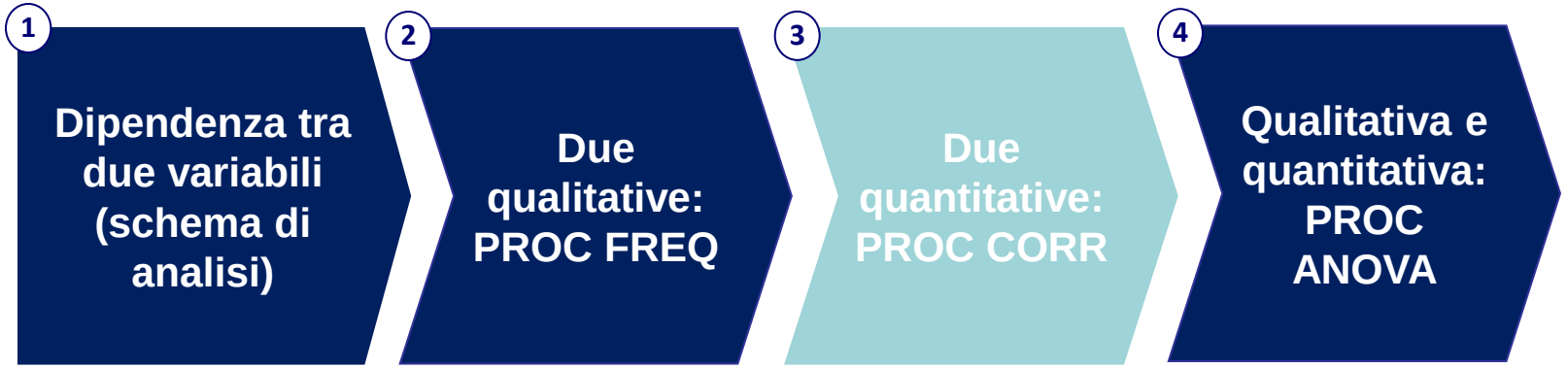
Solo con tabelle 2X2:
SAS utilizza una formula per il Cramer V leggermente modificata → l'indice varia tra -1 e 1

Come valutiamo la presenza di indipendenza a partire dagli indici calcolati?

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

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Riepilogo teorico

X e Y due variabili quantitative

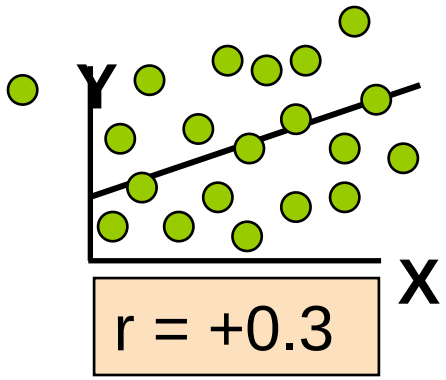
Indaghiamo la presenza di una relazione lineare tra le due variabili

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

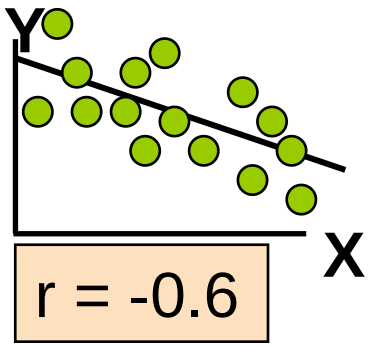
$$\rho = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

$\rho = 0$ → non c'è relazione lineare tra X e Y

$\rho > 0$ → relazione lineare positiva tra X e Y



$\rho < 0$ → relazione lineare negativa tra X e Y



PROC CORR - Descrizione

La PROC CORR permette di

- calcolare la correlazione tra due o più variabili quantitative

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

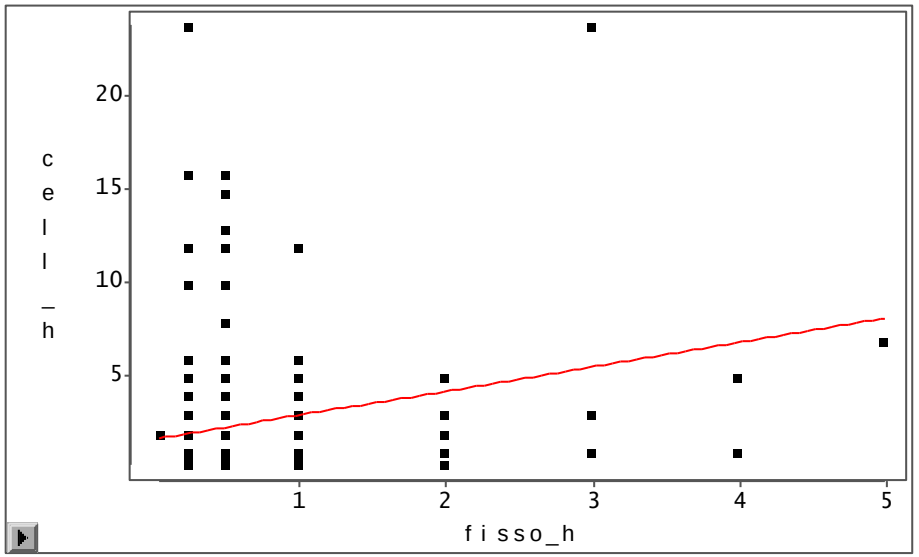
PROC CORR - Esempio

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;
```

Output PROC CORR - Esempio

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

PROC CORR - Esempio

Correlazione tra la durata media delle chiamate effettuate [durata_chiamate_e] e:

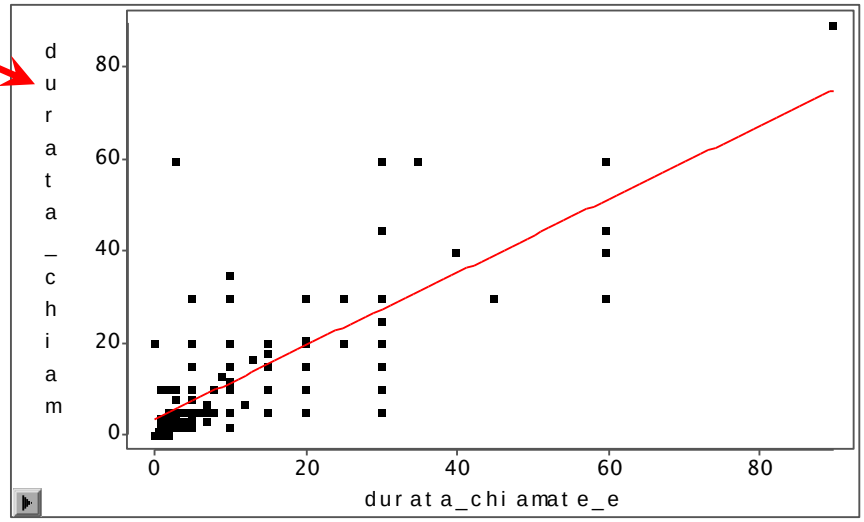
- durata media delle chiamate ricevute [durata_chiamate_r]
- numero medio di ore di utilizzo del telefono cellulare al giorno [cell_h]
- numero medio di ore di utilizzo del telefono fisso al giorno [fisso_h]

```
proc corr data=corso.telefonia;

var durata_chiamate_e durata_chiamate_r

cell_h fisso_h;

run;
```

Output PROC CORR - Esempio

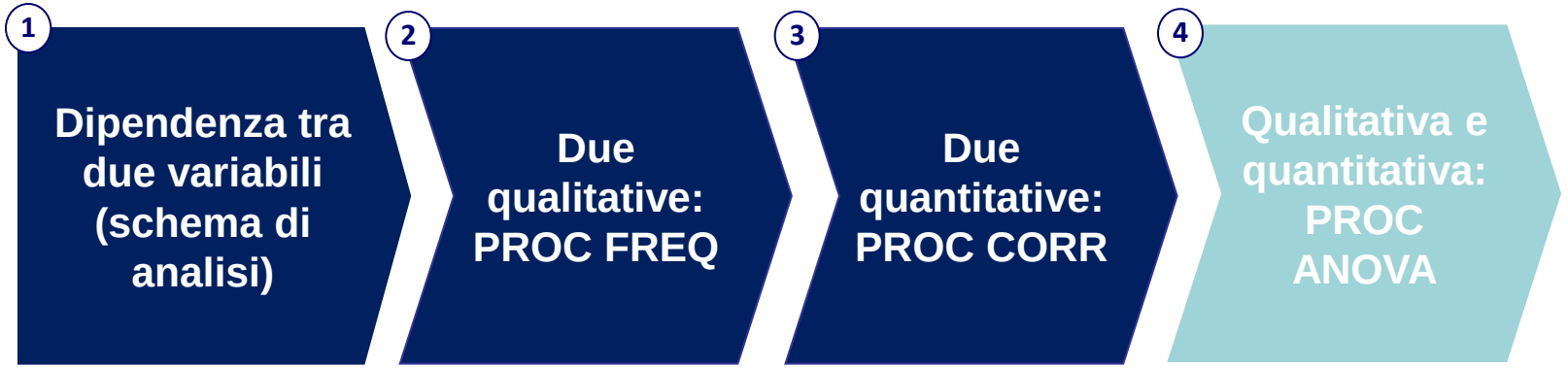
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	durata_chiamate_e	durata_chiamate_r	cell_h	fisso_h
durata_chiamate_e durata_chiamate_e	1 236	0.78645 <.0001 236	0.23099 0.0003 236	0.24568 0.0003 208
durata_chiamate_r durata_chiamate_r	0.78645 <.0001 236	1	0.31703	0.27686
cell_h cell_h	0.23099 0.0003 236			
fisso_h fisso_h	0.24568 0.0003 208			



Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Riepilogo teorico (1/3)

X variabile qualitativa e Y variabili 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 (2/3)

X variabile qualitativa e Y variabili 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 (3/3)

X variabile qualitativa e Y variabili 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

η^2 è sempre compreso tra 0 e 1.

PROC ANOVA – Sintassi generale

Sia Y una variabile quantitativa e X una variabile qualitativa

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

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;
```

Esempio: Output

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

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 libreria ESER4, in modo che punti alla cartella fisica dove è contenuta la tabella SAS «DENTI_NEW»
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.