Analisi Univariata
&
Esercizi

Metodi Quantitativi per Economia, Finanza e Management

Esercitazione n°3

Lavoro di Gruppo

- Inviare **entro oggi 16/10/2015**, a gmagistrelli@liuc.it e gdeppieri@liuc.it:
  - nome, cognome e numero di matricola dei partecipanti (min 3 – max 4 componenti per gruppo)
  - nome del gruppo
  - titolo/argomento del lavoro di gruppo

- entro **30/10/2015** inviare via e-mail il questionario da validare

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

SAS on Demand for Academics

Nella sezione **Varie** della pagina di insegnamento sono pubblicate le istruzioni di utilizzo di SAS on Demand for Academics:

- **SASOnDemandForAcademics_registrazione_1516**
  Procedure di registrazione e accesso a SAS on Demand

- **SASOnDemandForAcademics_SASStudio_1516**
  Breve manuale di utilizzo di SAS Studio (corrispettivo web del tool utilizzato a lezione) predisposto per gli obiettivi del corso

- **SASOnDemandForAcademics_ELearnings_1516**
  Illustrazione delle procedure di attivazione dei corsi
  SAS E-Learnings

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\esercitazione3*** e copiare il contenuto nella cartella C:\corso

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

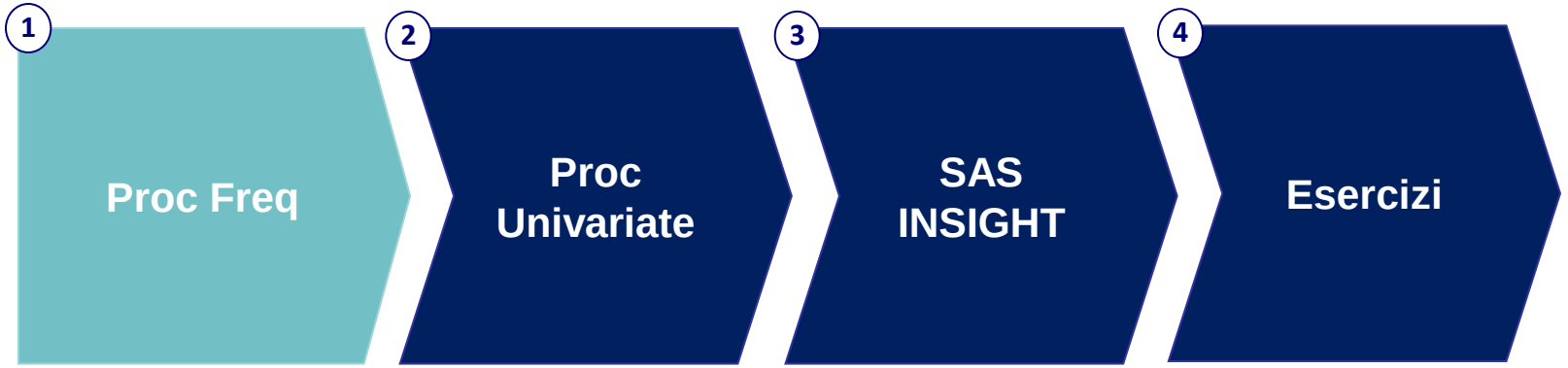
- Allocare la libreria ***corso,*** puntando il percorso fisico C:\corso, utilizzando l'istruzione:

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

- Nella libreria dovreste visualizzare la tabella TELEFONIA

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:



Analisi Univariata: Procedure SAS

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

Procedure SAS per l'analisi univariata di una variabile:

PROC SAS	TIPO VARIABILE	FUNZIONE
PROC FREQ	Variabili qualitative o quantitative discrete	Distribuzione di frequenze (frequenze assolute, relative e cumulate)
PROC UNIVARIATE	Variabili quantitative	Calcolo misure di sintesi di tipo univariato

PROC FREQ – Sintassi generale 1/2

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

```
proc freq data= dataset;
   tables variabile /options;
run;
```

OPTIONS:

- /missing considera anche i missing nel calcolo delle frequenze

PROC FREQ: Esempio 1

Variabile qualitativa: operatore telefonico

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

Output PROC FREQ

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

***Frequenza relativa*:** rapporto tra la frequenza assoluta ed il numero complessivo delle osservazioni effettuate

Frequenze cumulate

operatore	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Tim	55	23.31	55	23.31
Tre	12	5.08	67	28.39
Vodafone	154	65.25	221	93.64
Wind	15	6.36	236	100

PROC FREQ: Esempio 2

Variabile quantitativa discreta:

numero medio giorni utilizzo alla settimana telefono fisso

```
proc freq data=corso.telefonia;
tables fisso_g;
run;
```

Output PROC FREQ

fisso_g				
fisso_g	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	27	11.44	27	11.44
0.5	9	3.81	36	15.25
1	10	4.24	46	19.49
2	19	8.05	65	27.54
3	21	8.90	86	36.44
4	14	5.93	100	42.37
5	19	8.05	119	50.42
6	9	3.81	128	54.24
7	108	45.76	236	100.00

Fare attenzione al numero di modalità della variabile

PROC FREQ: Esempio 3

Variabile qualitativa:

secondo motivo di utilizzo mezzi di comunicazione

```
proc freq data=corso.telefonia;
tables motivo_utilizzo_2 / missing;
run;
```

OPZIONE missing: considera anche i missing nel calcolo delle frequenze

Output PROC FREQ

MISSING

Output con l'utilizzo dell'opzione MISSING

motivo_utilizzo_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	24	10.17	24	10.17
Altro	2	0.85	26	11.02
Famigliari	40	16.95	66	27.97
Partner	22	9.32	88	37.29
Piacere/Tempo libero	128	54.24	216	91.53
Studio	20	8.47	236	100.00

Output senza l'utilizzo dell'opzione MISSING

motivo_utilizzo_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Altro	2	0.94	2	0.94
Famigliari	40	18.87	42	19.81
Partner	22	10.38	64	30.19
Piacere/Tempo libero	128	60.38	192	90.57
Studio	20	9.43	212	100.00

Frequency Missing = 24

PROC FREQ – Sintassi generale 2/2

Distribuzione di frequenza univariata con variabile di classificazione

```
proc freq data = dataset;
   by variabile_1;
   tables variabile_2  / options;
run;
```

NOTA BENE: è necessario ordinare il dataset secondo la variabile di classificazione PRIMA di eseguire la PROC FREQ!

PROC FREQ: Esempio 4

Distribuzione di frequenza univariata con variabile di classificazione

PROC SORT: ordinare le osservazioni in base alla variabile di by

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

Output PROC FREQ

sesso=F

operatore	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Tim	27	27.00	27	27.00
Tre	7	7.00	34	34.00
Vodafone	63	63.00	97	97.00
Wind	3	3.00	100	100.00

sesso=M

operatore	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Tim	28	20.59	28	20.59
Tre	5	3.68	33	24.26
Vodafone	91	66.91	124	91.18
Wind	12	8.82	136	100.00

Metodi Quantitativi per Economia, Finanza e Management

Obiettivi di questa esercitazione:

Analisi Univariata: Procedure SAS

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

Procedure SAS per l'analisi univariata di una variabile:

PROC SAS	TIPO VARIABILE	FUNZIONE
PROC FREQ	Variabili qualitative o quantitative discrete	Distribuzione di frequenze (frequenze assolute, relative e cumulate)
PROC UNIVARIATE	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 dispersione:

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

Misure di forma della distribuzione:

- Skewness
- Kurtosis

PROC UNIVARIATE – Sintassi (1/2)

La PROC UNIVARIATE permette di calcolare per variabili **quantitative** misure di sintesi:

- di posizione
- di variabilità
- di forma della distribuzione

```
proc univariate data= dataset;
   var variabile;
run;
```

PROC UNIVARIATE – Esempio 1

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

```
proc univariate data=corso.telefonia;

var num_sms_e;

run;
```

Output PROC UNIVARIATE (1/7)

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

Basic Statistical Measures			
Location		**Variability**	
Mean	24.31356	**Std Deviation**	28.46175
Median	10.00000	**Variance**	810.07147
Mode	10.00000	**Range**	100.00000
		Interquartile Range	25.00000

Output PROC UNIVARIATE (2/7)

Misure di Variabilità

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

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

- ***Scarto Quadratico Medio*** [Std Deviation]:
  mostra la variabilità rispetto alla media (radice quadrata della varianza)

Basic Statistical Measures			
Location		**Variability**	
Mean	24.31356	**Std Deviation**	28.46175
Median	10.00000	**Variance**	810.07147
Mode	10.00000	**Range**	100.00000
		Interquartile Range	25.00000

Output PROC UNIVARIATE (3/7)

Misure di Variabilità

- ***Campo di Variazione*** [Range]:
  differenza tra il massimo e il minimo dei valori osservati

- ***Differenza Interquartile*** [Interquartile Range]:
  3° quartile – 1° quartile

Basic Statistical Measures			
Location		**Variability**	
Mean	24.31356	**Std Deviation**	28.46175
Median	10.00000	**Variance**	810.07147
Mode	10.00000	**Range**	100.00000
		Interquartile Range	25.00000

Output PROC UNIVARIATE (4/7)

Quantiles (Definition 5)	
Quantile	**Estimate**
100% Max	100
99%	100
95%	100
90%	70
75% Q3	30
50% Median	10
25% Q1	5
10%	2
5%	2
1%	1
0% Min	0

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

Output PROC UNIVARIATE (5/7)

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

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

Moments			
N	236	**Sum Weights**	236
Mean	24.3135593	**Sum Observations**	5738
Std Deviation	28.4617546	**Variance**	810.071475
Skewness	1.59619131	**Kurtosis**	1.44200254
Uncorrected SS	329878	**Corrected SS**	190366.797
Coeff Variation	117.061242	**Std Error Mean**	1.85270242

Output PROC UNIVARIATE (6/7)

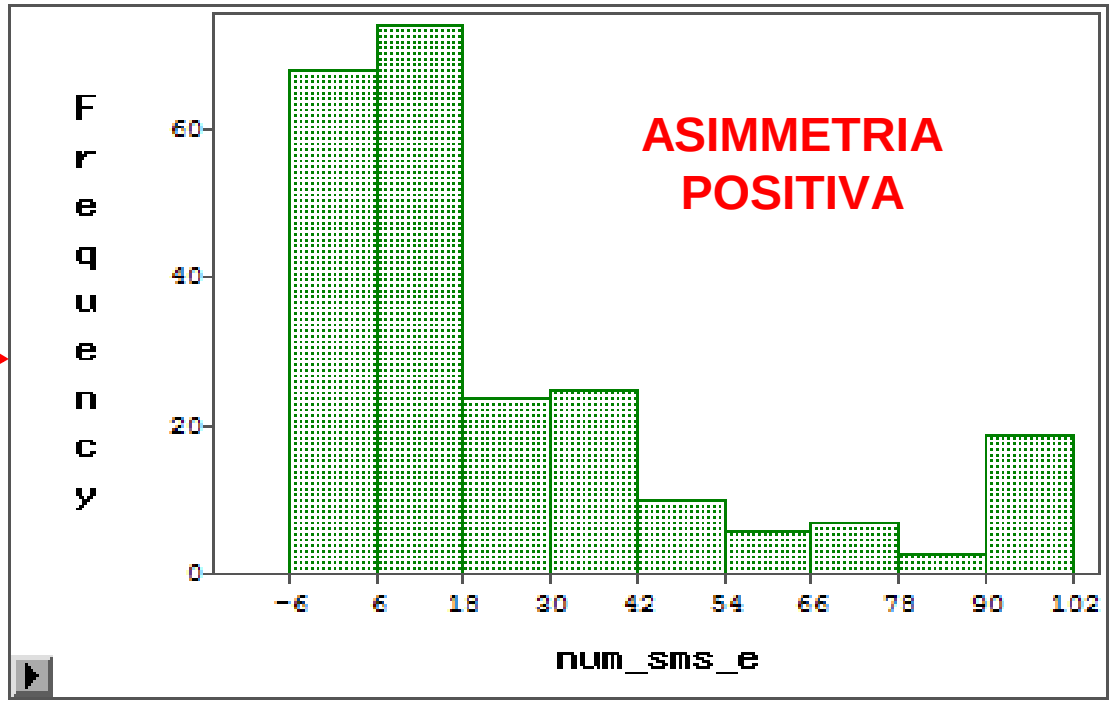
Misure di Forma della Distribuzione

- ***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)
- ***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

Moments			
N	236	**Sum Weights**	236
Mean	24.3135593	**Sum Observations**	5738
Std Deviation	28.4617546	**Variance**	810.071475
Skewness	1.59619131	**Kurtosis**	1.44200254
Uncorrected SS	329878	**Corrected SS**	190366.797
Coeff Variation	117.061242	**Std Error Mean**	1.85270242

Output PROC UNIVARIATE (7/7)

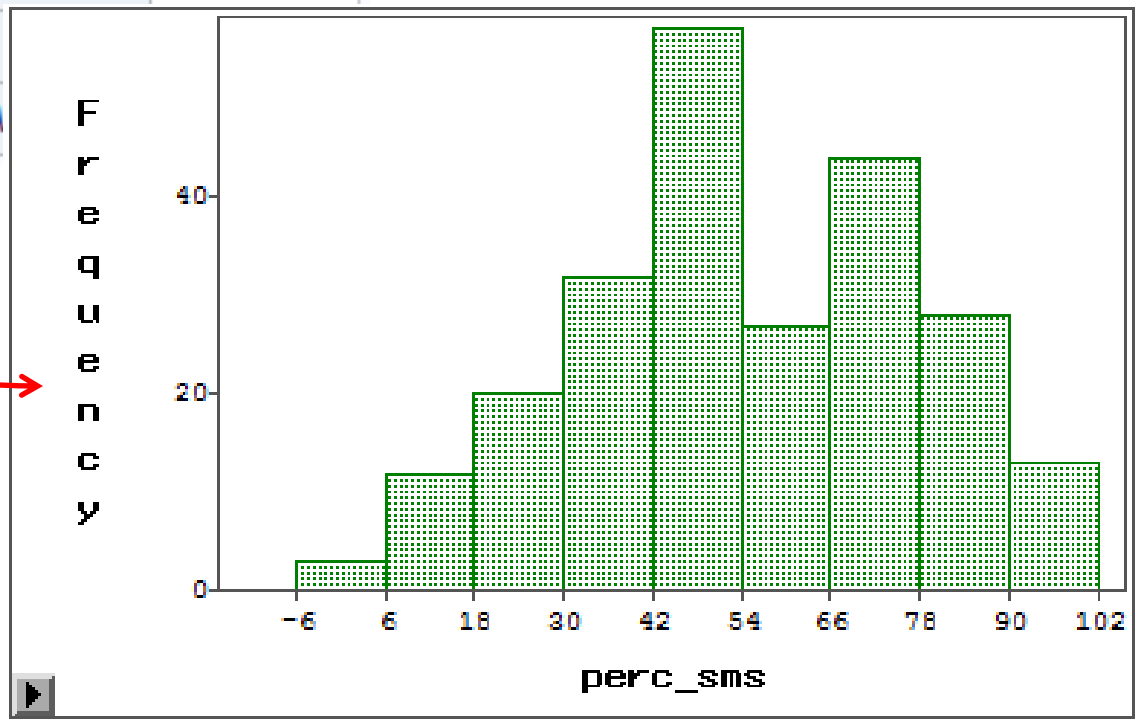
Moments			
N	236	**Sum Weights**	236
Mean	24.3135593	**Sum Observations**	5738
Std Deviation	28.4617546	**Variance**	810.071475
Skewness	1.59619131	**Kurtosis**	1.44200254
Uncorrected SS	329878	**Corrected SS**	190366.797
Coeff Variation	117.061242	**Std Error Mean**	1.85270242



Skewness: altro esempio

Variabile PERC_SMS del dataset TELEFONIA

Moments			
N	236	Sum Weights	236
Mean	53.7224576	Sum Observations	12678.5
Std Deviation	22.667539	Variance	513.817323
Skewness	-0.2791588	Kurtosis	-0.6556038
Uncorrected SS	801867.25	Corrected	
Coeff Variation	42.1937863	Std Error M	



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

PROC UNIVARIATE – Esempio 2

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

```
proc univariate data=corso.telefonia;
var cell_h;
run;
```

PROC UNIVARIATE – Sintassi 2/2

Distribuzione di frequenza univariata con variabile di classificazione

```
proc univariate data= dataset;
   class variabile_1 (options);
   var variabile_2;
run;
```

OPTIONS:

- (missing) considera anche la categoria “missing” (contenente tutti i valori mancanti) della variabile di classificazione

PROC UNIVARIATE – Esempio 3

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

```
proc univariate data=corso.telefonia;
class sesso;
var cell_h;
run;
```

PROC UNIVARIATE – Esempio 4

Misure di sintesi della variabile:
numero medio ore utilizzo al giorno telefono cellulare suddivisa per hobby con opzione “missing”

```
proc univariate data=corso.telefonia;
class hobby_3(missing);
var cell_h;
run;
```

Metodi Quantitativi per Economia, Finanza e Management

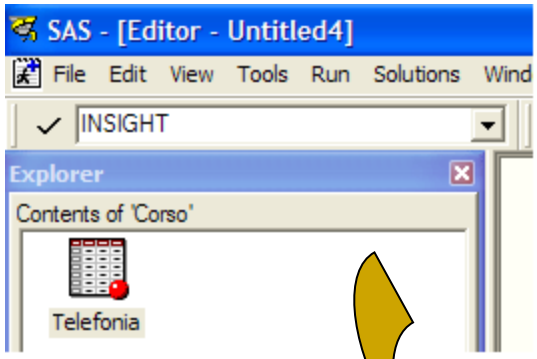
Obiettivi di questa esercitazione:

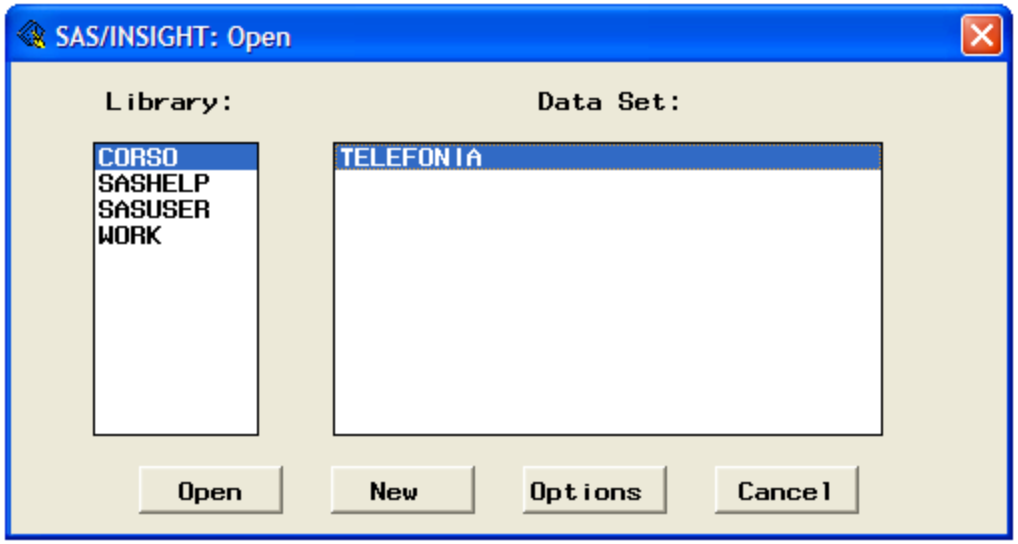
Analisi Univariata: SAS INSIGHT

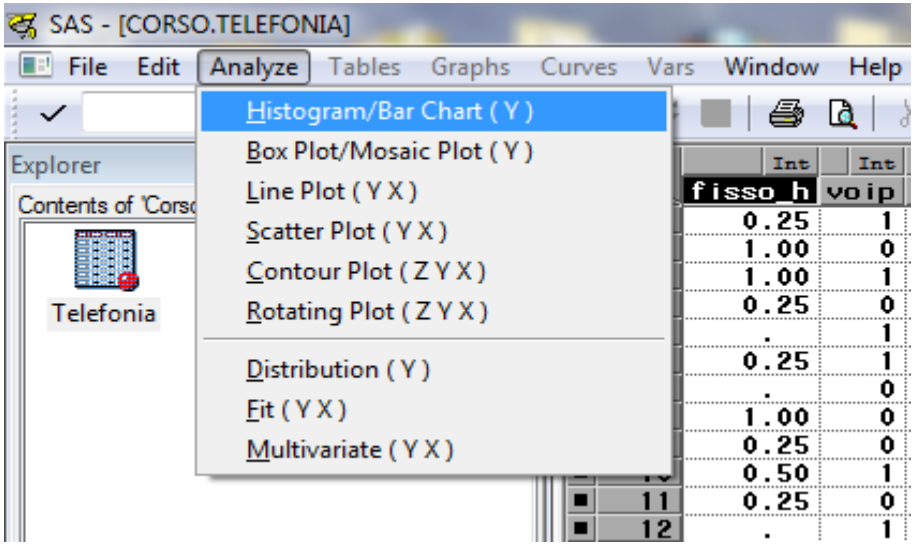
Rappresentazioni grafiche del modulo SAS INSIGHT per l'analisi univariata di una variabile:

SAS INSIGHT	TIPO VARIABILE	FUNZIONE
HISTOGRAM / BAR CHART	Sia variabili qualitative che quantitative	Istogramma (variabili numeriche) Bar chart o diagramma a barre (variabili alfanumeriche)
BOX PLOT	Solo per variabili quantitative	Rappresentazione grafica di alcune misure di sintesi

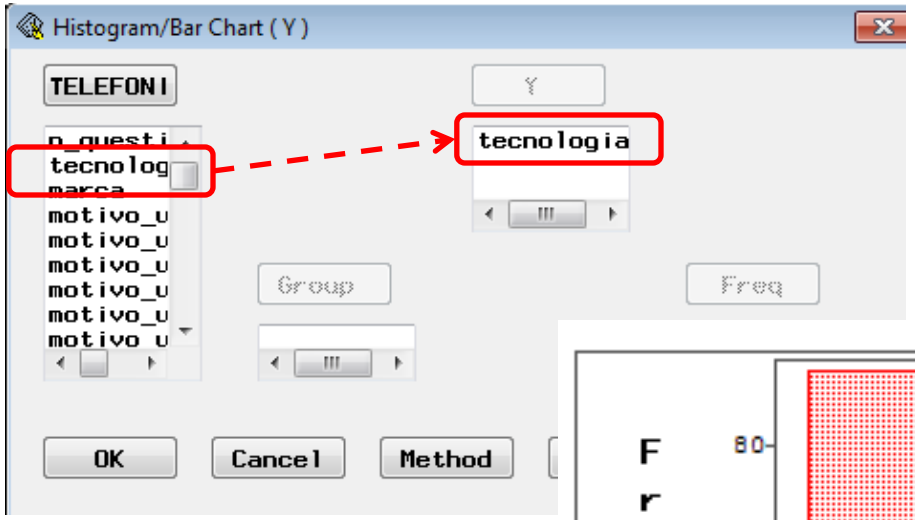
SAS INSIGHT: Histogram/Bar chart

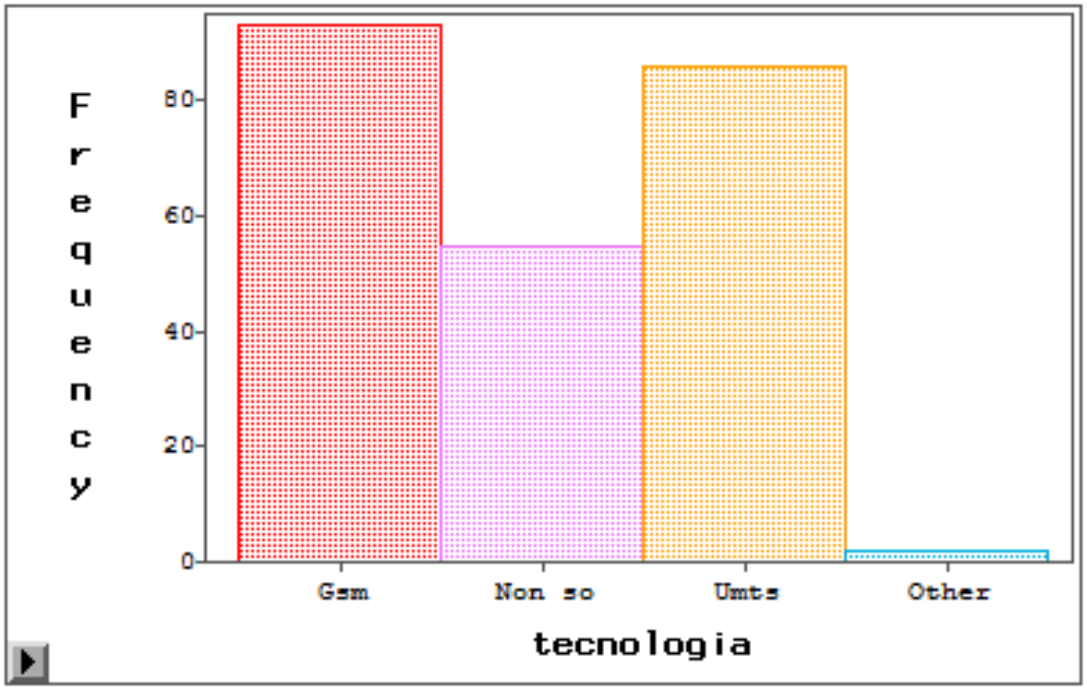




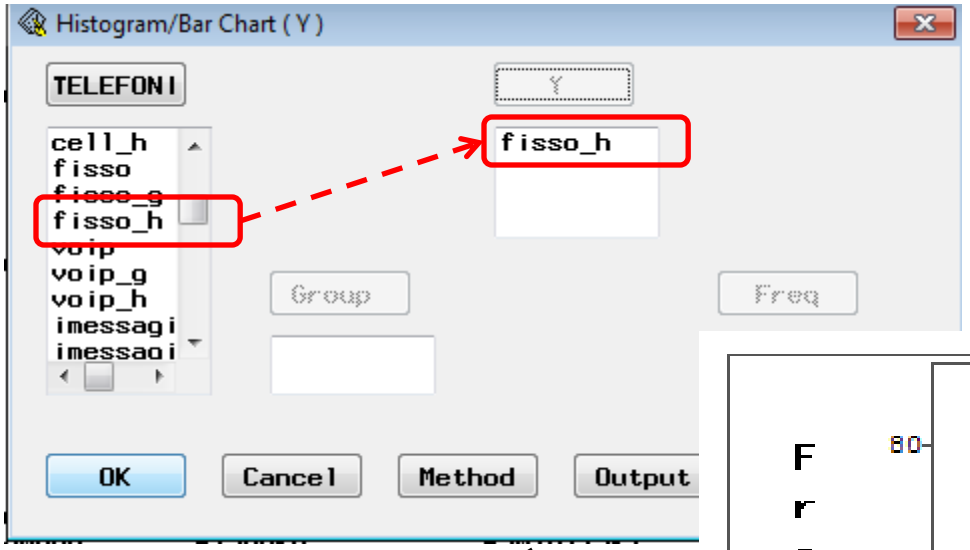


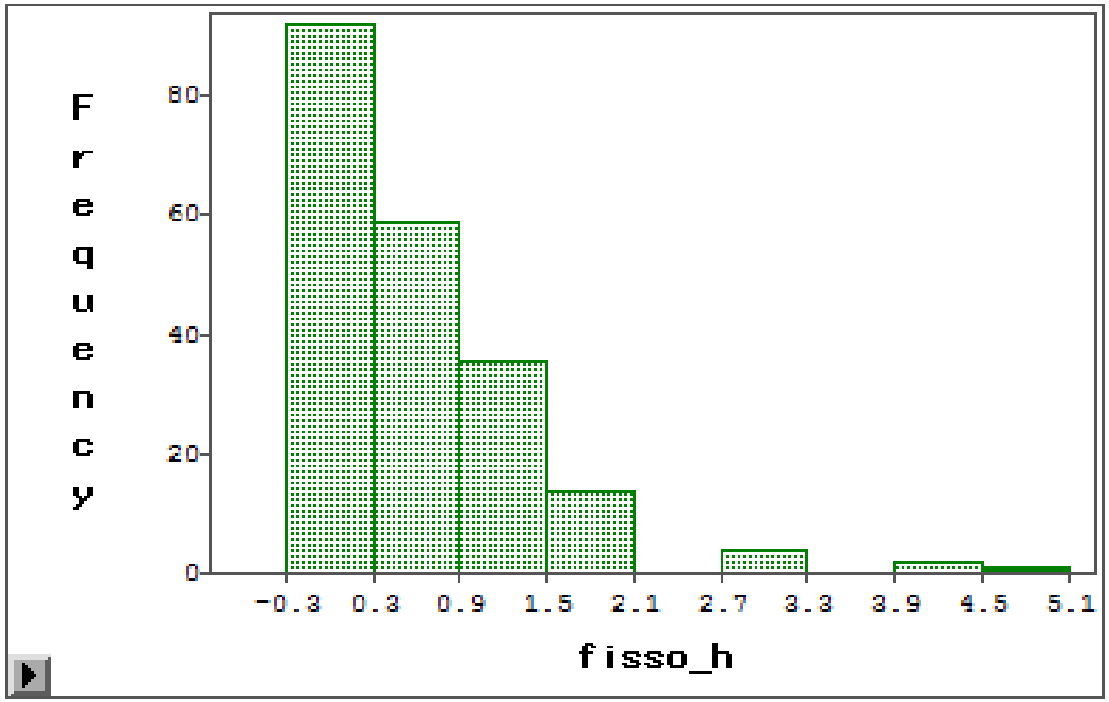
SAS INSIGHT: Histogram/Bar chart
Esempio 1





SAS INSIGHT: Histogram/Bar chart
Esempio 2



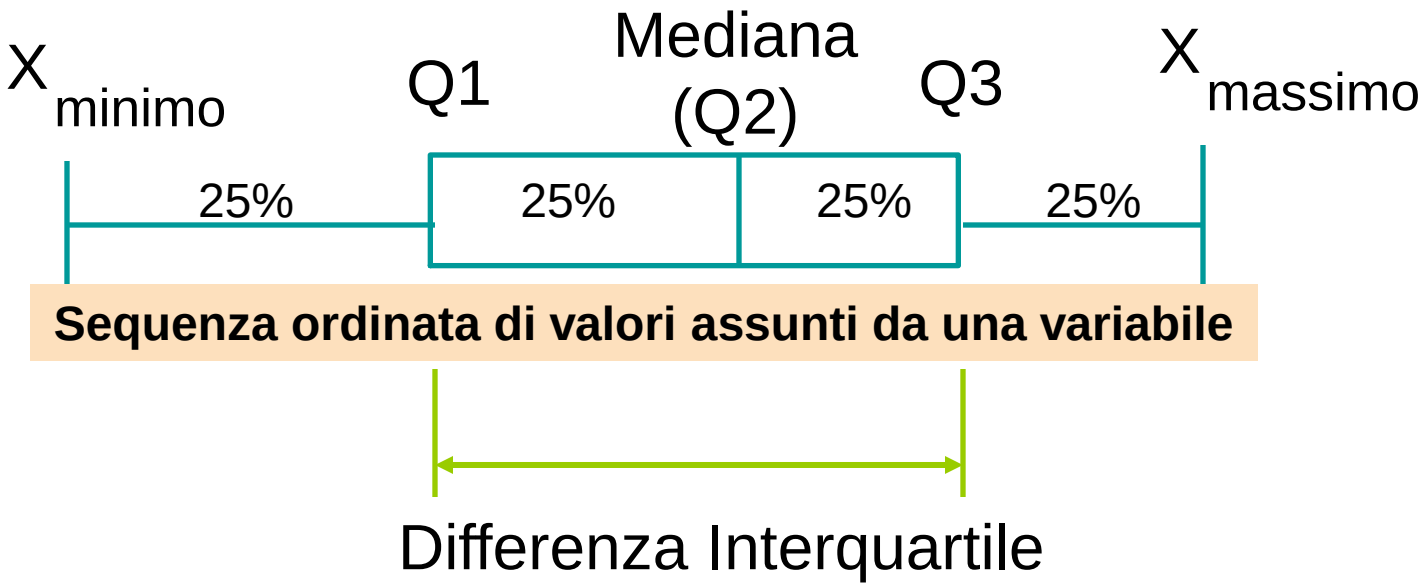


Analisi Univariata: SAS INSIGHT

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

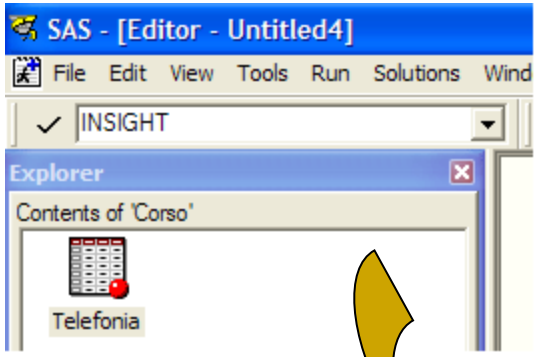
SAS INSIGHT	TIPO VARIABILE	FUNZIONE
HISTOGRAM / BAR CHART	Sia variabili qualitative che quantitative	Istogramma (variabili numeriche) Bar chart o diagramma a barre (variabili alfanumeriche)
BOX PLOT	Solo per variabili quantitative	Rappresentazione grafica di alcune misure di sintesi

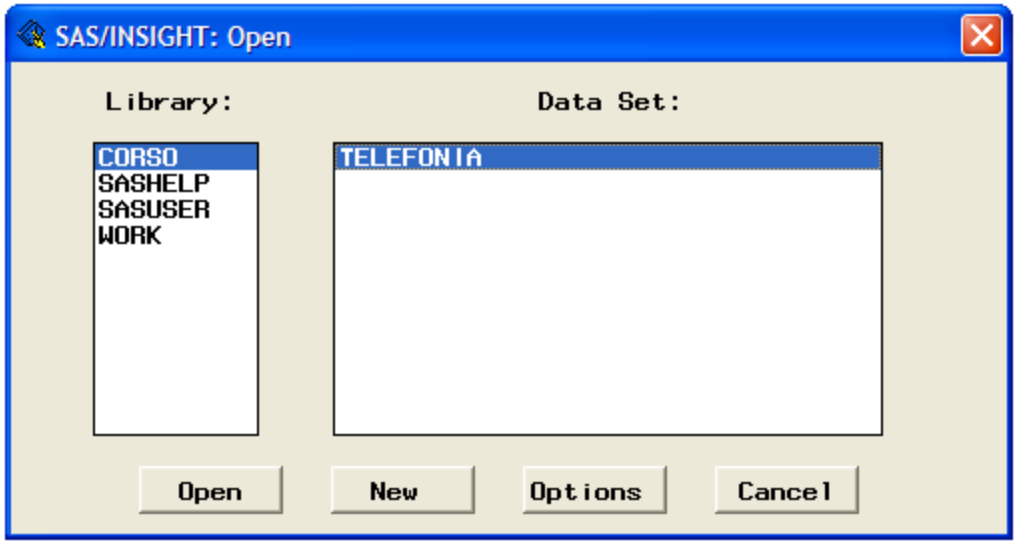
SAS INSIGHT: Box Plot (1/3)

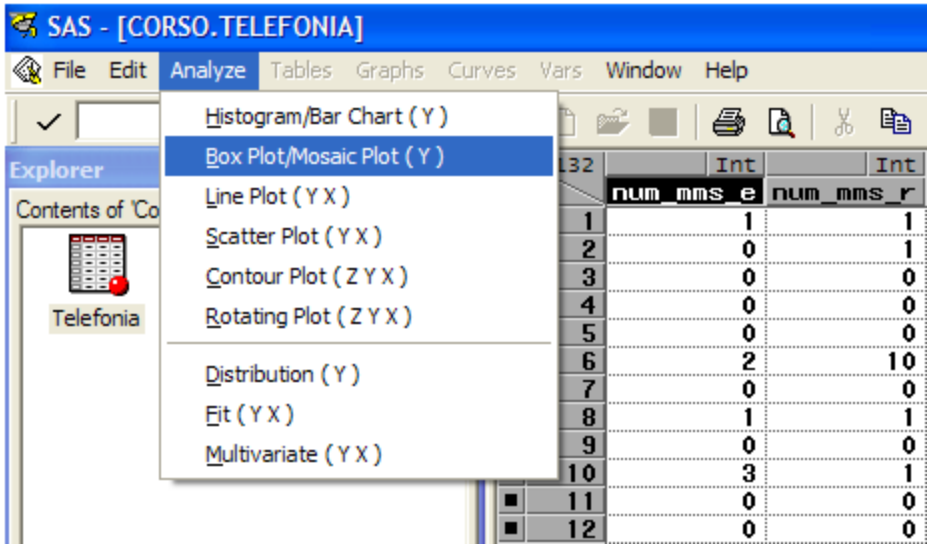


OUTLIERS:	Q1 - 1,5 * Differenza interquartile
	Q3 + 1,5 * Differenza interquartile

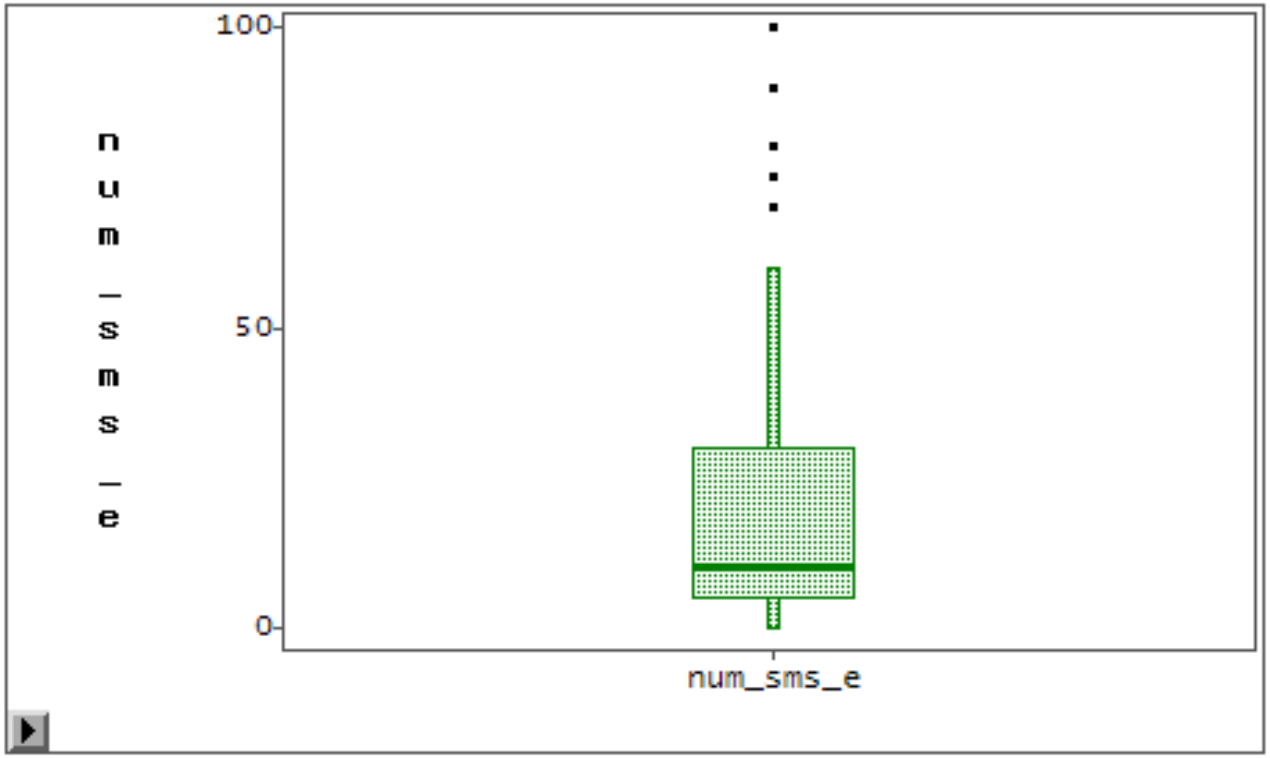
SAS INSIGHT: Box Plot (2/3)





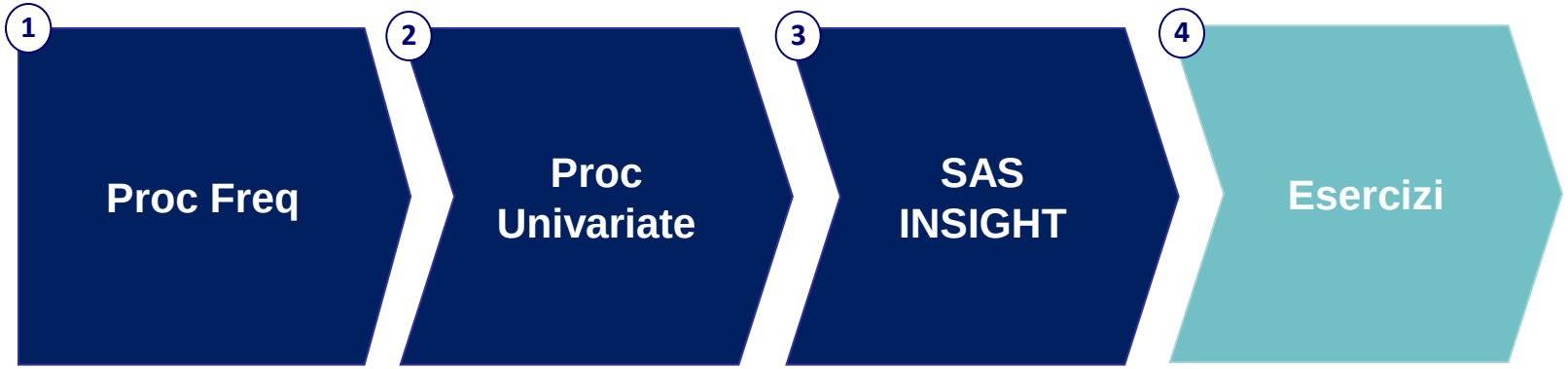


SAS INSIGHT: Box Plot (3/3)



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 libreria ESER3 (che punta alla cartella che contiene il file DENTI.XLS)

2. Importare in formato SAS la tabella Excel DENTI.XLS e chiamarla 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 procedura 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. Determinare la percentuale di clienti che hanno ricevuto meno di 11 contatti pubblicitari

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