Elementi di Statistica Descrittiva

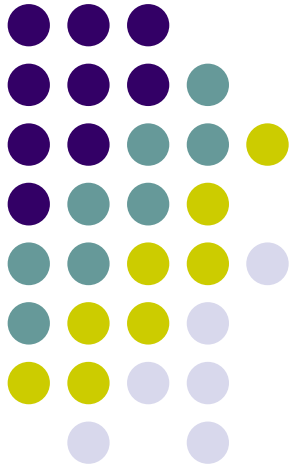
Misure di centralità

V Scuola Estiva AISV

La statistica come strumento di analisi nelle scienze umanistiche e comportamentali

Soriano nel Cimino (VT), 5 Ottobre 2009

Pier Francesco Perri
Dipartimento di Economia e Statistica - UNICAL
pierfrancesco.perri@unical.it



Che cos'è la Statistica?

Sai ched'è la statistica? E' 'na cosa che serve pe' fa' un conto in generale de la gente che nasce, che sta male, che more, che va in carcere e che sposa (Trilussa)

- Insieme di descrizioni numeriche di determinati fenomeni
  - ***statistiche economiche***
  - ***statistiche culturali***
  - ***statistiche del turismo***
  - ***statistiche del commercio***
  - ***statistiche meteorologiche***
  - ***statistiche del commercio***
  - ***statistiche mediche***
  - ***statistiche della popolazione***

Che cos'è la Statistica?

- Disciplina che fornisce una metodologia per:
  - *raccolta*
  - *classificazione*
  - *sintesi*
  - *analisi*
  - *interpretazione*

  dei dati osservati nelle scienze empiriche

Un metodo per ricavare informazione fruibile a partire da una mole di dati

Ambiti

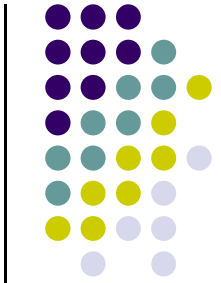
- Economia
- Marketing
- Ricerche di Mercato
- Finanza
- Fisica
- Genetica
- Medicina
- Fonetica
- Psicologia
- Giurisprudenza
- Studi storiografici e letterari

Supporto alle decisioni
in condizioni di
incertezza

L'Indagine Statistica

L'indagine statistica è un processo di acquisizione di informazioni su un fenomeno oggetto di studio attraverso la conoscenza delle sue manifestazione su un collettivo di unità statistiche

Fasi di un'indagine



- Progettazione
  - definizione delle finalità conoscitive dello studio
  - individuazione della popolazione oggetto di studio
  - creazione di un questionario
- Rilevazione
  - somministrazione del questionario
- Codifica dei dati
- Elaborazione
  - spoglio dei dati (accorpamento dei casi simili)
  - sintesi tabellari e grafiche
  - costruzione di indici
- Validazione dei risultati
- Diffusione dei risultati

Qualche definizione

- **Unità**
  sono le entità reali (individui, oggetti, aziende, ecc.) oggetto di studio

- **Popolazione o Collettivo**
  l'insieme delle unità statistiche

- **Caratteri o variabili**
  aspetti rilevati in corrispondenza di ciascuna unità statistica che si ritengono rilevanti ai fini dell'indagine

- **Modalità**
  diverse manifestazioni che un carattere presenta sulle unità statistiche.

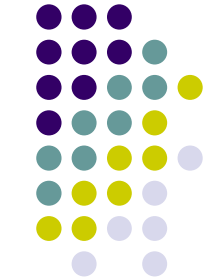
Classificazione dei caratteri

- **Qualitativi:** esprimibili tramite sostantivi, avverbi, aggettivi
  - ✓ Nominali o sconnessi: non è possibile individuare un ragionevole criterio di ordinamento tra le modalità (categorie). E' possibile solo confrontare tra di loro le categorie e stabilire se sono uguali o diverse (*stato civile, religione professata, categorie grammaticali, tipologia di opera letteraria, ecc.*)
  - ✓ **Ordinali:** è possibile istituire un ordinamento logico (crescente o decrescente) tra le modalità (*livello di istruzione, posizione occupata in una graduatoria, livello di abilità, ecc.*)

- **Quantitativi:** esprimibili tramite valori numerici
  - ✓ **Discreti:** rilevabili tramite un conteggio (*numero di parole per documento, numero di risposte corrette, numero di disturbi dell'apprendimento, ecc.*)
  - ✓ **Continui:** sono quei caratteri le cui modalità possono variare per quantità piccole a piacere. Generalmente, la rilevazione avviene tramite uno **strumento di misura**. Le modalità possono assumere valori su tutto l'insieme dei numeri reali, ma all'atto pratico, vengono discretizzate a causa della taratura dello strumento di misurazione (*ritardo nell'attacco della sonorità (VOT), la frequenza fondamentale (F0), ecc.*)

La metodologia statistica prevede strumenti diversi a seconda della tipologia dei caratteri considerati

L'organizzazione dei dati



- Il processo di rilevazione dei dati si realizza usualmente tramite la compilazione di questionari o schede

- Per ogni unità statistica si dispone, in generale, di una mole di informazioni che occorre organizzare sistematicamente al fine di renderne agevole l'elaborazione

- I dati acquisiti vengono archiviati sotto forma di *database.* Un *database* può essere assimilato ad una tabella formata da *R* righe e *C* colonne
  - ogni **riga** riporta le informazioni alfa/numeriche riferite alla singola unità statistica
  - ogni **colonna** riporta i valori delle variabili statistiche osservati sulle diverse unità statistiche

La sintesi dei dati

Il database è un serbatoio "senza fondo" di informazioni grezze e non immediatamente fruibili.

Al fine di cogliere gli aspetti più rilevanti del fenomeno oggetto di studio occorre organizzare i dati in maniera sintetica.

COME?

Accorpando in classi omogenee le diverse modalità del carattere e associando ad ognuna di essa il numero di volte che è stata rilevata sulle unità statistiche

Distribuzioni univariate

La *ripetizione di non-parole* è una prova con cui si può valutare la capacità di un bambino di analizzare il segnale acustico, costruire un piano articolatorio e immagazzinarlo nella memoria fonologica a breve termine.

Essa rappresenta una misura della memoria fonologica a breve termine.

La prova consiste nel chiedere al bambino di ripetere, il più fedelmente possibile, delle non-parole ascoltate tramite un registratore (esempio: deccarello < pennarello).

Distribuzioni univariate

Esempio: numero di non-parole prodotte erroneamente (X) da un gruppo di 20 bambini in età scolare

Dati

4, 3, 4, 3, 3, 1, 3, 1, 4,
2, 2, 3, 3, 4, 2, 3, 4, 4,
1, 3 **(elenco di modalità)**

Distribuzione di frequenza univariata

X	*Frequenze Assolute (n)*	*Frequenze Relative (f)*
1	3	0.15
2	3	0.15
3	8	0.40
4	6	0.30
Totale	**20**	**1**

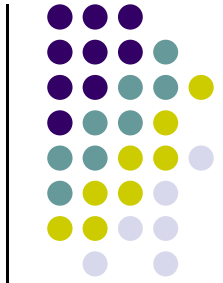
freq. assolute: numero di volte che nella popolazione è stata osservata una determinata modalità

freq. relative =

freq. ass. / num. collettivo

- non dipendono dalla numerosità del collettivo
- consentono di valutare l'importanza di ogni modalità

Distribuzioni univariate



Distribuzione di frequenza univariata

X	*Freq. Assolute (n)*	*Freq. Relative (f)*	*Freq. Cumulate Assolute (N)*	*Freq. Cumulate Relative (F)*
1	3	0.15	3	0.15
2	3	0.15	6	0.30
3	8	0.40	14	0.70
4	6	0.30	20	1
Totale	**20**	**1**		

Se moltiplichiamo le frequenze relative, otteniamo le frequenze percentuali

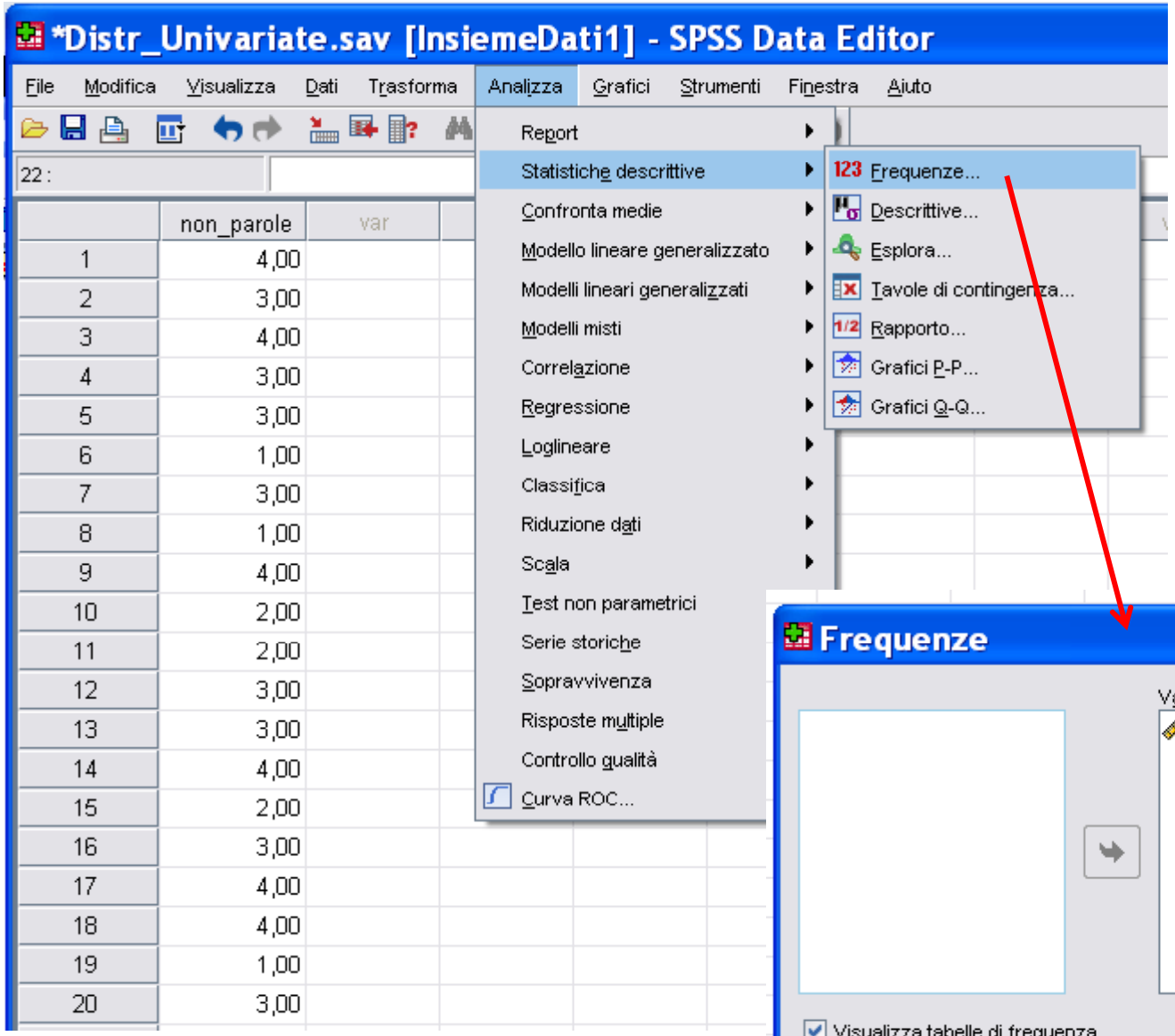
Distribuzioni univariate

In generale, una distribuzione di frequenze per un carattere con *k* **modalità distinte** si presenta nella forma:

X	n
x_1	n_1
x_2	n_2
...	
x_i	n_i
...	
x_k	n_k
Tot.	**n**

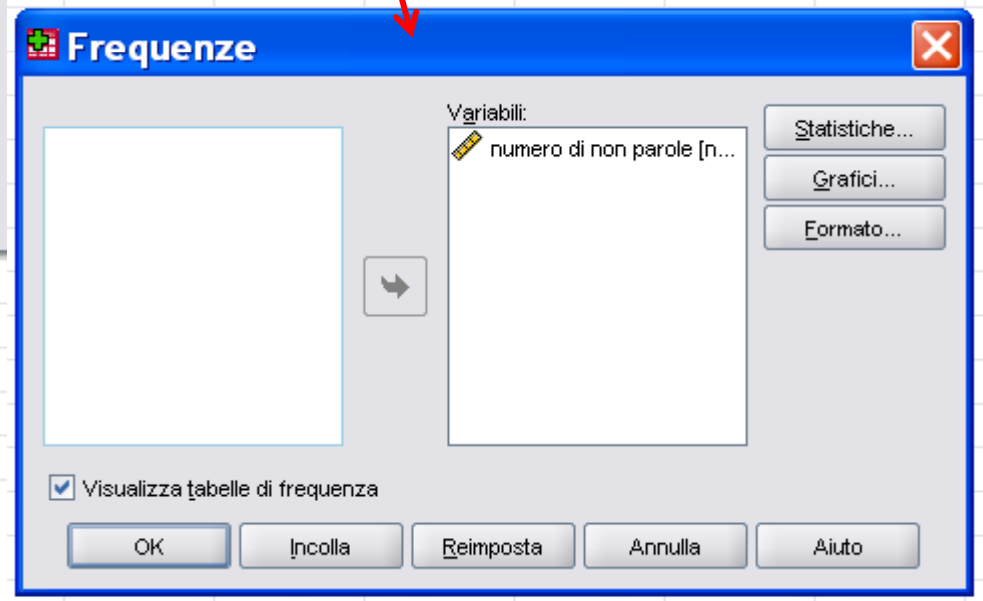
i-esima modalità del carattere → x_i

Numerosità del collettivo → **n**


*Distr_Univariate.sav [InsiemeDati1] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
22 :
non_parole
var
1 4,00
2 3,00
3 4,00
4 3,00
5 3,00
6 1,00
7 3,00
8 1,00
9 4,00
10 2,00
11 2,00
12 3,00
13 3,00
14 4,00
15 2,00
16 3,00
17 4,00
18 4,00
19 1,00
20 3,00
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Frequenze...
Descrittive...
Esplora...
Tavole di contingenza...
Rapporto...
Grafici P-P...
Grafici Q-Q...

Dati


SPSS


Frequenze
Variabili:
numero di non parole [n...
Statistiche...
Grafici...
Formato...
Visualizza tabelle di frequenza
OK
Incolla
Reimposta
Annulla
Aiuto

Distribuzioni in classi

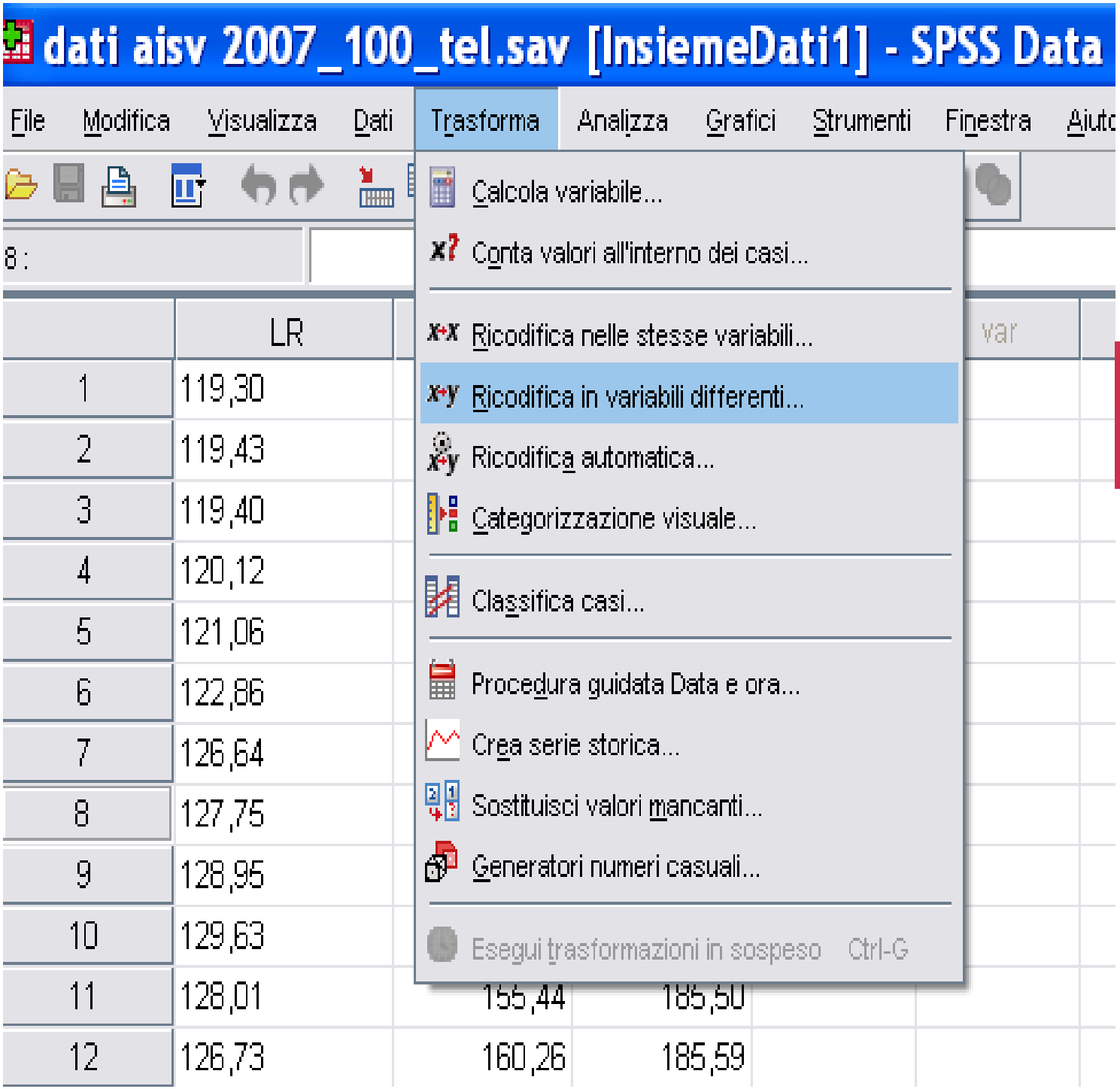
In presenza di caratteri quantitativi continui o discreti con numerose modalità occorre creare classi di modalità

Esempio: Frequenza Fondamentale (F0 per telefono)

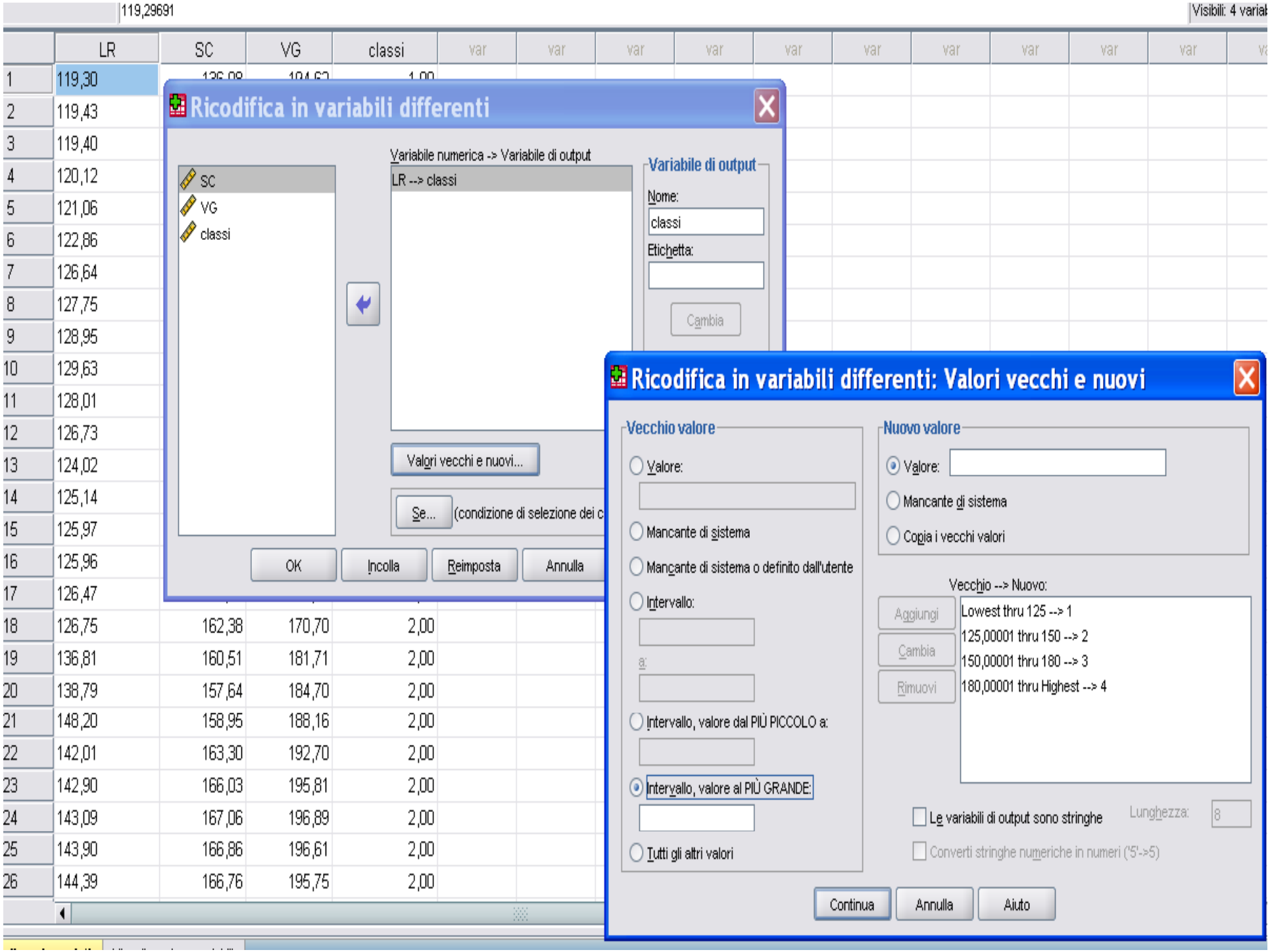
F0	n	f%	N	F%
110 –| 125	7	7.1%	7	7.1%
125 –| 150	43	43.9%	50	51.0%
150 –| 180	27	27.6%	77	78.6%
180 –| 200	21	21.4%	98	100%
Totale	**98**	**100**		

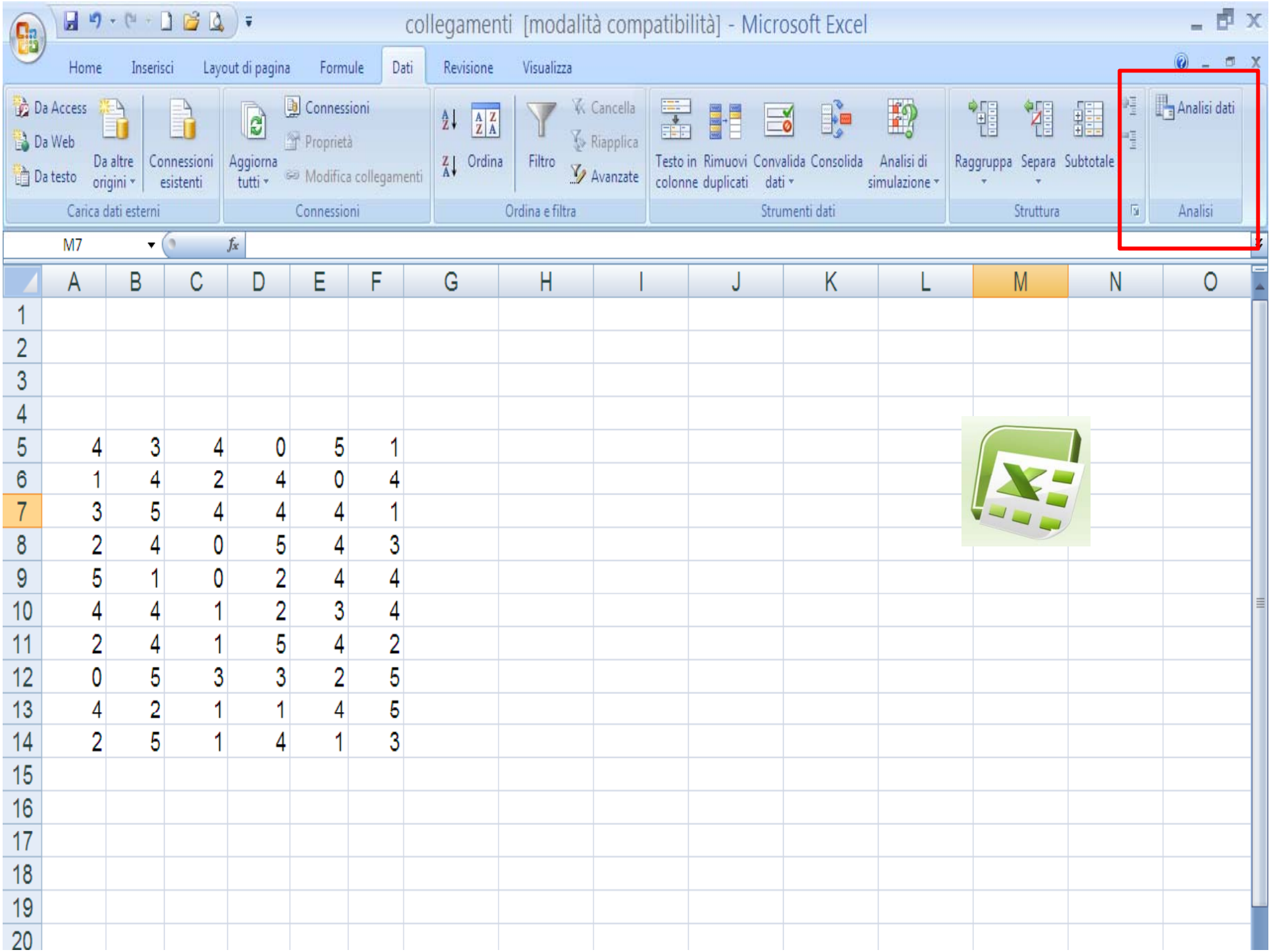
SPSS Dati

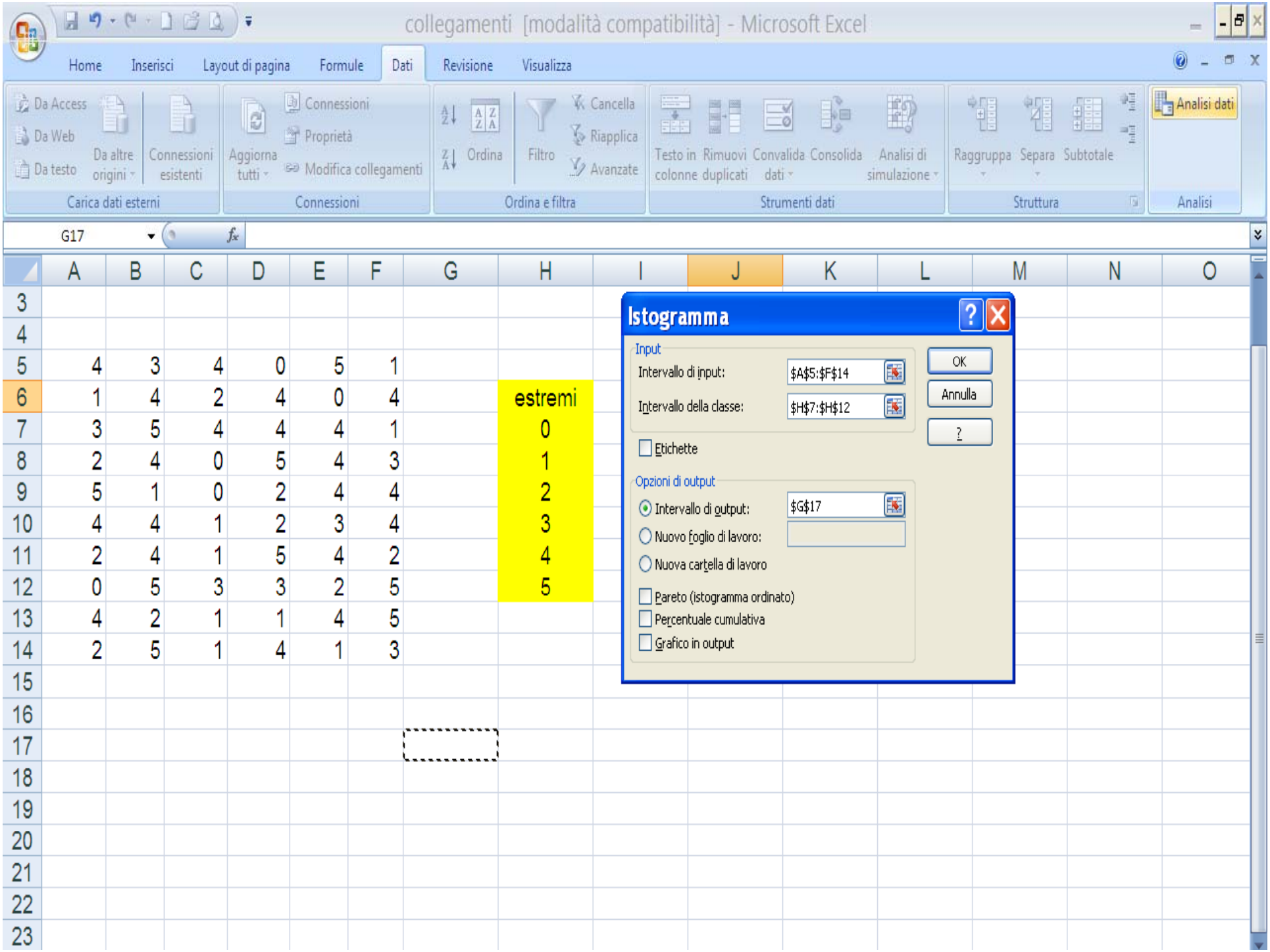
- ✓ Classi esaustive e disgiunte
- ✓ Quante classi definire?
- ✓ Quale ampiezza adottare?
- ✓ Come dimensionare la due classi estreme?


dati aisv 2007_100_tel.sav [InsiemeDati1] - SPSS Data
File
Modifica
Visualizza
Dati
Trasforma
Analizza
Grafici
Strumenti
Finestra
Aiuto
8 :
Calcola variabile...
Conta valori all'interno dei casi...
Ricodifica nelle stesse variabili...
Ricodifica in variabili differenti...
Ricodifica automatica...
Categorizzazione visuale...
Classifica casi...
Procedura guidata Data e ora...
Crea serie storica...
Sostituisci valori mancanti...
Generatori numeri casuali...
Esegui trasformazioni in sospeso Ctrl-G
LR
var
1 119,30
2 119,43
3 119,40
4 120,12
5 121,06
6 122,86
7 126,64
8 127,75
9 128,95
10 129,63
11 128,01
12 126,73
155,44
185,50
160,26
185,59

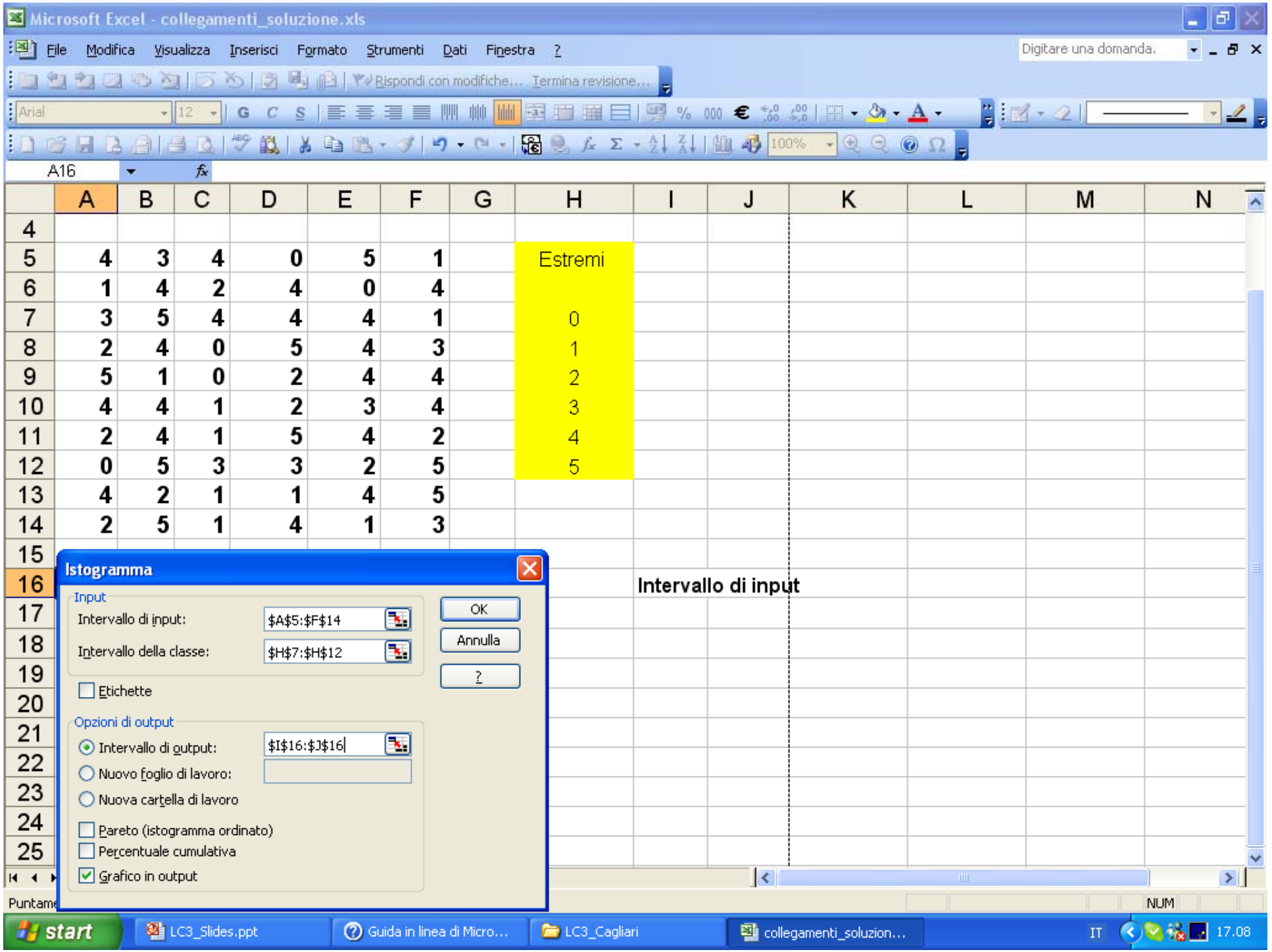
SPSS Dati


119,29691
Visibili: 4 variab
LR
SC
VG
classi
var
1 119,30
2 119,43
3 119,40
4 120,12
5 121,06
6 122,86
7 126,64
8 127,75
9 128,95
10 129,63
11 128,01
12 126,73
13 124,02
14 125,14
15 125,97
16 125,96
17 126,47
18 126,75 162,38 170,70 2,00
19 136,81 160,51 181,71 2,00
20 138,79 157,64 184,70 2,00
21 148,20 158,95 188,16 2,00
22 142,01 163,30 192,70 2,00
23 142,90 166,03 195,81 2,00
24 143,09 167,06 196,89 2,00
25 143,90 166,86 196,61 2,00
26 144,39 166,76 195,75 2,00
Ricodifica in variabili differenti
SC
VG
classi
Variabile numerica -> Variabile di output
LR --> classi
Variabile di output
Nome:
classi
Etichetta:
Cambia
Valori vecchi e nuovi...
Se...
(condizione di selezione dei c
OK
Incolla
Reimposta
Annulla
Ricodifica in variabili differenti: Valori vecchi e nuovi
Vecchio valore
Valore:
Mancante di sistema
Mancante di sistema o definito dall'utente
Intervallo:
a:
Intervallo, valore dal PIÙ PICCOLO a:
Intervallo, valore al PIÙ GRANDE:
Tutti gli altri valori
Nuovo valore
Valore:
Mancante di sistema
Copia i vecchi valori
Vecchio --> Nuovo:
Aggiungi
Cambia
Rimuovi
Lowest thru 125 --> 1
125,00001 thru 150 --> 2
150,00001 thru 180 --> 3
180,00001 thru Highest --> 4
Le variabili di output sono stringhe
Lunghezza:
8
Converti stringhe numeriche in numeri ('5'->5)
Continua
Annulla
Aiuto


collegamenti [modalità compatibilità] - Microsoft Excel
Home
Inserisci
Layout di pagina
Formule
Dati
Revisione
Visualizza
Da Access
Da Web
Da testo
Da altre origini
Connessioni esistenti
Carica dati esterni
Aggiorna tutti
Connessioni
Proprietà
Modifica collegamenti
Connessioni
Ordina
Filtro
Cancella
Riapplica
Avanzate
Ordina e filtra
Testo in colonne
Rimuovi duplicati
Convalida dati
Consolida
Analisi di simulazione
Strumenti dati
Raggruppa
Separa
Subtotale
Struttura
Analisi dati
Analisi
M7
A B C D E F G H I J K L M N O
1
2
3
4
5 4 3 4 0 5 1
6 1 4 2 4 0 4
7 3 5 4 4 4 1
8 2 4 0 5 4 3
9 5 1 0 2 4 4
10 4 4 1 2 3 4
11 2 4 1 5 4 2
12 0 5 3 3 2 5
13 4 2 1 1 4 5
14 2 5 1 4 1 3
15
16
17
18
19
20

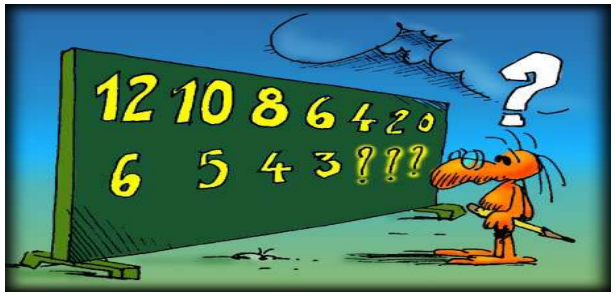


	A	B	C	D	E	F
5	4	3	4	0	5	1
6	1	4	2	4	0	4
7	3	5	4	4	4	1
8	2	4	0	5	4	3
9	5	1	0	2	4	4
10	4	4	1	2	3	4
11	2	4	1	5	4	2
12	0	5	3	3	2	5
13	4	2	1	1	4	5
14	2	5	1	4	1	3



	A	B	C	D	E	F
5	4	3	4	0	5	1
6	1	4	2	4	0	4
7	3	5	4	4	4	1
8	2	4	0	5	4	3
9	5	1	0	2	4	4
10	4	4	1	2	3	4
11	2	4	1	5	4	2
12	0	5	3	3	2	5
13	4	2	1	1	4	5
14	2	5	1	4	1	3

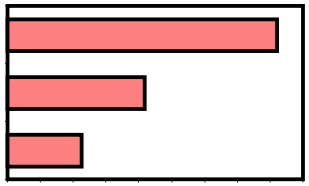
La sintesi grafica



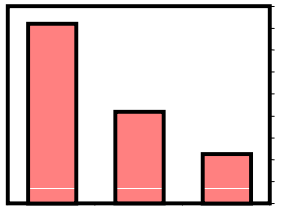
- Le rappresentazioni grafiche costituiscono uno strumento per comunicare in maniera immediata i risultati di un'indagine statistica ai "non addetti ai lavori"
  - *Infatti l'uomo percepisce più facilmente le relazioni tra le figure geometriche non quelle tra i dati numerici*

Rappresentazioni grafiche

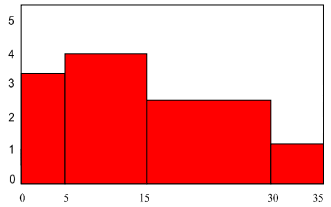
- Grafici a nastri: per caratteri qualitativi sconnessi



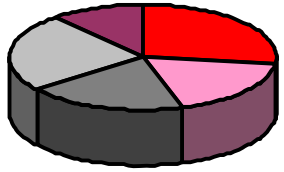
- Grafici a barre: per caratteri qualitativi ordinali o quantitativi discreti



- Istogramma: par caratteri continui



- Grafici a torta: per caratteri qualitativi sconnessi



Distribuzioni bivariate

Esempio: su un collettivo di bambini viene rilevato il numero di non-parole prodotte erroneamente (X) e il sesso (Y)

(M,0), (M,2), (F,0), (M,2), (F,0), (F1), (M,2), (M,1), (F,2), (D,0),
(F,0), (M,1), (F,2), (F,1), (M,2), (F,2), (M,2), (M,0), (F2), (D, 1)

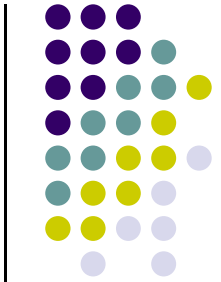
Distribuzione congiunta dei caratteri X e Y

	0	*1*	*2*	**Totale**
M	2	2	5	**9**
F	4	3	4	**11**
Totale	**6**	**5**	**9**	**20**

Distribuzione marginale del carattere Y

Distribuzione marginale del carattere X

Distribuzioni condizionate



- Distribuzione condizionata del carattere "Sesso" rispetto alle modalità del carattere "Numero di non-parole"

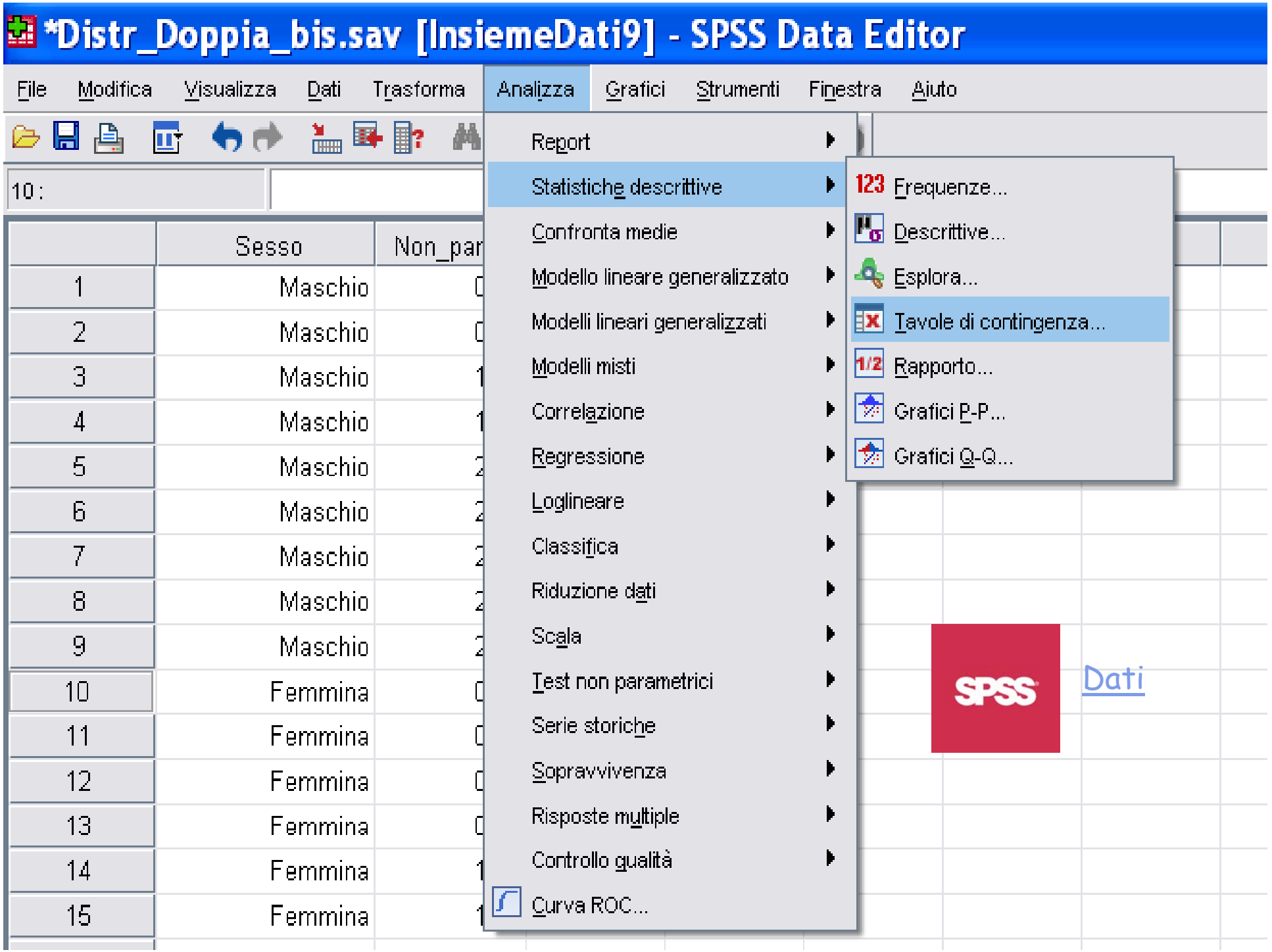
	0	*1*	*2*
M	0.333	0.4	0.556
F	0.667	0.6	0.444
Totale	**1**	**1**	**1**

tra coloro che hanno prodotto 2 non-parole il 55.6% sono bambini

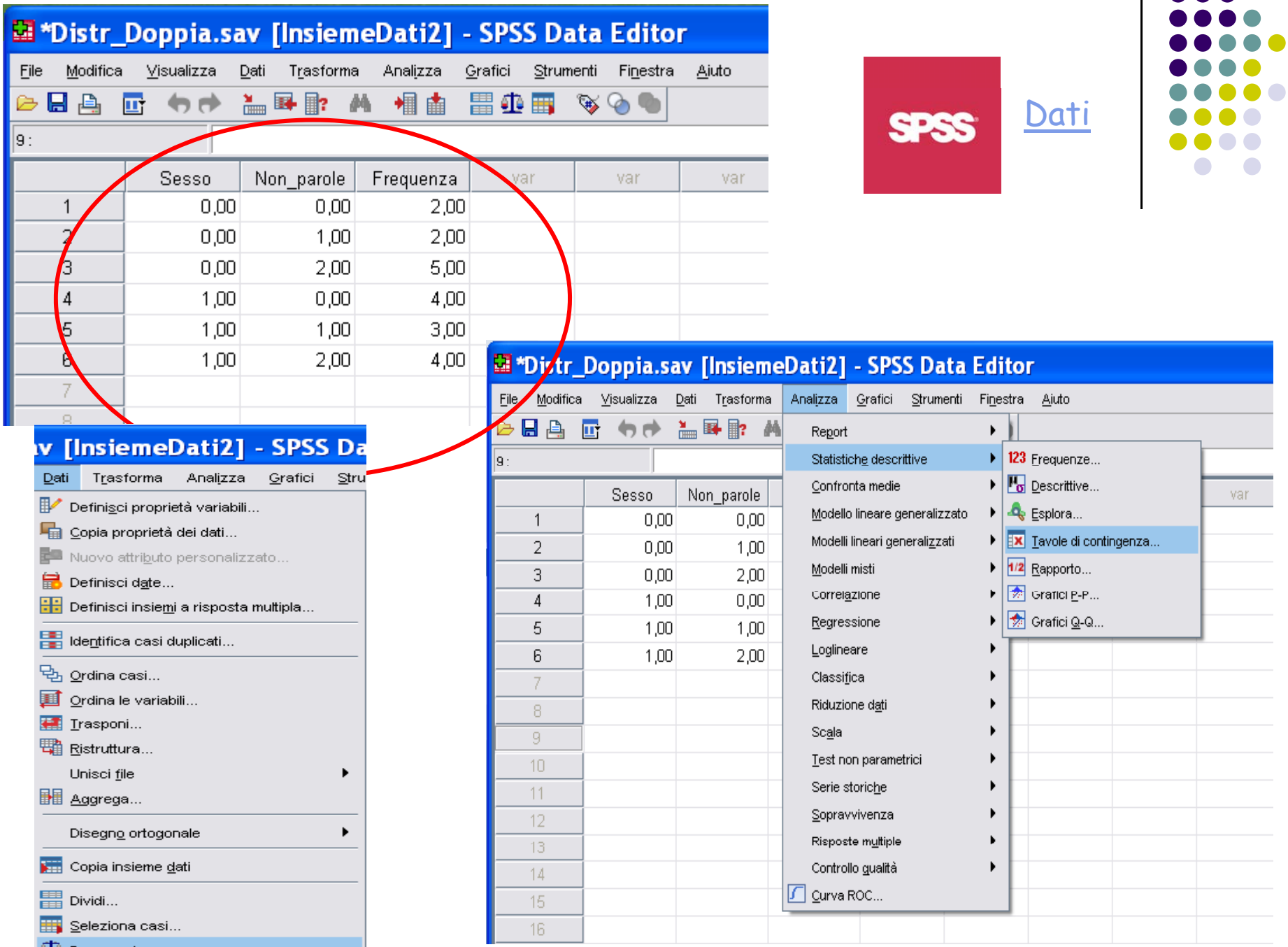
- Distribuzione condizionata del carattere "Numero di non-parole" rispetto alle modalità del carattere "Sesso"

il 36.4% delle bambine ha prodotto erroneamente 2 non-parole

	0	*1*	*2*	**Totale**
M	0.222	0.222	0.556	**1**
F	0.364	0.273	0.364	**1**


*Distr_Doppia_bis.sav [InsiemeDati9] - SPSS Data Editor
File
Modifica
Visualizza
Dati
Trasforma
Analizza
Grafici
Strumenti
Finestra
Aiuto
10 :
Sesso
Non_par
1 Maschio
2 Maschio
3 Maschio
4 Maschio
5 Maschio
6 Maschio
7 Maschio
8 Maschio
9 Maschio
10 Femmina
11 Femmina
12 Femmina
13 Femmina
14 Femmina
15 Femmina
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Frequenze...
Descrittive...
Esplora...
Tavole di contingenza...
Rapporto...
Grafici P-P...
Grafici Q-Q...
SPSS
Dati

Dati


*Distr_Doppia.sav [InsiemeDati2] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
9 :
Sesso Non_parole Frequenza var var var
1 0,00 0,00 2,00
2 0,00 1,00 2,00
3 0,00 2,00 5,00
4 1,00 0,00 4,00
5 1,00 1,00 3,00
6 1,00 2,00 4,00
7
Dati Trasforma Analizza Grafici Stru
Definisci proprietà variabili...
Copia proprietà dei dati...
Nuovo attributo personalizzato...
Definisci date...
Definisci insiemi a risposta multipla...
Identifica casi duplicati...
Ordina casi...
Ordina le variabili...
Trasponi...
Ristruttura...
Unisci file
Aggrega...
Disegno ortogonale
Copia insieme dati
Dividi...
Seleziona casi...
Pesa casi...
*Distr_Doppia.sav [InsiemeDati2] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
Report
Statistiche descrittive
Confronta medie
Modello lineare generalizzato
Modelli lineari generalizzati
Modelli misti
Correlazione
Regressione
Loglineare
Classifica
Riduzione dati
Scala
Test non parametrici
Serie storiche
Sopravvivenza
Risposte multiple
Controllo qualità
Curva ROC...
Frequenze...
Descrittive...
Esplora...
Tavole di contingenza...
Rapporto...
Grafici P-P...
Grafici Q-Q...
Sesso Non_parole var
1 0,00 0,00
2 0,00 1,00
3 0,00 2,00
4 1,00 0,00
5 1,00 1,00
6 1,00 2,00
7
8
9
10
11
12
13
14
15
16

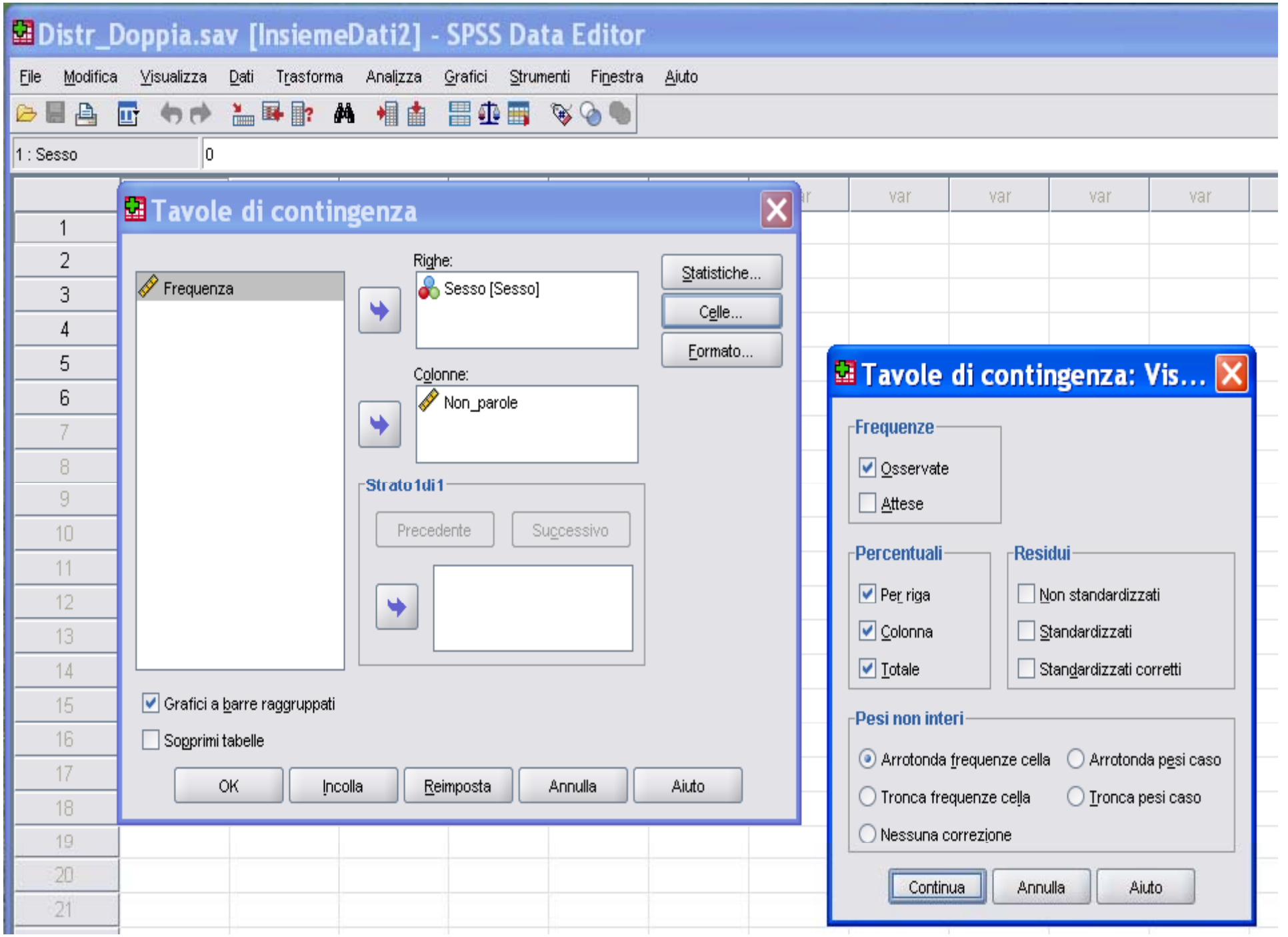

Distr_Doppia.sav [InsiemeDati2] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
1 : Sesso
0
var
Tavole di contingenza
Frequenza
Righe:
Sesso [Sesso]
Colonne:
Non_parole
Statistiche...
Celle...
Formato...
Strato 1 di 1
Precedente
Successivo
Grafici a barre raggruppati
Sopprimi tabelle
OK
Incolla
Reimposta
Annulla
Aiuto
Tavole di contingenza: Vis...
Frequenze
Osservate
Attese
Percentuali
Per riga
Colonna
Totale
Residui
Non standardizzati
Standardizzati
Standardizzati corretti
Pesi non interi
Arrotonda frequenze cella
Arrotonda pesi caso
Tronca frequenze cella
Tronca pesi caso
Nessuna correzione
Continua
Annulla
Aiuto

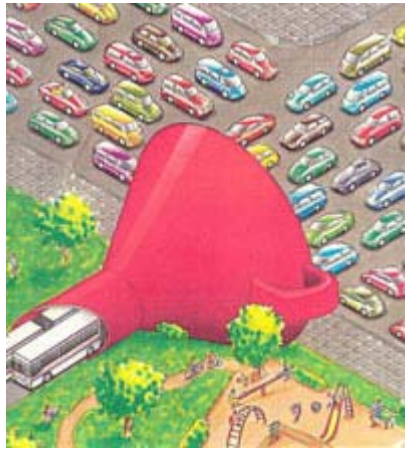
Tavola di contingenza Sesso * Non_parole

			Non_parole			
			0	1	2	Totale
Sesso	Maschio	Conteggio	2	2	5	9
		% entro Sesso	22,2%	22,2%	55,6%	100,0%
		% entro Non_parole	33,3%	40,0%	55,6%	45,0%
		% del totale	10,0%	10,0%	25,0%	45,0%
	Femmine	Conteggio	4	3	4	11
		% entro Sesso	36,4%	27,3%	36,4%	100,0%
		% entro Non_parole	66,7%	60,0%	44,4%	55,0%
		% del totale	20,0%	15,0%	20,0%	55,0%
	Totale	Conteggio	6	5	9	20
		% entro Sesso	30,0%	25,0%	45,0%	100,0%
		% entro Non_parole	100,0%	100,0%	100,0%	100,0%
		% del totale	30,0%	25,0%	45,0%	100,0%

Misure di sintesi

- Il processo di sintesi non può limitarsi solo alle distribuzioni di frequenze ma deve spingersi oltre fino a sintetizzare in un unico dato numerico una particolare caratteristica delle popolazione

- L'idea è quella di sostituire tutte le modalità del carattere in esame con un'unica modalità che le rappresenti



Misure di sintesi

Questa finalità può essere perseguita attraverso la determinazione di opportuni indici sintetici del fenomeno considerato.

Questi indici, data la loro funzione, possono essere impiegati per:

- ***Confrontare tra di loro situazioni diverse***
  - *Stesso fenomeno rilevato su collettivi diversi*
  - *Stesso fenomeno rilevato su tempi e/o luoghi diversi*
  - *Fenomeni diversi tra di loro*
- ***Valutare il dato ottenuto confrontandolo con un valore standard noto***

Misure di sintesi

Quali sono le misure sintetiche che possono essere calcolate su un insieme di dati?

La scelta dipende dalle caratteristiche che descrivono un collettivo statistico

Misure di sintesi

Tra le possibili caratteristiche di una popolazione, due sono di gran lunga le più importanti:

- *Centralità*
- *Variabilità*

✓ **Misure di centralità**
*Esprimono sinteticamente il centro ideale della distribuzione, ovvero il valore intorno al quale tendono a **gravitare** i dati*

✓ **Misure di variabilità**
Consentono di valutare il grado di diversità delle modalità del carattere, ovvero forniscono informazioni sul grado di dispersione dei dati intorno al loro centro ideale

Le medie

Non è possibile definire un'unica misura di centralità (**media**). La scelta dipende dalla natura dei caratteri in esame e dalle finalità dello studio

Indipendentemente dalla media adoperata, il valore di sintesi non è detta che coincida esattamente con una delle modalità osservate. Accade di frequente che tale valore è solo **fittizio**.

Le medie

Sai ched'è la statistica? E' 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.

Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che, lì, la media è sempre eguale
puro co' la persona bisognosa.

Me spiego, da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:

e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perchè c'è un antro che se ne magna due.

(Trilussa)

ignorava
la
variabilità !!!!!!!

Morale della favola:

- *Sintetizzare in maniera opportuna i dati*
- *Affiancare una misura della bontà della sintesi realizzata*

Le medie

❑ Medie di Posizione (o Medie Lasche)

Gli indici che rientrano in questa categoria si identificano in un valore della distribuzione che risulta "privilegiato" rispetto agli altri, o perché più frequente oppure perché occupa una determinata posizione.

Possono essere determinate in linea di massima per tutti i tipi di caratteri

❑ Medie Algebriche (o Medie Ferme)

Gli indici che rientrano in questa categoria possono essere determinate solo per i caratteri quantitativi poiché sono il risultato di una serie di operazioni algebriche effettuate su tutte le modalità del carattere

La moda

Definizione: la moda di una distribuzione è la modalità a cui è associata la frequenza (assoluta o relativa) più elevata

In altre parole, la moda rappresenta il valore prevalente nell'insieme dei dati, ovvero quello che si presenta con maggiore frequenza e può essere determinato per tutti i tipi di carattere

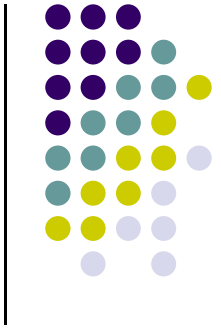
Definizione: la classe modale è la classe a cui è associata la frequenza più elevata se le classi presentano la stessa ampiezza, ovvero la classe a cui compete la **densità di frequenza** più elevata se la classi presentano ampiezza diversa.

$$h_i = \frac{n_i}{Es - Ei}$$

Convenzionalmente si assume come moda della distribuzione il valore centrale della classe modale.

$$Mo = \frac{x_i + x_{i+1}}{2}$$

La mediana



Definizione: è il centro ordinale di un insieme di valori, ovvero il valore che bipartisce il collettivo statistico in due gruppi di uguale numerosità

La determinazione della mediana richiede, come prerequisito, che il carattere in esame sia almeno ordinale. Pertanto potrà essere determinata per tutti i tipi di caratteri tranne quelli sconnessi

La determinazione della mediana segue procedimenti diversi a seconda di come sono organizzati i dati

La mediana: elenco di modalità

Consideriamo una successione di *n* valori ordinati in senso non decrescente

$$x_1 \leq x_2 \leq \ldots \leq x_i \ldots \leq x_n$$

La mediana (Me) è definita come il valore centrale della successione, cioè come quel valore che è preceduto e seguito dallo stesso numero di dati

$$Me = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{se n è pari} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{se n è dispari} \end{cases}$$

La mediana: elenco di modalità, n dispari

Consideriamo i seguenti valori relativi al punteggio conseguito ad un test da un gruppo di studenti

8, 7, 6, 5, 8, 9, 9, 4, 3

e ordiniamoli in senso non decrescente

Mediana

Valori	**3**	**4**	**5**	**6**	**7**	**8**	**8**	**8**	**9**
Rango	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*

posizione centrale

Poiché *n=9*, esiste un'unica posizione centrale: la posizione *(9+1)/2=5*.
Pertanto Me=2

La mediana: elenco di modalità, n pari

Consideriamo i seguenti valori

8, 7, 6, 5, 8, 9, 9, 4, 3, 8

e ordiniamoli in senso non decrescente

7.5=Mediana

Valori	**3**	**4**	**5**	**6**	**7**	**8**	**8**	**8**	**9**	**9**
Rango	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

Posizioni centrali

Poiché *n=10*, esistono due posizioni centrali, P1=5 e P2=6. Pertanto la mediana è la semisomma delle modalità che occupano queste posizioni

Me=(7+8)/2=7.5

La mediana: elenco di modalità, n pari

Osservazioni:

- ✓ *il valore mediano non coincide con nessun valore rilevato*
- ✓ *poiché le modalità di rango centrale sono diverse segue che esattamente il 50% dei valori è inferiore alla mediana, mentre il rimanente 50% è superiore alla mediana*
- ✓ *nel caso in cui si presenta la stessa situazione per caratteri qualitativi, la mediana risulta indeterminata*

La mediana: distribuzione di frequenza

Esempio: distribuzione di un gruppo di bambini per numero di non-parole prodotte erroneamente

Mediana ⟶

X	n	f	F
0	5	0,10	0,10
1	12	0,24	0,34
2	19	0,38	0,72
3	9	0,18	0,90
4	4	0,08	0,98
5	1	0,02	1,00
Totale	**50**		

La mediana: distribuzione in classi di modalità

La mediana è data dalla seguente espressione:

$$Me = Ei + (0.5 - Fi)\frac{Es - Ei}{Fs - Fi}$$

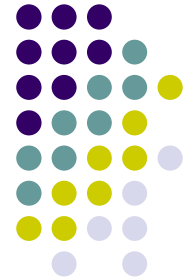
Es = estremo superiore della classe mediana

Ei = estremo inferiore della classe mediana

Fs = frequenza cumulata relativa della classe mediana

Fi = frequenza cumulata relativa della classe che precede quella mediana

La mediana: distribuzione in classi di modalità



Esempio: distribuzione di un gruppo di persone classificate in base all'IQ

IQ	n	f%	F%	Es-Ei	densità
59 -|70	11	6.2	0.1	21	0.52
70 -| 90	42	26.7	33	20	2.1
90 -| 105	52	29.5	62.5	15	3.47
105-| 120	9	5.1	67.6	15	0.6
120 -| 130	11	6.2	73.9	10	1.1
130-|150	46	26.1	100	20	2.3
Totale	176	**1**			

Classe Modale: 90-105

Moda=(90+105)/2=97.5

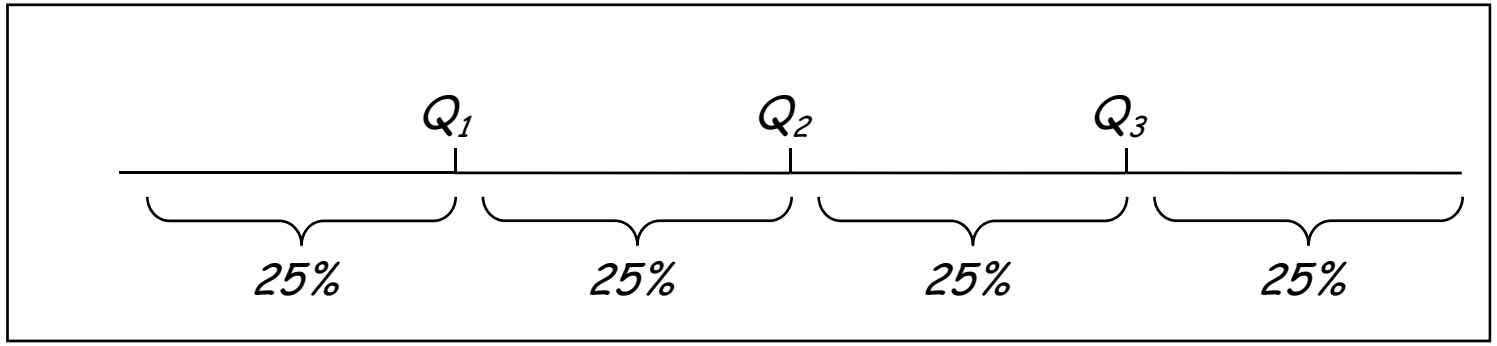
Classe Mediana: 90-105 $Me = 90 + (0.5 - 0.33)\frac{105 - 90}{0.625 - 0.33} = 98.64$

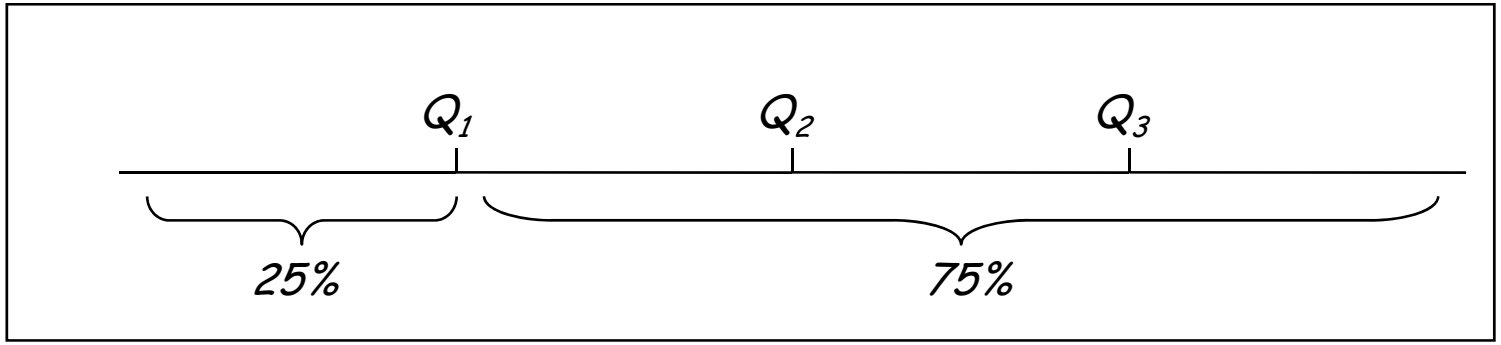
Percentili e quartili

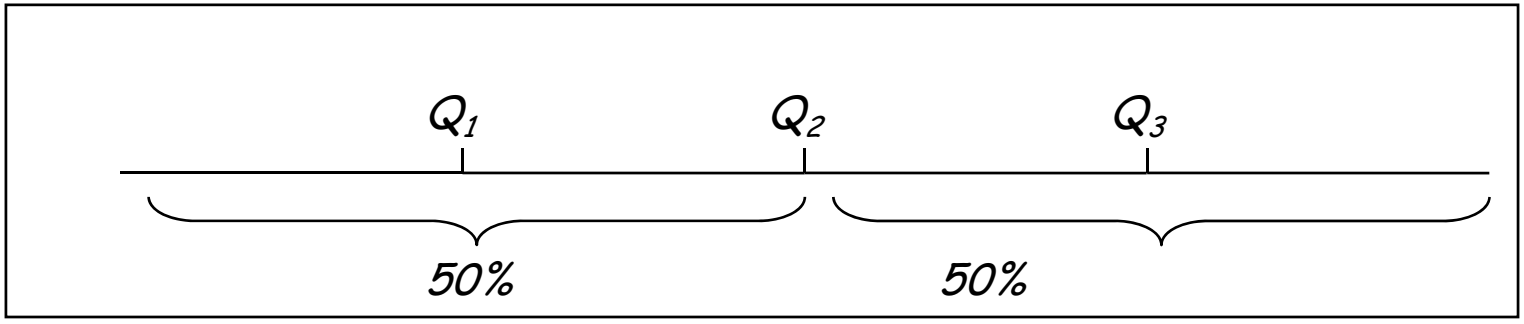
Immaginiamo ora di suddividere il collettivo in 100 parti, ognuna delle quali contenente lo stesso numero di unità.
I valori che suddividono la distribuzione in 100 parti di uguale numerosità sono detti **percentili**.

I percentili che vengono maggiormente impiegati sono *25-esimo*, il *50-esimo* e il *75-esimo,* detti rispettivamente primo quartile (Q_1), secondo quartile (Q_2) e terzo quartile (Q_3).

Come è facile intuire, i **quartili** dividono il collettivo statistico in quattro parti uguali.


Q_1
Q_2
Q_3
25%
25%
25%
25%


Q_1
Q_2
Q_3
25%
75%


Q_1
Q_2
Q_3
50%
50%

L'individuazione dei percentili risulta particolarmente importante quando occorre valutare la "performance" di un soggetto.

Così, ad esempio, nella prova non-parole, la prestazione di un bambino viene considerata "bassa" rispetto alla sua età se il punteggio ottenuto si colloca intorno al 10° percentile.

Percentile	10°	25°	50°	75°	90°
scuola dell'infanzia	17	13	10	7	7
scuola elementare	12	9	6	4	2

Fonte: ***Orsolini M., Capriolo S., Santese A., Cerracchio S.: Un compito per valutare la memoria fonologica a breve termine. La prova di ripetizione di non-parole***

La media aritmetica: elenco di modalità

In nove tragedie di Racine viene conteggiato il numero di volte che compare l'aggettivo "heureux " (felice)

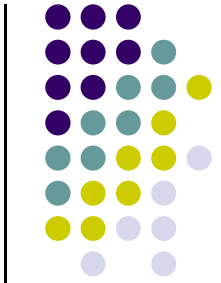
tragedia	I	II	III	IV	V	VI	VII	VIII	IX
occorrenza	10	11	13	15	16	18	18	13	23

Fonte: Cortellazzo M., Tuzzi A. : Metodi Statistici Applicati all'Italiano

In totale si osservano 143 occorrenze suddivise nelle nove tragedie. Se il numero complessivo venisse suddiviso in maniera equa tra le nove tragedie, quante occorrenze spetterebbero ad ogni tragedia?

La risposta è banale: $\frac{10+11+13+15+16+18+18+19+23}{9} = \frac{143}{9} = 15.89$

La media aritmetica: distribuzioni di frequenze



In molte situazioni, i dati si presentano sotto forma di distribuzioni di frequenza.

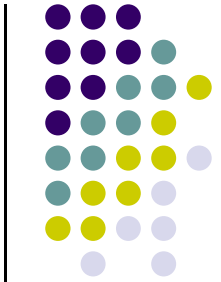
X	n	f%
1	65	21.67%
2	60	20.00%
3	50	16.67%
4	43	14.33%
5	27	9.00%
6	21	7.00%
7	15	5.00%
8	12	4.00%
9	5	1.67%
10	2	0.67%
tot	300	100%

Nel prospetto è riportata la lunghezza (X) e la frequenza di alcune forme grafiche (= successione di caratteri tra due separatori) presenti in un testo scritto.

Si è interessati ad individuare la lunghezza media della forma grafica (f.g.) .

In tal caso occorre "pesare" le lunghezze per il numero di volte che queste si presentano.

La media aritmetica: esempi di calcolo



X	n	f	x*n	x*f
1	65	0.217	65	0.217
2	60	0.200	120	0.400
3	50	0.167	150	0.500
4	43	0.143	172	0.573
5	27	0.090	135	0.450
6	21	0.070	126	0.420
7	15	0.050	105	0.350
8	12	0.040	96	0.320
9	5	0.017	45	0.150
10	2	0.007	20	0.067
tot	300	1	1034	3.447

Lunghezza totale Lunghezza media

$$M(X) = \frac{(1*6)+(2*65)+\ldots+(10*2)}{300} = \frac{1034}{300} = 3.447$$

$$M(X) = (1*0.217)+(2*0.2)+\ldots+(10*0.007) = 3.447$$

Significato: se la lunghezza complessiva delle f.g. (1034 caratteri) venisse ripartita in maniera equa (uniforme) tra le 300 f.g., ognuna di questa avrebbe una lunghezza di 3.447 caratteri

La media aritmetica: elenco di modalità

Definizione: la media aritmetica M di una successione di valori $x_1, x_2,\ldots,x_n$ si ottiene dividendo la somma degli stessi per il numero n di osservazioni

$$M = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Interpretazione: Se la somma dei valori ha un significato, allora la media aritmetica, essendo un rapporto, indica quella parte del totale che spetterebbe a ciascuna unità qualora questo venisse suddiviso in *n* parti uguali

La media aritmetica: distribuzione di frequenze e classi di modalità

Definizione: la media aritmetica M per una distribuzione di frequenze è data dalle espressioni equivalenti

$$M = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \qquad\qquad M = \sum_{i=1}^{k} x_i f_i$$

Definizione: la media aritmetica M per una distribuzione in classi di modalità è data dalle espressioni equivalenti

$$M = \frac{1}{n}\sum_{i=1}^{k} c_i n_i \qquad\qquad M = \sum_{i=1}^{k} c_i f_i$$

Valore centrale della classe $\longrightarrow$ $c_i = \dfrac{x_i + x_{i+1}}{2}$