Elementi di Psicometria

2-Grafici e distribuzioni di frequenza
vers. 1.1a (10 novembre 2011)
versione per stampa

Germano Rossi[1]
germano.rossi@unimib.it

[1]Dipartimento di Psicologia, Università di Milano-Bicocca

2011-2012

Obiettivo della statistica descrittiva

- L'obiettivo principale è quello di cercare (e trovare) un certo ordine/struttura nelle diverse informazioni rappresentate dai dati numerici
- Quest'obiettivo si può raggiungere tramite valori numerici o foeme di visualizzazione che rappresentano delle "sintesi" statistiche
  - distribuzioni di frequenza
  - indici della tendenza centrale
  - indici di variabilità
- oppure tramite rappresentazioni grafiche

- Vedrete queste cose praticamente durante le esercitazioni (sia in Excel, sia in R), sia in SPSS
- Trovate dei testi di introduzione a questi software sul sito di *tutoring*: `http://lab4.psico.unimib.it/tutoring/forum/viewforum.php?f=127`

Distribuzione di frequenza semplice

Si tratta di contare quanti elementi appartengono ad una stessa categoria presente in una variabile.

Esempio

Se abbiamo la seguente distribuzione di dati, la distribuzione di frequenza sarà:

M F M F F M M M F F M M F F M
M F F M M M M F F M F M F M M

File: `Esempio2-1.xls`

	f		
F	13	$\leftarrow$	f_f
M	17	$\leftarrow$	f_m
Tot	30	$\leftarrow$	N

- Contiamo le F; il loro numero è la frequenza delle femmine (f_f)
- Contiamo le M; il loro numero è la frequenza dei maschi (f_m)
- La somma di tutte le frequenze equivale al numero dei casi statistici (indicato con N)

Distribuzione di frequenza semplice

- La somma di tutte le frequenze, deve necessariamente equivalere a N (chiamata **numerosità** ovvero il numero di casi statistici).
- In termini matematici, equivale a scrivere

$$\sum_{i=1}^{N} f_i = \sum f = N$$

  dove f indica *frequenza* e *i* è un indice che indica tutte le categorie possibili, per cui (in questo caso) può valere *f* o *m* oppure 1 o 2
- la distribuzione di frequenza permette di riassumere velocemente molti dati
- si applica sicuramente a scale Nominali, Ordinali ed è possibile anche a livello Intervallo/Rapporto (ma non sempre è utile)

Distribuzione di frequenza cumulata

- È la somma delle frequenze che precedono una determinata categoria
- La frequenza cumulata della prima categoria è uguale a se stessa
- la frequenza cumulata delle categorie intermedie, si ottiene sommando i singoli valori di frequenza delle categorie precedenti
- la frequenza cumulata dell'ultima categoria è uguale a N (somma di tutte le frequenze)
- si applica a scale Ordinali e Intervallo/Rapporto

	f	fc		
Nessuna	4	4	4	
1 scelta	2	6	4+2	
2-4 scelte	52	58	6+52	4+2+52
5 o più scelte	18	76	58+18	4+2+52+18
Totale	76			

Intervalli di classe

- Quando a qualcuno chiedete “Quanti anni hai?” potete ricevere diverse risposte: 19; quasi 23; ho appena compiuto i 20...
- In tutti i casi l’età detta è in qualche modo “approssimata” ad un intero, perché chiunque avrà un certo numero di anni, qualche mese, un po’ di giorni, qualche ora, alcuni minuti, alcuni secondi...
- L’età, come numero intero di **anni** trascorsi dalla nascita, può essere pensato come un **intervallo di classe** con intervallo di ampiezza 1
- Il valore centrale di un intervallo diventa il punto verso cui gravitano i veri valori dell’età.
- 19 anni e 8 mesi è più vicino ai 20, mentre 19 anni e 3 mesi è più vicino ai 19.

Intervalli di classe

- Una qualsiasi variabile (discreta o continua, purché quantitativa), se ha molte categorie di valori, può essere raggruppata in intervalli di classe
- La classificazione in intervalli può essere utilizzata sia per costruire la distribuzione di frequenza e fare rappresentazioni grafiche sia per il calcolo delle principali statistiche (in questo caso si perdono informazioni ed è altamente sconsigliato)
- Se faccio gli intervalli di classe di uguale ampiezza, qualunque variabile intervallo/a rapporto (*anche se discreta*) viene considerata "come se fosse" continua
- Se faccio intervalli di classe di diversa ampiezza, **abbasso la scala a livello ordinale** (es, l'età in fasce: adolescenti, giovani, giovani adulti, adulti...)

Intervalli di classe

Per fare classi di ampiezza uguale, alcune regole sono:

- individuare il valore massimo e quello minimo
- calcolare la gamma di variazione (massimo - minimo)
- dividere la gamma per il numero degli intervalli che si desiderano (fra 8 e 15) ottenendo l'ampiezza dell'intervallo
- arrotondare l'ampiezza dell'intervallo all'intero (generalmente 3, 5 o 10)
- usando il valore minimo si sceglie un valore iniziale dell'intervallo che contenga il valore minimo
- si attribuiscono i valori a ciascun intervallo
- si considerano i valori di ciascun intervallo come "uniformemente distribuiti" all'interno dell'intervallo stesso

Intervalli di classe

per ogni intervallo si definisce:

- un limite inferiore (il valore corrispondente si include)
- un limite superiore (il valore corrispondente si esclude da questo intervallo e si include nel successivo)
- un valore centrale
- dei limiti reali

limiti indicati	valore centrale	limiti reali ($\vdash$)
85-89	87	84,5-89,5
90-94	92	89,5-94,5
95-99	97	94,5-99,5

Intervalli di classe

Esempio

- Consideriamo l'età di 25 studenti (`Esempio2-2.xls`): 22, 18, 20, 21, 26, 37, 28, 17, 30, 42, 37, 21, 19, 20, 18, 24, 28, 20, 30, 20, 31, 20, 33, 19, 22. Minimo: 17; Massimo: 42
- gamma: 42-17=25, 25/10=2.5 $\simeq$ 3; 25/5=5
- Scegliamo l'ampiezza di 5 e costruiamo gli intervalli

	f	lim inf ($\vdash$)	lim sup	vc	valori inclusi
16-20	10	15,5	20,5	18	17,18x2,19x2,20x5
21-25	5	20,5	25,5	23	21x2,22x2,24
26-30	5	25,5	30,5	28	26,28x2,30x2
31-35	2	30,5	35,5	33	31,33
36-40	2	35,5	40,5	38	37x2
41-45	1	40,5	45,5	43	42

Intervalli di classe

Esempio

- Usiamo ora un'ampiezza di 3 e costruiamo gli intervalli

	f	li	ls	vc
16-18	3	15,5	18,5	17
19-21	9	18,5	21,5	20
22-24	3	21,5	24,5	23
25-27	1	24,5	27,5	26
28-30	4	27,5	30,5	29
31-33	2	30,5	33,5	32
34-36	.	33,5	36,5	35
37-39	2	36,5	39,5	38
40-42	1	39,5	42,5	41

Intervalli di classe

Esempio

	f	fc	%	vc=x	fx
16-20	10	10	40	18	180
21-25	5	15	20	23	115
26-30	5	20	20	28	140
31-35	2	22	8	33	66
36-40	2	24	8	38	76
41-45	1	25	4	43	43
Somme	25		100		620
Media					24,8

Rappresentazioni grafiche

Con le attuali conoscenze, le rappresentazioni grafiche disponibili sono:

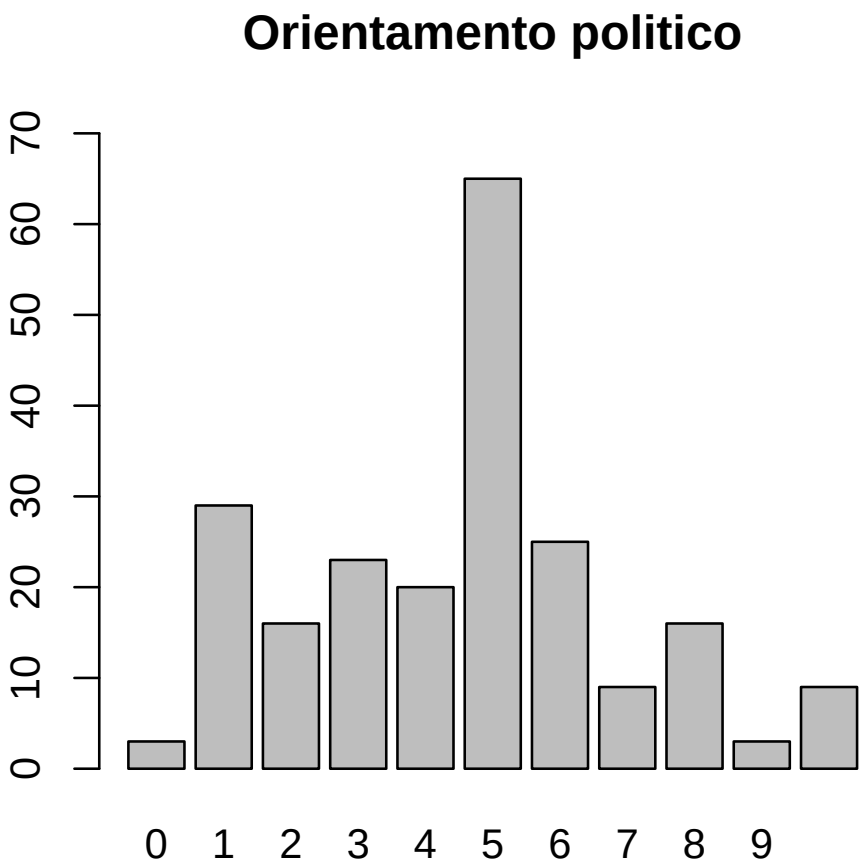
- Grafico a barre (verticali/orizzontali)
- Grafico a torta
- Istogramma a barre (verticali/orizzontali)
- Poligoni di frequenza (semplice e cumulata)
- Diagrammi ramo-foglia

- Alcuni di questi grafici si possono ottenere sia in Excel sia in R sia in SPSS
- Alcuni non si possono ottenere se non in R

Grafici/Istogrammi a barre

- Grafici e istogrammi a barre sono molti simili
- I **grafici a barre** sono indicati per variabili nominali e ordinali: le barre sono separate l'una dall'altra
- Gli **istogrammi a barre** sono indicati per variabili intervallo/rapporto: le barre sono contigue fra loro
- Ogni frequenza è rappresentata da una barra
- La lunghezza di ogni barra è proporzionale alla frequenza: barre più lunghe indicano frequenze più elevate
- Negli istogrammi anche l'area di una barra è proporzionale alla frequenza

Grafico a barre (Nominale/Ordinale)



- la variabile è misurata su una scala a 10 punti che vanno da 1=sinistra a 10=destra
- è quindi una variabile ordinale (*)

(*) ma viene spesso considerata a intervalli

Grafico a barre (Nominale/Ordinale)

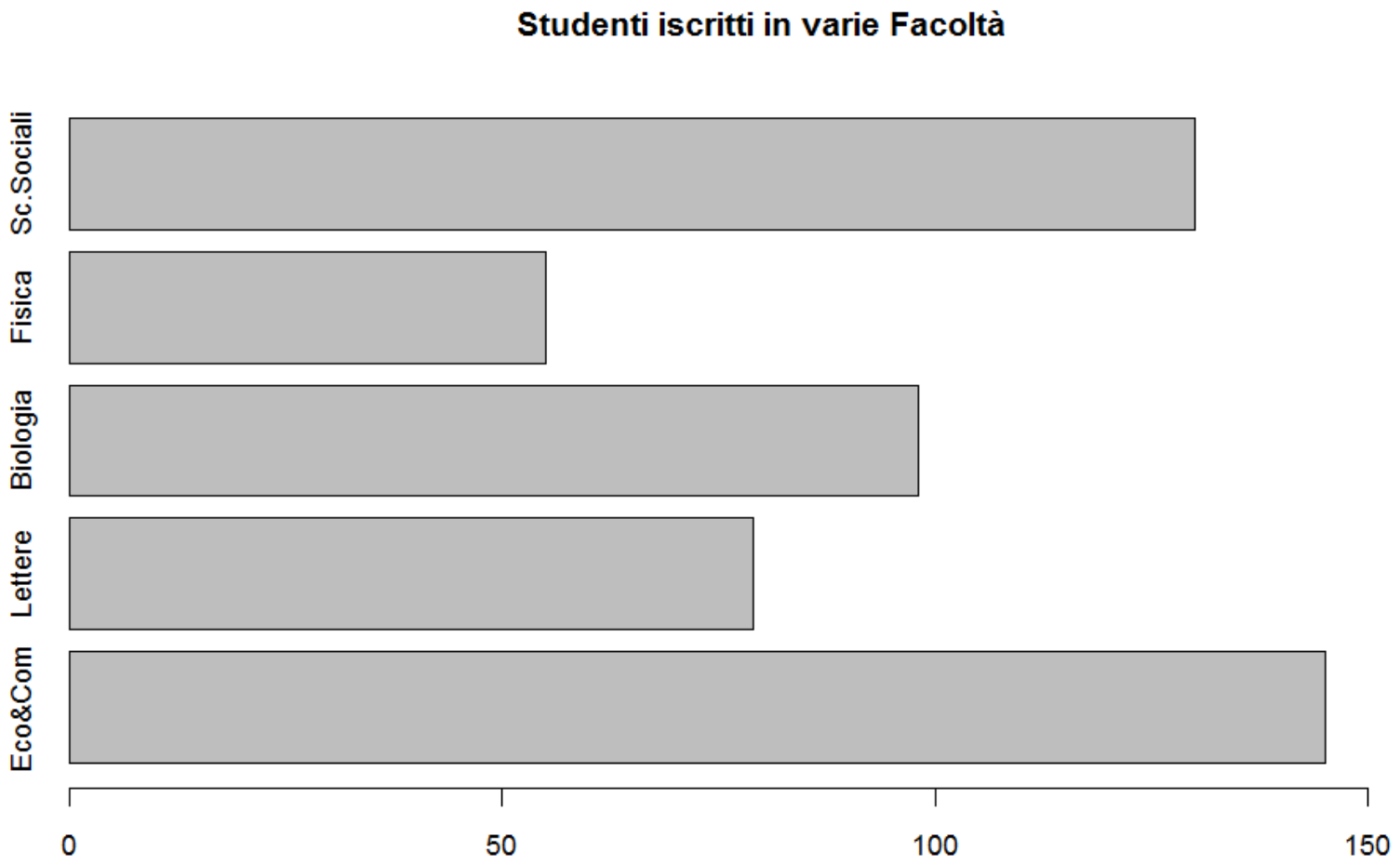
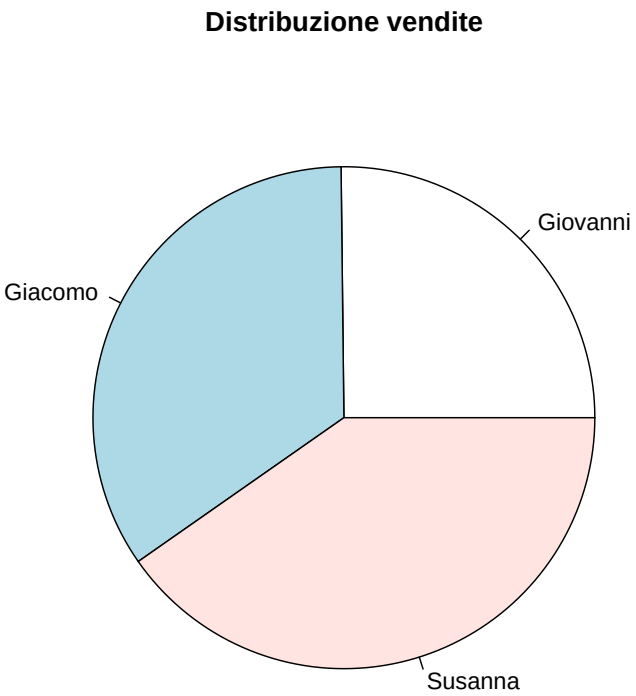
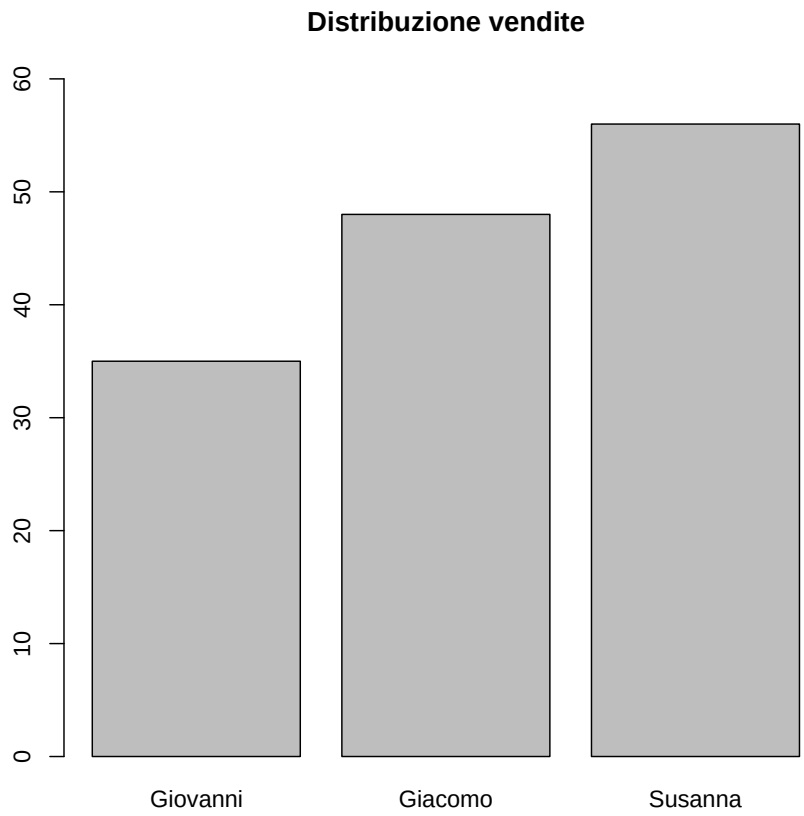


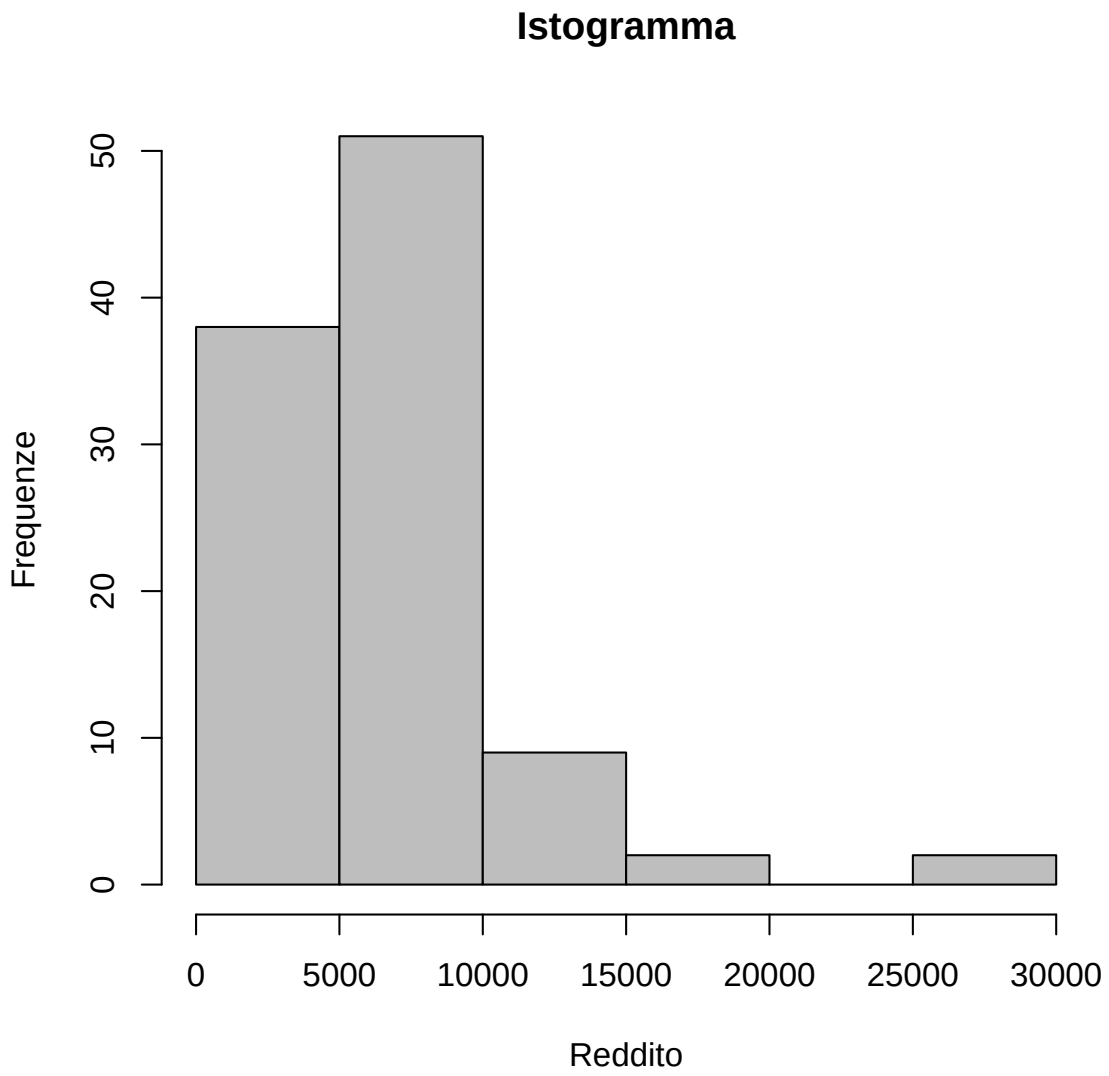
Grafico a torta (N/O)





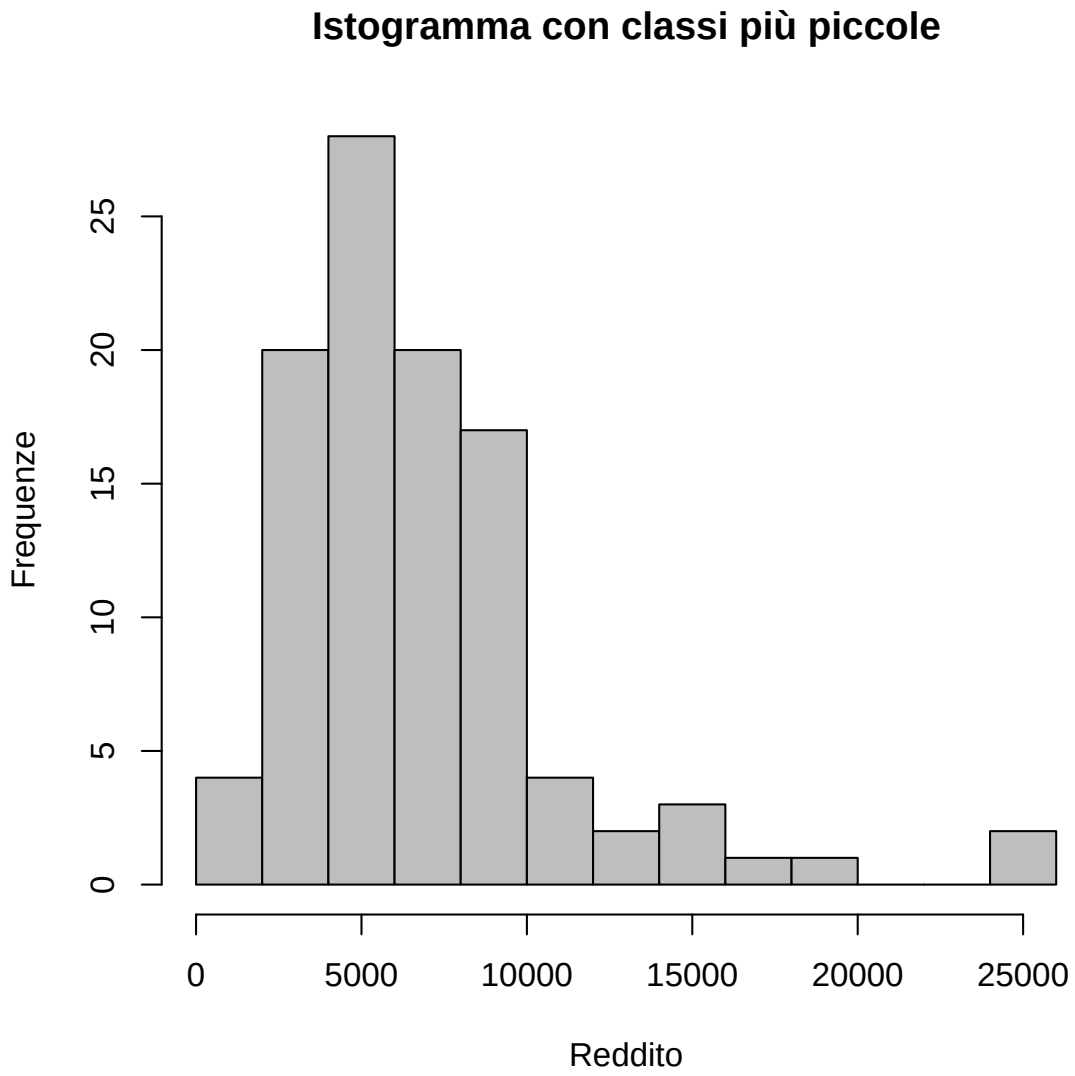
Le informazioni fornite sono analoghe; le torte fanno più scena, ma diventano complicate da interpretare se ci sono molte categorie con frequenze molto vicine fra loro

Istogramma (I/R)



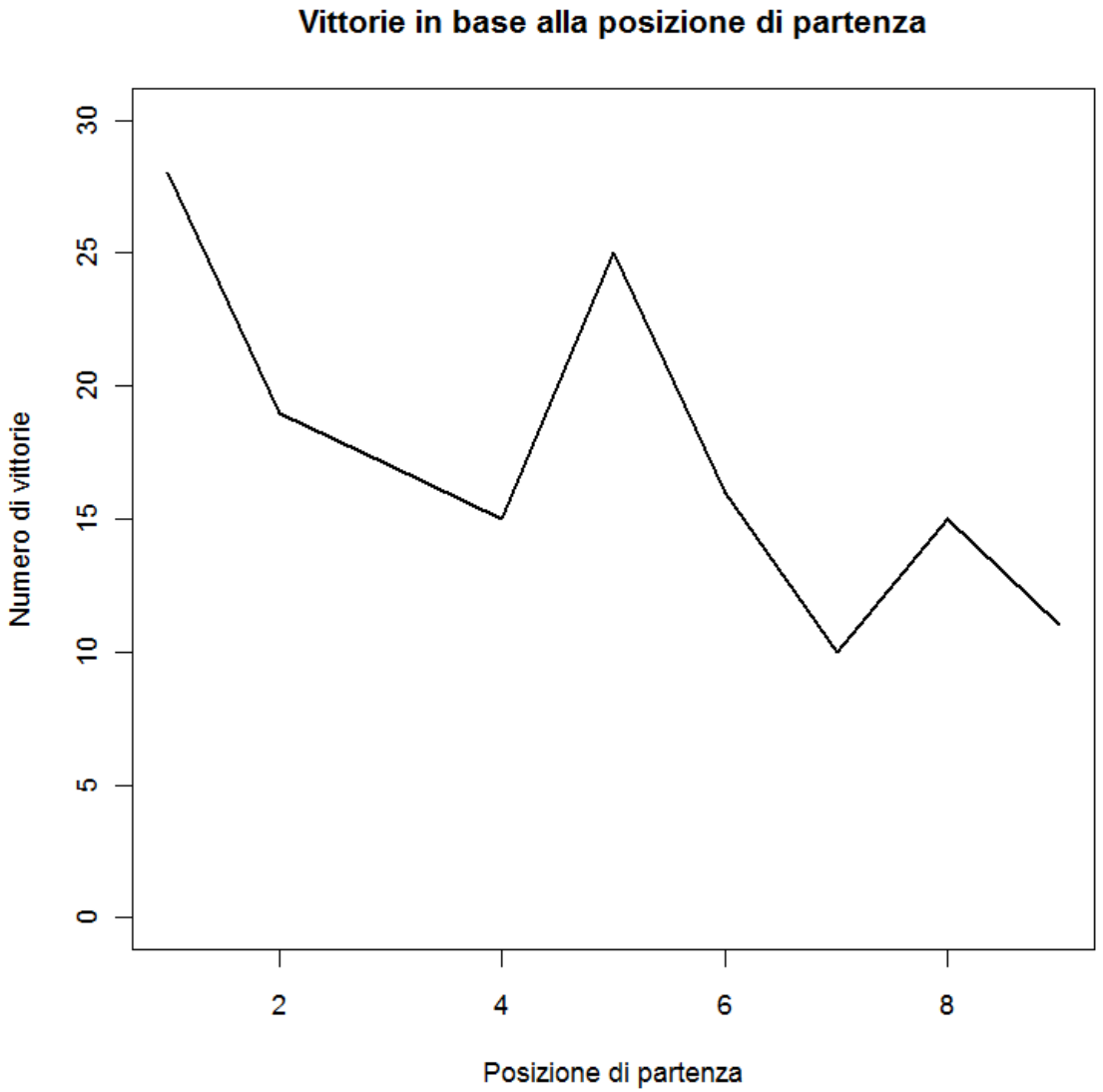
- la variabili quantitative (in particolare quelle continue) vengono rappresentate tramite intervalli di classe (SPSS lo fa automatica-mente)

Istogramma con classi più piccole (I/R)



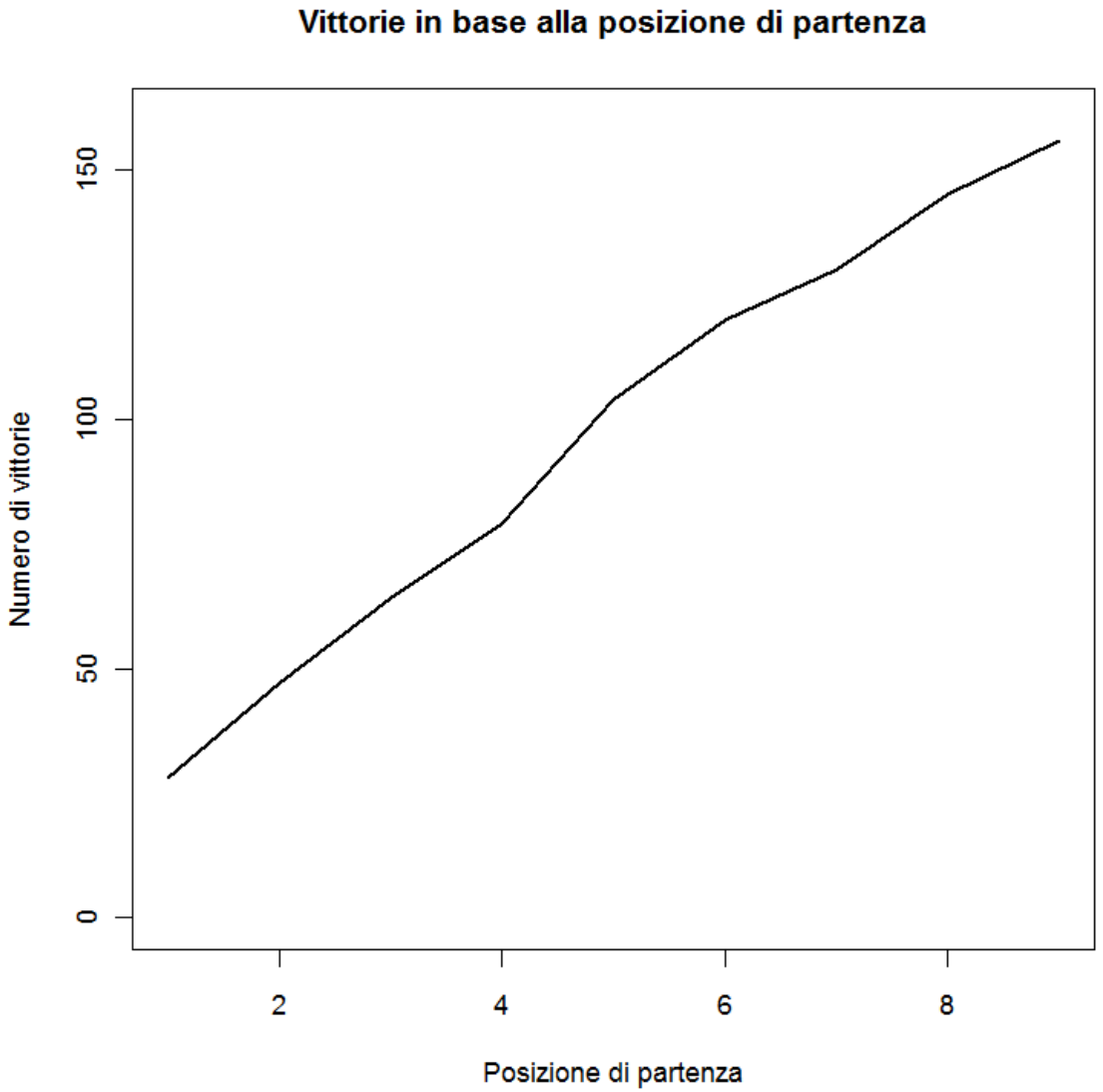
- se la variabile non è molto simmetrica, l'ampiezza degli intervalli può cambiare moltissimo la distribuzione raffigurata

Poligono di frequenze (I/R)



- Nei poligoni i punti corrispondenti alle frequenze sono uniti da una linea

Poligono di frequenze cumulate (I/R)



- Nei poligoni cumulati i punti uniti sono quelli delle frequenze cumulate

Ramo e foglia [Stem & leaf]

Consideriamo un insieme di dati (`EsempioG.xls`)
G=2 3 4 5 6 2 2 2 5 5 3 4 2 6 3 4 4 5 3 2
ordiniamo i dati: 2 2 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 6 6

```
2 | 000000
3 | 0000
4 | 0000
5 | 0000
6 | 00
```

- Per ogni valore attiviamo un “ramo” e usiamo poi lo 0 per indicare la “foglia”
- Abbiamo 6 volte il valore 2
- Sul ramo “2” avremo 6 “foglie”
- e via così

- In questo modo otteniamo un “grafico” (a caratteri) molto simile ai grafici a barre orizzontali

Complichiamo un po’ le cose

Ramo e foglia [Stem & leaf] (I/R)

A=33, 45, 39, 31, 37, 46, 34, 22, 30, 35, 27, 45, 42, 27, 31, 33, 44, 39, 36, 24, 27, 30, 24, 22, 33, 36, 54, 54, 46, 32, 33, 24, 24, 36, 35, 42, 24, 42, 45, 27, 41 (`EsempioA.xls`)
Dati ordinati: 22 22 24 24 24 24 24 27 27 27 27 30 30 31 31 32 33 33 33 33 34 35 35 36 36 36 37 39 39 41 42 42 42 44 45 45 45 46 46 54 54

```
2 | 2244444
2 | 7777
3 | 0011233334
3 | 55666799
4 | 12224
4 | 55566
5 | 44
```

- Se i valori utilizzano le decine, queste vengono usate per i "rami"
- e le unità per le "foglie"
- Se su un ramo ci sono molte foglie il ramo viene "spezzato in due" (per non avere pochi rami e troppe foglie)

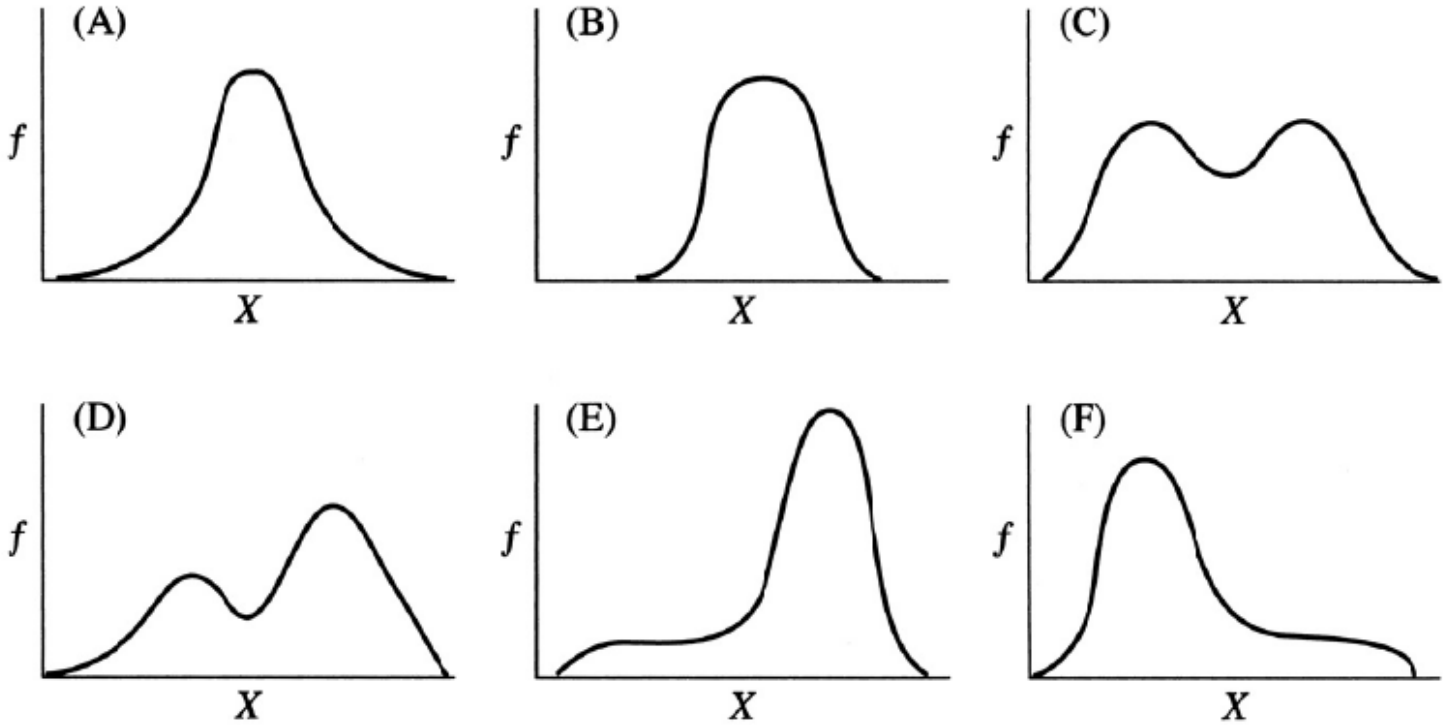
Ramo e foglia [Stem & leaf]

Con variabili più complesse come il `Reddito`

Min.	ramo-foglia	Max.	ramo-foglia	
611	0 e 6	25880	24 e 9	Esempio SPSS

```
Il punto decimale \`{e} 3 posizioni a destra del |
   0 | 6979
   2 | 4468900112555666799 9
   4 | 0122334567778811112345668 89
   6 | 012335566799011456 79
   8 | 000012334488999936
  10 | 4004
  12 | 45
  14 | 026
  16 | 5
  18 | 3
  20 |
  22 |
  24 | 39
```

Forme di distribuzione

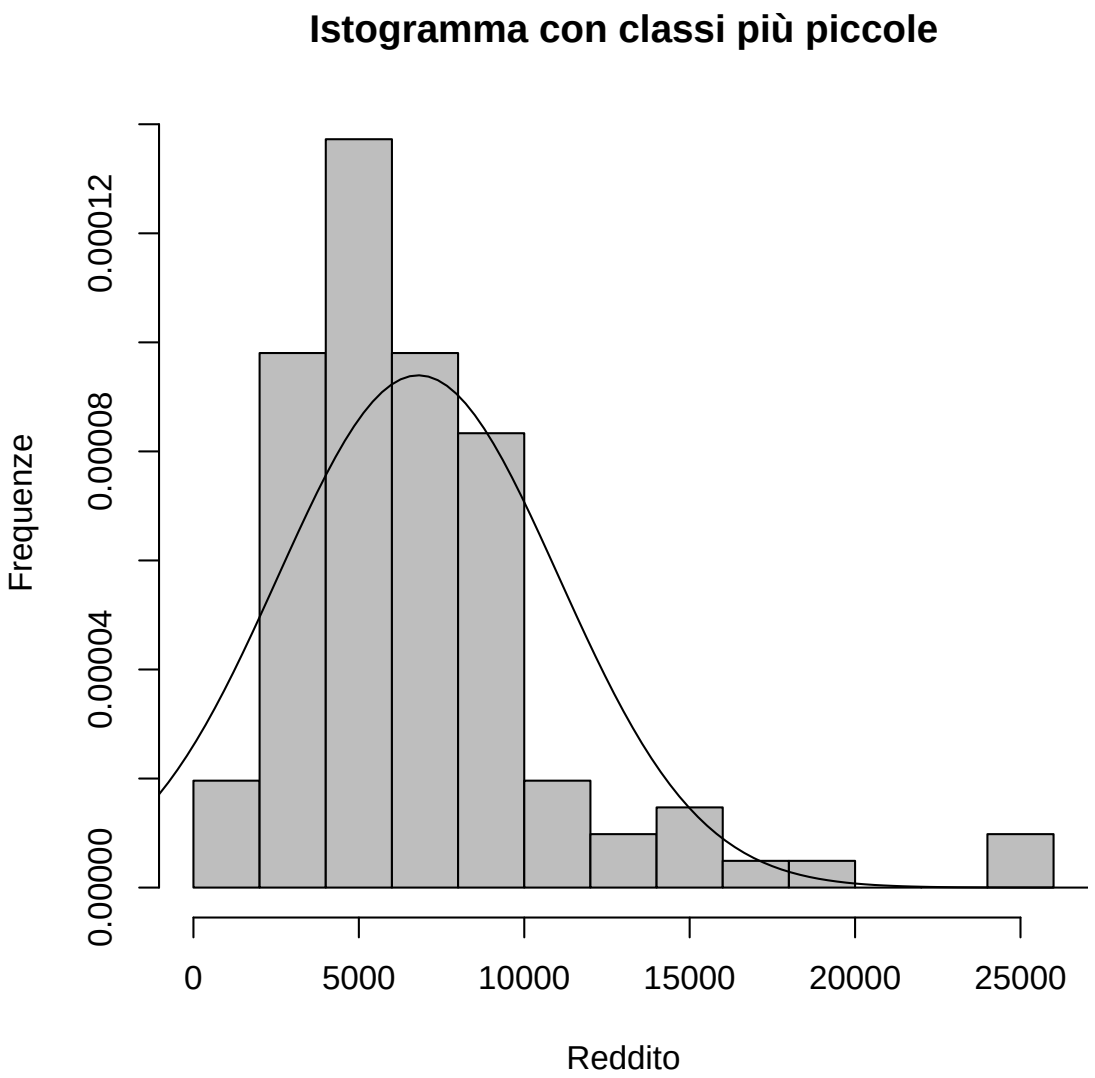


- una distribuzione è simmetrica se è speculare rispetto alla metà
- asimmetrica se non lo è (negativa=sinistra; positiva=destra)

Forme di distribuzione

- la forma grafica della distribuzione permette di identificare alcune sue caratteristiche
- i picchi rappresentano i valori più frequenti
- si classificano in *unimodali*, *bimodali* e multimodali
- si identificano facilmente le distribuzioni simmetriche o vicino alle simmetriche
- Una particolare curva simmetrica con forma a campana è chiamata “normale” o “gaussiana”
- Gaussiana perché studiata da Gauss
- Normale perché moltissime misurazioni di eventi fisici e/o naturali si distribuiscono con questa forma, che diventa un punto di riferimento in tutta la statistica

Istogramma con curva normale (I/R)



Cos'è un foglio elettronico

- Un foglio elettronico è una tabella in cui si possono fare calcoli in tempo reale: cambiando un valore in una cella, cambia il contenuto di eventuali celle collegate (ad es. una somma)
- In una cella di un foglio elettronico, oltre a inserire del testo o un numero, si possono inserire "istruzioni" che indicano come usare il contenuto di altre celle per fare calcoli o altre "manipolazioni" (ad es. `=SOMMA(A1:a5)`)
- Excel, OpenCalc, LibreCalc sono alcuni fogli elettronici che fanno parte di *suite* come "Office", "OpenOffice", "LibreOffice"
- OpenOffice e LibreOffice sono software "open" (quindi gratuiti).

Cos'è R

- R è un software statistico "open source", molto potente e sofisticato, programmabile
- È come avere a disposizione una mega calcolatrice che permette di svolgere sia calcoli semplici e banali sia analisi molto complesse
- Si può scaricare dal sito "The Comprehensive R Archive Network" (abbreviato in CRAN, `http://cran.r-project.org/` o da un suo "mirror")
- Oltre alle guide disponibili in CRAN, potete utilizzare la mia dispensa "L'uso di R in psicologia" (versione provvisoria) disponibile a `http://www.germanorossi.it/mi/file/pdf/RxPsi.pdf`

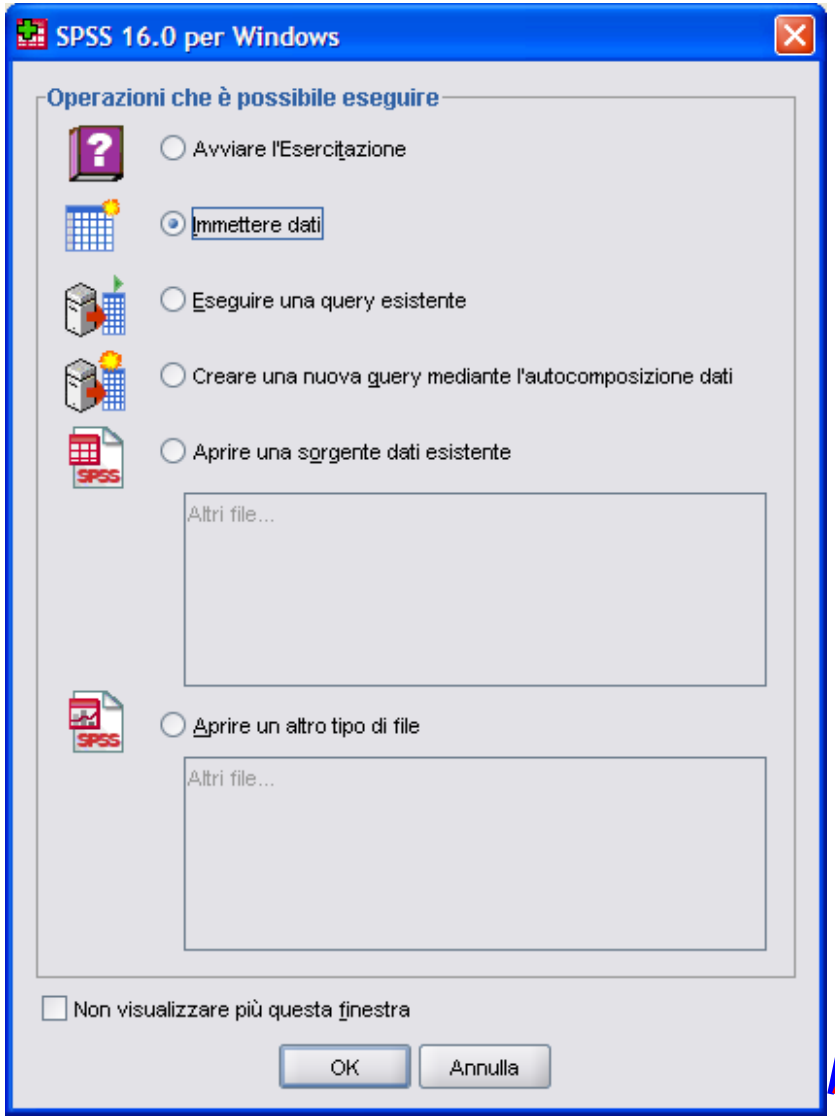
Cos'è SPSS

SPSS significa **Statistical Package for Social Sciences** (Pacchetto statistico per le scienze sociali) ed è un programma per gestire dati e calcolare le statistiche. Ma adesso di chiama PASW

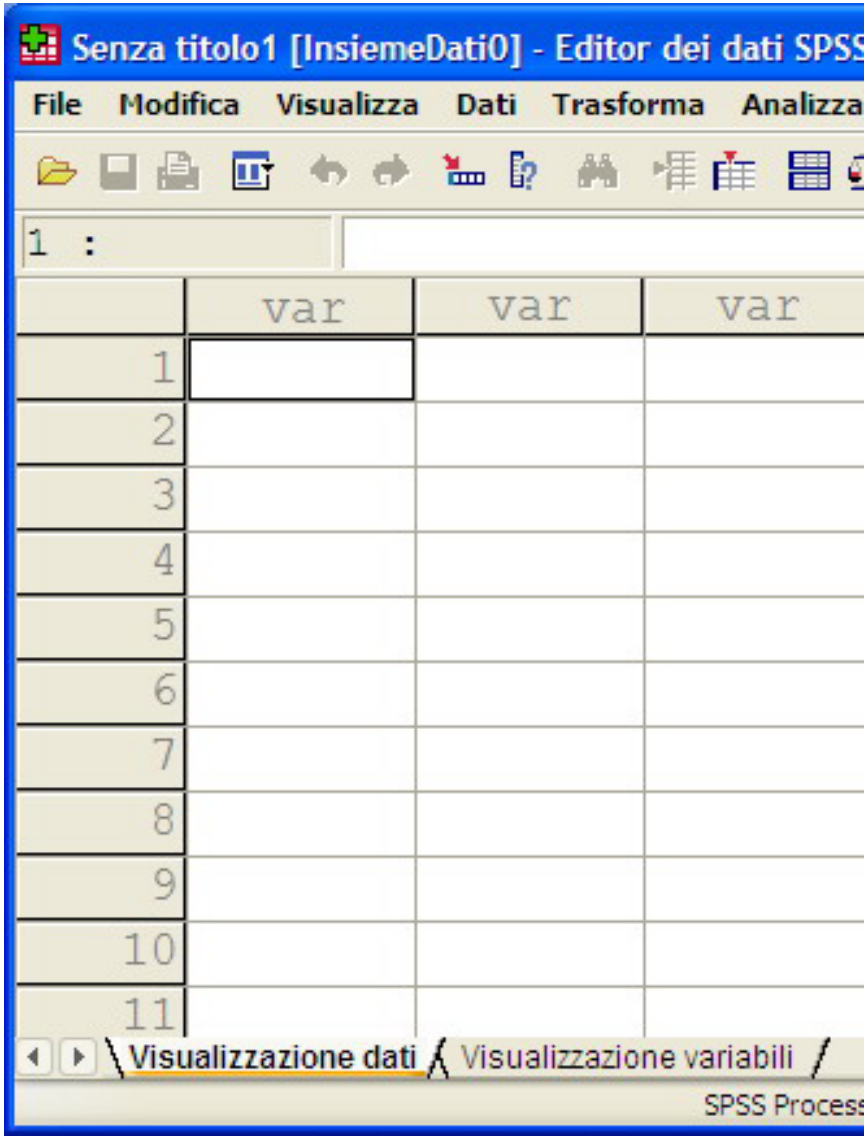
- Permette di inserire, nominare e gestire le misurazioni statistiche
- Permette di manipolare le variabili (in modo analogo ai fogli elettronici
- Permette di calcolare le statistiche descrittive, di fare grafici
- Permette di fare l'analisi dei dati

Primo avvio

- Per eseguire SPSS (Win XP)
  `Start | Tutti i programmi | Spss per Windows | Spss`
- Selezionate “Immettere dati”
- Alla prima esecuzione compare una finestra di dialogo che permette di effettuare alcune scelte. Per disattivarla, click-ate su “Non visualizzare più questa finestra”
- poi OK



Finestra di base



- La finestra principale è formata da due pannelli
- uno per la visualizzazione dei dati
- uno per la descrizione delle variabili

Finestra di base

Le principali voci di menù sono:

- Dati: permette di agire sul file dei dati (ordinare, selezionare, filtrare...)
- Trasforma: permette di manipolare le variabili (calcolare nuove variabili, ricodificare, contare valori...)
- Analizza: È il menù più utilizzato perché contiene tutte le procedure statistiche disponibili

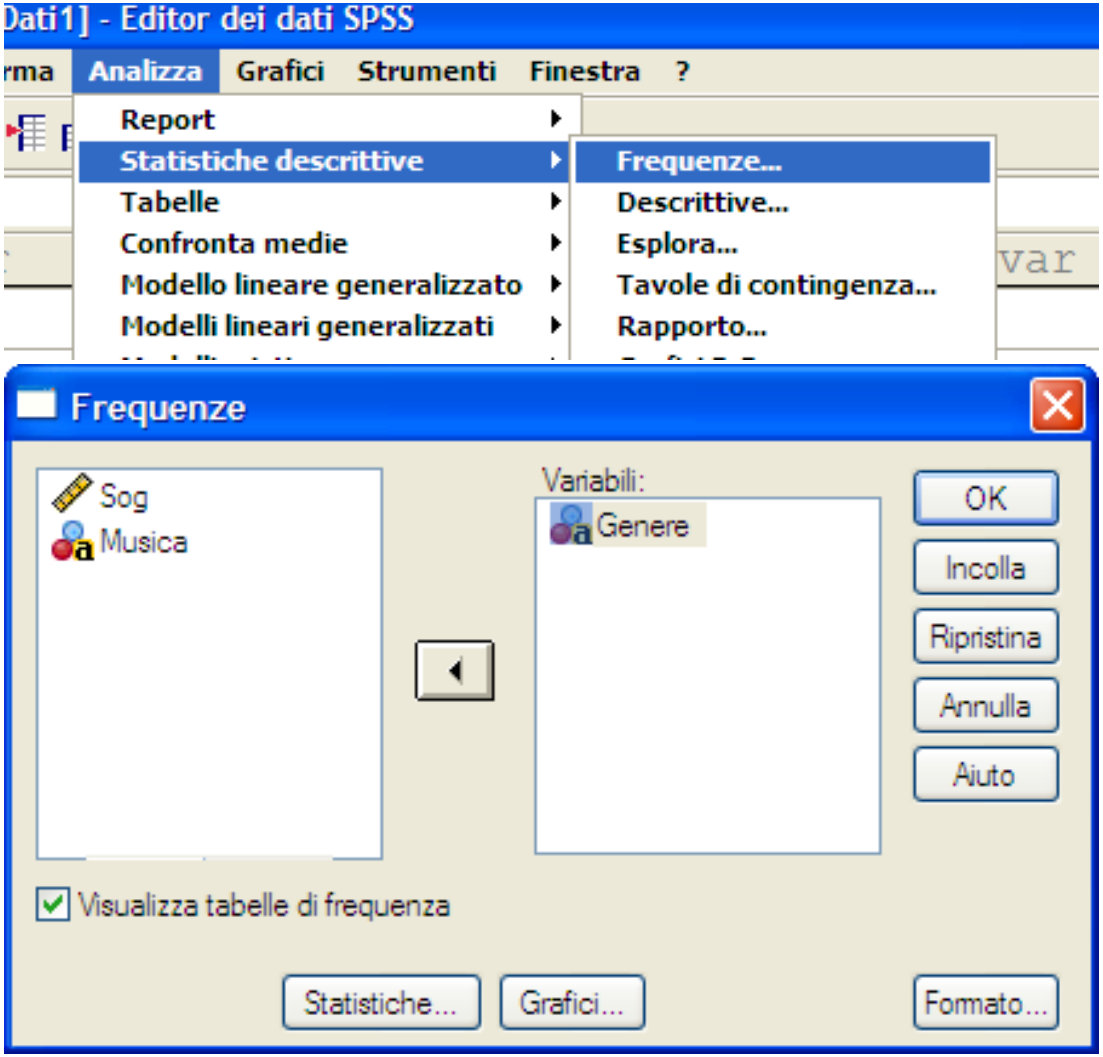
La prima volta che usate SPSS vi conviene fare l'**esercitazione** disponibile nell'Help.
Per indicare un percorso di menù, userò questa convenzione `voce principale | sottomenù | sotto-sottomenù`:
ad es. `Aiuto | Esercitazione`

Distribuzione e grafici in SPSS

- La maggior parte delle statistiche che abbiamo visto e che vedremo e i relativi grafici, si ottengono in SPSS tramite due comandi
  - `Analizza | Statistiche descrittive | Frequenze...`
  - `Analizza | Statistiche descrittive | Esplora...`
- In Spss si può fare solo quello che è previsto dal software; i grafici non sono bellissimi
- In Excel si possono fare molti più grafici (anche più “belli”)
- In R (usato in questi lucidi) si può fare praticamente tutto, ma è molto più complicato

Spss: frequenze

- Calcoliamo le frequenze con il comando `Analizza | Statistiche descrittive | Frequenze...`
- Poi spostiamo una variabile nominale/ordinale fra le Variabili
- E premiamo OK



Spss: frequenze

Frequenze

```
[InsiemeDati1] C:\Documenti\TeX\lucidi\Elem\Fig\esempio_dati_x_lucidi.sav
```

Statistiche

Genere

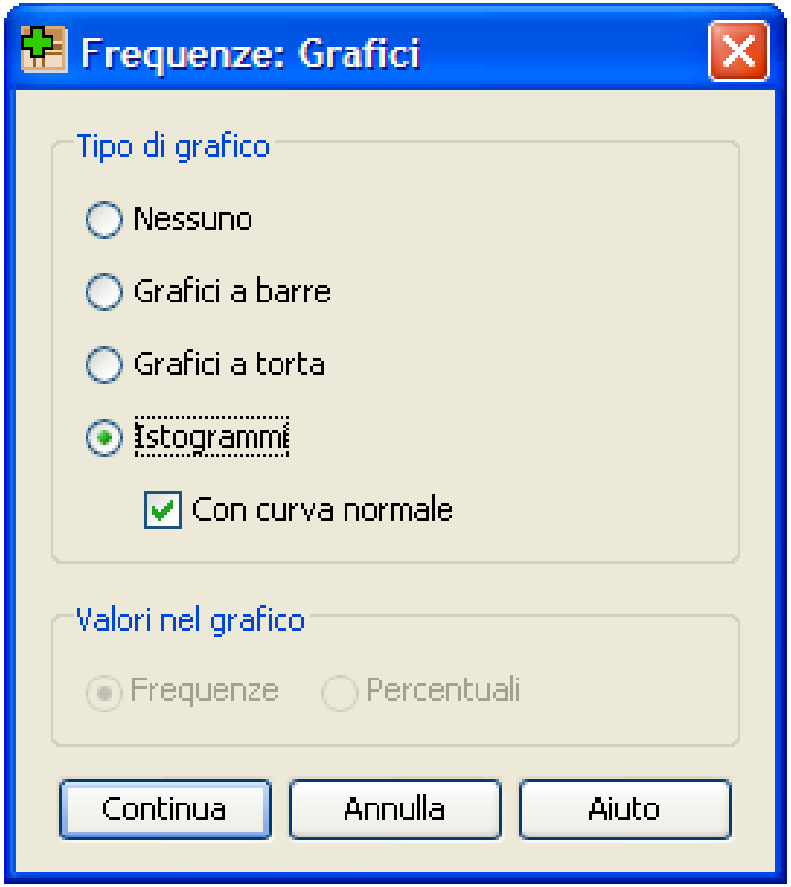
N	Validi	30
	Mancanti	0

Genere

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	F	13	43,3	43,3	43,3
	M	17	56,7	56,7	100,0
	Totale	**30**	**100,0**	**100,0**	

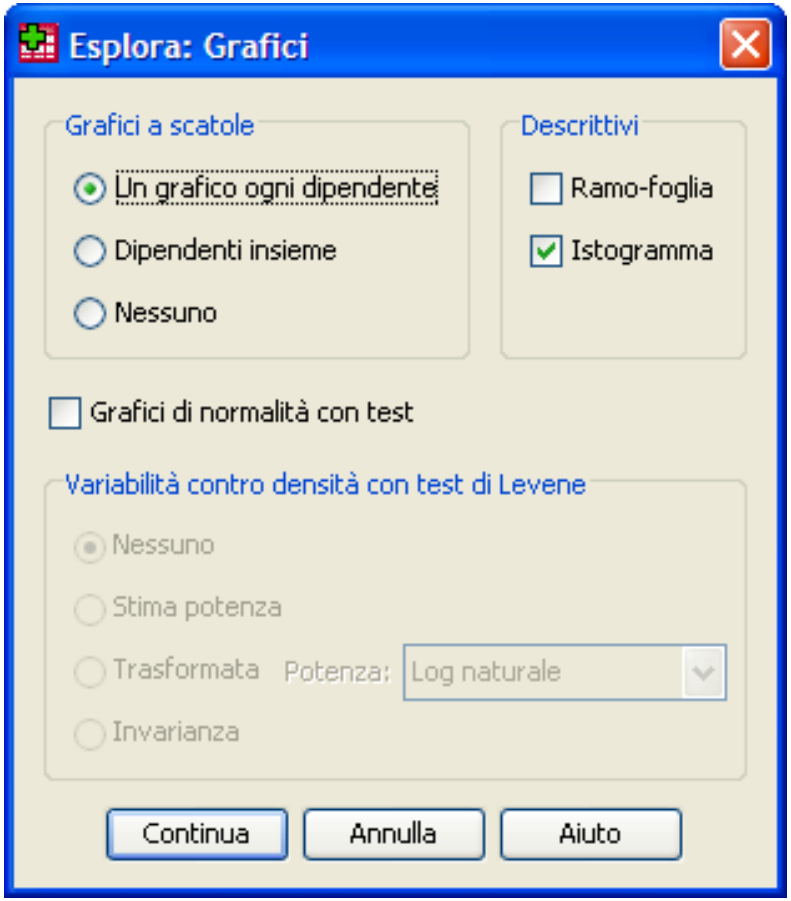
Spss: istogramma (Frequenze)

- `Analizza | Statistiche descrittive | Frequenze...`
- Pulsante `Grafici...`
- Selezioniamo `Istogramma`
- eventualmente anche `Con curva normale`
- Pulsante `Continua`
- e `OK`

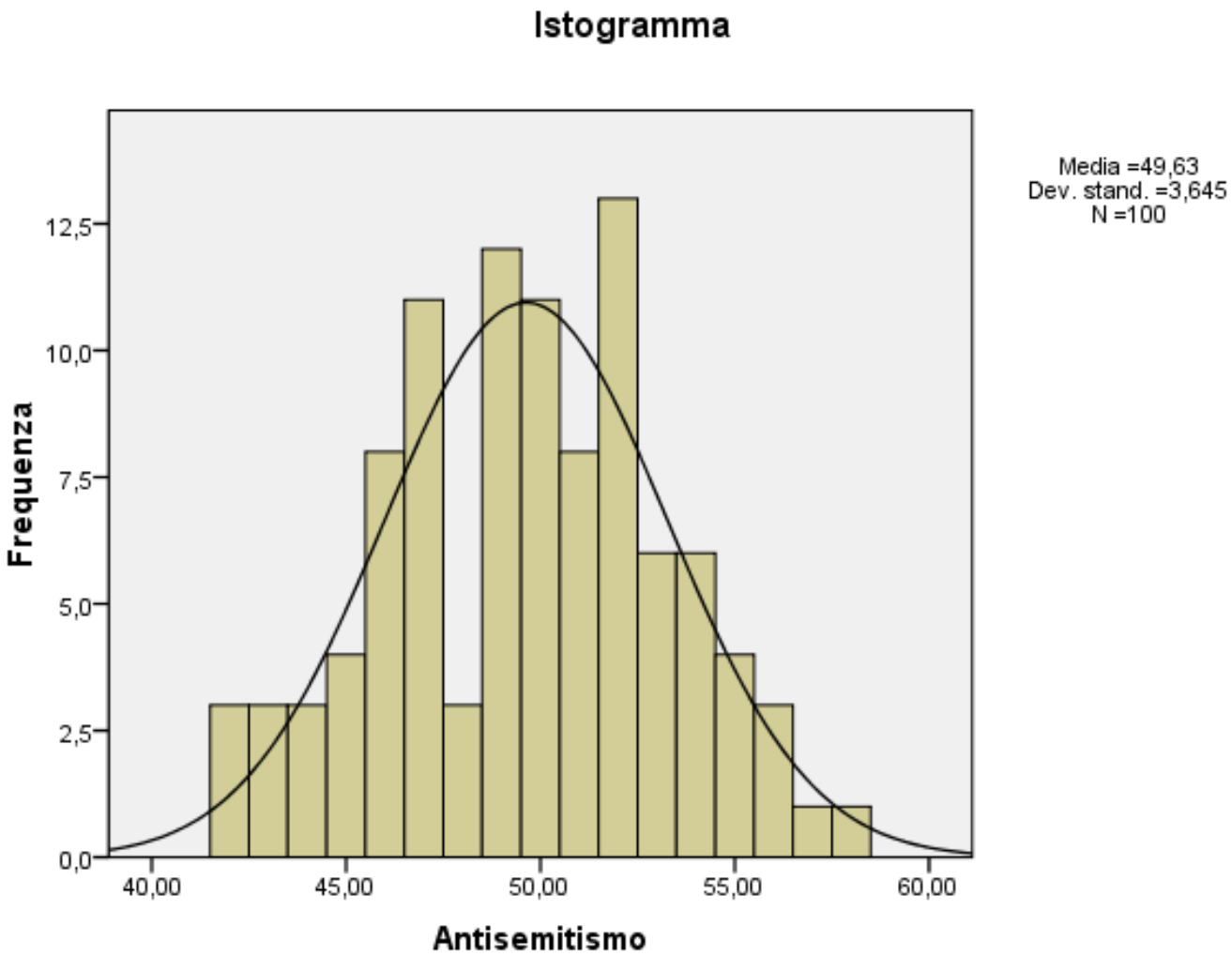


Spss: istogramma (Esplora)

- `Analizza | Statistiche descrittive | Esplora...`
- Pulsante `Grafici...`
- Selezioniamo `Istogramma`
- (non è possibile avere la curva normale)
- Pulsante `Continua`
- e `OK`



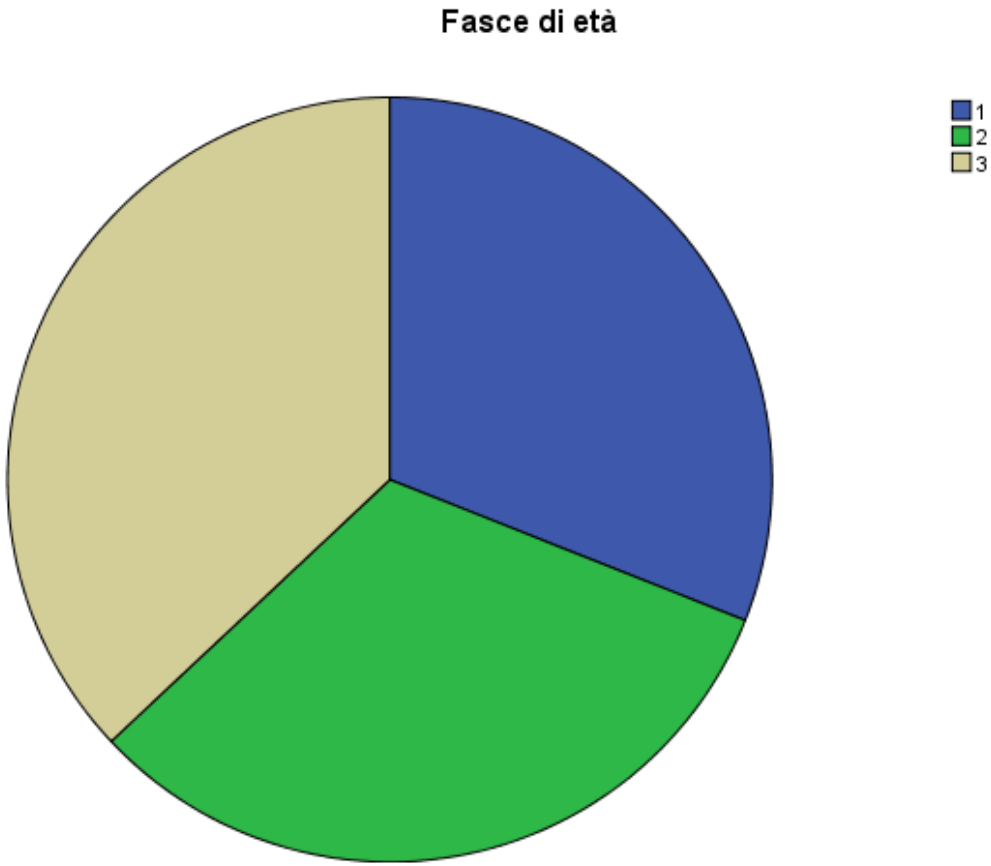
Spss: istogramma (Frequenze/Esplora)



- Spss usa automaticamente gli intervalli di classe
- non è possibile modificare l'ampiezza di classe
- La curva normale non c'è con `Esplora`

Spss: torta

- `Analizza | Statistiche descrittive | Frequenze...`
- Pulsante `Grafici...`
- Selezioniamo `Grafici a torta`
- Pulsante `Continua`
- e `OK`



Spss: Steam-leaf (Esplora)

- `Analizza | Statistiche descrittive | Esplora...`
- Pulsante `Grafici...`
- Selezioniamo `Ramo-foglia`
- Pulsante `Continua`
- e `OK`