

# Elementi di Psicometria con Laboratorio di SPSS 1

E1-Trattamento dei dati  
(vers. 1.3, 17 novembre 2015)  
versione per stampa

Germano Rossi<sup>1</sup>

`germano.rossi@unimib.it`

<sup>1</sup>Dipartimento di Psicologia, Università di Milano-Bicocca

2015-16

# Dalla ricerca all'analisi dei dati

- Finora abbiamo visto diversi metodi di analisi dei dati
- Ma come nasce una ricerca?
- In genere, il ricercatore si interessa di un certo argomento
- Raccoglie la bibliografia su quell'argomento
- Formula un'ipotesi generale
- Sulla base dell'ipotesi, seleziona gli strumenti da usare per le misurazioni delle variabili
- Quindi formula le ipotesi specifiche e pensa alle analisi dei dati che possono dare risposta alle sue ipotesi
- Raccoglie i dati
- Effettua l'analisi dei dati
- Scrive un report della sua ricerca (articolo, tesina, tesi. . .)

# Le ipotesi di ricerca

- La statistica, oltre a descrivere i dati, serve per fare la verifica delle ipotesi
- In psicologia, la ricerca “indaga” le “regole” tramite “esperimenti” o “indagini”
- Ogni ricerca si basa su un’ipotesi di lavoro che “ipotizza” determinati legami fra determinate variabili
- In quasi tutte le ricerche ci sono variabili dipendenti e variabili indipendenti
- Si assume che le variabili dipendenti misurino l’oggetto della ricerca
- e che le indipendenti siano variabili che hanno o potrebbero avere una qualche forma di influenza sulle dipendenti
- Le misure delle variabili possono essere di tipo molto diverso

# Misurazione delle variabili

- Certe ricerche usano poche variabili indipendenti (ricerche più “sperimentali”) e cercano di tenere sotto controllo tutte le altre possibili variabili influenti
- Altre utilizzano svariate variabili indipendenti (perché non è sempre possibile controllarle e quindi si fanno variare il più possibile)
- I questionari sono particolarmente utilizzati nelle “indagini” (o *survey*) in cui si raccolgono molte variabili conoscitive
- I questionari possono essere cartacei o raccolti a computer (ad es. domande poste da un PC, tempi di reazione a stimoli. . . )
- Ci occuperemo qui dei questionari cartacei.

# Misurazione delle variabili

- Dal momento che le variabili qualitative (nominali e ordinali) sono meno informative di quelle quantitative (intervallo/rapporto) si cerca sempre di costruire un questionario che possa essere associato a variabili quantitative.
- Ma non sempre è possibile o opportuno
- Non ha senso considerare il **genere** come una variabile quantitativa
- ma la propria **percezione di mascolinità o femminilità** può essere una variabile quantitativa
- ma **misurano concetti diversi**

---

## 1. Genere:

- Maschio
  - Femmina
- 

---

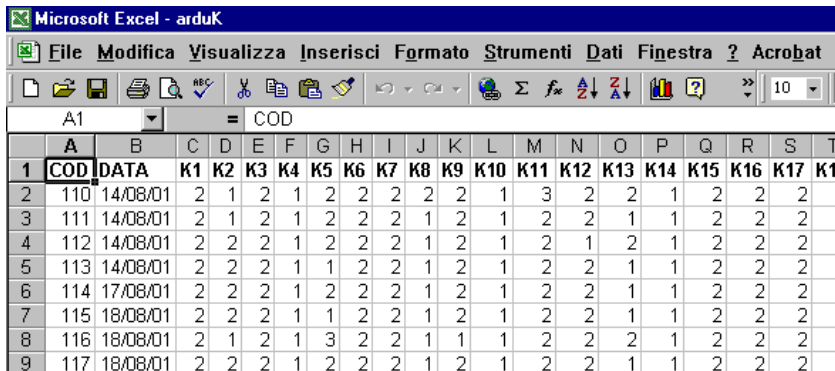
## 7. Quanto ti senti “maschile” o “femminile”?

maschio   5   4   3   2   1   0   1   2   3   4   5   femmina

---

# Inserire i dati raccolti

- Il modo migliore di inserire i dati è quello di usare un foglio elettronico (ad es. Excel o LibreCalc) in cui ogni questionario è rappresentato da una riga e ogni variabile da una colonna
- La prima riga conterrà i nomi delle variabili



The screenshot shows a Microsoft Excel window titled "Microsoft Excel - arduK". The menu bar includes File, Modifica, Visualizza, Inserisci, Formato, Strumenti, Dati, Finestra, and Acrobat. The toolbar contains various icons for file operations, editing, and data analysis. The active cell is A1, which contains the formula "= COD". The spreadsheet displays data for 9 rows (1-9) and 20 columns (A-T). The first row (row 1) contains the following data: A: COD, B: DATA, C: K1, D: K2, E: K3, F: K4, G: K5, H: K6, I: K7, J: K8, K: K9, L: K10, M: K11, N: K12, O: K13, P: K14, Q: K15, R: K16, S: K17, T: K1. The subsequent rows (2-9) contain numerical data for each variable.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	COD	DATA	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12	K13	K14	K15	K16	K17	K1
2	110	14/08/01	2	1	2	1	2	2	2	2	2	1	3	2	2	1	2	2	2	
3	111	14/08/01	2	1	2	1	2	2	2	1	2	1	2	2	1	1	2	2	2	
4	112	14/08/01	2	2	2	1	2	2	2	1	2	1	2	1	2	1	2	2	2	
5	113	14/08/01	2	2	2	1	1	2	2	1	2	1	2	2	1	1	2	2	2	
6	114	17/08/01	2	2	2	1	2	2	2	1	2	1	2	2	1	1	2	2	2	
7	115	18/08/01	2	2	2	1	1	2	2	1	2	1	2	2	1	1	2	2	2	
8	116	18/08/01	2	1	2	1	3	2	2	1	1	1	2	2	2	1	2	2	2	
9	117	18/08/01	2	2	2	1	2	2	2	1	2	1	2	2	1	1	2	2	2	

# Inserire i dati raccolti

- La maggior parte dei software preferisce lavorare con valori numerici e limita i valori “stringa” (alfanumerici) alle variabili “nominali”
- Tendenzialmente si utilizzano quindi solo numeri
- Il modo migliore è di inserire i dati secondo un approccio *dall'alto al basso e da sinistra a destra* per ogni variabile
- Per cui, 1=Nubile/celibe, 2=Sposato, . . . 5=Divorziato, 6=Convivente. . .
- Per le “non risposte” si lascia la cella vuota

---

## 10. Qual è il tuo stato civile?

- |   |   |
|---|---|
| <input type="radio"/> Nubile/celibe                             | <input type="radio"/> Divorziata/o                |
| <input type="radio"/> Sposata/o                                 | <input type="radio"/> Convivente ma non sposata/o |
| <input type="radio"/> Sposata/o ma separata/o                   | <input type="radio"/> Vedova/o                    |
| <input type="radio"/> Sposata/o ma convivente con altra persona |   |
-

# Inserire i dati raccolti

- Per alcune variabili (ad es. la professione) si può decidere di
  - 1 lasciare uno spazio per poter scrivere la professione
  - 2 classificare *a priori* le professioni in categorie, obbligando il rispondente a scegliere solo fra queste
- nel primo caso, si trascrive esattamente quello che è stato scritto (tutto in minuscolo o tutto in maiuscolo, perché per molti software “Casa”, “casa” e “CASA” sono diversi)
- nel secondo si segue la regola generale (dall’alto al basso. . . )

---

## 6. Qual è la tua professione?

---

---

## 6. Qual è la tua professione?

- ☐ Pensionato/Casalinga/Disoccupato
  - ☐ Operaio
  - ☐ Impiegato/Insegnante
  - ☐ ...
-



# Pulizia dei dati

- Prima di effettuare qualunque analisi dei dati (statistiche descrittive, statistiche inferenziali. . . ) è bene verificare che i dati siano corretti e adeguati per quella tecnica
- Fra le cose da controllare ci sono:

slide	<b>Tipo di controllo</b>	<b>valido per</b>
10	Errori di immissione	tutte le variabili
13	Coerenza dei dati	
16	Valori mancanti	
25	Valori poco frequenti	variabili qualitative
26	Normalità	variabili quantitative
39	Valori anomali	

# Errori di immissione

- Nell'inserire i dati di un questionario si possono fare errori di immissione
- Prima di iniziare l'analisi dei dati è meglio fare un controllo
- Un primo controllo andrebbe effettuato a mano, prendendo casualmente un 10% dei questionari e controllando domanda per domanda.
- Per ogni questionario con un errore di immissione, si controlla un questionario aggiuntivo
- Se i dati sono stati inseriti tramite Excel, controllare che il tipo dei dati sia coerente (numeri=allineati a destra, testo=stringa=allineati a sinistra)

# Errori di immissione

- Una seconda possibilità (spesso usata per prima) è di controllare i valori minimi e massimi delle variabili
- Gli errori di immissione più frequenti (soprattutto con le scale Likert) sono la doppia pressione di un tasto
- Come conseguenza della doppia pressione potrebbe esserci un valore mancante in fondo allo strumento

---

## Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione std.
ATG1	454	1	7	3.84	1.938
ATG2	453	1	22	3.43	2.238
ATG3	452	1	7	3.41	1.922
ATG4	451	0	7	2.52	1.729
ATG5	455	1	11	2.00	1.532
ATG6	447	1	7	4.33	1.999

---

# Errori di immissione in SPSS

- Per verificare i principali errori di immissione si possono chiedere i valori minimo e massimo di tutte le variabili:

```
DESCRIPTIVES VARIABLES=ALL
/STATISTICS=MEAN STDDEV
MIN MAX.
```

Analizza | Statistiche  
descrittive | Descrittive

- Per identificare i casi con valori balordi, si seleziona temporaneamente il campione per controllare il codice-caso:

```
TEMPORARY.
SELECT IF ATG2>7.
LIST VARIABLES=cod ARG2.
```

Dati | Seleziona casi...

- Se la condizione è soddisfatta

Se..

ATG4=0

OK

OK

Scorro i dati in cerca dei valori non barrati (/)

# Coerenza dei dati

- La coerenza delle risposte è importante!
- Se ci sono domande del tipo “Se hai risposto. . . allora. . .”, bisognerebbe controllare che non siano state date risposte incoerenti
- Con questionari lunghi possono esserci effetti di stanchezza che portano ad usare il *response-set*, cioè i rispondenti possono aver risposto in modo più o meno casuale oppure aver indicato lo stesso valore per molti item
- Se il questionario conteneva scale con item pro- e contro-tratto, bisogna verificare che le risposte pro- siano coerenti con quelle contro-
- Alcuni di questi controlli sono fattibili direttamente in SPSS

# Coerenza dei dati in SPSS

- Si possono trovare i casi che violano i pro/contro:

```
COMPUTE sdG=SD(G1 TO G7).
EXECUTE.
TEMPORARY.
SELECT IF (sdG<=0.5).
list var=cod sdG G1 TO G7.
```

Trasforma | Calcola  
variabile...

sdG SD(G1 TO G7)

Dati | Seleziona casi...

- Se la condizione è soddisfatta

Se.. sdG<.5 OK OK

- Una sequenza numerica tipo 1 2 1 2 1 2 ha una deviazione standard piccola
- Se alcuni item sono contro-tratto la sequenza dovrebbe essere 1 6 1 2 7 (con una deviazione standard più alta)

# Coerenza dei dati in SPSS

- E si possono azzerare i valori (o ricodificarli come “mancanti”):

```
DO IF (sdG=0).  
RECODE G1 TO G7 (ELSE=SYSMIS).  
END IF.  
EXECUTE.
```

Trasforma | Ricodifica nelle  
stesse variabili...

G1 G2 G3 G4 G5 G6 G7

Valori vecchi e nuovi

Tutti gli altri valori -> Mancante di sistema

Continua

Se

Includi se il caso soddisfa la  
condiziona sdG=0

Continua

OK

# Valori mancanti

- Si definiscono **valori mancanti** (o *missing*) le misurazioni che non sono state raccolte per i più svariati motivi
- Se una *variabile* ha molti mancanti (in genere più del 5-6%), ci si dovrebbe chiedere perché
- Se un *soggetto* ha molti mancanti, varrebbe la pena di eliminare tutte le sue risposte
- SPSS distingue fra
  - *mancanti di sistema (SYSMIS, system missing)*: quando l'informazione non si conosce proprio e non è stata inserita (la cella è vuota; Spss la rappresenta con un puntino)
  - *mancanti definiti dall'utente (MISSING, user missing)*: quando il valore esiste ed è stato inserito, ma noi vogliamo trattarlo come se fosse mancante (ad es. risposte particolari come “non voglio rispondere”)
  - *mancanti (missing)*: entrambi i tipi senza distinguerli



# Valori mancanti

- Da un punto di vista teorico i mancanti vengono generalmente classificati in tre categorie:
  - 1 **Valori mancanti completamente casuali** (MCAR): quando i mancanti non dipendono né dalla variabile in sé né dalle altre variabili misurate
  - 2 **Valori mancanti casuali** (MAR): quando i mancanti non dipendono dalla variabile, ma dipendono da una o più delle altre variabili
  - 3 **Non mancanti a caso** (NMAR) ovvero valori mancanti inaccessibili o non ignorabili perché dipendono dalla variabile stessa (e forse anche dalle altre variabili): quando vi è un motivo più o meno evidente perché il dato manchi
- Per ognuno di queste categorie ci sarebbe un modo per ovviare. . .
- Purtroppo questi tre tipi di mancanti si presentano spesso mischiati fra loro e quindi poco utili

# Mancanti in Spss

- I mancanti di sistema corrispondono alle celle vuote
- È possibile definire un valore come mancante se la frequenza di quella categoria è molto bassa oppure se il valore inserito è assurdo (oltre i limiti teorici) e non si può confrontarlo con il protocollo di raccolta
- Per inserire un valore come mancante, andare in `Visualizzazione variabili`, avanzare fino alla colonna `Mancante/i` e inserirlo
- In alternativa, si può usare la sintassi `MISSING VALUES nome (valore)` ad es. `MISSING VALUE Rif (3)`.
- si possono avere 3 valori mancanti oppure un intervallo di valori continui

# Valori mancanti: strategie

Premesso che sarebbe meglio evitare il più possibile i valori mancanti. . .

- Se la variabile è la somma di x item, si può evitare il mancante con una ponderazione delle risposte
- In SPSS, usando il + se una delle variabili è mancante, il totale sarà mancante.

```
COMPUTE nuova1 = Altm_1 + Altm_2 + Altm_3 + Altm_4.
```

- In SPSS, usando la funzione SUM, se una della variabili è mancante non verrà usata quella variabile (ma la somma sarà calcolata su una gamma più piccola).

```
COMPUTE nuova2 = SUM(Altm_1, Altm_2, Altm_3, Altm_4).
```

- Ponderare

```
COMPUTE nuova3 = MEAN(Altm_1, Altm_2, Altm_3, Altm_4)*4.
```

Ci sono altre strategie per gestirli.

# Valori mancanti: strategie

Ognuna di queste strategie ha dei pregi e dei difetti

- metodo *listwise*: si “buttano” tutti i casi con valori mancanti
- Il campione potrebbe ridursi drasticamente!

COD	ATG24	ATG25	ATG26		COD	ATG24	ATG25	ATG26
504	2	4	1		504	2	4	1
505	4	4						
506	2	2	4		506	2	2	4
507		1	1					
508	4	1	1		508	4	1	1
509	4		4					
510	4	4	1		510	4	4	1
511	4	4	3	⇒	511	4	4	3
512	5	3	2		512	5	3	2
513	7	1	3		513	7	1	3
514	3	5						
515	4	1	1		515	4	1	1

# Valori mancanti: strategie

- metodo *pairwise* (esclusione casi test per test, a coppie, analisi per analisi. . .): si “ignorano” i soggetti con valori mancanti, limitatamente alle statistiche calcolate di volta in volta
- con certe tecniche d'analisi, si perde la concomitanza delle risposte

COD	ATG24	ATG25	ATG26	Usati
504	2	4	1	tutti
505	4	4		solo con 24 e 25
507		1	1	solo con 25 e 26
509	4		4	solo con 24 e 26
515	4	1	1	tutti

# Valori mancanti: strategie

- si sostituisce il *missing* con la media della variabile (diminuisce la dev.st.) o del soggetto (se le variabili misurate hanno la stessa metrica) o con altri metodi, come la media o la mediana dei punti vicini (in Spss, `Trasforma | Sostituisci valori mancanti`)
- si sostituisce con una stima ottenuta da una regressione o da interpolazione lineare sui casi (in Spss, come sopra)
- si sostituisce per somiglianza con altri casi che hanno risposte analoghe sulle altre variabili [potrebbe non esistere una soluzione] (non si può fare in Spss)
- Si usa la procedura EM (*Expectation, Maximization*) basata sull'iterazione fra stima dei valori mancanti e la verifica di massimaverosimiglianza (in Spss, `Analizza | Analisi dati mancanti`).

# Valori mancanti: in SPSS

- Analizza | Analisi dei valori mancanti, in Modelli  
**selezionare** Casi con valori mancanti, ordinati per  
 modello

	N	Media	Dev. std.	Mancante		N. di estremi <sup>(a)</sup>	
				Max	Conteggio	Perc.	Min
x1	196	3,24	1,286	4	2,0	0	0
x2	189	2,96	1,279	11	5,5	0	0
x3	180	2,83	1,156	20	10,0	0	0
x4	186	3,42	1,330	14	7,0	0	0
x5	194	3,72	1,136	6	3,0	0	0
x6	179	3,21	1,498	21	10,5	0	0
x7	178	3,25	1,139	22	11,0	0	0
x8	171	3,12	1,280	29	14,5	0	0
x9	178	2,39	1,208	22	11,0	0	0
x10	182	3,70	1,113	18	9,0	0	0

<sup>a</sup>. Numero di casi non compresi nell'intervallo ( $Q1 - 1.5 * IQR$ ,  $Q3 + 1.5 * IQR$ ).

# Valori mancanti: in SPSS

Caso	# Manc.	% Manc.	Modelli di valori mancanti ed estremi(a)									
			x5	x4	x10	x3	x9	x7	x8	x6	x1	x2
13	1	10,0									S	
17	1	10,0									S	
...	...	...										
104	5	50,0			S	S	S	S			S	
97	6	60,0		S		S	S	S	S	S		
26	9	90,0	S	S		S	S	S	S	S	S	S
43	10	100,0	S	S	S	S	S	S	S	S	S	S
70	7	70,0	S	S		S		S	S		S	S

- Se la scala x1-x10 fosse fondamentale, il caso 43 andrebbe buttato
- eventualmente anche il caso 26



# Valori poco frequenti

- Quando si usano determinate variabili come indipendenti, bisogna prima verificare che i valori raccolti siano tutti utilizzabili
- Es. se una variabile avesse 4 categorie (N=160) e una di queste categorie avesse frequenza 2, questa categoria sarebbe inutilizzabile in qualunque tipo di analisi successiva (chi quadro, t-test, ecc.)
- Meglio quindi dichiarare quel valore come “mancante” ed eliminarlo

Valore	Frequenza
1	38
2	73
3	2
4	47

```
MISSING VALUE
nomevar (3) .
```

# Normalità

- La distribuzione normale di una variabile (univariata) permette di avere una serie di informazioni sulla distribuzione della variabile stessa
- Media, mediana e moda coincidono
- La deviazione standard coincide con i punti di flesso
- Fra -1 e +1 dev. st. abbiamo il 68% (circa) dei dati
- Fra -2 e +2 dev. st. abbiamo il 95% (circa) dei dati
- Se una variabile da noi misurata si distribuisse in modo “normale” avremmo molte informazioni a partire dalla sola indicazione di media e dev.st.
- Inoltre molte statistiche inferenziali assumono che le variabili quantitative si distribuiscano normalmente (ad es. il t-test)

# Verificare la normalità

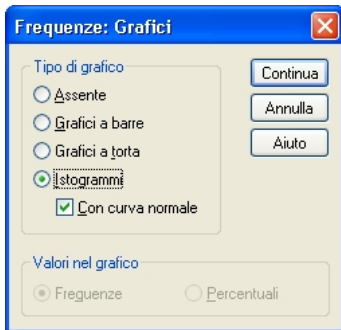
- È possibile valutare la normalità di una variabile misurata, tramite la rappresentazione grafica dei dati o tramite alcuni indici statistici.
- Se la mediana è parecchio distante dalla media, la distribuzione è asimmetrica (a destra o a sinistra della media)
- Se la dev. st. è molto grande, la distribuzione è piatta; se è molto piccola, è acuta
- Anziché andare a senso logico (o peggio, “ad occhio”), possiamo usare:
  - due indici statistici: **curtosi** (o *kurtosis*) e **asimmetria** (o *skewness*)
  - due test statistici: test di Kolmogorov-Smirnov e test di Shapiro-Wilk
  - un criterio pragmatico

# Istogramma con normale

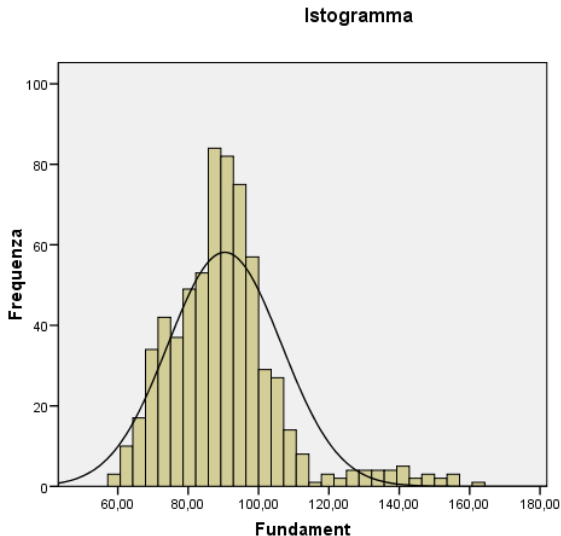
- Guardando un istogramma (con i valori raccolti in intervalli) e sovrapponendo poi la curva normale costruita con la media e la dev.st. della variabile
- possiamo avere una idea

- In Spss usare

Analizza | Statistiche descrittive | Frequenze, poi **Grafici**, scegliere Istogrammi e click-are su Con curva normale, eventualmente spuntare Visualizza tabelle di frequenza nel dialogo precedente



# Istogramma con normale



- La curva è vagamente asimmetrica, ma abbastanza?

# Esplora di SPSS

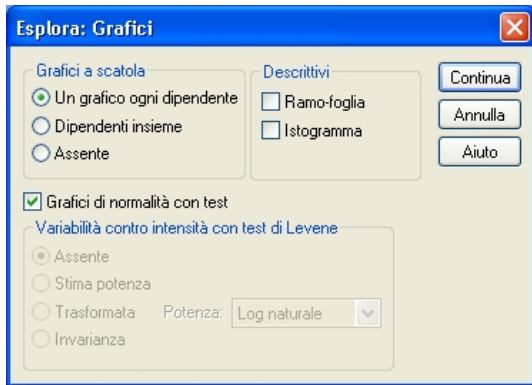
- Usando il comando `Analizza | Statistiche descrittive | Esplora` possiamo ottenere altri grafici e statistiche che permettono di *stimare* la normalità
- Il grafico dei quantili
- Il test di Kolmogorov-Smirnov
- Il test di Shapiro-Wilk

# Grafici di normalità

- Usando il comando Analizza | Statistiche descrittive | Esplora, scegliendo Grafici e poi Grafici...

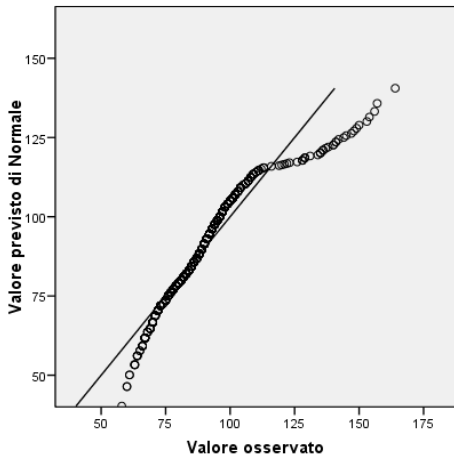
- togliere

Ramo-foglia  
e scegliere  
Grafici di  
normalità  
con test



# Grafico dei quantili

Grafico Q-Q Normale di Fundament



- Sulla X i valori osservati e sulla Y quelli attesi in base alla normale
- La diagonale rappresenta la distribuzione attesa
- La variabile si distribuisce normalmente se i punti sono lungo la diagonale
- È necessario avere un po' di esperienza per interpretarla



# Curtosi e asimmetria

- Questi due indici dovrebbero essere entrambi uguali a 0
- Ma la maggior parte delle volte sono diversi da 0 (poco o molto)
- Fino a che punto possono essere diversi da 0?
- Dividendo ciascun indice per il suo errore standard c
  - si ottiene un punto  $z$
  - che si confronta con il valore critico di  $z$  per un determinato  $\alpha$
  - se il suo valore assoluto è superiore a 1.96 è significativo al 5%

alfa	valore critico
.05	1.96
.01	2.57, 2.58

# Curtosi e asimmetria

	Fundament	Errore std	z	ES*2
N	659			
Media	90,3915			
Deviazione std.	16,14585			
Asimmetria	1,350	0,095	14.21	$0,095*2=0.19$
Curtosi	3,336	0,190	17.58	$0,190*2=0.38$

- Dividendo l'indice per il suo errore standard ( $1,350/0,095 = 14.21$  e  $3,336/0,190 = 17.58$ ), otteniamo un valore standardizzato che si interpreta come un punto z.
- l'errore standard dipende dalla numerosità (all'aumentare di N, diminuisce di valore)
- per questo motivo, il test **tende a sottostimare la normalità** nei campioni con N grande
- vedi criteri pragmatici

# Test di Kolmogorov-Smirnov e test di Shapiro-Wilk

- Se significativi, la variabile **non** si distribuisce normalmente

## Test di normalità

	Kolmogorov-Smirnov <sup>(a)</sup>			Shapiro-Wilk		
	Statistica	df	Sig.	Statistica	df	Sig.
Fundament	0,115	659	0,000	0,906	659	0,000

a. Correzione di significatività di Lilliefors

- Anche questi test tendono a rifiutare la normalità (soprattutto con grandi campioni) anche quando la violazione non sarebbe eccessiva per le statistiche parametriche
- vedi criteri pragmatici

# Criteri pragmatici

- Siccome i criteri statistici (punti z di asimmetria e curtosi, test di normalità) tendono a sovrastimare la a-normalità, si preferisce usare un *criterio pragmatico*
- Criterio **molto restrittivo**: si considerano come “accettabili” valori di asimmetria o di curtosi non superiori a 0.5 (in valore assoluto)
- Criterio **meno restrittivo** (il più utilizzato): si considerano come “accettabili” valori di asimmetria o di curtosi non superiori a 1 (in valore assoluto)
- Criterio **più lasso** (in genere in ambito clinico): si considerano come “accettabili” valori di asimmetria non superiori a 2 (in valore assoluto) e di curtosi non superiori a 5/7 (in valore assoluto)

Quando una variabile viola i criteri di asimmetria, può essere sottoposta a delle trasformazioni lineari che dovrebbero migliorare la distribuzione (rispetto ai criteri della normalità)

# Trasformazione dei dati

- Se una variabile è asimmetrica si può cercare di “aggiustarla” tramite delle trasformazioni lineari
- usando la statistica di asimmetria (e in base ai criteri utilizzati)

asimmetria	positiva	negativa
0 – 0.5	si accetta così	
0.5 – 1 (*)	$\log_{10}x, \sqrt{x}$	$\log_{10}(k - x), \sqrt{k - x}$
1 – 2	$\log_{10}x, \sqrt{x}$	$\log_{10}(k - x), \sqrt{k - x}$
> 2	$1/x$	$1/(k - x)$

(\*) si può accettare tranquillamente (senza trasformare) in base al criterio usato

$k$  è il valore massimo, che la variabile può assumere, più 1 ( $k=\max+1$ )

# Esempi di trasformazioni in SPSS

Ipotizzando `Fundament` come variabile da trasformare

- `COMPUTE flog=LG10 (Fundament)`
- `COMPUTE frad=SQRT (Fundament)`
- `COMPUTE frecipr=1/Fundament`

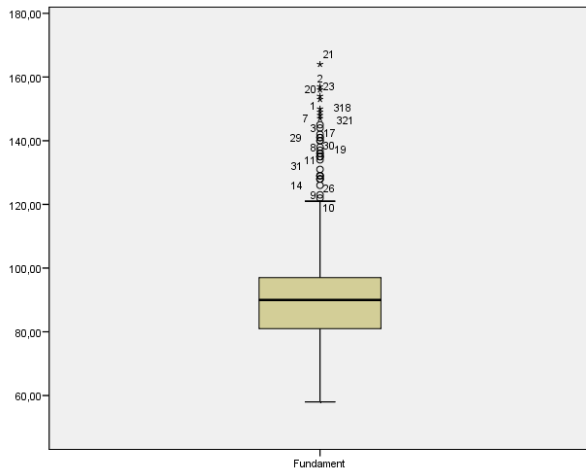
`max=164`

- `COMPUTE flog=LG10 (165-Fundament)`
- `COMPUTE frad=SQRT (165-Fundament)`
- `COMPUTE frecipr=1/ (165-Fundament)`

# Valori anomali “outliers”

- Gli *outliers* sono valori particolarmente anomali ovvero i valori molti inferiori o molto superiori alla media
- Possono essere considerati anomali i valori grezzi che corrispondono a  $|z| > 2$  (ma se la variabile non è normale)
- cioè i valori che (trasformati in punti z) superano 2 o 3 (in valore assoluto)
- Bisogna trasformare in punti z tutti i punteggi e quindi verificare quanti sono maggiori di 2 e/o di 3
- Se sono molti, potrebbe esserci un sottocampione “anomalo”
- È più semplice usare i `box-plot` o grafici a scatola. In SPSS, Analizza | Esplora

# Box-Plot



- La scatola è costruita con valori che tendono a Q1 e Q3 per grandi campioni (la riga al centro è la mediana)
- Le barre corrispondono a  $IQR * 1.5$
- Oltre, gli *outliers*: un pallino per i valori fra 2 e 3, un asterisco per quelli  $\geq 3$
- i numeri indicano i casi (la riga o una variabile identificativa)



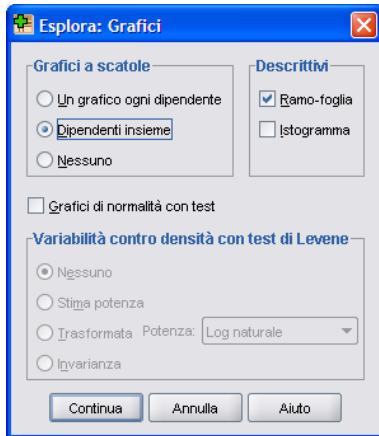
# Outliers in SPSS

- Se in **Analizza** | **Esplora** apriamo **Statistiche**, possiamo chiedere di **elencare gli Anomali**
- Mentre nel grafico visualizza tutti gli *outliers*, chiedendo di elencare i valori anomali, Spss visualizza i 5 valori più bassi e i 5 più alti (anche se non sono *outlier*)
- Per eliminare gli *outlier*, bisogna trasformare in punti z e quindi filtrare o ricodificare i valori superiori a 2 (meglio superiori a 3).
- ma bisogna anche fare attenzione alle caratteristiche del campione, perché la *normalità* dovrebbe essere relativa alle variabili indipendenti che si vogliono usare

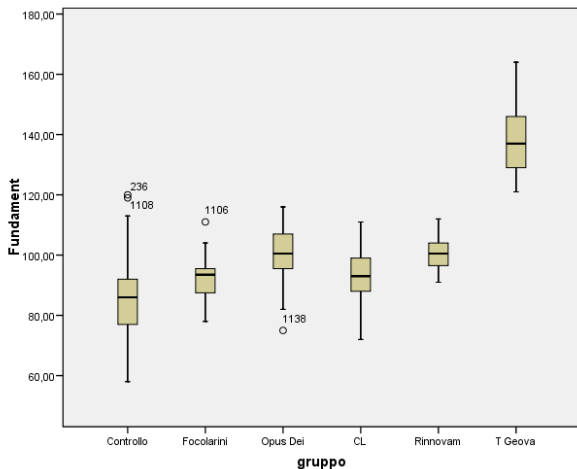


# Box-Plot suddiviso per una indipendente

- Sempre in *Analizza | Esplora*, possiamo inserire una variabile indipendente in *Fattori*
- e scegliendo il pulsante **Grafici** possiamo chiedere che i box-plot siano disegnati insieme
- questo permette di vedere come una variabile si distribuisce entro dei sottogruppi



# Box-Plot suddiviso per una indipendente



- Suddividendo per una variabile indipendente, gli outliers di prima diminuiscono drasticamente e acquistano un senso
- La maggior parte degli outliers di prima, coincide con uno specifico gruppo della indipendente

# Altre trasformazioni

- Con variabili di tipo Likert (ovvero i cui valori sono costituiti da numeri discreti compresi fra 0 e n) che poi vengono sommati per formare il punteggio di una "scala" (ad es. Autoritarismo)
- Spesso le scale sono costituite da item costruiti in modo tale che la risposta massimamente congruente con il costrutto misurato non coincida sempre con il valore massimo dell'item (ad es. alcuni item sono espressi in forma negativa, item contro-tratto)
- Prima di sommarli per formare la scala, questi item vanno "ribaltati"

# Altre trasformazioni: ribaltare

vecchio	nuovo	somma
1	5	6
2	4	6
3	3	6
4	2	6
5	1	6

vecchio	nuovo	somma
0	4	4
1	3	4
2	2	4
3	1	4
4	0	4

- Quindi basta sottrarre il valore osservato dell'item alla somma di massimo e minimo (i valori teorici).
- Ipotizzando un item a 5 gradini (da 1 a 5) che chiamiamo  $x_1$ , il ribaltamento si ottiene con

`COMPUTE x1r=6-x1.`

- Ipotizzando un item a 5 gradini (da 0 a 4) che chiamiamo  $x_2$ , il ribaltamento si ottiene con

`COMPUTE x2r=4-x2.`

- Considerate di non modificare la variabile originale, perché poi non sapete più se è pro- o contro-tratto

- C'è sempre la possibilità di ribaltare al volo

`COMPUTE nuova3 = MEAN(6-x1, 6-x2, x3, x4)*4.`