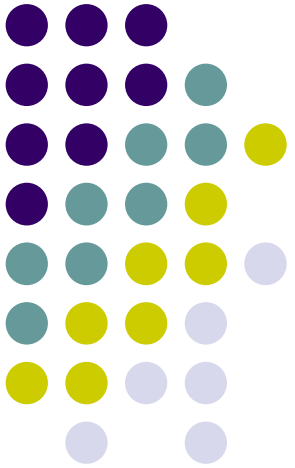
Elementi di Statistica Descrittiva

La Variabilità

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

… per iniziare

Consideriamo la distribuzione dei punteggi conseguiti in tre diversi test psico-attitudinali da un gruppo di persone

A	22	22	23	23	24	25	26	27	27	28	28
B	22	22	22	22	22	25	28	28	28	28	28
C	25	25	25	25	25	25	25	25	25	25	25

Il punteggio medio e quello mediano coincidono nei tre test.

Le due misure di sintesi ci poterebbero a concludere erroneamente che i tre test hanno prodotto gli stessi risultati.

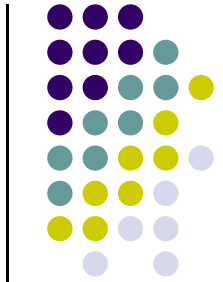
Tuttavia, essi sono profondamente diversi!!!!!

E allora, in che modo possiamo confrontare i tre test?

… per iniziare

Definizione: con il termine variabilità [mutabilità] di suole indicare l'attitudine di un carattere quantitativo [qualitativo] ad assumere modalità diverse

Lo studio della variabilità



Lo studio della variabilità/mutabilità è importante per almeno due motivi:

- **Valore intrinseco**

  La conoscenza della variabilità è alla base della Statistica, nel senso che se tutte le manifestazioni di un fenomeno fossero uguali tra loro, la rilevazione di una singola modalità sarebbe equivalente alla conoscenza della totalità delle informazioni e, quindi, non avrebbe più senso uno studio statistico. Analizzare e misurare l'attitudine a variare di un fenomeno rappresenta una delle finalità che si vuole perseguire con l'analisi statistica.

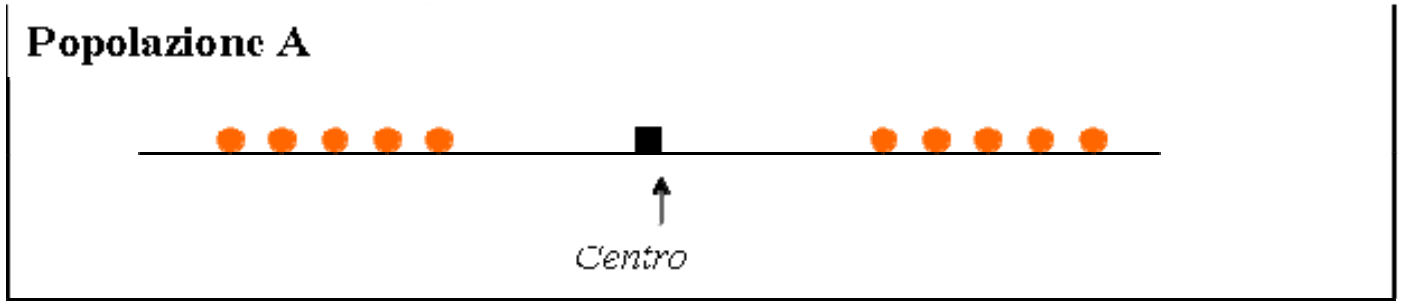
- **Accuratezza della sintesi dei dati**

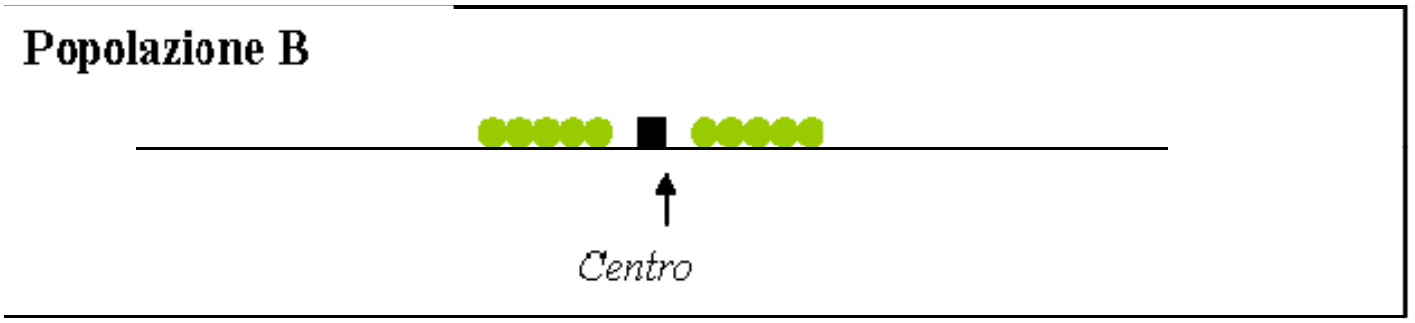
  L'impiego delle medie (sia di posizione che algebriche) non è sufficiente a sintetizzare le informazioni rilevate sulla popolazione oggetto di studio, specialmente quando occorre confrontare tra di loro popolazioni diverse.

  Dopo aver individuato il centro della distribuzione appare del tutto naturale valutare la dispersione dei dati osservati intorno ad esso. Questo compito è affidato alle misure di variabilità o di dispersione.

Lo studio della variabilità

Le misure di dispersione consentono di valutare il grado di dispersione delle modalità e la bontà della sintesi della distribuzione operata tramite gli indici di centralità





E' immediato rendersi conto che la sintesi effettuata tramite l'indice di centralità è più significativa nella popolazione B, perché le osservazione sono maggiormente addensate intorno al centro

Diverse misure di variabilità

A seconda degli aspetti che della variabilità che si vuole mettere in evidenza, è possibile raggruppare gli indici di variabilità in tre categorie:

- indici che si basano sulla differenza fra i valori che occupano determinate posizioni in un dato ordinamento
- indici che si basano sugli scostamenti delle osservazioni da una media
- indici che si basano sulle differenze tra tutte le modalità osservate

Un'altra classificazione che viene spesso adottata è quella fra indici assoluti e indici relativi/adimesionali

Requisiti di un indice di variabilità

Un indice di variabilità deve soddisfare almeno due requisiti:

1. deve assumere il valore minimo se e solo se tutte le unità della distribuzione presentano uguale modalità del carattere
2. deve aumentare all'aumentare della "diversità" tra le modalità assunte dalle varie unità

Il Campo di Variazione

Una prima idea della variabilità di un carattere quantitativo è fornita dal campo di variazione (range):

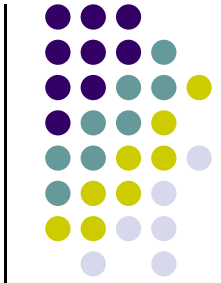
$$\Delta = x_{max} - x_{min}$$

E' un indice di variabilità semplice da calcolare e di immediata interpretazione in quanto rappresenta l'ampiezza dell'intervallo in cui si è manifestato il fenomeno

Difetti

- dipende solo da due osservazioni e non tiene conto delle altre
- essendo espressione dell'osservazione più grande e di quella più piccola è poco stabile in quanto estremamente sensibile agli outliers
- presenta difficoltà di calcolo in presenza di classi aperte

Distanza Interquartilica



La ***distanza interquartilica*** (*DI*) è la differenza tra il terzo e il primo quartile:

$$DI = Q_3 - Q_1$$

e rappresenta l'ampiezza dell'intervallo centrale (quello intorno alla mediana) nel quale si collocano il 50% dei valori.

Tanto più DI è piccola tanto più la metà delle osservazioni risulterà addensata intorno alla mediana.

In tal senso, la distanza interquartilica risulta un indice di variabilità interno, nel senso che si riferisce solo al 50% delle unità che presentano valori intorno alla mediana.

Peculiarità della DI

- è un indice più stabile del campo di variazione perché non si basa sulle osservazioni estreme
- potrebbe essere nulla senza che il carattere risulti degenere

Consideriamo, ad esempio, la seguente distribuzione:

X	f	F
0	0.1	0,1
1	0.70	0,80
2	0.20	1
Tot	1	

$Q_3 = Q_1 = 1$ ma il carattere non è degenere!!

La varianza

Un indice basato sugli scostamenti dalla media aritmetica che dipende da tutte le modalità è la varianza.

Definizione: per una distribuzione di frequenze, la varianza è la media aritmetica (semplice o ponderata) degli scarti al quadrato delle modalità dalla loro media aritmetica:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{k}(x_i - M)^2 n_i$$

Peculiarità

- *è espressa nel quadrato dell'unità di misura del carattere investigato* (→ scarto quadratico medio o deviazione standard)
- *assume solo valori non negativi*
- *è nulla se e solo se il carattere è degenere*
- *è un indice assoluto di variabilità*

Deviazione standard e coefficiente di variazione

La varianza è un indice assoluto di variabilità.

Ciò significa che, oltre ad essere espressa tramite l'unità di misura del carattere in esame, non è riferita né al massimo valore che può assumere, né a qualche altro valore standard.

Per questo motivo, la varianza non può essere utilizzate per effettuare confronti di variabilità tra:

- due o più collettivi sui quali si manifesta uno stesso fenomeno ma con un diverso ordine di grandezza (esempio: numero di non-parole negli adulti e nei bambini)

- due o più fenomeni diversi, ovvero espressi in diverse unità di misura (esempio: FO vs IQ)

Deviazione standard e coefficiente di variazione

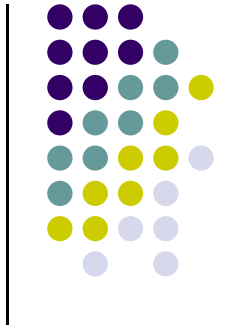
Per poter realizzare confronti di variabilità è possibile utilizzare il **coefficiente di variazione** definito come:

$$CV(X) = \frac{\sqrt{Var(X)}}{|M(X)|}$$

Tale rapporto dà come risultato un numero che non è espresso in nessuna unità di misura (adimensionale) e non risente dell'ordine di grandezza del fenomeno.

La radice quadrata della varianza è detta deviazione standard o scarto quadrato medio. Indica di quanto mediamente le modalità del carattere differiscono dalla loro media aritmetica

Esempio: lunghezza delle f.g.



X	n	$X-3.447$	$(X-3.447)^2$	$(X-3.447)^2*n$
1	65	-2.447	5.988	389.208
2	60	-1.447	2.094	125.629
3	50	-0.447	0.200	9.990
4	43	0.553	0.306	13.150
5	27	1.553	2.412	65.119
6	21	2.553	6.518	136.874
7	15	3.553	12.624	189.357
8	12	4.553	20.730	248.758
9	5	5.553	30.836	154.179
10	2	6.553	42.942	85.884
tot	**300**			**1418.147**

$$Var(X) = \frac{1418.147}{300} = 4.73 \text{ caratteri}^2$$

$$SQR(X) = \sqrt{4.727} = 2.174 \text{ caratteri}$$

$$CV(X) = \frac{2.174}{3.447} = 0.631$$

La standardizzazione di una variabile

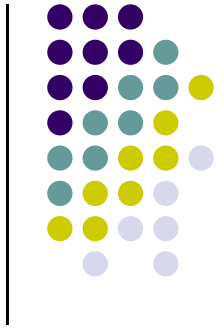
In ambito psicologico accade spesso di dover confrontare due o più prestazioni di uno stesso soggetto misurate con scale diverse (in termini di punteggio medio e scarto quadratico medio).

Così, ad esempio, supponiamo di aver sottoposto 20 soggetti ad un test di personalità: il *California Psychological Inventory (CPI)*. Il test consente di attribuire punteggi su 18 diverse dimensioni della personalità.

Supponiamo di voler stabilire in base al test se un determinato soggetto è più "*dominante*" o "*spontaneo*".

Supponiamo che i punteggi grezzi sulle due dimensioni siano entrambi pari a 25. Saremo pertanto tentati ad affermare che le due caratteristiche di personalità sono presenti in ugual misura.

La conclusione è errata in quanto non si è tenuto conto che le scale con cui vengono misurate le due caratteristiche sono diverse in termini di media e deviazione standard.



Da qui l'esigenza di avere una scala comune che consenta di effettuare i confronti. La standardizzazione dei punteggi risponde a questa esigenza

$$\longrightarrow Z = \frac{X - M(X)}{\sigma(X)}$$

Punteggio standardizzato "dominante" $z_d = \frac{25 - 24.5}{3.57} = 0.14$

Punteggio standardizzato "spontaneo" $z_s = \frac{25 - 28.9}{3.78} = -1.03$

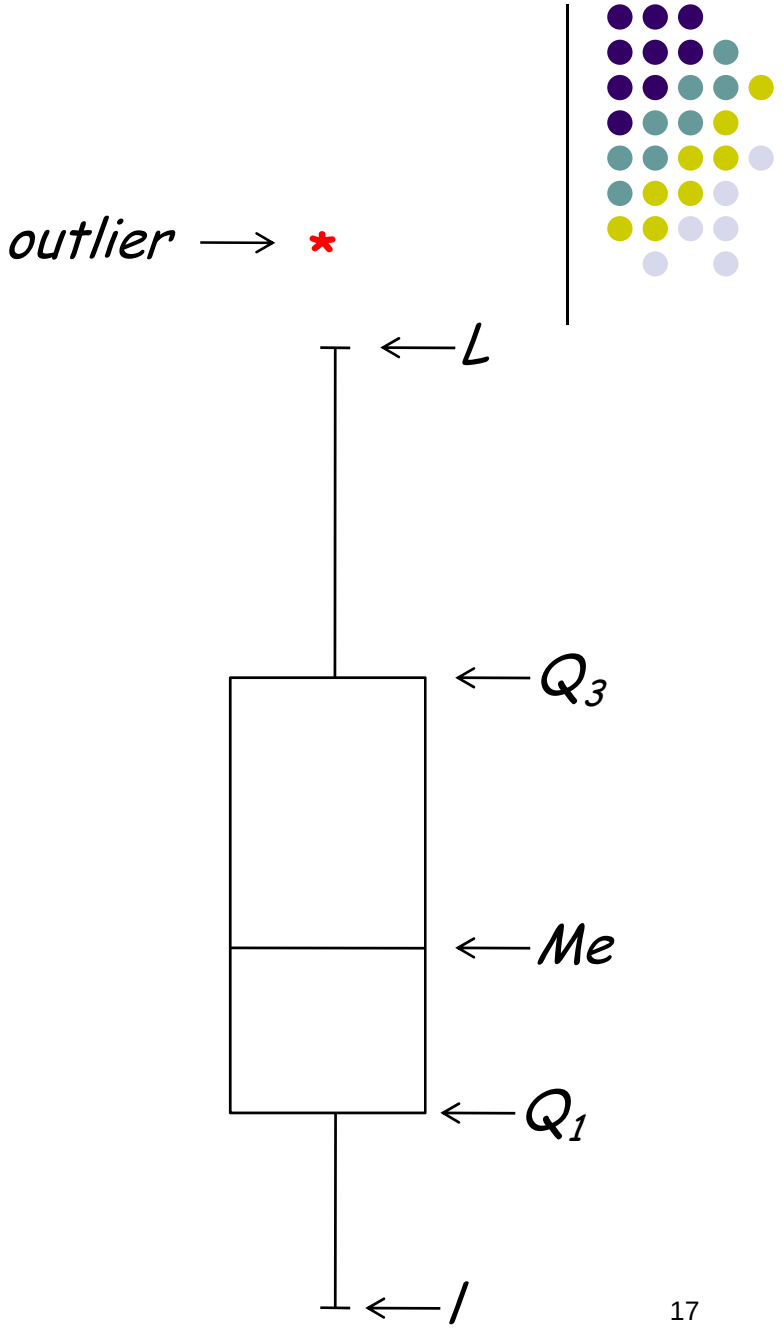
Il Box-Plot

Il box-plot (grafico a scatola) racchiude in una sola rappresentazione grafica alcuni aspetti sintetici di una distribuzione di frequenza.

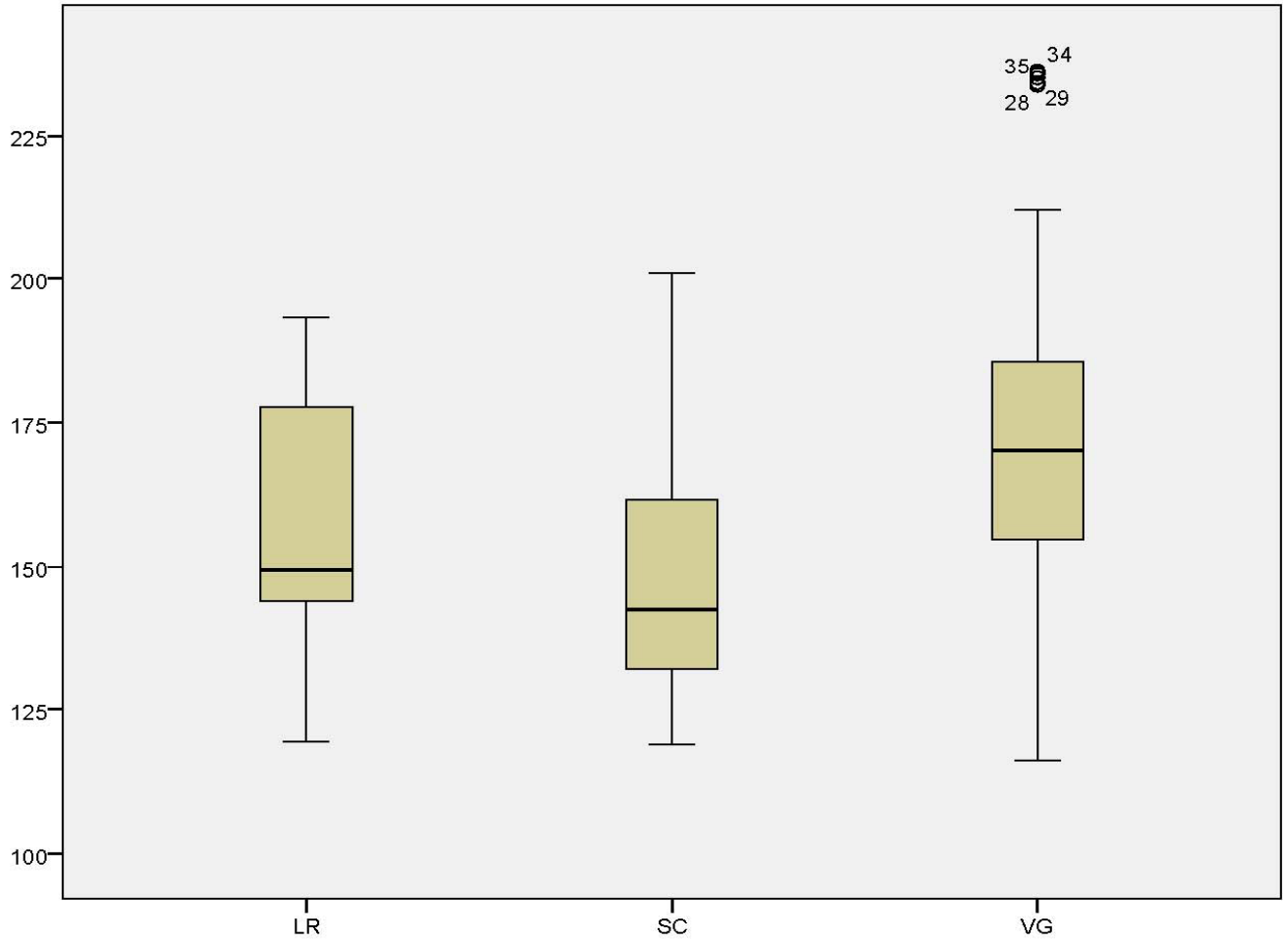
Nella sua forma originaria è un grafico basato sui quartili

$L=\min \{Q_3+1.5DI\ ,\ x_{max}\}$

$l=\max \{Q_1-1.5DI\ ,\ x_{min}\}$



Esempio: FO per tra soggetti



L'indice di eterogeneità di Gini

Siamo interessati a valutare il grado di diversità delle categorie grammaticali presenti nei sette discorsi di fine anno dell'ex Presidente della Repubblica Carlo Azeglio Ciampi

Categorie	n	f
Aggettivi	1762	0.14
Avverbi	571	0.05
Congiunzioni	628	0.05
Articoli	1210	0.1
Nomi	3187	0.25
Preposizioni	2354	0.19
Pronomi	767	0.06
Verbi	1912	0.15
Altro	178	0.01
Totale	**12569**	**1**

Gli indici di variabilità discussi in precedenza non possono essere calcolati, essendo il carattere di tipo qualitativo

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

L'indice di eterogeneità di Gini

Per poter valutare il grado di omogeneità/eterogeneità delle k modalità di un carattere qualitativo è possibile impiegare l'indice di Gini

$$G = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k} f_i^2\right)$$

- Se $G=0$, il carattere non varia. Tutte le unità presentano lo stessa modalità del carattere **(omogeneità)**
- Se $G=1$, le unità della popolazione si distribuiscono equamente tra le k distinte modalità del carattere **(massima eterogeneità)**

Esempio di calcolo

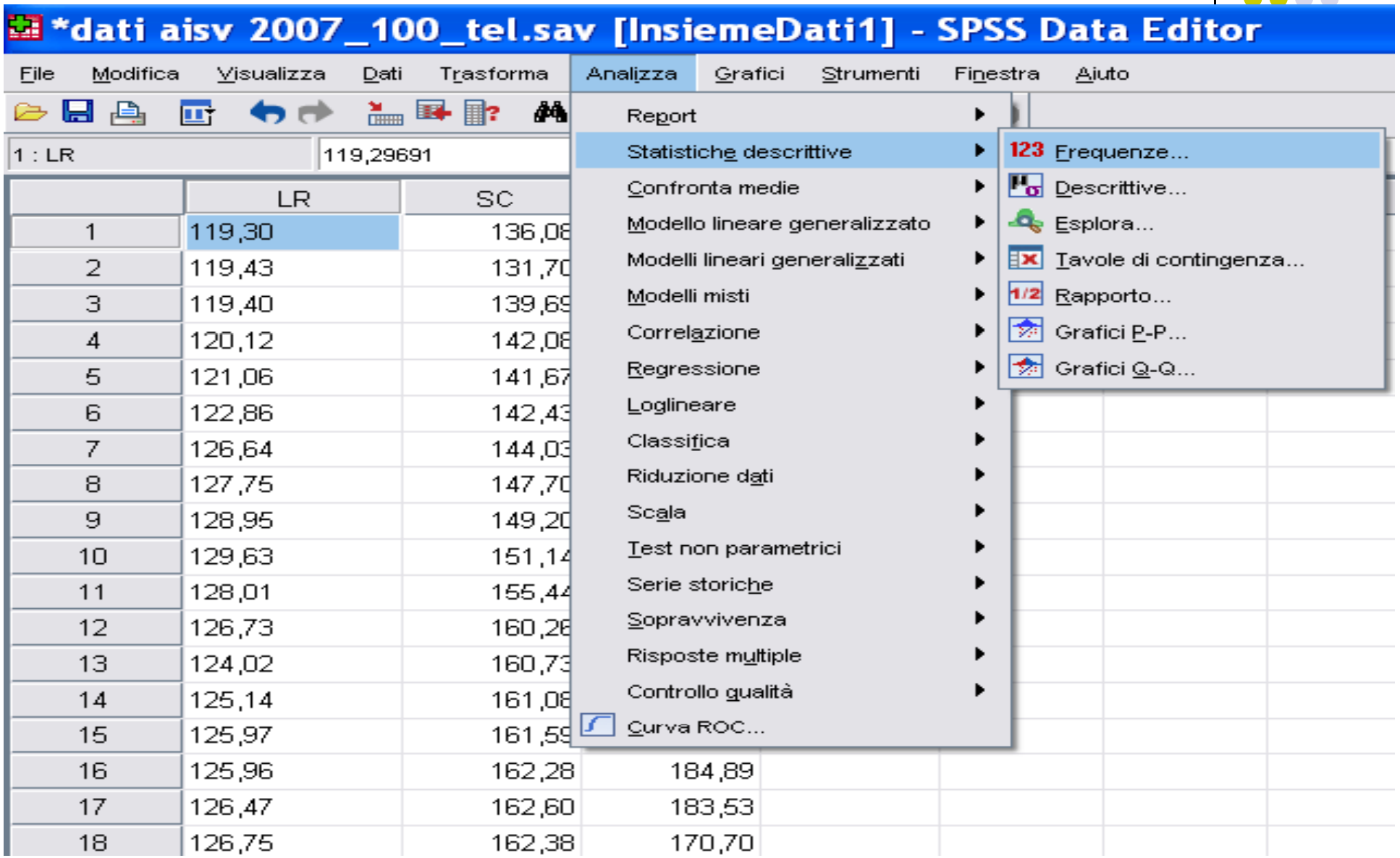
Categorie	f	f^2
Aggettivi	0.14	0.0196
Avverbi	0.05	0.0025
Congiunzioni	0.05	0.0025
Articoli	0.1	0.01
Nomi	0.25	0.0625
Preposizioni	0.19	0.0361
Pronomi	0.06	0.0036
Verbi	0.15	0.0225
Altro	0.01	0.0001
Totale	**1**	**0.1594**

$$G = \frac{9}{9-1}(1-0.1594)=0.946$$

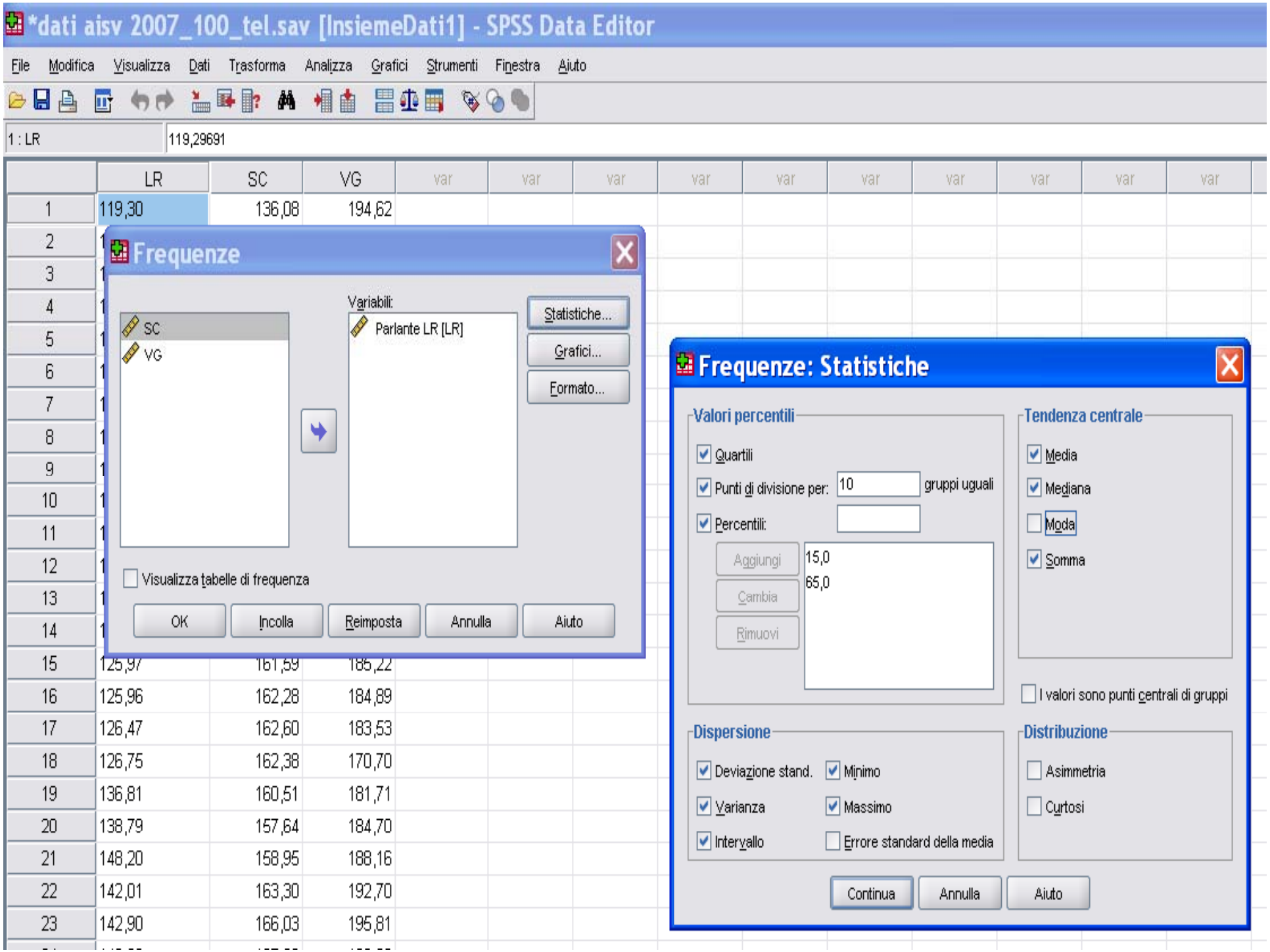
La mutabilità osservata è pari al 94.6% di quella massima

Misure di centralità e variabilità: elenco grezzo

Dati



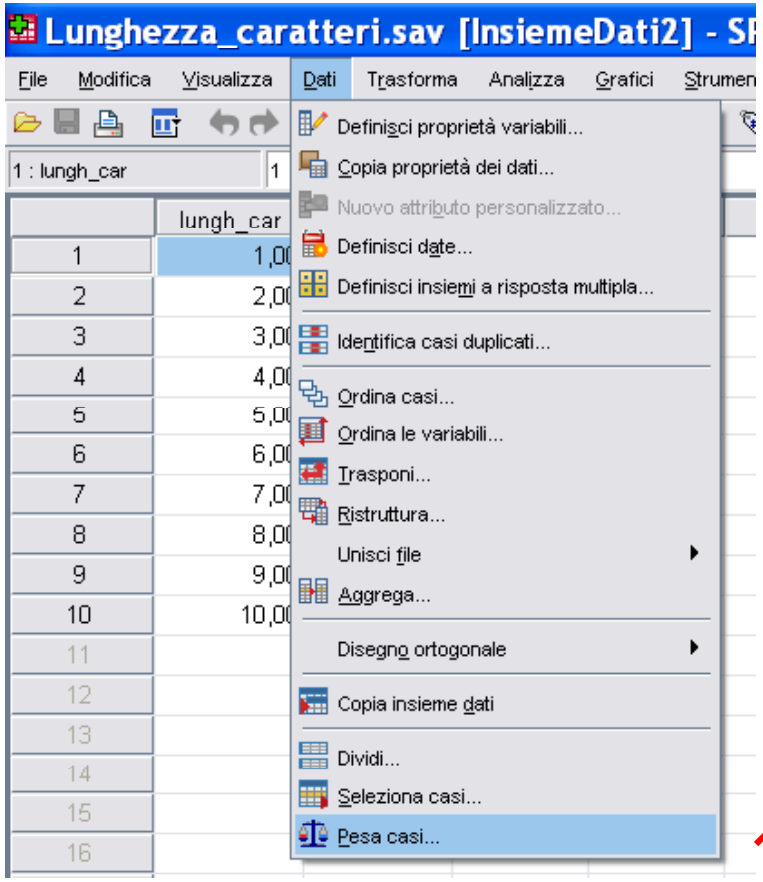
	LR	SC	
1	119,30	136,08	
2	119,43	131,70	
3	119,40	139,69	
4	120,12	142,08	
5	121,06	141,67	
6	122,86	142,43	
7	126,64	144,03	
8	127,75	147,70	
9	128,95	149,20	
10	129,63	151,14	
11	128,01	155,44	
12	126,73	160,26	
13	124,02	160,73	
14	125,14	161,08	
15	125,97	161,59	
16	125,96	162,28	184,89
17	126,47	162,60	183,53
18	126,75	162,38	170,70

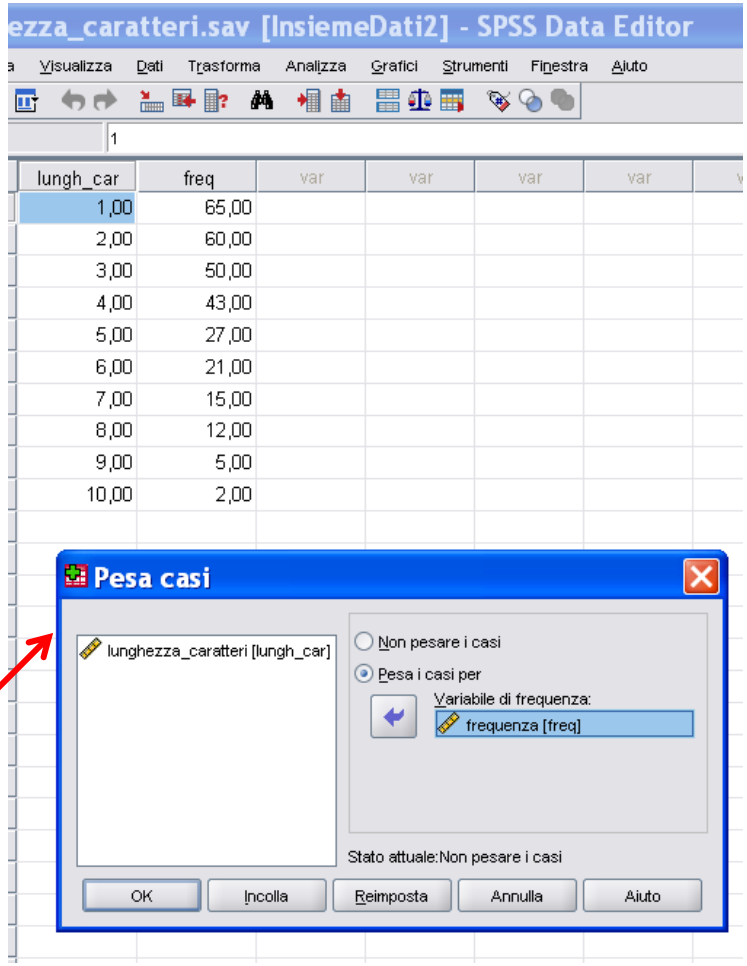

*dati aisv 2007_100_tel.sav [InsiemeDati1] - SPSS Data Editor
File Modifica Visualizza Dati Trasforma Analizza Grafici Strumenti Finestra Aiuto
1 : LR
119,29691
LR SC VG
1 119,30 136,08 194,62
2 125,97 161,59 185,22
3 125,96 162,28 184,89
4 126,47 162,60 183,53
5 126,75 162,38 170,70
6 136,81 160,51 181,71
7 138,79 157,64 184,70
8 148,20 158,95 188,16
9 142,01 163,30 192,70
10 142,90 166,03 195,81
Frequenze
SC
VG
Variabili:
Parlante LR [LR]
Statistiche...
Grafici...
Formato...
Visualizza tabelle di frequenza
OK
Incolla
Reimposta
Annulla
Aiuto
Frequenze: Statistiche
Valori percentili
Quartili
Punti di divisione per: 10 gruppi uguali
Percentili:
Aggiungi
Cambia
Rimuovi
15,0
65,0
Tendenza centrale
Media
Mediana
Moda
Somma
I valori sono punti centrali di gruppi
Dispersione
Deviazione stand.
Minimo
Varianza
Massimo
Intervallo
Errore standard della media
Distribuzione
Asimmetria
Curtosi
Continua
Annulla
Aiuto

Misure di centralità e variabilità: distribuzione di frequenza

Dati

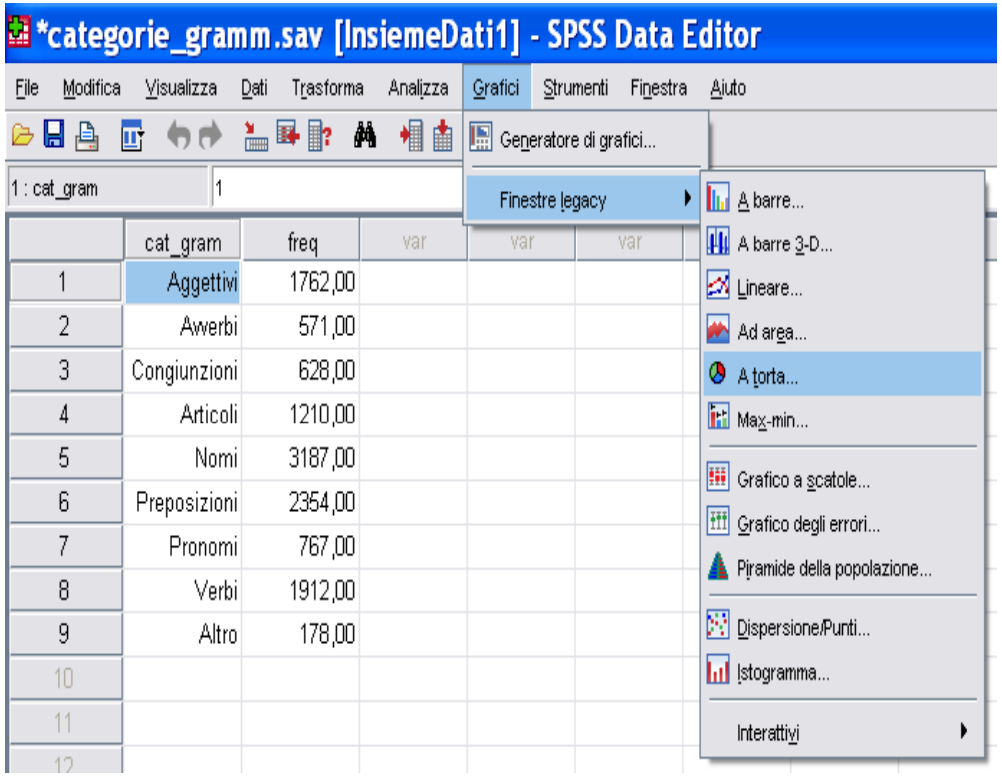
La procedura è uguale alla precedente con l'unica eccezione che occorre "comunicare" le frequenze, ovvero "pesare i casi"

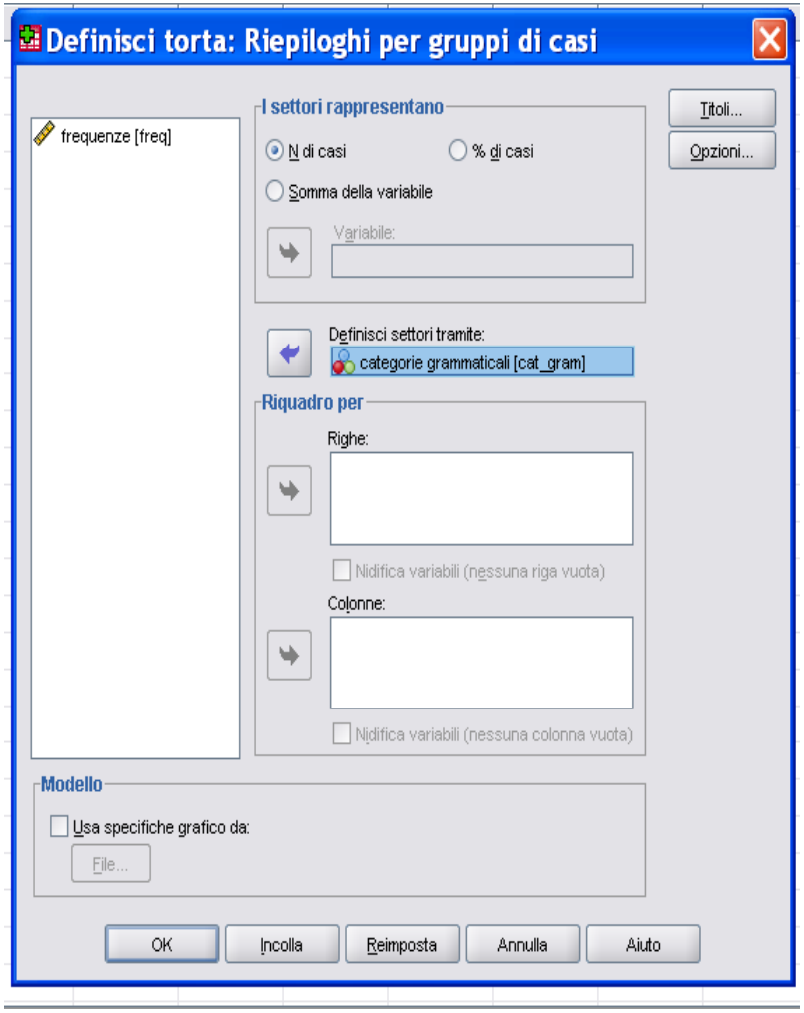




Grafici: carattere qualitativo

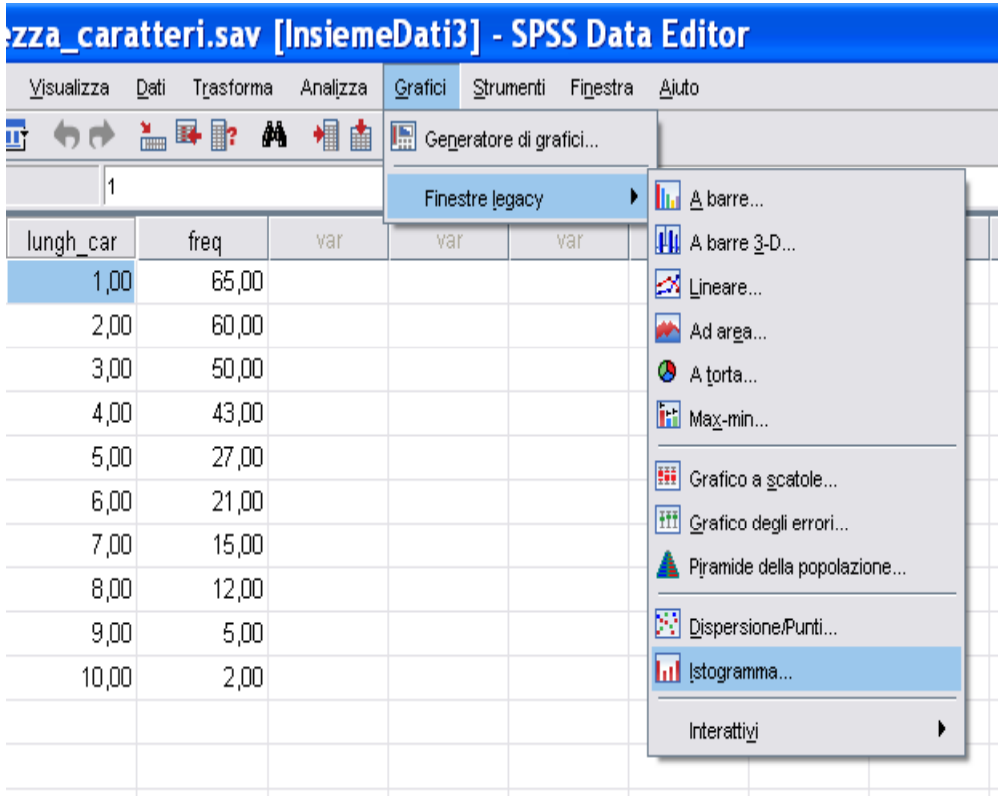
Dati

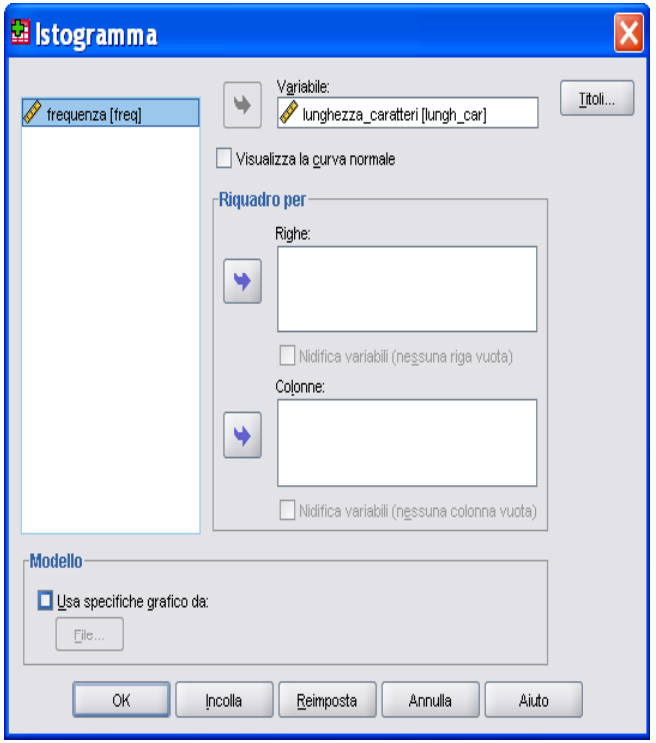




Grafici: carattere discreto

Dati





Il Box-Plot

Dati