

# Tecniche di ricampionamento

# Ricampionamento

- Supponiamo di voler stimare un parametro di interesse  $\theta$  sulla base del campione a disposizione.
- Diverse possibilità tra cui
  - metodo bootstrap
  - metodo jackknife

## PROBLEMA

Sia  $X_1, \dots, X_n$  un campione casuale  
di dimensione  $n$  della distribuzione  
 $F$  (micrograte) si vuole stimare  
un parametro  $\theta = t(F)$   
sulle basi del campione

Il bootstrap è in sintesi una tecnica  
di ricampionamento che fa uso presolen-  
tamente dell'elaboratore elettronico e  
fornisce una soluzione numerica a  
problemi la cui complessità impedisce  
l'utilizzo di tecniche statistiche standard.

# Ricampionamento: Bootstrap method

La tecnica è stata introdotta da B. Efron (1979) per ricavare misure di incertezza e distorsione sulla accuratezza delle stime di parametri incogniti.

La tecnica si applica quando la forma funzionale della VA non è nota e si è quindi costretti ad utilizzare la relativa forma funzionale non parametrica (vedi la FDR empirica  $\hat{F}$ ).

Supponiamo di avere un campione casuale semplice e di voler stimare un certo parametro di interesse:

$$X_1, X_2, \dots, X_n = \underline{X} \leftrightarrow \theta = t(F) \quad \text{var i.i.d.}$$

Allo scopo introduciamo uno stimatore del tipo:  $\hat{\theta} = s(\underline{X})$

La "compila"  
Le distub. de esegue prob  $\frac{1}{4}$  ed ogni OSS.  
E' le stime più opportune nei pedie  
sotto l'aspetto teorico. Costituisce la  
stima non parametrica alle massime  
verosimiglianze della f. di m.p. teorica  
nei pedie sotto l'aspetto interpretativo  
compone alle istituzioni dello pop.  
An il campione nelle I<sup>e</sup> fase boot. inf.

# Ricampionamento: Bootstrap method

Supponiamo di avere un campione casuale semplice e di voler stimare un certo parametro di interesse:

$$X_1, X_2, \dots, X_n = \underline{X} \leftrightarrow \theta = t(F) \quad \text{var i.i.d.}$$

Introduciamo uno stimatore del tipo:

- $\hat{\theta} = s(\underline{X})$

Sia dato un campione casuale bernoulliano  $\underline{X} = (X_1, X_2, \dots, X_n)$  e sia  $F$  (incognita)

la ~~fn~~ di ripartizione della popolazione da cui proviene il campione casuale.

Si vuole stimare un parametro di interesse  $\theta = t(F)$  sulla base di  $\underline{X}$

A tale scopo si ~~cerca~~ <sup>considera</sup> sulla base di

$\underline{X}$  uno stimatore

$$\hat{\theta} = s(\underline{X})$$

Qual'è l'eccezione di  $\hat{\theta}$ ?

Il bootstrap è stato introdotto per stimare lo standard error di  $\hat{\theta}$ .



3 metodi bootstrap dipendono dalla

definizione di CAMPLONE BOOTSTRAP.

Sia  $\bar{F}^1$  la distribuzione empirica  
di campione bootstrap e un campione  
casuale di empirie  $n$

$$\underline{X}^* = (X_1^*, \dots, X_n^*)$$

estratto dalla  $\bar{F}$   
ovvero  $n$  estrazioni ricampionate  
casualmente con ripetizione e  
origine

ES.

$$n = 7$$

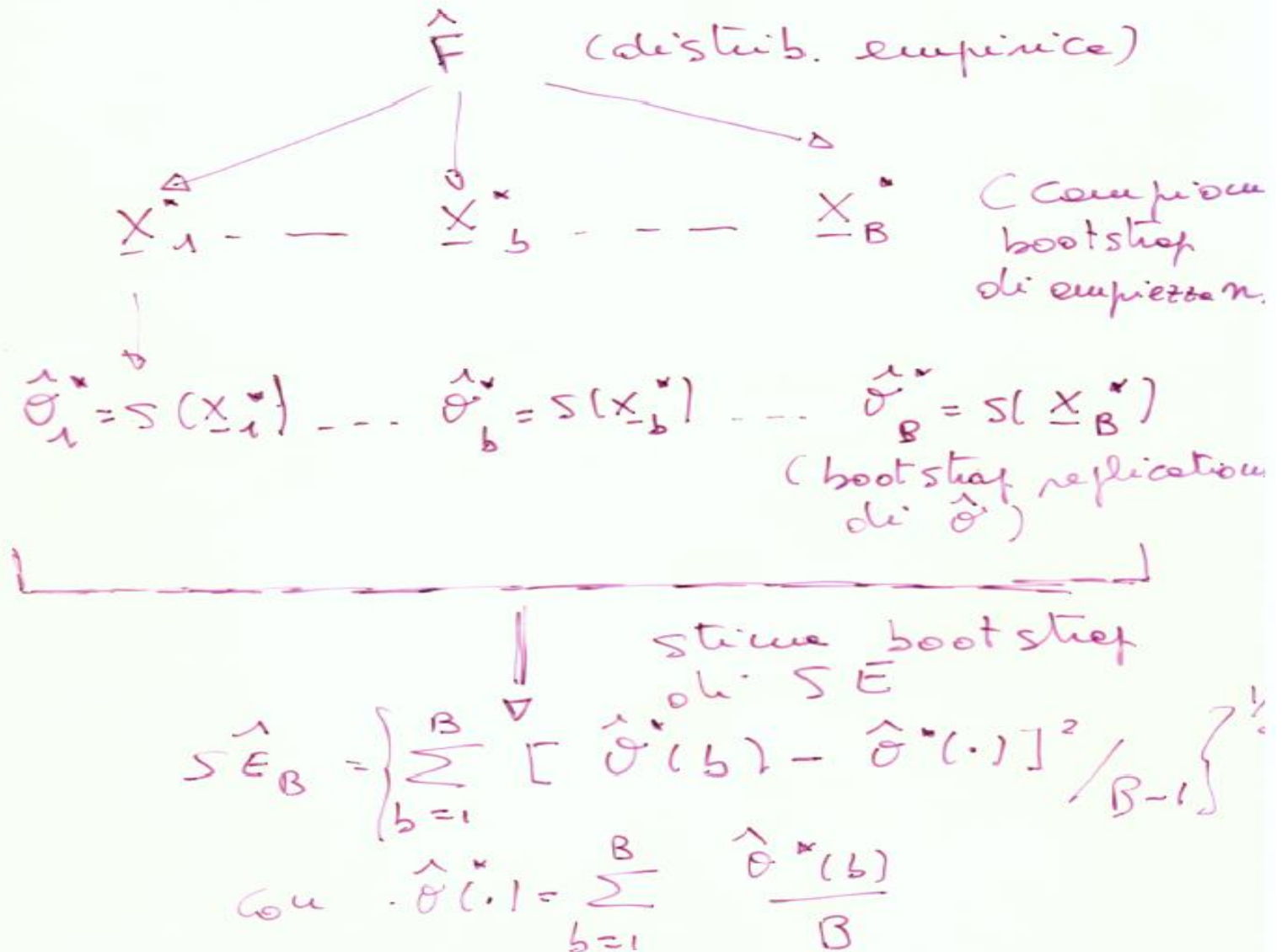
$$\underline{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)$$

$$\underline{X}^* = (X_5, X_7, X_5, X_4, X_7, X_3, X_1)$$

# Ricampionamento: Bootstrap standard error

- Il primo passo fondamentale è ottenere un campione bootstrap su cui lavorare. Un campione di questo tipo si ottiene da quello di partenza ricampionando casualmente con ripetizione, prendendo in considerazione la FDR empirica.
- Dal nuovo campione ottenuto  $\underline{X}^*$  passiamo a calcolare la **replica bootstrap** dello stimatore prescelto.  $\hat{\theta}^* = s(\underline{X}^*)$
- Valutiamo l'accuratezza della stima tramite misure come l'errore standard e la distorsione.

# ALGORITMO BOOTSTRAP PER LA STIMA DELLO SE DI UNA STATISTICA $\hat{\theta} = S(\underline{X})$



Il lim  
 $B \rightarrow \infty$

$$\Sigma_{\hat{F}}^{-1} B = \Sigma_{\hat{F}}^{-1} = \Sigma_{\hat{F}}^{-1} (\hat{\sigma}^2)^{-1} \textcircled{E}$$

APPROXIMAZIONE

è ideal boot strap

estimate of  $\Sigma_{\hat{F}}^{-1} (\hat{\sigma}^2)^{-1}$

È DETTO

NON PARAMETRIC

BOOTSTRAP ESTIMATE

oss: stime boot strap non parametrica  
perché basate sulle  $\hat{F}$  che è  
le stime non parametrica  
della  $F$ .

# Ricampionamento: Bootstrap standard error

- Ripetendo questo ragionamento per un numero elevato di volte diciamo  $B$  possiamo calcolare la stima bootstrap dell'errore standard nel modo seguente:

- $$SE_{\hat{F}}(\hat{\theta}^*) = \sqrt{\frac{\sum_i (\hat{\theta}_i^* - \hat{\theta}_{(\cdot)}^*)^2}{B-1}}$$

- essendo  $\hat{\theta}_{(\cdot)}^*$  la media delle repliche bootstrap.
- Se  $SE_F(\hat{\theta})$  è l'errore standard dello stimatore  $\hat{\theta}$ , allora il legame tra questa quantità e la corrispondente stima bootstrap è:

- $$SE_F(\hat{\theta}) = \lim_{B \rightarrow \infty} SE_{\hat{F}}(\hat{\theta}^*) = \lim_{B \rightarrow \infty} \sqrt{\frac{\sum_i (\hat{\theta}_i^* - \hat{\theta}_{(\cdot)}^*)^2}{B-1}}$$

- Il nome completo della stima bootstrap dello SE è stima bootstrap non parametrica ideale dello SE, poiché si basa sulla FDR empirica  $\hat{F}$  che è la stima non parametrica della FDR  $F$ .

# Ricampionamento: Bootstrap bias

- L'accuratezza dello stimatore  $\hat{\theta}$  può essere valutata anche in termini di distorsione.
- Ragionando al di fuori della tecnica bootstrap la distorsione dello stimatore  $\hat{\theta}$  è data dalla relazione:
- $$B_F(\hat{\theta}, \theta) = \mathbb{E}_F(S(\underline{X})) - t(F)$$
- La **stima bootstrap della distorsione** utilizzando la Fdr empirica è:
- $$B_{\hat{F}}(\hat{\theta}^*) = \mathbb{E}_{\hat{F}}(\hat{\theta}^*) - t(\hat{F}) = \hat{\theta}_{(\cdot)}^* - \hat{\sigma}$$
- essendo  $\hat{\sigma}$  la stima plugin di  $\sigma$ .
- Per molti stimatori la stima bootstrap della distorsione deve venir approssimata utilizzando metodi Monte Carlo.

# Ricampionamento: Bootstrap bias

- A livello pratico lo schema consiste nell'individuare un certo numero di campioni bootstrap a partire dai quali calcolare le corrispondenti repliche bootstrap.
- Poi si approssima la quantità  $\mathbb{E}_{\hat{F}}(\hat{\theta}^*)$  con la media delle repliche bootstrap già vista in precedenza  $\hat{\theta}_{(\cdot)}^*$ .
- Si individuano le stime plug-in  $\hat{\sigma}$  ed infine si passa a calcolare la stima bootstrap della distorsione:
- $B_{\hat{F}}(\hat{\theta}^*) = \hat{\theta}_{(\cdot)}^* - \hat{\sigma}$

# Ricampionamento: Jackknife

Contesto di nascita del metodo è l'analisi delle serie temporali (Quenouille 1959). Due anni dopo Tukey lo utilizza per sviluppare un metodo generale di costruzione di intervalli di confidenza approssimati.

- Il jackknife è una tecnica per valutare l'accuratezza di stime tramite misure quali:
- l'errore standard;
- la distorsione.



# Ricampionamento: Jackknife

- Anche per questa tecnica consideriamo una VA  $X$  da cui otteniamo una sequenza di VA  $X_1, \dots, X_n$  i.i.d. che costituisce il campione  $\underline{X}$ .
- Se siamo interessati ad un parametro  $\theta = t(F)$  che caratterizza la distribuzione della VA, allora andremo a considerare uno stimatore:
- $\hat{\theta} = s(\underline{X})$

Si consideri un campione  $\underline{X} = (X_1, \dots, X_n)$   
e una statistica  $\hat{\theta} = S(\underline{X})$

Per stimare il bias e lo standard  
error si introduce il concetto di

CAMPIONI JACKKNIFE: sono

campioni che partendo dal campione  
originario esistono fuori una osserva-  
zione alla volta.

Il  $j$ -esimo campione jackknife  
consiste di un insieme di dati:

in cui manca l' $j$ -esima osservazione  
(Variable)

$$\underline{X}_{(j)} = (X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_n)$$

# Ricampionamento: Jackknife

- Il primo passo fondamentale è ottenere un campione jackknife su cui lavorare. Un campione di questo tipo si ottiene da quello di partenza eliminando ad ogni passo l'elemento relativo a quel passo (p.es. nel primo campione non sarà presente il primo elemento) .
- Per valutare l'accuratezza della stima dobbiamo partire costruendo i campioni jackknife. Se il campione di partenza ha dimensione  $n$ , i campioni jackknife si ottengono lasciando fuori una osservazione alla volta. Alla fine otterremo  $n$  campioni jack-knife di dimensione  $n - 1$ .

# Ricampionamento: Jackknife

- Dai campioni jackknife si calcolano le relative repliche e si procede a calcolare le stime di errore standard e distorsione:

- $$\widehat{se}_{jk} = \sqrt{\frac{(n-1) \sum_{i=1}^n (\hat{\theta}_i^* - \hat{\theta}_{(\cdot)}^*)^2}{n}};$$
- $$\hat{B}_{jk} = (n-1)(\hat{\theta}_{(\cdot)}^* - \hat{\sigma}).$$

Sia  $\hat{\sigma}_{(j)}^1 = s\left(\frac{x}{(j)}\right)$  la  $j$ -esima  
copia  
j-esima copia di  $\hat{\sigma}^1$

Le stime j-esime delle  
DISTORSIONI  $\hat{\epsilon}^1$ :

$$\hat{\text{bias}}_{j\text{ech}}^1 = (n-1) (\hat{\sigma}_{(j)}^1 - \hat{\sigma}^1)$$

$$\text{con } \hat{\sigma}_{(j)}^1 = \sum_{i=1}^n \hat{\sigma}_{(i)}^1 / n$$

$$\hat{\sigma}^1 = t(\hat{F})$$

Le stime dello STANDARD ERROR  
 $\hat{\epsilon}^1$  definite da:

$$\hat{\text{SE}}_{j\text{ech}}^1 = \left[ \frac{n-1}{n} \sum_{\neq} (\hat{\sigma}_{(j)}^1 - \hat{\sigma}_{(i)}^1)^2 \right]^{1/2}$$

Un altro tipo di raggruppamento è  
partire dagli PSEUDOVALORI.

A causa dello stimatore  $\hat{\sigma}^2 = S(\underline{x})$

si considerano gli  $n$  stimatori.

$$\hat{\sigma}_{(j)}^2 = S(\underline{x}_{(j)})$$

e gli pseudovalori:

$$\tilde{\sigma}_{(j)}^2 = n \hat{\sigma}^2 - (n-1) \hat{\sigma}_{(j)}^2$$

$\hat{S}_E$  può essere:

$$\hat{S}_E \text{ può essere } = \left\{ \sum_1^n (\tilde{\sigma}_i - \tilde{\sigma})^2 / \{(n-1)n\} \right\}^{1/2}$$

in cui  $\tilde{\sigma} = \sum \tilde{\sigma}_i / n$

# ESEMPIO

Score data di Mardia, Kent , Bibby (1979) ripreso  
da Efron (1993)



# Score data di Mardia, Kent , Bibby (1979) ripreso da Efron (1993)

- Stud. mec vec alg ana sta
- 1 77.00 82.00 67.00 67.00 81.00
- 2 63.00 78.00 80.00 70.00 81.00
- 3 75.00 73.00 71.00 66.00 81.00
- 4 55.00 72.00 63.00 70.00 68.00
- 5 63.00 63.00 65.00 70.00 63.00
- 6 53.00 61.00 72.00 64.00 73.00
- 7 51.00 67.00 65.00 65.00 68.00
- 8 59.00 70.00 68.00 62.00 56.00
- 9 62.00 60.00 58.00 62.00 70.00
- 10 64.00 72.00 60.00 62.00 45.00
- .....

Osservazioni	88
Variabili	5

*La procedura PRINCOMP*

<b>Osservazioni</b>	88
<b>Variabili</b>	5

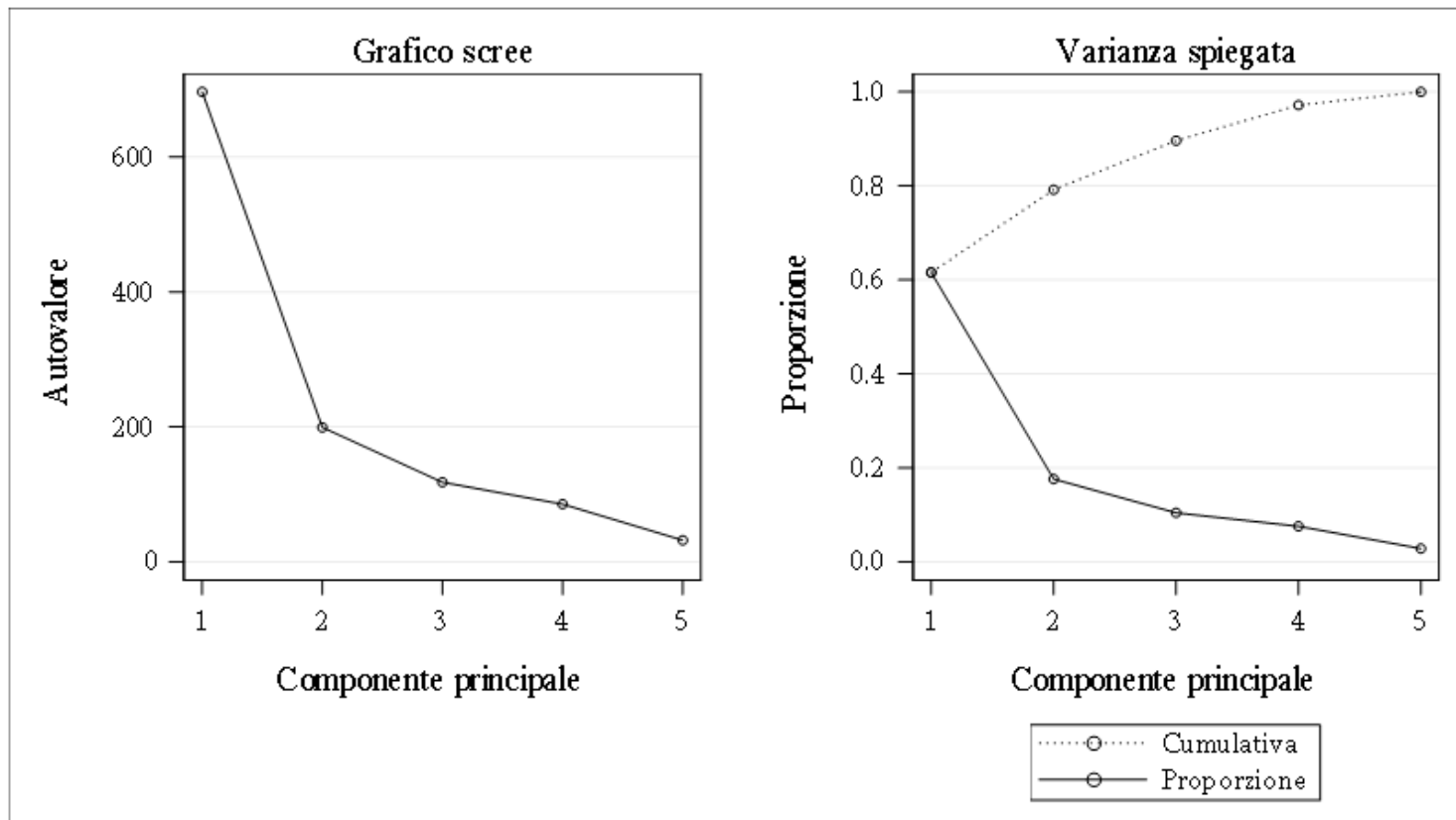
Statistiche semplici					
	<b>mec</b>	<b>vec</b>	<b>alg</b>	<b>ana</b>	<b>sta</b>
<b>Media</b>	38.95454545	50.37500000	50.60227273	46.27272727	42.30681818
<b>StD</b>	17.48622387	13.54117815	10.62478102	15.21624002	17.25558910

Matrice di covarianza					
	<b>mec</b>	<b>vec</b>	<b>alg</b>	<b>ana</b>	<b>sta</b>
<b>mec</b>	305.7680251	125.4195402	101.5794148	107.9090909	117.4049112
<b>vec</b>	125.4195402	183.3635057	86.2772989	108.3333333	101.5272989
<b>alg</b>	101.5794148	86.2772989	112.8859718	116.0867294	121.8705590
<b>ana</b>	107.9090909	108.3333333	116.0867294	231.5339603	152.3521421
<b>sta</b>	117.4049112	101.5272989	121.8705590	152.3521421	297.7553553

<b>Autovalori della matrice di covarianza</b>				
	<b>Autovalore</b>	<b>Differenza</b>	<b>Proporzione</b>	<b>Cumulativa</b>
<b>1</b>	696.536386	497.248479	0.6157	0.6157
<b>2</b>	199.287907	81.299577	0.1762	0.7918
<b>3</b>	117.988330	32.485796	0.1043	0.8961
<b>4</b>	85.502533	53.510871	0.0756	0.9717
<b>5</b>	31.991662		0.0283	1.0000

<b>Autovettori</b>					
	<b>Prin1</b>	<b>Prin2</b>	<b>Prin3</b>	<b>Prin4</b>	<b>Prin5</b>
<b>mec</b>	0.497116	0.766144	-.304870	-.250560	-.100860
<b>vec</b>	0.380829	0.171385	0.456055	0.775634	-.126502
<b>alg</b>	0.344368	-.081129	0.097196	-.056867	0.928519
<b>ana</b>	0.462095	-.289830	0.568041	-.544052	-.289487
<b>sta</b>	0.525095	-.541357	-.605764	0.190749	-.166954

*La procedura PRINCOMP*



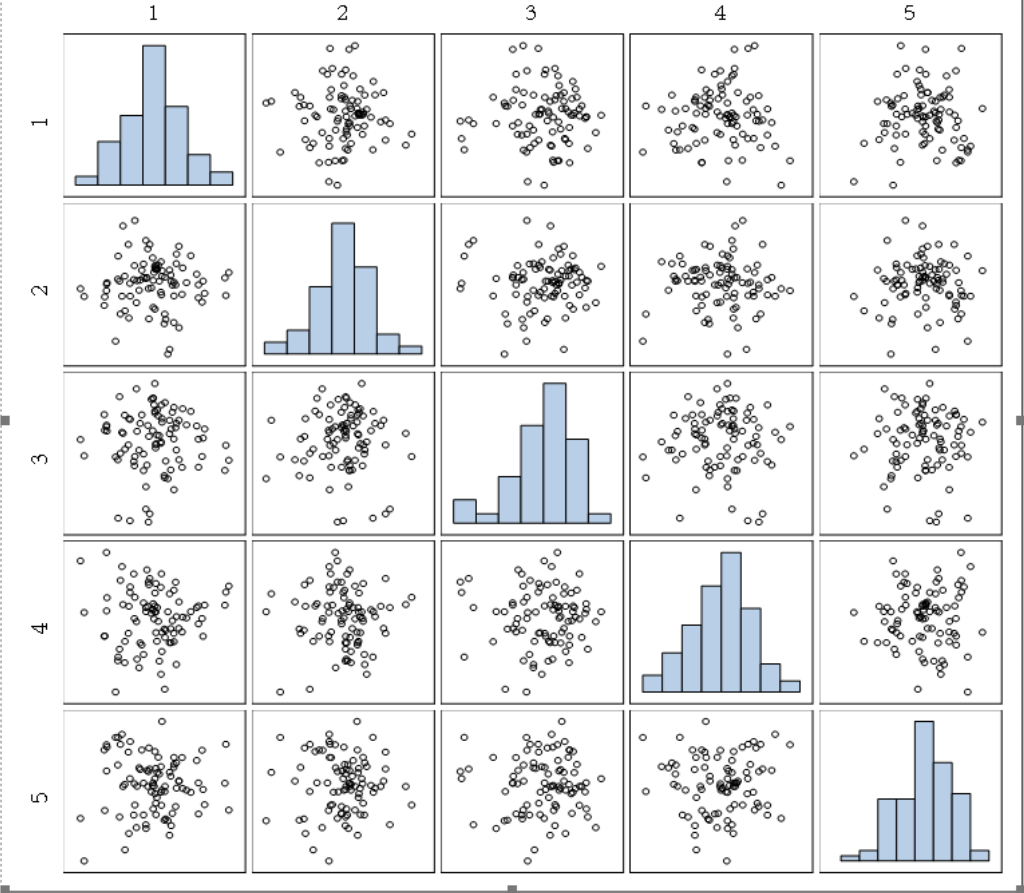
**Coefficienti di correlazione di Pearson, N = 88**  
**Prob > |r| sotto H0: Rho=0**

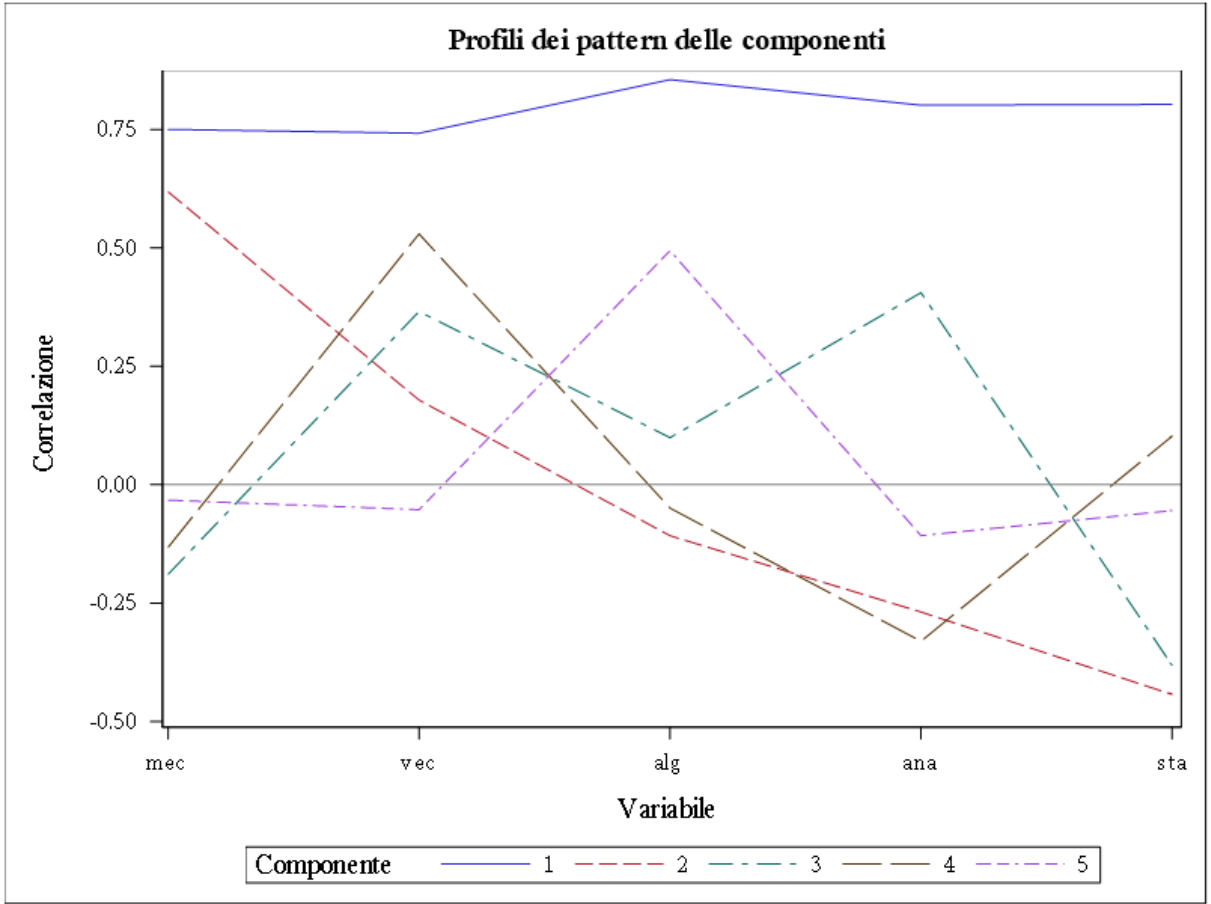
	mec	vec	alg	ana	sta	Prin1	Prin2	Prin3	Prin4	Prin5
mec	1.00000	0.52968 <.0001	0.54675 <.0001	0.40556 <.0001	0.38910 0.0002	0.75030 <.0001	0.61852 <.0001	-0.18938 0.0772	-0.13250 0.2185	-0.03262 0.7628
vec	0.52968 <.0001	1.00000	0.59968 <.0001	0.52577 <.0001	0.43451 <.0001	0.74224 <.0001	0.17867 0.0958	0.36583 0.0005	0.52965 <.0001	-0.05284 0.6249
alg	0.54675 <.0001	0.59968 <.0001	1.00000	0.71805 <.0001	0.66474 <.0001	0.85541 <.0001	-0.10779 0.3175	0.09937 0.3570	-0.04949 0.6470	0.49430 <.0001
ana	0.40556 <.0001	0.52577 <.0001	0.71805 <.0001	1.00000	0.58025 <.0001	0.80149 <.0001	-0.26889 0.0113	0.40550 <.0001	-0.33062 0.0017	-0.10761 0.3183
sta	0.38910 0.0002	0.43451 <.0001	0.66474 <.0001	0.58025 <.0001	1.00000	0.80312 <.0001	-0.44289 <.0001	-0.38132 0.0002	0.10222 0.3433	-0.05473 0.6126
Prin1	0.75030 <.0001	0.74224 <.0001	0.85541 <.0001	0.80149 <.0001	0.80312 <.0001	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin2	0.61852 <.0001	0.17867 0.0958	-0.10779 0.3175	-0.26889 0.0113	-0.44289 <.0001	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin3	-0.18938 0.0772	0.36583 0.0005	0.09937 0.3570	0.40550 <.0001	-0.38132 0.0002	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000
Prin4	-0.13250 0.2185	0.52965 <.0001	-0.04949 0.6470	-0.33062 0.0017	0.10222 0.3433	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000
Prin5	-0.03262 0.7628	-0.05284 0.6249	0.49430 <.0001	-0.10761 0.3183	-0.05473 0.6126	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000

Nome opzione	Descrizione
<b>Crea grafici scree e della varianza</b>	crea un grafico scree degli autovalori e un diagramma della varianza della proporzione. È possibile utilizzare un grafico scree per decidere quante componenti utilizzare nell'analisi. Ogni autovalore corrisponde a ciascuna delle componenti principali e rappresenta una quota della variabilità totale nel campione. Gli autovalori sono ordinati in modo decrescente, con i primi che formano la maggior parte della variazione. Il grafico scree solitamente mostra una ripida discesa per i primi autovalori, seguita da una stabilizzazione per i rimanenti. È possibile utilizzare il numero di autovalori in questa area di transizione per determinare il numero appropriato di componenti da includere.
<b>Diagramma degli score delle componenti principali a matrice a dispersione</b>	crea una matrice del grafico a dispersione degli score delle componenti principali. L'istogramma di ogni componente viene visualizzato nell'elemento diagonale della matrice.
<b>Crea diagramma dei profili a pattern</b>	crea un diagramma dei profili a pattern. Vi è un profilo per ogni componente. Il valore sull'asse Y è la correlazione fra la variabile e la componente principale.
<b>Crea grafico a dispersione degli score delle componenti principali</b>	crea un grafico a dispersione di ogni score delle componenti principali. Questi diagrammi possono essere personalizzati scegliendo di mostrare le ellissi di previsione per gli score delle componenti principali di una nuova osservazione. Per impostazione predefinita non sono visualizzate ellissi.
<b>Crea diagrammi delle componenti del pattern</b>	crea diagrammi dei pattern delle componenti accoppiati. Ogni osservazione sul diagramma è la correlazione fra la variabile e le due componenti corrispondenti sul diagramma. È possibile scegliere se rappresentare i pattern in un vettore. Se si sceglie di mostrare i vettori, allora per impostazione predefinita, viene rappresentato un cerchio unitario con un cerchio di varianza del 100% per il diagramma dei pattern del vettore.

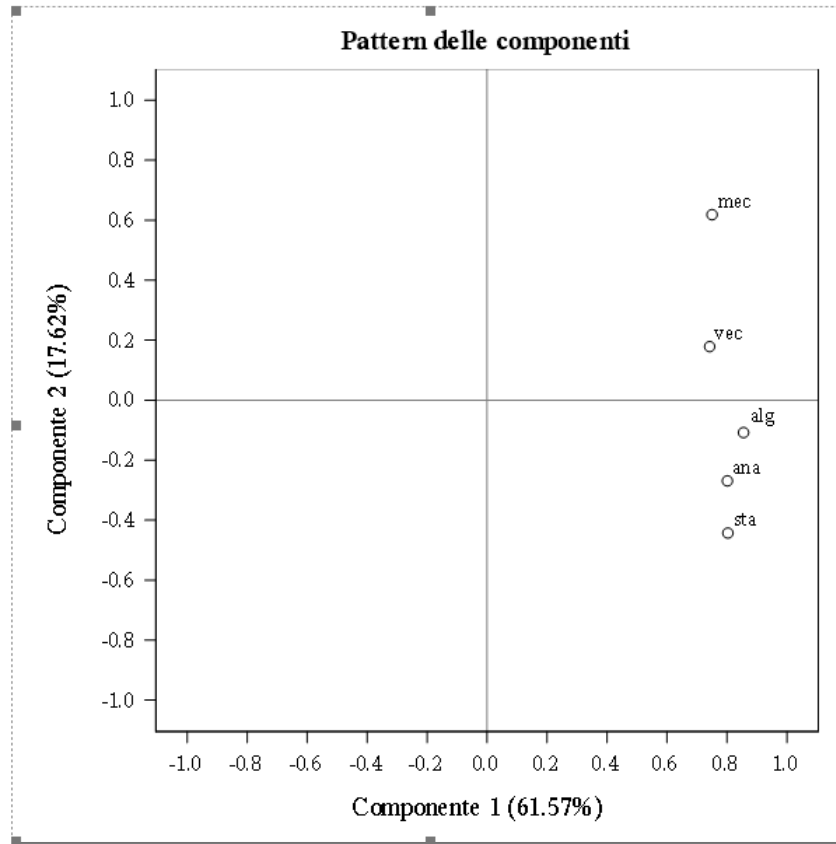
- (SAS Guide)

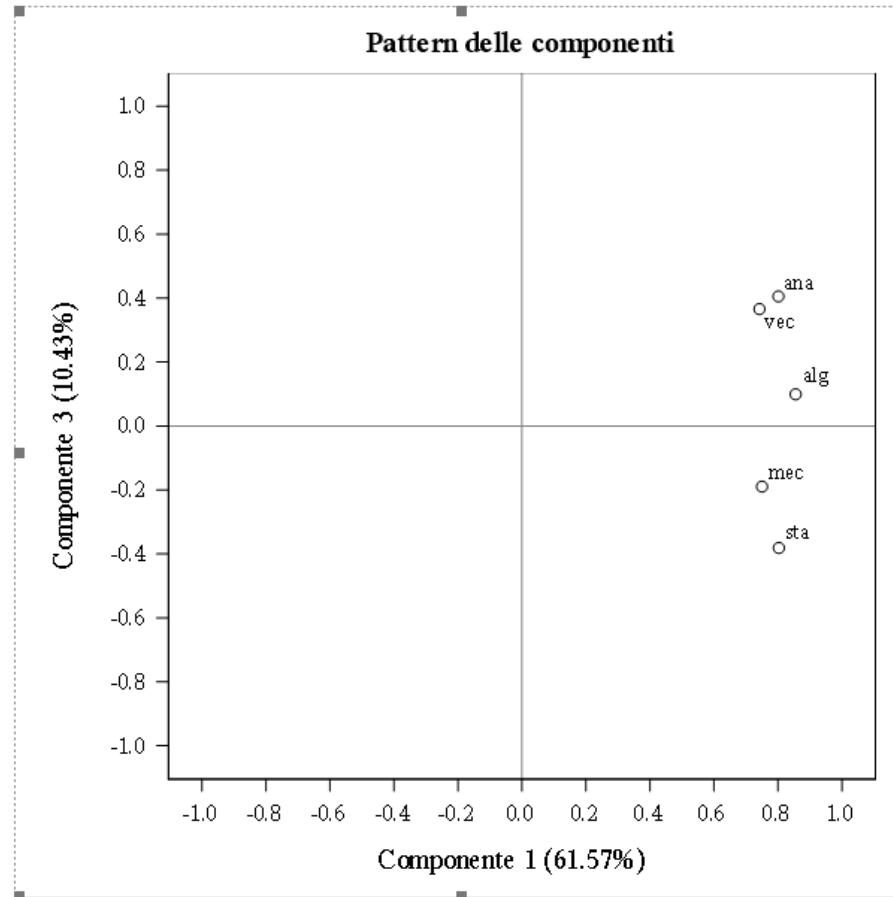
Matrice degli score delle componenti

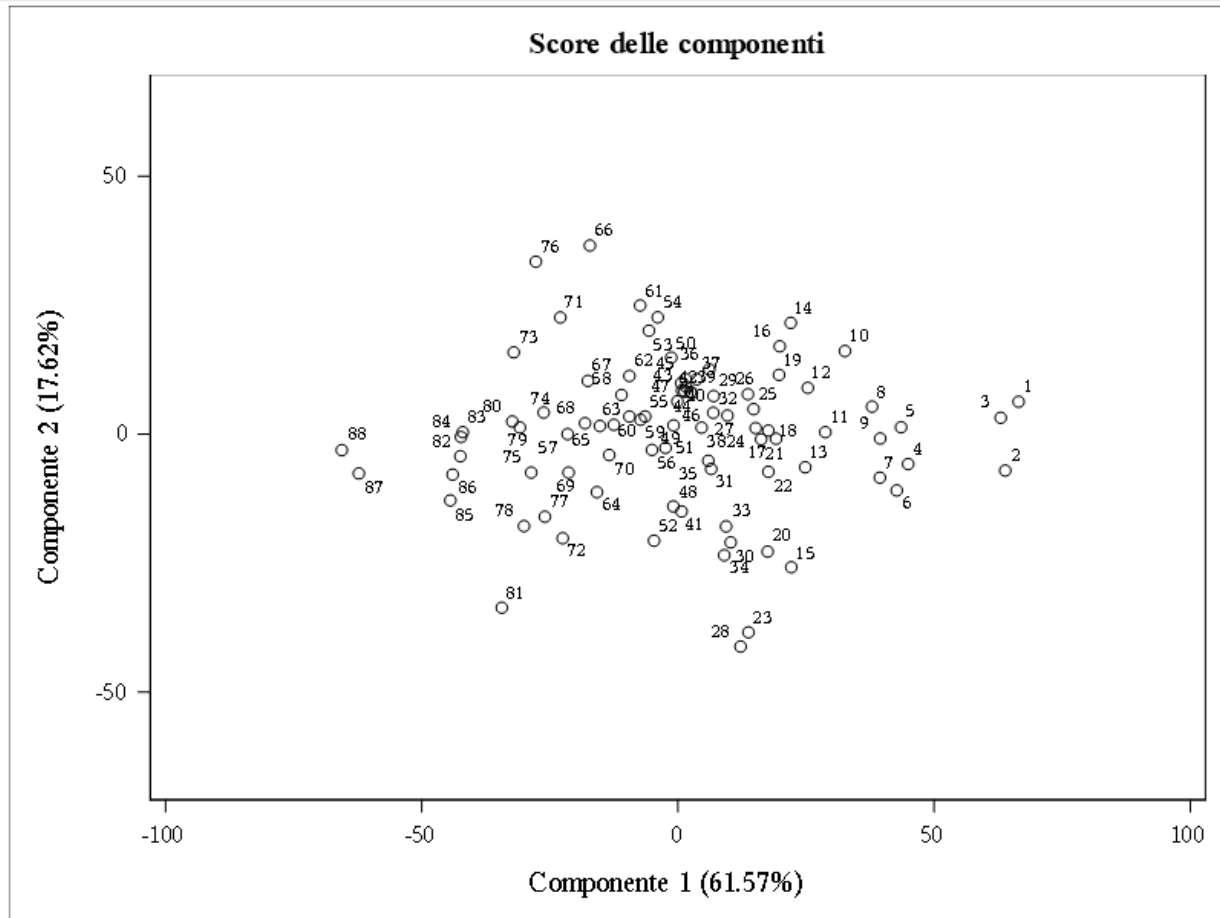












Eccetera..

Studente tests: mec (c) vec (c) alg (o) ana (o) sta (o)

(o) open book (c) closed book

Cinque numeri per rappresentare la capacità del generico studente

- Sintetizzando un unico numero può rappresentare la capacità del generico studente per es l'IQ (Intelligence Quotient) ?

Si nel caso il primo autovalore sia 1 e gli altri nulli .

Qui non succede.

Autovalori della matrice di covarianza				
	Autovalore	Differenza	Proporzione	Cumulativa
1	696.536386	497.248479	0.6157	0.6157
2	199.287907	81.299577	0.1762	0.7918
3	117.988330	32.485796	0.1043	0.8961
4	85.502533	53.510871	0.0756	0.9717
5	31.991662		0.0283	1.0000

Autovettori					
	Prin1	Prin2	Prin3	Prin4	Prin5
<b>mec</b>	0.497116	0.766144	-.304870	-.250560	-.100860
<b>vec</b>	0.380829	0.171385	0.456055	0.775634	-.126502
<b>alg</b>	0.344368	-.081129	0.097196	-.056867	0.928519
<b>ana</b>	0.462095	-.289830	0.568041	-.544052	-.289487
<b>sta</b>	0.525095	-.541357	-.605764	0.190749	-.166954

Coefficienti di correlazione di Pearson, N = 88  
 Prob > |r| sotto H0: Rho=0

	mec	vec	alg	ana	sta	Prin1	Prin2	Prin3	Prin4	Prin5
mec	1.00000 <.0001	0.52968 <.0001	0.54675 <.0001	0.40556 <.0001	0.38910 0.0002	0.75030 <.0001	0.61852 <.0001	-0.18938 0.0772	-0.13250 0.2185	-0.03262 0.7628
vec	0.52968 <.0001	1.00000	0.59968 <.0001	0.52577 <.0001	0.43451 <.0001	0.74224 <.0001	0.17867 0.0958	0.36583 0.0005	0.52965 <.0001	-0.05284 0.6249
alg	0.54675 <.0001	0.59968 <.0001	1.00000	0.71805 <.0001	0.66474 <.0001	0.85541 <.0001	-0.10779 0.3175	0.09937 0.3570	-0.04949 0.6470	0.49430 <.0001
ana	0.40556 <.0001	0.52577 <.0001	0.71805 <.0001	1.00000	0.58025 <.0001	0.80149 <.0001	-0.26889 0.0113	0.40550 <.0001	-0.33062 0.0017	-0.10761 0.3183
sta	0.38910 0.0002	0.43451 <.0001	0.66474 <.0001	0.58025 <.0001	1.00000	0.80312 <.0001	-0.44289 <.0001	-0.38132 0.0002	0.10222 0.3433	-0.05473 0.6126
Prin1	0.75030 <.0001	0.74224 <.0001	0.85541 <.0001	0.80149 <.0001	0.80312 <.0001	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin2	0.61852 <.0001	0.17867 0.0958	-0.10779 0.3175	-0.26889 0.0113	-0.44289 <.0001	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000

- La prima componente principale dà pesi positivi di circa pari ammontare a tutti i test per cui si può interpretare come prendere per ogni studente il punteggio totale o quello medio (Efron 1993)
- La seconda componente principale dà pesi positivi ai test a libri aperti e negativi ai test a libri chiusi per cui si può interpretare come una performance degli studenti a seconda che i test siano a libri aperti o a libri chiusi.

- $\lambda_1 = 696,536$
- $\hat{\theta}_1 = \lambda_1 / \sum \lambda_i = 0,6157$  (percentuale di varianza spiegata dalla prima componente)
- Qual è l'accuratezza di  $\hat{\theta}_1$ ? Si può utilizzare il metodo Bootstrap o il metodo Jackknife. (Vedi Efron 1993)

# PROBLEMA VALUTARE L'ACCURATEZZA DELLA STIMA BOOTSTRAP DELL'INDICE DI SKEWNESS DELLA VARIABILE VEC

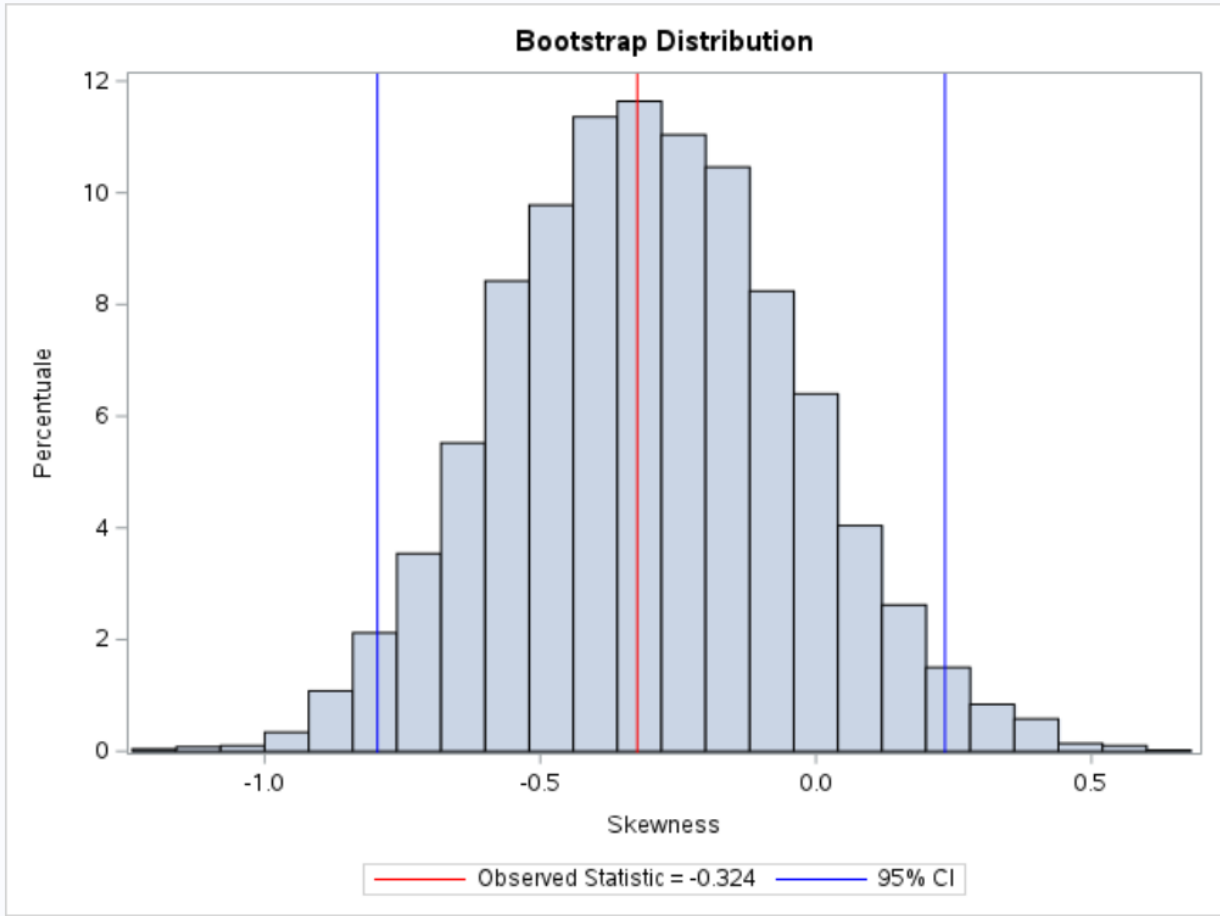
La procedura MEANS

Variabile di analisi: x
Skewness
-0.3244424

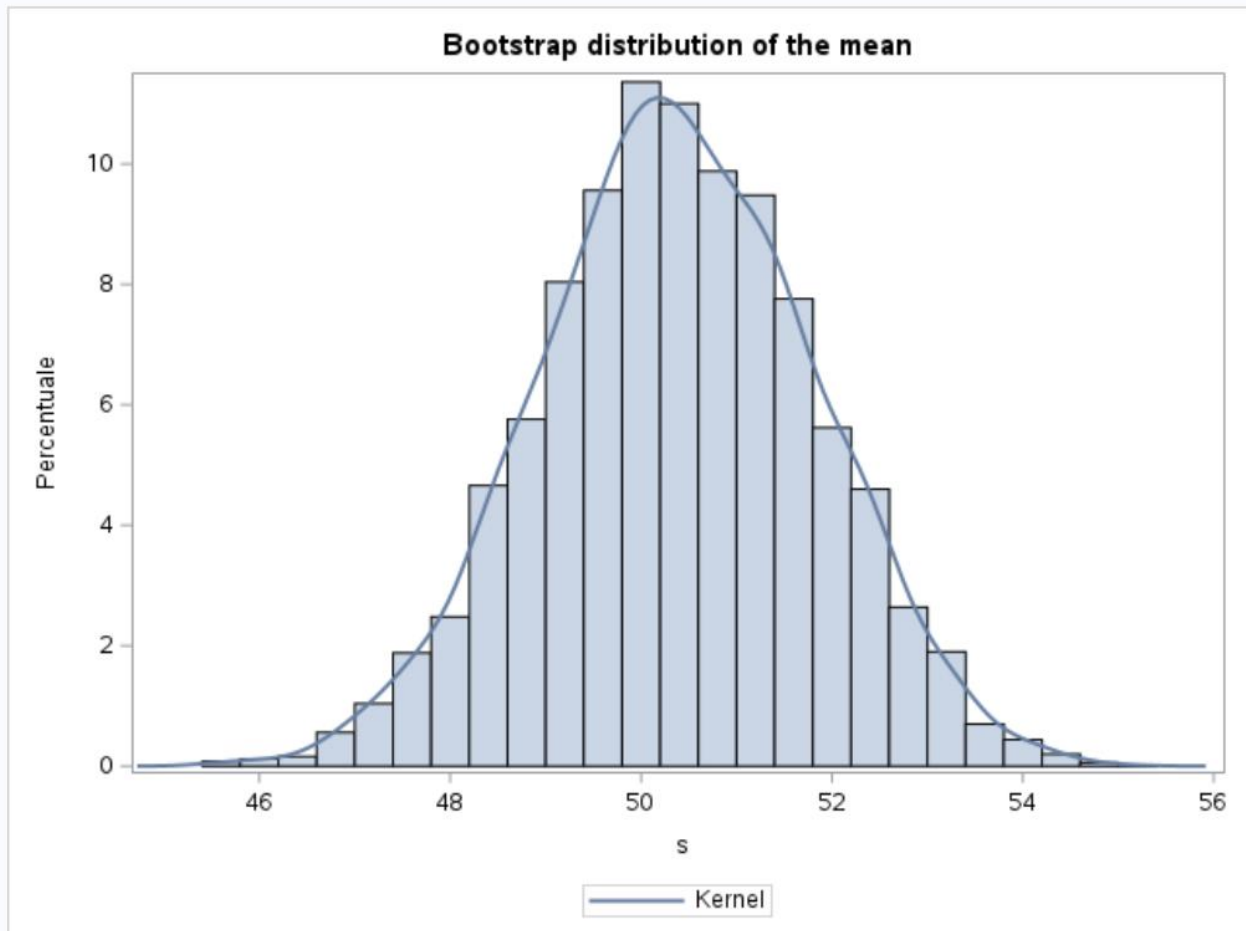
La procedura MEANS

Variabile di analisi: Skewness	
N	Dev std
5000	0.2640122

BootMean	BootStdErr	CI95_Lower	CI95_Upper
-0.30244	0.26401	-0.79536	0.23424







### Bootstrap distribution of the mean

Mean	MeanBoot	StdErrBoot	Lower 95% CL	Upper 95% CL
50.375	50.37195	1.4304284	47.528409	53.159091