Università Ca' Foscari di Venezia
Federica Giummolè

Statistica descrittiva

Probabilità e Statistica
A.A. 2014/2015

Variabili quantitative

In un reparto dove sono assemblati *lettori mp3* vengono provate in tre giorni diversi tre differenti organizzazioni delle linee di produzione. Le tre diverse organizzazioni sono chiamate nel seguito vecchia (quella in uso al momento dell'esperimento), nuova 1 e nuova 2.

Nei tre giorni, per ciascuno dei 288 addetti che lavorano nel reparto, viene rilevato

“il numero di operazioni completato”

che, ovviamente, può essere visto come una misura della produttività.

Domanda: qual è la migliore organizzazione del lavoro?

I dati

Vecchia organizzazione

725	724	710	724	700	724	713	692	683	712	684	707	703	691	709	702	705	715
704	705	697	725	692	719	694	717	696	707	726	703	705	712	710	697	698	694
701	715	701	707	706	701	687	708	719	713	699	702	694	708	712	704	703	687
709	693	715	707	710	700	718	702	718	705	723	718	701	698	692	684	716	710
708	707	695	726	710	709	692	707	717	709	710	718	708	720	705	714	687	707
707	723	695	676	705	684	717	719	715	710	711	696	696	715	686	702	708	713
701	692	713	700	704	726	702	706	706	700	700	687	696	694	699	709	704	704
715	706	688	724	713	686	697	710	704	724	721	717	690	707	713	685	706	699
687	702	701	708	704	705	702	701	699	699	685	712	678	706	706	695	707	718
706	716	703	721	714	704	697	693	711	697	710	713	702	715	714	716	698	714
704	717	700	692	718	699	698	690	710	703	702	719	710	725	721	713	699	703
698	712	714	707	691	711	712	718	702	711	709	700	719	692	716	700	707	714
717	714	703	709	711	704	689	712	714	711	692	720	697	698	700	689	693	707
699	704	696	708	713	714	712	708	704	720	705	703	712	719	713	716	712	703
717	695	711	697	693	701	699	697	724	713	706	705	704	707	704	719	711	700
694	706	705	698	697	697	700	705	722	712	703	688	694	708	703	690	706	704

Organizzazione ‘‘nuova 1’’

695	686	694	690	713	704	693	697	723	694	690	721	683	701	718	715	738	694
692	704	728	697	711	706	714	710	717	729	709	695	699	714	691	698	680	720
683	696	713	674	689	683	708	704	725	695	690	696	678	725	683	700	699	705
688	714	709	693	681	717	691	706	684	684	693	719	731	706	686	698	710	679
712	688	697	729	695	697	717	679	736	671	695	739	698	696	714	711	701	720
686	706	722	695	688	709	693	756	677	712	670	693	695	683	713	672	706	708
690	685	686	681	716	709	704	679	686	676	718	683	689	696	687	736	699	685
698	700	723	681	713	700	708	705	718	692	743	715	745	700	693	676	723	712
671	714	687	687	687	683	671	677	696	696	714	713	671	688	675	671	692	725
690	680	693	703	733	708	720	704	688	732	711	685	714	704	686	682	699	708
708	704	685	685	694	702	738	702	696	709	701	687	703	701	702	693	691	701
735	721	705	691	741	685	716	716	737	687	732	697	670	684	693	711	685	705
690	705	693	698	678	704	710	686	689	686	698	684	687	696	719	679	696	701
712	691	686	704	744	705	718	709	725	699	721	690	678	713	714	705	681	721
673	698	717	711	670	726	694	723	701	683	716	671	712	704	699	705	727	719
702	692	708	694	670	694	697	682	718	705	699	709	695	711	688	717	699	686

Organizzazione ‘‘nuova 2’’

698	715	675	710	731	721	705	718	693	702	713	730	707	710	744	725	724	701
737	715	704	723	705	702	698	729	698	723	716	698	732	724	721	722	728	740
727	709	724	746	704	740	729	708	721	714	739	713	752	732	713	692	734	727
725	690	749	706	758	722	697	722	705	723	748	730	706	688	709	739	709	744
704	716	748	713	744	721	723	733	707	723	702	734	690	715	711	705	718	702
706	742	742	736	740	712	722	731	713	704	704	735	700	717	746	735	717	718
691	696	720	735	716	745	714	698	709	704	704	684	749	747	715	717	731	700
747	709	705	749	704	697	694	715	737	734	705	726	710	716	740	731	714	733
726	752	714	710	714	753	749	728	696	733	731	728	686	706	710	729	729	730
722	707	716	702	728	716	743	750	715	735	710	734	712	706	719	709	702	712

710	729	728	720	721	752	715	712	717	692	724	720	739	719	712	713	734	734
710	711	722	743	707	729	712	681	739	699	721	706	703	708	719	708	724	730
726	731	734	739	727	759	718	716	715	719	693	729	738	710	730	726	719	726
733	717	701	723	720	744	730	698	729	696	717	713	705	700	715	710	735	726
732	701	707	724	708	730	721	720	706	700	735	706	725	725	735	695	709	705
702	737	688	727	717	708	720	724	731	706	730	714	703	721	712	748	734	724

Frequenze assolute

	vecchia	nuova 1	nuova 2
[670,675)	0	13	0
[675,680)	2	12	1
[680,685)	4	20	2
[685,690)	13	33	3
[690,695)	23	33	8
[695,700)	35	38	13
[700,705)	55	27	24
[705,710)	52	28	34
[710,715)	50	28	32
[715,720)	33	19	32
[720,725)	15	12	34
[725,730)	6	9	27
[730,735)	0	4	30
[735,740)	0	7	17
[740,745)	0	3	12
[745,750)	0	1	12
[750,755)	0	0	5
[755,760)	0	1	2
Totale	288	288	288

Frequenze relative

	vecchia	nuova 1	nuova 2
[670,675)	0,000	0,045	0,000
[675,680)	0,007	0,042	0,003
[680,685)	0,014	0,069	0,007
[685,690)	0,045	0,115	0,010
[690,695)	0,080	0,115	0,028
[695,700)	0,122	0,132	0,045
[700,705)	0,191	0,094	0,083
[705,710)	0,181	0,097	0,118
[710,715)	0,174	0,097	0,111
[715,720)	0,115	0,066	0,111
[720,725)	0,052	0,042	0,118
[725,730)	0,021	0,031	0,094
[730,735)	0,000	0,014	0,104
[735,740)	0,000	0,024	0,059
[740,745)	0,000	0,010	0,042
[745,750)	0,000	0,003	0,042
[750,755)	0,000	0,000	0,017
[755,760)	0,000	0,003	0,007
Totale	1,000	1,000	1,000

$$\text{frequenze relative} = \frac{\text{frequenze assolute}}{\text{numero totale di osservazioni}}$$

Tabelle di frequenza: notazioni

y_i : modalità/classe i del carattere y, $i = 1, 2, \ldots, k$ (k modalità/classi)

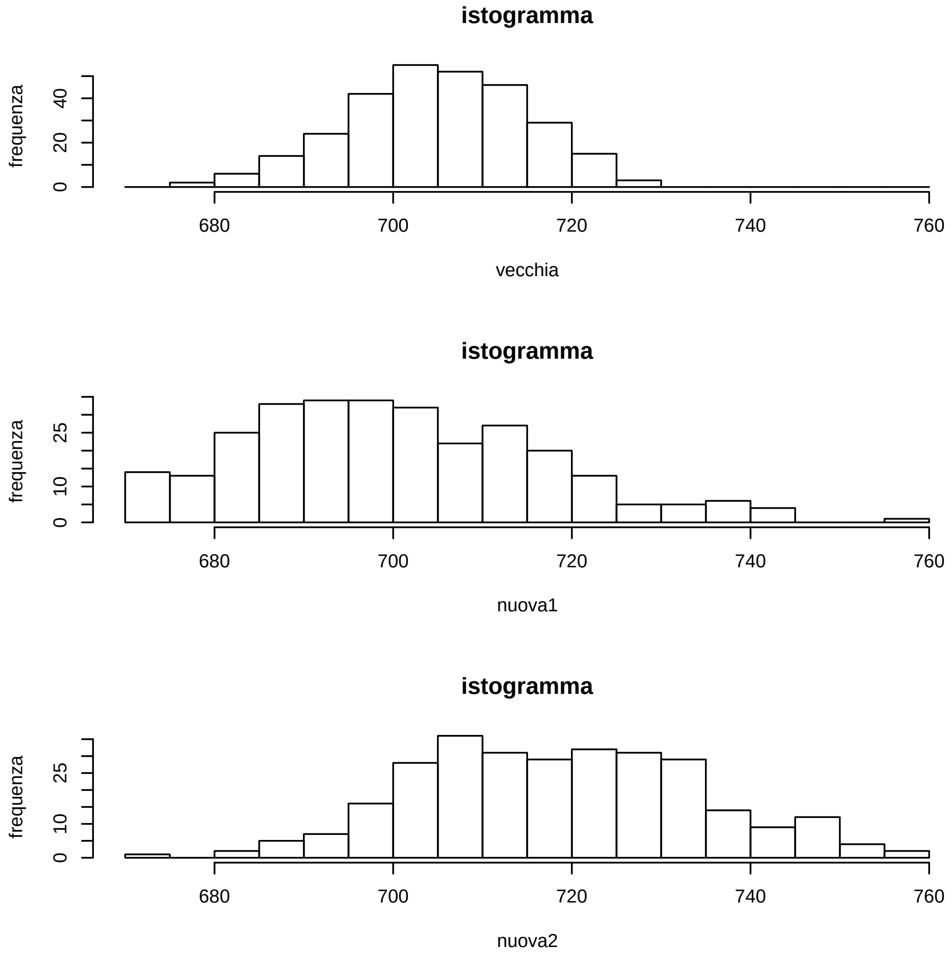
f_i : frequenza assoluta, numero di unità statistiche che possiedono la modalità/classe y_i

n : numero totale di osservazioni ($n = f_1 + f_2 + \cdots + f_k$)

p_i : frequenza relativa ($p_i = f_i/n$)

modalità/classe	freq. assolute	freq. relative
y_1	f_1	$p_1 = f_1/n$
y_2	f_2	$p_2 = f_2/n$
$\vdots$	$\vdots$	$\vdots$
y_k	f_k	$p_k = f_k/n$
Totale	n	1

Istogramma



Gli istogrammi in questo grafico sono stati costruiti ponendo:

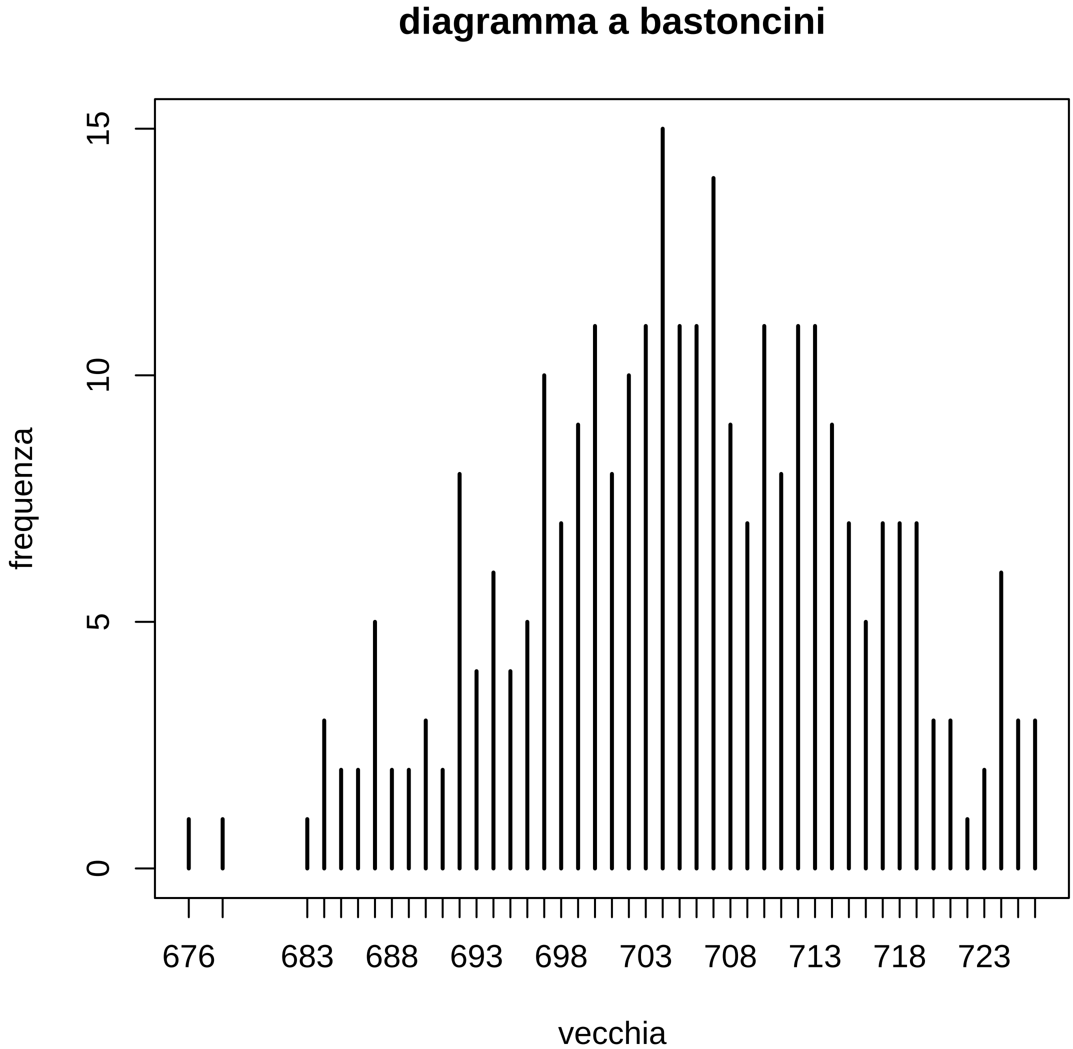
1. la base dei rettangoli pari agli intervalli riportati nella 1° colonna delle tabelle precedenti;

2. l'altezza dei rettangoli pari alle frequenze assolute.

Attenzione! questa regola è valida perché tutti gli intervalli hanno la stessa ampiezza...

Diagrammi a bastoncini

Il diagramma a bastoncini (da non confondere con l'i-stogramma!) è costruito disegnando in corrispondenza di ogni valore osservato un bastoncino di lunghezza uguale alla frequenza assoluta con cui quel valore è stato osservato.



Intervalli di differenti lunghezze

Può capitare o per scelta (si vuole fornire informazioni più dettagliate su parte della distribuzione) o per necessità (i dati sono già stati raggruppati in classi da qualcuno) di costruire degli istogrammi utilizzando intervalli di lunghezza differente.

In questo caso, le altezze dei rettangoli che compongono l'istogramma non devono essere proporzionali alle frequenze osservate ma alla **densità** delle osservazioni nelle singole classi:

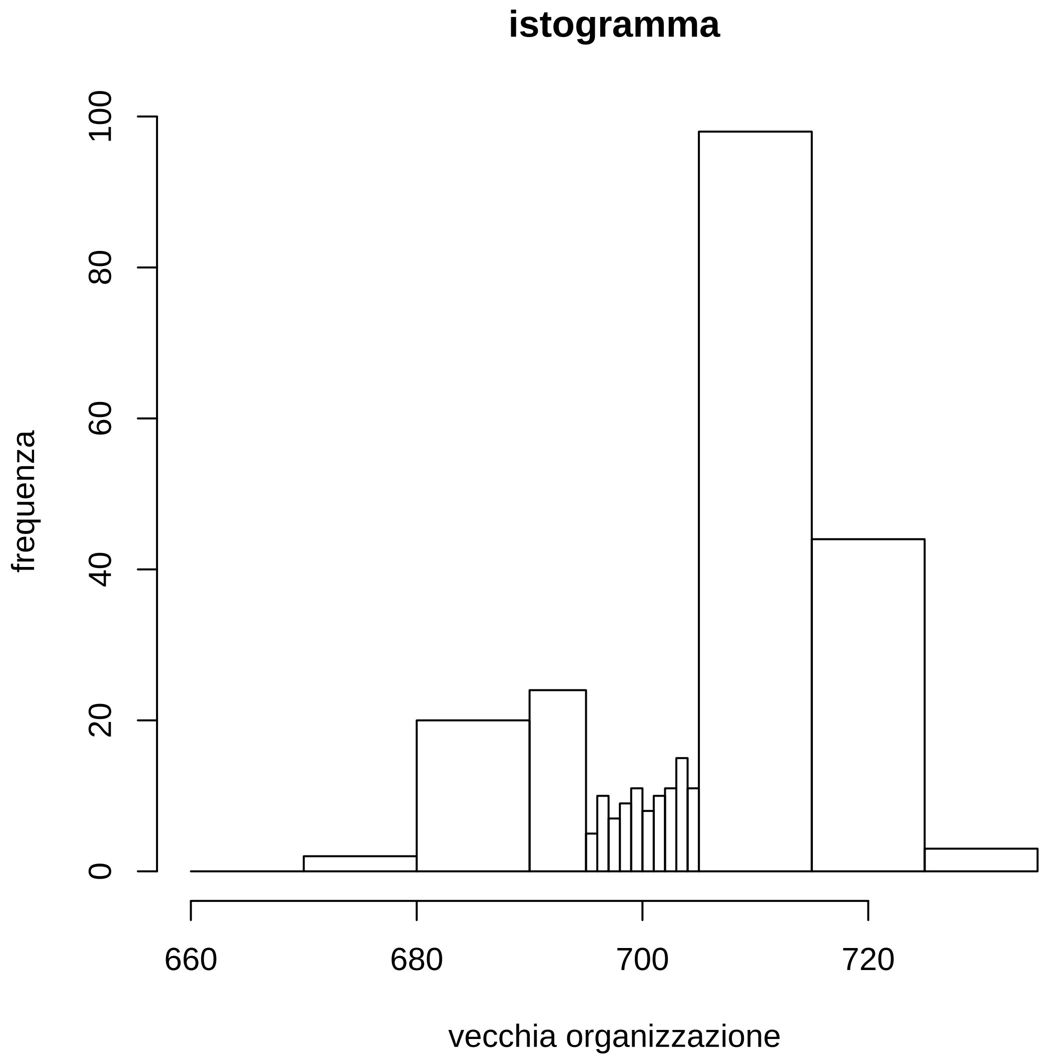
$$\begin{array}{c}\text{densità}\\ \text{di un intervallo}\end{array} = \frac{\text{frequenza dell'intervallo}}{\text{lunghezza dell'intervallo}}.$$

Per capire la definizione si pensi alla popolazione. E' la densità della popolazione non il numero totale di abitanti che ci dice quanto gli individui sono *addensati* in una certa regione geografica.

Istogramma per organizzazione "vecchia" costruito con:

1) intervalli più piccoli nella parte centrale;

2) altezze dei rettangoli proporzionali alle frequenze.

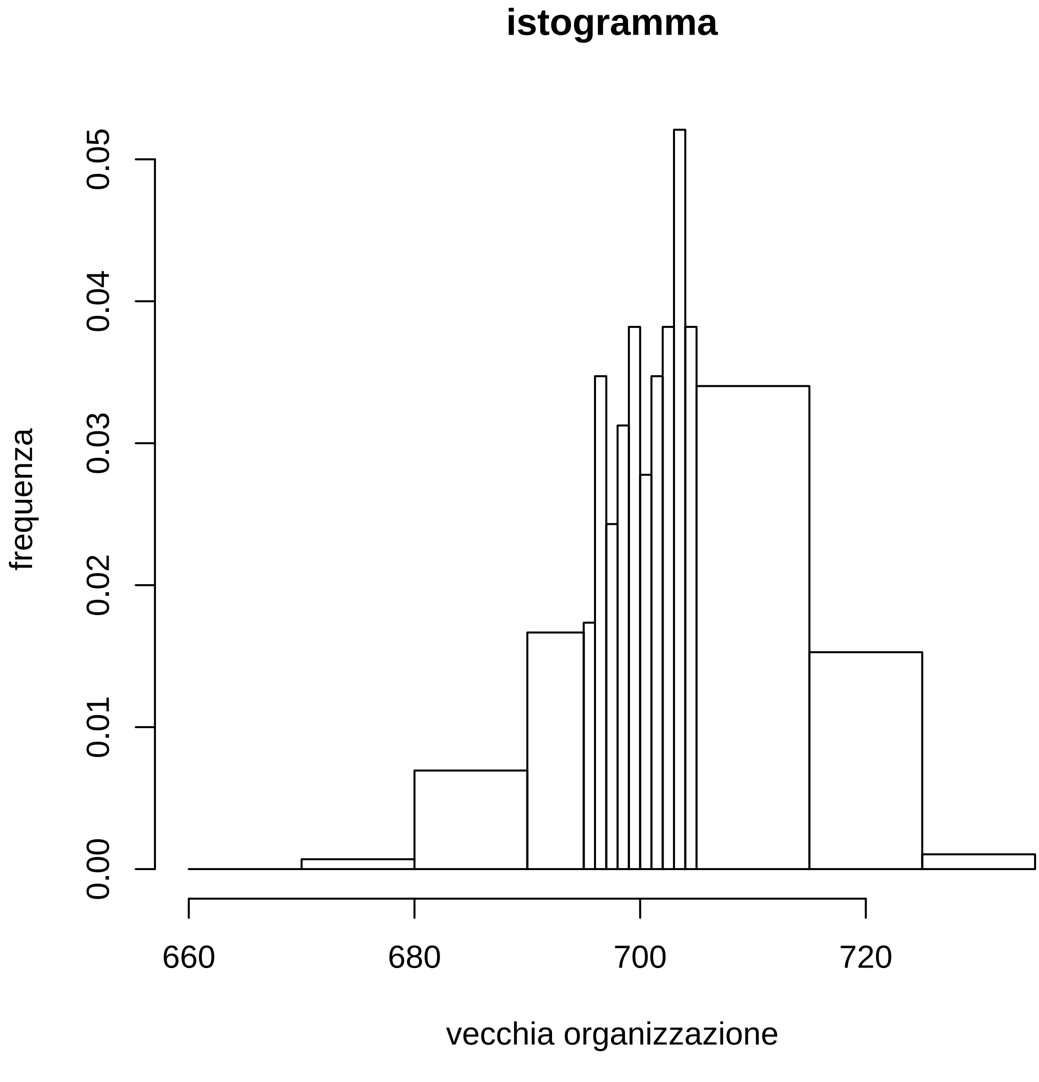


Sembra esserci un buco al centro, esattamente dove le osservazioni sono più *addensate*.

Istogramma per organizzazione “vecchia” costruito con:

1) intervalli più piccoli nella parte centrale;

2) altezze dei rettangoli proporzionali alle densità.



Il buco al centro è sparito. Il grafico correttamente ci dice che le osservazioni sono *addensate* intorno a 705.

Frequenze cumulate

Si ottengono "cumulando" progressivamente le frequenze.

Possono essere "assolute" o "relative".

Esempio di calcolo per organizzazione "nuova 1":

fine int.	freq. ass.	freq. cum. ass.	freq. cum. rel.
675	13	13	13/288=0.045
680	12	25=13+12	25/288=0.087
685	20	45=13+12+20	45/288=0.156
⋮	⋮	⋮	⋮
755	0	$287=13+12+\cdots+0$	287/288=0.997
760	1	$288=13+12+\cdots+0+1$	288/288=1

Funzione di ripartizione empirica

Osservazioni ordinate

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$$

Quindi

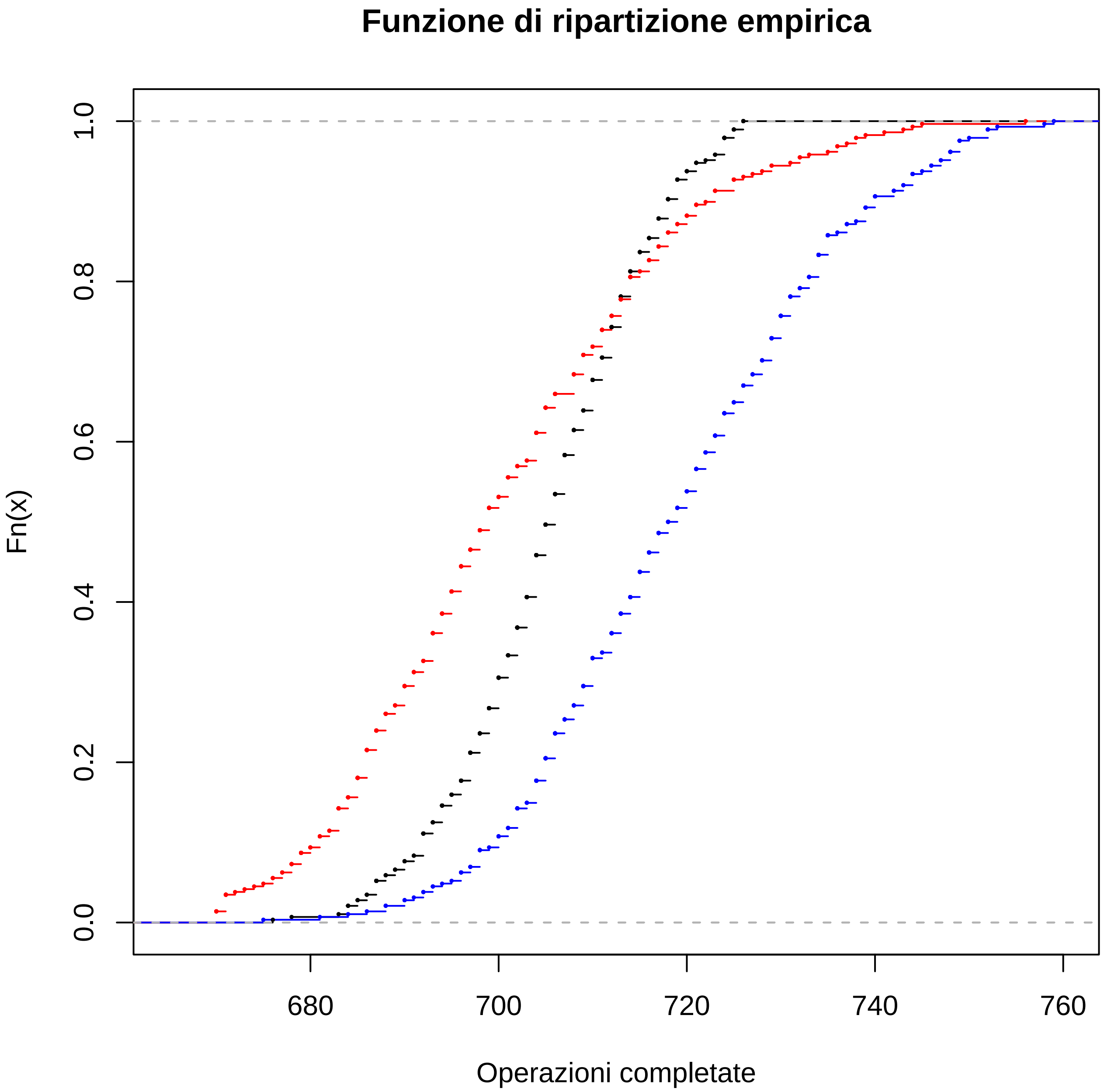
- la frazione $1/n$ di unità statistiche assumono valori della variabile Y inferiori o uguali ad $y_{(1)}$;
- la frazione $2/n$ di unità statistiche assumono valori della variabile Y inferiori o uguali ad $y_{(2)}$;
- ...
- la frazione i/n di unità statistiche assumono valori della variabile Y inferiori o uguali ad $y_{(i)}$;
- ...
- la frazione $n/n = 1$ di unità statistiche assumono valori della variabile Y inferiori o uguali ad $y_{(i)}$.

Funzione di ripartizione empirica:

$$\begin{aligned}\widehat{F}(y) &= \text{freq. rel. di unità che assumono valore} \leq y \\ &= \frac{\text{frequenza assoluta di unità che assumono valore} \leq y}{n}.\end{aligned}$$

Proprietà:

1. $0 \leq \widehat{F}(y) \leq 1$;
2. $\widehat{F}(-\infty) = 0$;
3. $\widehat{F}(\infty) = 1$;
4. $\widehat{F}(y)$ è una funzione (“a gradini”) non decrescente;
5. $\widehat{F}(y)$ è continua da destra.


Funzione di ripartizione empirica
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
680
700
720
740
760
Operazioni completate

Misure o parametri di posizione

Le distribuzioni dei pezzi prodotti differiscono soprattutto per la diversa posizione.

Nuova 2 sembra migliore di vecchia. Ma quanto migliore?

Un modo per rispondere a questa domanda consiste in:

1. sintetizzare le singole distribuzioni in un unico numero, detto **misura (o parametro o indice) di posizione**, che indichi in qualche modo dove la distribuzione stessa è posizionata.

2. confrontare gli indici calcolati al punto precedente.

Noti parametri di posizione sono: la **media aritmetica**, la **mediana** e i **quantili**.

La media aritmetica

Supponiamo di aver rilevato un certo fenomeno (esprimibile numericamente) su n unità statistiche diverse. Indichiamo con $y_1, y_2, \ldots, y_n$ i valori osservati (ovvero, i nostri dati).

La **media aritmetica** dei dati è

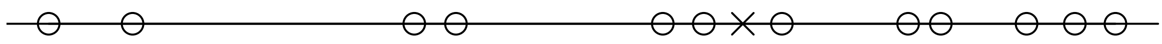
$$\overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Esistono altri tipi di medie. Quella aritmetica è senza ogni dubbio quella di utilizzo più comune. Per questo motivo, viene comunemente indicata come la media, senza nessuna ulteriore aggettivazione.

La mediana

L'idea che è alla base della **mediana** è di cercare un numero che sia più grande di un 50% delle osservazioni e più piccolo del restante 50%.

Ad esempio nel grafico seguente, supponendo che le osservazioni corrispondano ai punti disegnati con una 'o', un possibile valore per la mediana è stato indicato con una 'x'. Infatti, il punto così marcato lascia sia a sinistra che a destra 6 osservazioni.



Media e mediana: il caso delle tre organizzazioni del lavoro

	Vecchia	Nuova 1	Nuova 2
media	705,5	700,8	719,2
mediana	706	699	718,5

Come si vede, risulta confermato quanto osservato già graficamente: l'organizzazione nuova 2 potrebbe far aumentare la produzione di circa il 2%.

I quantili

I quantili generalizzano la mediana.

L'idea alla base di un **quantile** di ordine p, dove $p \in (0,1)$, indicato con y_p, è di cercare un numero che sia più grande del $100 \times p\%$ dei dati osservati e più piccolo del restante $100 \times (1-p)\%$. Ad esempio, un quantile di ordine 0,1 deve essere un valore che lascia a sinistra il 10% delle osservazioni ed a destra il restante 90%.

I quantili con p uguale a 0,25, 0,50 e 0,75 vengono chiamati rispettivamente il primo, il secondo e il terzo **quartile**. Dividono la popolazione in quattro parti uguali. Si osservi che il 2° quartile coincide con la mediana.

I quantili con $p = 0{,}01, \ldots, 0{,}99$ si chiamano **percentili**.

Alcune proprietà della media aritmetica

1. Se i dati sono tutti uguali ad una costante, diciamo a, allora anche la media è uguale ad a.

   Infatti, se

   $$y_1 = y_2 = \cdots = y_n = a$$

   allora

   $$\overline{y} = \frac{\overbrace{a + \cdots + a}^{n \text{ volte}}}{n} = \frac{na}{n} = a$$

La media è sempre compresa tra il più piccolo e il più grande dei valori osservati.

In simboli,

$$y_{(1)} \leq \overline{y} \leq y_{(n)}$$

dove

$$y_{(1)} = \min \{y_1, \ldots, y_n\}$$

e

$$y_{(n)} = \max \{y_1, \ldots, y_n\}$$

Infatti, ad esempio, per quanto riguarda la prima disuguglianza

$$y_{(1)} = \frac{\overbrace{y_{(1)} + \cdots + y_{(1)}}^{n \text{ volte}}}{n} \leq \frac{y_1 + y_2 + \cdots + y_n}{n} = \overline{y}$$

2. La media di una trasformazione lineare dei dati è la stessa trasformazione lineare applicata alla media dei dati.

   Ovvero, se $z_1 = a + by_1$, $z_2 = a + by_2, \ldots, z_n = a + by_n$ dove a e b sono due numeri qualsiasi, allora

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = a + b\overline{y}.$$

   Si osservi come la relazione precedente permetta di calcolare agevolmente la media delle z_i *senza* dover calcolare le z_i stesse.

   La dimostrazione è anche in questo caso immediata. Infatti

$$\begin{aligned}
\overline{z} &= \frac{z_1 + z_2 + \cdots + z_n}{n} = \\
&= \frac{(a + by_1) + (a + by_2) + \cdots + (a + by_n)}{n} = \\
&= \frac{\overbrace{a + \cdots + a}^{n \text{ volte}}}{n} + b\frac{y_1 + y_2 + \cdots + y_n}{n} \\
&= a + b\overline{y}.
\end{aligned}$$

3. La somma delle differenze dei dati dalla media (i cosiddetti **scarti**) è sempre uguale a zero:

$$\sum_{i=1}^{n}(y_i - \overline{y}) = (y_1 - \overline{y}) + (y_2 - \overline{y}) + \cdots + (y_n - \overline{y}) = 0.$$

Si tratta di una conseguenza della proprietà precedente (basta porre $a = -\overline{y}$ e $b = 1$).

4. La somma dei quadrati degli scarti da una costante è minima se e solo se la costante è posta uguale alla media.

   Ciò è conseguenza del fatto che, se a è un numero qualsiasi, allora

$$\sum_{i=1}^{n}(y_i - a)^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{y} - a)^2 \qquad (1)$$

   Infatti (tutte le sommatorie vanno da 1 a n)

$$\begin{aligned}
\sum(y_i - a)^2 &= \sum(y_i - a + \overline{y} - \overline{y})^2 = \\
&= \sum\left[(y_i - \overline{y}) + (\overline{y} - a)\right]^2 \\
&= \sum\left[(y_i - \overline{y})^2 + (\overline{y} - a)^2 + 2(\overline{y} - a)(y_i - \overline{y})\right] \\
&= \sum(y_i - \overline{y})^2 + \sum(\overline{y} - a)^2 + 2(\overline{y} - a)\sum(y_i - \overline{y}) \\
&= \sum(y_i - \overline{y})^2 + n(\overline{y} - a)^2 + 2(\overline{y} - a) \times 0.
\end{aligned}$$

   La (1) garantisce che

$$\sum_{i=1}^{n}(y_i - a)^2 > \sum_{i=1}^{n}(y_i - \overline{y})^2 \text{ se } a \neq \overline{y}.$$

Un difetto della media aritmetica

Non è del tutto infrequente trovare degli insiemi di dati contenenti una piccola frazione di **osservazioni anomale**, ovvero osservazioni che assumono valori lontani da quelli assunti dalla maggior parte delle altre osservazioni e che, quindi, sembrano provenire da una popolazione diversa o essere state generate da un meccanismo differente.

In una situazione del tipo descritto, bisogna tenere presente che la media aritmetica può essere molto sensibile alla presenza delle osservazioni anomale potendo anche, a volte, fornire risultati non molto sensati.

Infatti una sola osservazione molto grande o molto piccola può *dominare* nel calcolo della media tutte le altre osservazioni.

Esercizio: Si supponga di avere 10.000 osservazioni, $y_1, \ldots, y_{10.000}$, tali che $y_i \in [0, 1]$ quando $2 \leq i \leq 10.000$ (ovvero, tutte le osservazioni con la possibile eccezione della prima sono comprese tra 0 e 1). Mostrare che

$$\lim_{y_1 \to -\infty} \frac{1}{n} \sum_{i=1}^{n} y_i = -\infty$$

e commentare il risultato.

Alcune proprietà della mediana

1. Siano $y_1, \ldots, y_n$ dei numeri reali qualsiasi e sia m un valore tale che

   $$(\text{numero dati} < m) = (\text{numero dati} > m).$$

   Allora
   $$\sum_{i=1}^{n} |y_i - m| \leq \sum_{i=1}^{n} |y_i - a|$$
   per qualsivoglia costante a.
   Ovvero, la mediana è il numero che minimizza la somma dei valori assoluti degli scarti di un insieme di dati da una costante.

2. La mediana è, come si usa dire, **resistente**, ovvero non molto sensibile alla presenza di valori anomali.

Esempi di calcolo della mediana

Minori problemi di calcolo possono sorgere dato che
(i) non è detto che esista un valore maggiore di un 50% esatto dei dati e minore dei restanti
(ii) può esistere ma non essere unico.
Illustriamo i vari casi e delle *ragionevoli* soluzioni con semplici esempi numerici.

1. Dati: 1, 4, 2, 9, 3.
   Dati ordinati: 1, 2, 3, 4, 9.
   5 osservazioni, non esiste un numero che lascia esattamente un 50% di osservazioni sulla destra ed un 50% sulla sinistra; però la terza osservazione dal basso lascia a sinistra e a destra lo stesso numero di dati. Sembra quindi *sensato* porre (mediana) = 3.

2. Dati: 1, 2, 1, 5.
   Dati ordinati: 1, 1, 2, 5.
   4 dati; qualsiasi numero tra 1 e 2 lascia a sinistra e a destra esattamente un 50% delle osservazioni; tipicamente si pone

$$\text{mediana} = \begin{pmatrix} \text{punto centrale} \\ \text{dell'intervallo} \end{pmatrix},$$

   in questo caso, mediana $= (1 + 2)/2 = 1{,}5$.

3. Dati: 4, 3, 2, 2, 5, 2, 6, 5, 1, 3.
   Dati ordinati: 1, 2, 2, 2, 3, 3, 4, 5, 5, 6
   Il numero di osservazioni è pari come nel caso 2 precedente. La presenza di osservazioni ripetute rende però la situazione simile a quella dell'esempio 1. Sembra in questo caso *sensato* porre (mediana) $= 3$.

4. Supponiamo in questo caso di avere i seguenti dati *raggruppati*:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
freq. ass.	1	4	4	2	1

I dati sono 12. La mediana dovrebbe essere scelta tra la 6° e la 7° osservazione dal basso. Sulla base dei dati disponibili possiamo quindi affermare che la mediana in questo caso appartiene all'intervallo $(2,3]$. Volendo assegnarle un valore numerico preciso, potremmo supporre che i quattro dati appartenenti al terzo intervallo siano equidistribuiti ed, ad esempio, uguali a 2,25, 2,50, 2,75, 3,00*. Sotto questa assunzione, ricordiamoci arbitraria, la 6° e la 7° osservazione dal basso sarebbero rispettivamente uguali a 2,25 e a 2,50. Potremmo quindi porre (mediana) = 2,375.

*Si osservi che è facile *inventarsi* altri ipotetici valori equidistribuiti. Ad esempio 2,2, 2,4, 2,6, 2,8

Ambiguità nel calcolo dei quartili (e, quindi, di un quantile)

Un valore con esattamente la proprietà richiesta ad un quantile può non esistere o, viceversa, non essere unico. Per il calcolo si vedano, i seguenti esempi, oltre a quelli sulla mediana.

Dati (già ordinati):

6,4 6,7 6,8 7,0 7,3 7,5 7,5 7,6 7,9 8,1

La mediana deve cadere tra $7,3$ e $7,5$. Tradizionalmente, si sceglie il punto centrale dell'intervallo, ovvero si pone mediana $=$ 7,4.

La determinazione del primo (e del terzo) quartile è più ambigua. Il primo quartile dovrebbe lasciare sulla sinistra il 25% delle osservazioni, ovvero in questo caso 2,5 osservazioni. Questo è ovviamente impossibile da raggiungere esattamente. Esistono vari ragionamenti che possono essere utilizzati per *sciogliere* l'ambiguità. Ad esempio,

1. potremmo *decidere* di interpretare "lasciare a sinistra 2,5 osservazioni" come "posizionarsi sul punto intermedio tra la seconda e la terza osservazione (dal basso)" ovvero di *assegnare* al primo

quartile il valore di 6,75. Allora, in maniera analoga potremmo *assegnare* al terzo quartile il valore di 7,75 (= punto intermedio tra l'ottava e la nona osservazione).

2. oppure, potremmo *decidere* che il primo quartile deve dividere le osservazioni alla sinistra della mediana in due parti uguali. Quindi, poiché abbiamo alla sinistra della mediana 5 osservazioni, decidere di *porre* il primo quartile uguale al terzo dato dal basso (ovvero a 6,8). Argomentando in maniera analogo assegneremo al terzo quartile il valore 7,6 (= terza osservazione dal basso nel gruppo a destra della mediana).

Nessuna delle due scelte è migliore dell'altra. Si tenga comunque presente che, a meno di casi particolari, più il numero di osservazioni diventa grande, più le varie possibilità tendono ad avvicinarsi. Ad esempio, supponiamo di avere 99 dati già ordinati in senso crescente

$$y_1, \ldots, y_{24}, y_{25}, \cdots y_{49}, y_{50}, y_{51}, \ldots, y_{99}.$$

Allora il primo quartile dovrebbe lasciare $(25 \times 99)/100 =$ 24,75 osservazioni a sinistra. Questo è impossibile. Le due "soluzioni" viste prima continuano a dare "soluzioni" diverse:

1. nel primo caso infatti potremmo interpretare "lasciare 24,75 osservazioni a destra" come "posizionarsi a tre quarti dell'intervallo $[y_{24}, y_{25}]$ ovvero calcolare il primo quartile come $0{,}25 y_{24} + 0{,}75 y_{25}$;

2. nel secondo caso, viceversa, calcoleremmo il primo quartile come la mediana di $y_1, \ldots, y_{49}$ e quindi gli assegneremmo il valore di y_{25}.

Però più è elevato il numero di osservazioni più ci aspettiamo che l'intervallo in cui ha senso scegliere il primo quartile sia piccolo. Infatti, più osservazioni abbiamo più ce le aspettiamo *addensate*.

Dati raggruppati: approssimazione della media

Supponiamo di non conoscere i dati individuali (ovvero riferiti alle singole unità statistiche) ma solo una distribuzione di frequenza per intervalli del tipo

intervalli	$[a_0, a_1)$	$[a_1, a_2)$	$\cdots$	$[a_{k-1}, a_k)$
frequenze assolute	f_1	f_2	$\cdots$	f_k

dove k indica il numero degli intervalli.

La media non può essere calcolata esattamente.

Un'*approssimazione* spesso usata in questi casi è

$$\frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i} = \frac{1}{n} \sum_{i=1}^{k} m_i f_i$$

dove m_i è il punto centrale dell'intervallo i-simo, ovvero

$$m_i = \frac{a_{i-1} + a_i}{2}$$

Esercizio-Interpretazione

Si mostri come l'approssimazione per la media appena vista possa essere ottenuta *facendo finta* o che (i) tutte le osservazioni nell'intervallo i-simo siano tutte uguali a m_i o che (ii) le osservazioni appartenti all'intervallo i-simo siano *equidistribuite* nell'intervallo stesso (equidistribuite = uguale distanza tra le osservazioni successive).

Si dica inoltre quale delle seguenti due affermazioni è vera e quale è falsa:

1. Più gli intervalli sono grandi (lunghi) più l'approssimazione è accurata.

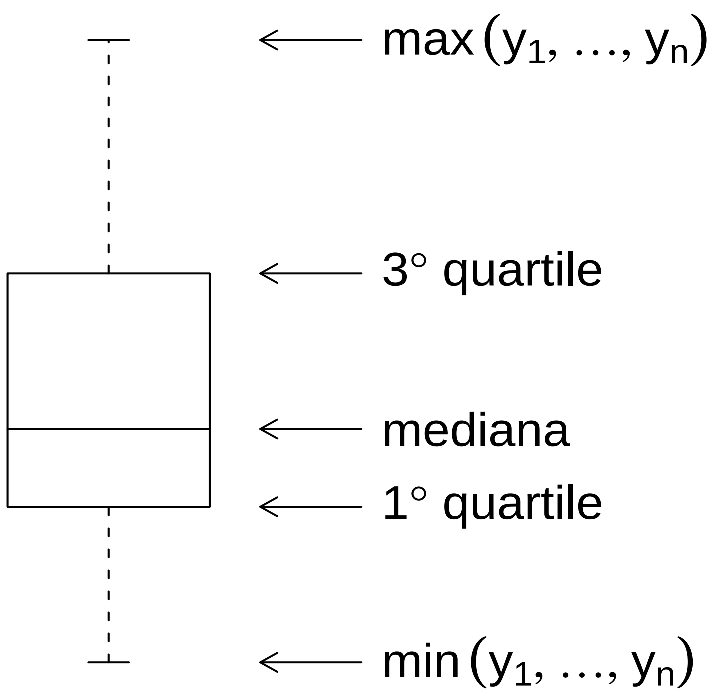
2. Più piccoli (corti) sono gli intervalli più l'approssimazione è accurata.

Diagrammi a scatola con baffi

Il nome deriva dall'inglese (box and whiskers plot, spesso abbreviato in **boxplot**).

Forniscono un'idea schematica di un insieme di dati basata sui quantili.

Sono costituiti, come dice il nome, da una scatola e da due baffi costruiti in accordo al disegno sottostante:



Una variante

Variante comunemente usata del boxplot:

1) la scatola è costruita come descritto precedentemente a partire dai tre quartili;

2) i baffi si estendono fino ai dati più lontani che siano però non più distanti di k volte lo scarto interquartile dalla scatola. Lo **scarto interquartile** è la differenza tra il terzo e il primo quartile (ossia l'ampiezza della scatola), k è una costante arbitraria tipicamente scelta uguale a 1.5. Ovvero, non accettiamo baffi esageratamente lunghi;

3) le osservazioni che sono oltre i baffi sono disegnate opportunamente sul grafico (ad esempio utilizzando un pallino).

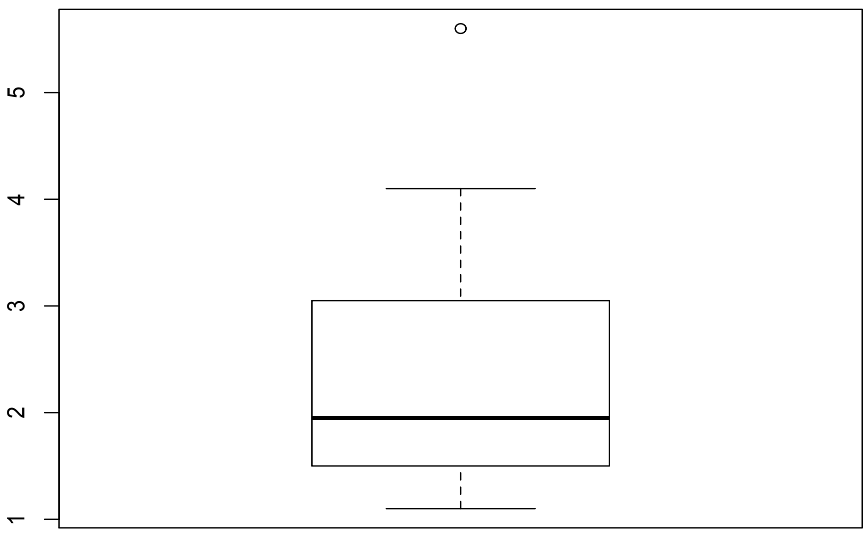
Esempio di costruzione di un boxplot

Dati (già ordinati):

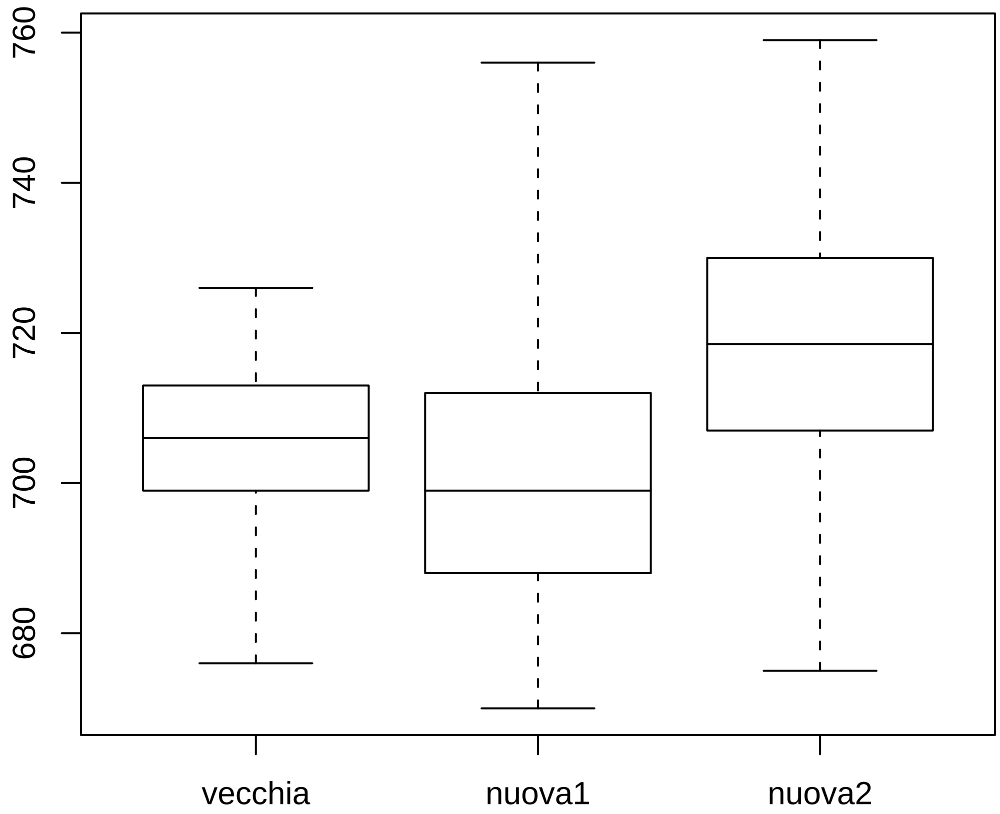
1.1 1.3 1.4 1.6 1.8 1.9 2.0 2.5 2.9 3.2 4.1 5.6

Perciò $y_{0.25} = 1.5$, $y_{0.5} = 1.95$ e $y_{0.75} = 3.05$. Quindi 1.5×(scarto interquartile)= $1.5 \times 1.55 = 2.325$. Di conseguenza:

1. la scatola si estende da 1.5 a 3.05;
2. il baffo inferiore si estende fino all'osservazione più piccola tra quelle maggiori di $y_{0.25} - 2.325 = -0.825$, ovvero fino a 1.1;
3. il baffo superiore si estende fino all'osservazione più grande tra quelle minori di $y_{0.75} + 2.325 = 5.375$, ovvero fino a 4.1;
4. vanno disegnate esplicitamente nel diagramma le osservazioni più piccole di 1.1 o più grandi di 5.375; in questo caso l'osservazione pari a 5.6.



Le tre organizzazioni della produzione

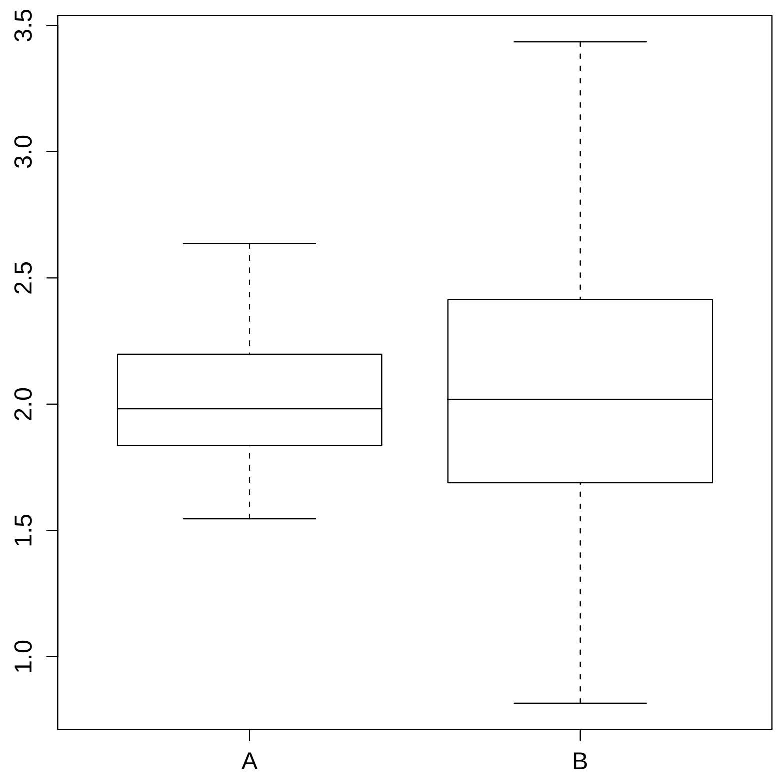


La variabilità

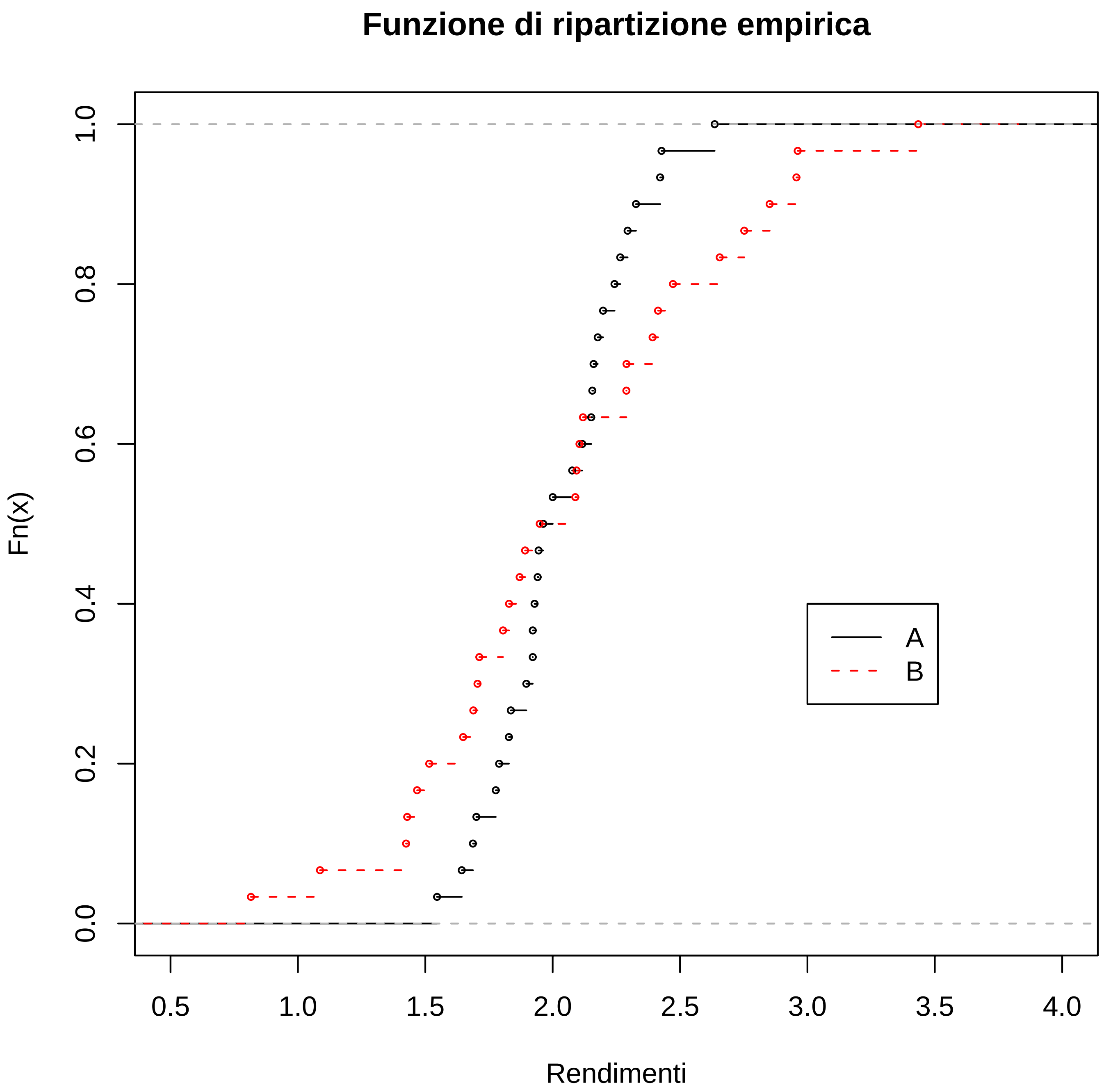
Per confrontare le *performance* di due tipologie di fondi, etichettate come A e B abbiamo preso in considerazione i rendimenti di 30 fondi per ciascuna tipologia. Riportiamo di seguito i diagrammi a scatola dei rendimenti.

Gruppo A
1.643 2.117 1.897 1.836 2.294 1.929 2.243 1.777 1.922 1.945
2.156 2.265 2.177 1.941 2.198 1.922 1.828 2.422 2.151 1.790
2.427 1.687 2.000 2.327 1.700 2.160 1.963 2.636 1.546 2.077

Gruppo B
2.752 1.805 2.290 2.105 2.472 1.087 3.435 0.816 1.705 1.516
2.094 2.957 1.689 1.468 1.829 1.949 2.289 2.414 2.656 2.089
2.852 1.712 1.649 1.870 2.962 1.892 1.429 2.392 1.424 2.119



e le rispettive funzioni di ripartizione



Commento

1. Ambedue le tipologie sembrano produrre *in media* lo stesso rendimento, visto che i due insiemi di dati si distribuiscono intorno al valore 2%.

2. Però i rendimenti della tipologia B sembrano essere più difformi tra di loro. Infatti in questo caso i dati sono più *dispersi* intorno al valore 2%. Ovvero, come si usa dire, mostrano una **variabilità** superiore.

Nota: E' importante capire che l'incrocio delle due funzioni di ripartizione empiriche è dovuto alla differente variabilità dei due insiemi di dati.

La varianza

Così come per la posizione, è interessante disporre di indici che ci permettano di valutare in maniera sintetica la variabilità di un insieme di dati.

Il più usato prende il nome di **varianza**:

$$\text{varianza}(y_1, \ldots, y_n) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

dove con $y = (y_1, \ldots, y_n)$ abbiamo indicato i dati osservati, con n il loro numero e con $\overline{y}$ la loro media aritmetica, ovvero

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

In seguito varianza$(y_1, \ldots, y_n)$ verrà abbreviato in var(y).

La varianza è quindi una misura di quanto i dati siano *distanti* dalla media aritmetica. La distanza è valutata usando i quadrati delle differenze.

Formula per il calcolo

Si osservi che

$$
\begin{aligned}
\mathrm{var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = & \\
= & \; \frac{1}{n}\sum_{i=1}^{n} y_i^2 + \frac{1}{n}\sum_{i=1}^{n} \overline{y}^2 - \frac{1}{n}\sum_{i=1}^{n} 2\overline{y}y_i = \\
= & \; \frac{1}{n}\sum_{i=1}^{n} y_i^2 + \frac{n\overline{y}^2}{n} - \frac{2\overline{y}}{n}\sum_{i=1}^{n} y_i = \\
= & \; \frac{1}{n}\sum_{i=1}^{n} y_i^2 + \overline{y}^2 - 2\overline{y}^2
\end{aligned}
$$

e quindi che possiamo scrivere

$$\mathrm{var}(y) = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \overline{y}^2$$

ovvero

$$(\text{varianza}) = \left(\begin{array}{c}\text{media dei}\\ \text{quadrati}\end{array}\right) - \left(\begin{array}{c}\text{quadrato}\\ \text{della media}\end{array}\right).$$

Esempio di utilizzo

dati: $1, 3, 2, 5$.

media: $\dfrac{1+3+2+5}{4} = 2{,}75$.

media dei quadrati: $\dfrac{1^2+3^2+2^2+5^2}{4} = 9{,}75$.

varianza: $9{,}75 - 2{,}75^2 = 2{,}19$.

Esercizio: Si dia una formula generale della varianza nel caso di una tabella di frequenza. Si verifichi che nella tabella seguente

y_i	f_i
4	2
6	8
7	3

$\mathrm{var}(y) \simeq 0.840$.

Varianza di trasformazioni lineari dei dati

Dati: $y = (y_1, \ldots, y_n)$.

Dati trasformati: $z = (z_1, \ldots, z_n)$, con $z_i = a + by_i$, $i = 1, \ldots, n$, dove a e b sono due costanti qualsiasi.

Allora

$$\text{var}(z) = b^2 \text{var}(y).$$

Sappiamo infatti che

$$\bar{z} = a + b\bar{y}.$$

Quindi,

$$\text{var}(z) = \frac{1}{n} \sum_{i=1}^{n} (a + by_i - a - b\bar{y})^2 =$$
$$= \frac{b^2}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 = b^2 \text{var}(y).$$

Esercizio: La formula mostra che la varianza delle z_i non dipende da a ("l'intercetta" della trasformazione). Si spieghi perché il contrario sarebbe stato quantomeno bizzarro e, per molti versi, preoccupante.

Lo scarto quadratico medio

La radice quadrata della varianza è usualmente chiamata **scarto quadratico medio**. Useremo l'abbreviazione $\text{sqm}(y)$. Quindi

$$\text{sqm}(y) = \sqrt{\text{var}(y)}.$$

Si osservi che mentre l'unità di misura della varianza è uguale al quadrato dell'unità di misura dei dati originali, l'unità di misura dello scarto quadratico medio coincide con l'unità di misura dei dati.

Altre misure di variabilità

In aggiunta alla varianza sono stati suggeriti e sono utilizzati una molteplicità di indici (misure) di variabilità.

Ne elechiamo tre tra i più diffusi:

1. **Campo di variazione**

   $$\begin{pmatrix}\text{Campo di}\\ \text{variazione}\\ (\textit{range})\end{pmatrix} = \max(y_1, \ldots, y_n) - \min(y_1, \ldots, y_n).$$

   Veloce da calcolare ma *pericoloso* perché troppo sensibile a possibili valori anomali.

2. **Scarto interquartile**

   $$\text{Scarto interquartile} = (3^\circ \text{ quartile}) - (1^\circ \text{ quartile}).$$

   E' molto più *resistente* della varianza in presenza di poche osservazioni estreme. Per questo motivo è usato soprattutto nelle situazioni in cui si sospetta la possibile presenza di osservazioni anomale.

3. **MAD**

$$\text{MAD} = \text{mediana}(|y_1 - y_{0,5}|, \ldots, |y_n - y_{0,5}|)$$

dove $y_{0,5}$ indica la mediana dei dati. L'acronimo deriva dall'inglese (*Median Absolute Deviations*). Anche questo indice è poco sensibile alla presenza di valori anomali.

Indici di variabilità per due tipologie di fondi

	A	B
varianza	0,06	0,34
scarto quadratico medio	0,25	0,58
campo di variazione	1,09	2,62
scarto interquartile	0,34	0,72
MAD	0,31	0,58

La tabella mostra chiaramente come tutti gli indici considerati evidenzino la maggiore variabilità dei rendimenti (leggi 'rischio') dei fondi di tipo B.

Variabili qualitative

I dati si riferiscono ad un'indagine ISTAT condotta nel 2001 sugli esercizi ricettivi, ovvero alberghi, campeggi e villaggi turistici, alloggi agro-turistici ed altri esercizi (ostelli, case per ferie, rifugi alpini, .etc.), divisi per area geografica.

I dati prendono la forma di una lunga tabella di questo tipo:

esercizio	tipo	area geografica
1	albergo	Nord
2	camp. e vill. tur.	Sud
⋮	⋮	⋮

Per ogni esercizio (*unità statistica*) sono state rilevate due variabili: il *tipo* di esercizio e l'*area geografica* dell'esercizio.

Tabelle di frequenza

La variabile *tipo* ha la seguente distribuzione di frequenze

tipo	freq.	freq. rel.
Alberghi	33.338	0,314
Campeggi e villaggi turistici	2.371	0,022
Alloggi agro-turistici	7.769	0,073
Altri esercizi	62.727	0,591
TOTALE	106.205	1,00

La variabile *area geografica* ha invece la seguente distribuzione di frequenze

area geografica	freq.	freq. rel.
Nord	83.732	0,788
Centro	13.454	0,127
Sud	9.019	0,085
TOTALE	106.205	1,00

La moda

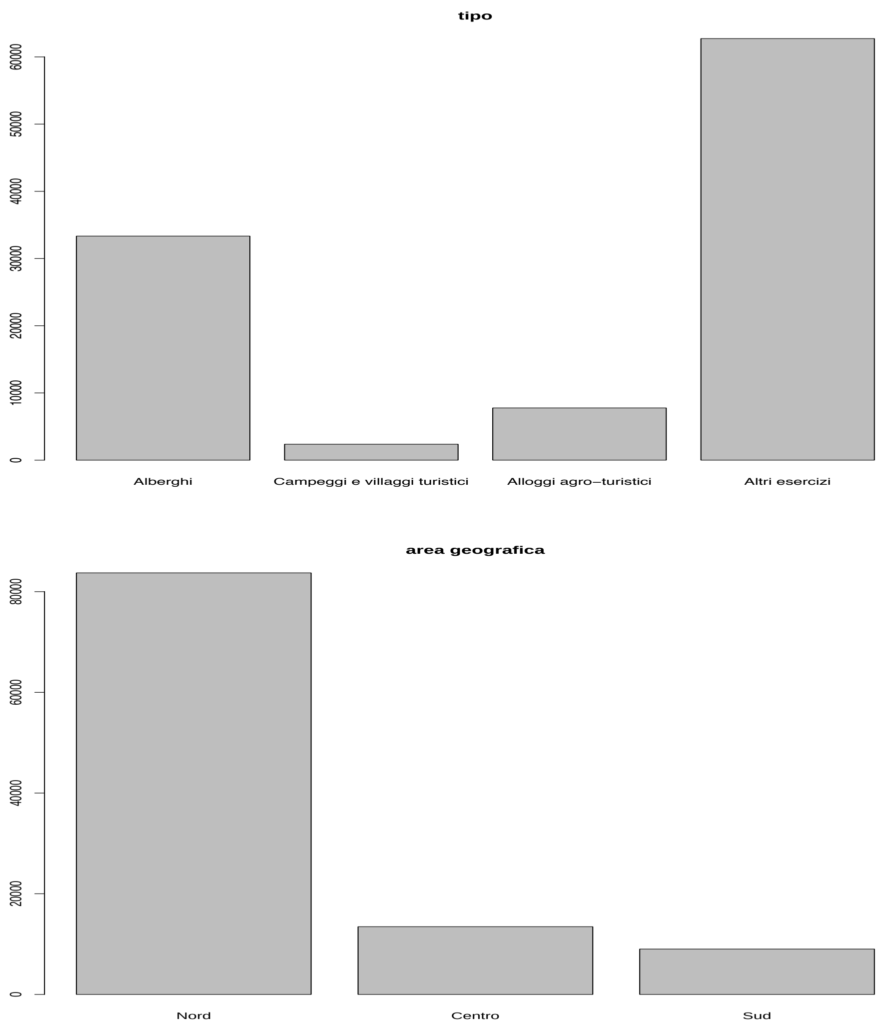
- Nei precedenti esempi avevamo dati **numerici**. In questo caso le caratteristiche rilevate sono espresse da aggettivi. Sono dei dati **qualitativi** o **categoriali**.

- Questo cambia quello che possiamo e non possiamo fare. Ad esempio, non ha senso chiederci quanto vale la media aritmetica dell'area geografica per gli esercizi. O quanto è grande la varianza.

- Volendo sintetizzare ogni variabile in un unico valore probabilmente useremo la **moda** della variabile. Definiamo la moda come la modalità con la più alta frequenza. In questo caso, la moda della variabile *tipo* è la modalità *Altri esercizi*, con frequenza relativa pari a 0,591. La moda della variabile *area geografica* è invece la modalità *Nord*, con frequenza relativa pari a 0,788.

- Si osservi che la moda può essere usata per qualsiasi distribuzione di frequenza. Anche per quelle basate su dati numerici.

Diagramma a barre: frequenze assolute



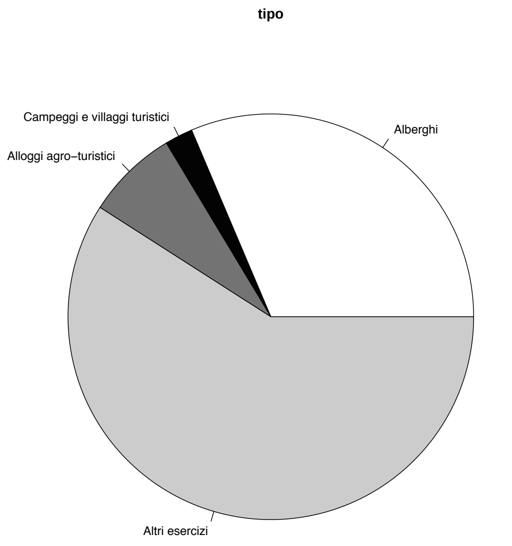
La rappresentazione grafica più utilizzata è il diagramma a barre, in cui ogni modalità è rappresentata da una barra di altezza pari alla frequenza (assoluta o relativa) della modalità. Si osservi che i rettangoli, contrariamente al caso di un istogramma, sono disegnati *staccati*.

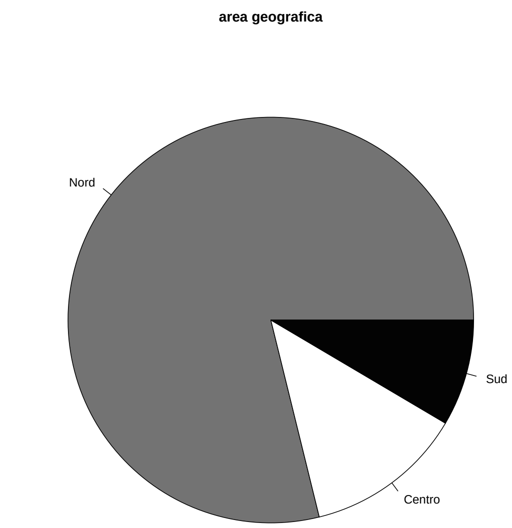
Notiamo che, se la variabile non è ordinale, l'ordine delle modalità nell'asse delle ascisse del grafico è arbitrario.

Diagramma a torte: frequenze relative

Una diversa rappresentazione grafica per variabili qualitative è data dal diagramma a torta, in cui ogni modalità è rappresentata da una fetta di torta proporzionale alla sua frequenza relativa:

$$\textit{angolo} = 360 \cdot \textit{frequenza relativa}$$





Mutabilità (idea di)

Il concetto di **mutabilità** è l'analogo per dati qualitativi della variabilità.

Nel caso di variabili qualitative, non possiamo guardare alle differenze tra i valori osservati. Possiamo però guardare alle differenze tra le frequenze.

Una situazione di *minima mutabilità* è quella in cui tutte le unità statistiche si *concentrano* nella stessa modalità. In questo caso le unità statistiche sono perfettamente omogenee rispetto al fenomeno considerato. La distribuzione delle frequenze relative si presenta come

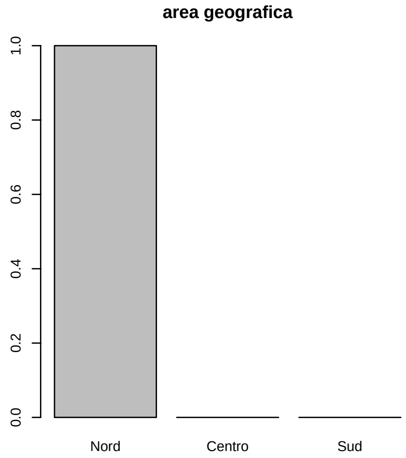
modalità	c_1	$\cdots$	c_i	$\cdots$	c_k
frequenza relativa	0	$\cdots$	1	$\cdots$	0

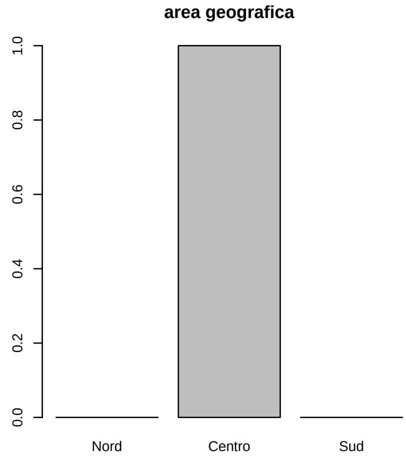
dove abbiamo supposto che le modalità siano k e che la i-sima sia quella in cui le unità statistiche si sono concentrate.

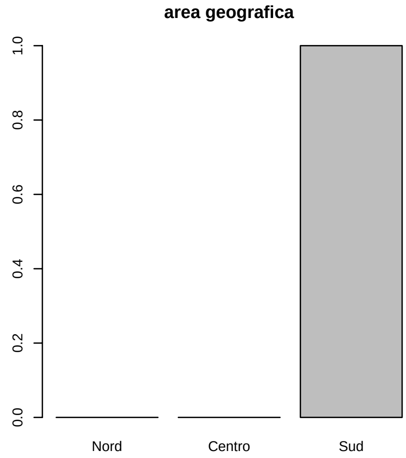
La situazione opposta (*massima mutabilità*) la troviamo invece quando le unità statistiche si ripartiscono in maniera uguale tra le varie modalità. In questo caso la distribuzione delle frequenze relative diventa

modalità	c_1	$\cdots$	c_i	$\cdots$	c_k
frequenza relativa	$\frac{1}{k}$	$\cdots$	$\frac{1}{k}$	$\cdots$	$\frac{1}{k}$

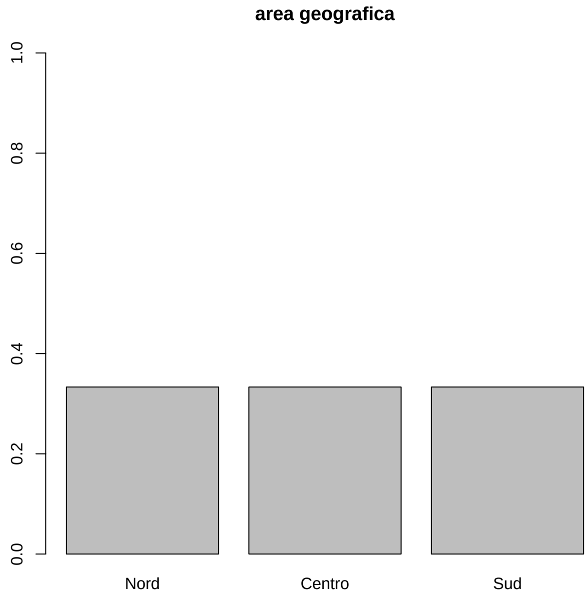
Ad esempio, per la variabilie *area geografica*, le tre situazioni di **minima** mutabilità sono rappresentate nei seguenti grafici







mentre la situazione di **massima** mutabilità corrisponde all'equidistribuzione tra le diverse modalità



Alcuni indici di mutabilità

Tabella delle frequenze relative:

modalità	c_1	$\cdots$	c_i	$\cdots$	c_k
frequenza relativa	p_1	$\cdots$	p_i	$\cdots$	p_k

- **Indice di Gini.**

$$G = \sum_{i=1}^{k} p_i(1 - p_i)$$

  – Si annulla in corrispondenza di una tabella di minima mutabilità.

  – Assume valore massimo nelle situazioni di massima mutabilità. Ovvero qualsiasi siano le frequenze relative,

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

– Spesso si usa la versione *normalizzata* di G

$$G_{norm} = \frac{G}{\max(G)} = \frac{k}{k-1}G$$

L'indice normalizzato varia tra 0 ed 1. In particolare, assume valore 0 in presenza di minima mutabilità e valore 1 in presenza di massima mutabilità.

– Nel caso in cui sia disponibile la tabella delle frequenze assolute

modalità	c_1	$\cdots$	c_i	$\cdots$	c_k	totale
freq ass	f_1	$\cdots$	f_i	$\cdots$	f_k	$n = \sum f_i$

può essere calcolato utilizzando la formula

$$G = 1 - \frac{1}{n^2}\sum_{i=1}^{k} f_i^2$$

(esercizio per lo studente).

– Per la variabile *area geografica* si ha $G = 0{,}3557$ e $G_{norm} = 0{,}5335$.

- **Entropia di Shannon.**

$$H = -\sum_{i=1}^{k} p_i \log(p_i)$$

dove, se $p_i = 0$ poniamo $p_i \log(p_i) = 0$.

  - Proviene dalla *teoria dell'informazione* dove viene utilizzato per misurare la complessità di un messaggio.

  - Si annulla, come è facile verificare, nelle situazioni di minima mutabilità.

  - Assume valore massimo nelle situazioni di massima mutabilità:

$$H \leq -\sum_{i=1}^{k} \frac{1}{k} \log\left(\frac{1}{k}\right) = -\log\left(\frac{1}{k}\right) = \log(k).$$

  - Può quindi essere eventualmente *normalizzato* ponendo $H_{norm} = H/\log(k)$.

  - Se sono note le frequenze assolute possiamo calcolare H utilizzando la formula

$$H = \log(n) - \frac{1}{n}\sum_{i=1}^{k} f_i \log(f_i).$$

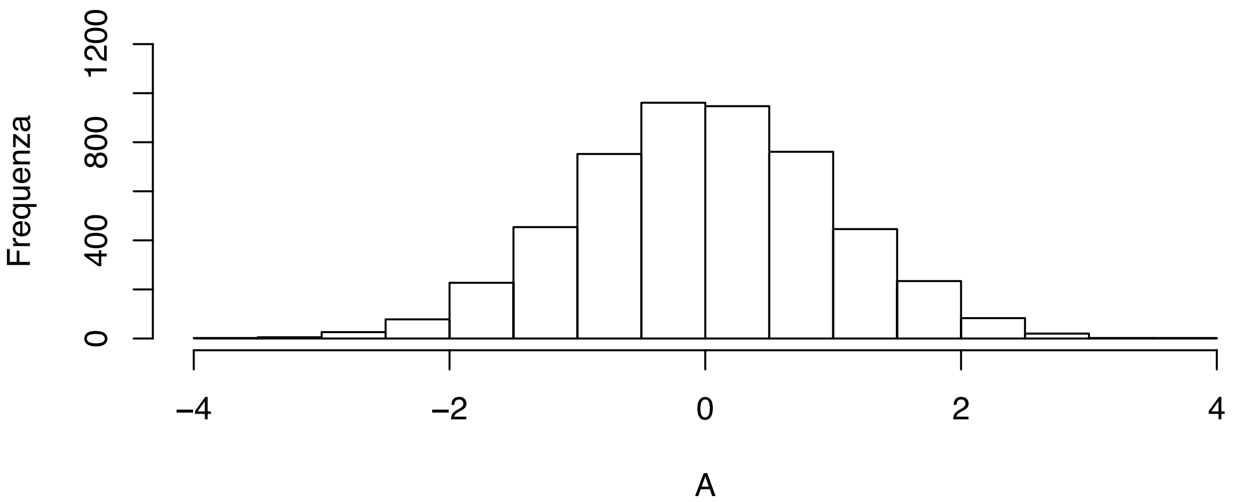
  - Per la variabile *area geografica* si ha $H = 0{,}6594$ e $H_{norm} = 0{,}6002$.

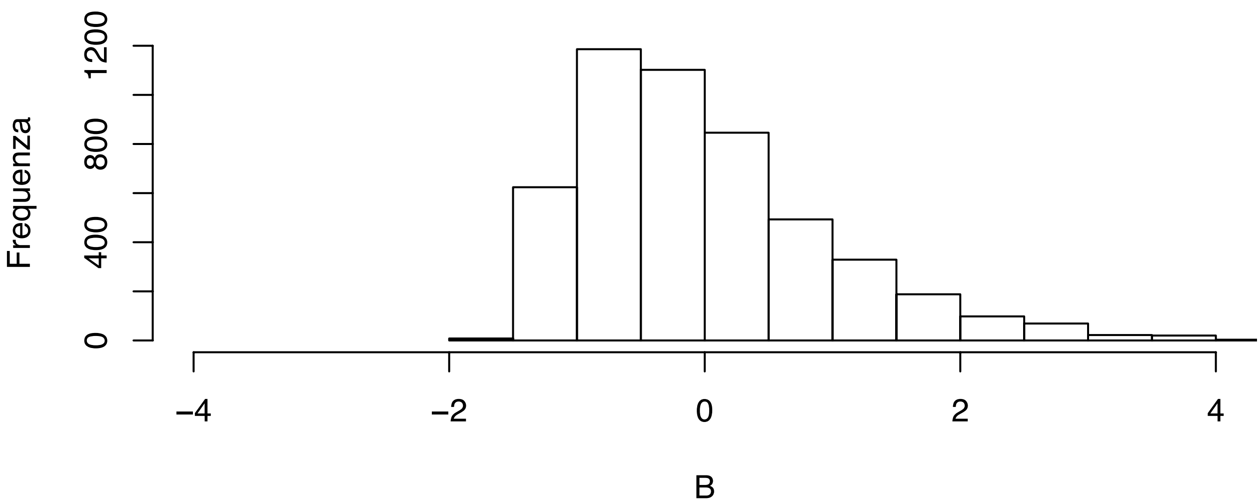
Esercizio: Provate a calcolare gli indici normalizzati di mutabilità per la variabile *tipo* di esercizio.

Simmetria e curtosi

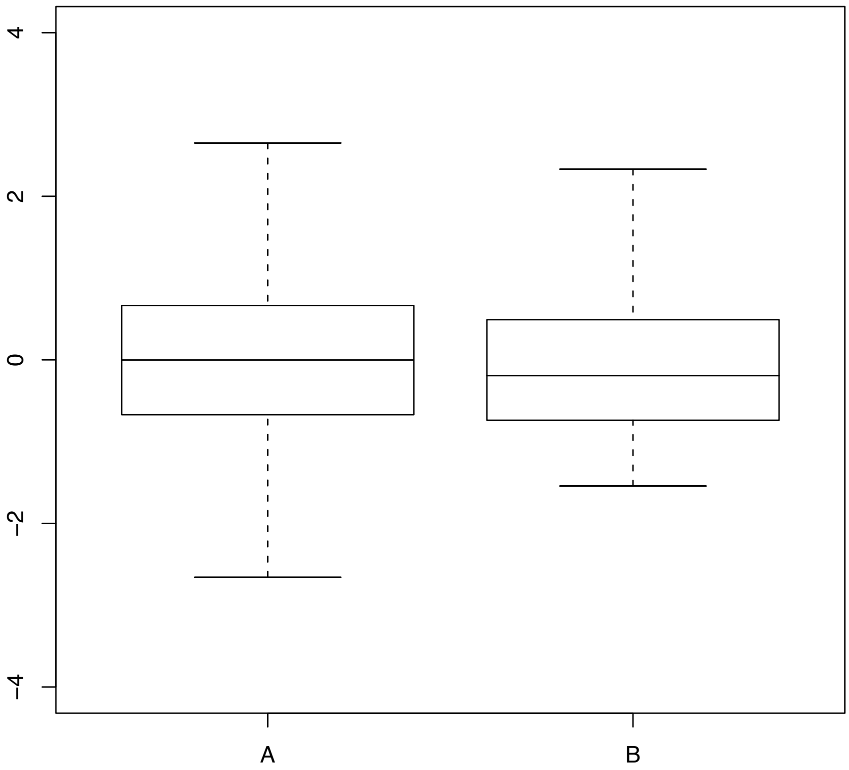
Consideriamo brevemente due aspetti di una distribuzione di frequenza, la **simmetria** e la **curtosi**, a volte interessanti di per sè ma soprattutto utili nella la scelta di un appropriato *modello statistico*.

Due insiemi di dati standardizzati: istogramma





Due insiemi di dati standardizzati: boxplot



Simmetria

I grafici precedenti mostrano rispettivamente gli istogrammi e i boxplot costruiti a partire da due insiemi di dati (A e B) *standardizzati*.

I due insiemi di dati sono perciò almeno approssimativamente omogenei per quanto riguarda posizione e variabilità. Hanno ambedue media nulla e varianza unitaria.

Nonostante questo le due distribuzioni sono diverse. La prima è più o meno **simmetrica** rispetto allo zero. Viceversa, la *coda verso i valori alti* della seconda è molto più lunga della *coda verso i valori bassi*. Si parla in questo caso di **asimmetria positiva**. Ovviamente, nel caso opposto (coda sinistra più lunga di quella destra) parleremo di **asimmetria negativa**.

Indice di asimmetria

La misura di asimmetria di uso più comune è il cosidetto **indice di asimmetria standardizzato** definito come

$$\frac{1}{n\text{sqm(y)}^3}\sum_{i=1}^{n}(y_i-\overline{y})^3$$

dove, come al solito $y = (y_1, \ldots, y_n)$ indica i dati osservati, n il loro numero e sqm(y) lo scarto quadratico medio.

L'interpretazione è agevole. Nei casi in cui i dati si distribuiscano in maniera esattamente simmetrica intorno alla media i termini positivi e negativi nella sommatoria si compenseranno tra di loro e quindi l'indice sarà nullo. Viceversa, nei casi di asimmetria positiva i termini positivi predomineranno e quindi l'indice assumerà valori positivi. Opposta la situazione nei casi di asimmetria negativa.

Nel nostro esempio, l'indice è pari a -0.012 per l'insieme di dati A e a 1.300 per l'insieme di dati B.

L'indice, per costruzione, è invariante rispetto a trasformazioni lineari dei dati. Ovvero, otteniamo lo stesso risultato sia lavorando con i dati originali che con dati trasformati del tipo $z_i = a + by_i,\ \ i = 1, \ldots, n$ (esercizio).

Curtosi

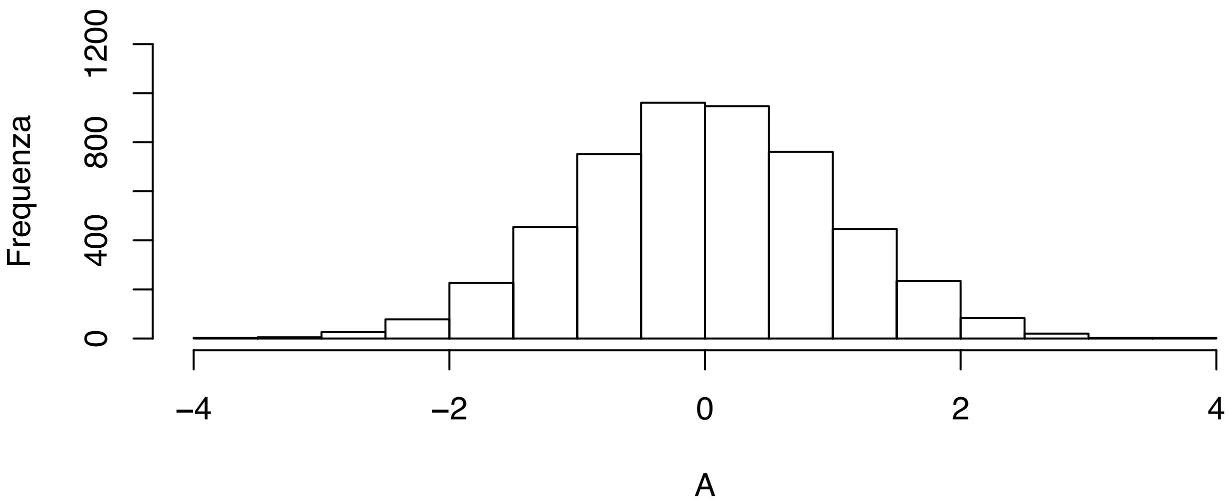
Anche i grafici nelle seguenti due pagine confrontano dati standardizzati (l'insieme A e un nuovo insieme C). In questo caso, ambedue le distribuzioni sono (almeno approssimativamente) simmetriche. Però, nonostante l'uguaglianza delle varianze, la prima distribuzione ha delle code più pesanti della seconda. Questa caratteristica (maggiore o minore peso delle code non dovuto ad una maggiore o minore variabilità) è spesso indicata con il termine **curtosi**.

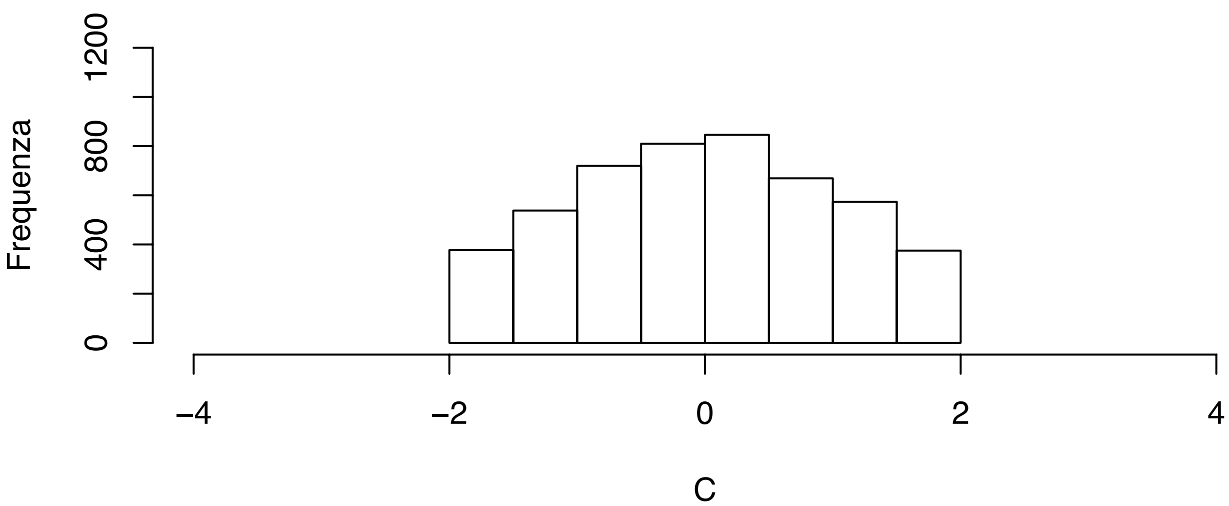
L'**indice di curtosi standardizzato** è definito come

$$\frac{1}{n\text{sqm}(\text{y})^4}\sum_{i=1}^{n}(y_i-\overline{y})^4.$$

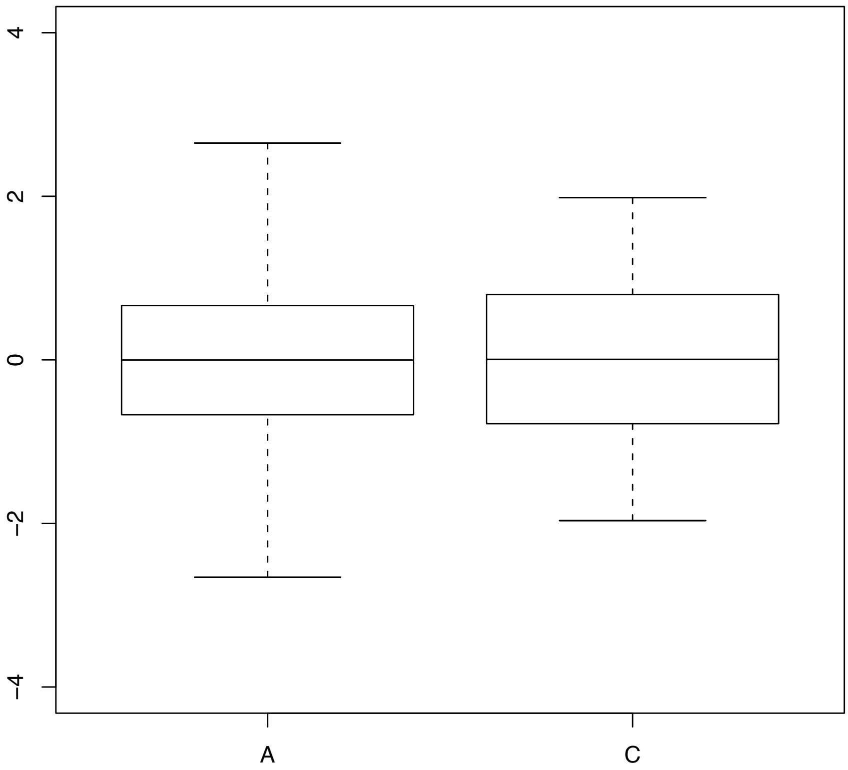
Può essere visto come un rapporto tra due indici di variabilità. L'indice a numeratore (la media delle potenze quarte degli scarti dalla media aritmetica) è scelto in maniera tale da essere più sensibile alla presenza di code pesanti dell'indice a denominatore (la potenza quarta dello scarto quadratico medio).

Due insiemi di dati standardizzati: istogramma





Due insiemi di dati standardizzati: boxplot



L'indice di asimmetria per i due insiemi di dati è pari a -0.012 e a 0.003 rispettivamente (le due distribuzioni sono sostanzialmente simmetriche).

L'indice di curtosi è invece pari a 2.956 e a 2.085, rispettivamente per l'insieme A e C, indicando che la distribuzione dell'insieme A ha code più pesanti di quella dell'insieme C.

Relazione fra variabili

Fino ad ora abbiamo considerato alcuni metodi per rappresentare, descrivere e sintetizzare le caratteristiche principali di una distribuzione.

Spesso però quello che interessa realmente è la relazione che lega due (o più) variabili e come al variare di una grandezza si modificano i valori dell'altra.

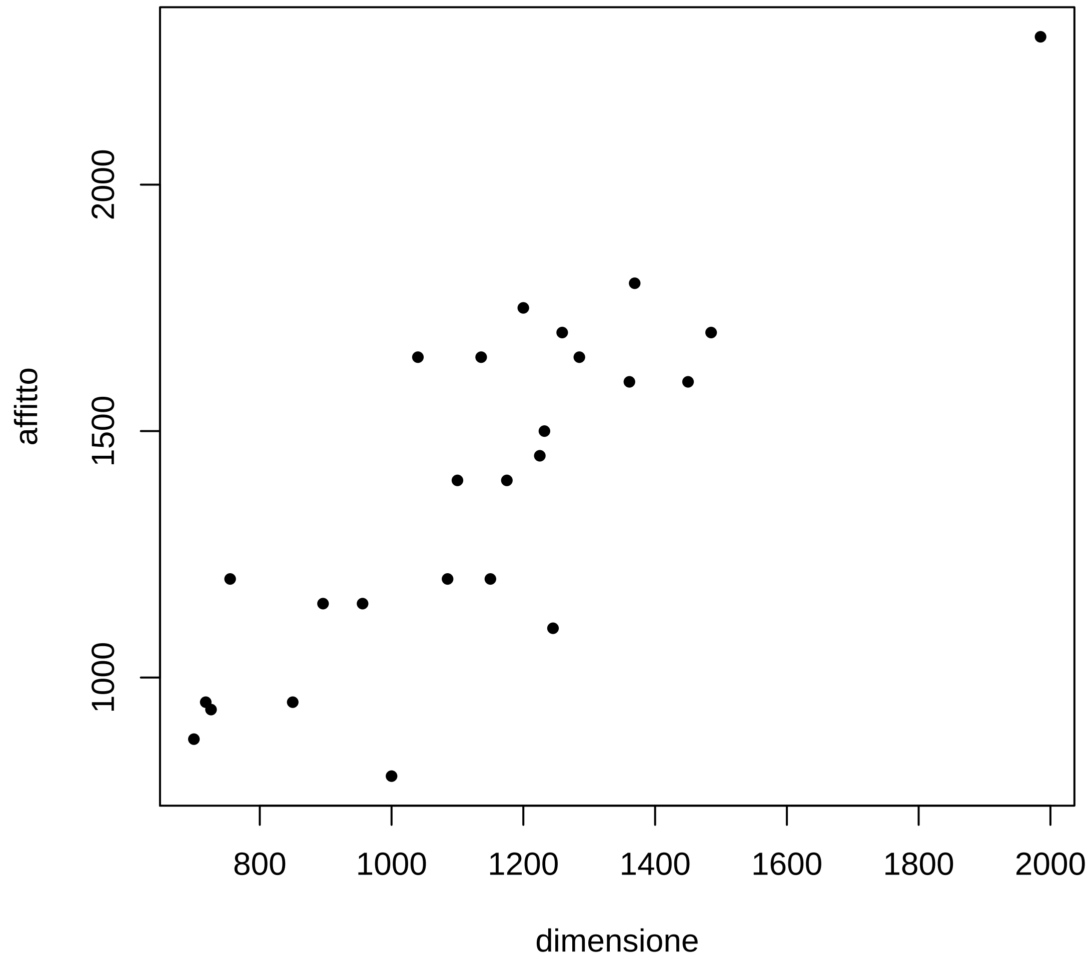
Anche nello studio congiunto di due distribuzioni, gli strumenti più adatti dipendono dal tipo di variabili che stiamo considerando.

Due variabili numeriche

Un agente immobiliare intende prevedere gli affitti mensili degli appartamenti sulla base della loro dimensione. Per questo conduce un'indagine e reperisce i dati su 25 appartamenti in una zona residenziale. La seguente tabella mostra i dati ottenuti per i 25 appartamenti. L'affitto è l'affitto mensile in dollari e la dimensione è espressa in piedi al quadrato.

	affitto	dimensione
1	950	850
2	1600	1450
3	1200	1085
4	1500	1232
5	950	718
6	1700	1485
7	1650	1136
8	935	726
9	875	700
10	1150	956
11	1400	1100
12	1650	1285
13	2300	1985
14	1800	1369
15	1400	1175
16	1450	1225
17	1100	1245
18	1700	1259
19	1200	1150
20	1150	896
21	1600	1361
22	1650	1040
23	1200	755
24	800	1000
25	1750	1200

Diagramma di dispersione



Abbiamo semplicemente disegnato i punti osservati sul piano. E' evidente una forte relazione, certamente crescente come ci si poteva attendere.

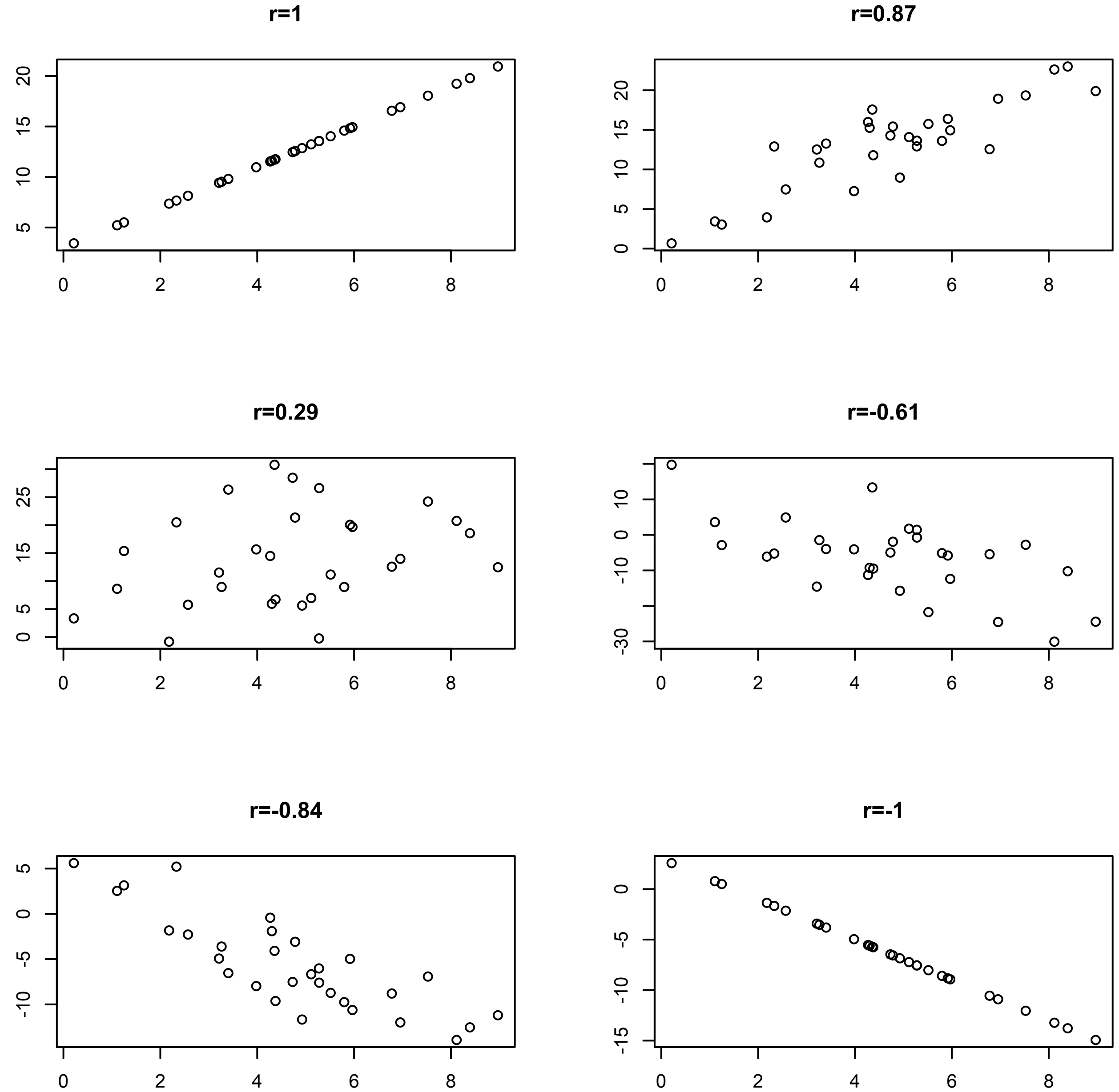
Covarianza e correlazione

La **covarianza** e la **correlazione** misurano l'intensità del legame lineare fra due variabili. Si definiscono così:

$$\begin{aligned} \mathrm{cov}(x,y) &= \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y} \end{aligned}$$

$$\begin{aligned} \mathrm{cor}(x,y) &= \frac{\mathrm{cov}(x,y)}{\mathrm{sqm}(x)\cdot \mathrm{sqm}(y)} \\ &= \mathrm{cov}\left(\frac{x-\bar{x}}{\mathrm{sqm}(x)}, \frac{y-\bar{y}}{\mathrm{sqm}(y)}\right). \end{aligned}$$

La correlazione



Dati sugli affitti delle case

Calcoliamo covarianza e correlazione per i nostri dati, ponendo x =dimensione e y =affitto:

$$\begin{aligned}
\mathrm{cov}(x,y) &= \frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y} \\
&= \frac{1}{25} \cdot 41480210 - 1135.32 \cdot 1386.4 \\
&= 85200.75
\end{aligned}$$

$$\begin{aligned}
\mathrm{cor}(x,y) &= \frac{\mathrm{cov}(x,y)}{\mathrm{sqm}(x) \cdot \mathrm{sqm}(y)} \\
&= \frac{85200.75}{282.8249 \cdot 354.3854} \\
&= 0.85
\end{aligned}$$

Tabelle di contingenza: il Titanic

Dopo il disastro, una commissione d'inchiesta del *British Board of Trade* ha compilato una lista di tutti i 1316 passeggeri con alcune informazioni aggiuntive riguardanti: se è stato salvato (SI, NO), la classe (I, II, III) in cui viaggiavano, il sesso, l'età,. . . .

Ci limitiamo a considerare le informazioni sull'esito e la classe. Quindi dal nostro punto di vista i dati sono costituiti da una lunga lista del tipo

Passeggero	Classe	Salvato
nome 1	II	SI
nome 2	III	NO
nome 3	I	NO
⋮	⋮	⋮
nome 1316	III	SI

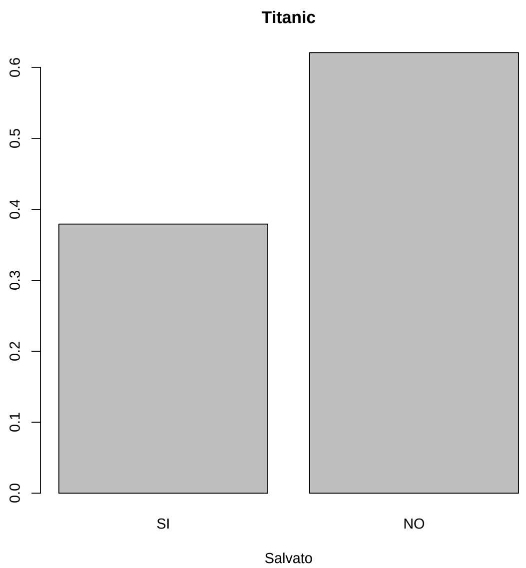
Una variabile alla volta

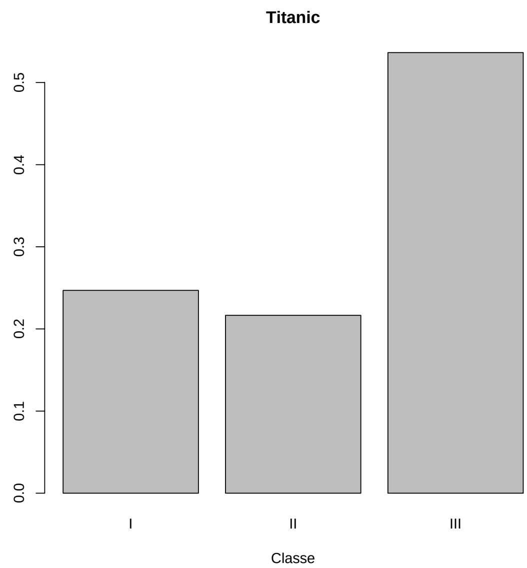
La variabile Salvato ha la seguente distribuzione di frequenze

Salvato	Freq. assolute	Freq. relative
SI	499	0,379
NO	817	0,621
	1316	1,000

La variabile Classe ha invece la seguente distribuzione

Classe	Freq. assolute	Freq. relative
I	325	0,247
II	285	0,216
III	706	0,537
	1316	1,00





Le due variabili assieme: frequenze congiunte

La prima sintesi che possiamo operare consiste nel costruire una tabella del tipo

	Classe			
Salvato	I	II	III	totale
SI	203	118	178	499
NO	122	167	528	817
totale	325	285	706	1316

dove consideriamo tutti i possibili incroci di modalità delle due variabili ($2 \times 3 = 6$).

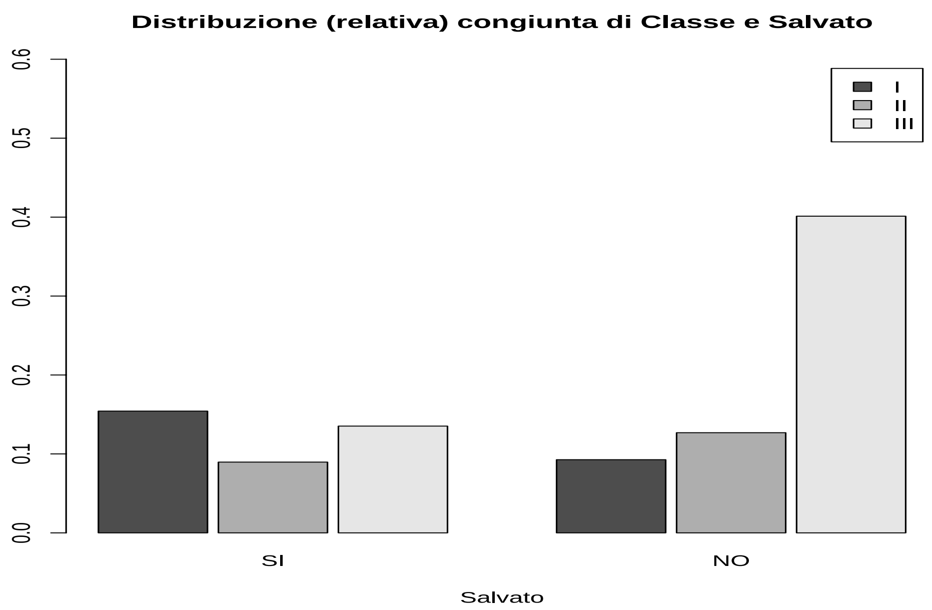
Possiamo anche considerare le frequenze relative, ottenute semplicemente dividendo le frequenze assolute per il numero totale $n = 1316$ di unità

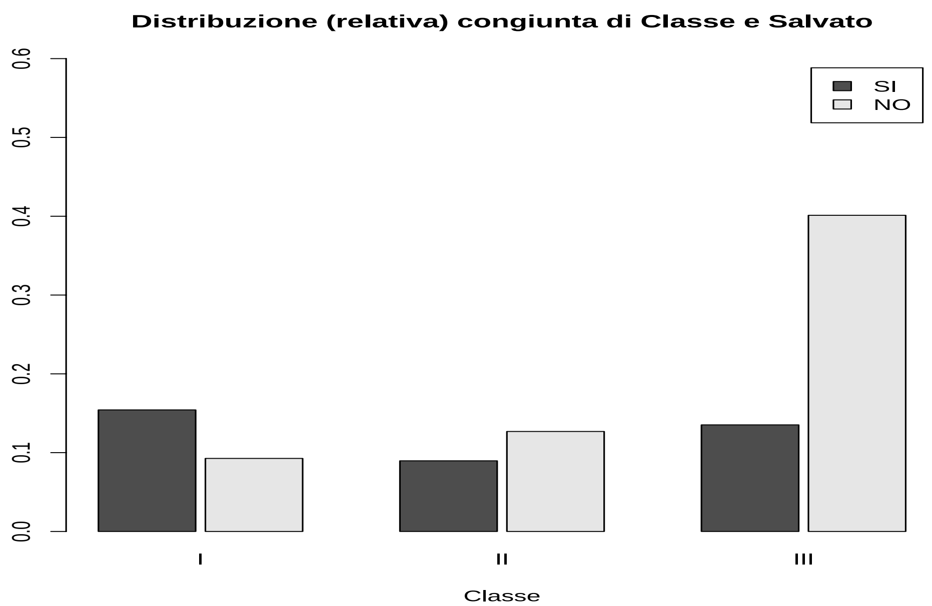
	Classe			
Salvato	I	II	III	totale
SI	0,154	0,090	0,135	0,38
NO	0,093	0,127	0,401	0,62
totale	0,247	0,217	0,536	1,000

Frequenze congiunte: rappresentazione grafica

Possiamo rappresentare le frequenze (sia assolute che relative) della tabella attraverso un appropriato diagramma a barre.

La stessa informazione può essere rappresentata in due modi diversi ("per riga" o "per colonna"):



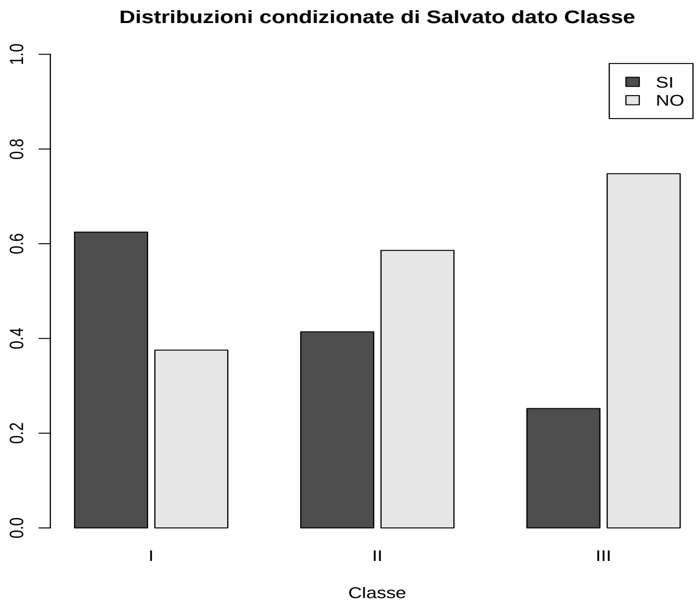


Distribuzioni condizionate di Salvato dato Classe

Ci sono tre distribuzioni condizionate di Salvato dato Classe (le tre colonne), una per ogni modalità di Classe (I, II, II).
Le distribuzioni condizionate relative si ottengono dividendo ogni colonna per il totale di colonna

	Classe		
Salvato	I	II	III
SI	203	118	178
NO	122	167	528
totale	325	285	706

	Classe		
Salvato	I	II	III
SI	0,62	0,41	0,25
NO	0,38	0,59	0,75
totale	1,00	1,00	1,00

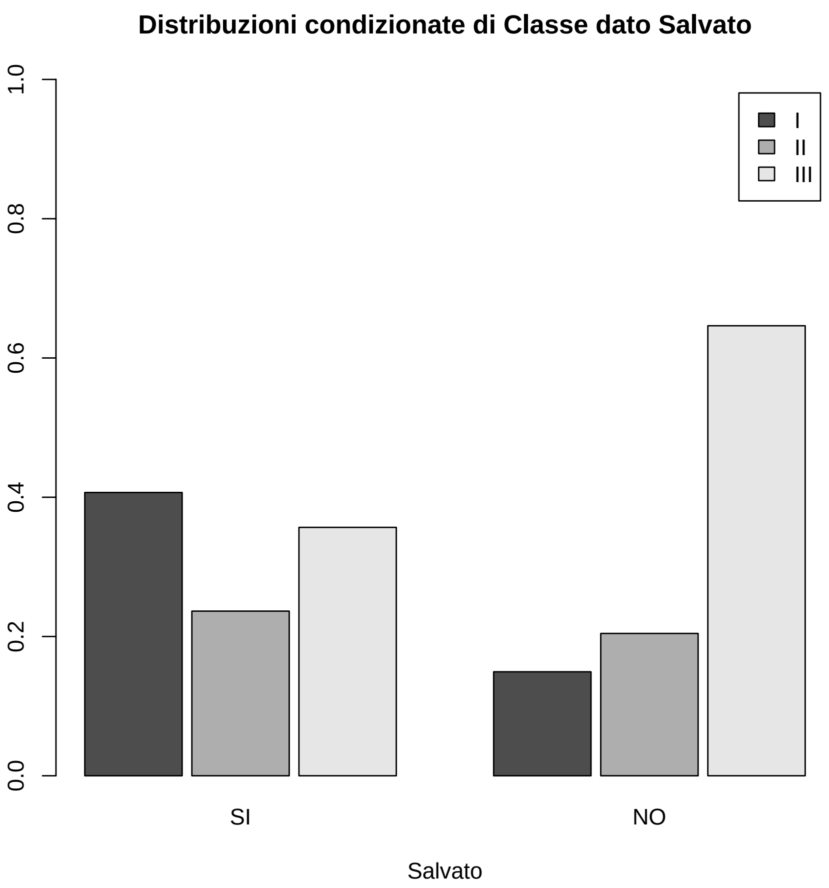


Distribuzioni condizionate di Classe dato Salvato

Ci sono due distribuzioni condizionate di Classe dato Salvato (le due righe), una per ogni modalità di Salvato (SI, NO).
Le distribuzioni condizionate relative si ottengono dividendo ogni riga per il totale di riga

	Classe			
Salvato	I	II	III	totale
SI	203	118	178	499
NO	122	167	528	817

	Classe			
Salvato	I	II	III	totale
SI	0,41	0,24	0,36	1,00
NO	0,15	0,20	0,65	1,00



Frequenze attese

La tabella delle frequenze attese è quella che si osserverebbe se fra le due variabili non ci fosse nessun tipo di dipendenza:

	classe			
salvato	I	II	III	totale
SI	123,2	108,1	267,7	499
NO	201,8	176,9	438,3	817
totale	325	285	706	1316

Il confronto con le frequenze osservate è particolarmente istruttivo.

	Classe			
Salvato	I	II	III	totale
SI	203	118	178	499
NO	122	167	528	817
totale	325	285	706	1316

Ad esempio, ci indica che, senza la preferenza accordata ai passeggeri di I classe, si sarebbero salvati un centinaio di passeggeri di III classe in più.

Quindi, sembra esserci evidenza contro l'ipotesi di indipendenza tra le due variabili.

L'indice X^2 di Pearson

E' una *misura della distanza* fra le frequenze osservate e le frequenze attese.

$$
\begin{aligned}
X^2 &= \frac{(203-123,2)^2}{123,2}+\frac{(118-108,1)^2}{108,1}+\ldots \\
&\quad \cdots+\frac{(528-438,3)^2}{438,3} \\
&= 133,05
\end{aligned}
$$

$$\tilde{X}^2 = \frac{133,05}{1316\cdot\min(1,2)}=0,1011\,.$$

Purtroppo, per sapere se il valore che abbiamo ottenuto è grande o piccolo, abbiamo bisogno del calcolo delle probabilità...