Università Ca' Foscari di Venezia
Federica Giummolè

Informazioni generali

Probabilità e Statistica
A.A. 2013/2014

Informazioni generali

• Docente del corso: Federica Giummolè
email: `giummole@unive.it`
web: `http://www.dst.unive.it/~giummole`

• Ricevimento studenti: Martedì 12:00–13:00, via Torino stanza ospiti
`http://www.dst.unive.it/~giummole/avvisi.html`

• Lezioni: Martedì e Giovedì 8:45–10:15, via Torino aula 2

• Tutorato (Marco Fiorucci): Martedì 12:15–13:45, via Torino aula 2 o lab 5

• Moodle: `http://moodle.unive.it/` per materiale didattico e test di autovalutazione

- Testi di riferimento:

Ross, S.M. (2007). *Calcolo delle probabilità*, seconda edizione, Apogeo.
Ross, S.M. (2009). *Probabilità e statistica per l'ingegneria e le scienze*, seconda edizione, Apogeo.

- L'esame consiste in una prova scritta con 10 domande a risposta multipla con sbarramento e alcuni esercizi teorici e pratici, anche sull'uso di R.

Laboratorio con R

R è un ambiente di sviluppo specifico per l'analisi statistica dei dati che utilizza un linguaggio di programmazione derivato e in larga parte compatibile con S.

- R è open-source e può essere scaricato gratuitamente dal sito `http://cran.r-project.org/`

- R funziona sotto UNIX, Windows e Mac

- R ha un *help* approfondito e dettagliato

- R ha eccellenti capacità grafiche

- R è un linguaggio di programmazione con molte funzioni predefinite e la possibilità di costruirne di nuove

- R è mantenuto e aggiornato da una squadra internazionale di esperti. Tutti possono contribuire con *packages* sempre nuovi

Università Ca' Foscari di Venezia
Federica Giummolè

Introduzione alla statistica

Probabilità e Statistica
A.A. 2013/2014

La statistica nella società dell'informazione

- Tutti dicono che viviamo nella società dell'informazione. Ma molti si lamentano che le informazioni sono troppe. E' facile raccoglierle, memorizzarle, distribuirle. E' difficile verificarle ed interpretarle.

- La statistica è la *tecnologia* necessaria per risolvere queste difficoltà.

- Uno statistico sa combinare informazioni di tipo differente, valutarne l'affidabilità, sintetizzare e presentare molti dati in maniera tale da evidenziare la storia che raccontano, costruire modelli (=visioni stilizzate di una parte di mondo) che facilitano l'interpretazione, e, ad esempio, permettono di calcolare previsioni o di formulare ipotesi di decisione.

Un po' di terminologia

- Un insieme (di individui o animali o oggetti o aziende o. . .) costituisce la parte del mondo che interessa, quella su cui dobbiamo produrre nuove conoscenze, quella che è coinvolta nelle decisioni da prendere. Questo insieme viene chiamato convenzionalmente la **popolazione di riferimento**. Gli elementi della popolazione sono chiamati genericamente **unità statistiche**.

- Alcune caratteristiche di tutte o di una parte delle unità statistiche vengono rilevate/misurate. Il risultato di questo rilevare/misurare costituisce quello che chiamiamo i **dati**. Le unità statistiche sono disomogenee rispetto ai fenomeni rilevati.

- L'obbiettivo è quello di trasformare i dati in nuove conoscenze o ipotesi di decisione. Ovvero, di trasformare i dati in affermazioni sul mondo (sulla popolazione di riferimento).

Un po' di terminologia

- Le caratteristiche rilevate sulle unità statistiche vengono chiamate **variabili**.

- I valori distinti assunti da una variabile sono chiamate le **modalità** della variabile stessa.

- Se le variabili di interesse non sono rilevate su tutte le unità statistiche, il sottoinsieme della popolazione oggetto della rilevazione è chiamato il **campione**.

Tipi di variabili

In statistica si parla di variabili:

- **qualitative** o **categoriali** quando le modalità utilizzate per descrivere il fenomeno analizzato prendono la forma di aggettivi o di altre espressioni verbali. Le variabili qualitative possono essere

  - **sconnesse** se non esiste nessun ordinamento naturale tra le modalità; esempi di variabili sconnesse sono: (i) il sesso, (ii) il tipo di servizio offerto da un albergo;

  - **ordinali** nel caso in cui un ordinamento naturale esiste; esempi di variabili qualitative ordinali sono: (i) il titolo di studio, (ii) il parere di un intervistato (ad es. classificato come "mediocre", "discreto", "buono").

  Quando le modalità sono solamente due (esempi (i) maschio vs. femmina, (ii) vivo vs. morto; (iii) buono vs. difettoso) si parla di variabili **dicotomiche** o **binarie**.

- **numeriche** quando le modalità sono espresse da numeri. Dal punto di vista dei modelli e delle tecniche utilizzate le variabili numeriche si suddividono a loro volta in

  - **discrete** o **intere** quando le modalità sono esprimibili da numeri interi; esempi sono: (i) il numero di clienti, (ii) il numero di pezzi prodotti;

  - **continue** o **reali** quando le modalità sono esprimibili da numeri reali; esempi sono: (i) il tempo d'attesa ad uno sportello, (ii) il peso di un manufatto.

Piccolo esempio (per fissare la terminologia)

Vogliamo avere un'idea sul numero di clienti e sul volume di vendite dei negozi di una città per tre categorie merceologiche ritenute simili. La popolazione di riferimento è l'insieme di tutti i negozi secondo le tre categorie merceologiche. Le unità statistiche sono i negozi. I dati si presentano in questa forma:

negozio	clienti	vendite	categoria
1	907	11.2	A
⋮	⋮	⋮	⋮
10	420	6.12	B
11	679	7.63	B
⋮	⋮	⋮	⋮
19	1010	11.77	C
20	621	7.41	A

Le variabili considerate nello studio sono tre:

clienti le cui *modalità* sono numeriche e discrete;

vendite (in migliaia di euro) le cui *modalità* sono numeriche e (con approssimazione) continue.

categoria le cui *modalità* sono sconnesse (A, B e C.)

Il modo in cui sono raccolti i dati può condizionare il loro tipo

Si consideri una macchina che deve forare delle lastre di metallo. Il diametro nominale dei fori è di 1mm con una tolleranza di 0,06mm. Ovvero un foro è *ben fatto* se il suo diametro è compreso tra 0,94mm e 1,06mm.

Allora, dati sulla *qualità* della produzione della macchina, potrebbero essere disponibili nella forma

1. "buono" vs. "difettoso" (dati dicotomici);

2. "troppo piccolo", "buono", "troppo grande" (dati qualitativi ordinali);

3. lunghezza del diametro (dati numerici continui).

Si osservi che le differenze non sono semplicemente dovute a come i dati vengono registrati ma possono anche essere dovute a come *i diametri vengono effettivamente misurati*. Ad esempio, raccogliere dati sui diametri nella forma (2) è più rapido e richiede strumenti meno costosi (bastano due bastoncini metallici di diametro rispettivamente uguale ai due estremi dell'intervallo di tolleranza) di quanto richiesto dalla forma (3).

Dati sperimentali vs dati osservazionali

Nell'analizzare dei dati è bene poi tenere presente il tipo di studio in cui sono stati rilevati. In particolare, è importante la distinzione tra:

- **studi sperimentali** ovvero situazioni in cui i dati sono stati raccolti in situazioni replicabili e controllate (esempio classico sono gli esperimenti di laboratorio);

- **studi osservazionali** ovvero situazioni in cui il ricercatore semplicemente rileva dei dati già esistenti (esempio: il numero di presenze alberghiere in una stagione, il prezzo di un'azione,...).

Il problema principale degli studi osservazionali è che non controllando i fattori che possono influenzare il fenomeno sotto indagine risulta difficile essere certi di averli individuati appropriatamente.

Metodi di raccolta dei dati

1. Esperimenti in laboratorio
2. Interviste telefoniche
3. Questionari inviati per posta
4. Social network
5. Carte fedeltà
6. ...

Il **campionamento** e il **disegno degli esperimenti** si occupano delle problematiche connesse con la raccolta dei dati.

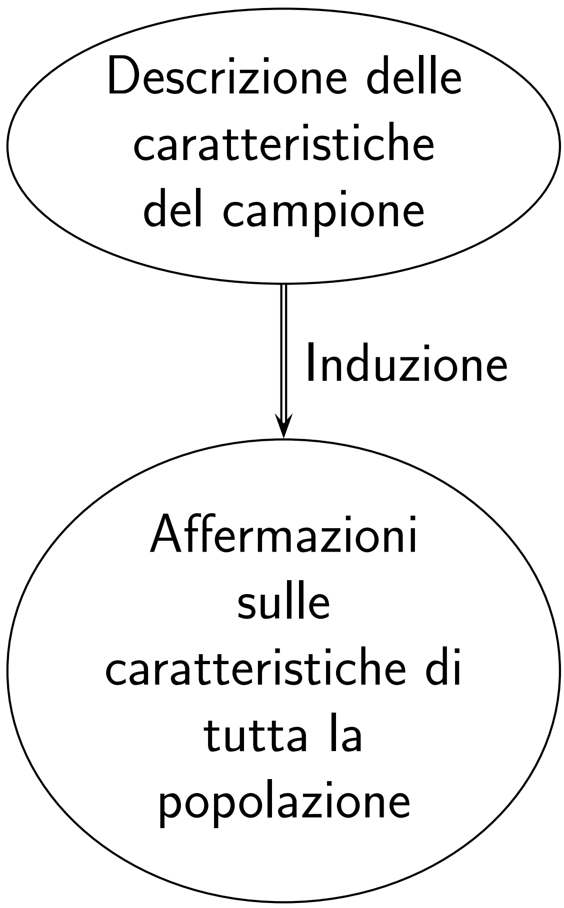
Statistica Descrittiva e Inferenza Statistica

Descrittiva: metodi per rappresentare, sintetizzare ed evidenziare le caratteristiche più significative di un insieme di dati. Usualmente si dispone di dati su tutta la popolazione di riferimento.

Inferenza: i dati disponibili sono stati rilevati solamente su una parte delle unità statistiche (il campione, da cui *indagini campionarie*). Si vogliono utilizzare le informazioni del campione per fare delle affermazioni sulle caratteristiche generali di tutta la popolazione.

Statistica Descrittiva e Inferenza Statistica

Statistica Descrittiva ed **Inferenza Statistica** nelle applicazioni non sono facilmente separabili. Infatti i problemi di *inferenza* vengono normalmente affrontati in accordo allo schema



La statistica descrittiva viene dunque utilizzata per un'analisi preliminare delle caratteristiche del campione.

Calcolo delle probabilità

Perché l'inferenza porti a risultati sensati, bisogna che sia noto il legame fra popolazione e campione. In particolare, il campione deve essere scelto in modo che rappresenti, in piccolo, la popolazione.

La relazione fra campione e popolazione si descrive attraverso il **calcolo delle probabilità**.

E' il calcolo delle probabilità che fornisce gli strumenti per l'inferenza e che permette di quantificare gli errori che commettiamo nel passaggio dal particolare (campione) al generale (popolazione).

Università Ca' Foscari di Venezia
Federica Giummolè

Statistica descrittiva

Probabilità e Statistica
A.A. 2013/2014

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

$$\begin{array}{c}\text{frequenze}\\ \text{relative}\end{array} = \frac{\begin{array}{c}\text{frequenze}\\ \text{assolute}\end{array}}{\begin{array}{c}\text{numero totale}\\ \text{di osservazioni}\end{array}}$$

Tabelle di frequenza: notazioni

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

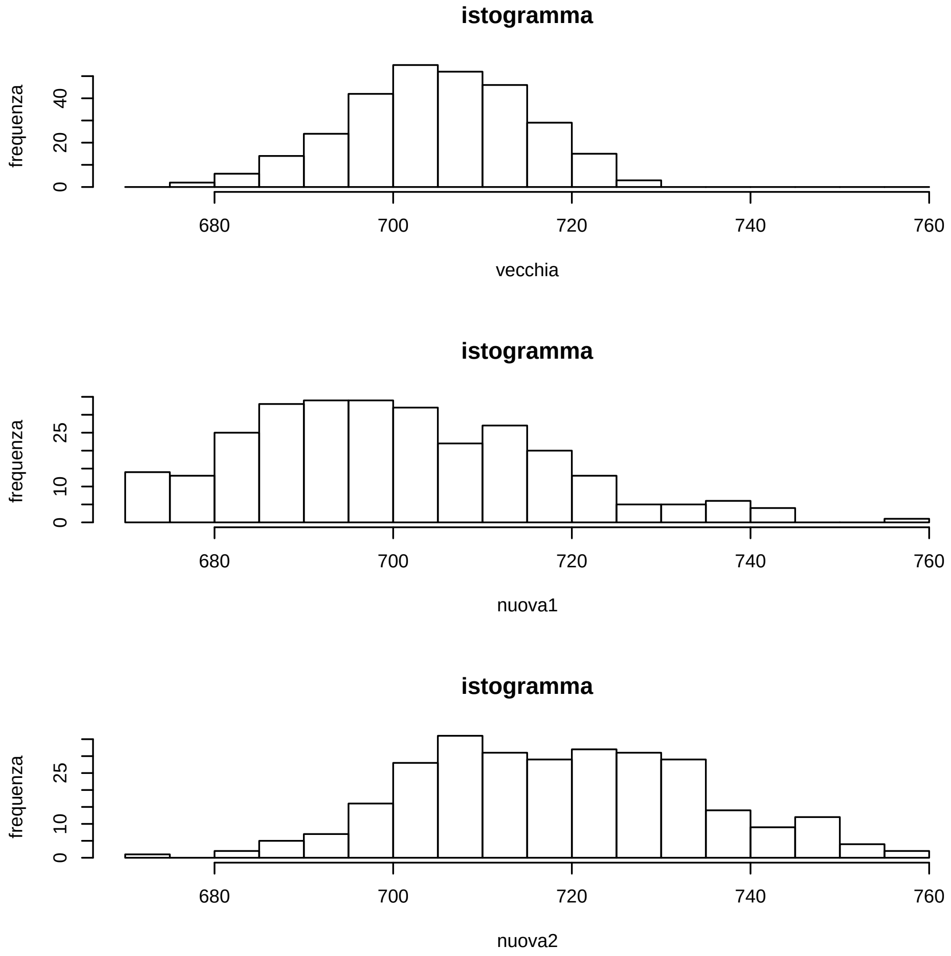
f_i : frequenza assoluta, numero di unità statistiche che possiedono la modalità/classe y_i

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:

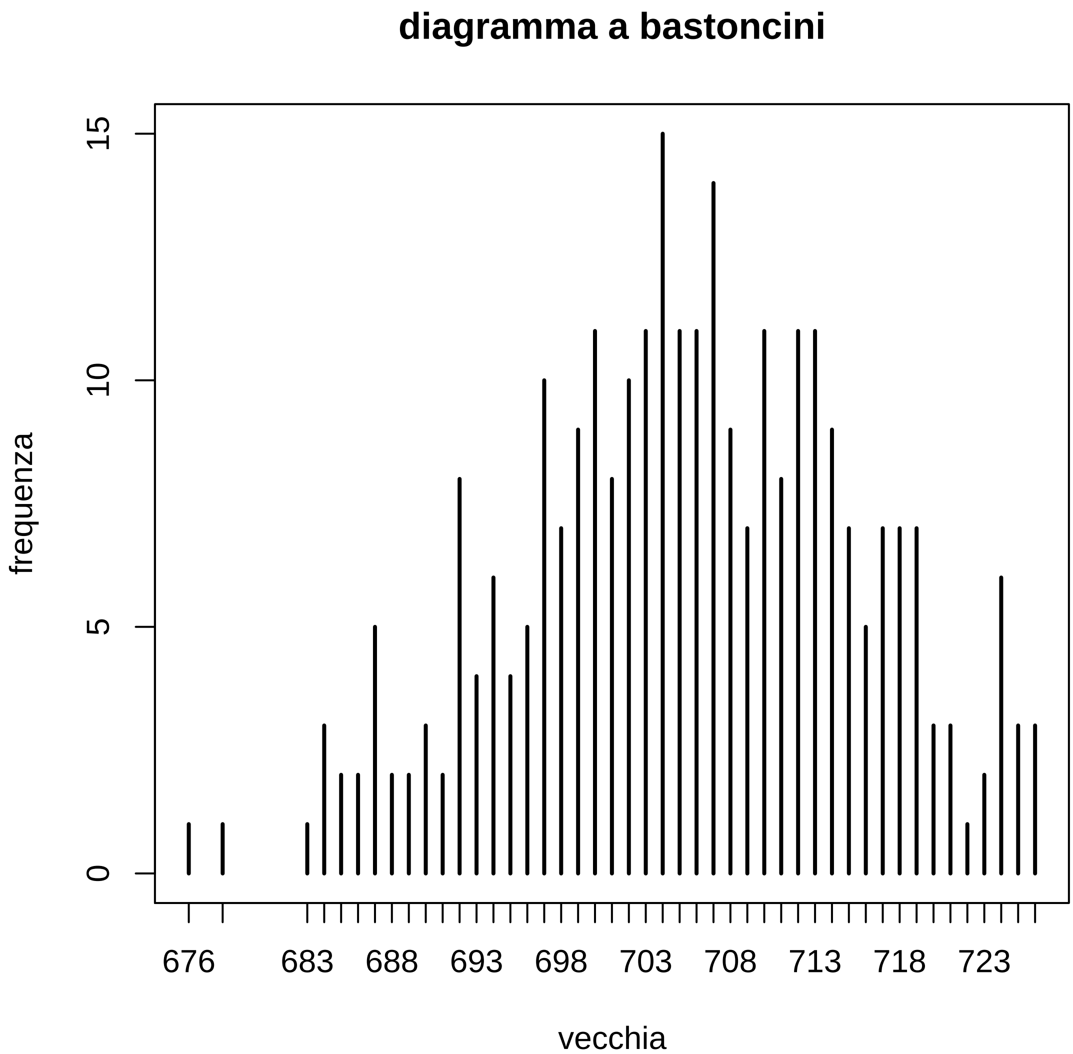
1. la base dei rettangoli pari agli intervalli riportati nella 1° colonna delle tabelle precedenti;

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'istogramma!) 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:

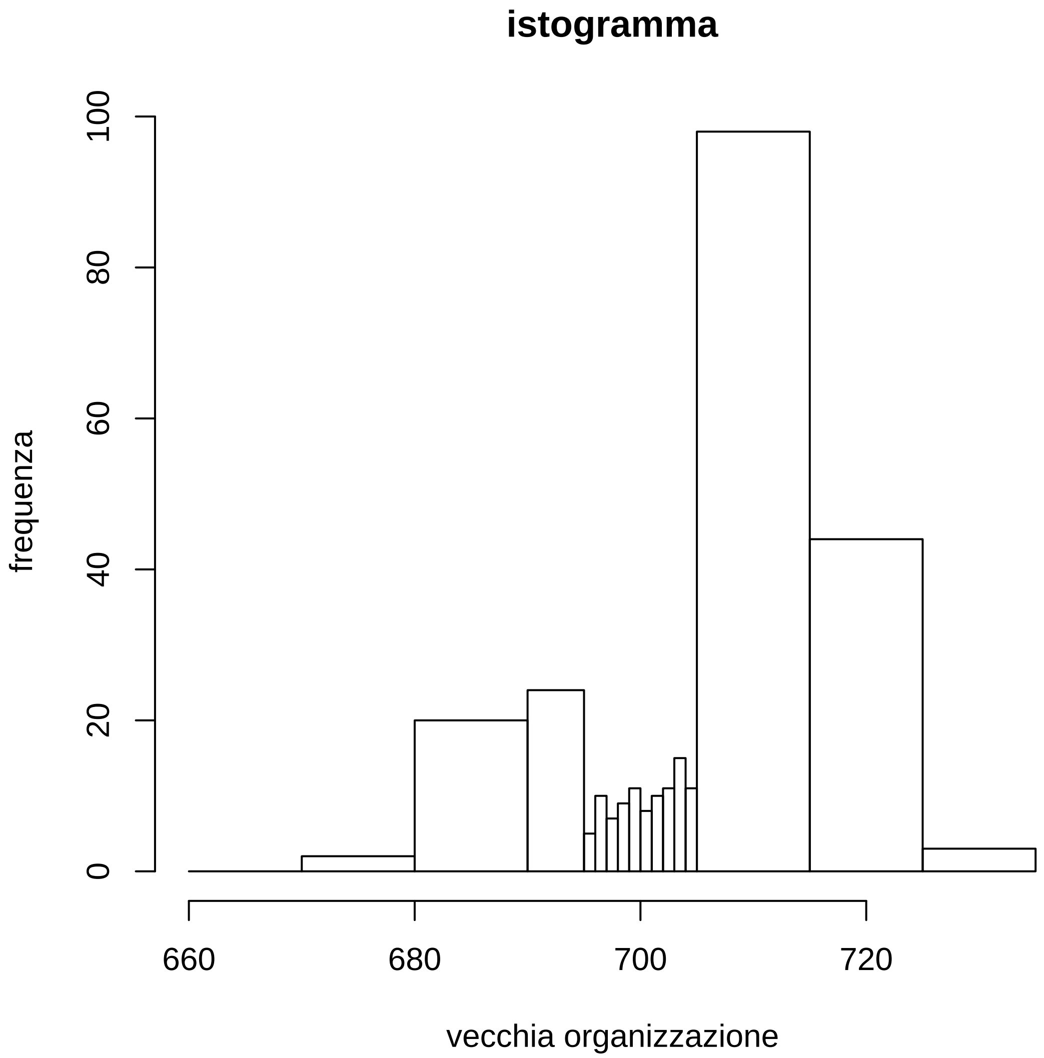
$$\begin{array}{c}\text{densità}\\ \text{di un intervallo}\end{array} = \frac{\text{frequenza dell'intervallo}}{\text{lunghezza dell'intervallo}}.$$

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.

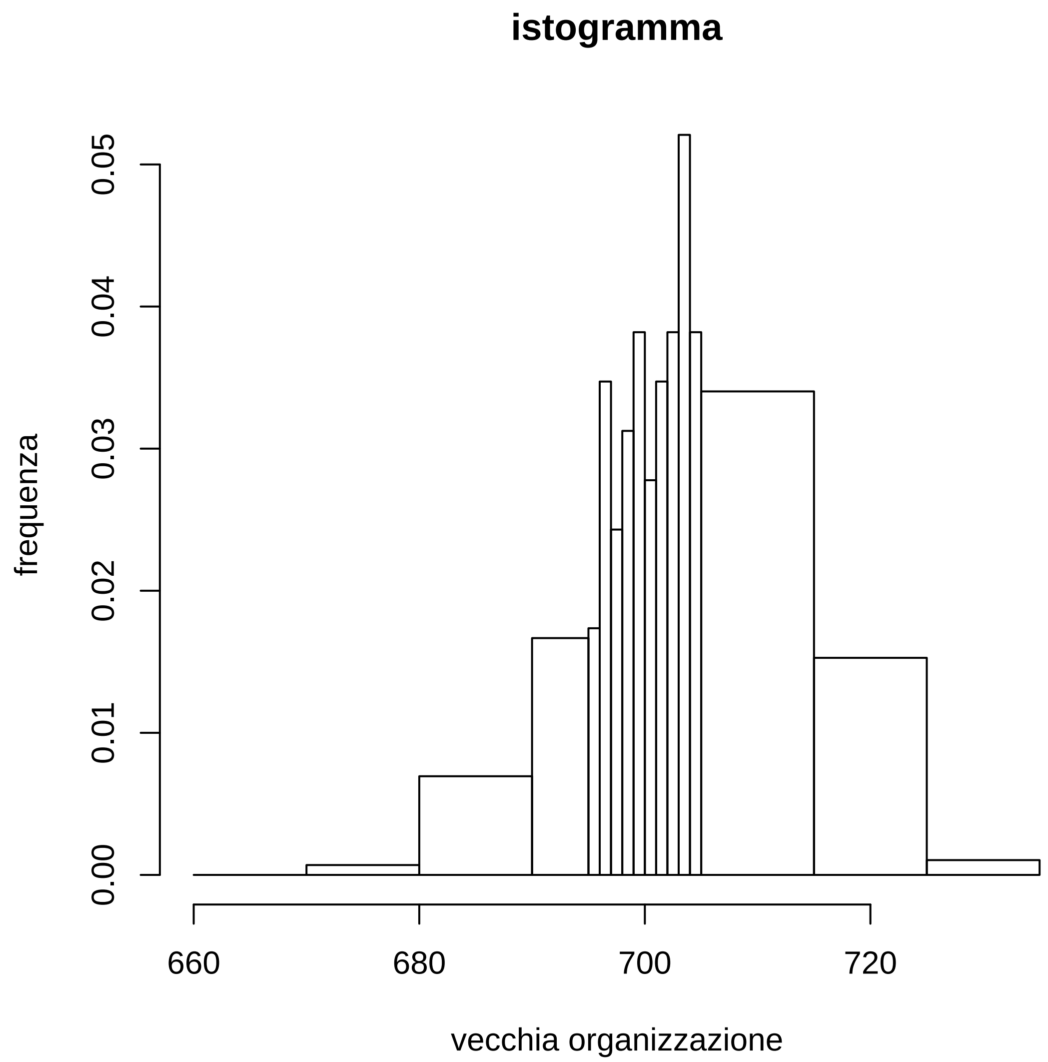


Sembra esserci un buco al centro, esattamente dove le osservazioni sono più *addensate*.

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+⋯+0	287/288=0.997
760	1	288=13+12+⋯+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:

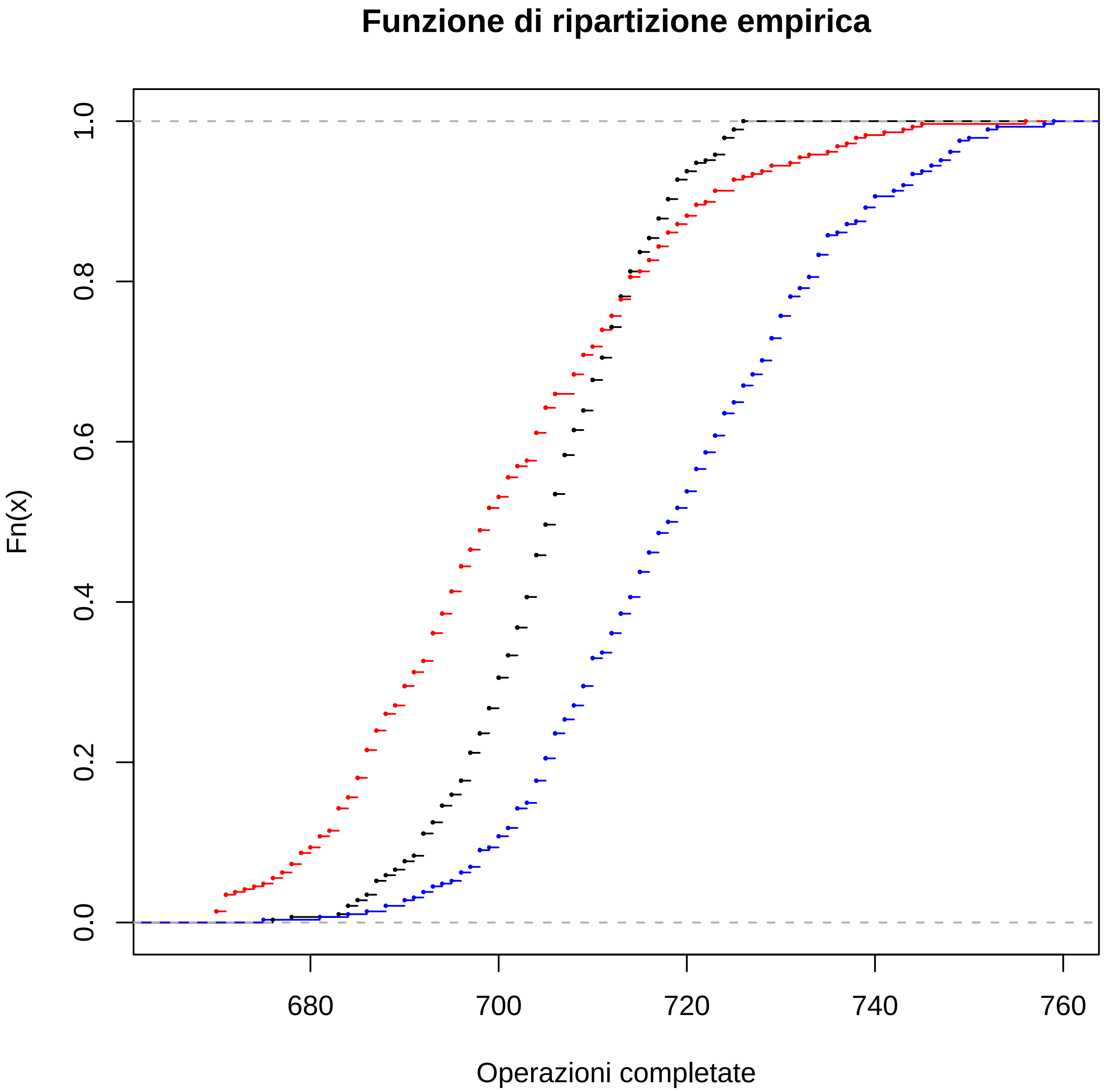
1. $0 \leq \widehat{F}(y) \leq 1$;

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.

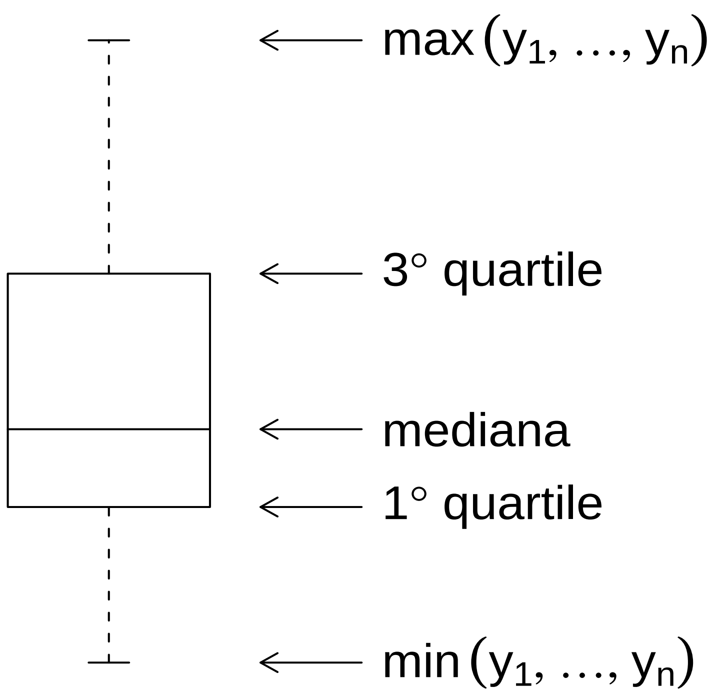

Funzione di ripartizione empirica
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
680
700
720
740
760
Operazioni completate

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).

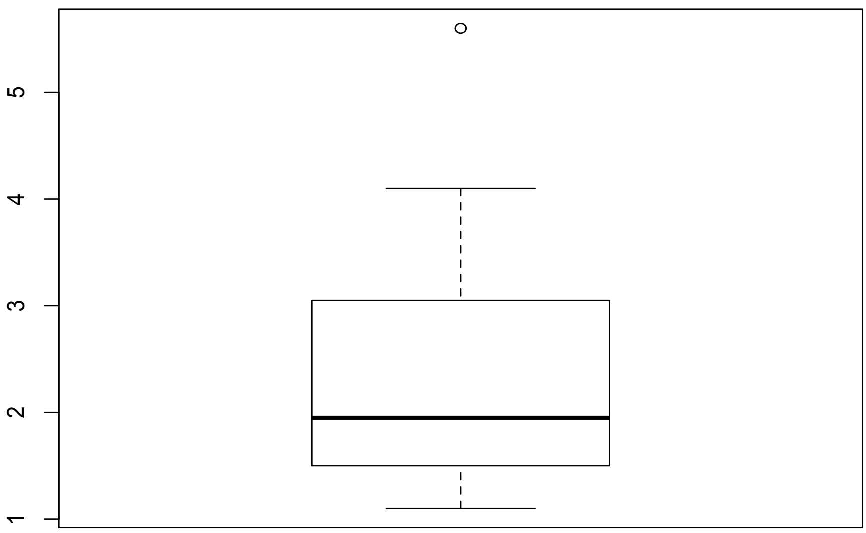
Esempio di costruzione di un boxplot

Dati (già ordinati):

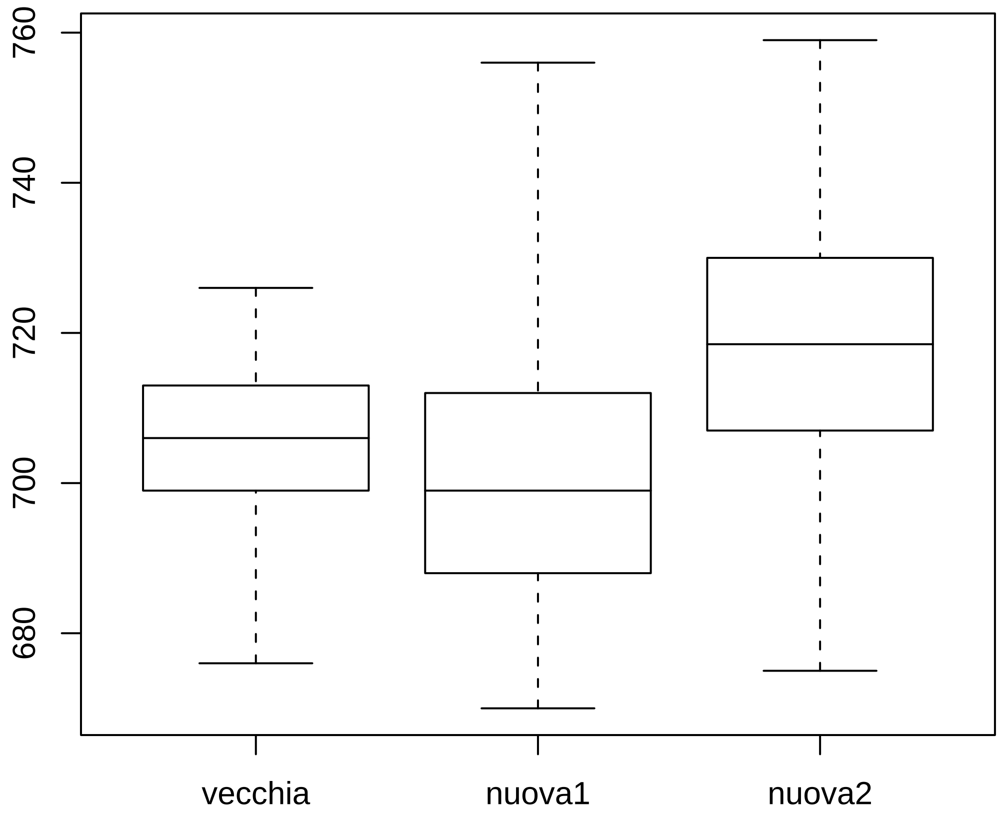
$$1.11.31.41.61.81.92.02.52.93.24.15.6$$

Perciò $y_{0.25} = 1.5$, $y_{0.5} = 1.95$ e $y_{0.75} = 3.05$. Quindi $1.5\times$(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



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
$\vdots$	$\vdots$	$\vdots$

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

Tabelle di frequenza

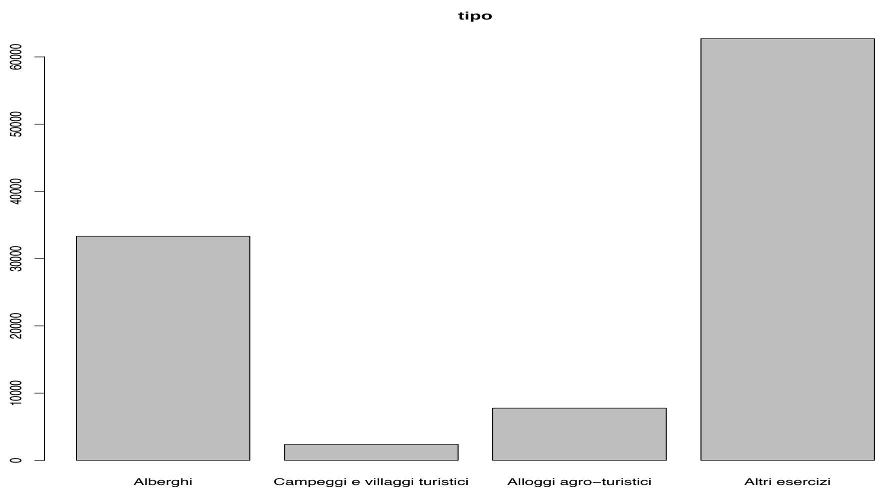
La variabile *tipo* ha la seguente distribuzione di frequenze

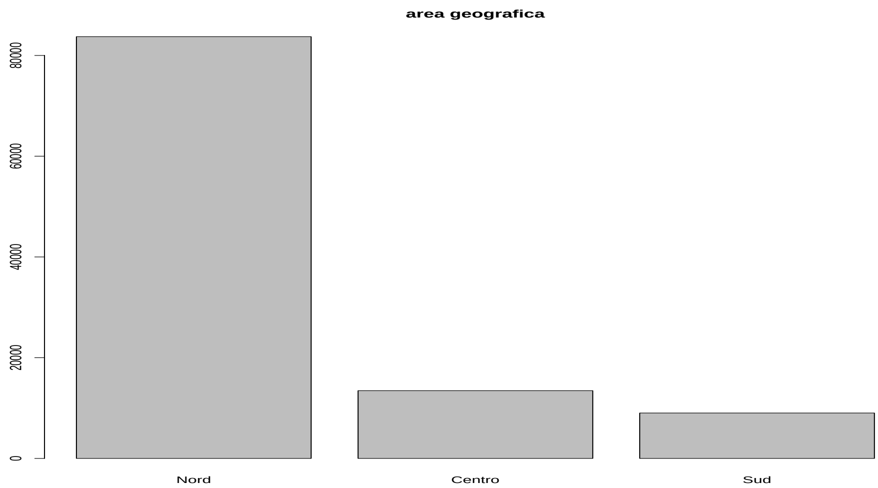
tipo	freq.	freq. rel.
Alberghi	33.338	0,314
Campeggi e villaggi turistici	2.371	0,022
Alloggi agro-turistici	7.769	0,073
Altri esercizi	62.727	0,591
TOTALE	106.205	1,00

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

Diagramma a barre: frequenze assolute





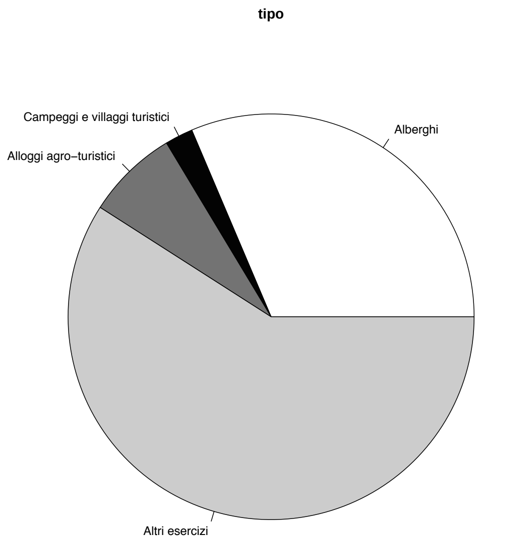
La rappresentazione grafica più utilizzata è il diagramma a barre, in cui ogni modalità è rappresentata da una barra di altezza pari alla frequenza (assoluta o relativa) della modalità. Si osservi che i rettangoli, contrariamente al caso di un istogramma, sono disegnati *staccati*.

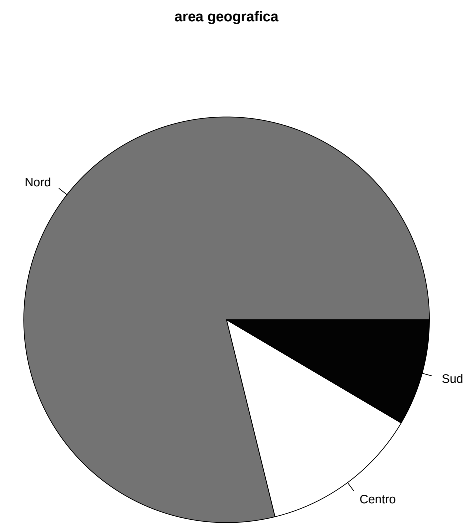
Notiamo che, se la variabile non è ordinale, l'ordine delle modalità nell'asse delle ascisse del grafico è arbitrario.

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:

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



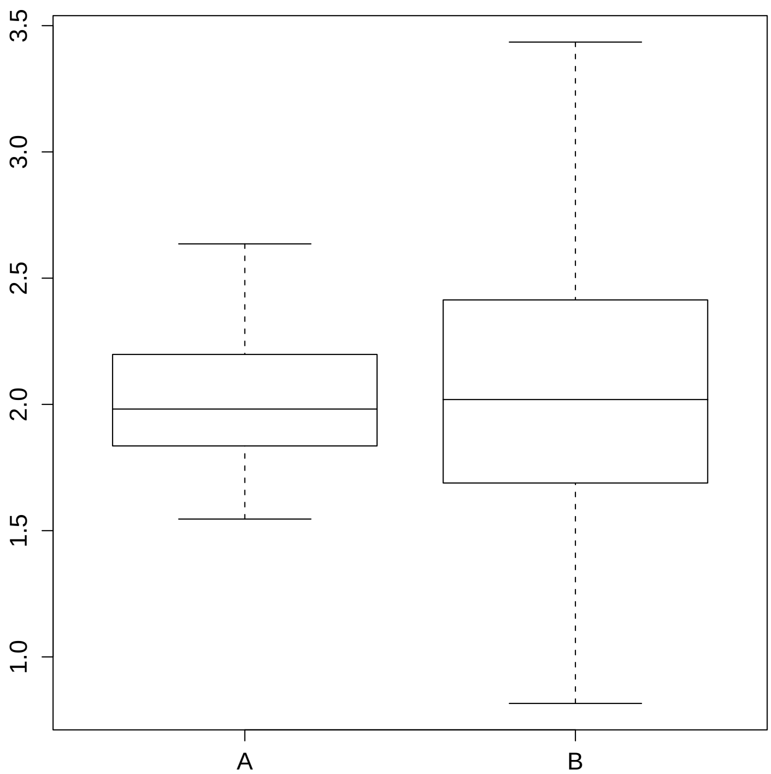


La variabilità

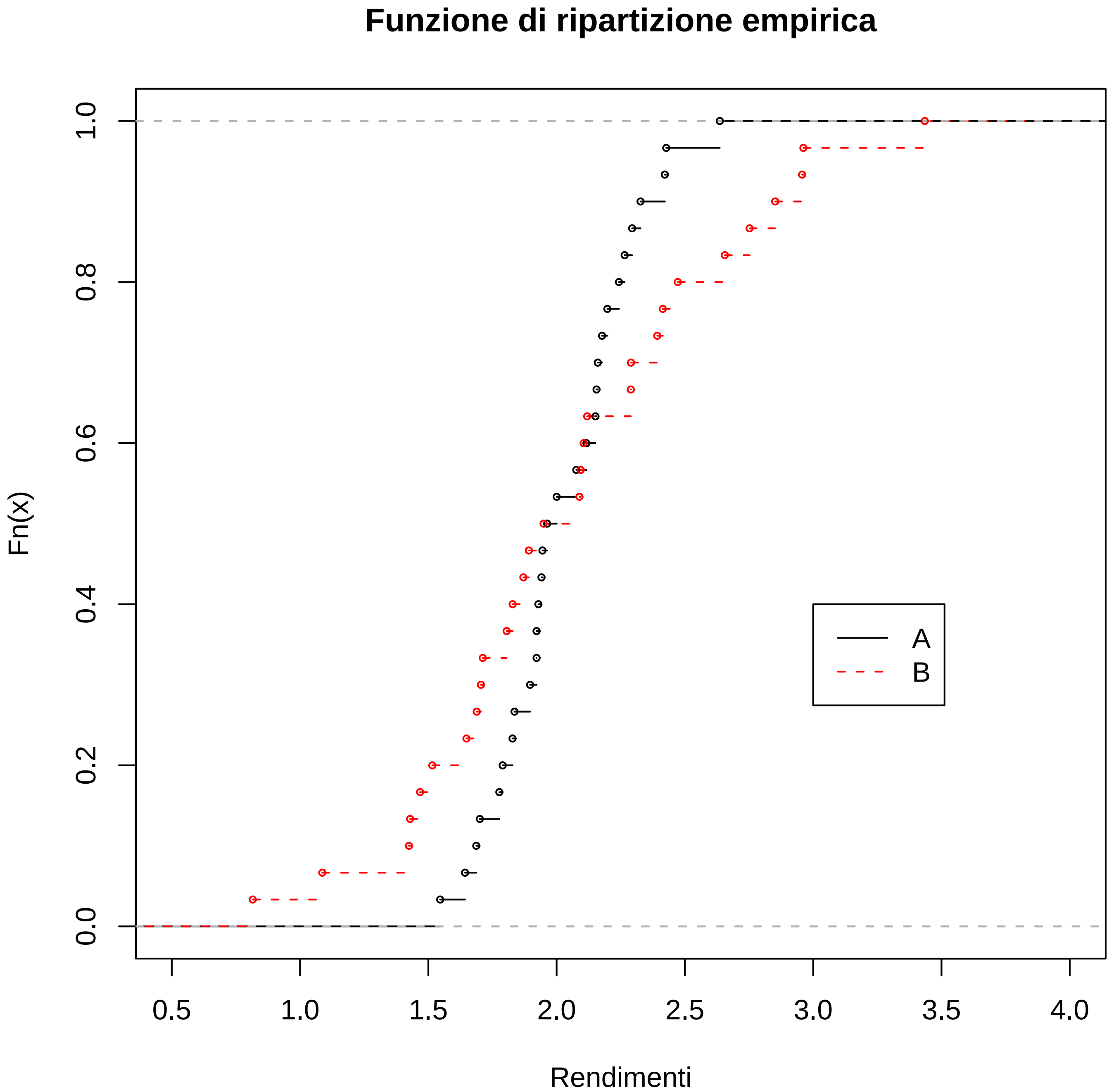
Per confrontare le *performance* di due tipologie di fondi, etichettate come A e B abbiamo preso in considerazione i rendimenti di 30 fondi per ciascuna tipologia. Riportiamo di seguito i diagrammi a scatola dei rendimenti.

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

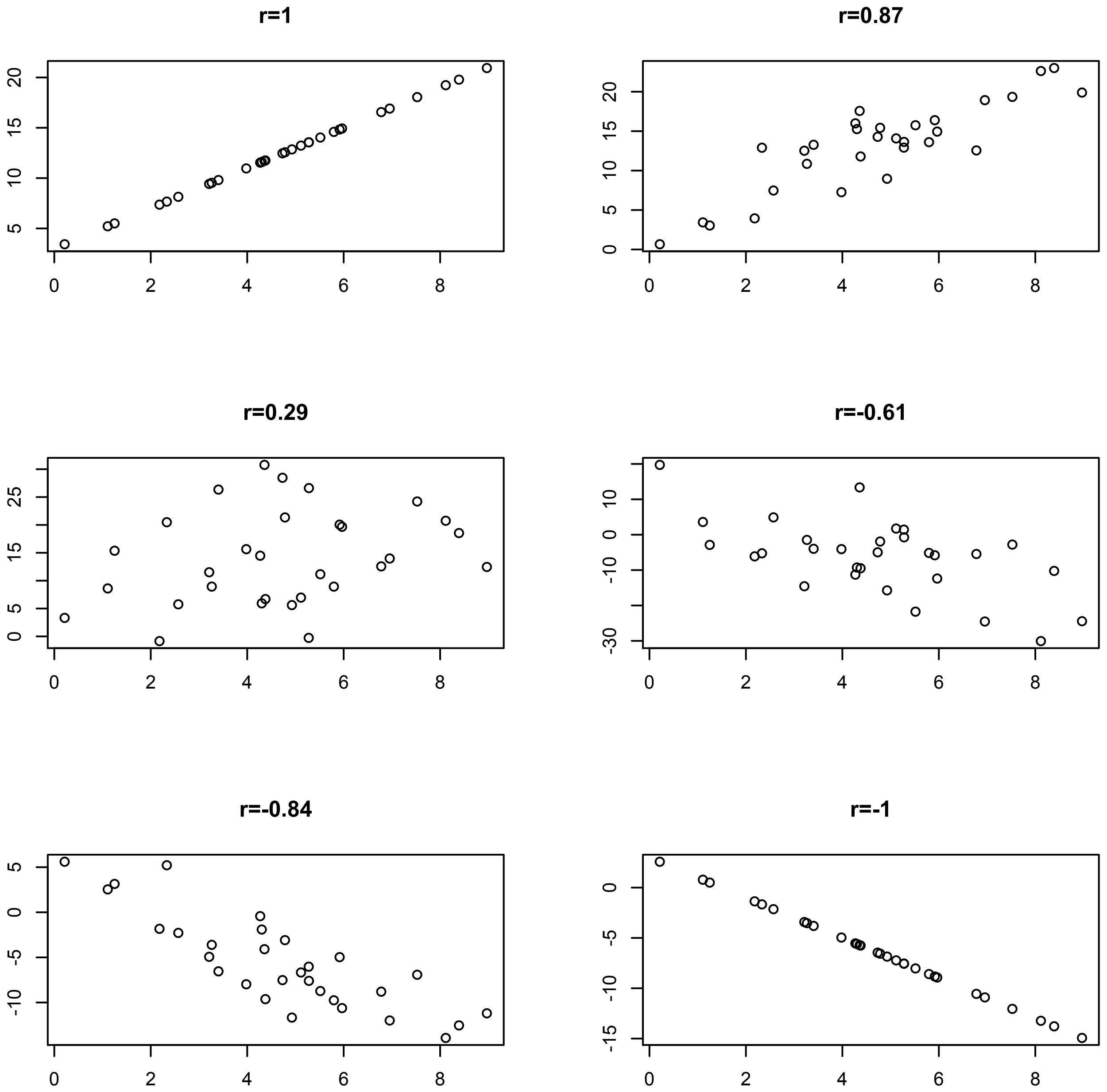


Indici di variabilità

	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.

La correlazione

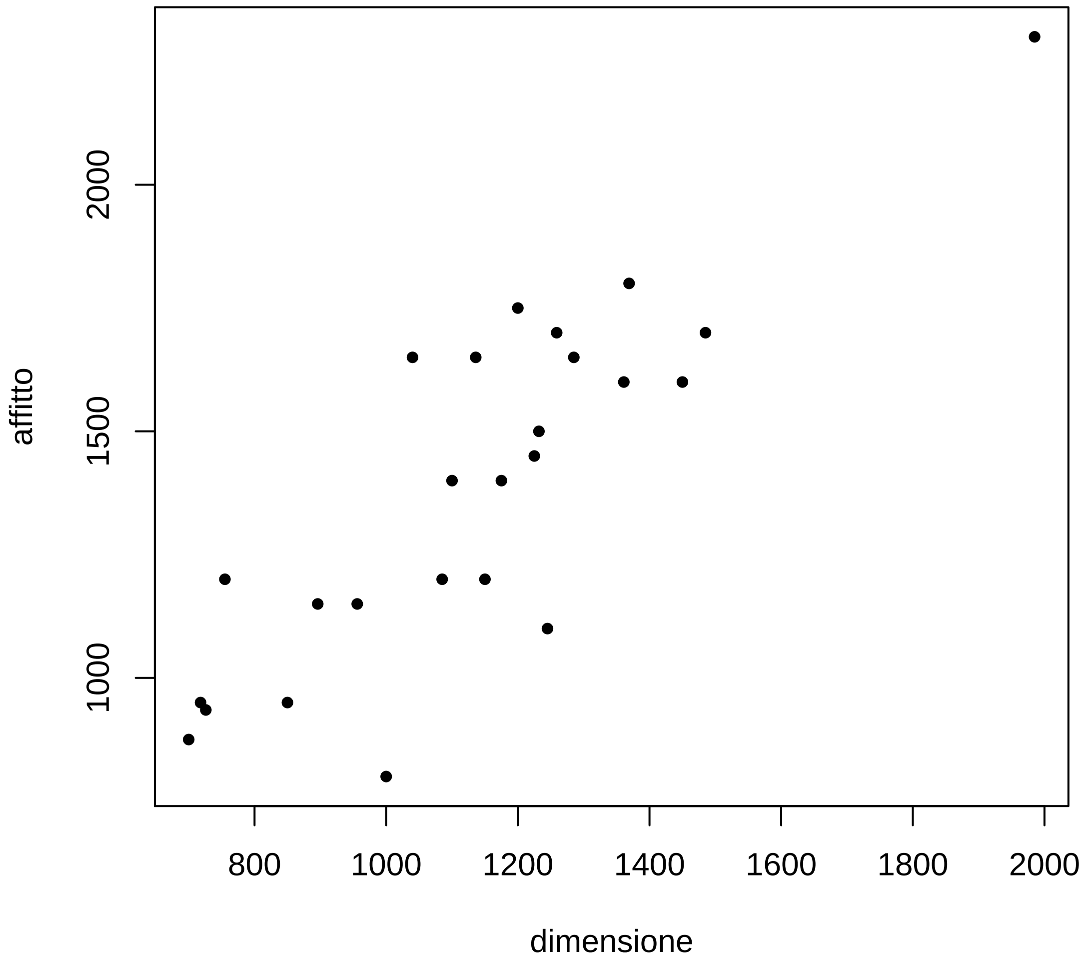


I dati

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

Calcoliamo la covarianza e la correlazione dei nostri dati, ponendo x =dimensione e y =affitto:

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

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

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

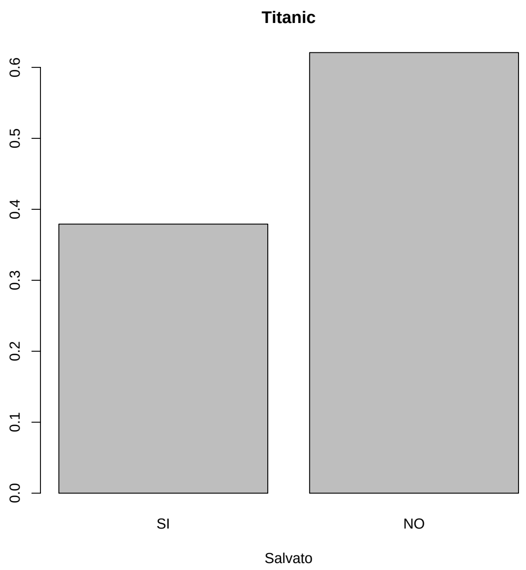
Una variabile alla volta

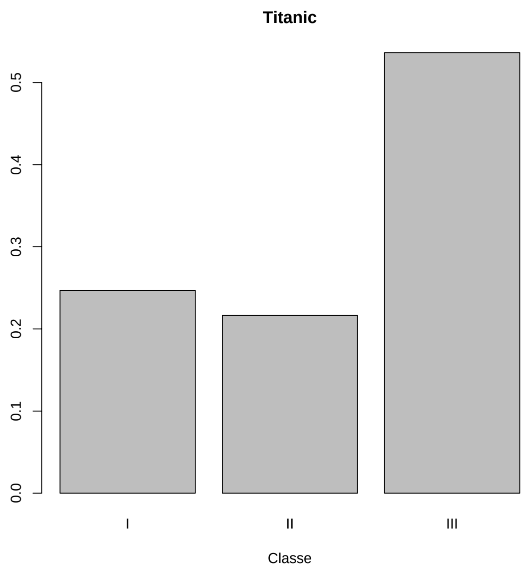
La variabile Salvato ha la seguente distribuzione di frequenze

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 ($2 \times 3 = 6$).

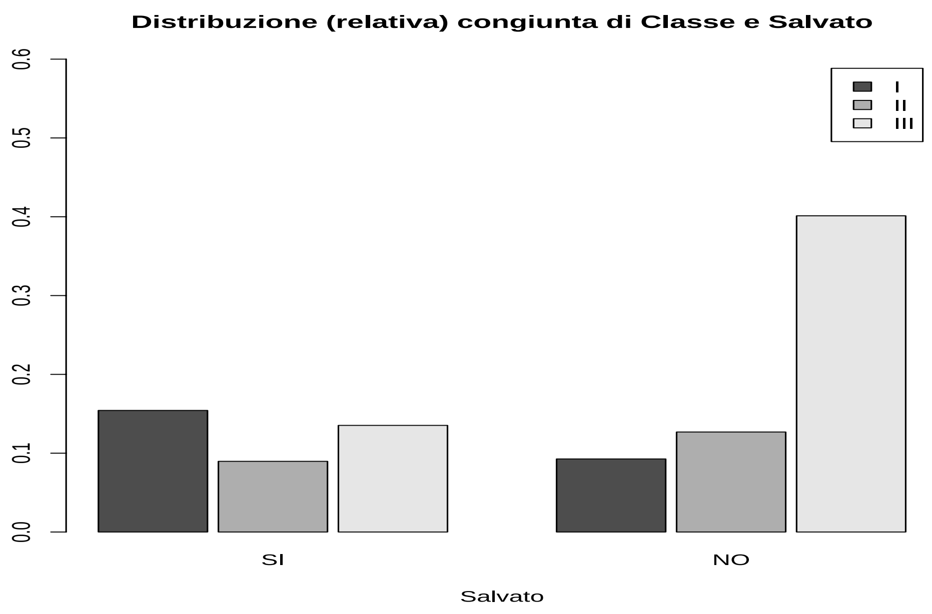
Possiamo anche considerare le frequenze relative, ottenute semplicemente dividendo le frequenze assolute per il numero totale $n = 1316$ di unità

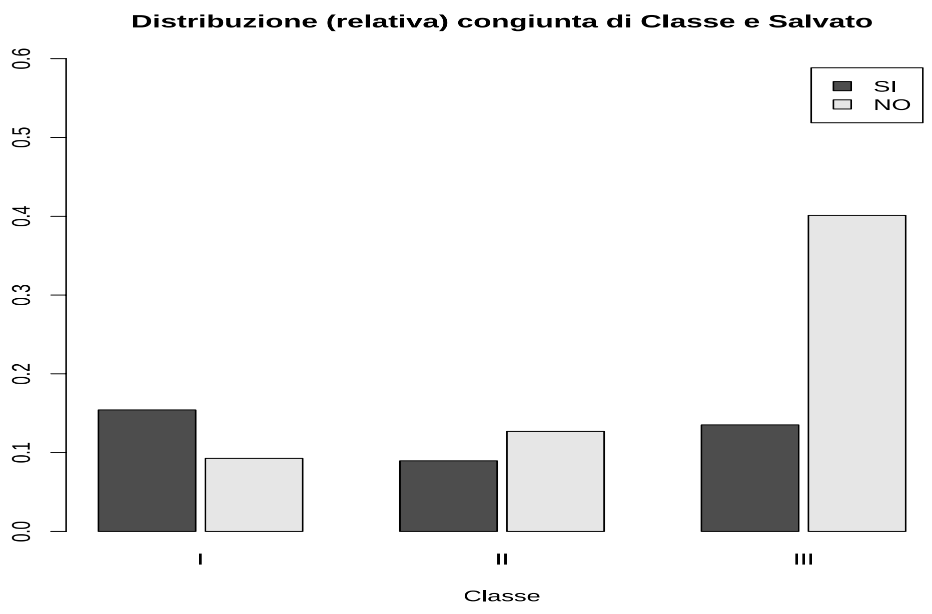
	Classe			
Salvato	I	II	III	totale
SI	0,154	0,090	0,135	0,38
NO	0,093	0,127	0,401	0,62
totale	0,247	0,217	0,536	1,000

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"):



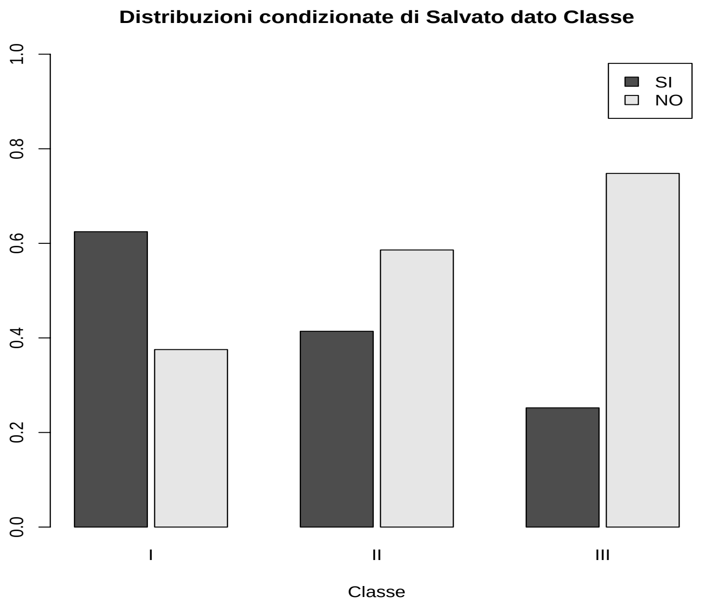


Distribuzioni condizionate di Salvato dato Classe

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

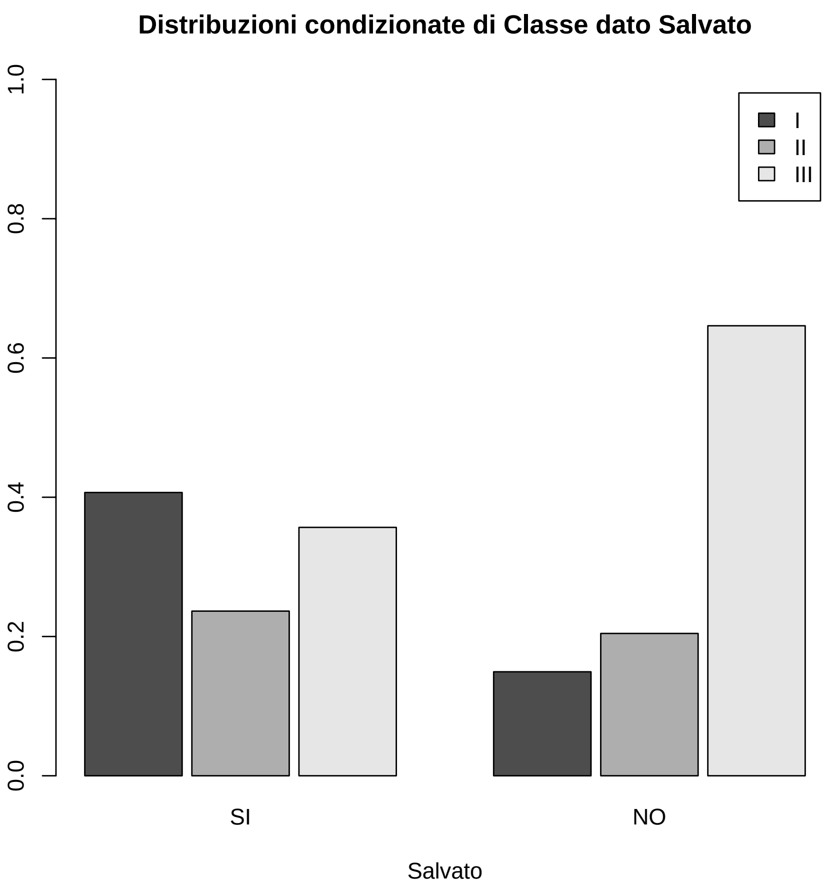


Distribuzioni condizionate di Classe dato Salvato

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.

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

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

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