Statistica Descrittiva
(lucidi a.a. 2001/2002)

Guido Masarotto
Facoltà di Scienze Statistiche
Università di Padova
guido@sirio.stat.unipd.it

2 dicembre 2001

Indice

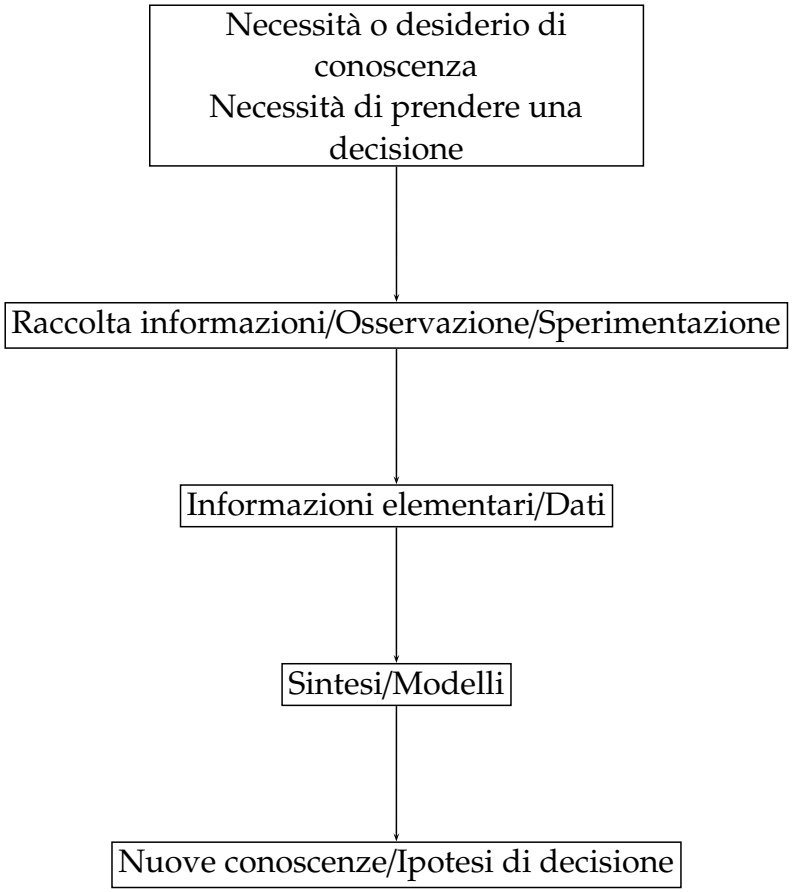
Unità A

Introduzione al corso

La statistica nella società dell'informazione

- Tutti dicono che viviamo nella *società dell'informazione*.
- Ma molti (tutti?) si lamentano che le informazioni sono troppe. E' facile raccoglierle, memorizzarle, distribuirle. E' difficile verificarle ed interpretarle.
- La statistica è, in molte situazioni, la *tecnologia* necessaria per risolvere queste difficoltà.
- Uno statistico, ad esempio, sa combinare informazioni di tipo differente, è in grado di valutarne l'affidabilità, sa sintetizzare e presentare molti dati in maniera tale da evidenziare la storia che raccontano, sa costruire modelli (=visioni stilizzate di una parte di mondo) che facilitano l'interpretazione, e, per esempio, permettono di calcolare previsioni o di formulare ipotesi di decisione.

Informazioni, nuove conoscenze, decisioni



Statistici, informazioni, nuove conoscenze, decisioni

Stilizzazione dei problemi di cui si occupa la statistica ed un po' di terminologia

- Un insieme (di individui o animali o oggetti o squadre di pallavolo 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 **polazione 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 od ipotesi di decisione. Ovvero, di trasformare i dati in affermazioni sul mondo (sulla popolazione di riferimento).

Ulteriore terminologia

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

Piccolo esempio (per fissare la terminologia)

Vogliamo sapere quale tra due trattamenti, chiamiamoli A e B, per una certa patologia, è migliore.

La *popolazione di riferimento* è l'insieme di tutti i pazienti che hanno quella particolare patologia *oggi ma anche domani,. . . .* Le unità statistiche sono i pazienti.
In questo caso, la popolazione è **virtuale**

25 pazienti con quella patologia sono trattati con A e 25 con B. Alla fine della 1° settimana, rileviamo per ogni paziente trattato se i sintomi sono scomparsi.

Il *campione* è costituito dai 50 pazienti trattati e per cui è nota la risposta (dopo una settimana) al trattamento.

I dati sono del tipo

paziente	trattamento	risposta
1	A	SI
...	...	...
25	A	NO
26	B	NO
...	...	...
50	B	NO

dove "risposta=SI" se il paziente ha risposto al trattamento entro la prima settimana (non ha sintomi alla fine della prima settimana); "risposta=NO" indica il caso opposto.

Le variabili considerate nello studio sono due:

trattamento le cui *modalità* sono A e B;

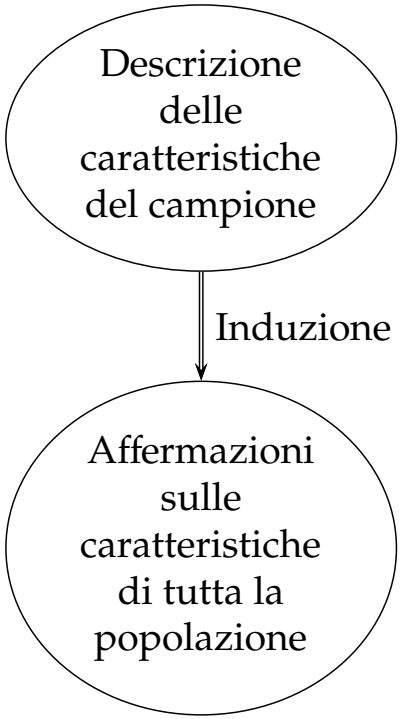
risposta dopo una settimana le cui *modalità* sono "SI" e "NO".

"Statistica Descrittiva" vs "Inferenza Statistica"

Descrittiva: ("quasi" sinonimi: esplorazione statistica dei dati, statistica senza modello probabilistico) Disponiamo di dati riferiti a 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*). Vogliamo utilizzare le informazioni del campione per fare delle affermazioni sulle caratteristiche di tutta la popolazione.

Tra *Statistica Descrittiva* ed *Inferenza Statistica* esiste una ovvia "fratellanza" ed, in realtà, nelle applicazioni, non sono facilmente separabili anche perchè i problemi di *inferenza* vengono normalmente affrontati in accordo allo schema



Contenuto del corso

Oggetto del corso sono alcune idee e tecniche di base utili per la descrizione di insiemi di dati.

Poichè essere uno statistico vuol dire essere in grado di affrontare e risolvere *problemi concreti* con metodi statistici, tutto il corso è organizzato intorno allo studio di alcuni casi reali.

I casi riguardano *deliberatamente* campi del sapere e della società diversificati (organizzazione aziendale, cronaca, medicina, botanica,. . .) per far percepire l'ampia "spendibilità".

Lo statistico, l'artigiano e le macchine ovvero breve istruzione per l'uso

- Ogni problema ha caratteristiche peculiari, ogni domanda arriva con conoscenze pregresse diverse,ogni volta ci sono scelte da fare,. . .
- Uno statistico è quindi come un vecchio artigiano: tutti i pezzi sono unici, i risultati dipendono dalla sua capacità, in determinato momento, di scegliere e controllare gli stumenti che usa, successi ed insuccessi sono frutto delle sue scelte. Del resto, perchè dovrebbero esistere delle lauree in Statistica se gli statistici fossero sostituibili con delle macchine?
- Quello che si può fare in un corso universitario è mostrarvi idee, tecniche e strategie che si sono rivelate utili in una certa gamma di situazioni. Ma voi dovrete affrontare altre situazioni, rispondere ad altre domande, . . .
- Tenetene già conto. Non cercate ricette. Guardate ai contenuti dei corsi (di Statistica) solo come ad esempi di quello che si può fare. Domandatevi sempre: "Sono convinto? Cosa farei di diverso?".

Esercizio (serio o scherzoso?)

Uno studente che durante un esame protesta dicendo

"Professore, questo esercizio non poteva darlo nel compito! Non è mai stato fatto a lezione!"

viene immediatamente bocciato. Spiegare perchè.

Unità B

Tre organizzazioni di un reparto di produzione

- Frequenze assolute, relative e cumulate
- Istogramma
- Diagramma a bastoncini
- Funzione di ripartizione empirica

Il problema

- In un reparto dove sono assemblati *walkman* vengono in tre giorni diversi provate 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: qualè l'organizzazione del lavoro migliore?

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

Organizzazione dei dati in una distribuzione di frequenza

I dati non sono "tantissimi" rispetto ad altre situazioni che si possono incontrare nelle applicazioni. Sono però troppi per cercare di capire qualcosa solamente "guardandoli". Dobbiamo quindi cercare di "sintetizzarli" in qualche modo.

Un primo tentativo in questo senso consiste nel suddividere l'intervallo che contiene tutti i valori osservati ($[670, 759]$) in un certo numero di "sotto-intervalli" e poi semplicemente nel "contare" quante osservazioni cadono nei vari "sotto-intervalli".

Utilizzando "sotto-intervalli" di lunghezza 5 ed aperti a destra, si ottiene la tabella della pagina seguente.

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

Nota 1 alla tabella: La prima colonna mostra i sotto-intervalli utilizzati. Le altre mostrano il numero di addetti che hanno "completato un numero di operazioni" appartenti al sotto-intervallo considerato. Ad esempio, il 13 che compare nella prima riga alla colonna 3 indica che esattamente 13 dei 288 addetti hanno, nel giorno in cui è stata sperimentata l'organizzione nuova 1, completato un numero di operazioni maggiore od uguale di 670 e minore (strettamente) di 675.

Nota 2 alla tabella: Le ultime tre colonne contengono delle **frequenze assolute**. In generale, si usa questo termine per indicare *il numero di unità statistiche che soddisfano una certa condizione* (nel nostro caso, che "appartengono" alla classe (intervallo) della prima colonna). Le prime due colonne (quella che mostra gli intervalli e quella contenente le *frequenze assolute* di vecchia) mostrano come gli addetti si sono *distribuiti* nei vari intervallini nel giorno in cui è stato utilizzato vecchia. Per questo motivo quando prese congiuntamente sono chiamate la **distribuzione di frequenza** di vecchia.

Commento alla tabella: nuova 2 sembra l'organizzazione migliore; nuova 1 è probabilmente l'organizzazione peggiore.

Frequenze relative

Dividendo una frequenza assoluta per il numero totale di unità statistiche (288 nel nostro caso) otteniamo le cosidette **frequenze relative**, ovvero

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

Hanno il vantaggio, rispetto alle frequenze assolute, di permettere di confrontare distribuzioni di frequenza basate su numeri differenti di unità statistiche.

	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

Frequenze relative per i dati considerati (consiglio: ricalcolatene almeno un paio).

Esercizio (semplice ma importante)

Si consideri una generica distribuzione delle frequenze relative

c_1	c_2	...	c_k
p_1	p_2	...	p_k

dove le c_i indicano le possibili modalità o classi di modalità mentre le p_i indicano le frequenze relative. Dimostrare che

$$\sum_{i=1}^{k} p_i = p_1 + p_2 + \cdots + p_k = 1$$

Istogramma

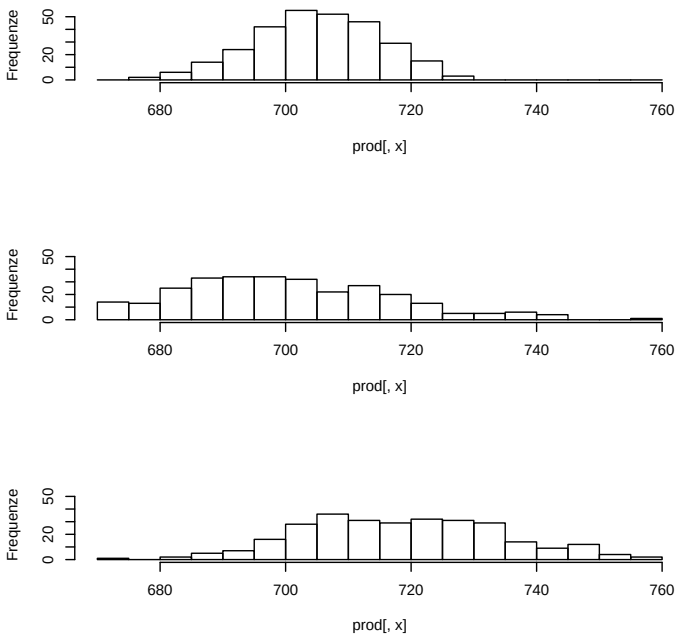
Le differenze tra le tre distribuzioni di frequenza si colgono ancora meglio se le rappresentiamo graficamente. Una possibilità è nella pagine seguente.

Il grafico è stato costruito ponendo

$$\text{(base rettangoli)} = \left(\begin{array}{c} \text{intervallini riportati} \\ \text{nella 1° colonna delle} \\ \text{tabelle precedenti} \end{array} \right)$$

$$\text{(altezza rettangoli)} = \text{(frequenze assolute)}$$

I diagrammi del tipo mostrato sono chiamati **istogrammi**.



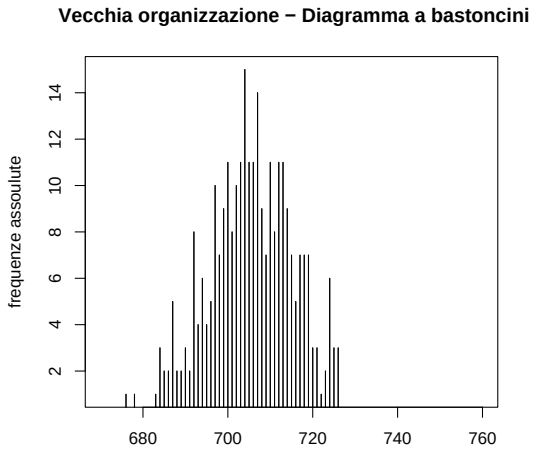
Commento al grafico: Il grafico suggerisce le stesse considerazioni fatte sulla base della tabella. La distribuzione di nuova 2 è, rispetto alle altre, quella più spostata verso destra (ovvero verso livelli di maggiore produttività). nuova 2 è quindi l'organizzazione migliore (sulla base di questi dati).

Diagrammi a bastoncini

Torneremo nel seguito sulla scelta delle classi (intervallini) e del loro numero.

Osserviamo comunque che avendo a che fare con dati che assumono solo valori interi possiamo in questo caso anche evitare del tutto la formazione delle classi.

Il grafico seguente (**diagramma a bastoncini**) è costruito disegnando in corrispondenza di ogni valore osservato un bastoncino (perpendicolare all'asse delle x) di lunghezza uguale alla frequenza assoluta con cui quel valore è stato osservato.

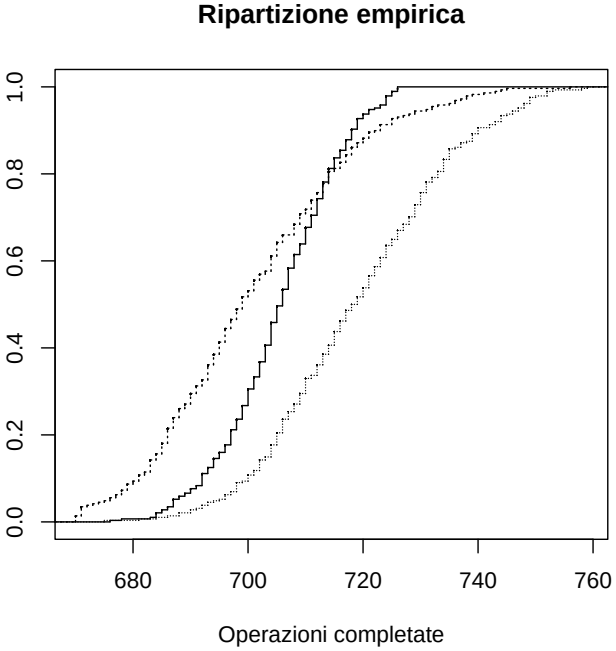


Funzione di ripartizione empirica

Una terza rappresentazione grafica di uso frequente è offerta dalla **funzione di ripartizione empirica** (che, tra l'altro, ha altre importanti applicazioni).

$$\left(\begin{array}{c}\text{funzione di}\\ \text{ripartizione}\\ \text{empirica}\\ \text{calcolata in } x\end{array}\right) = \frac{\left(\begin{array}{c}\text{numero di}\\ \text{osservazioni}\\ \text{minori di } x\end{array}\right)}{\left(\begin{array}{c}\text{numero totale}\\ \text{di}\\ \text{osservazioni}\end{array}\right)}$$

Per le tre organizzazioni del lavoro, il grafico di queste funzioni è riportato nella pagina seguente. Il "messaggio" può forse sembrare "a prima vista" meno evidente di quello contenuto negli istogrammi visti prima. Lo studente guardi però la definizione precedente e il grafico fino a che non si convince che il "messaggio" è il medesimo.



La curva centrale, tracciata con un tratto continuo, è riferita a vecchia. La curva "più alta" è quella riferita a nuova 1. La curva "più bassa" è quella di "nuova 2".

Frequenze cumulate

- Sono essenzialmente analoghe alla funzione di ripartizione empirica.
- Si ottengono "cumulando" progressivamente le frequenze.
- Possono essere "assolute" o "relative". Quelle relative coincidono con la funzione di ripartizione empirica alla fine di ogni intervallo.

Esempio di calcolo (nuova 1)

fine intervallo	frequenza assoluta	frequenza cumulata
675	13	13
680	12	25 = 13 + 12
685	20	45 = 13 + 12 + 20
⋮	⋮	⋮
755	0	287 = 13 + 12 + ⋯ + 0
760	1	288 = 13 + 12 + ⋯ + 0 + 1

Unità C

Misure di posizione

- Media aritmetica
- Mediana
- Quantili, quartili e percentili
- Diagramma a scatola con baffi

Misure o parametri di posizione

Le distribuzioni dei pezzi prodotti differiscono, come visto, sopratutto per la diversa "posizione".

Una domanda che sembra naturale è "di quanto?". Ad esempio, "Nuova 2" sembra con i dati disposizione migliore di "Vecchia". Ma quanto migliore?

Una possibile maniera per rispondere a questo tipo di domande si concretizza nel

1. Sintetizzare le singole distribuzioni in un unico numero che, in una qualche senso, indichi dove la distribuzione stessa è "posizionata". Ovvero, calcolare per ogni distribuzione una **misura (o parametro o indice) di posizione**.
2. Rispondere confrontando gli indici calcolati al punto precedente.

"Famosi" 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

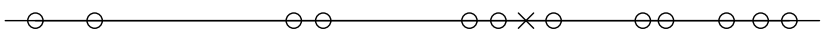
$$\overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

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 riorganizzazioni del lavoro

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

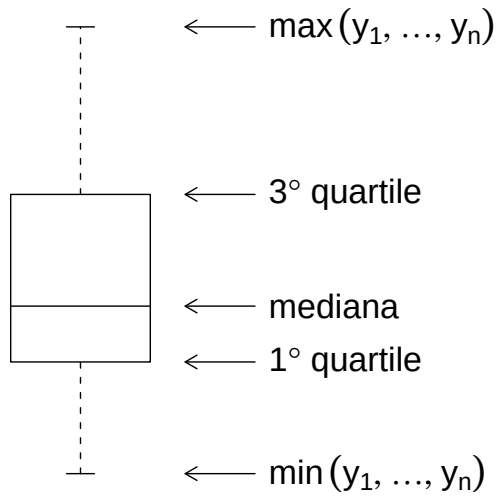
Come si vede risulta confermato i risultati precedenti. Indicano che nuova 2 potrebbe far aumentare la produzione di circa un 2%.

Quantili

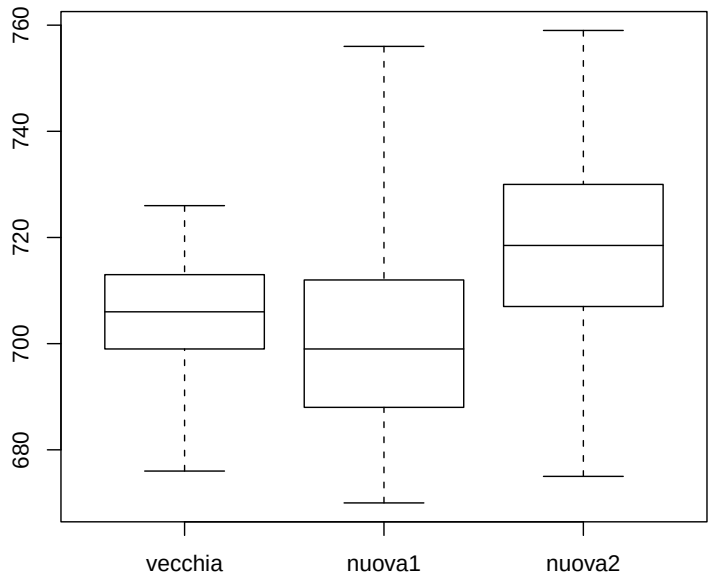
- Generalizzano la mediana.
- L'idea alla base di un **quantile-p** dove $p \in [0,1]$ è di cercare un numero che sia più grande del $100 \times p\%$ dei dati osservati e più piccolo del restante $100 \times (1-p)\%$. Ad esempio, un quantile−0,1 deve essere un valore che lascia a sinistra il 10% delle osservazioni ed a destra il restante 90%. Si osservi che, per costruzione, $\hat{F}(y_p) \approx p$ dove con $\hat{F}(\cdot)$ abbiamo indicato la funzione di ripartizione empirica.
- 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**.

Diagrammi a scatola con baffi

- Il nome deriva dall'inglese (*box and whiskers plot* spesso, anche in italiano, abbreviato in *boxplot*).
- Forniscono una 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.



Tre organizzazioni della produzione: diagrammi a scatola con baffi



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.

Una *approssimazione* spesso usata in questi casi è

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

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.

La media aritmetica: alcune proprietà

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

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

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

Ovvero, se $z_1 = a + by_1$, $z_2 = a + by_2$,. . . ,$z_n = a + by_n$ 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

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

La somma, e quindi la media, delle differenze dei dati dalla media (i cosidetti **scarti**) è sempre uguale a zero

Ovvero, in simboli,

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

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

Siano a un numero qualsiasi. Allora

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

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

Si tratta di una conseguenza banale ma importante della (C.1). Infatti la (C.1) garantisce che

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

Una non-proprietà (da non dimenticare) della media aritmetica

La media di una trasformazione non-lineare dei dati *non* è, in genere, uguale alla stessa trasformazione applicata alla media.

In formule, se $f(\cdot)$ è una qualsiasi funzione non lineare che trasforma numeri in numeri, allora, in generale, **non è vero** che

$$\frac{1}{n}\sum_{i=1}^{n} f(y_i) = f\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

Ad esempio, se $f(x) = x^2$, in generale, non è vero che

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2$$

ovvero che la media dei quadrati dei dati è uguale al quadrato della media. Lo si verifichi ad esempio ponendo $n = 3$, $y_1 = -1$, $y_2 = 0$ e $y_3 = 1$.

Un difetto della media aritmetica

Non è del tutto infrequente trovare degli insiemi di dati contenenti una piccola frazione di osservazioni anomale o atipiche, 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, come è facile capire dalla definizione stessa, una sola osservazione molto grande o molto piccola può *dominare* il valore assunto dalla media.

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 che (i) non è detto che esista un valore maggiore di un 50% esatto dei dati e minore dei restanti oppure (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
   $$\text{mediana} = \left(\begin{array}{c} \text{punto centrale} \\ \text{dell'intervallo} \end{array} \right),$$
   ovvero, 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*:

	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
frequenze assolute	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[1]. 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.

[1]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 la settima e l'ottava 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 già ordinati in senso crescente

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

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 $0{,}25y_{24} + 0{,}75y_{25}$;
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*.

Unità D

Analisi di un esperimento su due sonniferi

Piccolo insieme di dati utile come esercizio.

Descrizione dell'esperimento

Per verificare se una certa molecola aveva un qualche effetto come sonnifero è stato condotto il seguente esperimento:

1. A 10 volontari, senza una storia pregressa di insonnia, è stato somministrato in una notte un *placebo* (= una pillola contenente una polvere innoqua) e in un altra notte una pillola contenente il sonnifero. L'individuo ignorava quale pillola aveva assunto.
2. Per ogni individuo e per tutte e due le notti sono state cronometrate le ore di sonno.
3. E' stata poi calcolata la variabile, denominata sonno extra,

$$\begin{pmatrix}\text{ore di sonno}\\ \text{nella notte con}\\ \text{sonnifero}\end{pmatrix} - \begin{pmatrix}\text{ore di sonno}\\ \text{nella notte con}\\ \text{placebo}\end{pmatrix}$$

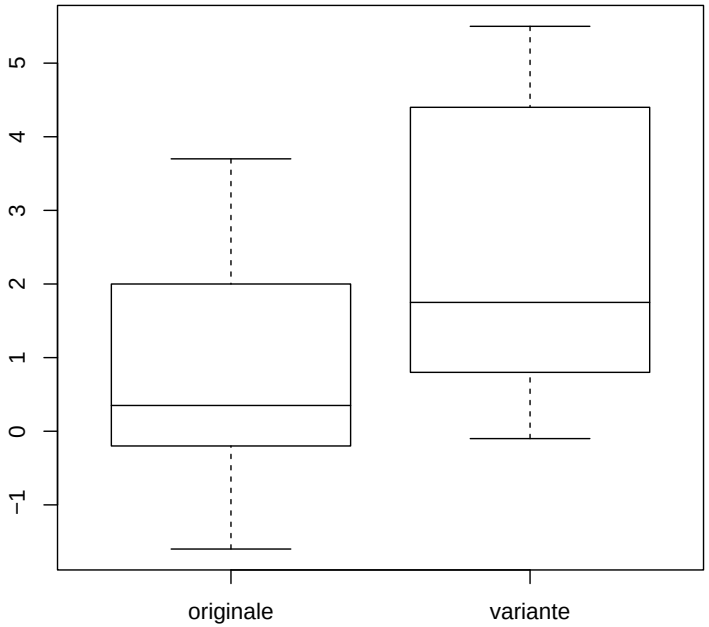
L'esperimento è stato poi ripetuto per una variante della molecola sotto studio.

I dati

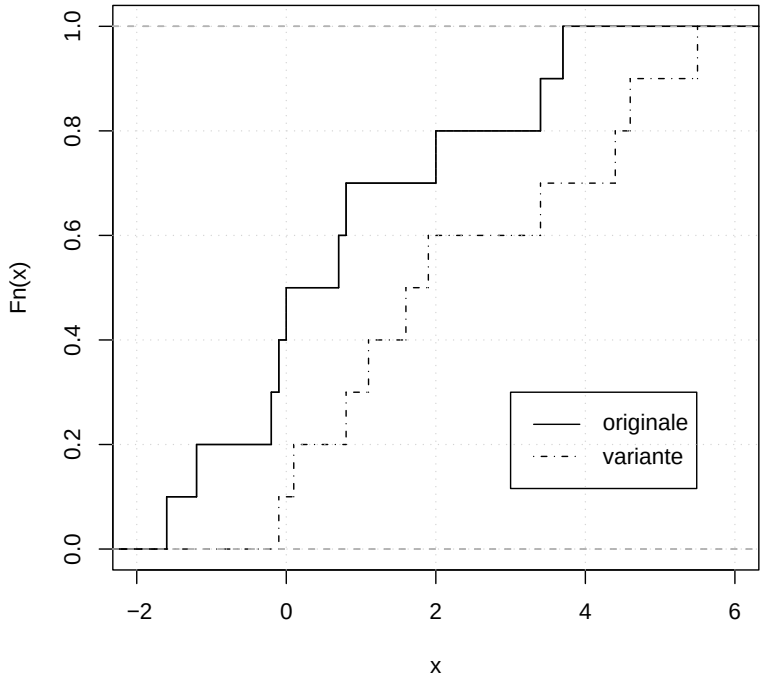
Ore di sonno extra

soggetto	molecola originaria	variante
1	0,7	1,9
2	-1,6	0,8
3	-0,2	1,1
4	-1,2	0,1
5	-0,1	-0,1
6	3,4	4,4
7	3,7	5,5
8	0,8	1,6
9	0,0	4,6
10	2,0	3,4

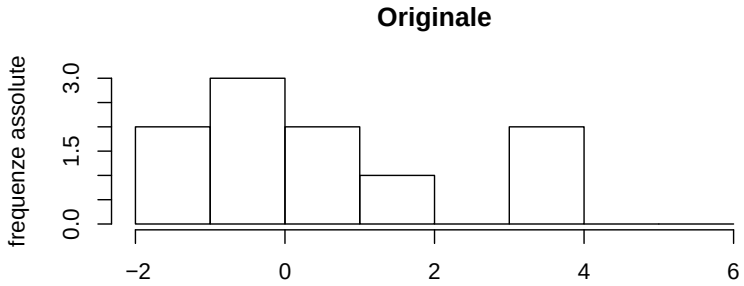
Diagramma a scatola con baffi

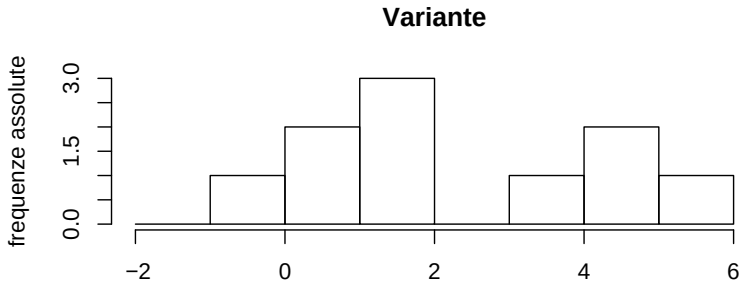


Funzione di ripartizione empirica



Istogrammi





Indici di posizione

sonnifero	media	mediana
originale	0,75	0,35
variante	2,33	1,75

Commento

L'effetto del farmaco originale non è del tutto chiaro visto che il 40% dei soggetti ha "dormito di meno". Nello stesso tempo però un 20% dei soggetti ha sperimentato un aumento di sonno superiore alle tre ore. Gli indici di posizione per il farmaco originale indicano "una moderata attività sonnifera".

La variante sembra decisamente più attiva. L'aumento del sonno è valutabile (almeno sulla base di questi dati) in circa due ore.

Indici di variabilità

Questa pagina fa riferimento a concetti e indici descritti nell'unità successiva.

	metodo A	metodo B
varianza	2,8	3,6
scarto quadratico medio	1,7	1,9
campo di variazione	5,3	5,6
scarto interquartile	3,4	4,1
MAD	1,1	1,7

Commento

L'effetto della molecola modificata sembra essere piú variabile tra i soggetti di quello della molecola originaria.

Unità E

Due metodi per il dosaggio dell'emoglobina

- Concetto di variabilità
- Varianza e scarto quadratico medio
- Altre misure di variabilità (campo di variazione, scarto interquartile, MAD)
- Standardizzazione
- Il coefficiente di variazione

Descrizione dei dati

Per confrontare l'efficacia di due diverse metodiche, diciamo A e B, per la misurazione dell'emoglobina si è proceduto nella seguente maniera:

1. Si è creato in laboratorio del *sangue artificiale* con un contenuto nominale di 15 grammi per $100cm^3$ di emoglobina.
2. Dal composito sono stati estratti 360 campioni. Su 180 campioni l'emoglobina è stata misurata utilizzando la metodica A mentre per i restanti campioni è stata usata la metodica B.

I dati sono riportati nel seguito. In questo caso le differenze tra le diverse misurazioni sono da attribuire in parte piccolissima (trascurabile da un punto di vista pratico) alle differenze intercorrenti tra i campioni di *sangue artificiale*. Sono, viceversa, dovute agli errori di misura della due metodiche.

I dati

Metodica A:

14.98654 15.14828 15.15741 14.78573 15.22160 14.86475 14.94835 14.64653
15.00364 15.06475 14.99282 14.92189 14.99575 15.64405 15.33492 14.73466
14.93331 14.94189 15.22719 14.64697 15.29087 14.90708 15.02215 14.97803
14.85369 15.35937 15.09510 14.70682 14.77203 15.28139 14.97825 14.93426
14.88053 15.11486 15.01449 15.10965 15.06394 14.79222 14.76657 15.15009
14.72711 14.98090 14.75410 14.90115 15.17730 14.89586 14.96936 15.00198
15.21905 14.97432 15.04769 14.90602 14.98397 15.33279 15.23659 15.07793
15.11311 14.78668 14.94863 15.05302 14.66592 15.13632 15.05976 14.98333
14.94836 14.99909 14.89628 15.19492 14.58320 14.83856 15.02708 15.03368
14.77780 15.40490 15.34432 14.71978 15.01237 14.73821 15.15963 15.22495
14.72350 15.31984 15.15886 15.49579 14.92472 15.32498 14.75861 15.13450
15.32448 14.98809 15.10179 15.29890 14.98695 14.96983 15.41413 15.12123
15.26597 15.09788 14.72764 15.14286 15.46952 15.13055 15.00559 14.83167
14.71813 15.03142 14.99039 14.89292 15.38937 14.76792 15.28734 15.20681
15.10421 15.45162 15.00441 14.90440 14.88044 15.26423 14.73072 14.60118
15.02576 14.84693 15.14960 14.55898 14.98658 15.23278 15.12612 15.09414
15.08752 15.09658 15.00574 15.03512 15.14976 14.93821 14.60090 14.97181
15.46100 15.05996 15.15891 15.08800 15.18066 14.64953 14.78898 14.82059
15.10667 15.00653 14.86939 15.16498 14.73762 14.58239 14.88403 15.19361
14.97908 15.00001 14.67176 15.01569 14.84952 14.79514 14.83009 15.11935
14.88455 14.51907 15.30290 14.93205 15.04080 15.11013 14.98879 14.91813
14.44163 14.80356 14.95201 15.11953 14.70853 15.15046 14.86291 15.34034
14.75090 14.86912 14.92036 14.93270

Metodica B:

14.62067 15.26097 14.87602 15.45027 15.32593 14.74834 15.19424 14.97163
15.09104 15.31831 15.06252 15.19373 14.72724 14.90797 14.75423 14.99820
14.31944 15.36786 15.48341 15.01780 14.58473 15.12630 15.03021 15.01365
14.34351 14.58493 14.97563 15.29785 15.44040 14.76965 14.53352 15.27177
15.28055 15.53123 13.82846 15.12360 15.24214 14.92428 14.65803 15.18722
14.83586 15.60325 14.85619 15.01115 15.30457 14.63107 15.13094 15.01104
14.64376 14.95360 15.53356 15.69041 14.82695 15.29568 15.09679 15.28104
15.10458 14.56744 15.19535 15.12521 14.46866 14.71408 14.93787 14.96020
14.78997 14.90338 15.45271 15.21145 15.38612 15.65470 15.06021 14.93443
15.02647 14.81672 15.38435 15.33135 15.11671 14.88797 15.03316 15.44367
15.43368 14.49792 14.82239 14.67267 14.59706 14.47115 15.19742 14.99877
15.06001 15.38270 14.60696 15.22274 15.63418 14.93377 15.36101 15.33171
15.32998 15.22515 14.42690 15.04781 15.64006 15.00974 14.60602 14.99329
15.33431 15.21277 14.87049 14.51086 15.80354 14.87940 14.90026 14.88420
14.44305 14.97657 14.78323 14.88690 15.10098 15.17093 14.91092 15.11042
14.59041 15.53938 14.58149 15.02527 14.86442 15.11187 14.85419 14.71329
14.75993 14.91381 15.09344 14.98552 15.30405 15.21515 14.99674 14.82704
15.06626 15.07602 14.96841 15.25176 14.84417 15.22224 15.42500 15.54409
15.54403 15.04272 14.74762 14.99012 15.15673 14.71890 15.15306 15.40174
14.37135 14.83015 15.27419 14.97257 14.59326 15.17822 15.15458 15.41121
15.16800 15.15758 14.58378 14.74625 14.85794 15.67266 14.99509 14.78124
15.21816 14.68988 14.84746 14.96116 14.96338 15.07308 14.85328 15.13676
14.89107 15.18928 14.74481 15.13047

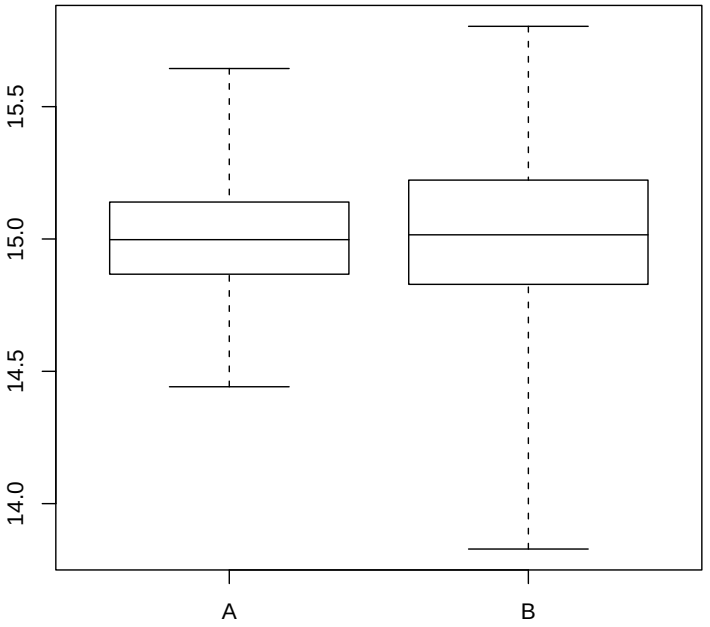
Una prima analisi

Applichiamo le cose che sappiamo.

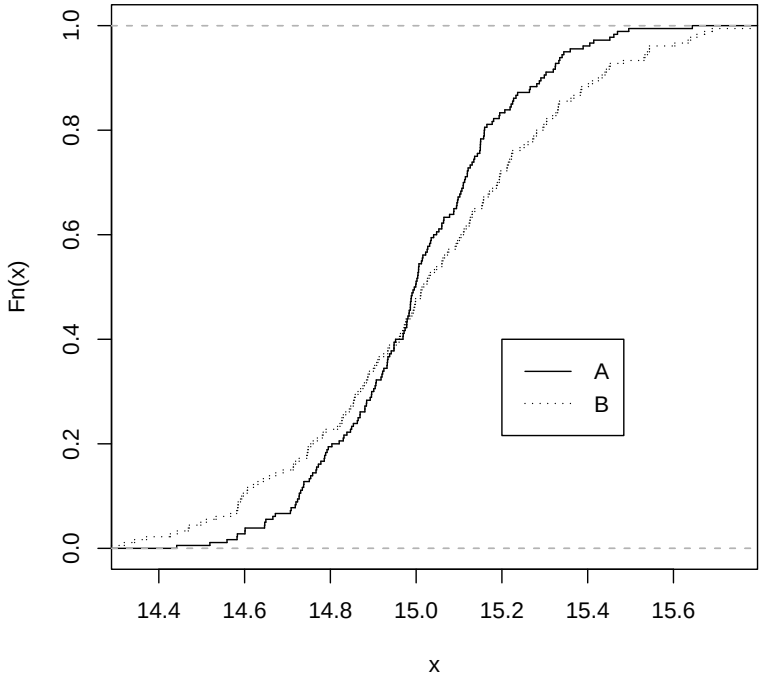
- Nei seguenti tre lucidi sono riportati rispettivamente i *boxplot*, le funzioni di ripartizione empirica e gli istogrammi.
- La seguente tabella riporta la media, i massimi e minimi ed i quartili dei due insiemi di dati

metodica	A	B
minimo	14,40	13,83
1° quartile	14,87	14,83
mediana	15,00	15,02
media	15,00	15,02
3° quartile	15,14	15,22
massimo	15,64	15,80

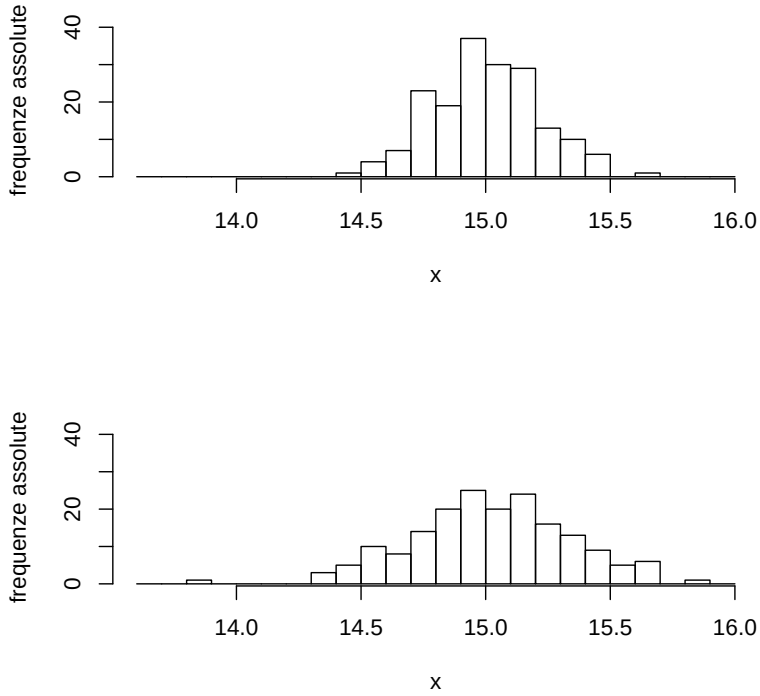
Diagramma a scatola con baffi



Funzione di ripartizione empirica per le due metodiche



Istogrammi per le due metodiche



Legenda: Primo grafico metodica A, secondo metodica B. Gli intervalli hanno ampiezza $0{,}1g$ per $100cm^3$.

Commento

1. Ambedue le metodiche sembrano essere state *tarate* accuratamente visto che i due insiemi di dati si distribuiscono intorno al valore nominale.
2. Però gli errori di misura della metodica B sembrano essere più grandi. Infatti in questo caso i dati sono più **dispersi** intorno al valore nominale. Ovvero, come si usa dire, mostrano una **variabilità** superiore.

Nota: E' importante che lo studente cerchi di 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

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

Nel 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. Può comunque anche essere interpretata come una media delle differenze al quadrato tra tutte le possibili coppie di dati. Infatti

$$\mathrm{var}(Y) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2$$

Dimostrazione.

$$\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2 = \\
=\ &\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(y_i - \overline{y}) - (y_j - \overline{y})]^2 = \\
=\ &\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - \overline{y})^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_j - \overline{y})^2 - \\
&-\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - \overline{y})(y_j - \overline{y}) = \\
=\ &\frac{2n}{n^2}\sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})\right]^2 = 2\mathrm{var}(Y).
\end{aligned}$$

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

$$(\text{varianza}) = \begin{pmatrix}\text{media dei}\\ \text{quadrati}\end{pmatrix} - \begin{pmatrix}\text{quadrato}\\ \text{della media}\end{pmatrix}.$$

Esempio di utilizzo

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

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

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

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

Varianza di una trasformazione lineare dei dati

Dati: $y_1, \ldots, y_n$.

Dati trasformati: $z_1 = a + by_1, \ldots, z_n = a + by_n$ dove a e b sono due costanti qualsiasi.

Allora

$$\text{varianza}(z_1, \ldots, z_n) = b^2 \text{varianza}(y_1, \ldots, y_n).$$

Sappiamo infatti che

$$\text{media}(z_1, \ldots, z_n) = a + b\text{media}(y_1, \ldots, y_n) = a + b\overline{y}.$$

Quindi,

$$\text{varianza}(z_1, \ldots, z_n) = \frac{1}{n}\sum_{i=1}^{n}(a + by_i - a - b\overline{y})^2 =$$

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

Esercizio: La formula mostra che la varianza delle z_i non dipende da a ("l'intercetta" della trasformazione). Si spieghi perchè 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 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* perchè troppo sensibile a possibili valori anomali.

2. **Scarto interquartile**.

   Scarto interquartile = (3° quartile)−(1° 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.

Due metodiche per la misurazione dell'emoglobina: indici di variabilità

	metodo A	metodo B
varianza	0,046	0,099
scarto quadratico medio	0,213	0,315
campo di variazione	1,200	1,970
scarto interquartile	0,270	0,393
MAD	0,135	0,198

La tabella mostra chiaramente come tutti gli indici considerati evidenzino la maggiore variabilità delle misure ottenute con la metodica B.

Il coefficiente di variazione

La variabilità guarda alle differenze tra le unità sperimentali. E' però evidente che il significato pratico delle differenze può dipendere dal livello del fenomeno considerato. Si pensi, ad esempio, al reddito. Una differenza di 30 milioni nel reddito annuo è importante se stiamo confrontando il reddito di due individui, uno con un reddito di 20 milioni e l'altro con un reddito di 50 milioni. La stessa differenza è praticamente irrilevante se stiamo confrontando il reddito di due ultra miliardari.

Può quindi essere interessante disporre di una qualche misura di variabilità *aggiustata* in qualche maniera per tenere conto del livello del fenomeno.

Il più diffuso prende il nome di **coefficiente di variazione** ed è definito come

$$\left(\begin{array}{c}\text{coefficiente}\\ \text{di variazione}\end{array}\right) = \frac{\text{(scarto quadratico medio)}}{\text{(media aritmetica)}}$$

Standardizzazione dei dati

A volte è utile trasformare un insieme di dati $y_1, \ldots, y_n$ in maniera tale che i dati trasformati, indichiamoli $z_1, \ldots, z_n$, abbiano media nulla e varianza unitaria.

E' facile verificare (lasciamo la dimostrazione come esercizio; si usino le proprietà della media e della varianza) che una trasformata appropriata consiste nel porre per $i = 1, \ldots, n$,

$$z_i = \frac{y_i - \text{media}(y_1, \ldots, y_n)}{\text{scarto quadratico medio}(y_1, \ldots, y_n)}.$$

I dati così trasformati vengono usualmente chiamati **standardizzati**.

Unità F

Ancora su istogrammi e diagrammi a scatola con baffi

- Numero di intervalli in un istogramma.
- Intervalli di ampiezza diversa: densità non frequenze.
- Una variante dei diagrammi a scatola con baffi.

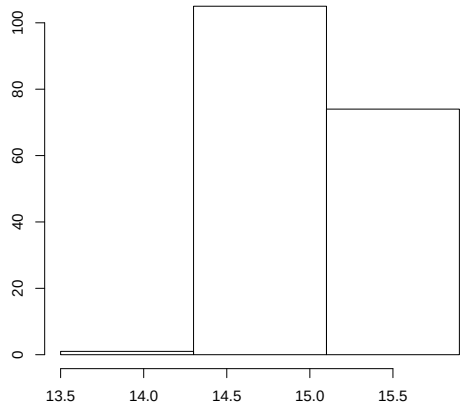
Numero degli intervalli

Nella costruzione di un istogramma esiste un elemento di arbitrarietà: la scelta di quanti e quali intervalli utilizzare.

E' prematuro a questo punto affrontare il problema della scelta ottima (ed in parte inutile visto che andando avanti avremmo strumenti migliori per fare quello che l'istogramma fa).

E' comunque necessario fare un po' di attenzione. Vediamo alcuni esempi.

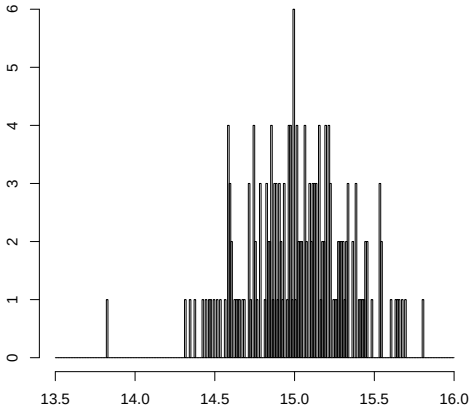
Pochi intervalli, poche informazioni



Misurazioni dell'emoglobina, metodica B (vedi unità F).

Questo istogramma *ci dice di meno* di quello mostrato nella unità E. Ad esempio non ci mostra che le frequenze diminuiscono *dolcemente* ma con regolarità quando ci si allontana da 15. In maniera essenzialmente erronea ci indica che non c'è una grande differenza tra, ad. es., 15.1 e 15.8.

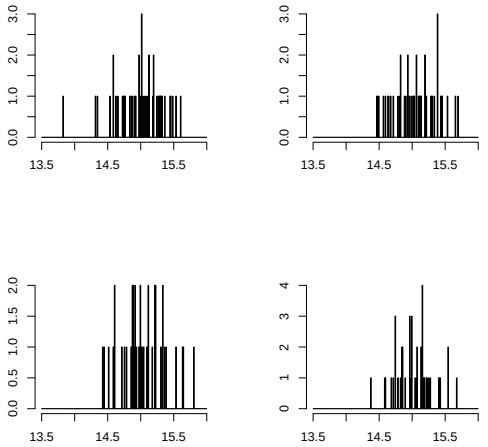
Troppi intervalli, troppi dettagli



Misurazioni dell'emoglobina, metodica B (vedi unità E).

Usando troppi intervalli mostriamo molti dettagli. Forse troppi. Ad esempio, le oscillazioni anche del 100% delle frequenze in intervalli adiacenti sono probabilmente *rumore*, caratteristiche particolari dei dati disponibili più che della metodica utilizzata per il dosaggio.

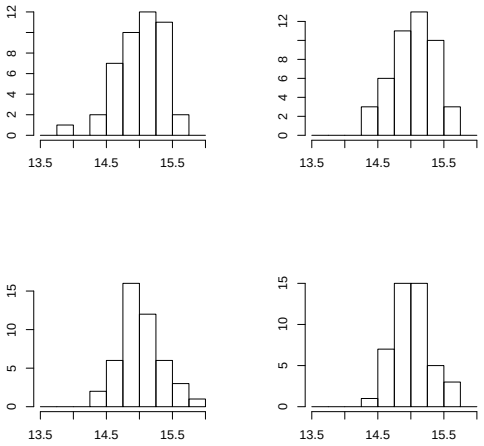
Troppi intervalli, non tanti dati, troppo rumore



Misurazioni dell'emoglobina, metodica B. Ogni istogramma si riferisce ad un sottogruppo di 45 dati.

Intervalli troppo piccoli e non tante osservazioni enfatizzano il *rumore*.

Un numero ragionevole di intervalli, non tanti dati



Misurazioni dell'emoglobina, metodica B. Ogni istogramma si riferisce ad un sottogruppo di 45 dati. I sottogruppi sono gli stessi della pagina precedente.

Meno *rumore*. Si confronti, ad esempio, l'istogramma in basso a destra con l'analogo della pagina precedente.

Suggerimenti pratici

- Quasi sempre è conveniente fare più di un grafico. Provare differenti lunghezze per gli intervalli e poi scegliere.
- Si tenga presente che il numero degli intervalli deve dipendere dal numero dei dati: ripartire 1000 osservazioni in 40 intervalli può anche dare risultati sensati, usare gli stessi 40 intervalli per 20 dati non può che dare un risultato erratico.
- Sono state suggerite varie *regolette*. Due tra le più usate sono:

  1. *Sturges*: (num. intervalli) = $1+\log_2$(num. dati)
  2. *Freedman & Diaconis*: (lunghezza intervalli) = 2(scarto interquartile)(num. dati)$^{-1/3}$

  Vanno però usate non in maniera automatica. Sono solo un punto di partenza.

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 è importante capire che le altezze dei rettangoli che compongono l'istogramma non devono essere proporzionali alle frequenze osservate ma alla **densità** delle osservazioni nelle singole classi. La densità è definita come

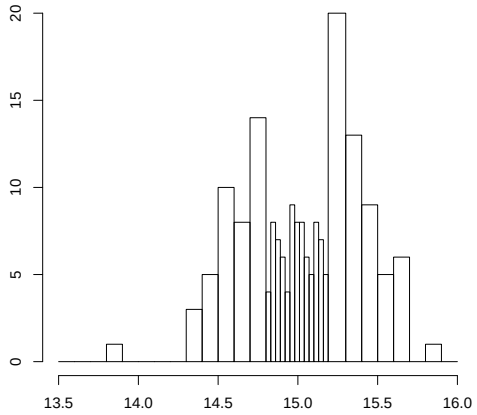
$$\left(\begin{array}{c}\text{densità}\\ \text{di un intervallo}\end{array}\right) = \frac{\text{frequenza dell'intervallo}}{\text{lunghezza dell'intervallo}}.$$

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.

L'uso della densità è anche legato al nostro cervello. In un istogramma percepiamo *alto* come sinonimo di *tanti*.

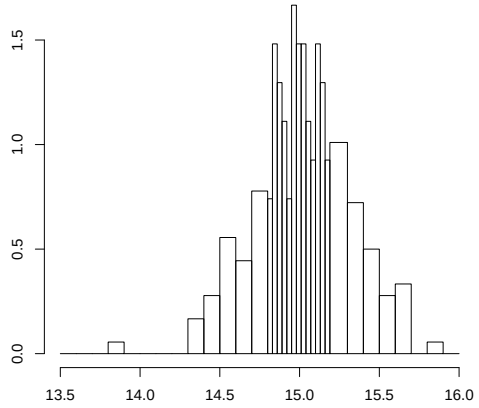
Due esempi sono presentati nelle seguenti due figure.

Misurazioni dell'emoglobina, metodica B. Intervalli più piccoli nella parte centrale. Altezze dei rettangoli proporzionali alle frequenze.



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

Misurazioni dell'emoglobina, metodica B. Intervalli più piccoli nella parte centrale. Altezze dei rettangoli proporzionali alle densità.



Il *buco* al centro è sparito. Il grafico correttamente ci dice che le osservazioni sono *addensate* intorno a $15g$ per $100cm^3$.

La variante più usata dei diagrammi a scatola con baffi

Spesso con un diagramma a scatola con baffi si vuole: (i) descrivere in maniera stilizzata la distribuzione dei dati (per noi, in questo momento, la posizione e la variabilità) e anche (ii) evidenziare eventuali valori *estremi*.

Una variante del diagramma usata a questo scopo può essere costruita come segue:

1. la *scatola* è costruita come descritto nell'unità C a partire dai tre quartili.
2. i *baffi* si estendono fino ai dati più lontani che siano però non più distanti di $k \times$ (scarto interquantile) dalla *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*

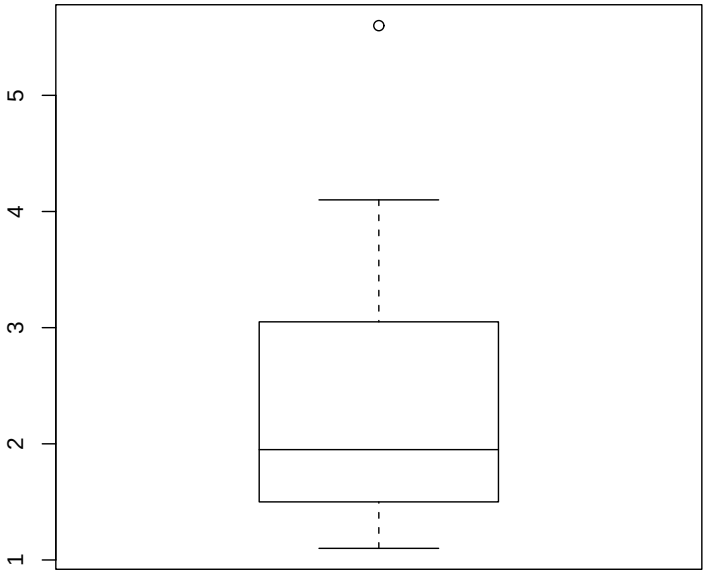
Dati (già ordinati):

1,1 1,3 1,4 1,6 1,8 1,9 2,0 2,5 2,9 3,2 4,1 5,6

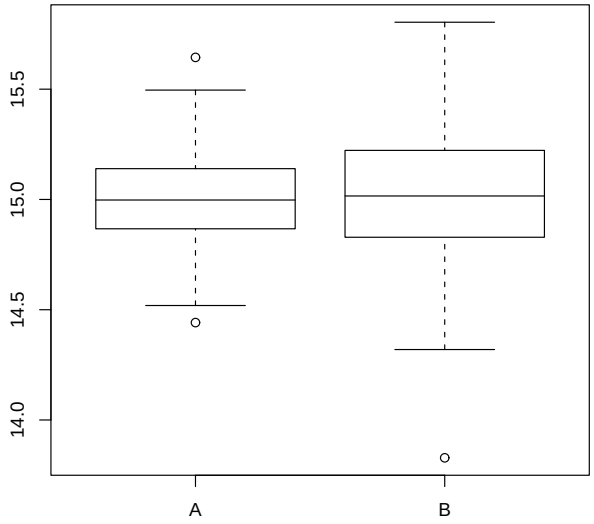
Perciò (1° quartile) = 1,5, (2° quartile) = mediana = 1,95 e (3° quartile) = 3,05. Quindi $1,5 \times$ (scarto interquartile) $= 1,5 \times 1,55 = 2,325$. Allora:

1. la *scatola* si estende da 1,5 a 3,05 con la mediana indicata da una linea a 1,95.
2. il *baffo* inferiore si estende fino all'osservazione più bassa tra quelle maggiori di (1° quartile) $-$ $2,325 = -0,825$, ovvero fino a 1,1.
3. il *baffo* superiore si estende fino all'osservazione più alta tra quelle minori di (3° quartile) + $2,325 = 5,375$, ovvero fino a 4,1.
4. sono da disegnare esplicitamente nel diagramma le osservazioni più piccole di 1,1 o più grandi di 4,1; in questo caso solamente l'osservazione risultata uguale a 5,6.

Diagramma a scatola con baffi (esempio precedente)



Due metodiche per la misurazione dell'emoglobina (vedi unità E): esempio di *boxplot* con possibili osservazioni anomale evidenziate



Unità G

Cenno a simmetria e curtosi

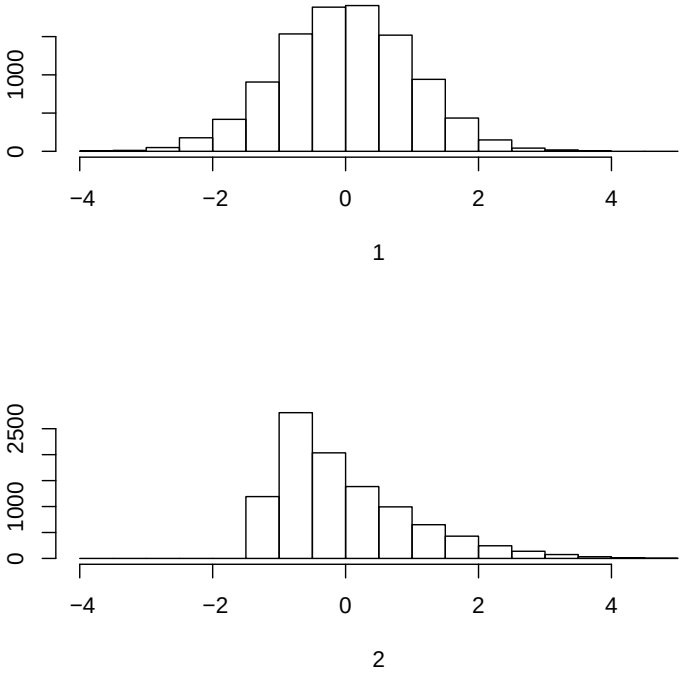
Consideriamo brevemente in questa unità due aspetti di una distribuzione di frequenza a volte interessanti di per sè ma sopratutto, che saranno utili nella la scelta di un appropriato *modello statistico*.

Simmetria

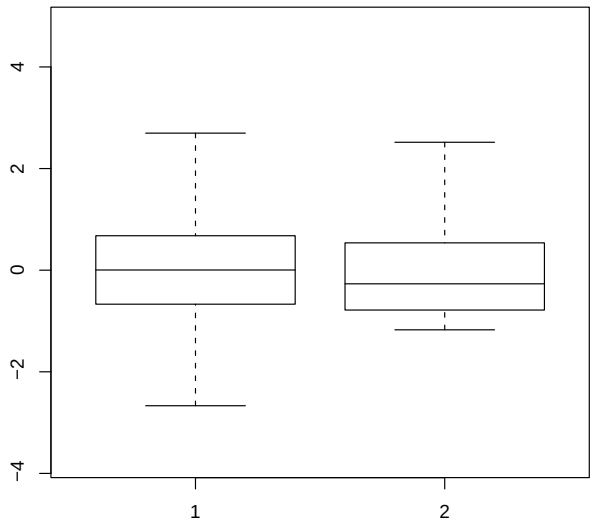
Le seguenti due pagine mostrano rispettivamente gli istogrammi e i *boxplot* costruiti a partire da due insiemi di dati *standardizzati* nella maniera brevemente descritta alla fine della unità E. I due insiemi di dati sono perciò almeno approssimativamente omogenei per quanto riguarda posizione e variabilità. Quantomeno, 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**.

Due insiemi di dati standardizzati: istogramma



Due insiemi di dati standardizzati: *boxplot*



Nota: la lunghezza massima dei baffi è stata posta uguale a 1,5 × (scarto interquartile) ma **non** sono state evidenziate le osservazioni esterne ai baffi stessi.

Indice di asimmetria

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

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

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.

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 $z_i = a + by_i, \quad i = 1, \ldots, n$. Lo studente lo verifichi come esercizio.

Curtosi

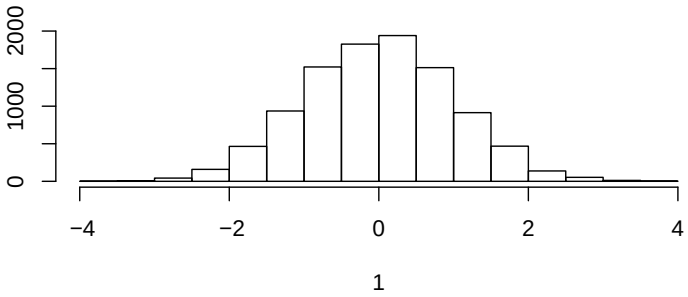
Anche i grafici nelle seguenti due pagine confrontano dati standardizzati. In questo caso, ambedue le distribuzioni sono (almeno approssimativamente) simmetriche. Però, nonostante l'uguaglianza delle varianze, la prima distribuzione a delle **code più pesanti** della seconda. Questa caratteristica (maggiore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

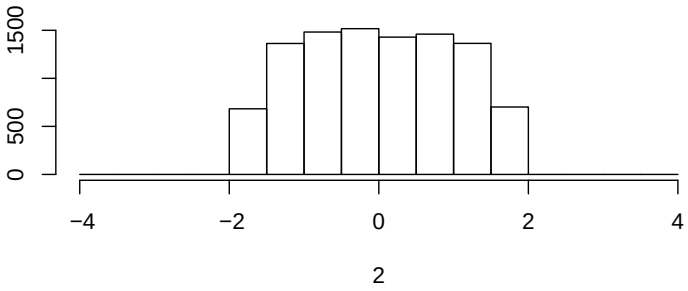
Il principale indice usato è l'**indice di curtosi standardizzato** definito come

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

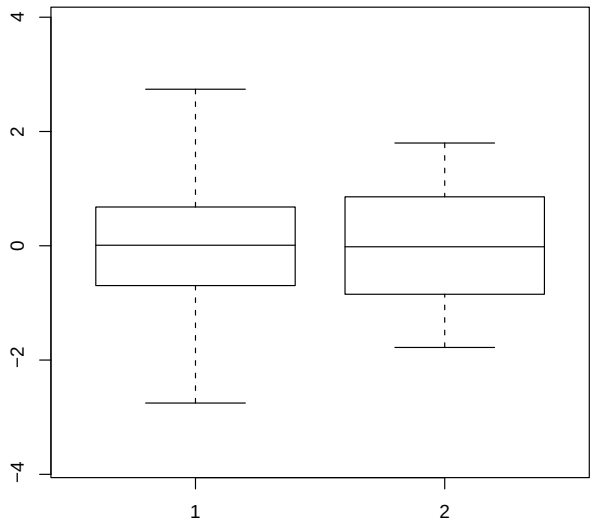
Si osservi che questo indice può essere visto come un rapporto tra due indici di variabilità. L'indice a numeratore (la media delle potenze quarte degli scarti dalla media aritmetica) è scelto in maniera tale da essere più sensibile alla presenza di code pesanti dell'indice a numeratore (la potenza quarta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

Due insiemi di dati standardizzati: istogramma





Due insiemi di dati standardizzati: *boxplot*



Nota: la lunghezza massima dei baffi è stata posta uguale a 1,5 × (scarto interquartile) ma **non** sono state evidenziate le osservazioni esterne ai baffi stessi.

Unità H

Trattamento della calcolosi uretrale mediante litotripsia extracorporea

- Moda
- Grafico a barre
- Mutabilità

I dati

La litrotripsia extracorporea è un trattamento relativamente poco gravoso per il paziente (può essere fatto in *day hospital*) per la calcolosi. Per valutarne l'efficiacia nel caso della calcolosi uretrale la risposta di 80 pazienti è stata rilevata utilizzando la seguente scala di modalità che si riferisce al grado di frammentazione dei calcoli dopo la prima seduta di trattamento:

1. buono: tutti i frammenti sono più piccoli di 3*mm*.
2. medio: nessun frammento sopra i 5*mm*, almeno uno maggiore di 3*mm*.
3. scarso: frammenti maggiori di 5*mm*.
4. assente: nessun segno di frammentazione dei calcoli originari.

Per ogni paziente è poi noto l'uretere (*lombare*, *pre-sacrale* o *pelvico*) dove si erano formati i calcoli.

I dati prendono quindi la forma di una lunga tabella del tipo:

paziente	uretere	grado di frammentazione
1	pelvico	buono
2	pre-sacrale	assente
3	pelvico	scarso
⋮	⋮	⋮
80	lombare	buono

Frequenze assolute e relative

La tabella della pagina precedente è *poco maneggevole*. La seguente mostra le frequenze assolute.

	Grado di frammentazione				
Uretere	buono	medio	scarso	assente	Tot.
Lombare	12	26	3	1	42
Pre-Sacrale	2	8	0	0	10
Pelvico	12	13	2	1	28

Ad esempio, 12 è il numero dei pazienti che avevano una calcolosi all'uretere lombare e che hanno avuto una buona risposta.

Dividendo ogni riga per il suo totale otteniamo le frequenze relative (sede per sede).

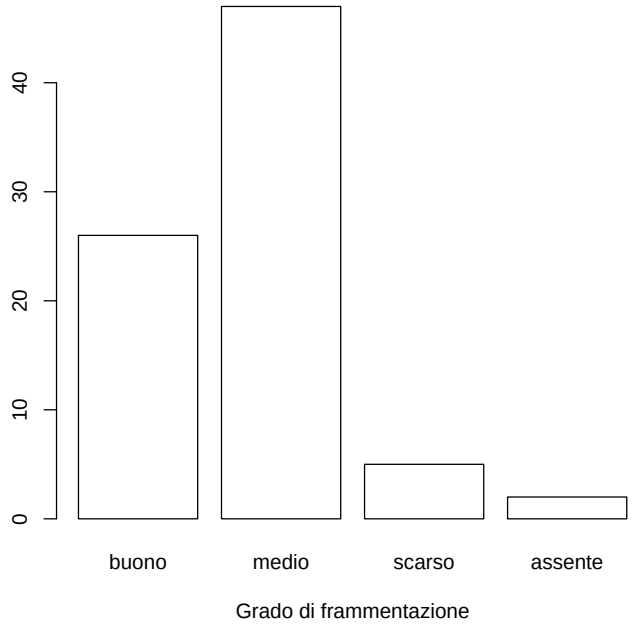
	Grado di frammentazione			
Uretere	buono	medio	scarso	assente
Lombare	0,286	0,619	0,071	0,023
Pre-Sacrale	0,200	0,800	0	0
Pelvico	0,429	0,464	0,071	0,036

Commento: I medici considerano soddisfacente sia le risposte "buone" che quelle "medie". Quindi, la tabella mostra una ottima efficacia della metodica considerata. Per quanto riguarda le differenze tra le sedi si osserva: (i) una probabile maggiore risposta delle calcolosi pelviche (quasi il 43% dei pazienti hanno avuto una risposta "buona") e (ii) l'assenza di risposte non soddisfacenti per le calcolosi all'uretere pre-sacrale (forse però dovuta alla scarsa numerosità del gruppo).

La natura di questi dati è diversa da quelli visti in precedenza

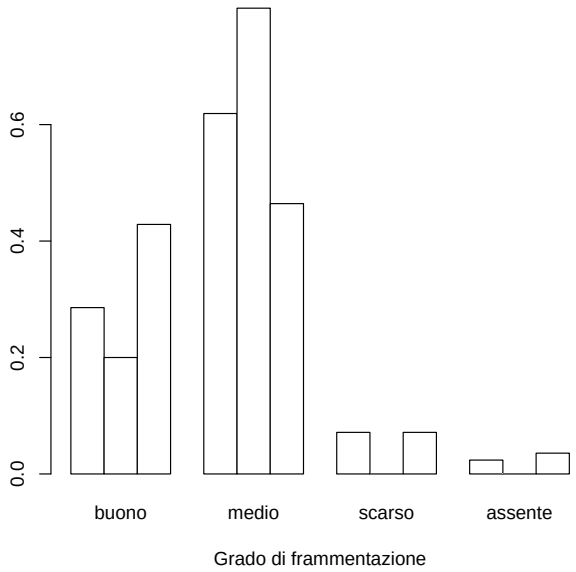
- Nei precedenti esempi avevamo dati **numerici**. In questo caso sono espressi da aggettivi. Sono dei dati **qualitativi** o **categoriali**.
- Questo cambia quello che possiamo e non possiamo fare. Ad esempio, non ha senso chiederci quanto vale la media aritmetica della risposta per i pazienti con una calcolosi all'uretere lombare. O quanto è grande la varianza.
- Le risposte sono ordinate. Possiamo quindi, almeno in teoria definire la risposta mediana. La mediana infatti non guarda ai valori ma solo all'ordine delle osservazioni.
- Volendo sintetizzare ogni riga in un unico valore probabilmente useremo la **moda** della riga. La moda la definiamo come la modalità con la più alta frequenza. In questo caso, per tutte e tre le regioni la moda è *medio*. Si osservi che la moda può essere usata per qualsiasi distribuzione di frequenza. Anche per quelle delle unità precedenti basate su dati numerici.

Diagramma a barre: tutte le sedi insieme, frequenze assolute



Anche la rappresentazione grafica più usata è leggermente diversa. Si osservi che i rettangoli, contrariamente al caso di un istogramma, sono disegnati *staccati*. Quello che mostriamo nel grafico sono frequenze di categorie in una qualche forma *separate*.

Diagramma a barre: sedi distinte, frequenze relative



Le frequenze relative per le tre sedi sono rappresentate affiancate. L'ordine (da sinistra) è lombare, pre-sacrale e pelvico. Ovviamente ci si può sbizzarrire con i colori. Si osservi che perchè abbia senso il confronto dobbiamo utilizzare le frequenze relative.

Mutabilità (idea di)

- Analogo della variabilità per dati qualitativi.
- Non possiamo guardare alle differenze tra i valori osservati. Possiamo però guardare alle differenze tra le frequenze.
- Si definisce come situazione di *minima mutabilità* una situazione in cui tutte le unità statistiche si *concentrano* nella stessa modalità. In questo caso le unità statistiche sono perfettamente omogenee rispetto al fenomeno considerato. Si osservi che in questo caso 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 supposte 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	$\frac{1}{k}$	$\cdots$	$\frac{1}{k}$	$\cdots$	$\frac{1}{k}$

Esempio di un ambito applicativo in cui la mutabilità costituisce una caratteristica importante di una popolazione

Gli studiosi di ecologia considerano la diversificazione delle specie che popolano un certo territorio come una proprietà fondamentale. Infatti, più le specie sono diversificate più è grande il patrimonio genetico e quindi più il sistema sarà capace di adattarsi a cambiamenti di qualsiasi origine. Viceversa, un territorio popolato da una sola specie è intrinsicamente fragile.

L'idea è fondamentalmente quella che è alla base della mutabilità. Si pensi, ad esempio, ad un lago e ai pesci che lo popolano. Se i pesci appartengono tutti alla stessa specie, allora la distribuzione dei pesci tra le varie specie assume la forma prevista dalla tabella di minima mutabilità. Viceversa, se il lago è popolato da più specie di pesci senza una specie particolarmente predominante (ovvero se la popolazione dei pesci è ben diversificata) allora la tabella che ci mostra come i pesci si ripartiscono tra le varie specie si avvicinerà a quella di massima mutabilità.

Cenno agli 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. Lo studente se ne convinca.

– Si dimostra che assume valore massimo nelle situazioni di massima mutabilità. Ovvero che, 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

$$G_{norm} = \frac{G}{\text{massimo valore possibile per } G} = \frac{k}{k-1}G$$

L'indice normalizzato varia tra 0 ed 1. In particolare, assume valore 0 in presenza di minima mutabilità e, viceversa, 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
frequenza assoluta	f_1	$\cdots$	f_i	$\cdots$	f_k	$n = \sum f_i$

può essere calcolato utilizzando la formula

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

Lasciamo la semplice verifica di questa formula come esercizio allo studente.

• **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.

– E' possibile inoltre dimostrare che anche questo indice assume valore massimo nelle situazioni di massima mutabilità. Ovvero che, qualsiasi siano le frequenze relative,

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

– Può quindi essere eventualmente *normalizzato* in maniera analoga a quanto visto per l'indice di Gini, ovvero, ponendo $H_{norm} = H/\log(k)$.

– Se sono note le frequenze assolute possiamo

calcolare H utilizzando la formula

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

- **Indici di mutabilità per i dati sulla litotripsia extracorporea**

	lombare	pre-sacrale	pelvico
G_{norm}	0,71	0,43	0,79
H_{norm}	0,67	0,36	0,74

Commento: gli indici normalizzati indicano una maggiore variabilità della risposta per le sedi "lombare" e "pelvica" rispetto alla sede "pre-sacrale". I valori per le prime due sedi non sono *piccoli*. Questo è dovuto alla dispersione dei pazienti sulle prime due categorie. Infatti, se le *aggregassimo* ovvero se decidessimo di lavorare con la seguente tabelle delle frequenze assolute

	Grado di frammentazione		
Uretere	buono o medio	scarso	assente
Lombare	38	3	1
Pre-Sacrale	10	0	0
Pelvico	25	2	1

gli indici di mutabilità diventerebbero

	lombare	pre-sacrale	pelvico
G_{norm}	0,26	0	0,29
H_{norm}	0,33	0	0,37

Unità I

Tipi di dati

- Dati qualitativi o categoriali (ordinali, sconnessi). Dati dicotomici o binari.
- Dati numerici (interi, continui). Scala intervallo e rapporto.
- Dati univariati vs. multivariati.
- Studi sperimentali vs. osservazionali.

Dati qualitativi, dati numerici, dati . . .

Abbiamo visto nell'unità precedente che differenti tipi di dati possono condizionare i metodi utilizzabili già al livello elementare a cui siamo. Come si vedrà nel seguito (e soprattutto nei corsi successivi) il tipo dei dati disponibili condizionerà anche la scelta di possibili *modelli statistici*. E' quindi conveniente, se non altro per avere a disposizione una terminologia adeguata, tentare una classificazione.

In *statistica* si parla di dati:

- **qualitativi** o **categoriali** quando le modalità utilizzate per descrivere il fenomeno analizzato prendono la forma di aggettivi o di altre espressioni verbali. A loro volta i dati qualitativi possono essere
  - **sconnessi** se non esiste nessun ordinamento naturale tra le modalità; esempi di dati sconnessi sono: (i) la religione, (ii) la modalità di somministrazione di un farmaco (ad es., per via orale, parentale o sottocutanea);
  - **ordinali** nel caso in cui un ordinamento naturale esiste; esempi di dati qualitativi ordinali sono: (i) il titolo di studio, (ii) la risposta ad un trattamento (ad es. classificata come "assente", "parziale", "ottima").

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

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

  – **interi** quando le modalità sono esprimibili da numeri interi; esempi sono: (i) il numero di figli, (ii) il numero di metastasi polmonari, (ii) il numero di pezzi prodotti;

  – **continui** o **reali** quando le modalità sono esprimibili da numeri reali; esempi sono: (i) il diametro del tronco di un albero, (ii) il volume di una massa tumorale.

Sempre per quanto riguarda i dati numerici si dice che si è utilizzata una

  – **scala intervallo** quando l'origine della scala stessa è arbitraria, ovvero, quando lo zero ha un'interpretazione convenzionale (esempio: la temperatura);

  – **scala rapporto** nel caso contrario ovvero quando l'origine non è arbitraria (esempio: la lunghezza di un uovo).

Per comprendere quest'ultima suddivisione, trasversale alla precedente e importante più nella fase di interpretazione dei risultati che nel momento dell'analisi, si pensi ai due esempi e si osservi che mentre possiamo dire che un uovo di 30*mm è lungo il doppio* di un uovo di 15*mm* non possiamo, viceversa, dire che quando ci sono 30° Celsius la temperatura è doppia rispetto a quando ce ne sono 15. Ad esempio, proprio per la differente origine scelta, l'affermazione sarebbe falsa se usassimo una scala Fahrenheit. Infatti 30 e 15 sulla scala Celsius corrispondono a 86 e 59 sulla scala Fahrenheit.

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).

Una variabile, due variabili, . . .

Possiamo incontrare situazioni in cui su ogni unità statistica è rilevata una sola variabile, oppure sono rilevate due variabili, oppure .... Si parla in questo caso di dati **univariati**, **bivariati**, . . . , **multivariati**[1].

Nel caso generale (m variabili rilevate, $m \geq 1$), possiamo pensare di organizzare i dati grezzi in una matrice (chiamata usualmente la **matrice dei dati**) del tipo

$$\begin{array}{c} 1^\circ\text{unità} \\ \cdots \\ i^\circ\text{unità} \\ \cdots \\ n^\circ\text{unità} \end{array} \begin{pmatrix} y_{11} & \cdots & y_{1j} & \cdots & y_{1m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{i1} & \cdots & y_{ij} & \cdots & y_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{n1} & \cdots & y_{nj} & \cdots & y_{nm} \end{pmatrix}$$

in cui ogni riga corrisponde ad una unità statistica mentre ogni colonna ad una variabile.

[1] Ovviamente "multivariato" include "bivariato"

Dati sperimentali verso 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, ad esempio, lo studio visto sulle due metodiche per la valutazione dell'emoglobina),
- e **studi osservazionali** ovvero situazioni in cui il ricercatore semplicemente rileva dei dati *già esistenti* (esempio: il numero di ricoveri per malattie legate all'asma nell'Azienda Ospedaliera di Padova).

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

Unità J

Diametro del tronco e volume del legno nei ciliegi neri: un primo modello

---

- Diagramma di dispersione
- Modello di regressione lineare semplice
- Minimi quadrati
- Covarianza
- Media e varianza residua; coefficiente di determinazione multipla (R^2).

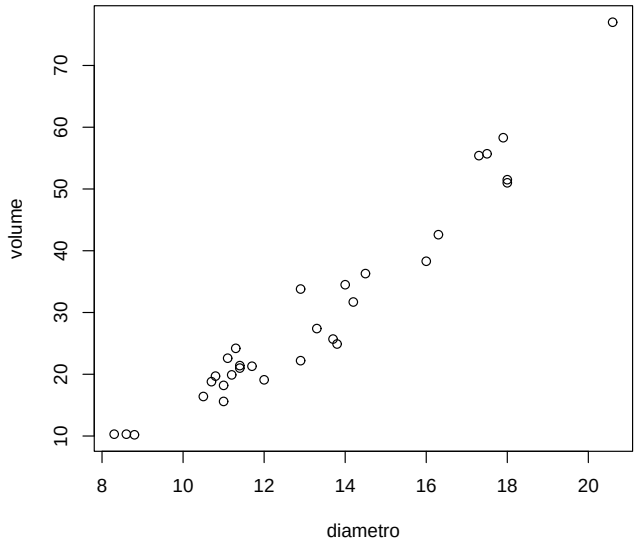
I dati

La seguente tabella mostra per 31 alberi di ciliegio nero il diametro del tronco (misurato a circa 1m dal suolo) e il volume del legno ricavato dall'albero dopo l'abbattimento.

diametro	volume	diametro	volume	diametro	volume
8,3	10,3	11,3	24,2	14,0	34,5
8,6	10,3	11,4	21,0	14,2	31,7
8,8	10,2	11,4	21,4	14,5	36,3
10,5	16,4	11,7	21,3	16,0	38,3
10,7	18,8	12,0	19,1	16,3	42,6
10,8	19,7	12,9	22,2	17,3	55,4
11,0	15,6	12,9	33,8	17,5	55,7
11,0	18,2	13,3	27,4	17,9	58,3
11,1	22,6	13,7	25,7	18,0	51,5
11,2	19,9	13,8	24,9	18,0	51,0
20,6	77,0				

Si vogliono utilizzare i dati per ottenere una equazione che permetta di prevedere il volume (ottenibile solo dopo l'abbattimento dell'albero) dal diametro (facilmente misurabile). Una simile equazione ha differenti utilizzi (dal decidere quanti e quali alberi tagliare per ricavare un certo ammontare di legno al fissare il prezzo per un bosco).

Diagramma di dispersione



Abbiamo semplicemente disegnato i punti osservati sul piano. E' evidente una forte relazione. Sostanzialmente lineare con forse un po' di problemi agli estremi.

Un primo modello

Adottiamo per il momento l'ipotesi di una relazione lineare.

Possiamo allora pensare ad un modello del tipo

$$(\text{volume}) = \alpha + \beta(\text{diametro}) + (\text{errore}) \qquad \text{(J.1)}$$

dove l'ultima componente esprime la parte delle oscillazioni del volume non legate al diametro (o, forse più precisamente, che una funzione lineare del diametro non riesce a spiegare).

Modelli di regressione lineare semplice: caso generale e terminologia

Un modello del tipo (J.1) viene usualmente chiamato **modello di regressione lineare semplice**.

Nel caso generale, cerchiamo di **spiegare** una variabile, diciamo y, utilizzando un'altra variabile, diciamo x, mediante un modello del tipo

$$y = \alpha + \beta x + \text{errore}.$$

y viene usualmente indicata come la **variabile risposta** o la **variabile dipendente** mentre x come il **regressore** o la **variabile esplicativa** o la **variabile indipendente**. α e β sono i **parametri** del modello.

Per quanto riguarda il nome, regressione viene dalla storia, lineare perchè è lineare, semplice perchè si tenta di "spiegare" la risposta utilizzando una sola variabile esplicativa.

Minimi quadrati: idea

Il problema è come "determinare" α e β. Infatti, se riusciamo a calcolare un valore "ragionevole" per questi due parametri, diciamo $\hat{\alpha}$ e $\hat{\beta}$, possiamo poi pensare di "prevedere" il volume del legno utilizzando

$$\hat{\alpha} + \hat{\beta}(\text{diametro}). \tag{J.2}$$

Sembra ragionevole cercare di "calcolare" $\hat{\alpha}$ e $\hat{\beta}$ in modo tale che la (J.2) fornisca buone "previsioni" sull'insieme di dati osservato. Al proposito, indichiamo con n il numero delle osservazioni (in questo caso $n = 31$), e poniamo y_i = (volume legno albero i-simo) e x_i = (diametro tronco albero i-simo). Quello che vorremmo è trovare dei valori per i parametri tali che

$$\begin{aligned} y_1 &\approx \hat{\alpha} + \hat{\beta} x_i \\ y_2 &\approx \hat{\alpha} + \hat{\beta} x_2 \\ &\vdots \\ y_n &\approx \hat{\alpha} + \hat{\beta} x_n \end{aligned} \tag{J.3}$$

Per rendere "operativa" la (J.3), dobbiamo decidere (i) in che senso interpretiamo gli $\approx$ che abbiamo scritto e (ii) come combiniamo tra di loro le varie linee della (J.3) stessa. La soluzione più usata si concretizza nello scegliere i due parametri minimizzando

$$s^2(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

ovvero scegliendo $\hat{\alpha}$ e $\hat{\beta}$ in maniera tale che

$$s^2(\hat{\alpha}, \hat{\beta}) \leq s^2(\alpha, \beta)$$

per qualsivoglia $\alpha \in R$ e $\beta \in R$. In questo caso si dice che "i parametri sono stati calcolati utilizzando il **metodo dei minimi quadrati**".

Minimi quadrati: determinazione dei parametri

Osserviamo, in primo luogo, che per ogni prefissato β, conosciamo già la soluzione del seguente problema

$$\inf_{\alpha \in R} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Infatti sappiamo dall'unità C che assegnati n numeri, diciamo $z_1, \ldots, z_n$, la media aritmetica delle z_i minimizza in a $\sum (z_i - a)^2$. Nel problema di minimizzazione precedente α gioca il ruolo di a e $(y_i - \beta x_i)$ quello di z_i. Quindi, per qualsivoglia β, la soluzione del problema la troviamo in corrispondenza di

$$\alpha(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta x_i) = \overline{y} - \beta \overline{x}$$

dove $\overline{y}$ e $\overline{x}$ indicano rispettivamente la media delle y_i e quella delle x_i.

Dalla definizione di $\alpha(\beta)$ segue che, per qualsivoglia α e β,

$$s^2(\alpha, \beta) \geq s^2(\alpha(\beta), \beta).$$

Quindi, $\hat{\beta}$ può essere cercato risolvendo il problema di ottimizzazione

$$\inf_{\beta \in R} s^2(\alpha(\beta), \beta)$$

mentre

$$\hat{\alpha} = \alpha(\hat{\beta})$$

Ora,

$$s^2(\alpha(\beta), \beta) = \sum_{i=1}^{n} [y_i - \overline{y} - \beta(x_i - \overline{x})]^2.$$

Derivando rispetto a β e mettendo a zero la derivata si ottiene l'equazione (per β)

$$-2 \sum_{i=1}^{n} (x_i - \overline{x})[(y_i - \overline{y}) - \beta(x_i - \overline{x})] = 0,$$

che possiamo riscrivere come

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \beta \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Se $\sum_{i=1}^{n} (x_i - \overline{x})^2 > 0$, l'equazione precedente ammette l'unica soluzione

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Lasciamo allo studente il compito di vericare che questa soluzione corrisponde ad un punto di minimo (e non, ad esempio, ad un massimo).

La soluzione trovata può quindi essere scritta compattamente come

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \quad \text{(J.4)}$$

$$\hat{\beta} = \frac{\text{cov}(X, Y)}{\text{var}(X)} \quad \text{(J.5)}$$

dove $\overline{y}$, $\overline{x}$ e var(X) sono rispettivamente la media della variabile risposta, la media e la varianza della variabile esplicativa mentre

$$\text{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

indica una quantità che usualmente viene chiamata **covarianza** tra X e Y.

Le (J.4-J.5) forniscono la soluzione del problema che ci si era proposti solamente se var$(X) > 0$. Questo è molto ragionevole: β ci dice come varia la risposta al variare della esplicativa, ma se var$(X) = 0$ l'esplicativa non è variata affatto nei dati disponibili.

Calcolo della covarianza

Per il calcolo della covarianza è conveniente utilizzare la seguente relazione

$$\text{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}.$$

ovvero

$$(\text{covarianza}) = \begin{pmatrix}\text{media dei}\\ \text{prodotti}\end{pmatrix} - \begin{pmatrix}\text{prodotto}\\ \text{delle medie}\end{pmatrix}.$$

Infatti, abbiamo

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \overline{y}) - \frac{\overline{x}}{n}\sum_{i=1}^{n}(y_i - \overline{y})$$

Il secondo addendo è nullo poichè la somma degli scarti dalla media vale zero. Espandendo il primo addendo troviamo

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{y}\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}.$$

Esercizio: Abbiamo visto una formula analoga per la varianza. Spiegare che connessione esiste tra le due formule (ad esempio, una è più generale dell'altra? Se si quale e in che senso?).

Calcolo dei parametri nel caso degli alberi di ciliegio

In questo caso,

$$\sum y_i = 935{,}3 \quad \sum x_i = 410{,}7$$
$$\sum x_i^2 = 5736{,}5 \quad \sum x_i y_i = 13887{,}86.$$

Perciò

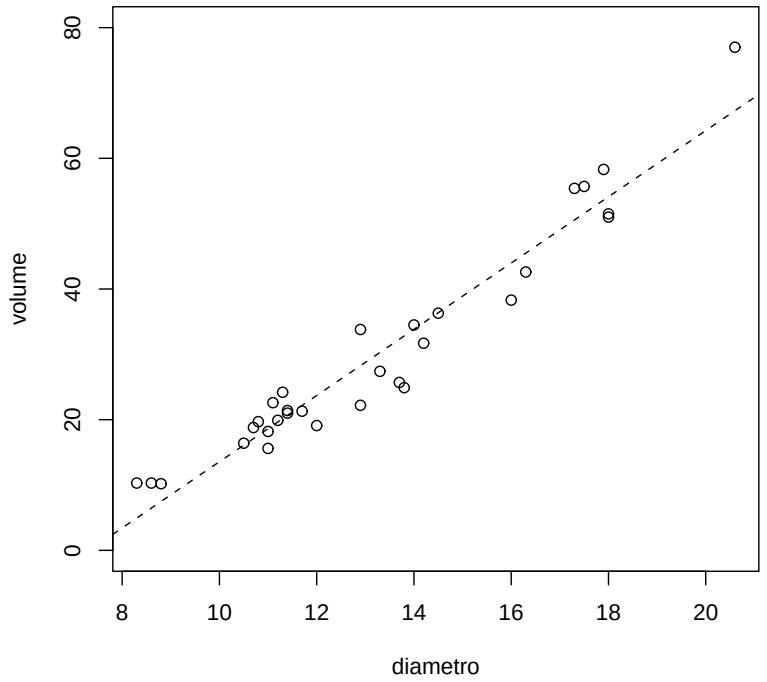
$$\overline{y} = 935{,}3/31 = 30{,}2$$
$$\overline{x} = 410{,}7/31 = 13{,}2$$
$$\text{var}(X) = (5736{,}5/31) - 13{,}2^2 = 9{,}5$$
$$\text{cov}(X, Y) = (13887{,}86/31) - 13{,}2 \times 30{,}2 = 48{,}3.$$

Quindi

$$\hat{\beta} = 48{,}3/9{,}5 = 5{,}1$$
$$\hat{\alpha} = 30{,}2 - 5{,}1 \times 13{,}2 = -37{,}1.$$

Il grafico nella pagina seguente mostra i dati osservati con la retta di regressione calcolata. La capacità di descrivere le variazione del volume sembra discreta con l'eccezione forse delle osservazioni più "esterne".

Diagramma di dispersione con retta di regressione



I residui: media e varianza

Le differenze tra i valori osservati della risposta ed i valori "previsti" dal modello, ovvero,

$$r_i = y_i - \hat{\alpha} - \hat{\beta}x_i \quad (i = 1, \ldots, n)$$

sono usualmente chiamati **residui**.

E' facile verificare che la media dei residui è nulla. Infatti

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} x_i = \\ = n\overline{y} - n(\overline{y} - \hat{\beta}\overline{x}) - n\hat{\beta}\overline{x} = 0$$

La varianza dei residui, che, per quanto appena detto, coincide con la media dei quadrati dei residui, può essere utilizzata per avere una "idea numerica" della bontà di adattamento del modello ai dati. Infatti, più la varianza dei residui sarà piccola, più la retta di regressione "spiega" le variazioni della risposta.

Si osservi che la varianza dei residui è sempre non più grande della varianza della risposta. Infatti

$$\mathrm{var}(Y) = \inf_{\alpha \in R} \sum (y_i - \alpha)^2/n \geq \\ \geq \inf_{(\alpha,\beta)\in R^2} \sum (y_i - \alpha - \beta x_i)^2/n = \mathrm{var}(r_1, \ldots, r_n).$$

Inoltre, può agevolmente essere calcolata come

$$\mathrm{var}(r_1, \ldots, r_n) = \mathrm{var}(Y) - \mathrm{cov}^2(X, Y)/\mathrm{var}(X).$$

Infatti

$$\begin{aligned} \mathrm{var}(r_1, \ldots, r_n) &= \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}[(y_i - \overline{y}) - \hat{\beta}(x_i - \overline{x})]^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 + \frac{\hat{\beta}^2}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \\ &\quad -\frac{2\hat{\beta}}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \\ &= \mathrm{var}(Y) + \hat{\beta}^2\mathrm{var}(X) - 2\hat{\beta}\mathrm{cov}(X, Y) = \\ &= \mathrm{var}(Y) + \mathrm{cov}^2(X, Y)/\mathrm{var}(X) - 2\mathrm{cov}^2(X, Y)/\mathrm{var}(X) = \\ &= \mathrm{var}(Y) - \mathrm{cov}^2(X, Y)/\mathrm{var}(X) \end{aligned}$$

Coefficiente di determinazione

E' definito come

$$R^2 = 1 - \frac{\text{var}(r_1, \ldots, r_n)}{\text{var}(Y)}$$

e quindi, come si usa dire, misura "la frazione della varianza della risposta (spesso indicata come varianza totale) spiegata dal modello". Infatti, R^2 varia tra 0 e 1, vale 0 quando $\text{var}(r_1, \ldots, r_n) = \text{var}(Y)$ ovvero quando il modello non "spiega per niente" la risposta, mentre vale 1 quando la varianza dei residui è nulla, ovvero quando il modello spiega perfettamente la risposta.

R^2 per i ciliegi neri

Abbiamo visto prima che

$\overline{y} = 935{,}3/31 = 30{,}2$
$\text{var}(X) = (5736{,}5/31) - 13{,}2^2 = 9{,}5$
$\text{cov}(X, Y) = (13887{,}86/31) - 13{,}2 \times 30{,}2 = 48{,}3.$

Inoltre

$$\sum y_i^2 = 36324{,}99$$

Quindi

$$\text{var}(Y) = 36324{,}99/31 - 30{,}2^2 = 261{,}5$$

e perciò

$$\text{var}(r_1, \ldots, r_n) = 261{,}5 - 48{,}3^2/9{,}5 = 15{,}9.$$

Il coefficiente di determinazione vale

$$R^2 = 1 - 15{,}9/261{,}5 = 0{,}94,$$

ovvero il modello spiega appena meno del 95% della varianza della risposta.

Unità K

Gli alberi sono solidi con i buchi!

Esempio di un modello linearizzabile.

Problemi

Il modello costruito nell'unità precedente *non è male* ma ha due caratteristiche *disturbanti*:

1. Non sembra cogliere in maniera del tutto appropriata l'andamento ai due estremi (le osservazioni sembrano come "curvare", il modello non lo fa);

2. Il comportamento per diametri piccoli è fisicamente insensato; l'intercetta della retta è molto diversa da zero. Viveversa, sembrerebbe naturale che per un diametro nullo il modello preveda un volume nullo. Inoltre, il valore determinato a minimi quadrati per l'intercetta è negativo. Quindi, per valori piccoli del diametro l'equazione prevede un *valore minore di zero per il volume*.

Un po' di geometria

Possiamo ragionare nella seguente maniera:

1. un albero è un "solido con dei buchi"; i buchi sono lo spazio tra i rami e, per quello che riguarda il volume di legno ricavabile, anche i rami troppo piccoli per essere utilizzati e le foglie;

2. i volumi dei solidi (almeno di quelli abbastanza regolari) sono del tipo $k \times$ (area della base) $\times$ altezza dove k dipende dalla forma del solido; indichiamo con h la "frazione del solido" non costituita da "buchi" e supponiamo che sia approssimativamente costante da albero ad albero (tutto sommato sono alberi della stessa specie); allora il volume del legno ricavabile è approssimativamente $h \times k \times$ (area della base) $\times$ altezza;

3. La "area della base" probabilmente ha a che fare con il diametro del tronco: ogni anno il tronco diventa più largo e anche l'albero complessivamente diventa più largo. Però l'area della base **non** dovrebbe variare linearmente con il diametro. Ad esempio, l'area di un cerchio varia con il quadrato del diametro. Possiamo tentare quindi di migliorare il modello precedente ipotizzando un qualche cosa del tipo

$$(\text{area base}) \approx \gamma_1 (\text{diametro})^{\gamma_2}$$

dove γ_1 e γ_2 sono delle costanti.

4. L'altezza non è nota. L'ipotesi più semplice che possiamo tentare di fare è l'albero cresca sia in altezza che in larghezza mantenendo la proporzionalità tra altezza e diametro, ovvero che (altezza) $\approx \delta$(diametro) per qualche δ.

Combinando tutti le assunzioni arriviamo al seguente modello

$$(\text{volume del legno}) \approx h \times k \times \delta \times \gamma_1 \times (\text{diametro})^{1+\gamma_2}$$

ovvero del tipo

$$(\text{volume del legno}) \approx \eta (\text{diametro})^{\lambda}$$

per appropriati valori di η e λ.

In maniera analoga a quanto fatto nell'unità J, possiamo determinare appropriati valori per η e λ minimizzando, nei due parametri, la somma dei quadrati degli scarti (differenze) tra i volumi osservati e i valori previsti dal modello, ovvero, minimizzando in η e λ

$$\sum_{i=1}^{n} \left[\begin{pmatrix} \text{volume albero} \\ i\text{-simo} \end{pmatrix} - \eta \begin{pmatrix} \text{diametro albero} \\ i\text{-simo} \end{pmatrix}^{\lambda} \right]^2$$

Questo problema di minimo non ammette una soluzione in forma chiusa. Può essere risolto utilizzando appropriate tecniche *numeriche*. Esiste però una alternativa più semplice che verrà illustrata nelle prossime pagine.

Linearizzazione del modello

Tenendo presente che $\log(\cdot)$ è una funzione continua, se

$$(\text{volume}) \approx \eta(\text{diametro})^{\lambda}$$

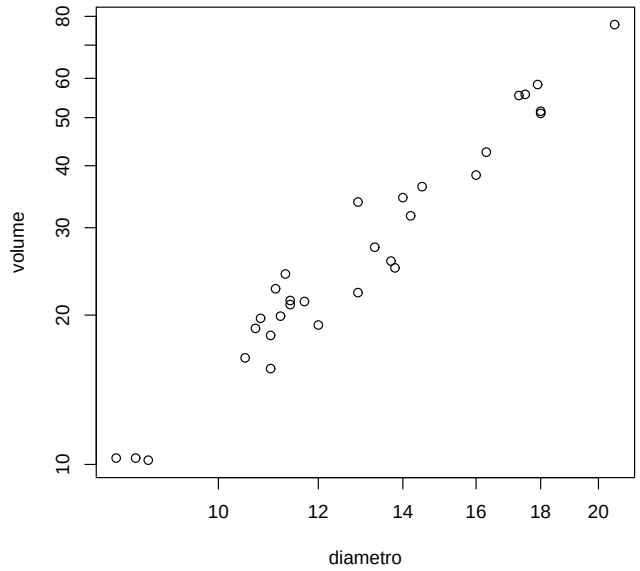
allora dovrebbe risultare

$$\log(\text{volume}) \approx \log(\eta) + \lambda \log(\text{diametro}) \qquad \text{(K.1)}$$

ovvero, la relazione *non lineare* tra (volume) e (diametro) dovrebbe corrispondere ad una relazione *lineare* tra i logaritmi delle due variabili.

Nella pagina seguente le due variabili (volume e diametro) sono disegnate su di un grafico con *scala logaritmica*. E la relazione infatti sembra lineare. Agli estremi più lineare di quanto indicava il grafico (volume) verso (diametro) non trasformati mostrato nell'unità J.

Diagramma di dispersione su scala logaritmica



Il grafico mostra il logaritmo del volume disegnato verso il logaritmo del diametro. Gli assi però sono *etichettati* utilizzando i valori delle variabili originali. Si osservi in particolare l'asse delle y dove il tutto è più evidente.

La (K.1) descrive un modello di regressione lineare semplice in cui la variabile risposta è il logaritmo del volume e la variabile esplicativa è il logaritmo del diametro. In particolare, ponendo

$$z = \log(\text{volume}),\ w = \log(\text{diametro}),$$

ed introducendo esplicitamente il termine di errore, possiamo riscrivere la (K.1) come

$$z = \alpha + \beta w + (\text{errore}).$$

dove[1]

$$\alpha = \log(\eta),\ \beta = \lambda. \qquad \text{(K.2)}$$

Possiamo quindi pensare di determinare α e β nella maniera descritta nell'unità J, ovvero utilizzando il metodo dei minimi quadrati applicati questa volta ai logaritmi delle variabili, e poi di calcolare η e λ *invertendo* le (K.2).

[1]Stiamo supponendo che $\eta > 0$. Ma questo non è assolutamente restrittivo. In caso contrario, il nuovo modello "stimerebbe" sempre un volume negativo

Calcolo dei parametri (modello linearizzato)

Tutti i calcoli sono basati sui logaritmi dei dati.

$$\sum z_i = 101.45,\ \sum z_i^2 = 340.34,$$
$$\sum w_i = 79.28,\ \sum w_i^2 = 204.37,$$
$$\sum z_i w_i = 263.06.$$

Perciò

$$\overline{z} = 101{,}45/31 = 3{,}27$$
$$\text{var}(Z) = (340{,}34/31) - 3{,}27^2 = 0{,}27$$
$$\overline{w} = 79{,}28/31 = 2{,}56$$
$$\text{var}(W) = (204{,}37/31) - 2{,}27^2 = 0{,}05$$
$$\text{cov}(W, Z) = (263{,}06/31) - 3{,}27 \times 2{,}56 = 0{,}12.$$

Quindi

$$\hat{\beta} = 0{,}12/0{,}05 = 2{,}20$$
$$\hat{\alpha} = 3{,}27 - 2{,}20 \times 2{,}56 = -2{,}35.$$
$$\text{var. residua} = 0{,}27 - 0{,}12^2/0{,}05 = 0{,}012$$
$$R^2 = 1 - 0{,}012/0{,}27 = 0{,}95$$

Ritorniamo alla scala originale

Ritornando sulla scala dei dati originali, abbiamo

$$\hat{\eta} = \exp(\hat{\alpha}) = 0{,}10,$$
$$\hat{\lambda} = \hat{\beta} = 2{,}20.$$

Ha inoltre senso, più che considerare il valore della varianza residua e dell R^2 della pagina precedente che sono riferiti agli errori commessi dal modello nella scala trasformata, calcolare come indice di adattamento la media dei quadrati dei residui per la vera variabile che si vuole prevedere, ovvero

$$\frac{1}{n}\sum_{i=1}^{n}\left[\begin{pmatrix}\text{volume}\\ i\text{-simo}\end{pmatrix} - \hat{\eta}\begin{pmatrix}\text{diametro}\\ i\text{-simo}\end{pmatrix}^{\hat{\lambda}}\right]^2$$

che, calcolata con i dati disponibili, vale 10,2. Questa quantità è direttamente confrontabile con la varianza residua per il modello lineare calcolata nell'unità J e che valeva 15,9. Quindi, possiamo dire che il nuovo modello riduce il quadrato degli *errori di previsione* di circa il 30%.

Anche da un punto di vista puramente grafico il nuovo modello sembra migliore, in particolare, perchè "coglie" la curvatura agli estremi che avevamo osservato.

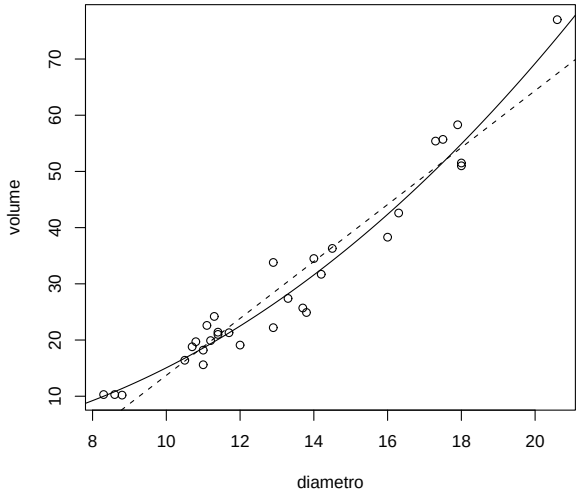


Diagramma di dispersione con (i) retta di regressione (linea tratteggiata) e (ii) curva descritta dai valori previsti dal nuovo modello (linea continua).

Commenti

1. I metodi basati sui minimi quadrati descritti nell'unità J possono essere applicati non solo a modelli del tipo

$$y = \alpha + \beta x + (\text{errore})$$

ma anche a modelli, più generali, ad esempio del tipo

$$g(y) = \alpha + \beta h(x) + (\text{errore})$$

dove $g(\cdot)$ e $h(\cdot)$ sono appropriate funzioni. Quello che è importante è che, come si usa dire, *il modello sia lineare nei parametri non nelle variabili*. Ad esempio, risulta trattabile senza problemi un modello del tipo

$$y = \alpha + \beta \sin^{27}(x) + (\text{errore})$$

2. Spesso modelli lineari nelle variabili possono essere visti al più come approssimazioni di relazioni non lineari (si pensi alla formula di Taylor). In questo situazioni, ottenere estrapolazioni dal modello è pericoloso e può dare luogo a risultati insensati (nel caso considerato, *previsioni* negative per il volume).

3. Non bisogna mai *buttare via* quello che si sa. Ad esempio, in questo caso, poche conoscenze di geometria al livello III media ci hanno portato ad un modello che sembra adattarsi meglio ai dati osservati e sopratutto che è più ragionevole da un punto di vista fisico. In generale, lo statistico ha il **dovere** di recuperare le conoscenze sul fenomeno che sta analizzando. Inoltre, e' spesso utile (e, tra l'altro, quasi sempre "divertente") che lo statistico "vada sul campo" (nel laborario, nello stabilimento di produzione,...) per "vedere dal vivo" come i dati sono effettivamente raccolti.

Unità L

Agricoltura, fertilità ed istruzione nella Svizzera francese del 1888

- Interpretazione della covarianza
- Coefficiente di correlazione
- Matrice delle varianze e covarianze (di dispersione)
- Matrice di correlazione

I dati

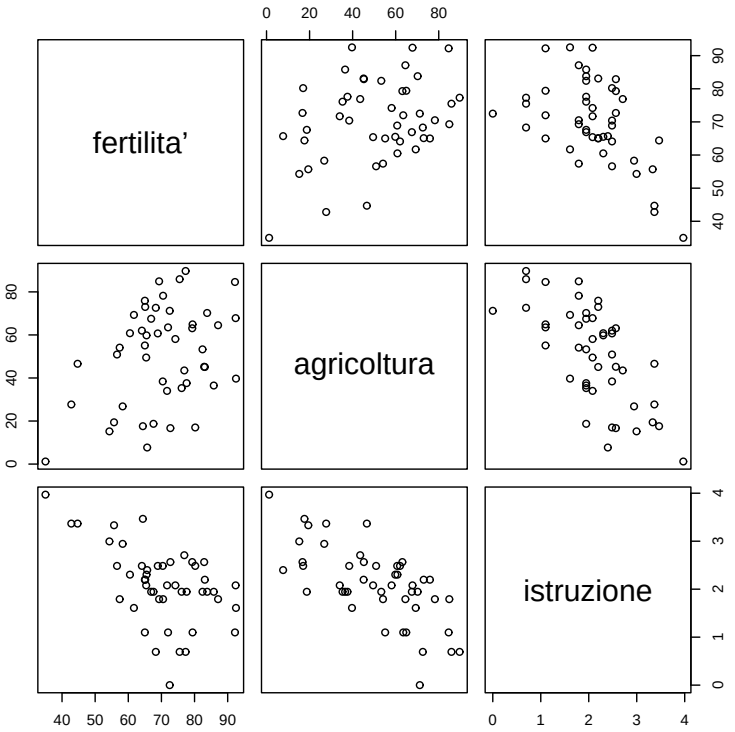
Consideriamo in questa unità tre indicatori *socio-economici* disponibili per 47 province svizzere di lingua francese. I dati sono storici. Sono infatti riferiti al 1888. I tre indicatori considerati sono

1. una misura di fertilità (nati per donna) standardizzata in maniera tale che vari tra 0 e 100.
2. la percentuale degli occupati in agricoltura sul totale degli occupati (che può anche essere vista come un indicatore di quanto è *urbanizzata* la provincia).
3. il logaritmo della percentuale della popolazione con un istruzione superiore alla scuola primaria (il logaritmo è stato scelto perchè ci occuperemmo di relazioni lineari e la linearità sembra maggiore utilizzando questa trasformazione).

Il problema che ci poniamo è di cercare di descrivere le relazioni esistenti tra i tre indicatori.

fertilita'	agricoltura	educazione	fertilita'	agricoltura	educazione
80,2	17,0	2,485	57,4	54,1	1,792
83,1	45,1	2,197	74,2	58,1	2,079
92,5	39,7	1,609	72,0	63,5	1,099
85,8	36,5	1,946	60,5	60,8	2,303
76,9	43,5	2,708	58,3	26,8	2,944
76,1	35,3	1,946	65,4	49,5	2,079
83,8	70,2	1,946	75,5	85,9	0,693
92,4	67,8	2,079	69,3	84,9	1,792
82,4	53,3	1,946	77,3	89,7	0,693
82,9	45,2	2,565	70,5	78,2	1,792
87,1	64,5	1,792	79,4	64,9	1,099
64,1	62,0	2,485	65,0	75,9	2,197
66,9	67,5	1,946	92,2	84,6	1,099
68,9	60,7	2,485	79,3	63,1	2,565
61,7	69,3	1,609	70,4	38,4	2,485
68,3	72,6	0,693	65,7	7,7	2,398
71,7	34,0	2,079	72,7	16,7	2,565
55,7	19,4	3,332	64,4	17,6	3,466
54,3	15,2	2,996	77,6	37,6	1,946
65,1	73,0	2,197	67,6	18,7	1,946
65,5	59,8	2,303	35,0	1,2	3,970
65,0	55,1	1,099	44,7	46,6	3,367
56,6	50,9	2,485	42,8	27,7	3,367
72,5	71,2	0,000			

Disegnamo i dati



Il grafico mostra la *matrice* dei diagrammi di dispersione di tutte le possibili coppie di variabili.

Commenti

I grafici precedenti mostrano che:

- percentuale di occupati in `agricoltura` e `fertilità` sono "positivamente associati": provincie con una alta percentuale di occupati in agricoltura hanno anche una alta fertilità, viceversa, basse percentuali di occupati in agricoltura si osservano in provincie con bassi livelli di fertilità;
- esiste una "associazione negativa" tra `istruzione` e `fertilità`; ovvero provincie con un alto livello di istruzione hanno una fertilità più bassa delle provincie con un basso livello di istruzione.
- lo stesso (associazione negativa) può essere detto per la relazione tra `agricoltura` e `istruzione`
- almeno in prima approssimazione le relazioni sembrano lineari.

- la relazione tra `agricoltura` e `fertilità` sembra più debole della relazione esistente tra `agricoltura` ed `istruzione` (si pensi, ad esempio, alla dispersione intorno a delle ipotetiche rette di regressione). Meno facile è valutare la forza relativa delle relazioni intercorrenti tra `istruzione` e, rispettivamente, `agricoltura` e `fertilità`. La prima (`istruzione` verso `agricoltura`) sembra però in una qualche misura più forte della seconda (si osservi in particolare l'*allargarsi* del grafico `fertilità` verso `istruzione` per valori bassi dell'`istruzione`).

- in situazioni tipo quella che stiamo considerando può essere interessante essere in grado di descrivere compattamente sia la direzione che la forza delle relazioni intercorrenti tra le varie variabili.

La covarianza come misura della direzione e della forza della relazione tra due variabili

Ricordiamo che nella unità J abbiamo visto che la covarianza è definita come

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{k} (x_i - \overline{x})(y_i - \overline{y}) \qquad \text{(L.1)}$$

dove (x_i, y_i), $i = 1, \ldots, n$, sono i dati disponibili su due variabili *numeriche* mentre $\overline{x}$ e $\overline{y}$ indicano le due medie aritmetiche.

Si osservi che

1. Nei caso in cui a valori crescenti di X corrispondano valori crescenti di Y, ci aspettiamo che valori maggiori della media di X corrispondano a valori maggiori della media per Y; in questo caso quindi la covarianza risulterà positiva.

2. Completamente simmetrico è quello che accade nel caso in cui al crescere della X la Y tendenzialmente descresce. Quindi, in questo caso, ci aspettiamo una covarianza negativa.

3. Più è forte la relazione tra le due variabili più ci aspettiamo che la covarianza diventi grande in valore assoluto. Infatti, più è forte la relazione più il numero di addendi concordi nella (L.1) dovrebbe crescere ed inoltre un certo numero di addendi sarà il prodotto di scarti dalle media grandi in valore assoluto.

4. In assenza di una qualche forma di relazione *monotona* tra le due variabili, viceversa, gli addendi della (L.1) saranno in parte positivi ed in parte negativi. Quindi in questi casi ci aspettiamo che la covarianza risulti nulla o comunque vicina allo zero.

Le considerazioni precedenti suggerisono l'uso della covarianza per *misurare* la direzione e la forza delle relazioni esistenti tra le variabili (quantomeno monotone, in realtà come vedremo essenzialmente lineari).

La matrice delle varianze e covarianze

Nel caso in esame, troviamo,

$$\begin{aligned} \text{cov}(\texttt{fertilità},\texttt{agricoltura}) &= 98{,}0, \\ \text{cov}(\texttt{fertilità},\texttt{istruzione}) &= -5{,}1, \\ \text{cov}(\texttt{agricoltura},\texttt{istruzione}) &= -11{,}9. \end{aligned}$$

Tipicamente, le covarianze per tutte le coppie di variabili vengono organizzate, insieme alle varianze delle singole variabili, in una matrice, chiamata **matrice delle varianze e covarianze** o **matrice di dispersione**, che nel caso in esame sarebbe

$$\begin{pmatrix} 152{,}7 & 98{,}0 & -5{,}1 \\ 98{,}0 & 504{,}8 & -11{,}9 \\ -5{,}1 & -11{,}9 & 0{,}61 \end{pmatrix}. \qquad \text{(L.2)}$$

Nella (L.2), l'elemento (i, j) è dato dalla covarianza tra la variabile i-sima e la variabile j-sima. Poichè, come è immediato verificare, $\text{cov}(Y, Y) = \text{var}(Y)$, sulla diagonale troviamo le varianze. Ad esempio, in questo caso 157,7 è la varianza della `fertilità`. Si osservi che, poichè $\text{cov}(X, Y) = \text{cov}(Y, X)$, la matrice di dispersione è per costruzione simmetrica.

Grande quanto?

L'esempio illustra uno dei problemi connessi con l'utilizzo della covarianza.

L'interpretazione del segno non pone nessuno problema. Le covarianze riportate ci indicano una relazione tendenzialmente crescente tra `fertilità` ed `agricoltura` ed una relazione tendenzialmente decrescente tra queste due variabili e l'`istruzione`.

Però, ad esempio, che cov(`fertilità`,`agricoltura`) risulti uguale a 98,0 indica un debole od un forte legame tra le due variabili? Per rispondere alla domanda avremmo bisogno di conoscere un estremo superiore, possibilmente con una chiara interpretazione, per il valore assoluto della covarianza.

Fortunatamente la risposta praticamente la conosciamo già. Infatti sappiamo che il coefficiente di determinazione introdotto nell'unità J è sempre minore od uguale a 1, ovvero sappiamo che

$$R^2 = 1 - \frac{\text{var}(Y) - \text{cov}^2(X,Y)/\text{var}(X)}{\text{var}(Y)} \leq 1.$$

Quindi

$$1 - \frac{\text{var}(Y)}{\text{var}(Y)} + \frac{\text{cov}^2(X,Y)}{\text{var}(Y)\text{var}(X)} \leq 1,$$

che, semplificando, diventa

$$\frac{\text{cov}^2(X,Y)}{\text{var}(Y)\text{var}(X)} \leq 1,$$

ovvero

$$\text{cov}^2(X,Y) \leq \text{var}(Y)\text{var}(X).$$

Ricordando che lo scarto quadratico medio è la radice quadrata della varianza e scrivendo l'ultima diseguaglianza in termini della covarianza e non del suo quadrato troviamo infine

$$-\text{sqm}(Y)\text{sqm}(X) \leq \text{cov}(X,Y) \leq \text{sqm}(Y)\text{sqm}(X).$$

Il coefficiente di correlazione (lineare)

I limiti trovati per la covarianza suggeriscono che per affermare se la covarianza è "piccola" o è "grande" dobbiamo confrontarla con il prodotto degli scarti quadratici medi.

Per semplificare il lavoro, è usuale presentare i risultati utilizzando non direttamente la covarianza ma una sua versione normalizzata nota come **coefficiente di correlazione (lineare)**[1] e definito come

$$\mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\mathrm{sqm}(X)\mathrm{sqm}(Y)}.$$

Ci si ricordi che il coefficiente di correlazione è spesso indicato con la lettera `r` (erre minuscola).

[1] Il "lineare" tra parantesi indica che a volte l'aggettivo è omesso

Coefficienti di correlazione delle tre variabili considerare. La matrice di correlazione

Nel caso in esame, è immediato dalla matrice delle varianze e covarianze data, calcolare

$$\begin{aligned} \mathrm{cor}(\texttt{fertilità},\texttt{agricoltura}) &= 0{,}35, \\ \mathrm{cor}(\texttt{fertilità},\texttt{istruzione}) &= -0{,}52, \\ \mathrm{cor}(\texttt{agricoltura},\texttt{istruzione}) &= -0{,}68. \end{aligned}$$

Similmente a quanto visto per le covarianze spesso, i coefficienti di correlazione sono organizzati in una matrice, detta **matrice di correlazione**, del tipo

$$\begin{pmatrix} 1{,}00 & 0{,}35 & -0{,}52 \\ 0{,}35 & 1{,}00 & -0{,}68 \\ -0{,}52 & -0{,}68 & 1{,}00 \end{pmatrix}$$

dove l'elemento (i, j) è dato dal coefficiente di correlazione tra la i-sima e la j-sima variabile. Si osservi, che $\mathrm{cor}(Y, Y) = 1$ come è facile verificare. Questo spiega la diagonale della matrice precedente.

Interpretazione di cor(X, Y)

Per quanto detto, il coefficiente di correlazione varia tra -1 e 1. La sua interpretazione è a grande linee la seguente.

Se $\boxed{\text{cor}(X, Y) < 0}$ allora i dati indicano una associazione negativa tra le due variabili (al crescere di una l'altra decresce). Questa associazione è man mano più forte più cor(X, Y) si avvicina a -1. Se $\boxed{\text{cor}(X, Y) = -1}$ allora i dati sono perfettamente allineati su di una retta con coefficiente angolare negativo.

Se $\boxed{\text{cor}(X, Y) = 0}$, ed in realtà da un punto di vista pratico, se cor$(X, Y) \approx 0$, allora non esiste una relazione lineare (e più in generale una associazione monotona) tra le due variabili.

Se $\boxed{\text{cor}(X, Y) > 0}$ l'interpretazione è simmetrica a quando detto per il caso "coefficiente di correlazione negativo". La relazione è crescente e se $\boxed{\text{cor}(X, Y) = 1}$ i dati sono perfettamente allineati su di una retta con coefficiente angolare positivo.

Due limiti di cor(X, Y) da tenere presente

- Dati posti perfettamente su di una curva *monotona*, pensiamola crescente, ma *non lineare* indicano una dipendenza perfetta tra le due variabili ma non risultano in cor$(X, Y) = 1$. Si pensi ad esempio, a dei dati posti sulla curva $Y = \exp(X)$. In questo caso, i dati non risulteranno allineati. Quindi la varianza dei residui nella regressione di Y su X non sarà nulla e quindi risulterà cor$(X, Y) < 1$. In definitiva, il coefficiente di correlazione misura *accuratamente* la forza della relazione esistente solo se questa è lineare. Osservando che $\text{cor}^2(X, Y) = R^2$ possiamo dire che, in generale, misura la parte della relazione spiegabile in termini lineari.

- $\text{cor}(X, Y) = 0$ non implica che non esiste nessuna relazione tra X e Y. Ad esempio, lo studente verifichi che se le coppie di dati sulle due variabili sono $(x_1, y_1) = (-2, 4)$, $(x_2, y_2) = (1, -1)$, $(x_3, y_3) = (0, 0)$, $(x_4, y_4) = (1, 1)$ e $(x_5, y_5) = (2, 4)$ allora $\text{cor}(X, Y) = 0$. Nonostante questo però i dati sono esattamente posti sulla parabola $Y = X^2$. Il fatto è che il coefficiente di correlazione è, per costruzione, inutile nel valutare l'esistenza e la forza di relazioni non monotone.

Regressione e correlazione

Le analogie tra l'unità J, dove abbiamo iniziato a studiare la regressione, e questa sono molte. Il problema di base è lo stesso (studio delle relazioni tra variabili). Gli "ingredienti" che abbiamo maneggiato pure (medie, varianze e covarianze). La trattazione è stata in parte sovrapposta.

Nonostante questo si noti che esiste una differenza:

- Nell'unità J (primi rudimenti sulla regressione) ci siamo posti il problema di modellare l'effetto di una possibile variabile esplicativa su una variabile risposta. Trattavamo le variabili in maniera *asimmetrica* visto che pensavamo ad una relazione con una precisa direzione del tipo `diametro` $\rightarrow$ `volume`.
- Viceversa in questa unità (primi rudimenti sulla correlazione) ci siamo posti in maniera *simmetrica* rispetto alle variabili. Non abbiamo cercato di spiegarne una sulla base di un altra ma abbiamo semplicemente cercato di valutare le relazioni intercorrenti.

Unità M

Ancora sulla Svizzera francese del 1888

Cenno alla correlazione parziale

Una congettura

Si supponga che un sociologo faccia le seguenti ipotesi sulle relazioni intercorrenti tra i tre indicatori socio-economici[1].

1. Tra `agricoltura` e `istruzione` esiste una sostanziale interdipendenza. Nelle province "molto agricole" i bimbi vanno meno a scuola perchè servono braccia per lavorare i campi, l'istruzione è percepita come inutile per fare il contadino, la minore urbanizzazione rende più difficile il raggiungimento della scuola stessa,.... Quindi le province "molto agricole" rimangono associate a bassi livelli di istruzione. Dall'altra parte, possiamo pensare che un buon livello di `istruzione` faciliti la "transizione" verso attività secondarie e terziarie. Quindi, esiste anche un effetto, diciamo di ritorno, dall'`istruzione` all'`agricoltura`.

[1]Per i dati e la definizione delle variabili si veda l'unità precedente

2. L'`istruzione` ha un effetto diretto sulla `fertilità`. Coppie con buona scolarità vogliono (e riescono a) controllare la natalità. Simultaneamente, famiglie con pochi figli hanno più disponibilità di reddito e quindi sono più "portate" a mandare i figli a scuola.

3. Per quanto riguarda la `fertilità` province "molto agricole e colte" si comportano come le "province poco agricole e colte". E simultaneamente province "molto agricole e poco colte" si comportano come le province "poco agricole e poco colte". Ovvero, non esiste nessuna relazione diretta tra `agricoltura` e `fertilità`. La relazione osservata precedentemente è, come si dice comunemente, **spuria**. E' una conseguenza delle relazioni descritte ai punti 1 e 2 precendenti. Ovvero osservo una associazione positiva tra `agricoltura` e `fertilità` semplicemente perchè "tanta agricoltura" risulta in "bassa istruzione" e "bassa istruzione" risulta in "alta fertilità".

In definitiva, il tipo di relazione che il sociologo ipotizza tra le tre variabili può essere rappresentato schematicamente come

$$\texttt{agricoltura} \Longleftrightarrow \texttt{istruzione} \Longleftrightarrow \texttt{fertilità}$$

dove le freccie indicano un effetto diretto. Il punto cruciale della congettura è l'inesistenza di una freccia che metta in relazione diretta `agricoltura` e `fertilità` senza "passare" per l'`istruzione`.

Il problema è: cosa possiamo fare per dire se i dati disponibili "votano" a favore o contro la congettura. Ovvero, cosa possiamo fare per capire se *eliminata* la dipendenza tra `agricoltura` e `fertilità` attribuibile alle relazione che ambedue le variabili hanno con `istruzione` rimane ancora qualcosa.

Reinterpretazione della congettura

Una possibile interpretazione è che possiamo guardare a `agricoltura` e `fertilità` come divise in due parti

$$\texttt{agricoltura} = \begin{pmatrix}\text{parte } \textbf{legata}\\ \text{all'}\texttt{istruzione}\end{pmatrix} + \begin{pmatrix}\text{parte } \textbf{non}\\ \textbf{legata}\\ \text{all'}\texttt{istruzione}\end{pmatrix},$$

$$\texttt{fertilità} = \begin{pmatrix}\text{parte } \textbf{legata}\\ \text{all'}\texttt{istruzione}\end{pmatrix} + \begin{pmatrix}\text{parte } \textbf{non}\\ \textbf{legata}\\ \text{all'}\texttt{istruzione}\end{pmatrix},$$

e che la congettura postuli l'inesistenza di una relazione tra le due *parti* ***non legate*** all'`istruzione`.

Accettata questa interpretazione della congettura, possiamo allora *verificarla* "estraendo" le due *parti* ***non legate*** all'`istruzione` e studiandone le relazioni.

Attuazione pratica del programma precedente

Se assumiamo che le relazioni intercorrenti tra le variabili sono lineari possiamo procedere nella seguente maniera:

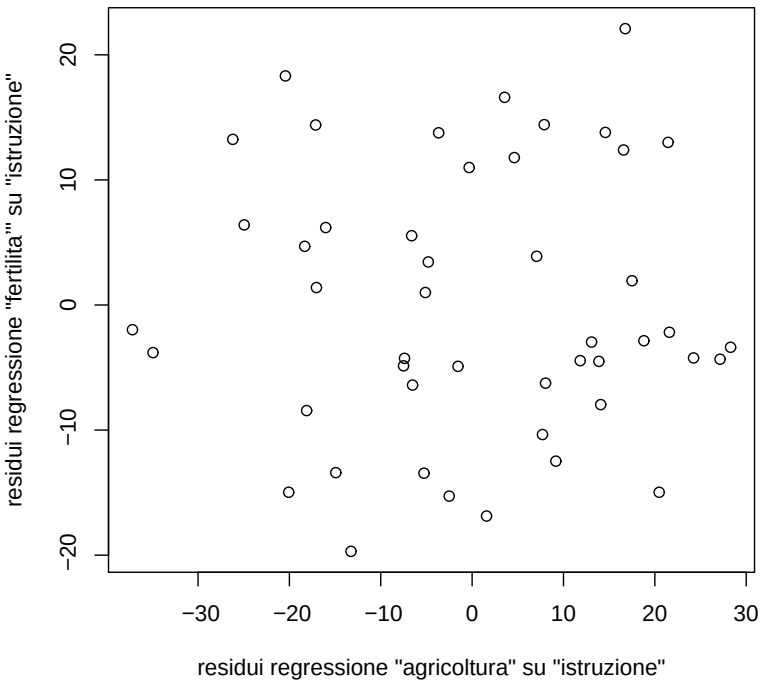
1. Determiniamo la retta di regressione di `agricoltura` su `istruzione`. Ovvero, costruiamo un modello di regressione lineare semplice in cui `agricoltura` è la variabile risposta e `istruzione` è la variabile esplicativa. Possiamo poi identificare con i residui del modello[2] la parte dell'`agricoltura` non legata all'`istruzione`.
2. In maniera analoga, "estraiamo" la parte della `fertilitá` non legata all'`istruzione` calcolando i residui di un modello di regressione lineare semplice in cui `fertilitá` gioca il ruolo di variabile risposta e `istruzione` quello di variabile esplicativa.
3. Infine, valutiamo la relazione esistente tra le due parti non legate all'`istruzione` semplicemente calcolando il coefficiente di correlazione tra i residui calcolati ai passi 1 e 2.

Nella terminologia statistica, il coefficiente suggerito al passo 3 viene usualmente chiamato il **coefficiente di correlazione parziale** tra `agricoltura` e `fertilitá` **data** l'`istruzione`.

Calcolato con i dati disponibili vale −0,0021. E' molto vicino allo zero e quindi ci indica che tra i residui dei due modelli di regressione non esiste una relazione lineare importante. In realtà come ci mostra la figura della pagina seguente, tra i residui dei due modelli non sembra esistere nessuna relazione rilevante. La conclusione è quindi che i dati sembrano andare d'accordo con la congettura fatta.

[2] Per la definizione di residui si veda l'unità J

Diagramma di dispersione dei residui dei due modelli di regressione



Esercizio

Siano X, Y e Z tre variabili numeriche rilevate sulle stesse unità statistiche. Si indichino $(x_1, \ldots, x_n)$ i dati su X e si proceda in maniera analoga per Y e Z.

Si dimostri allora che

$$\mathrm{cov}(\tilde{X}, \tilde{Y}) = \mathrm{cov}(X, Y) - \frac{\mathrm{cov}(X, Z)\mathrm{cov}(Y, Z)}{\mathrm{var}(Z)} \quad \text{(M.1)}$$

dove $\tilde{X}$ e $\tilde{Y}$ indicano i residui della regressione di, rispettivamente, X su Z e Y su Z, ovvero, per qualsivoglia $i = 1, \ldots, n$

$$\begin{aligned} \tilde{x}_i &= x_i - \overline{x} - \frac{\mathrm{cov}(X, Z)}{\mathrm{var}(Z)}(z_i - \overline{z}) \\ \tilde{y}_i &= y_i - \overline{y} - \frac{\mathrm{cov}(Y, Z)}{\mathrm{var}(Z)}(z_i - \overline{z}). \end{aligned}$$

Commento anche per chi non fa l'esercizio: La (M.1) permette, insieme alle formule viste nell'unità J per il calcolo della varianza dei residui in un modello di regressione lineare semplice, di calcolare agevolmente il coefficiente di correlazione parziale tra X e Y dato Z una volte che siano note le varianze e le covarianze delle tre variabili.

Unità N

Il disastro del Titanic

- Tabelle di contingenza.
- Distribuzioni congiunta, marginali e condizionate.
- Indipendenza in distribuzione
- Frequenza attese. X^2 di Pearson.

Alcuni dati sul 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: l'esito (salvato, non salvato), la classe (I, II, III) in cui viaggiavano, il sesso, l'età,. . . .

In questa unità 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	Esito
nome 1	II	salvato
nome 2	III	non salvato
nome 3	I	non salvato
⋮	⋮	⋮
nome 1316	III	salvato

Frequenze

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, ad esempio, 203 è il numero di passeggeri che viaggiavano in I classe e che sono sopravissuti al disastro.

La prima cosa che salta agli occhi è che i viaggiatori della I classe hanno ricevuto al momento del disastro un *trattamento preferenziale*. Ad esempio, la frazione degli individui che si sono salvati è 203/325, ovvero approssimativamente il 63% se consideriamo i viaggiatori della I classe ma scende a 178/706, ovvero più o meno al 25%, per la III classe.

Tabelle di contingenza

Una tabella del tipo visto viene usualmente chiamata di **contingenza** (un altro esempio l'abbiamo visto quando abbiamo parlato della *litotripsia*). In generale, una tabella di contingenza mostra la distribuzione delle unità statistiche classificate sulla base di due o più variabili.

Si osservi che una tabella di contingenza *contiene* varie distribuzioni di frequenza. Infatti:

- Se consideriamo il "cuore" della tabella (in questo caso le 2 righe e le 3 colonne centrali) la tabella ci mostra il numero di individui che presentano una particolare modalità della prima variabile **congiuntamente** ad una particolare modalità della seconda variabile. Ad esempio, 122/1316 è la frazione di passeggeri che *simultaneamente* viaggiavano in I classe **e** sono periti nel disastro.

- Se concentriamo l'attenzione sulla 1° colonna, vediamo l'esito del disastro per i passeggeri della I classe. Ad esempio, 122/325 è la frazione di viaggiatori della prima classe periti nel disastro. Un discorso analogo possiamo fare per la 2° e per la 3° colonna. Quindi, queste colonne mostrano l'esito del disastro **condizionatamente** al fatto di considerare *solamente* individui che viaggiavano in una *determinata* classe.

- Se, viceversa ci concentriamo sulla 1° (o sulla 2°) riga, vediamo la distribuzione tra le varie classi dei viaggiatori sopravvissuti (o non sopravvissuti) al disastro. Ad esempio, guardando alla 2° riga possiamo fare affermazioni del tipo "il 15% ($\approx (100\times122)/817$) dei passegeri periti nel disastro viaggiava in I classe". Ovvero, guardiamo alla classe **condizionatamente** all'esito.

- l'ultima colonna (riga), invece, mostra la distribuzione dei passeggeri rispetto all'esito (alla classe) a prescindere dall'altra variabile. Possiamo quindi fare affermazioni del tipo "solo il 38% (= (100 × 499)/1316) dei passeggeri del Titanic è sopravvissuto all'incidente". Si osservi che in questo caso guardiamo a tutti i passeggeri del Titanic non a quelli che viaggiavano in I (o in II o in III) classe.

Struttura generale

In generale, una tabella di contingenza (con due variabili) si presenta nella forma

			X			
Y	x_1	$\cdots$	x_j	$\cdots$	x_c	totale
y_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1c}	f_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ic}	f_{i+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rc}	f_{r+}
totale	f_{+1}	$\cdots$	f_{+j}	$\cdots$	f_{+c}	n

dove (i) X e Y sono le due variabili considerate, (ii) $\{x_1, \ldots, x_c\}$ e $\{y_1, \ldots, y_r\}$ indicano le modalità rispettivamente di X e di Y, (iii) f_{ij} è il numero di unità statistiche che presentano simultaneamente la modalità x_j di X e la modalità y_i di Y, (iv) f_{+j}, $j = 1, \ldots, c$, e f_{i+}, $i = 1, \ldots, r$ sono i totali rispettivamente delle colonne e delle righe, ovvero, $f_{+j} = \sum_{i=1}^{r} f_{ij}$ e $f_{i+} = \sum_{j=1}^{c} f_{ij}$.

Un po' di terminologia

Per le varie distribuzioni di frequenza evidenziate nel caso dei dati sul Titanic, sono in uso i seguenti termini (evidenziati in grassetto):

• La tabella nel suo complesso ci mostra la **distribuzione congiunta** di X e Y. Le f_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, c$, sono chiamate le **frequenze congiunte**.

• La j-sima colonna mostra la **distribuzione di Y condizionata a** $X = x_j$ o, equivalentemente, la **distribuzione di Y dato** $X = x_j$. Può essere indicata compattamente con la notazione $(Y|X = x_j)$ dove | si legge "dato". Si osservi che esiste una distribuzione condizionata (di Y dato X) per ogni modalità di X.

• In maniera simmetrica, la i-sima riga mostra la **distribuzione di X condizionata a** $Y = y_i$ o, equivalentemente, la **distribuzione di X dato** $Y = y_i$. Può essere indicata compattamente con la notazione $(X|Y = y_i)$.

• L'ultima colonna (riga) viene chiamata la **distribuzione marginale** di Y (X). Ci fornisce la distribuzione di Y (X) a prescindere da X (Y).

Dipendenza, indipendenza e distribuzioni condizionate

Riguardiamo la tabella sul disastro del Titanic. Abbiamo notato che l'esito **dipende** dalla classe in cui viaggiava il passeggero visto che la frazione di sopravvissuti all'incidente varia al variare della classe.

Indichiamo con X la classe (I, II, III) e con Y l'esito (sopravissuto, non sopravissuto). Allora, usando la terminologia appena introdotta, una affermazione sostanzialmente analoga a quella contenuta nel precedente paragrafo è:

> Poichè le distribuzioni di Y condizionate ad X sono tra di loro diverse, Y **dipende** da X

L'affermazione va nella direzione giusta. Deve però essere precisata meglio.

Supponiamo infatti, per un momento, che la distribuzione congiunta non sia quella già mostrata ma, viceversa, la seguente

	classe			
salvato	I	II	III	totale
SI	150	200	300	650
NO	300	400	600	1300
totale	450	600	900	1950

Le varie colonne (ovvero le distribuzioni di Y dato X) sono in un certo senso diverse visto che le frequenze assolute sono diverse. In questo caso non sembrerebbe però sensato affermare che l'esito dipende dalla classe. Infatti, per tutte e tre le classi è esattamente 1/3 la frazione dei sopravvissuti. Sembra quindi ragionevole affermare che quella mostrata dalla tabella è una situazione in cui non esiste dipendenza di Y da X.

Si osservi che passare da frasi del tipo "Si sono salvati 150 passegeri di prima classe" a "Si sono salvati un terzo dei passeggeri della prima classe" equivale a guardare non le frequenze assolute ma quelle condizionate delle distribuzioni condizionate.

La frase prima evidenziata deve quindi essere precisata nella seguente maniera:

Y (l'esito) **dipende** da X (la classe in cui viaggiava il passegero) poichè le distribuzioni di Y condizionate ad X sono diverse nel senso che hanno *frequenze relative* diverse

Questo discorso giustifica la seguente definizione generale (ci si ricordi che f_{ij}/f_{+j} è la frequenza relativa di y_i nella distribuzione di Y condizionata a $X = x_j$):

Diciamo che Y è **indipendente in distribuzione** o **stocasticamente** da X se, per qualsivoglia i,

$$\frac{f_{i1}}{f_{+1}} = \frac{f_{i2}}{f_{+2}} = \cdots = \frac{f_{ij}}{f_{+j}} = \cdots = \frac{f_{ic}}{f_{+c}}. \qquad \text{(N.1)}$$

Se la (N.1) non è vera diremo che Y **dipende in distribuzione** o **stocasticamente** da X.

Distribuzione marginale, distribuzioni condizionate e indipendenza

Dalla (N.1) discende immediatamente che

> Se le distribuzioni condizionate di Y dato X sono uguali tra di loro, allora sono anche uguali alla distribuzione marginale di Y.

L'uguaglianza, al solito, deve essere intesa nel senso delle frequenze relative.

Per dimostrare la proposizione ci basta far vedere che la (N.1) implica

$$\frac{f_{i+}}{n} = \frac{f_{i1}}{f_{+1}}, \quad i = 1, \ldots, r.$$

Ora, dalla (N.1) segue che $f_{ij} = (f_{i1} f_{+j})/f_{+1}$. Quindi,

$$\begin{aligned} \frac{f_{i+}}{n} &= \frac{\sum_{j=1}^{c} f_{ij}}{n} = \frac{\sum_{j=1}^{c} f_{i1} f_{+j}}{n f_{+1}} = \\ &= \frac{f_{i1} \sum_{j=1}^{c} f_{+j}}{n f_{+1}} = \frac{n f_{i1}}{n f_{+1}} = \frac{f_{i1}}{f_{+1}}. \end{aligned}$$

Y indipendente da X è equivalente a X indipendente da Y

Per quanto detto nella pagina precedente, se Y è indipendente da X allora

$$\frac{f_{ij}}{f_{+j}} = \frac{f_{i+}}{n}, \; i = 1, \ldots, r; \; j = 1, \ldots, c. \qquad \text{(N.2)}$$

La (N.2) può essere riscritta nella forma

$$\frac{f_{ij}}{f_{i+}} = \frac{f_{+j}}{n}, \; i = 1, \ldots, r; \; j = 1, \ldots, c$$

ovvero, mostra che l'indipendenza in distribuzione di Y da X implica l'uguaglianza di tutte le distribuzioni condizionate di X dato Y alla distribuzione marginale di X. Quindi, tutte le distribuzioni condizionate di X dato Y sono tra di loro uguali. Possiamo perciò parlare tranquillamente di indipendenza in distribuzione tra X e Y senza indicare la "direzione".

Frequenze attese.

Poniamo

$$\hat{f_{ij}} = \frac{f_{i+}f_{+j}}{n}.$$

Sempre dalla (N.2) segue che se esiste indipendenza tra le due variabili, $f_{ij} = \hat{f_{ij}}$ per qualsivoglia i e per qualsivoglia j, ovvero, le $\hat{f_{ij}}$ sono le frequenze che ci aspettiamo di trovare quando esiste indipendenza. Si osservi che, come è ovvio, dipendono dalle frequenze marginali. Ovvero le $\hat{f_{ij}}$ ci mostrano come i totali marginali dovrebbero "distribuirsi" tra le varie celle della tabella nel caso di indipendenza stocastica. Per questo motivo, le $\hat{f_{ij}}$ sono chiamate le **frequenze attese (sotto l'ipotesi di indipendenza in distribuzione)**.

X^2

Sul confronto tra frequenze attese e frequenze osservate è anche basato l'indice di uso più comune per *misurare* la dipendenza in distribuzione. Si tratta del cosidetto X^2 di Pearson che è definito come

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij} - \hat{f_{ij}})^2}{\hat{f_{ij}}}.$$

X^2 è uguale a zero in caso di indipendenza e cresce man mano le frequenze osservate si allontano da quelle attese.

Per la giustificazione del perchè questa e non altre (ad esempio, $\sum_{ij}|f_{ij} - \hat{f_{ij}}|$) misura della distanza tra frequenze osservate e frequenze attese ha "preso piede" rinviamo a corsi più avanzati.

Il caso del Titanic

Per il Titanic, le frequenza attese e X^2 valgono

	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

$X^2 = 133{,}05$

Il confronto con le frequenze osservate è particolarmente istruttivo. 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.

Esercizi

1. Lo studente ricalcoli frequenze attese e X^2 a partire dai dati della tabella data all'inizio.

2. Potrebbe venire il dubbio che la preferenza accordata alla I classe sia dipesa dal fatto che in I classe viaggiava un numero più elevato di donne e di bambini e quindi che quello che abbiamo osservato era semplicemente una manifestazione di una "politica di salvataggio" del tipo *prima le donne e i bambini*.

   La seguente tabella si riferisce solo alle donne e ai bimbi.

	classe		
salvato	I	II	III
SI	146	105	103
NO	4	13	141

   Lo studente prima commenti questa nuova tabella, poi calcoli

   (a) le distribuzioni marginali;

   (b) le frequenze attese;

   (c) l'X^2 di Pearson.

3. Lo studente ricostruisca dai dati già forniti in questa unità la distribuzione congiunta di esito

e classe riferita solo ai maschi e la analizzi con le tecniche studiate.

Unità O

I cuculi e Darwin (per non parlare di pettirossi, scriccioli e maiali)

Cenno a dipendenza in media ed altre modalità di dipendenza. La funzione di regressione.

Il problema e i dati

• E' noto che i cuculi depongono le proprie uove nei nidi di altri uccelli a cui viene poi lasciato il compito della cova.
• E' possibile osservare una certa associazione tra territorio e uccello scelto come "ospite", ovvero, in certi territori i cuculi sembrano preferire una specie di uccello come "ospite", in altri un altra.
• Sulla base della teoria della selezione naturale, ci si aspetta quindi una qualche forma di adattamento dell'uovo del cuculo a quella dell'uccello "ospite".
• Per verificare questa idea sono state misurate le lunghezze (in *mm*) di alcune uova di cuculo trovate in nidi di pettirossi e di scriccioli in due territori, uno in cui i cuculi "preferiscono" i pettirossi, l'altro in cui "preferiscono" gli scriccioli.

Pettirossi

```
21,05 21,85 22,05 22,05 22,05 22,25 22,45 22,45 22,65
23,05 23,05 23,05 23,05 23,05 23,25 23,85
```

Scriccioli

```
19,85 20,05 20,25 20,85 20,85 20,85 21,05 21,05 21,05
21,25 21,45 22,05 22,05 22,05 22,25
```


Analisi

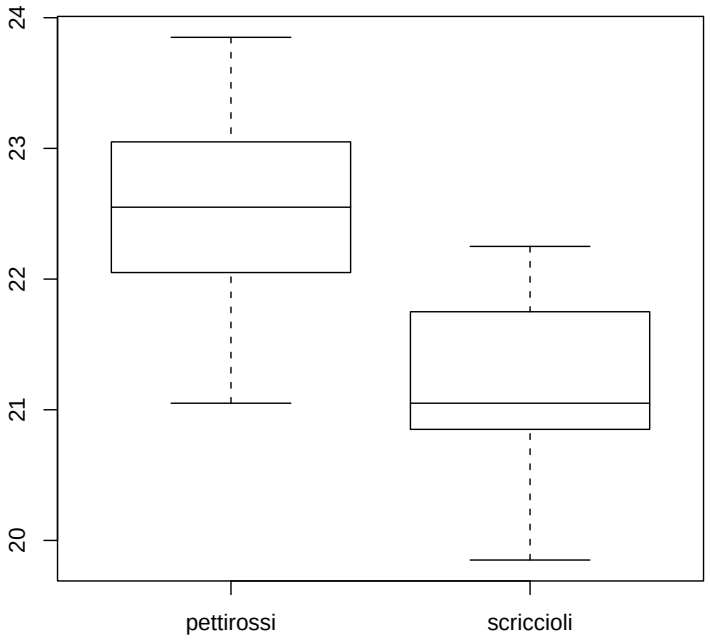
Nella pagina seguente è riportato il *boxplot* per i due gruppi. E' evidente che le uova deposte nei nidi di scricciolo sono tendenzialmente più piccole. Questo è confermato anche dalla media e dalla mediana (vedi tabella seguente). Poichè le uove degli scriccioli sono più piccole di quelle dei pettirossi sembra che i "cuculi votino per Darwin".

Si osservi inoltre che sia il *boxplot* che gli indici sintetici mostrano una sostanziale uguaglianza nella variabilità interna ai due gruppi.

Alcune misure di sintesi per le lunghezze

ospite	media	mediana	sqm	mad
pettirosso	22,57	22,55	0,66	0,5
scricciolo	21,13	21,05	0,72	0,4

Diagramma a scatola con baffi



Abbiamo studiato delle distribuzioni condizionate

• Le osservazioni possono essere viste come un insieme di dati bivariati. Le unità statistiche sono le uove dei cuculi, le due variabili "ospite" e "lunghezza". I dati individuali hanno la forma

uovo	"ospite"	"lunghezza"
1	pettirosso	21,05
2	pettirosso	21,85
	⋮	
16	pettirosso	23,85
17	scricciolo	19,85
	⋮	
31	scricciolo	22,25

• Possiamo poi costruire la distribuzione congiunta delle due variabili (vedi pagina seguente, lo studente la ricalcoli dai dati presentati all'inizio).

• Quello che abbiamo confrontato nell'analisi delle pagine 207-208 sono le due colonne centrali della tabella, ovvero, usando la terminologia dell'unità N, le distribuzioni della "lunghezza" *condizionate* all"'ospite".

Distribuzione congiunta “ospite” e “lunghezza”

	“ospite”		
lunghezza	pettirosso	scricciolo	totale
19,85	0	1	1
20,05	0	1	1
20,25	0	1	1
20,85	0	3	3
21,05	1	3	4
21,25	0	1	1
21,45	0	1	1
21,85	1	0	1
22,05	3	3	6
22,25	1	1	2
22,45	2	0	2
22,65	1	0	1
23,05	5	0	5
23,25	1	0	1
23,85	1	0	1
totale	16	15	31

Dipendenza in media, mediana,. . .

Possiamo guardare alla tabella di contingenza di pagina 210 utilizzando gli occhiali dell’unità N e, in un battibaleno, concludere che tra le due variabili **non** esiste indipendenza in distribuzione. Ad esempio, se “ospite” e “lunghezza” fossero indipendenti la frequenza della “cella” (19,85;pettirosso) dovrebbe essere $(16 \times 1)/31$. La frequenza osservata in quella cella è però nulla. A pagina 207 però eravamo stati più precisi. Non solo avevamo detto che le due distribuzioni della “lunghezza” condizionate all’“ospite” erano diverse (= esiste dipendenza in distribuzione) ma anche, ad esempio, che le due medie erano diverse, ovvero che tra le due variabili esiste **dipendenza in media**.

In generale, si dice che una variabile, per forza numerica, Y è **indipendente in media** da un'altra variabile X se le medie delle distribuzioni di Y condizionate alle varie modalità della X sono tutte uguali tra di loro. Sempre in generale, l'applicazione $x_j \to \text{media}(Y|X = x_j)$ viene chiamata **funzione di regressione di Y su X**. Quindi, possiamo anche dire che abbiamo indipendenza in media se e solo se la funzione di regressione è costante.

In maniera analoga possiamo definire altri concetti di dipendenza/indipendenza (ad es. indipendenza in mediana, indipendenza in varianza,...).

Una osservazione importante

Si noti che questi concetti di indipendenza sono *più deboli* di quello di indipendenza in distribuzione. Discutiamo questo punto con riferimento alla sola indipendenza in media.

Una tabella di contingenza generica è del tipo (si veda l'unità N)

			X			
Y	x_1	$\cdots$	x_j	$\cdots$	x_c	totale
y_1	f_{11}	$\cdots$	f_{1j}	$\cdots$	f_{1c}	f_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{ic}	f_{i+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_r	f_{r1}	$\cdots$	f_{rj}	$\cdots$	f_{rc}	f_{r+}
totale	f_{+1}	$\cdots$	f_{+j}	$\cdots$	f_{+c}	n

Supponiamo che Y sia una variabile numerica. Allora è immediato verificare che se esiste indipendenza in distribuzione tra Y e X allora esiste anche indipendenza in media. Per farlo vedere ci basta dimostrare che per qualsiasi j

$$\frac{\sum_{i=1}^{r} y_i f_{ij}}{f_{+j}} = \frac{\sum_{i=1}^{r} y_i f_{i1}}{f_{+1}}, \qquad \text{(O.1)}$$

ovvero, che la media della distribuzione condizionata di Y dato $X = x_j$ è uguale, per qualsiasi j alla media della distribuzione condizionata di Y dato $X = x_1$. Ma nelle nostre ipotesi la (O.1) è certamente vera poichè l'indipendenza in distribuzione ci garantisce che, anzi coincide con,

$$\frac{f_{ij}}{f_{+j}} = \frac{f_{i1}}{f_{+1}}$$

per $i = 1, \ldots, r$ e $j = 1, \ldots, c$.

Dall'altra parte è facile costruire tabelle in cui esiste indipendenza in media ma non indipendenza in distribuzione. Si verifichi, ad esempio, che questo è quello che accade con la seguente distribuzione congiunta.

	X		
Y	x_1	x_2	totale
-2	0	1	1
-1	1	0	1
0	1	1	1
1	1	0	1
2	0	1	1
totale	3	3	6

In definitiva abbiamo mostrato che
- l'indipendenza in distribuzione implica l'indipendenza in media
- ma che l'indipendenza in media non è sufficiente per concludere che esiste anche indipendenza in distribuzione.

Esercizio

Per verificare l'effetto della vitamina C sull'accrescimento dei maiali, a 30 maiali sono state somministrate dalla nascita dosi diverse di acido ascorbico. Ad una età prefissata è stata poi misurata la lunghezza media dei denti (usata come una misura della crescita). I dati sono i seguenti:

lunghezza (mm)	15,2	21,5	17,6	9,7	14,5	10,0	8,2	9,4
dose (mg)	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5
lunghezza (mm)	16,5	9,7	19,7	23,3	23,6	26,4	20	25,2
dose (mg)	0,5	0,5	1,0	1,0	1,0	1,0	1,0	1,0
lunghezza (mm)	25,8	21,2	14,5	27,3	25,5	26,4	22,4	24,5
dose (mg)	1,0	1,0	1,0	1,0	2,0	2,0	2,0	2,0
lunghezza (mm)	24,8	30,9	26,4	27,3	29,4	23		
dose (mg)	2,0	2,0	2,0	2,0	2,0	2,0		

Ad esempio, il primo maiale coinvolto nello studio ha ricevuto una dose di 0,5mg di vitamina C al giorno e, al controllo, la lunghezza media dei suoi denti era di 15,2mm.

• Si costruisca la tabella di contingenza congiunta e si dica se esiste o non esiste indipendenza in distribuzione.

• Si calcolino le medie della "lunghezza" condizionate alla "dose". Si dica se esiste dipendenza in media. Si disegni su di un grafico cartesiano la funzione di regressione.

• Utilizzando i seguenti risultati intermedi $\sum_i y_i = 619{,}9$, $\sum_i x_i = 35{,}0$, $\sum_i x_i y_i = 814{,}35$ e $\sum_i x_i^2 = 52{,}5$ dove y_i e x_i, $i = 1, \ldots, 30$, indicano rispettivamente le lunghezze e le dosi, determinare la retta di regressione a minimi quadrati della lunghezza sul diametro. La si disegni sul grafico già costruito.

• Si dimostri senza calcolarle che la somma dei quadrati degli scarti dalla funzione di regressione non può essere più grande della somma dei quadrati degli scarti dalla retta di regressione.