Statistica

Claudio Agostinelli
claudio@unive.it
Dipartimento di Scienze Ambientali, Informatica e Statistica
Università Ca' Foscari di Venezia
San Giobbe, Cannaregio 873, Venezia
Tel. 041 2347446, Fax. 041 2347444
http://www.dst.unive.it/~claudio

February 5, 2013

Indice

Prefazione

Questo materiale non è ancora nella sua forma definitiva.

La prima stesura di questi lucidi è stata curata da Guido Masarotto (Università di Padova) che li ha gentilmente resi disponibili. La seconda stesura ha visto contributi di Claudio Agostinelli, Carlo Gaetan e Nicola Sartori. La terza stesura vede aggiunti contributi gentilmente resi disponibili da Cristiano Varin. È stato utilizzato per l'insegnamento di Statistica dell'Università Ca' Foscari, Venezia, per i corsi di Laurea in Economia e Economia Aziendale negli Anni Accademici fino al 2011-2012.

Questo materiale non è un libro e neanche una dispensa! Viene attualmente distribuito agli studenti dell'insegnamento di Probabilità e Statistica 2 (Dipartimento DAIS, Ca' Foscari, Venezia) come materiale aggiuntivo al libro di testo per facilitarne l'apprendimento dei metodi classici di inferenza.

Venezia, 30 Gennaio 2013

Claudio Agostinelli

Informazioni generali

- Nome dell'insegnamento: **Probabilità e Statistica 2**
- Crediti: 6cfu
- Corsi di Laurea: Informatica
- Docente del corso: **Claudio Agostinelli**
  - email: `claudio@unive.it`
  - pagina personale di Ateneo:
  - `http://www.unive.it/nqcontent.cfm?aid=4151&persona=002894`
  - pagina personale:
  - `http://www.dst.unive.it/ claudio`
- Ricevimento studenti:
  - Gli orari sono pubblicati sulla pagina personale di Ateneo all'indirizzo
  - `http://www.unive.it/nqcontent.cfm?aid=415&persona=002894`. Le variazioni saranno pubblicate negli avvisi della medesima pagina.
  - Sono disponibili in formato iCal all'indirizzo
  - `tinyurl.com/9xzuf2v`.
- Lezioni: Orario e luogo delle lezioni sono disponibili in formato iCal all'indirizzo `tinyurl.com/9njmvej`
- Materiale scaricabile:
- `http://http://www.dst.unive.it/~claudio/index.php/Teaching/Courses` (con i link alle pagine dell'insegnamento)
- Testo di riferimento: Ross. Probabilità e Statistica.
- Il simbolo ∼ si ottiene nelle tastiere italiane con la successione ASCII 126 cioè ALT-GR 1 2 6.

Testi di Riferimento e di consultazione

Bibliografia

S. Ross. *Probabilità e statistica per l'ingegneria e le scienze.* Apogeo, 2a edizione, 2008. ISBN 978–88–503–2580–1 testo di riferimento per il corso.

G. Cicchitelli. *Probabilità e Statistica.* Maggioli editore, Rimini, seconda edizione, 2004. ISBN 88–387–1672–2.

A. Azzalini. *Inferenza Statistica.* Springer Italia, Milano, seconda edizione, 2002. ISBN 88–470–0130–7.

G. Casella e R.C. Berger. *Statistical Inference.* Duxbury Press, seconda edizione, 2002. ISBN 0–534–24312–6.

A.M. Mood, F.A. Graybill, e D.C. Boes. *Introduzione alla Statistica.* McGraw–Hill, 1991. ISBN 88–386–0661–7.

L. Pace e A. Salvan. *Introduzione alla Statistica - II. Inferenza, verosimiglianza, modelli.* Cedam, Padova, 2001.

G. Cicchitelli e M.A. Pannone. *Complementi ed esercizi di Statistica descrittiva e inferenziale.* Maggioli, Rimini, 1991.

M. Grigoletto e L. Ventura. *Statistica per le Scienze Economiche. Esercizi con richiami di teoria.* Giappichelli, Torino, 1998.

Modalità d'esame

- L'esame è composto da una parte scritta e da compiti per casa.
- La parte scritta è
  - Obbligatoria
  - Quattro esercizi da svolgere. Il punteggio massimo ottenibile con la prova scritta è 29/30;

  - La durata della prova scritta è di 2 ore

- I compiti per casa (Homeworks)
  - Circa ogni 15 giorni;
  - Da svolgere in due settimane in maniera individuale;

- Il voto finale sarà la media ponderata (con pesi uguali) del voto nella prova pratica e nei compiti per casa.

Materiale utilizzabile durante l'esame

- **Si può usare**
  - Un formulario con alcune formule e tavole delle principali distribuzioni, scaricabili dal sito del corso;
  - Calcolatrice non programmabile e/o Regolo calcolatore;
  - Penna nera o blu, matita e gomma.

- **Non si può usare** tutto quello non espressamente dichiarato nel punto precedente, in particolare
  - Telefonino;
  - Calcolatrici programmabili e/o Palmari;
  - Libri e/o appunti;
  - Fogli non forniti dal docente.

Capitolo 1

Statistica descrittiva

1.1 Introduzione al corso

'Among mathematicians and statisticians who teach introductory statistics, there is a tendency to view students who are not skillful in mathematics as unintelligent"
(McCabe and McCabe.)

Statistics are like a bikini. What they reveal is suggestive, but what they conceal is vital.
(Aaron Levenstein).

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

$$\boxed{\text{Necessità o desiderio di conoscenza Necessità di prendere una decisione}}$$
$$\downarrow$$
$$\boxed{\text{Raccolta informazioni/Osservazione/Sperimentazione}}$$
$$\downarrow$$
$$\boxed{\text{Informazioni elementari/Dati}}$$
$$\downarrow$$
$$\boxed{\text{Sintesi/Modelli}}$$
$$\downarrow$$
$$\boxed{\text{Nuove conoscenze/Ipotesi di decisione}}$$

Statistici, informazioni, nuove conoscenze, decisioni

Necessità o desiderio di conoscenza Necessità di prendere una decisione
↓
Cosa e come rilevare (con altri)
↓
Raccolta informazioni/Osservazione/Sperimentazione
↓
Quanti e quali dati?
↓
Informazioni elementari/Dati
↓
Analisi dei dati
↓
Sintesi/Modelli
↓
Interpretazione (con altri)
↓
Nuove conoscenze/Ipotesi di decisione

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

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

Dati qualitativi, dati numerici, dati ...

In *statistica* si parla di variabili:

- **qualitative** o **categoriale** 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, su scala
  - **sconnessa** (o **nominale**) se non esiste nessun ordinamento naturale tra le modalità; esempi di dati sconnessi sono: (i) il sesso, (ii) il tipo di servizio offerto da un albergo;
  - **ordinale** nel caso in cui un ordinamento naturale esiste; esempi di dati qualitativi ordinali sono: (i) il titolo di studio, (ii) il parere di un intervistato (ad es. classificato come "mediocre", "discreto", "buono").

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

Dati qualitativi, dati numerici, dati ...

Si parla di variabili:

- **quantitative** o **numeriche** 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
  - **discreti** (o **interi**) quando le modalità sono esprimibili da numeri interi; esempi sono: (i) il numero di clienti, (ii) il numero di pezzi prodotti;
  - **continui** (o **reali**)dc quando le modalità sono esprimibili da numeri reali; esempi sono: (i) il tempo d'attesa ad uno sportello, (ii) il peso di un manufatto.

Si noti che le variabili quantitative presentano sempre un ordinamento naturale e quindi sono sempre su scala ordinale.

Dati qualitativi, dati numerici, dati ...

Le variabili quantitative possono essere su scala

- **intervallo** quando l'origine della scala stessa è arbitraria, ovvero, quando lo zero ha un'interpretazione convenzionale (esempio: la temperatura);
- **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 $30mm$ *è lungo il doppio* di un uovo di $15mm$. Non possiamo, viceversa, dire che quando ci sono 30° Celsius la temperatura è doppia rispetto a quando ci 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 quantitativi 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.*

Esempio 1. *Ad esempio, raccogliere dati sui diametri nella forma (2, "troppo piccolo", "buono", "troppo grande") 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, lunghezza del diametro).*

Dati sperimentali vs dati osservazionali

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

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

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

Il problema principale degli studi osservazionali è che non controllando i fattori che possono influenzare il fenomeno sotto indagine risulta difficile essere *ragionevolmente certi* di averli individuati appropriatamente. E incontrerete questo tipo di dati continuamente nel corso dei vostri studi.

Piccolo esempio (per fissare la terminologia)

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

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

Le variabili considerate nello studio sono tre:

clienti le cui *modalità* sono numeriche e discrete;

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

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

“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.

“Statistica Descrittiva” vs “Inferenza Statistica”

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

Descrizione delle caratteristiche del campione
$\downarrow$ Induzione
Affermazioni sulle caratteristiche di tutta la popolazione

Metodi di raccolta dei dati

1. Esperimenti in laboratorio
2. Interviste telefoniche
3. Questionari inviati per posta
4. Porta a porta
5. Interviste per strada
6. ...

Esercizio 1. *Provate a pensare ad un modo di raccogliere informazioni*

Metodi di campionamento Probabilistici

1. **Campionamento casuale semplice**: è un metodo per selezionare individui da una popolazione in maniera tale che ogni possibile campione di una prefissata numerosità ha la medesima probabilità di essere selezionato. In questo caso il campionamento può avvenire con reinserimento oppure senza.

2. **Campionamento casuale stratificato**: è ottenuto selezionando dei campioni casuali semplici da alcuni strati (ovvero sottopopolazioni mutuamente esclusive). Alcuni criteri per dividere una popolazione in strati sono: sesso (maschio, femmina); età (under 18, 18 a 28, 29 a 39); tipologia di impiego (operaio, impiegato, quadro, dirigente).

3. **Campionamento a grappoli**: è un campionamento casuale semplice di grappoli di individui. Il campionamento a grappoli è utile quando è difficile o costoso costruire un campione casuale semplice Per esempio, per stimare il reddito medio familiare in una grande città si usa il campionamento a grappoli, poiché per un campionamento casuale semplice è necessario avere una lista completa delle famiglie da cui estrarre il campione. Un campionamento stratificato necessita ancora di una lista completa. Invece un modo meno dispendioso consiste nel dividere la città in blocchi. Un campione di blocchi viene selezionato e poi ogni famiglia all'interno del blocco viene intervistata.

Metodi di campionamento Non Probabilistici

1. **Campionamento selettivo**: In questo caso chi esegue il campionamento ha il diretto o indiretto controllo degli individui che intende selezionare.

2. **Campionamento per convenienza**: Il campionamento è guidato da criteri di semplicità, economicità.

3. **Campionamento per quote**: il decisore richiede un campione con un certo numero di individui con una prefissata caratteristica. Molti sondaggi concernenti temi politici sono di questo tipo.

A questo punto cosa potete dire di quei sondaggi organizzati durante delle trasmissioni televisive in cui vi chiedono di telefonare ad un numero magari a pagamento?

1.2 Tre organizzazioni di un reparto di produzione

One death is a tragedy, a million deaths is a statistic.
Stalin (Josef Vissarionovich Djugashvili, Stalin)

- 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: qual è 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

Commenti 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.

- 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.

- 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

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

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

Frequenze relative

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

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

Tabelle di frequenza: notazioni

- y_i modalità/classe i del carattere y, $i = 1, 2, \ldots, k$ (k modalità/classi)
- f_i frequenza assoluta numero di unità statistiche che possiedono la modalità/classe y_i
- n numero totale di osservazioni ($n = f_1 + f_2 + \cdots + f_k$)
- p_i frequenza relativa ($p_i = f_i/n$)

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

Scrivere in forma compatta: il simbolo $\sum$ (sommatoria)

Cosa intendiamo per

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

ovvero per 'Somma per i che va da 1 a k' ?

$$n = f_1 + f_2 + \cdots + f_k$$

Alcune proprietà

1. $\sum_{i=1}^{k}(y_i + x_i) = \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} x_i$ ('distribuire il simbolo $\sum$')
2. $\sum_{i=1}^{k} a y_i = a \sum_{i=1}^{k} y_i$ ('portare fuori dal simbolo $\sum$')

   fate attenzione: $\sum_{i=1}^{k} a = ak$

Esercizio 2. $\sum_{i=1}^{k} p_i = ?$

Istogramma

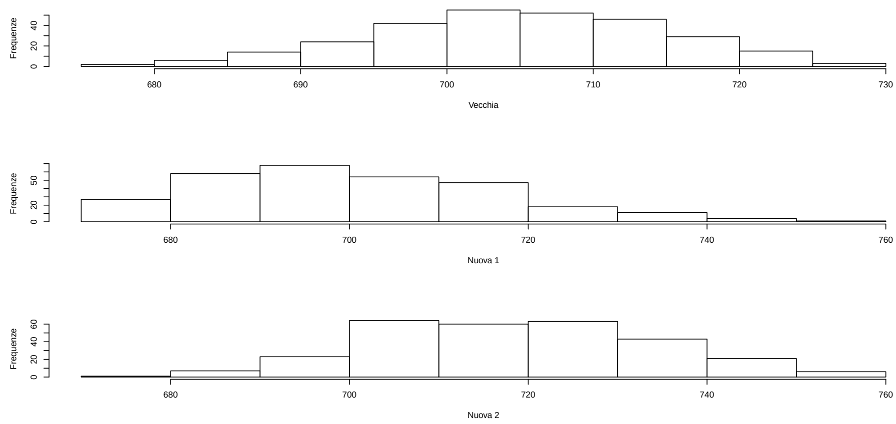
Le differenze tra le tre distribuzioni di frequenza si colgono ancora meglio se le rappresentiamo graficamente. Una possibilità è nella pagine seguente.

Il grafico è stato costruito ponendo

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

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

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



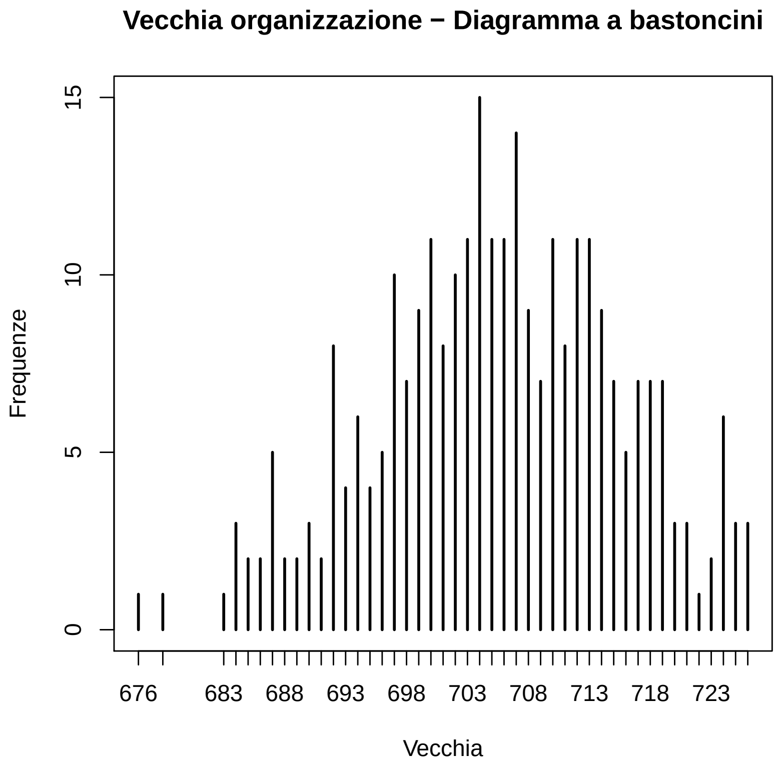
Il grafico suggerisce le stesse considerazioni fatte sulla base della tabella. La distribuzione di nuova 2 è, rispetto alle altre, quella più spostata verso destra (ovvero verso livelli di maggiore produttività). nuova 2 è quindi l'organizzazione migliore (sulla base di questi dati).

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.

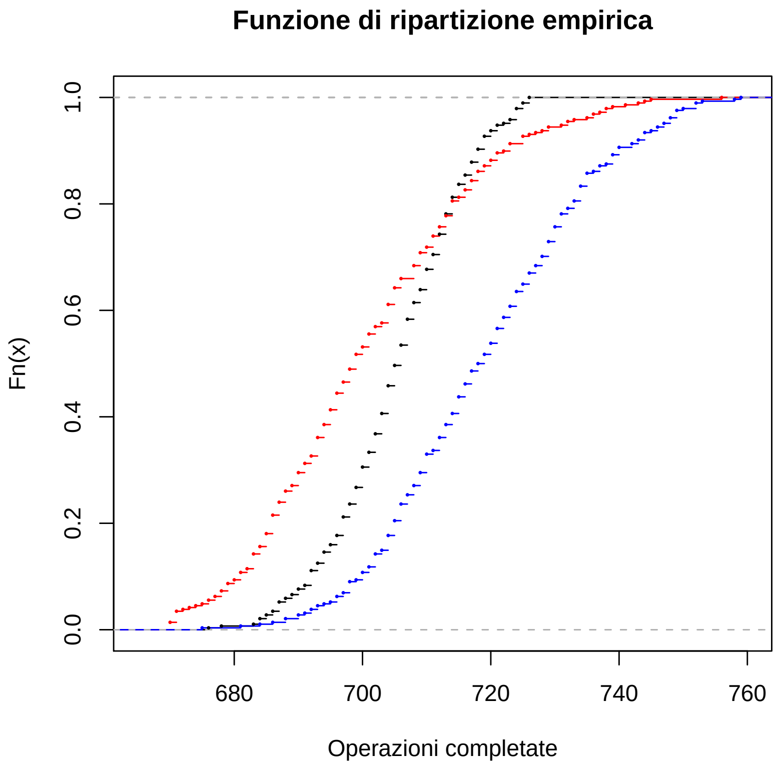


Funzione di ripartizione empirica

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

$$\begin{pmatrix} \text{funzione di} \\ \text{ripartizione empirica} \\ \text{calcolata in } x \end{pmatrix} = \frac{\begin{pmatrix} \text{numero di osservazioni} \\ \text{minori o uguali a } x \end{pmatrix}}{\begin{pmatrix} \text{numero totale di} \\ \text{osservazioni} \end{pmatrix}}$$

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" (rossa) è quella riferita a `nuova 1`. La curva "più bassa" (blu) 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 2 (Calcolo delle frequenze cumulate (nuova 1))**.**

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

1.3 Ancora sugli istogrammi

Se i dati vengono torturati alla fine confessano.
(Anonimo)

- Numero di intervalli in un istogramma.
- Intervalli di ampiezza diversa: densità non frequenze.

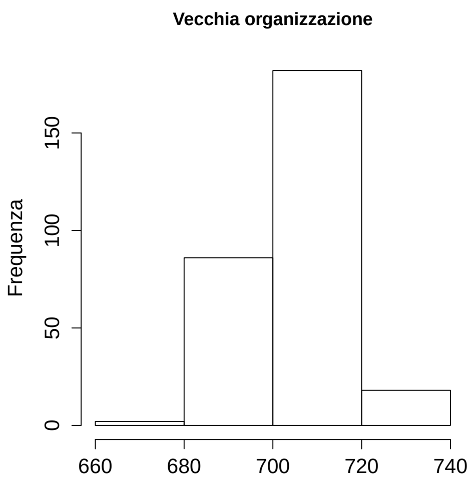
Numero degli intervalli

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

È 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).

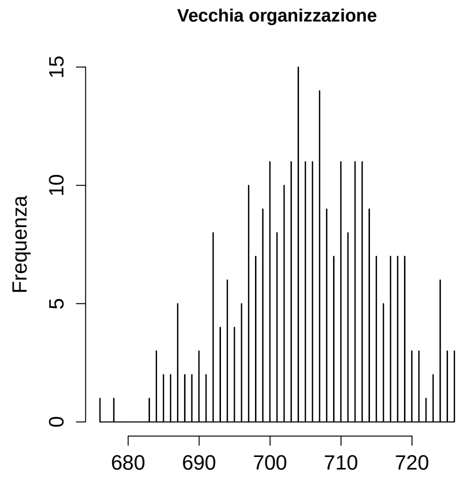
È comunque necessario fare un po' di attenzione. Vediamo alcuni esempi.

Pochi intervalli, poche informazioni



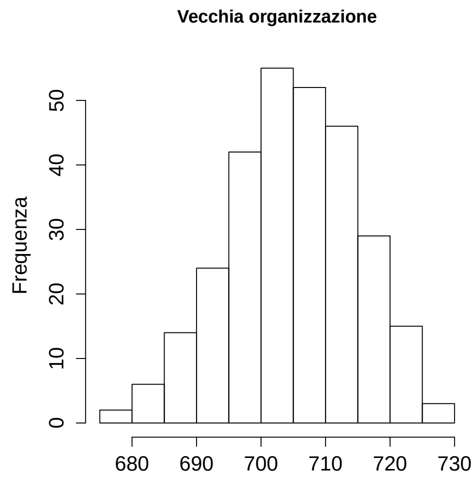
Questo istogramma ci mostra che le frequenze diminuiscono *bruscamente* quando ci si allontana da 720, in maniera essenzialmente erronea.

Troppi intervalli, troppi dettagli



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 del tipo di organizzazione.

Un numero ragionevole di intervalli



Meno *rumore*.

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

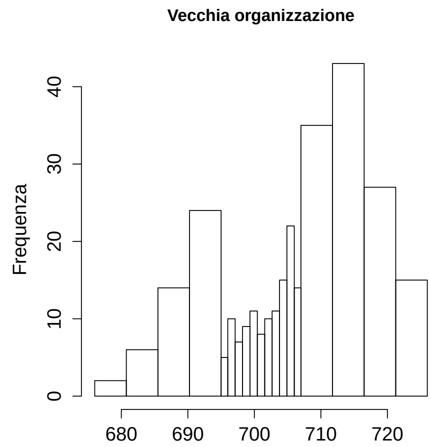
$$\begin{pmatrix} \text{densità} \\ \text{di un intervallo} \end{pmatrix} = \frac{\text{frequenza dell'intervallo}}{\text{lunghezza dell'intervallo}}.$$

Per capire la definizione si pensi alla popolazione. È 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*.

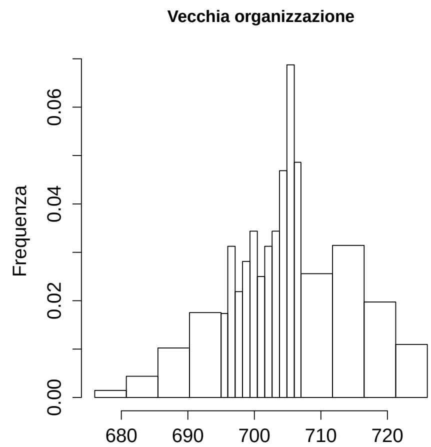
Due esempi (un po' esasperati!) sono presentati nelle seguenti due figure.

Organizzazione del lavoro. 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*.

Organizzazione del lavoro. 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 705.

1.4 Misure di posizione

Ci sono tre professori di statistica che vanno a caccia di lepri. Ad un tratto ne vedono una. Il primo spara, ma un metro a destra. Il secondo spara ma un metro a sinistra. Il terzo esclama "L'abbiamo presa".
(Anonimo)

- 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, soprattutto per la diversa "posizione".

Una domanda che sembra naturale è "di quanto?". Ad esempio, "Nuova 2" sembra con i dati a 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).

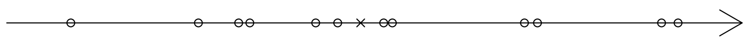
La media aritmetica dei dati è

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

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 risultano confermati i risultati precedenti. Indicano che nuova 2 potrebbe far aumentare la produzione di circa un 2%.

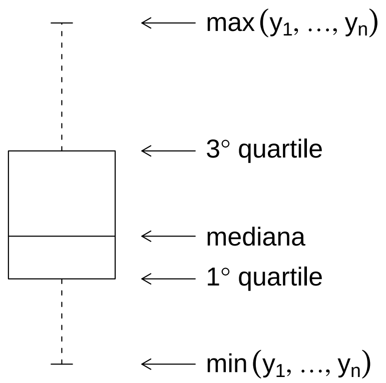
Quantili

- Generalizzano la mediana.
- L'idea alla base di un **quantile-p** dove $p \in [0, 1]$, indicato con y_p, è di cercare un numero che sia più grande del $100 \times p\%$ dei dati osservati e più piccolo del restante $100 \times (1 - p)\%$. Ad esempio, un quantile$-0{,}1$ 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**.

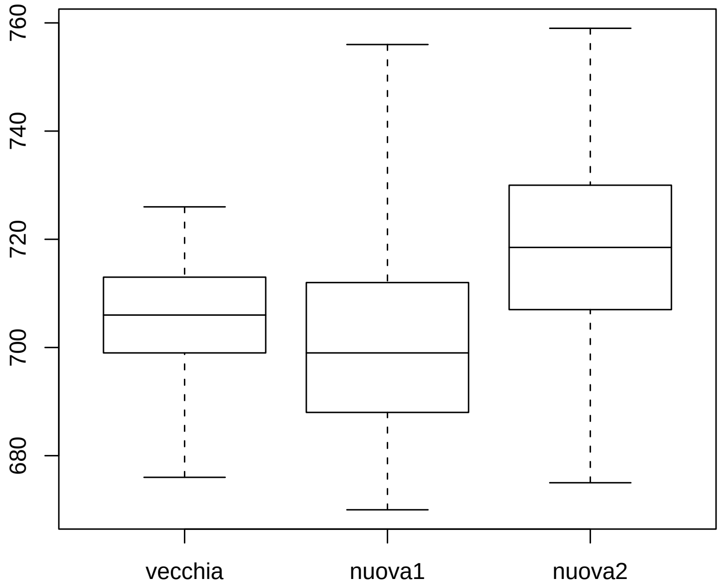
Diagrammi a scatola con baffi

- 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



Variante più usata dei diagrammi a scatola con baffi

Spesso con un diagramma a scatola con baffi si vuole:

- descrivere in maniera stilizzata la distribuzione dei dati (per noi, in questo momento, la posizione e la variabilità) e anche
- evidenziare eventuali valori *estremi*.

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

1. la *scatola* è costruita come descritto precedentemente a partire dai tre quartili.
2. i *baffi* si estendono fino ai dati più lontani che siano però non più distanti di $k \times$(scarto interquartile) dalla *scatola*; lo scarto interquartile è la differenza tra il terzo e il primo quartile (ossia è l'ampiezza della scatola); k è una costante arbitraria tipicamente scelta uguale a 1,5. Ovvero non accettiamo *baffi* esageratamente lunghi.
3. Le osservazioni che sono oltre i *baffi* sono disegnate opportunamente sul grafico (ad. esempio utilizzando un pallino).

Esempio di costruzione di un *boxplot*

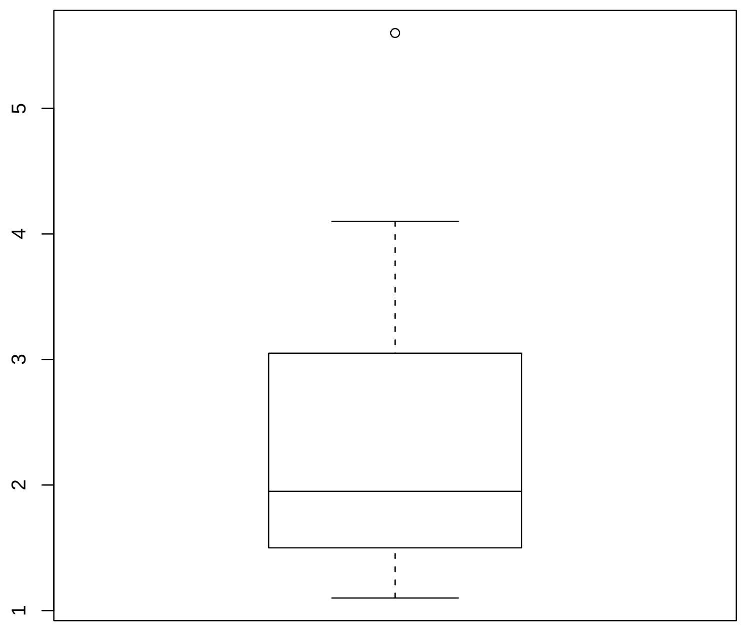
Dati (già ordinati):

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

Perciò (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 5,375; in questo caso solamente l'osservazione risultata uguale a 5,6.

Diagramma a scatola con baffi (esempio precedente)



La media aritmetica: alcune proprietà

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

Infatti, se

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

allora

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

Proprietà 2. *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

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

Proprietà 3. *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

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

Proprietà 4. *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 $b = 1$).

Proprietà 5. *Sia a un numero qualsiasi. Allora*

$$\sum_{i=1}^{n}(y_i - a)^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{y} - a)^2 \tag{1.1}$$

Infatti (tutte le sommatorie vanno da 1 a n)

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

Proprietà 6. *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 (1.1). Infatti la (1.1) 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

Proprietà 7. *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 3. *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

Proprietà 8. *1. Siano $y_1, \ldots, y_n$ dei numeri reali qualsiasi e sia m un valore tale che*

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

Allora

$$\sum_{i=1}^{n} |y_i - m| \leq \sum_{i=1}^{n} |y_i - a|$$

per qualsivoglia costante a. Ovvero, la mediana è il numero che minimizza la somma dei valori assoluti degli scarti di un insieme di dati da una costante.

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

Esempi di calcolo della mediana

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

Esempio 3. *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.

Esempio 4. *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*

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

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

Esempio 5. *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.

Esempio 6. *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*.*

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

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

Dati (già ordinati): 6,4 6,7 6,8 7,0 7,3 7,5 7,5 7,6 7,9 8,1

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

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

Ad esempio,

1. potremmo *decidere* di interpretare "lasciare a sinistra 2,5 osservazioni" come "posizionarsi sul punto intermedio tra la seconda e la terza osservazione (dal basso)" ovvero di *assegnare* al primo quartile il valore di 6,75. Allora, in maniera analoga potremmo *assegnare* al terzo quartile il valore di 7,75 (= punto intermedio tra l'ottava e la nona osservazione).

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

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

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

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

Allora il primo quartile dovrebbe lasciare $(25 \times 99)/100 = 24{,}75$ osservazioni a sinistra. Questo è impossibile.

Le due "soluzioni" viste prima continuano a dare "soluzioni" diverse:

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

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

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

Dati raggruppati: approssimazione della media

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

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

dove k indica il numero degli intervalli. La media non può essere calcolata esattamente. Una *approssimazione* spesso usata in questi casi è

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

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

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

Esercizio 4. *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.*

1.5 Numeri Indice

Xe tuta colpa del'euro!
(Anonimo veneziano)

- Medie ponderate
- Numeri indici
- Numeri indici dei prezzi al consumo

Medie ponderate

La media aritmetica calcolata per dati raggruppati è un esempio di media aritmetica ponderata

$$m_w = \frac{\sum_{i=1}^{k} y_i w_i}{\sum_{i=1}^{k} w_i}$$

dove ad ogni modalità y_i assegnamo un peso non negativo w_i. I pesi w_i possono essere di natura qualsiasi. Vediamo ora un'applicazione interessante della media aritmetica ponderata.

Numeri indici

I numeri indici rappresentano una soluzione al problema del confronto fra misure o gruppi di misure, ad esempio prezzi o produzioni, riferite a tempi, a luoghi e, in generale, a situazioni differenti.

Caratteristiche:

1. i numeri indici sono basati su rapporti (forniscono variazioni relative);
2. sono sempre positivi;
3. sono numeri puri (indipendenti dall'unita di misura in cui sono espresse le grandezze considerate).
4. è prassi moltiplicare il rapporto per 100, per facilitare il confronto in termini percentuali.

Con

$$_0I_t = \frac{y_t}{y_0}$$

indichiamo il numero indice semplice della grandezza nella situazione t, y_t rispetto alla grandezza nella situazione base, y_0, che per convenzione indichiamo con 0.

Indici dei prezzi al consumo

Tra i numeri indici in campo economico e sociale calcolati dall'Istituto Nazionale di Statistica (Istat) [2] quelli di cui accenneremo sono i numeri indici dei prezzi al consumo che misurano le variazioni nel tempo dei prezzi di un paniere di beni e servizi destinati al consumo finale delle famiglie presenti sul territorio economico nazionale e acquistabili sul mercato attraverso transazioni monetarie (sono escluse quindi le transazioni a titolo gratuito, gli autoconsumi, i fitti figurativi, ecc.). (Note Informative - Istat).

Gli indici dei prezzi calcolati con questo metodo sono tre:

1. l'indice Nazionale dei prezzi al consumo per l'Intera Collettività (NIC). E' l'indice utilizzato per misurare l'inflazione a livello di intero sistema economico.

2. l'indice dei prezzi al consumo per le Famiglie di Operai e Impiegati (FOI). Ad esso fa riferimento la maggior parte delle norme nazionali che prevedono l'adeguamento periodico di voci quali affitti e assegni dovuti al coniuge separato;

3. l'indice dei Prezzi al Consumo Armonizzato per i paesi dell'Unione Europea (IPCA)

Supponiamo di voler confrontare i prezzi di n beni o servizi nei due tempi 0 e t (anni, trimestri, mesi, ...), in modo da misurare la variazione media complessiva intervenuta tra i due istanti temporali, tenuto conto delle corrispondenti quantità scambiate. Indichiamo con p_{it} il prezzo del bene i al tempo t e con q_{it} la quantità scambiata del bene i al tempo t e

$$ {}_0I_t^{(i)} = p_{it}/p_{i0} $$

il numero indice elementare del bene i.

La media ponderata dei numeri indici con pesi pari ai valori $p_{i0}q_{i0}$ è detta numero indice dei prezzi di Laspeyres (${}_0^LI_t$)

$$ {}_0^LI_t = \frac{\sum_{i=1}^n {}_0I_t^{(i)} \cdot p_{i0}q_{i0}}{\sum_{i=1}^n p_{i0}q_{i0}} = \frac{\sum_{i=1}^n \frac{p_{it}}{p_{i0}} \cdot p_{i0}q_{i0}}{\sum_{i=1}^n p_{i0}q_{i0}} = \frac{\sum_{i=1}^n p_{it}q_{i0}}{\sum_{i=1}^n p_{i0}q_{i0}} $$

L'indice di Laspeyres richiede che nel tempo vengano rilevati solo i prezzi e non le quantita (fissate pari a quelle del tempo 0).

- L'uso di una ponderazione costante migliora la confrontabilita nel tempo degli indici.

[2] Per saperne di più si veda `www.istat.it/DATI/Prezzi/Aproposito/Main.htm`

- D'altra parte, il sistema di pesi si logora e perde di significato nel tempo. Soluzione possibile: cambio periodico della base.

- Si noti che l'indice di Laspeyres tende a dare un peso maggiore ai prezzi che registrano un aumento e minore a quelli che registrano una diminuzione. Infatti in un periodo di inflazione, ovvero in una fase in cui i prezzi aumentano, il consumatore tende a sostituire nel consumo i beni i cui prezzi crescono piu velocemente con quelli i cui prezzi crescono piu lentamente (cambiano le quantità). Questo significa che un indice di Laspeyres sovrastima il tasso di crescita dei prezzi, ovvero l'inflazione.

Inflazione e mezzi di comunicazione

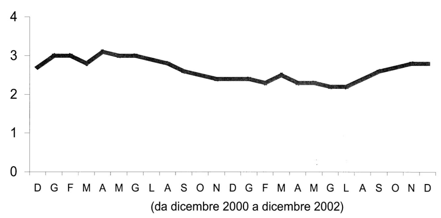
Ecco alcuni titoli (più o meno ad effetto)[3]

- *"Inflazione annuale al 2.21%"*

  ovvero più propriamente un tasso di variazione

  $$\frac{{}_0^L I_t - {}_0^L I_{t-1}}{{}_0^L I_{t-1}} = 0.0221$$

  moltiplicato per 100.

  - Variazione percentuale 2002/2001: 2.5%
  - Variazione percentuale Dic. 2002/Dic. 2001: 2.8%
  - Variazione percentuale Gen. 2003/Gen. 2002: 2.8%

- *"L'inflazione scende ma i prezzi aumentano!"* Infatti un tasso di variazione può diminuire nel tempo senza che questo implichi una diminuzione dei prezzi.

[3]Si rimanda a *Come si misura l'inflazione?* di Ugo Trivellato (`www.lavoce.info`) per altre considerazioni.

- *"Rovigo è risultata la città più cara perché ha avuto l'inflazione più elevata."*

  Pensate di comperare un panino a Milano e a Rovigo e dite quale ragionevolmente è la città più cara. Un numero indice misura il cambiamento in riferimento alla situazione base. É importante osservare che non dà nessuna informazione sull'ordine di grandezza del fenomeno nelle due situazioni a confronto.

1.6 Misure di variabilità

LA STATISTICA

Sai ched'è la statistica? E' 'na cosa che serve pe' fa' un conto in generale de la gente che nasce, che sta male, che more, che va in carcere e che sposa.

Ma pe' me la statistica curiosa è dove c'entra la percentuale, pe' via che, lì, la media è sempre eguale puro co' la persona bisognosa.

Me spiego, da li conti che se fanno seconno le statistiche d'adesso risurta che te tocca un pollo all'anno:

e, se nun entra ne le spese tue, t'entra ne la statistica lo stesso perché c'è un antro che se ne magna due.

(Trilussa)

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

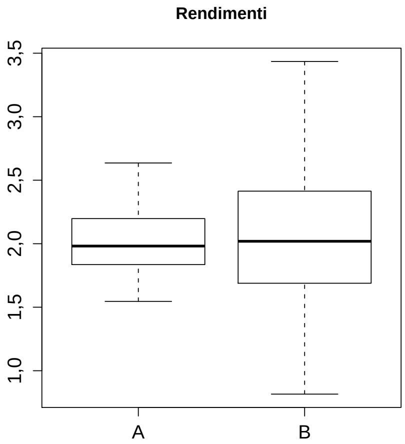
Esempio

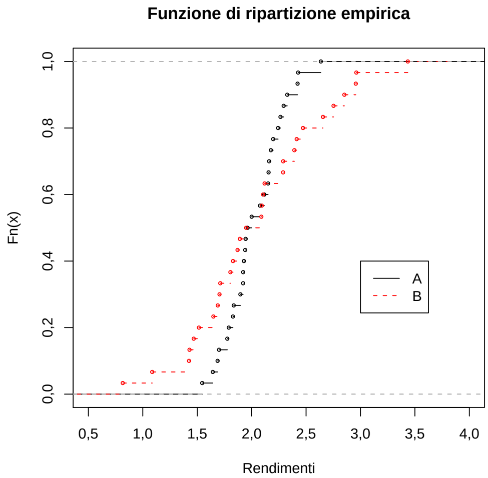
Per confrontare le *performance* di due tipologie di fondi, etichettate come A e B abbiamo preso in considerazione i rendimenti di 30 fondi per ciascuna tipologia.

Gruppo A 1,6432 2,1166 1,8967 1,8356 2,2942 1,929 2,2429 1,7766 1,9221 1,9451 2,1556 2,265 2,1769 1,941 2,1979 1,9218 1,8282 2,422 2,1513 1,7899 2,427 1,6872 2,0004 2,3271 1,7004 2,1604 1,9629 2,6358 1,546 2,0766

Gruppo B 2,7524 1,805 2,2897 2,1052 2,472 1,0867 3,4349 0,8157 1,7048 1,5156 2,0936 2,9569 1,6887 1,4679 1,8289 1,9493 2,2892 2,4138 2,6556 2,0892 2,8518 1,712 1,6486 1,8705 2,962 1,8919 1,4291 2,3924 1,4244 2,119

Riportiamo di seguito i diagrammi a scatola con baffi dei rendimenti e le rispettive funzioni di ripartizione.





Commenti

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

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

Nota: E' importante 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 sono *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

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

Dimostrazione.

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

Formula per il calcolo

Si osservi che

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

e quindi che possiamo scrivere

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

ovvero

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

Esempio 7. *dati:* $1, 3, 2, 5.$

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

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

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

Esercizio 5. *Si dia una formula generale della varianza nel caso di una tabella di frequenza. Si verifichi che nella tabella seguente*

y_i	f_i
4	*2*
6	*8*
7	*3*

var$(y) \simeq 0.840.$

Varianza di trasformazioni lineari dei dati

Proprietà 9. *Se* $y = (y_1, \dots, y_n)$ *e* $z = (z_1, \dots, z_n)$ *con* $z_1 = a+by_1, \dots, z_n = a + by_n$*, dove* a *e* b *sono due costanti qualsiasi. Allora*

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

Sappiamo infatti che

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

Quindi,

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

Esercizio 6. *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

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

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**.

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

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

3. **MAD**.

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

   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 tipologie di fondi: indici di variabilità

	A	B
varianza	0.06	0.34
scarto quadratico medio	0.25	0.58
campo di variazione	1.09	2.62
scarto interquartile	0.34	0.72
MAD	0.31	0.58

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

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

$$\begin{pmatrix} \text{coefficiente} \\ \text{di variazione} \end{pmatrix} = \frac{\text{(scarto quadratico medio)}}{\text{(media aritmetica)}}$$

Proprietà di un indice di variabilità

- Indici di variabilità **assoluta**
  - $V(Y) \geq 0$ (non negatività);
  - $V(Y) = 0$ quando Y è degenere;
  - $V(Y + c) = V(Y)$ per ogni $c \in \mathbb{R}$ (invarianza in traslazione);
  - $V(X) \geq V(Y)$ se le differenze interquartiliche $x_q - x_p \geq y_q - y_p$ per $0 \leq p < q \leq 1$.
- Indici di variabilità **relativa**
  - $W(Y) \geq 0$ definiti per $Y > 0$;
  - $W(cY) = W(Y)$ per ogni $c \neq 0$.

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**.

1.7 Cenno a simmetria e curtosi

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

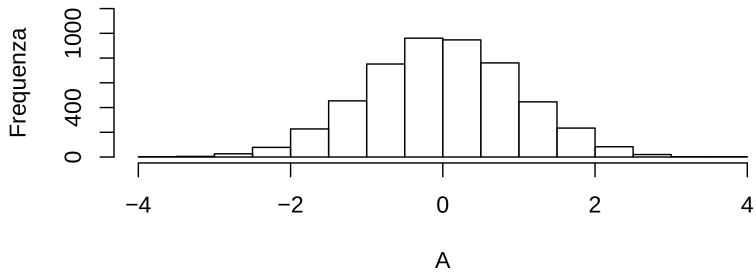
Simmetria

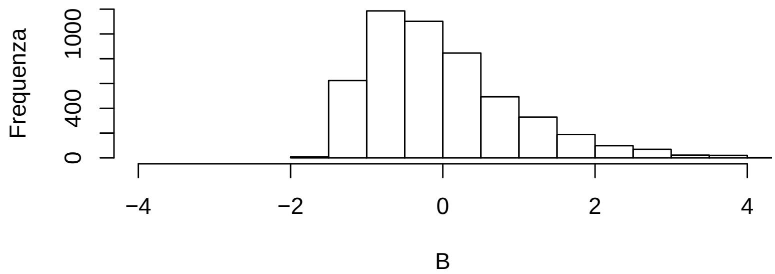
Le seguenti due pagine mostrano rispettivamente gli istogrammi e i *boxplot* costruiti a partire da due insiemi di dati (A e B) *standardizzati* nella maniera brevemente descritta alla fine della unità precedente.

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

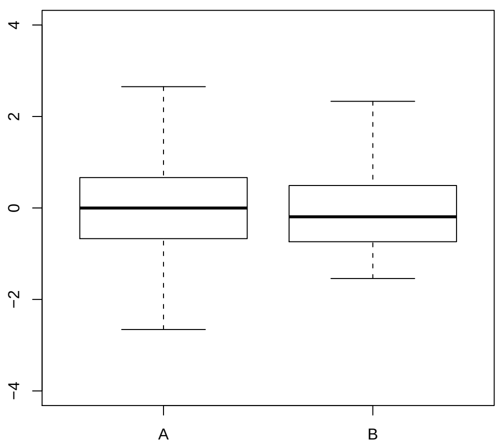
Nonostante questo le due distribuzioni sono diverse. La prima è più o meno **simmetrica** rispetto allo zero. Viceversa, la *coda verso i valori alti* della seconda è molto più lunga della *coda verso i valori bassi.* Si parla in questo caso di **asimmetria positiva**. Ovviamente, nel caso opposto (coda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

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

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

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

Nel nostro esempio, l'indice è pari a -0,012 per l'insieme di dati A e a 1,299 per l'insieme di dati B.

Proprietà 10. *L'indice, per costruzione, è invariante rispetto a trasformazioni lineari dei dati. Ovvero, otteniamo lo stesso risultato sia lavorando con i dati originali che con dati trasformati del tipo* $z_i=a+by_i,\ \ i=1,\dots,n$.

Esercizio 7. *Lo studente verifichi che l'indice di asimmetria è invariante a trasformazioni lineari dei dati.*

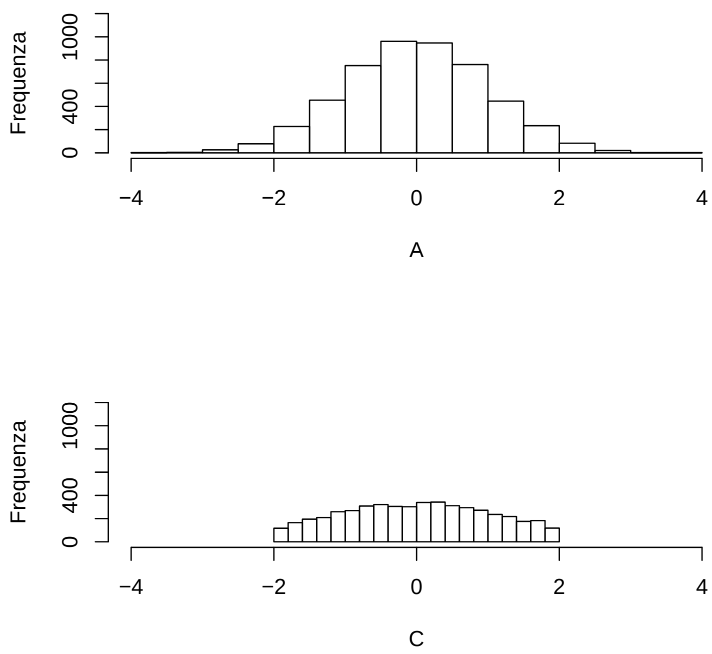
Curtosi

Anche i grafici nelle seguenti due pagine confrontano dati standardizzati (l'insieme A e un nuovo insieme C). In questo caso, ambedue le distribuzioni sono (almeno approssimativamente) simmetriche. Però, nonostante l'uguaglianza delle varianze, la prima distribuzione 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

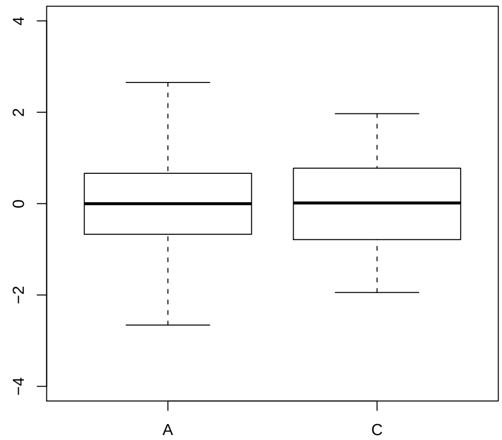
$$\frac{1}{n\mathrm{sqm(y)}^4}\sum_{i=1}^{n}(y_i-\overline{y})^4.$$

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.

L'indice di asimmetria per i due insiemi di dati è pari a -0,012 e a 0,005 rispettivamente (le due distribuzioni sono sostanzialmente simmetriche).

L'indice di curtosi è invece pari a 2,956 e a 2,059, rispettivamente per l'insieme A e B, indicando che la distribuzione dell'insieme A ha code più pesanti di quella dell'insieme B.

1.8 Misure di concentrazione

- Concetto di concentrazione
- Caratteri trasferibili
- Curva di Lorentz
- Indice di Gini
- Rapporto di concentrazione di Gini
- Indice di Boferroni

Aereoporti[4]

	Firenze, ADF SPA	Quota
A	BANCA MONTE DEI PASCHI DI SIENA SPA	5,79732467624791
B	AGA KHAN KARIM	17,8813047238777
C	PREMAFIN FINANZIARIA SPA HOLDING DI PARTECIPAZIONI	2,37082499022523
D	SAGAT SOCIETA' GESTIONE AEROPORTO TORINO SPA	36,77089134015
E	C.C.I.A.A. CAMERA COMMERCIO INDUSTRIA ARTIGIANATO AGRICOLTURA DI PRATO	4,71439911849386
F	CAMERA DI COMMERCIO INDUSTRIA ARTIGIANATO E AGRIC.DI FIRENZE	16,0152130898923
G	COMUNE DI FIRENZE	2,58764706579306
H	SO.G.IM. SPA	13,86239499532
	TOTALE	100

	Venezia, MARCO POLO SPA - SAVE	Quota
I	AEROPORTO DI VENEZIA MARCO POLO SPA - SAVE	2,5621473234711
J	KAIROS INVESTMENT MANAGEMENT LIMITED	3,72375578262417
K	COMUNE DI TREVISO	2,65619439784454
L	FONDAZIONE DI VENEZIA	2,75532509785979
M	FINANZIARIA INTERNAZIONALE HOLDING SPA	51,3420771694373
N	AMMINISTRAZIONE DELLA PROVINCIA DI VENEZIA	18,3709521630827
O	COMUNE DI VENEZIA	18,5895480656804
	TOTALE	100

- Per discriminare tra due popolazioni può non essere sufficiente basarsi su indici di variabilità perchè:
  - non vincolano allo zero del fenomeno descritto (dovuta all'invarianza per traslazione);
  - possono mostrare una dipendenza dalla scala del fenomeno.

[4] Quote degli azionisti rilevanti (con quote superiori al 2%) per le societè AEROPORTO DI FIRENZE - ADF SPA e AEROPORTO DI VENEZIA MARCO POLO SPA - SAVE alla data del 11 Settembre 2009. Fonte: `www.consob.it`.

- In particolare questo è vero quando le unità statistiche sono identificate con gli individui e la caratteristica oggetto di studio è un bene condivisibile come, per esempio, la ricchezza e le quote di mercato di un certo bene.
- In questi contesti, può aver senso chiedersi se il bene sia equamente distribuito tra gli individui oppure se esso è concentrato nelle mani di alcuni soggetti.

Un carattere è

- **equidistribuito** tra gli n individui della popolazione se ciascuno di essi ne possiede una quota pari a $1/n$;
- **concentrato** se uno o pochi degli n individui ne possiede tutte le quote.

Così diremo che il reddito di un paese è tanto più concentrato quanto più il reddito complessivo è posseduto da una frazione modesta delle unità statistiche, ovvero quanto più disparità vi è fra i soggetti in quel paese.

Il concento di concentrazione potrebbe erroneamente essere interpretato come l'inverso di quello della variabilità ma ciò non è vero

- la variabilità si manifesta nel modo in cui i valori di un dato carattere si disperdono attorno ad un centro comune (la media, la mediana, ecc.);
- la concentrazione si realizza quando un fenomeno si manifesta con maggiore intensità solo su un numero ridotto di osservazioni.

Chiameremo **precisione** il concetto inverso a quello di variabilità e **omogeneità** quello inverso alla concentrazione.

Ha senso parlare di concentrazione solo in presenza di fenomeni continui che sono intrinsecamente condivisibili tra gli individui di una popolazione. Per formalizzare questa idea si parla di trasferimento concentrativo.

Definizione 1 (Trasferimento concentrativo)**.** *Dati i seguenti valori ordinati in senso crescente, $y_1 \leq y_2 \leq \cdots \leq y_n$. Si definisce **trasferimento concentrativo** uno spostamento di una quantità $\varepsilon > 0$ da y_i a y_j per $i < j$.*

Un indice di concentrazione dovrà essere "sensibile" ai trasferimenti del carattere tra le unità statistiche, un trasferimento concentrativo deve portare ad un aumento dell'indice di concentrazione.

- **concentrazine minima**: si ottiene quando il carattere complessivo è ripartito in misura uguale tra tutte le unità statistiche e quindi lo possiedono tutte in misura pari alla media μ;

- **concentrazione massima**: quando il carattere è posseduto nella sua totalità da una sola unità statistica, che lo possiede quindi in misura pari a $n\mu$, mentre le rimanenti $n-1$ unità ne posseggono 0.

1.8.1 Curva di Lorenz

Consideriamo le modalità **ordinate** $y_1 \leq y_2 \leq \cdots \leq y_n$ e indichiamo con

- $p_i = i/n$: la frazione cumulata delle prime i unità statistiche ($i = 1, \cdots, n$);

- $$q_i = \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{n} y_j} \ ,$$

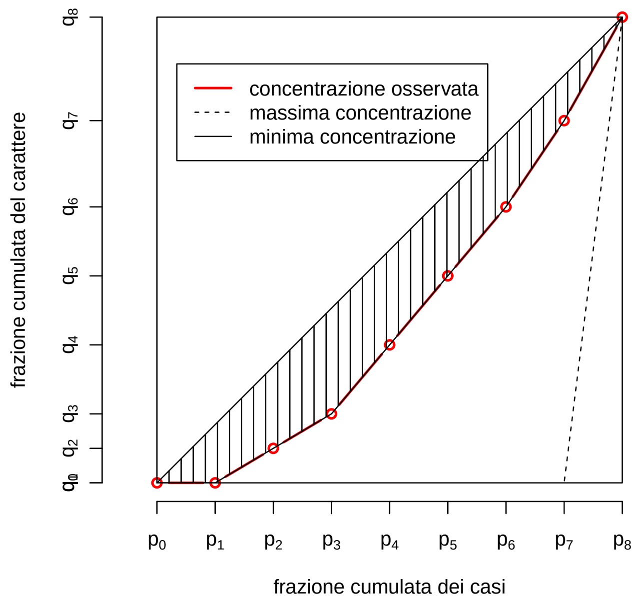
  la frazione cumulata del carattere posseduto dalle prime i unità statistiche ($i = 1, \cdots, n$);

- $p_0 = q_0 = 0$;

- Si noti che $p_i \geq q_i$ per $i = 1, \cdots, n$ e che $p_i = q_i$ solo quando la prima $g = i/n * 100$ frazione di unità statistiche possiede esattamente una frazione pari a g del carattere;

- Le coppie (p_i, q_i) posso essere rappresentate in un piano cartesiano. Il grafico è chiamato **curva di Lorenz**.

Curva di Lorenz



1.8.2 Indice di concentrazione di Gini

Indice di concentrazione di Gini

Definizione 2 (Indice di concentrazione di Gini)**.** *L'indice è definito come media ponderata, con pesi p_i, delle differenze normalizzate*

$$\frac{p_i - q_i}{p_i} \ .$$

$$G = \frac{\sum_{i=1}^{n-1}\left(\frac{p_i - q_i}{p_i}\right)p_i}{\sum_{i=1}^{n-1} p_i} = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} \ .$$

Essendo

$$\sum_{i=1}^{n-1} p_i = \sum_{i=1}^{n-1} \frac{i}{n} = \frac{\frac{n(n-1)}{2}}{n} = \frac{n-1}{2}$$

e quindi

$$G = \frac{2}{n-1}\sum_{i=1}^{n-1}(p_i - q_i) \ .$$

Ulteriori sviluppi algebrici portano a

$$\begin{aligned} G &= \frac{2}{n-1}\sum_{i=1}^{n-1}(p_i - q_i) \\ &= \frac{2}{n-1}\left(\frac{n-1}{2} - \sum_{i=1}^{n-1} q_i\right) \\ &= 1 - \frac{2}{n-1}\sum_{i=1}^{n-1} q_i \end{aligned}$$

Rapporto di concentrazione di Gini

Il rapporto di concentrazione di Gini è dato dal rapporto tra A: l'area sottesa dalla coppia di curve di equidistribuzione-concentrazione osservata e B: l'area di massima concentrazione conseguibile. Ricordando che

$$(p_{i+1} - p_i) = \frac{1}{n}\ ; \qquad q_0 = 0\ ; \qquad q_n = 1$$

otteniamo

$$\begin{aligned} A &= \frac{1}{2} - \frac{1}{2}\sum_{i=0}^{n-1}(q_i + q_{i+1})(p_{i+1} - p_i) \\ &= \frac{1}{2} - \frac{1}{2n}\sum_{i=0}^{n-1}(q_i + q_{i+1}) \\ &= \frac{1}{2} - \frac{1}{2n}\left((q_0 + q_1) + (q_1 + q_2) + (q_2 + q_3) + \cdots + (q_{n-1} + q_n)\right) \\ &= \frac{1}{2} - \frac{1}{2n}\left(1 + 2\sum_{i=1}^{n-1} q_i\right) \\ &= \frac{1}{2}\left(1 - \frac{1}{n}\right) - \frac{1}{n}\sum_{i=1}^{n-1} q_i \end{aligned}$$

$$B = \frac{1}{2} - \frac{1}{2\,n} = \frac{n-1}{2n} = \frac{1}{2}\left(1 - \frac{1}{n}\right) \stackrel{n\to\infty}{\rightarrow} \frac{1}{2} = B_\infty\ .$$

Quindi otteniamo

$$\begin{aligned}
R &= \frac{A}{B_\infty} \\
&= \frac{\frac{1}{2} - \frac{1}{2}\sum_{i=0}^{n-1}(q_i + q_{i+1})(p_{i+1} - p_i)}{\frac{1}{2}} \\
&= 1 - \sum_{i=0}^{n-1}(q_i + q_{i+1})(p_{i+1} - p_i) \\
&= \frac{n-1}{n} - \frac{2}{n}\sum_{i=1}^{n-1} q_i
\end{aligned}$$

Se invece consideriamo

$$\begin{aligned}
\frac{A}{B} &= \frac{\frac{1}{2} - \frac{1}{2}\sum_{i=0}^{n-1}(q_i + q_{i+1})(p_{i+1} - p_i)}{\frac{1}{2}\left(1 - \frac{1}{n}\right)} \\
&= \frac{\frac{1}{2}\left(1 - \frac{1}{n}\right) - \frac{1}{n}\sum_{i=1}^{n-1} q_i}{\frac{1}{2}\left(1 - \frac{1}{n}\right)} \\
&= 1 - \frac{2n}{n-1}\frac{1}{n}\sum_{i=1}^{n-1} q_i \\
&= 1 - \frac{2}{n-1}\sum_{i=1}^{n-1} q_i \\
&\equiv G \; .
\end{aligned}$$

Quindi abbiamo

$$R = \frac{n-1}{n} G \; .$$

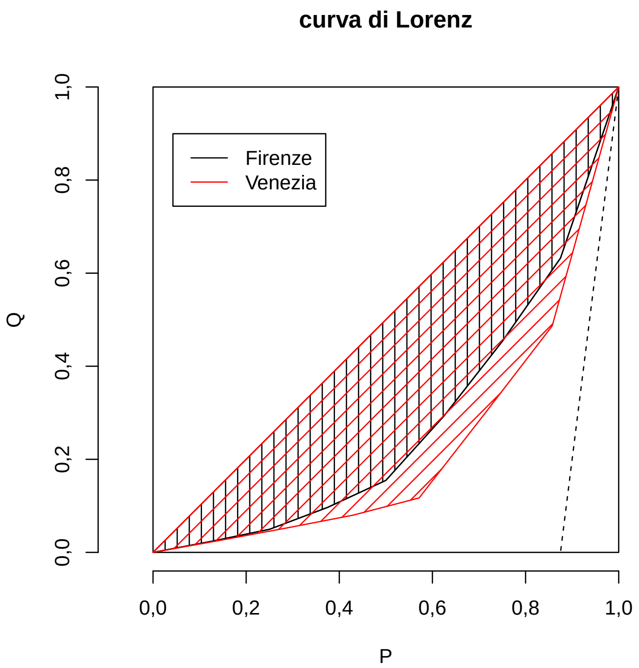
Aereoporti

Firenze	p_i	q_i	Venezia	p_i	q_i
2,371	0,125	0,024	2,562	0,143	0,026
2,588	0,250	0,050	2,656	0,286	0,052
4,714	0,375	0,097	2,755	0,429	0,080
5,797	0,500	0,155	3,724	0,571	0,117
13,862	0,625	0,293	18,371	0,714	0,301
16,015	0,750	0,453	18,590	0,857	0,487
17,881	0,875	0,632	51,342	1,000	1,000
36,771	1,000	1,000			

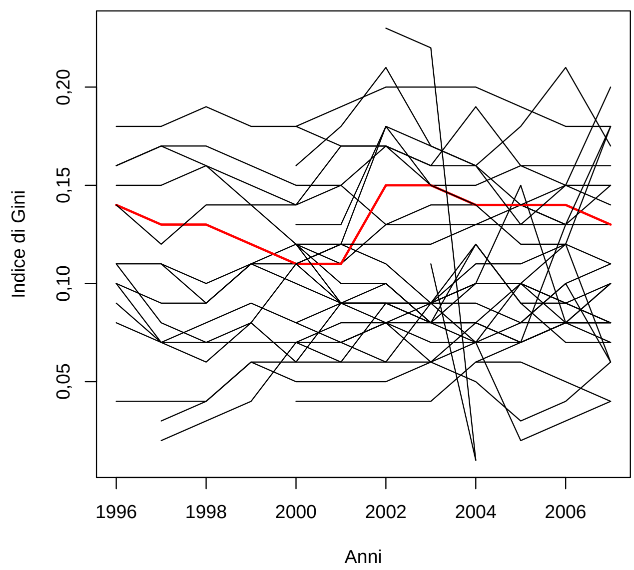
	G	R
Firenze	0,513	0,449
Venezia	0,646	0,554

Quindi la concentrazione è elevata in entrambi i casi, con una concentrazione maggiore nel caso dell'aereoporto di Venezia.

Curve di Lorenz per gli aereoporti

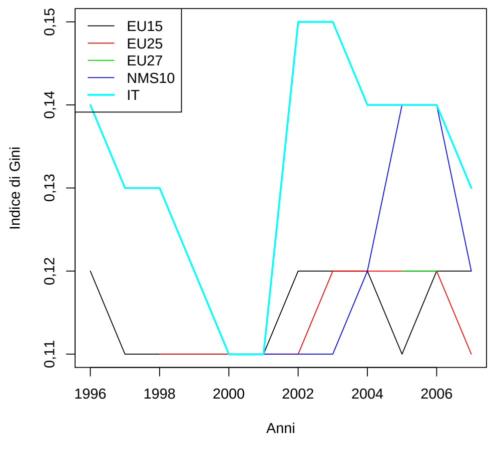


Indice di Gini per gli stati europei



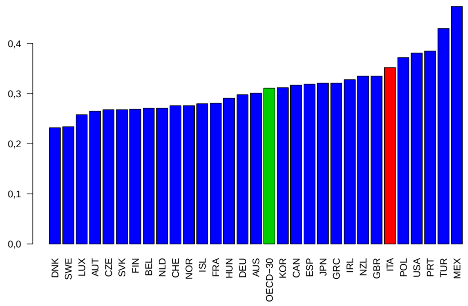
In rosso l'indice di Gini per l'Italia.

Indice di Gini per gli stati europei



EU15: Europa a 15, EU25: Europa a 25, EU27: Europa a 27, NMS10: Nuovi stati membri (CZ, EE, CY, LV, LT, HU, MT, PL, SI, SK).

Indice di Gini per gli stati in OECD-30



In rosso l'indice di Gini per l'Italia, in verde il valore medio dei 30 stati[5]. Fonte: OECD Income distribution questionnaire (Growing Unequal? Income Distribution and Poverty in OECD Countries - OECD 2008 - ISBN 9789264044180).

Calcolo dell'indice per tabelle di frequenza

Nel caso di dati già riassunti in tabella di frequenza possiamo utilizzare la seguente formula di calcolo

$$G = \frac{\sum_{i=1}^{n} (y_i C_i - T_i) f_i}{T \left(\sum_{i=1}^{n} f_i - 1 \right)}$$

dove

- y_i: modalità;
- f_i: frequenze assolute;
- $C_i = \sum_{j=1}^{i} f_j$: frequenze cumulate assolute;
- $T_i = \sum_{j=1}^{i} y_j f_j$: totale del carattere fino all'i–esima modalità;
- $T = \sum_{j=1}^{n} y_j f_j = \mu C_n = \sum_{j=1}^{n} f_j$: totale della variabile;
- $\mu = (\sum_{j=1}^{n} y_j f_j) / \sum_{j=1}^{n} f_j$: media della variabile.

[5]Note: Countries are ranked, from left to right, in increasing order in the Gini coefficient. Data refer to the mid-2000s for all countries except for Japan and Switzerland, where they refer to 2000. The income concept used is that of disposable household income in cash, adjusted for household size with an elasticity of 0.5.

Calcolo dell'indice per variabili continue suddivise in classi

Si possono considerare due casi

- quando per ciascuna classe è riportato il totale parziale t_i. In questo caso possiamo individuare in maniera esatta la media della classe che è

  $$\bar{x}_i = \frac{t_i}{f_i} \ .$$

  Il calcolo risulta esatto.

- quando il totale parziale non è riportato (la maggior parte dei casi). In questo caso dobbiamo approssimare la media della classe attraverso il valore centrale della classe. Il calcolo risulta approssimato.

Il calcolo prosegue come nel caso di dati già in forma di tabella di frequenza.

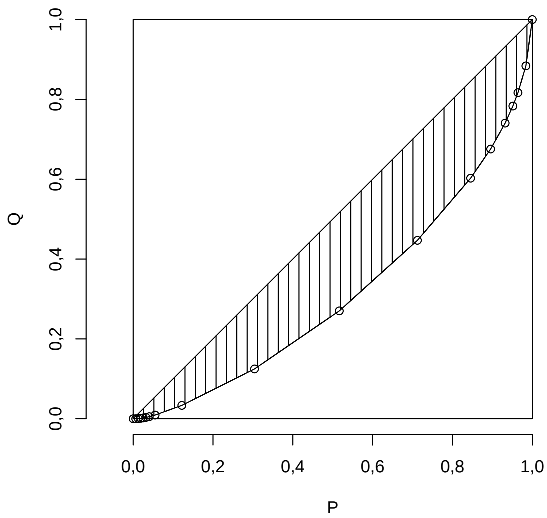
Reddito imponibile comune di Venezia[6]

[6] Reddito imponibile ai fini delle addizionali all'IRPEF per le persone fisiche nel comune di Venezia, anno d'imposta 2006. Fonte: Ministero dell'Economia e delle Finanze.

Classe	Frequenza (f_i)	Totale (t_i)	Totale cumulato (T_i)
0 $\dashv$ 1000	1059	506972	506972
1000 $\dashv$ 2000	1055	1491173	1998145
2000 $\dashv$ 3000	1007	2391164	4389309
3000 $\dashv$ 4000	974	3307305	7696614
4000 $\dashv$ 5000	1198	5207548	12904162
5000 $\dashv$ 6000	1346	7116853	20021015
6000 $\dashv$ 7500	2428	15635903	35656918
7500 $\dashv$ 10000	11059	93046184	128703102
10000 $\dashv$ 15000	30069	351178701	479881803
15000 $\dashv$ 20000	35076	559417951	1039299754
20000 $\dashv$ 26000	32306	679565583	1718865337
26000 $\dashv$ 33500	22007	598711407	2317576744
33500 $\dashv$ 40000	8283	280903581	2598480325
40000 $\dashv$ 50000	6003	249767815	2848248140
50000 $\dashv$ 60000	3197	164355875	3012604015
60000 $\dashv$ 70000	2103	128916108	3141520123
70000 $\dashv$ 100000	3279	258034989	3399555112
100000 $\dashv$ +	2657	446264468	3845819580

$\bar{y}_i = t_i/f_i$	f_i	t_i	T_i	$\bar{y}_i F_i$
479	1059	506972	506972	506972
1413	1055	1491173	1998145	2988000
2375	1007	2391164	4389309	7410946
3396	974	3307305	7696614	13904942
4347	1198	5207548	12904162	23007973
5287	1346	7116853	20021015	35103111
6440	2428	15635903	35656918	58389923
8414	11059	93046184	128703102	169332444
11679	30069	351178701	479881803	586232163
15949	35076	559417951	1039299754	1359964879
21035	32306	679565583	1718865337	2473264488
27205	22007	598711407	2317576744	3797452312
33913	8283	280903581	2598480325	5014652881
41607	6003	249767815	2848248140	6402094568
51409	3197	164355875	3012604015	8074721370
61301	2103	128916108	3141520123	9757288117
78693	3279	258034989	3399555112	12783630963
167958	2657	446264468	3845819580	27730877401

Curva di Lorenz per i redditi di Venezia



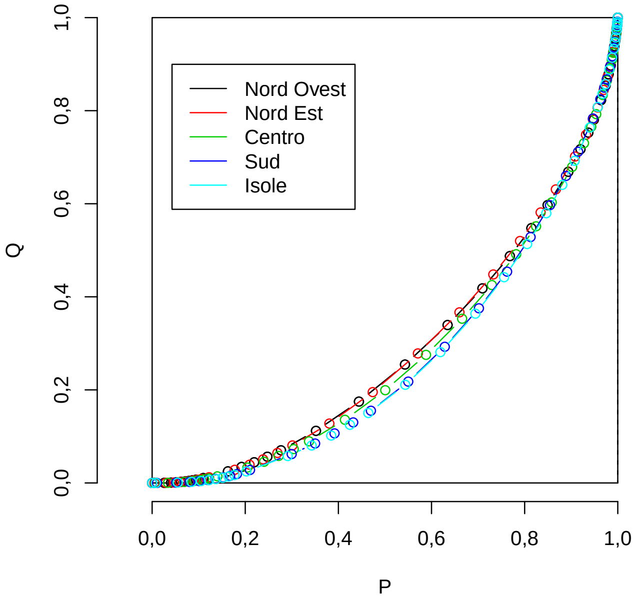
Indice di Gini: $G = 0,374$, Rapporto di concentranzione di Gini: $R = 0,374$.

Distribuzione del numero dei contribuenti per classi di reddito complessivo (in Euro) e area geografica. Persone Fisiche (Anno d'imposta 2006)

Classe	$\bar{y}$	Nord Ovest	Nord Est	Centro	Sud	Isole
0 ⊣ 0	0	64652	50195	50195	60575	86848
0 ⊣ 500	250	236149	203090	203090	220460	345865
500 ⊣ 1000	750	170319	148680	148680	164947	243380
1000 ⊣ 1500	1250	132692	109551	109551	120172	174200
1500 ⊣ 2000	1750	111454	91136	91136	97503	156145
2000 ⊣ 2500	2250	102429	84040	84040	90100	136824
2500 ⊣ 3000	2750	96249	80561	80561	86904	137280
3000 ⊣ 3500	3250	88931	71650	71650	80236	128232
3500 ⊣ 4000	3750	86896	70937	70937	80097	122744
4000 ⊣ 5000	4500	178400	138981	138981	153002	236122
5000 ⊣ 6000	5500	612552	474669	474669	530046	752259
6000 ⊣ 7000	6500	337680	272197	272197	295311	423260
7000 ⊣ 8000	7500	314376	250507	250507	256662	350990
8000 ⊣ 9000	8500	326448	263641	263641	266778	337011
9000 ⊣ 10000	9500	337721	276405	276405	270803	319639
10000 ⊣ 12500	11250	867752	680699	680699	634517	673902
12500 ⊣ 15000	13750	1065992	799758	799758	712837	659674
15000 ⊣ 17500	16250	1142654	832610	832610	720625	616787
17500 ⊣ 20000	18750	1056458	766144	766144	638136	508757
20000 ⊣ 22500	21250	865375	623561	623561	525527	423865
22500 ⊣ 25000	23750	680587	492707	492707	429881	349823
25000 ⊣ 27500	26250	533786	380329	380329	352153	290050
27500 ⊣ 30000	28750	401616	280185	280185	275677	219038
30000 ⊣ 35000	32500	513295	353843	353843	361331	266718
35000 ⊣ 40000	37500	302397	202842	202842	209733	133346
40000 ⊣ 45000	42500	194976	126899	126899	128481	70764
45000 ⊣ 50000	47500	137827	87255	87255	87729	46258
50000 ⊣ 60000	55000	177299	111817	111817	115139	59610
60000 ⊣ 70000	65000	111297	71081	71081	75146	41956
70000 ⊣ 80000	75000	75853	48881	48881	51505	29550
80000 ⊣ 90000	85000	52147	33000	33000	34136	18618
90000 ⊣ 100000	95000	37384	24001	24001	24592	12921
100000 ⊣ 125000	112500	55120	34521	34521	34393	17449
125000 ⊣ 150000	137500	27998	17160	17160	16203	7325
150000 ⊣ 175000	162500	15980	9341	9341	9003	3707
175000 ⊣ 200000	187500	9917	5763	5763	5392	2047
200000 ⊣ +	200000	29760	15357	15357	15484	5297

Fonte: Ministero dell'Economia e delle Finanze.

Curva di Lorenz per i redditi del 2006



Indici di concentrazione per il reddito del 2006

	G	R
Nord Ovest	0,428	0,428
Nord Est	0,427	0,427
Centro	0,45	0,45
Sud	0,468	0,468
Isole	0,472	0,472

1.8.3 Indice di concentrazione di Bonferroni

Consideriamo le osservazioni ordinate $y_1, y_2, \cdots, y_n$ e le rispettive medie parziali

$$\bar{y}_i = \frac{\sum_{j=1}^{i} y_j}{i} \qquad i = 1, \cdots, n$$

dove abbiamo $\bar{y}_n \equiv \bar{y}$.

Definizione 3 (Indice di concentrazione di Bonferroni)**.**

$$\begin{aligned} B &= \frac{\sum_{i=1}^{n-1}(\bar{y}-\bar{y}_i)}{(n-1)\bar{y}} \\ &= 1 - \frac{1}{\bar{y}(n-1)}\sum_{i=1}^{n-1}\bar{y}_i \\ &= 1 - \frac{n}{(n-1)}\sum_{i=1}^{n-1}\frac{q_i}{i} \end{aligned}$$

Gli indici di Gini e Bonferroni sono medie ponderate

Posto s_i gli scarti relativi dalle medie parziali rispetto alla media (generale)

$$s_i = \frac{\bar{y}-\bar{y}_i}{\bar{y}}$$

abbiamo

$$\begin{aligned} B &= \frac{\sum_{i=1}^{n-1} s_i}{n-1} \\ G &= \frac{\sum_{i=1}^{n-1} s_i\ i}{\sum_{i=1}^{n-1} i} = \frac{\sum_{i=1}^{n-1} s_i\ 2i}{n(n-1)} \end{aligned}$$

per cui entrambi gli indici sono medie ponderate di s_i con pesi $w_i = 1$ per l'indice di Bonferroni e $w_i = i$ per l'indice di Gini.

Proprietà di un indice di concentrazione

- $C(Y) \geq 0$;
- se X e Y sono due variabili con lo stesso totale $T = \sum_{i=1}^{n_X} x_i = \sum_{j=1}^{n_Y} y_j$ e se X si ottiene da Y con un trasferimento concentrativo, allora

  $$C(X) \geq C(Y)\ ;$$

- $C(dY) = C(Y)$ per ogni $d \neq 0$.

1.9 Esercizi ricettivi per area geografica

- Moda
- Diagramma a barre
- Diagramma a torta
- Mutabilità

I dati

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

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

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

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

Frequenze assolute e relative

La tabella della pagina precedente è *poco maneggevole* e non viene riportata dall'ISTAT che riporta la seguente tabella di frequenze assolute.

	Nord	Centro	Sud	TOTALE
Alberghi	21568	6324	5536	33428
Campeggi e villaggi turistici	992	494	885	2371
Alloggi agro-turistici	3194	3392	1183	7769
Altri esercizi	57978	3334	1415	62727
TOTALE	83732	13544	9019	106295

Ad esempio, 992 è il numero di campeggi e villaggi turistici censiti nel Nord.

Questa tabella ci permette lo studio congiunto delle due variabili. Vedremo nella prossima unità come fare questo tipo di analisi. Per il momento, limitiamoci a studiare le due variabili separatamente.

La variabile *tipo* ha la seguente distribuzione di frequenze

	frequenza	frequenza relativa
Alberghi	33428	0.314
Campeggi e villaggi turistici	2371	0.022
Alloggi agro-turistici	7769	0.073
Altri esercizi	62727	0.590
TOTALE	106295	1.000

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

	frequenza	frequenza relativa
Nord	83732	0.788
Centro	13544	0.127
Sud	9019	0.085
TOTALE	106295	1.000

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

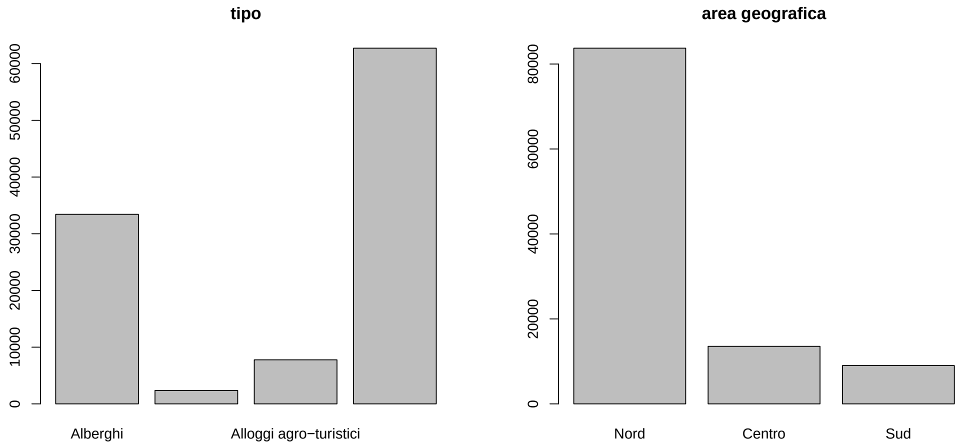
- Nei precedenti esempi avevamo dati **numerici**. In questo caso sono espressi da aggettivi. Sono **qualitativi** o **categoriali**.

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

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

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

Diagramma a barre: frequenze assolute



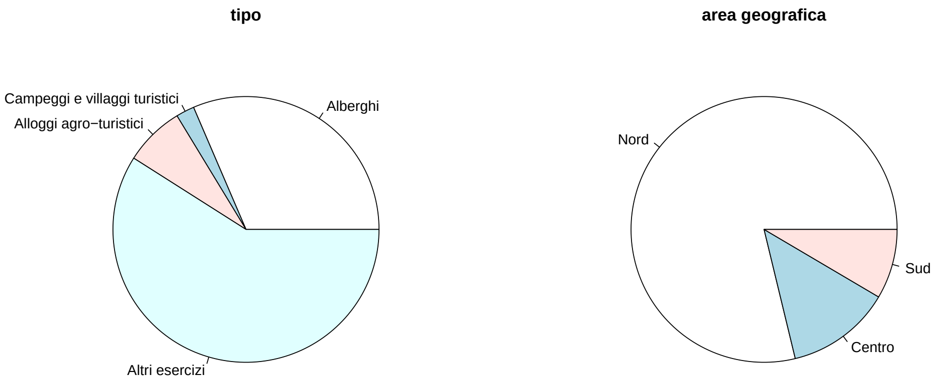
La rappresentazione grafica più utilizzata è il diagramma a barre, in cui ogni modalità è rappresentata da una barra di altezza pari alla frequenza (assoluta o relativa) della modalità. Si osservi che i rettangoli, contrariamente al caso di un istogramma, sono disegnati *staccati*.

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

Diagramma a torte: frequenze relative

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

$$angolo = 360 \cdot frequenza\ relativa$$



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

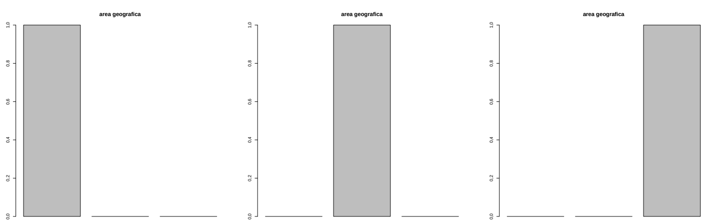
modalità	c_1	$\cdots$	c_i	$\cdots$	c_k
frequenza relativa	0	$\cdots$	1	$\cdots$	0

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

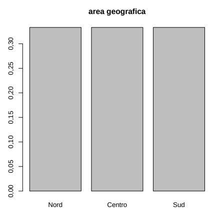
- La situazione opposta (*massima mutabilità*) la troviamo invece quando le unità statistiche si ripartiscono in maniera uguale tra le varie modalità. In questo caso la distribuzione delle frequenze relative diventa

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

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



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



Indice di Gini

Tabella delle frequenze relative.

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

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

- 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.

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

Entropia di Shannon

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

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

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

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

- E' possibile inoltre dimostrare che anche questo indice assume valore massimo nelle situazioni di massima mutabilità:

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

- Può quindi essere eventualmente *normalizzato* (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

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

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

Esercizio 8. *Provate a calcolare gli indici normalizzati di mutabilità per la variabile tipo di esercizio.*

Capitolo 2

Probabilità

2.1 Probabilità

La prova genera l'evento con una certa probabilità.
(Giuseppe Pompilj).

- Esperimento casuale, spazio campionario e eventi
- Richiami di insiemistica
- La probabilità
- La probabilità condizionata
- Teorema di Bayes
- Indipendenza stocastica
- Elementi di calcolo combinatorio

Cos'è la probabilità

- Tutti i fenomeni (riguardanti le scienze fisiche, naturali, umane,...) che sono caratterizzati da incertezza relativa ai possibili esiti di determinate azioni coinvolgono il concetto di probabilità.
- La probabilità è un *concetto primitivo* (come il tempo, lo spazio,...), cioè originario per l'essere umano perché innato e sempre presente nelle sue regole di comportamento.
- La probabilità è anche una *misura* perché associa al concetto primitivo una valutazione numerica.
- Concetto e misura sono due cose distinte: quasi tutti siamo in grado di dire che la probabilità di una cinquina al lotto è più bassa della probabilità di un ambo; tuttavia, poche persone sarebbero in grado di determinare con immediatezza la misura di queste probabilità.
- La necessità di prendere decisioni in condizioni di incertezza implica la necessità di "quantificare" (misurare) la probabilità e quindi di studiare la Teoria (assiomatica) della Probabilità (intesa come disciplina matematica).
- Nel seguito definiremo gli elementi costitutivi della Teoria delle Probabilità, le relazioni che legano questi elementi e infine introdurremo gli assiomi della probabilità.

Esperimento casuale

Definizione 4 (Esperimento casuale)**.** *Una qualunque azione il cui risultato non è certo.*

Esempi:

1. Un lancio di un dado
2. Un lancio di una moneta
3. La prima estrazione del lotto
4. Sostenere un esame
5. Il riuscire negli studi
6. Il rendimento di un titolo

Evento elementare

Definizione 5 (Evento elementare)**.** *Il risultato più semplice di un esperimento casuale*

Esempi:

1. Una delle facce di un dado
2. Testa o croce
3. Un numero da 1 a 90
4. Riuscire o non riuscire
5. Laurearsi o non laurearsi
6. Un qualunque numero reale compreso tra -1 e +1.

Spazio campionario

Definizione 6 (Spazio campionario)**.** *L'insieme di tutti i possibili eventi elementari. Viene indicato con $\mathcal{S}$.*

Esempi:

1. $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$

2. $\mathcal{S} = \{T, C\}$
3. $\mathcal{S} = \{1, 2, 3, \ldots, 90\}$
4. $\mathcal{S} = \{OK, KO\}$
5. $\mathcal{S} = \{OK, KO\}$
6. $\mathcal{S} = \{x | x \in [-1, 1]\}$

Esempio più complicato

1. Esperimento casuale: due lanci di un dado
2. Evento elementare: poichè i lanci sono due, il più semplice risultato dell'esperimento è una coppia di numeri (i, j) con $i = 1, \ldots, 6$ e $j = 1, \ldots, 6$. In ogni coppia il primo numero indica il risultato del primo lancio ed il secondo quello del secondo lancio.
3. Quindi lo spazio campionario è

$$\begin{array}{lllllll} \mathcal{S} = \{ & (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ & (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ & (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ & (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ & (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ & (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6)\} \end{array}$$

   $\mathcal{S}$ è quindi costituito da 36 eventi elementari.

   Attenzione: l'evento elementare $(1, 2)$ è considerato diverso da $(2, 1)$ perchè conta l'ordine di estrazione.

Evento

Definizione 7 (Evento)**.** *Un evento è un insieme di eventi elementari, cioè un qualunque sottoinsieme dello spazio campionario.*

Un evento si verifica quando si verifica un evento elementare che lo compone.

Nell'esempio dei due lanci del dado un evento è "almeno una faccia è pari", cioè l'insieme A costituito da:

$$\begin{array}{llllllllll} A = \{ & (1,2) & (1,4) & (1,6) & (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ & (3,2) & (3,4) & (3,6) & (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ & (5,2) & (5,4) & (5,6) & (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6)\} \end{array}$$

che è un sottoinsieme di $\mathcal{S}$, costituito da 27 eventi elementari.

L'evento A può essere visto come composto da altri eventi, cioè da altri sottoinsiemi di $\mathcal{S}$. Per esempio:

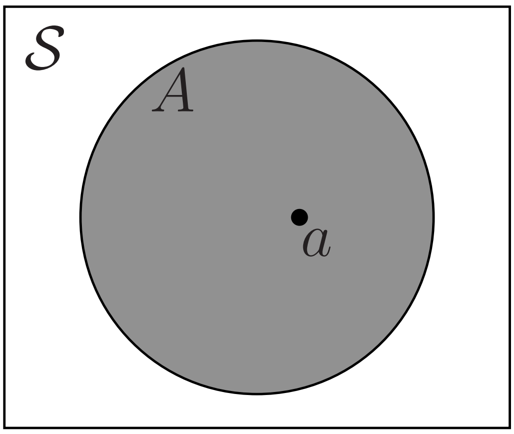
$$\begin{aligned}
B_1 &= \{\text{solo la prima faccia è pari}\} \\
&= \{(2,1)\quad(2,3)\quad(2,5)\quad(4,1)\quad(4,3)\quad(4,5)\quad(6,1)\quad(6,3)\quad(6,5)\} \\
B_2 &= \{\text{solo la seconda faccia è pari}\} \\
&= \{(1,2)\quad(1,4)\quad(1,6)\quad(3,2)\quad(3,4)\quad(3,6)\quad(5,2)\quad(5,4)\quad(5,6)\} \\
B_3 &= \{\text{entrambe le facce sono pari}\} \\
&= \{(2,2)\quad(2,4)\quad(2,6)\quad(4,2)\quad(4,4)\quad(4,6)\quad(6,2)\quad(6,4)\quad(6,6)\}
\end{aligned}$$

Nella terminologia dell'insiemistica l'insieme A è pari all'unione di B_1, B_2 e B_3.

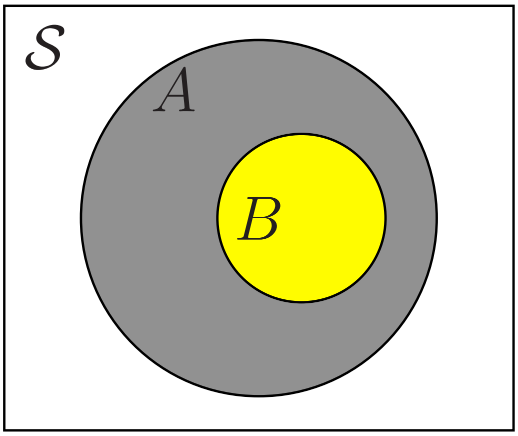
In definitiva "lavorare" con gli eventi è possibile mediante l'algebra dell'insiemistica.

Richiami di insiemistica (diagrammi di Venn)

- $a \in A$ significa che l'evento elementare a appartiene all'insieme A.



- B è un sottoinsieme di A, $B \subseteq A$, significa che tutti gli eventi elementari contenuti in B appartengono anche all'evento A. Ossia $x \in B \Rightarrow x \in A$ (per ogni evento elementare x in B)

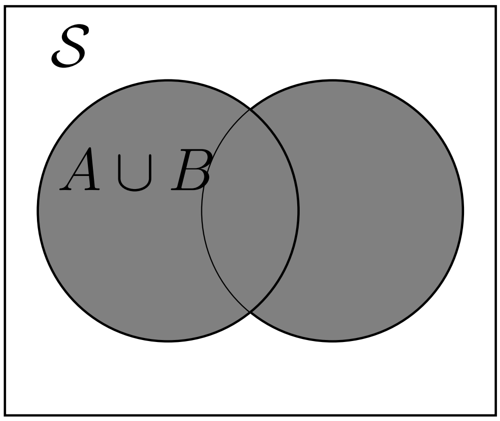


Notiamo che

$$A = B \Leftrightarrow A \subseteq B \quad \text{e} \quad B \subseteq A$$

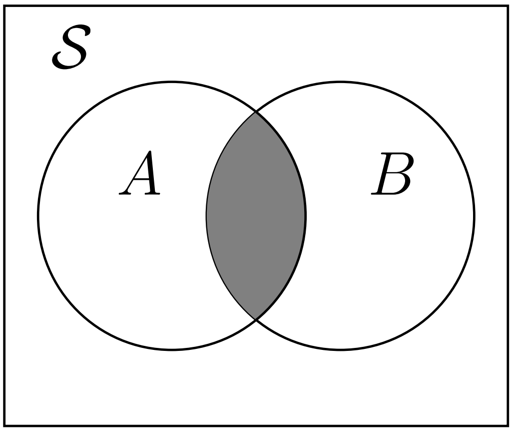
Operazioni tra insiemi

1. L'<u>unione</u> di due insiemi è l'insieme costituito dagli elementi appartenenti o ad un insieme o ad un altro e si indica con $A \cup B$.



$$A \cup B = \{x : x \in A \quad \text{o} \quad x \in B\}$$

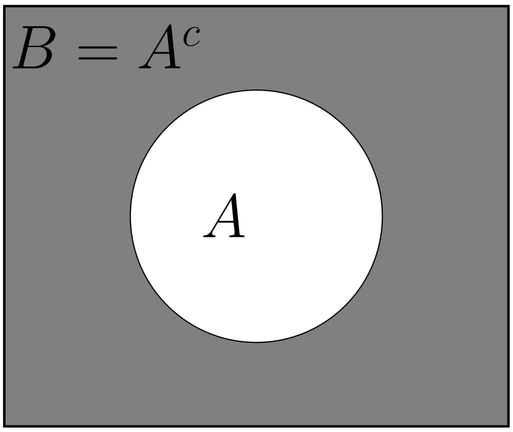
2. L'<u>intersezione</u> di due insiemi è l'insieme costituito dagli elementi appartenenti sia ad un insieme sia all'altro e si indica con $A \cap B$



$A \cap B = \{x : x \in A \quad \text{e} \quad x \in B\}$

Due insiemi (eventi) sono tra loro disgiunti (incompatibili) se $A \cap B = \emptyset$, dove $\emptyset$ rappresenta l'insieme vuoto (evento impossibile).

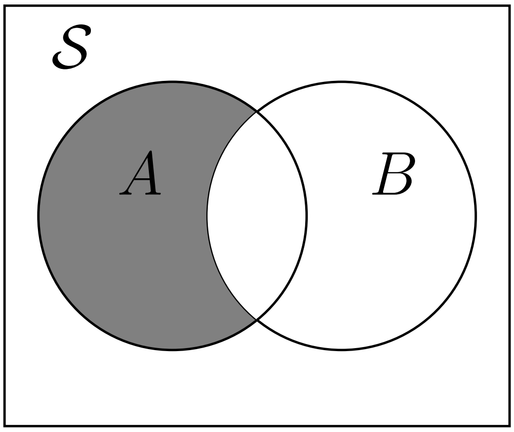
3. Insieme complementare (o negazione): il complementare dell'insieme A, indicato con $\bar{A}$ o A^c, è l'insieme degli elementi di $\mathcal{S}$ (spazio campionario) non appartenenti ad A.



$A^c = \bar{A} = \{x : x \notin A\}$

Valgono le seguenti proprietà: $A \cap A^c = \emptyset$ e $A \cup A^c = \mathcal{S}$.

4. Insieme differenza: è l'insieme degli elementi di A che non appartengono a B, indicato con $A - B$.



$$A - B = \{x : x \in A \quad \text{e} \quad x \notin B\}$$

Dalla definizione stessa deriva che $A - B = A \cap B^c$.

Proprietà delle operazioni tra insiemi

1. Idempotenza:

$$\begin{aligned} A \cup A &= A\,, \\ A \cap A &= A\,. \end{aligned}$$

2. Commutativa:

$$\begin{aligned} A \cup B &= B \cup A\,, \\ A \cap B &= B \cap A\,. \end{aligned}$$

3. Associativa:

$$\begin{aligned} A \cup (B \cup C) &= (A \cup B) \cup C\,, \\ A \cap (B \cap C) &= (A \cap B) \cap C\,. \end{aligned}$$

4. Distributiva:

$$\begin{aligned} A \cup (B \cap C) &= (A \cup B) \cap (A \cup C)\,, \\ A \cap (B \cup C) &= (A \cap B) \cup (A \cap C)\,. \end{aligned}$$

5. Evento impossibile

$$\begin{aligned} A \cup \emptyset &= A\,, \\ A \cap \emptyset &= \emptyset\,. \end{aligned}$$

6. Spazio campionario

$$\begin{aligned} A \cup \mathcal{S} &= \mathcal{S}\,, \\ A \cap \mathcal{S} &= A\,. \end{aligned}$$

7. Leggi di De Morgan

$$\begin{aligned} (A \cup B)^c &= A^c \cap B^c\,, \\ (A \cap B)^c &= A^c \cup B^c\,, \\ (A^c)^c &= A\,. \end{aligned}$$

Più in generale, dati gli insiemi $A_1, A_2, \ldots, A_n$, si ha che

$$\begin{aligned} (\bigcup_{i=1}^{n} A_i)^c &= \bigcap_{i=1}^{n} A_i^c\,, \\ (\bigcap_{i=1}^{n} A_i)^c &= \bigcup_{i=1}^{n} A_i^c\,, \end{aligned}$$

Esercizio 9. *Alcune di queste proprietà sono abbastanza intuitive. Altre, invece, sono meno ovvie: provate a dimostrare le proprietà c., d. e g. utilizzando i diagrammi di Venn.*

Catalogo di sottoinsiemi di $\mathcal{S}$

Costruiamo un "catalogo" (o "lista", o "algebra") di sottoinsiemi dello spazio campionario, mettendo in essa anche $\mathcal{S}$ stesso e l'insieme vuoto $\emptyset$.

Per esempio considerando l'esperimento casuale "lancio di una moneta", gli eventi elementari sono $\{T\}$ e $\{C\}$. Considerando anche $\mathcal{S} = \{T, C\}$ e l'insieme vuoto, allora il catalogo, indicato da $\mathcal{F}$ è dato da

$$\mathcal{F} = \{\emptyset, \{T\}, \{C\}, \{T, C\}\}.$$

Il catalogo $\mathcal{F}$ è, quindi, un insieme i cui elementi sono a loro volta degli insiemi: un insieme di insiemi.

Questo **catalogo** deve soddisfare le seguenti **proprietà**:

1. Se $A \in \mathcal{F}$, allora $A^c \in \mathcal{F}$ (chiusura rispetto al complementare).

2. $\emptyset \in \mathcal{F}$ e $\mathcal{S} \in \mathcal{F}$

3. Se si hanno n eventi $A_1, A_2, A_3, \ldots, A_n$ in $\mathcal{F}$ allora $\bigcup_{i=1}^{n} A_i \in \mathcal{F}$ (chiusura rispetto all'unione finita).

   Quest'ultima proprietà deve valere anche se gli eventi sono infiniti: $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ (chiusura rispetto all'unione numerabile).

Esercizio 10. *Utilizzando le leggi di De Morgan e le proprietà 1. e 3. di $\mathcal{F}$, mostrate che se $A_1, A_2, A_3, \ldots, A_n$ appartengono ad $\mathcal{F}$, allora $\bigcap_{i=1}^{n} A_i \in \mathcal{F}$ (chiusura rispetto all'intersezione finita).*

Inciso (culturale): Noi non possiamo dire molto di più su com'è fatto $\mathcal{F}$ senza entrare in un campo della matematica denominato teoria della misura. Una cosa che possiamo dire è che quando lo spazio campionario $\mathcal{S}$ è numerabile, cioè i suoi elementi sono in corrispondenza biunivoca (1 ad 1) con i numeri naturali, allora $\mathcal{F}$ è costituito da tutti i possibili sottoinsiemi di $\mathcal{S}$, $\mathcal{S}$ stesso e l'insieme vuoto. Se il numero degli eventi elementari è finito e pari ad n, allora il numero di elementi di $\mathcal{F}$ è pari a 2^n (si consideri l'esempio del lancio di una moneta).

Purtroppo ci sono molte situazioni empiriche in cui non è possibile considerare $\mathcal{F}$ come costituito da tutti i possibili sottoinsiemi di $\mathcal{S}$. Questo è vero in particolare quando il numero degli elementi di $\mathcal{S}$ è infinito e non è numerabile (si pensi ad un esperimento casuale, il cui risultato è un qualunque numero reale). In queste situazioni, se considerassimo tutti i possibili sottoinsiemi di $\mathcal{S}$, allora la funzione di probabilità che definiremo tra poco non soddisferebbe gli assiomi della probabilità (vedi dopo). In queste situazioni $\mathcal{F}$ contiene soltanto alcuni sottoinsiemi di $\mathcal{S}$.

Per i nostri scopi è sufficiente dire che il catalogo $\mathcal{F}$ è tale che gli assiomi della probabilità sono soddisfatti.

Probabilità

La probabilità è una funzione che assegna ad ogni elemento di $\mathcal{F}$ un numero reale appartenente all'intervallo chiuso $[0, 1]$

$$\begin{array}{rccc} P: & \mathcal{F} & \to & [0,1] \\ & A & \to & \Pr(A) \end{array}$$

La probabilità, quindi, deve essere vista come una funzione, con dominio $\mathcal{F}$ e codominio l'intervallo chiuso $[0, 1]$. Il modo con cui questa funzione viene costruita dipende dal problema che si sta analizzando.

Per esempio nel lancio di una moneta, se questa è perfettamente bilanciata, si può costruire la probabilità come il rapporto tra il numero totale di casi favorevoli ed il numero totale dei casi:

$$\Pr(A) = \frac{\sharp\text{di eventi elementari che portano alla realizzazione dell'evento } A}{\sharp\text{totale di eventi elementari}}$$

Si ha quindi:

$$
\begin{array}{rcl}
\emptyset & \longrightarrow & 0 \\
\{T\} & \longrightarrow & 1/2 \\
\{C\} & \longrightarrow & 1/2 \\
\{T, C\} & \longrightarrow & 1
\end{array}
$$

Questo approccio (detto *classico*) è senz'altro valido quando si trattano casi in cui il numero di elementi di $\mathcal{S}$ è finito (o è un'infinità numerabile).

Un diverso approccio nell'interpretazione della probabilità (e quindi anche nella sua assegnazione agli eventi) è il cosidetto approccio *frequentista*.

In particolare, si assume di poter ripetere l'esperimento un numero grande (infinito!) di volte e di contare il numero di volte in cui si verifica l'evento A. La probabilità dell'evento A sarà pari al rapporto (frequenza) tra il numero di volte in cui A si è verificato e il numero di esperimenti eseguiti:

$$\Pr(A) = \frac{\sharp\text{di esperimenti in cui } A \text{ si è verificato}}{\sharp\text{totale degli esperimenti svolti}}$$

È plausibile pensare che nell'esempio del lancio della moneta si arriverebbe ad assegnare le stesse probabilità ottenute secondo l'approccio classico.

Un ulteriore interpretazione della probabilità è dovuta all'approccio *soggettivista*, secondo il quale la probabilità di un determinato evento A va interpretata (assegnata) come la valutazione che il singolo individuo può coerentemente formulare, in base alle proprie conoscenze, del grado di avverabilità dell'evento.

Anche in questo caso, se il soggetto non ha ragione di pensare che la moneta sia truccata, sarà plausibile che assegni eguale probabilità al verificarsi dell'esito testa e dell'esito croce.

In determinati contesti, l'interpretazione soggettivista può essere più sensata degli altri due approcci. Si pensi, ad esempio, a come definire la probabilità che il Venezia batta il Real Madrid in un (MOLTO) ipotetico incontro di Champions League.

Gli assiomi della probabilità (di Kolmogorov)

Qualunque funzione, indipendentemente da come la costruisco, che ad ogni elemento di $\mathcal{F}$ associa un numero reale può essere considerata una probabilità, se essa soddisfa i seguenti assiomi:

1° **Assioma.** Ogni evento ha probabilità maggiore o uguale a 0:

$$\Pr(A) \geq 0$$

2º **Assioma.** La probabilità dello spazio campionario è pari ad 1

$$\Pr(\mathcal{S}) = 1$$

3º **Assioma.** La probabilità dell'unione di eventi disgiunti deve essere pari alla somma delle probabilità

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) \quad \text{se} \quad A \cap B = \emptyset$$

Il terzo assioma deve valere anche

1. per successioni finite di eventi mutualmente disgiunti, cioè

$$\Pr(\textstyle\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \Pr(A_i) \quad \text{se} \quad A_i \cap A_j = \emptyset \quad \text{per ogni } i \neq j$$

2. per successioni infinite di eventi mutualmente disgiunti, cioè

$$\Pr(\textstyle\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \Pr(A_i) \quad \text{se} \quad A_i \cap A_j = \emptyset \quad \text{per ogni } i \neq j$$

Alcune conseguenze degli assiomi

Gli assiomi della probabilità hanno una serie di conseguenze (facilmente dimostrabili):

1. $\Pr(\emptyset) = 0$
2. $\Pr(A) \leq 1$
3. $\Pr(A^c) = 1 - \Pr(A)$
4. Se $A \subseteq B$ allora $\Pr(A) \leq \Pr(B)$
5. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

   Notiamo che, se A e B sono incompatibili (cioè se $A \cap B = \emptyset$), riotteniamo il terzo assioma.

   Più in generale invece, $\Pr(A \cap B) \geq 0$, e quindi $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$.

Suggerimento: provate a verificare queste relazioni anche facendo ricorso ai diagrammi di Venn (l'area dell'insieme nel diagramma rappresenta la probabilità del corrispondente evento. . .)

Esempio

In un'urna vi sono quattro palline, ognuna contrassegnata con un numero, da 1 a 4. L'esperimento consiste nell'estrazione di una pallina. **Lo spazio campionario**:

$$\mathcal{S} = \{1, 2, 3, 4\} \; ;$$

Il catalogo:

$$\begin{aligned} \mathcal{F} = \quad & \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\} \\ & \{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}, \underbrace{\{1,2,3,4\}}_{= \mathcal{S}}\} \end{aligned}$$

Si hanno $16 = 2^4$ elementi di $\mathcal{F}$.

La funzione di probabilità:

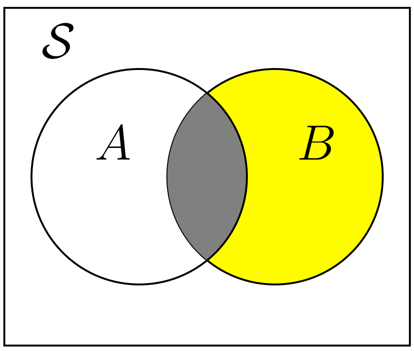
$$\Pr(A) = \frac{\sharp \text{di eventi elementari che portano alla realizzazione dell'evento } A}{\sharp \text{totale di eventi elementari}}$$

per cui, ad esempio:

$$\begin{aligned} \Pr(i) &= \frac{1}{4} \quad \text{per } i = 1, 2, 3, 4 \\ \Pr(\text{estrarre o 1 o 2}) &= \Pr(1,2) = \Pr(1) + \Pr(2) = \frac{2}{4} = \frac{1}{2} \\ \Pr(\text{estrarre o 1 o 2 o 3}) &= \Pr(1,2,3) = \Pr(1) + \Pr(2) + \Pr(3) = \frac{3}{4} \\ \Pr(\text{estrarre un numero pari}) &= \Pr(2,4) = \frac{1}{2} \end{aligned}$$

La probabilità condizionata

Si vuole valutare la probabilità di un evento sapendo che si è già realizzato un altro evento.



Sappiamo che si è già realizzato B. Vogliamo sapere che probabilità ha la realizzazione di A. Ma, se si è già realizzato B, gli unici eventi elementari che si possono ancora realizzare sono quelli di A contenuti anche in B.

Questo vuol dire che si ha un nuovo spazio campionario $\mathcal{S}^* = B$. Volendo valutare la probabilità di A dato che si è realizzato B, allora occorre considerare gli eventi elementari comuni ad A ed a B, cioè quelli appartenenti alla loro intersezione. Chiaramente non possiamo prendere *tout court* la probabilità dell'intersezione, perchè questa si riferisce allo spazio campionario $\mathcal{S}$ e non al nuovo. La probabilità dell'intersezione va, dunque, "normalizzata", in modo che la probabilità del nuovo spazio campionario si pari ad uno.

La formula della probabilità condizionata è data da:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \quad \text{se } \Pr(B) \neq 0$$

dove $\Pr(A|B)$ indica la probabilità di A dato B o subordinata o condizionata a B. L'evento B si dice "condizionante". Chiaramente la probabilità condizionata è definita se e solo se l'evento condizionante, B, ha probabilità diversa da zero di realizzarsi.

Esempio 8 (Il lancio di un dado)**.** *Sapendo che si è realizzata una faccia pari, si vuole calcolare la probabilità che si realizzi un 2.*

- *L'evento condizionante B è dato da: $B = \{2, 4, 6\}$ ed ha una probabilità di realizzarsi pari a $\Pr(B) = 3/6 = 1/2$.*

- *L'evento A è dato da: $A = \{2\}$.*

- *L'evento intersezione è*

$$\begin{aligned} A \cap B &= \{\textit{si realizza una faccia pari ed essa è uguale a 2}\} \\ &= \{2\} = A \end{aligned}$$

perchè $A \subset B$. La sua probabilità è

$$\Pr(A \cap B) = \Pr(A) = \frac{1}{6}$$

- *La probabilità di A dato B è*

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/6}{1/2} = \frac{1}{3}$$

D'altra parte, se sappiamo che si è realizzata una faccia pari, i possibili risultati sono soltanto $\{2, 4, 6\}$ e la probabilità che tra questi si realizzi il 2 è pari a 1/3.

Esempio 9 (Due estrazioni da un'urna)**.** *Un'urna contiene 5 palline: 3 bianche, indicate con w_1, w_2, w_3, e 2 rosse, r_1 e r_2. L'esperimento casuale consiste nell'estrarre due palline una alla volta, ma senza reinserire la prima estratta.*

Poichè le estrazioni sono due, gli eventi elementari sono le coppie e lo spazio campionario è dato da:

$$\begin{aligned}\mathcal{S} = \{ \quad & (w_1, w_2), (w_1, w_3), (w_1, r_1), (w_1, r_2), (w_2, w_1), \\ & (w_2, w_3), (w_2, r_1), (w_2, r_2), (w_3, w_1), (w_3, w_2), \\ & (w_3, r_1), (w_3, r_2), (r_1, w_1), (r_1, w_2), (r_1, w_3), \\ & (r_1, r_2), (r_2, w_1), (r_2, w_2), (r_2, w_3), (r_2, r_1)\}\end{aligned}$$

Si hanno quindi 20 casi possibili. Si noti che in $\mathcal{S}$

1. *non sono presenti le coppie con elementi uguali, tipo (w_1, w_1), perchè la seconda estrazione avviene senza il reinserimento della prima;*

2. *occorre tener conto dell'ordine, cioè ad esempio l'evento elementare (w_1, r_1) è diverso da (r_1, w_1), perchè si fanno due estrazioni.*

Si vuole determinare la probabilità di avere nella seconda estrazione una pallina rossa sapendo che la prima estratta è stata una pallina bianca.

1. *L'evento condizionante è*

$$\begin{aligned} B &= \{\textit{la prima estratta è una pallina bianca}\} \\ &= \{(w_1, w_2), (w_1, w_3), (w_1, r_1), (w_1, r_2), (w_2, w_1), \\ &\quad (w_2, w_3), (w_2, r_1), (w_2, r_2), (w_3, w_1), (w_3, w_2), (w_3, r_1), (w_3, r_2)\}\end{aligned}$$

   con 12 eventi elementari favorevoli. Dunque la sua probabilità è data da

$$\Pr(B) = \frac{12}{20} = \frac{3}{5}$$

2. *L'evento*

$$A = \{\textit{nella seconda estrazione si ottiene una pallina rossa}\}$$

   è dato da

$$\begin{aligned} A &= \{(w_1, r_1), (w_1, r_2), (w_2, r_1), (w_2, r_2), \\ &\quad (w_3, r_1), (w_3, r_2), (r_1, r_2), (r_2, r_1)\}\end{aligned}$$

3. *La sua intersezione con B è data da*

$$A \cap B = \{(w_1, r_1), (w_1, r_2), (w_2, r_1), (w_2, r_2), (w_3, r_1), (w_3, r_2)\}$$

   e la probabilità di realizzarsi di $A \cap B$ è:

$$\Pr(A \cap B) = \frac{6}{20} = \frac{3}{10}$$

4. *allora la probabilità condizionata di A dato B è pari a*

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{3/10}{3/5} = \frac{1}{2}$$

D'altra parte, se nella prima estrazione si è realizzata una pallina bianca, poichè questa non viene reinserita nell'urna, lo spazio campionario per la seconda estrazione è:

$$\begin{aligned} \mathcal{S}^* = & \{(w_1, w_2), (w_1, w_3), (w_1, r_1), (w_1, r_2), (w_2, w_1), (w_2, w_3), \\ & (w_2, r_1), (w_2, r_2), (w_3, w_1), (w_3, w_2), (w_3, r_1), (w_3, r_2)\} \\ = & B \end{aligned}$$

con 12 eventi elementari. I casi favorevoli ad ottenere in $\mathcal{S}^*$ una pallina rossa sono 6, quindi la probabilità di estrarre da $\mathcal{S}^*$ una pallina rossa è $6/12 = 1/2$.

Esempio 10 (Il gioco del lotto (semplificato))**.** *Due estrazioni da un'urna contenente 4 palline uguali, numerate da 1 a 4. Dopo la prima estrazione la pallina estratta non viene reinserita nell'urna.*

Gli eventi elementari sono le coppie di numeri (i, j) con $i = 1, 2, 3, 4$ e $j = 1, 2, 3, 4$.

Lo spazio campionario contiene tutti gli eventi elementari esclusi quelli per i quali $i = j$. Infatti non si possono realizzare gli eventi $(1, 1)$, $(2, 2)$, $(3, 3)$ e $(4, 4)$, poichè non viene reinserito il primo estratto.

$$\mathcal{S} = \left\{ \begin{array}{llll} & (1,2) & (1,3) & (1,4) \\ (2,1) & & (2,3) & (2,4) \\ (3,1) & (3,2) & & (3,4) \\ (4,1) & (4,2) & (4,3) & \end{array} \right\}$$

$\mathcal{S}$ contiene 12 elementi.

Vogliamo calcolare la probabilità che nella seconda estrazione si realizzi un 2, dato che nella prima si è realizzato un 1.

1. *L'evento $B = \{$il primo numero estratto è 1$\}$ ha tre eventi favorevoli e la sua proabilità è pari a:*

$$\Pr(B) = \frac{3}{12} = \frac{1}{4}$$

2. *L'intersezione tra l'evento B e l'evento*

$$A = \{\text{nella seconda estrazione si ha un } 2\}$$

è costituita dal'evento $(1,2)$ *per cui la sua probabilità è pari a:*

$$\Pr(A \cap B) = \frac{1}{12}$$

3. *la probabilità dell'evento* A *subordinata a* B *è allora pari a:*

$$\Pr(A|B) = \frac{1/12}{1/4} = \frac{1}{3}$$

Controllate che lo stesso risultato si ottiene considerando lo spazio campionario che si forma dopo la prima estrazione.

Teorema delle probabilità composte o regola moltiplicativa delle probabilità

Dalla formula della probabilità condizionata:

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B)$$

ma anche

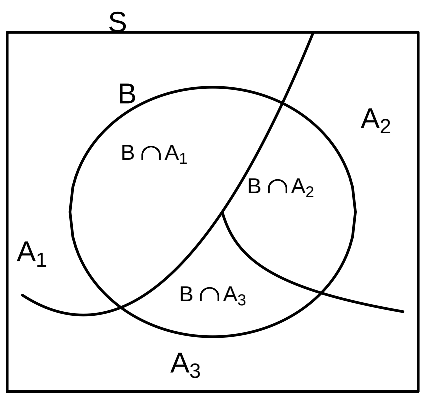
$$\Pr(A \cap B) = \Pr(B|A)\Pr(A).$$

• Legge delle probabilità totali

Se una successione di eventi A_i, con $i = 1, \cdots, n$, costituisce una partizione dello spazio campionario $\mathcal{S}$, cioè $\bigcup_{i=1}^{n} A_i = \mathcal{S}$ e $A_i \cap A_j = \emptyset$ per $i \neq j$, allora la probabilità di un qualunque evento B è data da:

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B|A_i) \cdot \Pr(A_i)$$

Nella figura il caso $n = 3$



$$\begin{aligned}\Pr(B) &= \Pr(B\cap A_1)+\Pr(B\cap A_2)+\Pr(B\cap A_3)\\ &= \Pr(B|A_1)\cdot\Pr(A_1)+\Pr(B|A_2)\cdot\Pr(A_2)+\Pr(B|A_3)\cdot\Pr(A_3)\,.\end{aligned}$$

Le formule di Bayes

Dalla formula della probabilità condizionata sappiamo che

$$\Pr(A\cap B)=\Pr(A|B)\Pr(B)=\Pr(B|A)\Pr(A)$$

da cui le due possibili formule

$$\left.\begin{array}{l}\Pr(A|B)=\dfrac{\Pr(B|A)\Pr(A)}{\Pr(B)}\\ \text{e}\\ \Pr(B|A)=\dfrac{\Pr(A|B)\Pr(B)}{\Pr(A)}\end{array}\right\}\text{Formule di Bayes}$$

entrambe valide se i denominatori sono diversi da zero.

Esempio 11 (Applicazione della formula di Bayes e di quella delle probabilità totale)**.** *Si hanno 4 scatole di pezzi di ricambio con la seguente composizione:*

Prima scatola:	*contiene 2000 pezzi di cui 100 difettosi*
Seconda scatola:	*contiene 500 pezzi di cui 200 difettosi*
Terza scatola:	*contiene 1000 pezzi di cui 100 di difettosi*
Quarta scatola:	*contiene 1000 pezzi di cui 100 difettosi*

Viene scelta a caso una scatola, da cui viene estratto un pezzo.

1. *Qual è la probabilità che questo componente sia difettoso?*

   Le quattro scatole rappresentano la partizione dello spazio campionario, i cui eventi elementari sono tutti i pezzi di ricambio.

   Tutte le scatole hanno la stessa probabilità di essere estratte ed essa è pari a:

   $$\Pr(Scatola_i)=\frac{1}{4}$$

   La percentuale di pezzi difettosi in ogni scatola costituisce la probabilità di estrarre un pezzo difettoso dato che si è scelta una particolare scatola, cioè

   $$\Pr(pezzo\ difettoso|Scatola_1)=100/2000=0{,}05$$
   $$\Pr(pezzo\ difettoso|Scatola_2)=200/500=0{,}40$$
   $$\Pr(pezzo\ difettoso|Scatola_3)=100/1000=0{,}10$$
   $$\Pr(pezzo\ difettoso|Scatola_4)=100/1000=0{,}10$$

Applicando la formula delle probabilità totali si ottiene la probabilità di estrarre un pezzo difettoso **qualunque sia la scatola da cui esso è stato estratto***:*

$$\begin{aligned}
\Pr(\textit{pezzo difettoso}) &= \Pr(\textit{pezzo difettoso}|\textit{Scatola}_1)\cdot\Pr(\textit{Scatola}_1)\\
&+ \Pr(\textit{pezzo difettoso}|\textit{Scatola}_2)\cdot\Pr(\textit{Scatola}_2)\\
&+ \Pr(\textit{pezzo difettoso}|\textit{Scatola}_3)\cdot\Pr(\textit{Scatola}_3)\\
&+ \Pr(\textit{pezzo difettoso}|\textit{Scatola}_4)\cdot\Pr(\textit{Scatola}_4)\\
&= (0{,}05)\cdot(0{,}25)+(0{,}4)\cdot(0{,}25)\\
&\quad +(0{,}1)\cdot(0{,}25)+(0{,}1)\cdot(0{,}25)\\
&= 0{,}1625
\end{aligned}$$

2. *Esaminiamo il pezzo e lo troviamo difettoso. Qual è la probabilità che provenga,ad esempio, dalla scatola 2.*

$$\begin{aligned}
&= \Pr(\textit{Scatola}_2|\textit{pezzo difettoso})\\
&= \frac{\Pr(\textit{pezzo difettoso}|\textit{Scatola}_2)\Pr(\textit{Scatola}_2)}{\Pr(\textit{pezzo difettoso})}\\
&= \frac{0{,}4\cdot 0{,}25}{0{,}1625} = 0{,}615 \quad \textit{(Formula di Bayes)}
\end{aligned}$$

Notiamo che, senza sapere che il pezzo è difettoso, la probabilità di aver estratto il pezzo dalla seconda scatola coincide con la probabilità di aver scelto la seconda scatola, e cioè $1/4 = 0{,}25$*. Invece, sapendo che il pezzo è difettoso, questa probabilità viene aggiornata e cresce notevolmente.*

Esercizio 11. *Provate a calcolare l'analoga probabilità anche per le altre tre scatole.*

Teorema di Bayes

Le formule di Bayes precedentemente visto possono essere riformulate nel modo seguente.

Proposizione 1. *Sia* $B \in \mathcal{F}$ *e siano* $A_1, A_2, \ldots, A_n$ *n eventi in* $\mathcal{F}$ *che costituiscono una partizione di* $\mathcal{S}$*. Se conosciamo le probabilità* $\Pr(A_i)$ *e* $\Pr(B|A_i)$*, con* $i = 1, \ldots, n$*, allora*

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_{j=1}^{n}\Pr(B|A_j)\Pr(A_j)}$$

Indipendenza stocastica

Due eventi sono stocasticamente indipendenti se e solo se

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B) \,.$$

L'indipendenza stocastica significa che il conoscere A non altera la probabilità di B (è vero anche il viceversa).

Infatti, applicando la formula della probabilità condizionata e tenendo conto della condizione di indipendenza, si ha

$$\begin{aligned}
\Pr(A|B) &= \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A) \cdot \Pr(B)}{\Pr(B)} = \Pr(A), \\
\Pr(B|A) &= \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(A) \cdot \Pr(B)}{\Pr(A)} = \Pr(B) \,.
\end{aligned}$$

Tre eventi sono stocasticamente indipendenti se e solo se

1. $\Pr(A_1 \cap A_2 \cap A_3) = \Pr(A_1) \Pr(A_2) \Pr(A_3)$
2. $\Pr(A_i \cap A_j) = \Pr(A_i) \cdot \Pr(A_j)$ per $i \neq j$.

Entrambe le condizioni devono essere soddisfatte. Quest'ultimo risultato è generalizzabile a più di tre eventi.

Ma come si calcola la probabilità? (digressione sul calcolo combinatorio)

La probabilità di un evento secondo l'*approccio classico* si definisce come il rapporto tra il numero di casi (eventi elementari) favorevoli al verificarsi dell'evento e il numero dei casi possibili.

Il calcolo combinatorio serve a contare i casi favorevoli e i casi possibili quando abbiamo a che fare con un numero finito di elementi.

Quindi vogliamo contare quante siano, a partire da un insieme $\mathcal{S}$ di n elementi, le composizioni che si possono formare con k elementi di $\mathcal{S}$.

Possiamo supporre, per semplicità, di avere un'urna con n palline e di estrarne k.

Il numero di possibili k-ple dipende dal tipo di estrazione: se dopo aver estratto una pallina la reinseriamo nell'urna prima di procedere ad una nuova estrazione si parla di *estrazioni con reinserimento*; viceversa, se non reinseriamo le palline, si parla di *estrazioni in blocco* o *senza reinserimento.*

Indipendentemente dal tipo di estrazione, il numero di k-ple dipende anche dal fatto di considerare due k-ple diverse se hanno gli stessi elementi ma in ordine diverso.

Disposizioni con ripetizione

Definizione 8 (Disposizioni con ripetizione di n oggetti in classe k)**.** *Numero di gruppi di k oggetti di $\mathcal{S}$ che si possono formare considerando diversi gruppi formati da elementi diversi o dai medesimi elementi purché disposti in modo diverso (l'ordine importa), ammettendo la possibilità che in essi un elemento possa ripetersi (estrazioni con reinserimento).*

$$D'_{n,k} = n^k$$

Numero di coppie di numeri che si possono ottenere dal lancio di due dadi: $6^2 = 36$

$$\mathcal{S} = \{A, B, C\}, \quad n = 3, \quad k = 2$$

$$\left\{ \begin{array}{cc} A & A \\ A & B \\ A & C \\ B & A \\ B & B \\ B & C \\ C & A \\ C & B \\ C & C \end{array} \right\}$$

Disposizioni senza ripetizione

Definizione 9 (Disposizioni senza ripetizione di n oggetti in classe k)**.** *Numero di gruppi di k oggetti di $\mathcal{S}$ che si possono formare considerando diversi gruppi formati da elementi diversi o dai medesimi elementi purché disposti in modo diverso (l'ordine importa), negando la possibilità che in essi un elemento possa ripetersi (estrazioni senza reinserimento).*

$$\begin{aligned} D_{n,k} &= n \cdot (n-1) \cdot \ldots \cdot (n-k+2) \cdot (n-k+1) \\ &= \prod_{i=0}^{k-1} (n-i) \end{aligned}$$

$$\mathcal{S} = \{A, B, C\}, \quad n = 3, \quad k = 2$$

$$\left\{ \begin{array}{cc} A & B \\ A & C \\ B & A \\ B & C \\ C & A \\ C & B \end{array} \right\}$$

Permutazioni

Definizione 10 (Permutazioni di n oggetti)**.** *Numero di gruppi di n oggetti di $\mathcal{S}$, distinti per l'ordine con cui essi si dispongono.*

$$\begin{aligned} P_n &= n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 \\ &= n! \quad (n \text{ fattoriale}) \\ (0! &= 1 \text{ per convenzione}) \end{aligned}$$

P_n è equivalente a $D_{n,n}$.

NOTA: le disposizioni semplici si possono scrivere in termini di fattoriali

$$\begin{aligned} D_{n,k} &= \frac{n!}{(n-k)!} \\ &= \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+2) \cdot (n-k+1) \cdot (n-k)!}{(n-k)!} \\ &= n \cdot (n-1) \cdot \ldots \cdot (n-k+2) \cdot (n-k+1) \,. \end{aligned}$$

$$\mathcal{S} = \{A, B, C\}, \quad n = 3$$

$$\left\{ \begin{array}{ccc} A & B & C \\ A & C & B \\ B & A & C \\ B & C & A \\ C & A & B \\ C & B & A \end{array} \right\}$$

Combinazioni

Definizione 11 (Combinazioni semplici di n oggetti in classe k)**.** *Numero di gruppi di k oggetti di $\mathcal{S}$ che si possono formare considerando diversi gruppi formati da elementi diversi, a prescindere dall'ordine con cui essi si dispongono (l'ordine non importa) e negando la possibilità che in essi un elemento possa ripetersi (estrazioni senza reinserimento).*

$$\begin{aligned} C_{n,k} &= \frac{n!}{k!(n-k)!} \\ &= \frac{D_{n,k}}{k!} \\ &= \frac{\prod_{i=0}^{k-1}(n-i)}{k!} \\ &= \binom{n}{k} \end{aligned}$$

$$\boxed{\begin{gathered} \mathcal{S} = \{A, B, C\}, \quad n = 3, \quad k = 2 \\ \left\{ \begin{array}{cc} A & B \\ A & C \\ B & C \end{array} \right\} \end{gathered}}$$

Esempio 12 (Carte)**.** *Supponiamo di giocare a scopone scientifico. Il mazzo ha* 40 *carte e ne vengono assegnate* 10 *ad ogni giocatore. Quanti sono i possibili gruppi di 10 carte che un giocatore può ricevere?*

L'estrazione delle carte è evidentemente in blocco (senza reinserimento) e, ai fini del gioco, non ha rilevanza in che ordine il giocatore riceve le 10 *carte. Quindi dobbiamo considerare le combinazioni di* 40 *elementi in classe* 10*:*

$$C_{40,10} = \binom{40}{10} = \frac{40!}{10!\,30!} = 847660528\,.$$

Se vogliamo calcolare la probabilità che un giocatore abbia tutti e quattro gli assi, il numero di casi favorevoli a questo evento sono tutti i gruppi di 10 *carte che hanno al loro interno i quattro assi e un qualsiasi sottogruppo di 6 carte scelto tra le rimanenti 36:*

$$C_{36,6} = \binom{36}{6} = \frac{36!}{6!\,30!} = 1947792\,.$$

La probabilità e quindi pari a:

$$\Pr(\text{“}avere\ 4\ assi\text{”}) = \frac{1947792}{847660528} = 0{,}002297844\,.$$

Esempio 13 (Lotto). *Si estraggono 5 numeri da un'urna contenente i primi 90 numeri. Nelle estrazioni del lotto le palline estratte non vengono reinserite nell'urna ed è importante sapere l'ordine in cui i vari numeri vengono estratti (si pensi al Superenalotto, che considera il primo estratto di certe estrazioni).*

Il numero di possibili cinquine ordinate è pari alle disposizioni di 90 *oggetti in classe* 5*:*

$$D_{90,5} = 90 \cdot 89 \cdot 88 \cdot 87 \cdot 86 = 5273912160\,.$$

Quindi la probabilità di estrarre una particolare cinquina (ordinata), ad esempio $(34, 53, 12, 1, 89)$*, è pari a*

$$\Pr\{(34, 53, 12, 1, 89)\} = \frac{1}{5273912160} = 1{,}896126e - 10\,,$$

e cioè prossima a zero.

2.2 Variabili casuali

Definizione di variabile casuale

Lo spazio campionario può non essere un insieme di numeri. Il lancio di una moneta non da luogo ad un numero.

Vogliamo definire una variabile Y che conta il numero di teste che si possono avere nel lancio di una moneta. Questa variabile può assumere i valori 0 o 1. Ad ogni elemento dello spazio campionario $\mathcal{S} = \{T, C\}$ associamo 0 se otteniamo Croce, 1 se otteniamo Testa.

$$Y(s) = \begin{cases} 0, & \text{se } s = C \\ 1, & \text{se } s = T \end{cases}$$

Definiamo ora una funzione che associa ad un qualunque spazio campionario un insieme numerico.

Definizione 12 (Variabile casuale)**.** *Una **variabile casuale** Y (indicata per brevità anche con **v.c.**) in uno spazio campionario $\mathcal{S}$ è una qualunque funzione definita in $\mathcal{S}$ e con valori nell'insieme dei numeri reali (e tale che l'immagine inversa di ogni intervallo di numeri reali sia un evento - ma questo discorso ci porterebbe lontano).*

Come assegnamo le probabilità ai valori di Y? Se ammettiamo che la moneta sia equilibrata, allora è naturale porre

$$\Pr(Y(s) = 1) = \Pr(T) = 1/2, \quad \Pr(Y(s) = 0) = \Pr(C) = 1/2.$$

Supponiamo ora di avere un diverso esperimento: lanciamo un dado (regolarmente bilanciato) e costruiamo la variabile casuale

$$Y(s) = \begin{cases} 0, & \text{se } s = \text{numero pari} \\ 1, & \text{se } s = \text{numero dispari} \end{cases}$$

In questo caso, $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$ e ci sono più eventi elementari che danno luogo allo stesso valore di Y. Quindi si ha

$$\begin{aligned} \Pr(Y(s) = 1) &= \Pr(\{1, 3, 5\}) = \Pr(1 \cup 3 \cup 5) \\ &= \Pr(1) + \Pr(3) + \Pr(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \\ \Pr(Y(s) = 0) &= \Pr(\{2, 4, 6\}) = \Pr(2 \cup 4 \cup 6) \\ &= \Pr(2) + \Pr(4) + \Pr(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}. \end{aligned}$$

Notiamo che due diversi esperimenti hanno portato alla costruzione della stessa variabile casuale Y. Nel secondo caso, a differenza del primo, non c'è corrispondenza biunivoca tra gli elementi di $\mathcal{S}$ e i valori di Y.

In generale, si avrà che a diversi elementi di $\mathcal{S}$ corrisponderà lo stesso numero a. Per cui definiremo la probabilità che Y sia uguale ad a, e scriveremo

$$\Pr(Y = a),$$

come la probabilità dell'unione di tutti gli eventi elementari che la funzione Y associa al valore a:

$$\Pr(Y = a) = \Pr(\{s \in \mathcal{S} : \ Y(s) = a\}) .$$

Più in generale, se $a < b$, si avrà ad esempio che

$$\Pr(a < Y \leq b) = \Pr(\{s \in \mathcal{S} : \ a < Y(s) \leq b\}) .$$

Variabili casuali discrete

Sia Y una variabile casuale e supponiamo per ora che Y assuma solo un numero finito di valori (reali):

$$Y(\mathcal{S}) = \{y_1, y_2, \ldots, y_k\}$$

L'insieme di valori che Y può assumere è detto anche **supporto** della variabile casuale.

Assegnamo ad ogni y_i le probabilità

$$\Pr(Y = y_i) = \Pr(\{s \in \mathcal{S} : \ Y(s) = y_i\})$$

quantità che scriviamo più semplicemente come $p(y_i)$

Questo definisce una funzione $p(\cdot)$ in $Y(\mathcal{S})$, funzione che ha queste proprietà (conseguenze degli assiomi della probabilità):

1. $0 \leq p(y_i) \leq 1$, (per $i = 1, \ldots, k$)
2. $\sum_{i=1}^{k} p(y_i) = 1$

che chiamiamo **funzione di probabilità** della variabile casuale (discreta) Y.

Funzione di probabilità

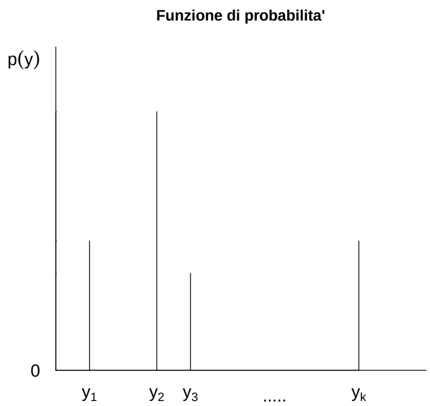
Una v.c. discreta è nota se si conoscono i valori che può assumere (supporto) e le rispettive probabilità (funzione di probabilità).

Possiamo rappresentare la funzione di probabilità con una tabella

y_1	y_2	$\cdots$	y_k
$p(y_1)$	$p(y_2)$	$\cdots$	$p(y_k)$

Notate qualche analogia con concetti già visti?

Graficamente si può rappresentare con un diagramma a bastoncini:



Il punteggio totale riportato nel lancio di due dadi

Il nostro esperimento casuale consiste nel lancio di due dadi equilibrati. Abbiamo pertanto uno spazio campionario finito con eventi elementari equiprobabili costituito da 36 coppie ordinate

$$\mathcal{S} = \{(1,1),(1,2),\ldots,(6,5),(6,6)\}$$

Sia Y la funzione che ad ogni elemento di $\mathcal{S}$ associa la somma dei punti riportati sulle due facce

$$Y(a,b) = a + b$$

Allora Y è una variabile casuale con valori

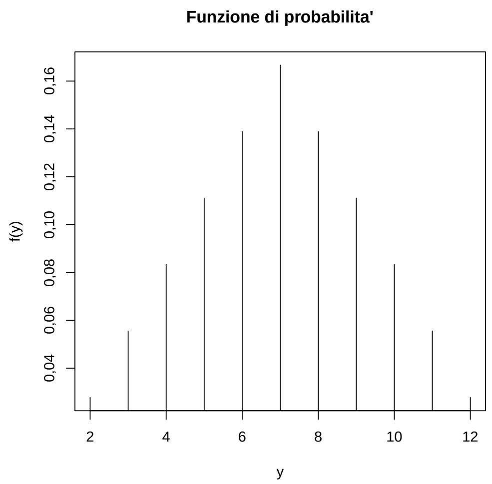
$$Y(\mathcal{S}) = \{2,3,\ldots,12\}$$

(si sarebbe potuto definire un'altra variabile casuale, ad esempio $Y(a,b) = a/\sqrt{b}$, ma forse non incontreremo mai questa variabile casuale...).

La funzione di probabilità è la seguente

$$\begin{aligned}
\Pr(Y=2) &= \Pr\{(1,1)\} = \frac{1}{36}\,, \\
\Pr(Y=3) &= \Pr\{(1,2)\cup(2,1)\} = \Pr\{(1,2)\}+\Pr\{(2,1)\} \\
&= \frac{1}{36}+\frac{1}{36} = \frac{2}{36}\,, \\
&\vdots \\
\Pr(Y=12) &= \Pr\{(6,6)\} = \frac{1}{36}\,.
\end{aligned}$$

y_i	2	3	4	5	6	7	8	9	10	11	12
$p(y_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



Una v.c. discreta che assume infiniti valori

Una v.c. discreta può assumere anche un numero infinito di valori, purché numerabile.

Consideriamo il seguente esperimento: un'urna contiene N palline nere e R palline bianche. Supponiamo di fare estrazioni con reinserimento da quest'urna e di considerare la v.c. $Y =$"**numero di estrazioni necessarie per ottenere una pallina nera la prima volta**".

Se indichiamo con E_i l'evento "pallina nera all'i-esima estrazione", si ha $\theta = \Pr(E_i) = N/(N+R),\ \forall i$.

La v.c. Y può assumere i valori $\{1, 2, 3, \ldots, i, \ldots\}$ con le seguenti probabilità:

$$\begin{aligned}
\Pr(Y = 1) &= \Pr(E_1) = \theta\,, \\
\Pr(Y = 2) &= \Pr(E_1^c \cap E_2) = \Pr(E_1^c)\Pr(E_2) = (1-\theta)\cdot\theta \\
\Pr(Y = 3) &= \Pr(E_1^c \cap E_2^c \cap E_3) = \Pr(E_1^c)\Pr(E_2^c)\Pr(E_3) = (1-\theta)^2\theta \\
&\vdots \\
\Pr(Y = i) &= \Pr(E_1^c \cap \cdots \cap E_{i-1}^c \cap E_i) = \Pr(E_1^c)\cdots\Pr(E_{i-1}^c)\Pr(E_i) \\
&= (1-\theta)^{i-1}\theta \\
&\vdots
\end{aligned}$$

y_i	1	2	3	$\cdots$	i	$\cdots$
$p(y_i)$	θ	$(1-\theta)\theta$	$(1-\theta)^2\theta$	$\cdots$	$(1-\theta)^{i-1}\theta$	$\cdots$

Essendo $0 < \theta < 1$, è ovvio che $p(y_i) \geq 0$.
Inoltre, sfruttando le proprietà della serie geometrica, è facile vedere che

$$\sum_{i=1}^{\infty} p(y_i) = \sum_{i=1}^{\infty}(1-\theta)^{i-1}\theta = \frac{\theta}{1-\theta}\sum_{i=1}^{\infty}(1-\theta)^i = 1\,.$$

Grazie a questo legame con la serie geometrica, la v.c. Y si dice avere **distribuzione geometrica**.

Possiamo calcolare la probabilità di qualsiasi evento di interesse; per esempio, si può calcolare la probabilità che siano necessarie più di 4 estrazioni:

$$\begin{aligned}
\Pr(Y > 4) &= 1 - \Pr(Y \leq 4) \\
&= 1 - [\Pr(Y = 1) + \Pr(Y = 2) + \Pr(Y = 3) + \Pr(Y = 4)] \\
&= 1 - [\theta + (1-\theta)\theta + (1-\theta)^2\theta + (1-\theta)^3\theta] \\
&= 1 - 4\theta + 6\theta^2 - 4\theta^3 + \theta^4 \\
&= (1-\theta)^4\,.
\end{aligned}$$

D'altra parte, questo risultato era ovvio, pensando che l'evento $(Y > 4)$ si verifica se e solo se non si estrae una pallina nera nelle prime 4 estrazioni. Quindi:

$$\begin{aligned}
\Pr(Y > 4) &= \Pr(E_1^c \cap E_2^c \cap E_3^c \cap E_4^c) \\
&= \Pr(E_1^c)\Pr(E_2^c)\Pr(E_3^c)\Pr(E_4^c) \\
&= (1-\theta)(1-\theta)(1-\theta)(1-\theta) \\
&= (1-\theta)^4\,.
\end{aligned}$$

La funzione di ripartizione

A questo punto risulta naturale introdurre la funzione di ripartizione della variabile casuale Y tramite la formula

$$F(y) = \Pr(Y \leq y) = \sum_{y_i \leq y} p(y_i).$$

Questa funzione è tale che

1. è non decrescente, ovvero se $a \leq b$, $F(a) \leq F(b)$.

2. $F(y) = 0$ se $y < \min(y_i)$ e $F(y) = 1$ se $y \geq \max(y_i)$.

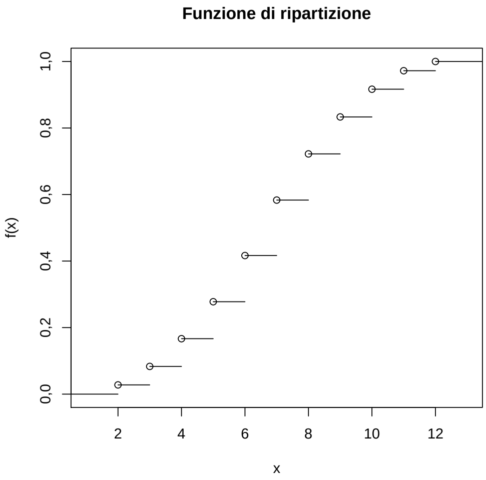
3. $F(y)$ è continua a destra: $\lim_{y \to y_0^+} F(y) = F(y_0)$[1].

La proprietà 1. è ovvia, la 2. è una conseguenza del fatto che comunque la funzione di ripartizione è una probabilità e quindi è compresa tra il limite inferiore 0 e quello superiore 1. Infine, la proprietà 3. ci dice che, se la v.c. è discreta, la funzione di ripartizione presenta delle discontinuità nei punti y in cui $\Pr(Y = y) > 0$.

Somma del punteggio di due dadi

$$F(y) = \begin{cases} 0, & y < 2 \\ 1/36, & 2 \leq y < 3 \\ 3/36, & 3 \leq y < 4 \\ 6/36, & 4 \leq y < 5 \\ 10/36, & 5 \leq y < 6 \\ 15/36, & 6 \leq y < 7 \\ 21/36, & 7 \leq y < 8 \\ 26/36, & 8 \leq y < 9 \\ 30/36, & 9 \leq y < 10 \\ 33/36, & 10 \leq y < 11 \\ 35/36, & 11 \leq y < 12 \\ 1, & y \geq 12 \end{cases}$$

[1] $y \to y_0^+$ vuol dire che y tende a y_0 da destra



Notiamo che, ad esempio,

$$\lim_{y \to 3^+} F(y) = F(3) = 3/36$$

ma

$$\lim_{y \to 3^-} F(y) = 1/36,$$

infatti $\Pr(Y = 3) = 2/36$.

A partire della funzione di ripartizione possiamo calcolare facilmente per $a < b$

$$\begin{aligned} F(b) = \Pr(Y \leq b) &= \Pr(\{Y \leq a\} \cup \{a < Y \leq b\}) \\ &= \Pr(\{Y \leq a) + \Pr(a < Y \leq b) \\ &= F(a) + \Pr(a < Y \leq b) \end{aligned}$$

e quindi

$$\Pr(a < Y \leq b) = F(b) - F(a)$$

Quindi, ad esempio,

$$\Pr(4 < Y \leq 10.2) = F(10.2) - F(4) = \frac{33}{36} - \frac{6}{36} = \frac{27}{36}\,.$$

Bisogna stare attenti ai “salti” di F(y)! Ad esempio,

$$\Pr(4 \leq Y \leq 10.2) = F(10.2) - F(4^-) = \frac{33}{36} - \frac{3}{36} = \frac{30}{36}\,.$$

Notiamo infine che $p(y)$ e $F(y)$ sono in corrispondenza biunivoca

$$p(y_0) = \Pr(Y = y_0) = \Pr(y_0 \leq Y \leq y_0) = F(y_0) - F(y_0^-)\,.$$

Valore atteso e varianza di una variabile casuale discreta

Data una v.c. discreta Y con supporto (per semplicità finito) $\{y_1, y_2, \ldots, y_k\}$ e con funzione di probabilità p, chiamiamo **valore atteso** di Y la quantità

$$\mathbb{E}(Y) = \sum_{i=1}^{k} y_i \, p(y_i)$$

ovvero $\mathbb{E}(Y)$ è una media pesata dei k valori y_i, con pesi $p(y_i)$, analogamente al caso della media aritmetica $\overline{y}$. Si suole porre $\mu = \mathbb{E}(Y)$.

Va osservata, tuttavia, la grande differenza tra i due concetti:

- $\overline{y}$ è il valore associato ad un particolare (e reale) insieme di dati;
- $\mathbb{E}(Y)$ esiste indipendentemente da qualsiasi rilevazione dei dati (prima di effettuare l'esperimento che genera Y possiamo calcolare $\mathbb{E}(Y)$).

Nel caso dell'esempio del lancio dei due dadi abbiamo

$$\begin{aligned} \mathbb{E}(Y) &= 2\,\frac{1}{36} + 3\,\frac{2}{36} + 4\,\frac{3}{36} + 5\,\frac{4}{36} + 6\,\frac{5}{36} + 7\,\frac{6}{36} \\ &+ 8\,\frac{5}{36} + 9\,\frac{4}{36} + 10\,\frac{3}{36} + 11\,\frac{2}{36} + 12\,\frac{1}{36} = 7 \end{aligned}$$

come si poteva dedurre dal grafico della funzione di distribuzione.

Analogamente a quanto succede per $\overline{y}$, si ha

Proposizione 2.

$$\mathbb{E}(a + bY) = a + b\mathbb{E}(Y)\,,$$

dove a e b sono due costanti.

Dimostrazione 1.

$$\sum_{i=1}^{k}(a + by_i)p(y_i) = a\sum_{i=1}^{k} p(y_i) + b\sum_{i=1}^{k} y_i p(y_i) = a + b\mathbb{E}(Y)$$

Più in generale, definiamo il **valore atteso** di $h(Y)$, dove h è una funzione qualsiasi, la quantità

$$\mathbb{E}(h(Y)) = \sum_{i=1}^{k} h(y_i)p(y_i)\,.$$

Esempio 14.

- $h(Y) = Y \Rightarrow \mathbb{E}(h(Y)) = \mathbb{E}(Y) = \mu$
- $h(Y) = Y^r$, $r = 1, 2, \ldots$ ***momento di ordine*** r

- $h(Y) = (Y - \mu)^r$, $r = 1, 2, \ldots$ ***momento centrato di ordine*** r

La varianza di Y corrisponde al caso $h(Y) = (Y - \mu)^2$.

$$\mathbb{V}ar(Y) = \sum_{i=1}^{k} (y_i - \mu)^2 p(y_i) = E((Y - \mu)^2)$$

La quantità $sqm(Y) = \sqrt{\mathbb{V}ar(Y)}$ è detta scarto quadratico medio di Y. Valgono qui considerazioni analoghe a quelle fatte a proposito della coppia $\overline{y}$, $\mathbb{E}(Y)$. Di solito si scrive σ^2 per indicare $\mathbb{V}ar(Y)$ e σ per $sqm(Y)$. Si dimostra facilmente che

$$\mathbb{V}ar(Y) = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2$$

e che

Proposizione 3.

$$\mathbb{V}ar(a + bY) = b^2 \mathbb{V}ar(Y)\,.$$

Trasformazione di una variabile casuale

Una funzione h di una variabile casuale Y è a sua volta una variabile casuale.

Infatti, ponendo $X = h(Y)$, questa è una funzione dai numeri reali nei numeri reali. Tuttavia, Y è una funzione che assegna ad ogni evento elementare ω un numero reale. Quindi, si ha

$$\omega \xrightarrow{\quad Y \quad} Y(\omega) \xrightarrow{\quad h \quad} h(Y(\omega))$$

che può essere riscritto come la composizione delle funzioni h e Y

$$\omega \xrightarrow{X = h \circ Y} X(\omega)\,.$$

Quindi X è una funzione che assegna ad ogni evento elementare un numero reale: X è una variabile casuale.

Conoscendo la funzione di probabilità di Y si può ottenere la funzione di probabilità di X nel modo seguente:

$$p_X(x) = \Pr(X = x) = \Pr(Y \in \{y : h(y) = x\})$$

e cioè sommando la probabilità di tutti i valori di Y che vengono assegnati al valore x dalla funzione h.

Ancora sul lancio dei due dadi

Abbiamo già definito la v.c. Y che rappresenta il punteggio totale ottenuto dal lancio di due dadi:

y	2	3	4	5	6	7	8	9	10	11	12
$p_Y(y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Definiamo la v.c. $X = h(Y) = Y - \mathbb{E}(Y) = Y - 7$ che è una trasformazione lineare (una semplice traslazione) della v.c. Y.

Dalla Proposizione [2] sappiamo che $\mathbb{E}(X) = \mathbb{E}(Y) - 7 = 7 - 7 = 0$. Determiniamo adesso l'intera funzione di probabilità di X; in questo caso, essendo X una funzione biunivoca di Y, è banale determinare $p_X(x)$:

y	2	3	4	5	6	7	8	9	10	11	12
x	-5	-4	-3	-2	-1	0	1	2	3	4	5
$p_X(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$$\begin{aligned} p_X(x) &= \Pr(X = x) = \Pr(Y - 7 = x) = \Pr(Y = 7 + x) \\ &= p_Y(7 + x)\,, \qquad x = -5, \ldots, 5 \end{aligned}$$

Quindi, ad esempio, $p_X(-3) = p_Y(7 - 3) = p_Y(4) = 3/36$.

Definiamo un'ulteriore v.c., considerando la seguente trasformazione della v.c. X appena definita: $Z = g(X) = |X|$, ossia la funzione valore assoluto di X.

In questo caso la funzione non è biunivoca, in quanto ci possono essere diversi valori di X che portano allo stesso valore di Z:

x	-5	-4	-3	-2	-1	0	1	2	3	4	5
z	5	4	3	2	1	0	1	2	3	4	5
$p_X(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

La funzione di probabilità di Z si ottiene nel modo seguente:

$$\begin{aligned} p_Z(0) &= \Pr(Z = 0) = \Pr(X = 0) = p_X(0) = 6/36 \\ p_Z(1) &= \Pr(Z = 1) = \Pr(\{X = -1\} \cup \{X = 1\}) \\ &= \Pr(X = -1) + \Pr(X = 1) = p_X(-1) + p_X(1) = 10/36 \\ &\vdots \\ p_Z(5) &= p_X(-5) + p_X(5) = 2/36 \end{aligned}$$

z	0	1	2	3	4	5
$p_Z(z)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

Esercizio 12. *Verificare che*

$$\mathbb{E}(Z) = \sum_{j=1}^{6} z_j p_Z(z_j) = \sum_{i=1}^{11} |x_i| p_X(x_i)\,.$$

La distribuzione binomiale

Consideriamo il seguente problema:

"qual è la probabilità di ottenere esattamente 3 volte il 4 in 15 lanci di un dado ?"

Il problema può sembrare poco interessante o perlomeno accademico, ma la sua risoluzione ci porta all'incontro con una variabile casuale di grande importanza nelle applicazioni.

Il problema rientra nella classe di problemi in cui un esperimento elementare con due possibili risultati, detti successo e insuccesso, viene ripetuto indipendentemente m volte e ci chiediamo qual è la probabilità di ottenere k successi sapendo che la probabilità di successo in ogni prova è p.

Nel nostro caso l'esperimento elementare è il lancio del dado, il successo è "appare il 4", l'insuccesso è "non appare il 4", m è 15 , k è 3 e p ammettendo equiprobabilità delle facce è 1/6.

Una precisazione: qui l'esperimento casuale è l'insieme dei 15 lanci (supposti indipendenti l'uno dell'altro: possiamo anche pensare di lanciare simultaneamente 15 dadi), cioè l'insieme dei 15 esperimenti elementari.

Lo spazio campionario è quindi costituito dalle sequenze del tipo

$$fffffsffffsffff$$

dove s sta per successo (esce il 4) e f sta per fallimento (non esce il 4). Il numero di queste sequenze è $2^{15} = 32768$.

Ritornando al caso generale, lo spazio campionario $\mathcal{S}$ è l'insieme costituito dalle 2^m sequenze di f e s di lunghezza m e noi definiamo una variabile casuale B, in questo modo: se sequenza $\in \mathcal{S}$

$$B(\text{sequenza}) = \text{numero di } s \text{ nella sequenza}$$

cioè, la variabile casuale B "conta" il numero di successi in m prove indipendenti. È chiaro che

$$B(\mathcal{S}) = \{0, 1, \ldots, m\}$$

e vogliamo ora determinare la funzione di probabilità di B.

Dobbiamo determinare $\Pr(B = k)$ con $0 \leq k \leq m$ e questo non è altro che risolvere nel caso generale il problema che ci eravamo posti per il dado.

B è uguale a k se si verifica una sequenza con k lettere s e $m - k$ lettere f. Ad esempio, se $m = 3$ e $k = 2$, le possibili sequenze sono: ssf, sfs, fss. Inoltre, ognuna di queste sequenze ha la stessa probabilità di verificarsi.

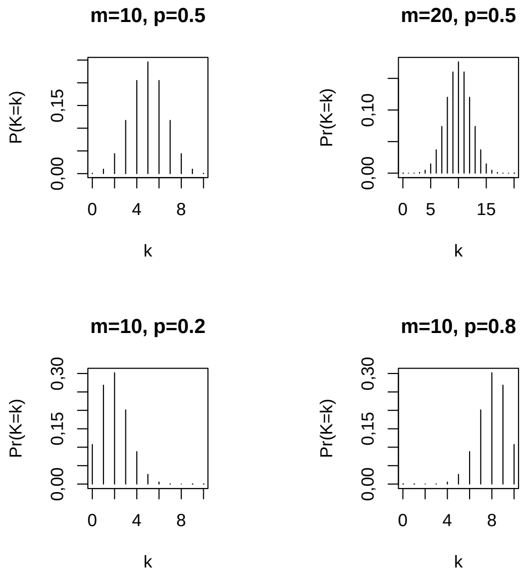
In generale, dobbiamo calcolare, per ogni k tra 0 e m, la probabilità dell'evento $A \subset \mathcal{S}$ costituito da tutte le sequenze con k lettere s e $m - k$

lettere f. Ora, A possiede $\binom{m}{k} = \frac{m!}{k!(m-k)!}$ sequenze e la probabilità di ognuna di esse è data da $p^k(1-p)^{m-k}$. Pertanto la probabilità cercata è

$$\Pr(B = k) = \binom{m}{k} p^k (1-p)^{m-k}.$$

In genere $\Pr(B = k)$ si indica con $Bin(k; m, p)$.

I due parametri m e p sono importanti.



Se B è una variabile casuale che si distribuisce in accordo alla distribuzione binomiale (più brevemente $B \sim \mathcal{B}(m,p)$) allora si può dimostrare che

$$\mathbb{E}(B) = mp, \qquad \mathbb{V}ar(B) = mp(1-p).$$

Nel caso del dado, $B \sim \mathcal{B}(15, 1/6)$ e la probabilità richiesta è

$$Bin(3; 15, 1/6) = \binom{15}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{12} = \frac{15!}{3!12!}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{12} \simeq 0.2363$$

Inoltre, il valore atteso è $\mathbb{E}(B) = 15 \cdot 1/6 = 2{,}5$ e lo scarto quadratico medio è pari a $\sqrt{15 \cdot 1/6 \cdot 5/6} = 1{,}4434$.

Possiamo anche calcolare, ad esempio,

$$\begin{aligned}\Pr(\text{“almeno un 4”}) &= 1 - \Pr(\text{“nessun 4”})\\ \Pr(B \geq 1) &= 1 - \Pr(B < 1) = 1 - \Pr(B = 0)\\ &= 1 - \binom{15}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^{15} = 0{,}9351\,.\end{aligned}$$

oppure

$$\begin{aligned}\Pr(2 \leq B < 5) &= \Pr(B = 2) + \Pr(B = 3) + \Pr(B = 4) \\ &= \text{provate da soli}\ldots\end{aligned}$$

La distribuzione bernoulliana

Nel caso particolare in cui $m = 1$, la v.c. binomiale ($B \sim \mathcal{B}(1, p)$) viene chiamata v.c. di Bernoulli (o bernoulliana): $B \sim Ber(p)$.

La v.c. bernoulliana può assumere solo due valori, 0 e 1, rispettivamente con probabilità p e $1 - p$.

Il valore atteso è pari a

$$\mathbb{E}(B) = 1\ p + 0\ (1 - p) = p$$

e la varianza è pari a

$$\mathbb{V}ar(B) = \mathbb{E}(B^2) - \mathbb{E}(B)^2 = [1^2 p + 0^2(1 - p)] - p^2 = p(1 - p)\,.$$

Dalla definizione di v.c. binomiale è immediato notare che la somma di m v.c. indipendenti $B_i \sim Ber(p)$, $i = 1, \ldots, m$, è una v.c. binomiale: $B = \sum_{i=1}^{m} B_i \sim \mathcal{B}(m, p)$.

La distribuzione di Poisson

Nell'esempio precedente abbiamo visto quale variabile aleatoria può essere utilizzata per descrivere il numero di successi in m prove bernoulliane di probabilità θ.

Cosa succede a questa v.a., qui indicata con P, quando il numero di prove m tende all'infinito e la probabilità θ_m decresce in modo che il prodotto $m\theta_m$ tenda a una costante, cioè $\theta_m = \lambda/m$?

La risposta, per un m fissato, viene ancora fornita dalla distribuzione binomiale:

$$\begin{aligned}\Pr(P = k) &= \binom{m}{k} \left(\frac{\lambda}{m}\right)^k \left(1 - \frac{\lambda}{m}\right)^{m-k} \\ &= \frac{m!}{m^k (m - k)!} \frac{\lambda}{k!}^k \left(1 - \frac{\lambda}{m}\right)^m \left(1 - \frac{\lambda}{m}\right)^{-k}\end{aligned}$$

Facendo tendere m all'infinito otteniamo

$$\lim_{m \to \infty} \frac{m!}{m^k (m - k)!} \frac{\lambda}{k!}^k \left(1 - \frac{\lambda}{m}\right)^m \left(1 - \frac{\lambda}{m}\right)^{-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

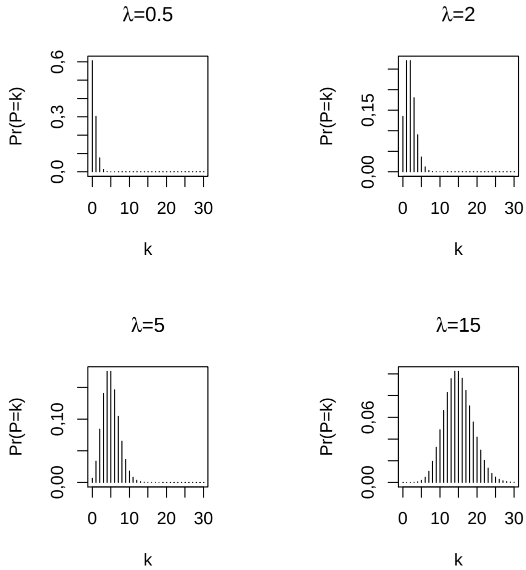
Infatti,

$$\begin{aligned}
\lim_{m\to\infty} \frac{m!}{m^k(m-k)!} &= \lim_{m\to\infty} \frac{m(m-1)\cdots(m-k+1)}{m^k} \\
&= \lim_{m\to\infty} \left[1\left(1-\frac{1}{m}\right)\cdots\left(1-\frac{k-1}{m}\right)\right] = 1\,, \\
\lim_{m\to\infty} \left(1-\frac{\lambda}{m}\right)^m &= e^{-\lambda}\,, \\
\lim_{m\to\infty} \left(1-\frac{\lambda}{m}\right)^{-k} &= 1\,.
\end{aligned}$$

L'espressione trovata è la funzione di probabilità di una variabile aleatoria discreta che può assumere i valori interi $0, 1, 2, \ldots$ (infatti in un numero infinito di prove bernoulliane possiamo avere un numero infinito di successi).

Tale variabile viene chiamata di Poisson di parametro $\lambda > 0$ e viene indicata con $Pois(\lambda)$. Riassumendo la funzione di probabilità è

$$\Pr(P = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k = 0, 1, 2, \ldots$$



Il parametro λ determina i momenti della distribuzione, in particolare

$$\mathbb{E}(P) = \lambda \qquad \mathbb{V}ar(P) = \lambda$$

così, λ è sia la media che la varianza della variabile.

La variabile di Poisson, come vedremo meglio nel corso di Statistica II, può essere utilizzata per descrivere molti fenomeni, ad esempio:

- Il numero di telefonate che arrivano ad una centralina telefonica in un'ora,
- Il numero di messaggi di posta elettronica che arrivano ad un utente in un giorno
- Il numero di decessi di titolari di una polizza di assicurazione sulla vita occorsi in un anno.
- Il numero di transazioni bancarie effettuate da uno sportello al giorno.

Il calcolo della funzione di probabilità può essere fatta in modo ricorsivo, infatti:

$$\frac{\Pr(P = k+1)}{\Pr(P = k)} = \frac{e^{-\lambda}\lambda^{k+1}/(k+1)!}{e^{-\lambda}\lambda^k/k!} = \frac{\lambda}{k+1}.$$

Iniziando da $\Pr(P = 0) = e^{-\lambda}$ allora abbiamo

$$\begin{aligned}
\Pr(P = 1) &= \lambda \Pr(P = 0) \\
\Pr(P = 2) &= \frac{\lambda}{2} \Pr(P = 1) \\
&\vdots \\
\Pr(P = k+1) &= \frac{\lambda}{k+1} \Pr(P = k)
\end{aligned}$$

Supponiamo, per esempio di dover calcolare la funzione di ripartizione in $k = 6$ di una variabile di Poisson di parametro $\lambda = 5$ allora

$$\begin{aligned}
F(6) &= \Pr(P \leq 6) \\
&= \sum_{k=0}^{6} \Pr(P = k) = \Pr(P = 0) + \cdots + \Pr(P = 6) \\
&= e^{-5} + 5e^{-5} + \frac{5}{2}5e^{-5} + \frac{5}{3}\frac{5}{2}5e^{-5} + \cdots + \frac{5}{6}\frac{5}{5}\frac{5}{4}\frac{5}{3}\frac{5}{2}5e^{-5} \\
&= 0.006738 + 0.033690 + \cdots + 0.146223 \\
&= 0.762
\end{aligned}$$

Variabile casuali continue

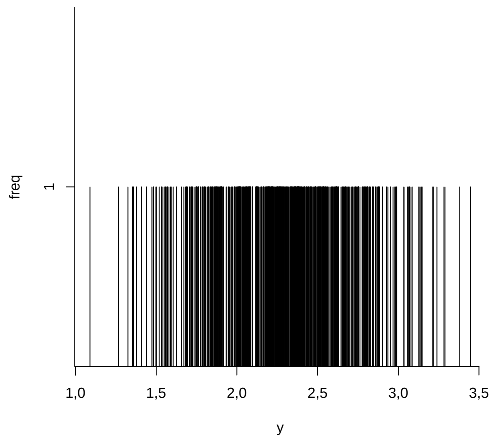
Consideriamo ora il caso in cui lo spazio campionario $\mathcal{S}$ è infinito e $Y(\mathcal{S})$ è l'insieme dei numeri reali o un intervallo di numeri reali, comunque un continuo di numeri (cioè un insieme non numerabile).

Per fissare le idee e per spiegare cosa intendiamo nella pratica con una situazione del genere, pensiamo alla misura del tempo impiegato per portare a termine una transazione finanziaria.

Supponiamo ora di osservare un milione di transazioni (sono tante e possono dare l'idea di un numero infinito, tuttavia questo numero non è una esagerazione in mercati telematici).

Supponiamo (questa sì è una esagerazione) che il nostro strumento di misura del tempo fornisca valori con 20 cifre decimali e che le transazioni si distribuiscano intorno ai 2,3 secondi.

Data la straordinaria precisione dello strumento di misura è quasi certo che non avremo due dati uguali e pertanto se rappresentiamo i nostri dati con un diagramma a bastoncini otteremo



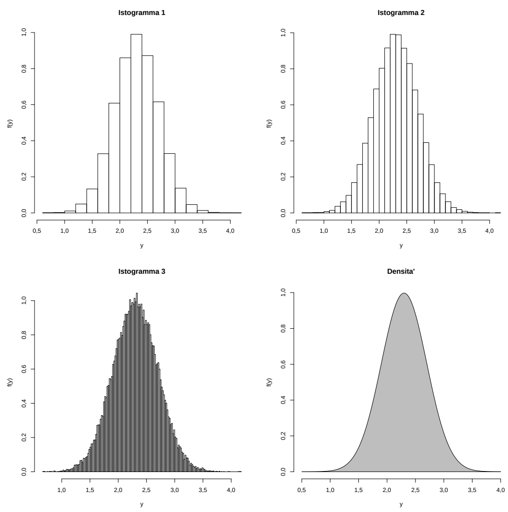
Se prendiamo un qualunque numero compreso tra 0.5 e 3.8 siamo quasi certi che questo numero non corrisponde ad un dato, ma siamo altrettanto certi che il nostro numero è vicino ad un dato. Anzi se il nostro numero è preso nella zona centrale esso trova molti dati vicini, o, il che è sostazialmente lo stesso trova un dato <u>molto</u> vicino.

Abbiamo quindi una variabile casuale $Y(\mathcal{S})$ che ad ogni transazione associa il valore (numerico) del tempo di transazione, ossia un numero nell'intervallo [0.5,3.8] e il ragionamento precedente ci "autorizza" a dire che ogni numero di questo intervallo è un dato, cioè corrisponde ad un campione.

Diciamo allora che siamo in presenza di una **variabile casuale continua**. Nella pratica potremmo avere un numero limitato di dati ma continuiamo a pensare in termini di variabili casuali continue supponendo che i dati in esame siano alcuni degli infiniti dati che avremmo potuto ottenere.

Funzione di densità

Supponiamo di rappresentare i dati mediante istogrammi con un numero sempre maggiore di classi. Non è difficile convincersi che facendo crescere indefinitamente il numero di classi si arriva (al limite) alla curva nel grafico in basso a destra

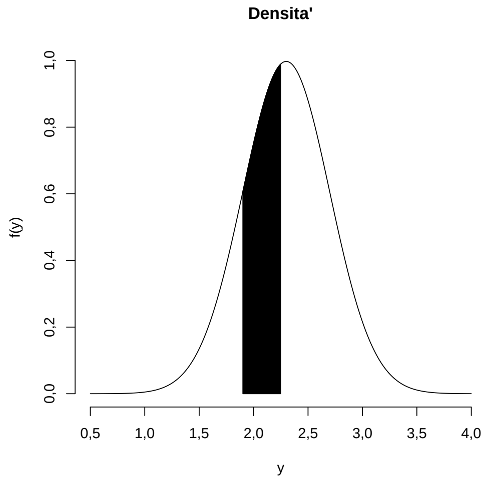


Indichiamo questa funzione con $f(y)$ e la chiamiamo **funzione di densità**.

Come facciamo a rendere l'intervallo $Y(\mathcal{S}) = [0.5, 3.8]$ uno spazio di probabilità?

È chiaro che la probabilità associata a ciascuno degli $y \in [0.5, 3.8]$ è zero, ma pensando alla genesi della curva ovvero che l'area sotto la curva è uguale a 1 e che l'area sotto la curva per $y \in [a, b]$, essendo la somma delle aree dei rettangolini con base contenuta in $[a, b]$ è pari alla frequenza relativa dei campioni con tempi tra a e b, otteniamo che

$$\begin{aligned}\Pr(a < Y \le b) &= \text{area sotto la curva } f(y) \text{ tra } a \text{ e } b \\ &= \int_a^b f(t)dt\end{aligned}$$



Analogamente al caso discreto, si ha:

1. $f(y) \geq 0$, $\forall y \in \mathbb{R}$.
2. $\int_{-\infty}^{+\infty} f(t)dt = 1$.

Funzione di ripartizione di una variabile continua

Possiamo definire quindi la **funzione di ripartizione** per una variabile casuale continua

$$F(y) = \Pr(Y \leq y) = \int_{-\infty}^{y} f(t)dt.$$

e otteniamo

$$\Pr(a < Y \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f(t)dt - \int_{-\infty}^{a} f(t)dt.$$

Se la funzione di ripartizione è una funzione continua allora

$$\begin{aligned} \Pr(Y = y_0) &= \lim_{\varepsilon \to 0} \Pr(y_0 < Y \leq y_0 + \varepsilon) \\ &= \lim_{\varepsilon \to 0} F(y_0 + \varepsilon) - F(y_0) = 0 \end{aligned}$$

e cioè ogni singolo punto ha probabilità nulla. Questo implica che

$$\Pr(a < Y \leq b) = \Pr(a \leq Y < b) = \Pr(a < Y < b) = \Pr(a \leq Y \leq b)\,.$$

Se la funzione di ripartizione è derivabile possiamo calcolare la funzione di densità semplicemente derivando

$$f(y) = \frac{dF(y)}{dy}.$$

e quindi

$$F(y) = \int_{-\infty}^{y} f(t)dt.$$

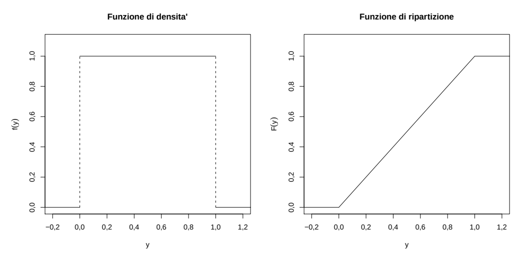
Esempi

Esempio 15 (Uniforme)**.** *Sia*

$$f(y) = \begin{cases} 1, & 0 < y < 1 \\ 0, & altrimenti \end{cases}$$

allora

$$F(y) = \int_{-\infty}^{y} f(t)dt = \begin{cases} 0, & y \leq 0 \\ y, & 0 < y < 1 \\ 1, & y \geq 1 \end{cases}$$

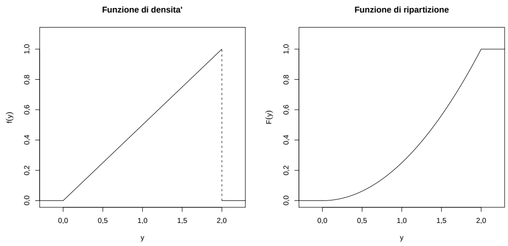


Esempio 16 (Lineare)**.** *Sia*

$$f(y) = \begin{cases} y/2, & 0 < y < 2 \\ 0, & altrimenti \end{cases}$$

allora

$$F(y) = \int_{-\infty}^{y} f(t)dt = \begin{cases} 0, & y \leq 0 \\ y^2/4, & 0 < y < 2 \\ 1, & y \geq 2 \end{cases}$$

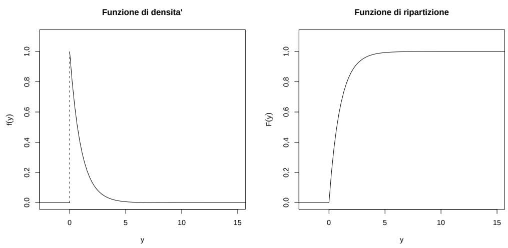


Esempio 17 (Esponenziale)**.** *Sia*

$$f(y) = \begin{cases} e^{-y}, & y > 0 \\ 0, & \textit{altrimenti} \end{cases}$$

allora

$$F(y) = \int_{-\infty}^{y} f(t)dt = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y}, & y > 0 \end{cases}$$



Valore atteso e varianza

Per il calcolo di queste quantità si procede in modo analogo a quanto fatto nel caso discreto (con le sommatorie che vengono “sostituite” da integrali. . .).

Quindi, ammesso che gli integrali esistano (non è sempre vero!), si ha

$$\mu = \mathbb{E}(Y) = \int_{-\infty}^{+\infty} y\, f(y)\, dy$$

$$\begin{aligned} \sigma^2 = \mathbb{V}ar(Y) &= \int_{-\infty}^{+\infty} (y-\mu)^2 f(y)dy \\ &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \\ &= \int_{-\infty}^{+\infty} y^2\, f(y)\, dy - \mu^2\,. \end{aligned}$$

Più in generale per una generica funzione h (sempre che l’integrale esista!), si ha

$$E(h(Y)) = \int_{-\infty}^{+\infty} h(y)f(y)dy\,.$$

Esempio 18. *Sia*

$$f(y) = \begin{cases} e^{-y}, & y > 0 \\ 0, & \textit{altrimenti} \end{cases}$$

allora, integrando per parti, si ha

$$\begin{aligned} \mu &= \int_{-\infty}^{+\infty} y f(y) dy = \int_0^{+\infty} y\, e^{-y} dy \\ &= \left[-y\, e^{-y}\right]_0^{+\infty} + \int_0^{\infty} e^{-y} dy = 1 \end{aligned}$$

e

$$\begin{aligned} \int_{-\infty}^{+\infty} y^2 f(y) dy &= \int_0^{+\infty} y^2\, e^{-y} dy \\ &= \left[-y^2\, e^{-y}\right]_0^{+\infty} + 2 \int_0^{\infty} y\, e^{-y} dy = 2 \end{aligned}$$

Quindi, $\mu = 1$ *e* $\sigma^2 = 2 - (1)^2 = 1$.

Esercizio 13. *Provate a calcolare valore atteso e varianza anche per le v.c. degli esempi [15] e [16] visti in precendenza.*

La variabile casuale uniforme

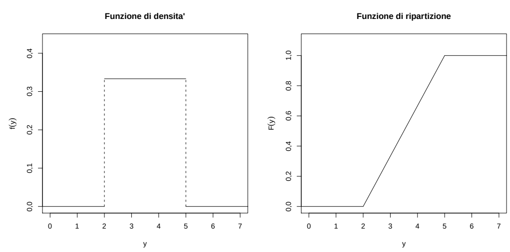
Sia Y una variabile casuale la cui funzione di densità è

$$f(y) = \begin{cases} \frac{1}{b-a} & \text{se } a \leq y \leq b \\ 0 & \text{altrimenti} \end{cases}$$

Diciamo allora che Y si distribuisce uniformemente in $[a, b]$.

È facile vedere (provate!) che

$$\mu = \frac{a+b}{2}, \qquad \sigma^2 = \frac{(b-a)^2}{12}$$



L'esempio [15] visto in precendenza era una v.c. uniforme con $a = 0$ e $b = 1$.

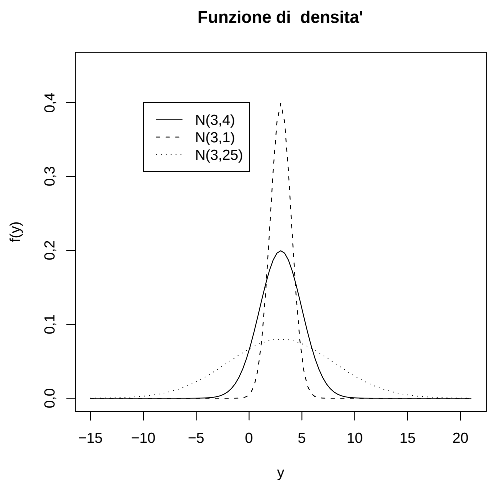
La variabile casuale normale o gaussiana

La più importante distribuzione continua è la distribuzione normale o gaussiana, che ha le seguente densità

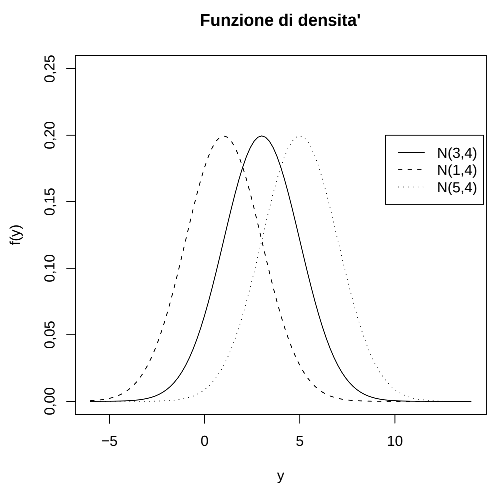
$$f(y; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

Il parametro μ rappresenta il valore atteso della variabile mentre il parametro σ^2 rappresenta la varianza della variabile. Noteremo con $Y \sim \mathcal{N}(\mu, \sigma^2)$ una variabile casuale normale di valore atteso μ e varianza σ^2.

Vediamo ora come cambia la forma densità al variare del parametro varianza (σ^2)



e al variare del parametro media (μ)



Osservazioni:

- la funzione di densità è simmetrica rispetto alla media;
- media ($\equiv$ valore atteso), moda e mediana coincidono;
- Una particolare distribuzione è la distribuzione normale **standard** ($Z \sim \mathcal{N}(0,1)$), ossia con media pari a 0 e varianza pari a 1 e densità

$$f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2}.$$

Calcolo della probabilità per una variabile casuale normale

In generale vorremmo calcolare degli integrali del tipo

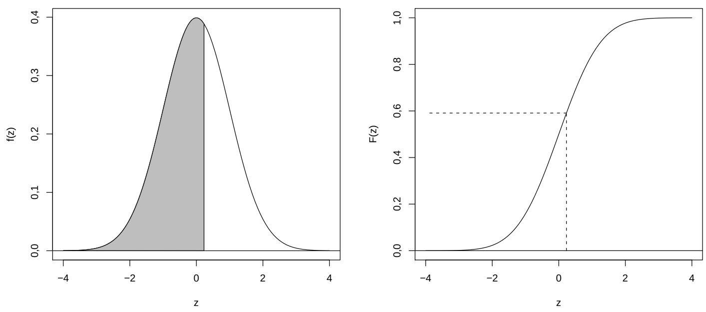
$$F(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

ma non è possibile dare una rappresentazione analitica. Per questo, facciamo ricorso a delle tavole del tipo

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
.2	0.5793	0.5832	0.5871	**0.5910**	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997

La tavola raccoglie le probabilità degli intervalli $-\infty < Z \leq z$, cioè l'area sotto la curva data dalla densità di una variabile casuale normale standard. Ad esempio

$$\Pr(-\infty < Z \leq 0.23) = \mathbf{0.5910}$$



Notiamo che i valori z_p relativi alle probabilità p della tavola sono i **quantili** di ordine p della variabile casuale Z:

$$\Pr(Z \leq z_p) = p\,.$$

Infatti z_p è il valore di Z che lascia esattamente probabilità p alla sua sinistra (e quindi $1 - p$ alla sua destra).

Esempi di utilizzo della tavole

1. $\Pr(Z > 0.23) = 1 - \Pr(Z \leq 0.23) = 1 - 0.5910 = 0.4090$

2. $\Pr(Z \leq -0.23) = \Pr(Z \geq 0.23) = 0.4090$ (abbiamo utilizzato la simmetria)

3. $\Pr(-0.15 < Z \leq 0.23) = \Pr(Z \leq 0.23) - \Pr(Z \leq -0.15) = 0.5910 - (1 - \Pr(Z \leq 0.15)) = 0.5910 - (1 - 0.5596) = 0.1506$

4. $\Pr(Z \leq -0.15 \cup Z > 0.23) = \Pr(Z \leq -0.15) + \Pr(Z > 0.23) = (1 - 0.5596) + (1 - 0.5910) = 0.8494$

5. Qual è il valore di $z_{0.975}$? Nelle tavole, il valore della probabilità p è crescente a partire dal primo elemento della prima riga e procedendo per riga. Quindi basta arrivare alla posizione in cui troviamo $p = 0.975$ e leggere che $z_{0.975} = 1.96$ (1.9 dalla riga + 0.06 dalla colonna).

6. Se vogliamo trovare $z_{0.005}$ dobbiamo tener conto che $z_p = -z_{1-p}$ (per la simmetria), poichè le tavole riportano solo valori di $p \geq 0.50$. Quindi $z_{0.005} = -z_{0.995} = -2.57$.

Standardizzazione

Da una qualunque variabile casuale Y normale di media μ e varianza σ^2 possiamo ottenere la variabile normale standard, attraverso la seguente semplice trasformazione lineare

$$Z = \frac{Y - \mu}{\sigma}.$$

In base a questo risultato ci è sufficiente conoscere la funzione di ripartizione della variabile Z per ottenere ogni probabilità desiderata. Infatti se indichiamo con

$$z_0 = \frac{y_0 - \mu}{\sigma}$$

$$\begin{aligned} \Pr(Y \leq y_0) &= \Pr\left(\frac{Y - \mu}{\sigma} \leq \frac{y_0 - \mu}{\sigma}\right) \\ &= \Pr\left(Z \leq \frac{y_0 - \mu}{\sigma}\right) = \Pr(Z \leq z_0) \end{aligned}$$

Ad esempio: $y_0 = 5$, $\mu = 3$, $\sigma^2 = 4 \Rightarrow z_0 = (5 - 3)/2 = 1$

$$\Pr(Y \leq 5) = \Pr(Z \leq 1) = 0.8413$$

2.3 Normal Probability Plot

QQ–plot

Definizione 13 (Statistica ordinata)**.** *Siano* $y_1, \dots, y_n$ *n osservazioni su di una variabile numerica. Una permutazione* $y_{(1)}, \dots, y_{(n)}$ *di* $y_1, \dots, y_n$ *tale che*

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

è detta statistica ordinata.

In parole semplici: la statistica ordinata è l'insieme dei valori osservati ordinati dal più piccolo al più grande. Quindi, ad esempio, $y_{(1)}$ è l'osservazione più piccola.

$y_{(j)}$ può essere visto come una stima del quantile-(j/n) della distribuzione che ha generato i dati almeno nei casi in cui la distribuzione dei dati non sia "troppo discreta" (ovvero, in cui i valori distinti tra le osservazioni non siano molto pochi).

Infatti,

- se solo una osservazione è uguale a $y_{(j)}$ allora la percentuale di osservazioni minore o uguale di $y_{(j)}$ è esattamente $100(j/n)$;
- se alcune osservazioni sono uguali a $y_{(j)}$ ma il loro numero non è grande rispetto a n allora la percentuale di osservazioni minore od uguale a $y_{(j)}$ non sarà esattamente uguale a $100(j/n)$ ma non potrà essere molto differente da questa percentuale.

È facile verificare che se $Y \sim N(\mu, \sigma^2)$ allora

$$(\text{quantile-}p \text{ di } Y) = \mu + \sigma z_p$$

dove z_p indica il quantile-p di una normale standard.

Si indichi infatti con $z_p(\mu, \sigma)$ il quantile-p di Y. Per definizione,

$$\Pr(Y \leq z_p(\mu, \sigma)) = p.$$

Quindi,

$$\Pr\left(\frac{Y-\mu}{\sigma} \leq \frac{z_p(\mu, \sigma) - \mu}{\sigma}\right) = p.$$

Ricordando che $(Y - \mu)/\sigma \sim N(0, 1)$, dalla ultima relazione segue che

$$\Pr\left(N(0,1) \leq \frac{z_p(\mu, \sigma) - \mu}{\sigma}\right) = p$$

ovvero, che

$$\frac{z_p(\mu,\sigma)-\mu}{\sigma} = z_p(0,1) \Rightarrow z_p(\mu,\sigma) = \mu + \sigma z_p(0,1).$$

Sintesi osservazioni precedenti:

- $y_{(j)}$ lascia a sinistra circa j osservazioni su n; quindi dovrebbe “cadere intorno” al quantile j/n della distribuzione da cui sono stati generati i dati;

- se questa distribuzione è normale questo quantile è $\mu + \sigma z_{j/n}$ dove μ e σ sono incogniti ma $z_{j/n}$ è una quantità nota.

Le due osservazioni insieme suggeriscono che se i dati sono normali disegnando su di un piano cartesiano i punti

$$\left(z_{\frac{j-0,5}{n}}, y_{(j)}\right)$$

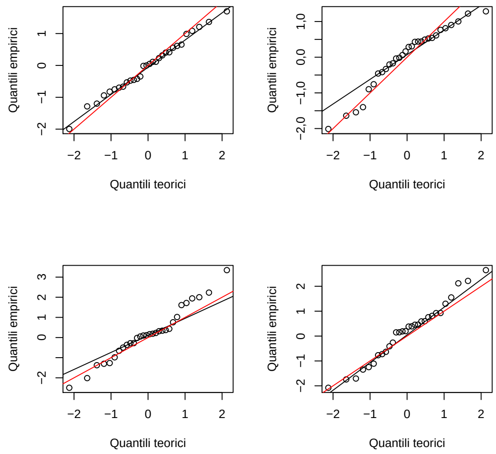
dovremmo osservare un andamento lineare. Viceversa se l’andamento non è lineare questo indica che i quantili della distribuzione dei dati non “si comportano” come quelli di una distribuzione normale ovvero che la distribuzione dei dati non è normale.

Il grafico ottenuto (o sue varianti in cui $z_{(j-0,5)/n}$ è sostituito da analoghe quantità “vicine” a $z_{j/n}$) è chiamato *normal probability plot.* Si osservi che si tratta di un grafico in cui disegnamo nella sostanza i quantili campionari verso i quantili di una distribuzione teorica.

Domanda: Perchè non usiamo $z_{j/n}$? *Risposta*: Perchè $z_1 = z_{n/n} = +\infty$ e quindi dovremmo disegnare l’osservazione più grande ad infinito.

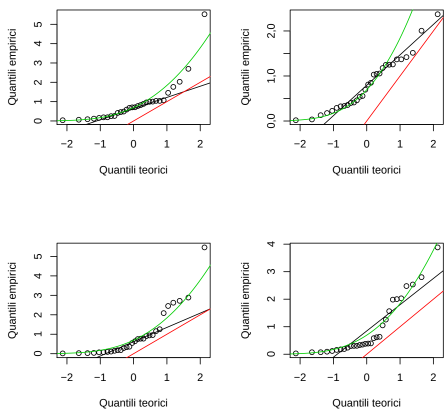
Spesso, per riferimento, si riporta la bisettrice del 1° e 3° quadrante o la retta passante per le coppie generate dal primo e terzo quartile empirico e teorico.

Esempio 19 (Campioni generati da una distribuzione normale)**.**



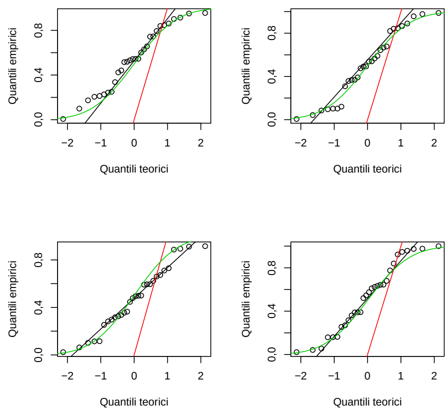
I grafici sono basati su 4 *campioni di numerosità pari a* 30 *simulati da una distribuzione normale standard. In questo caso i punti dovrebbero, stare intorno alla bisettrice del* 1° *e* 3° *quadrante (in rosso nel grafico).*

Esempio 20 (Campioni generati da una distribuzione esponenziale)**.**



I grafici sono basati su 4 *campioni di numerosità pari a* 30 *simulati da una distribuzione esponenziale di media* 1*. Si osservi che il quantile-p di questa distribuzione vale* $-\log(1-p)$*. Quindi, i punti li aspettiamo in questo caso intorno alla curva* $(z_p, -log(1-p))$, $0 < p < 1$ *(in verde). In rosso la bisettrice del primo quadrante.*

Esempio 21 (Campioni generati da una distribuzione uniforme)**.**



I grafici sono basati su 4 *campioni di numerosità pari a* 30 *simulati da una distribuzione con densità uniforme tra* 0 *e* 1*. Si osservi che il quantile-p di questa distribuzione vale p. Quindi, i punti li aspettiamo in questa caso intorno alla curva* (z_p, p), $0 < p < 1$ *(in verde). In rosso la bisettrice del primo quadrante.*

Si osservi come la relativamente lunga parte lineare centrale possa rendere difficile disciminare tra una distribuzione normale e una distribuzione uniforme. È ad esempio quello che accade nel grafico in basso a destra nella figura precedente.

2.4 Variabili casuali bivariate

Variabili casuali bivariate discrete

Consideriamo due variabili casuali X e Y con valori $\{x_1, \ldots, x_h\}$, $\{y_1, \ldots, y_k\}$[2]. La funzione di probabilità congiunta delle variabili X e Y, $p_{X,Y}(\cdot, \cdot)$, indica la probabilità che la variabile X assuma il valore x_i e **congiuntamente** la variabile Y assuma il valore y_j e può essere definita come segue

$$\begin{aligned} p_{X,Y}(x_i, y_j) &= \Pr(\{X = x_i\} \cap \{Y = y_j\}) \\ &= \Pr(X = x_i, Y = y_j) \\ & \quad i = 1, \ldots, h, \quad j = 1, \ldots, k \end{aligned}$$

La funzione di probabilità bidimensionale può essere rappresentata molto chiaramente mediante una tabella a doppia entrata

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_k
x_1	$p_{X,Y}(x_1, y_1)$	$\cdots$	$p_{X,Y}(x_i, y_j)$	$\cdots$	$p_{X,Y}(x_1, y_k)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_i	$p_{X,Y}(x_i, y_1)$	$\cdots$	$p_{X,Y}(x_i, y_j)$	$\cdots$	$p_{X,Y}(x_1, y_k)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_h	$p_{X,Y}(x_h, y_1)$	$\cdots$	$p_{X,Y}(x_h, y_j)$	$\cdots$	$p_{X,Y}(x_h, y_k)$

La **funzione di ripartizione** di due variabili casuali discrete è data da:

$$F_{X,Y}(x, y) = \Pr(X \leq x, Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{X,Y}(x_i, y_j).$$

Possiamo ricavare le funzioni di probabilità delle singole variabili (funzioni di probabilità **marginali**)

$$p_X(x_i) = \Pr(X = x_i) = \sum_{j=1}^{k} p_{X,Y}(x_i, y_j), \quad i = 1, \ldots, h$$

$$p_Y(y_j) = \Pr(Y = y_j) = \sum_{i=1}^{h} p_{X,Y}(x_i, y_j),, \quad j = 1, \ldots, k\,.$$

e quindi possiamo completare la tabella precedente con

[2]Per semplicità consideriamo v.c. con insiemi di valori finito. In generale i risultati restano validi anche per v.c. che assumono insiemi infinito numerabili di valori.

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_k	$p_X(x)$
x_1	$p_{X,Y}(x_1,y_1)$	$\cdots$	$p_{X,Y}(x_i,y_j)$	$\cdots$	$p_{X,Y}(x_1,y_k)$	$p_X(x_1)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	$p_{X,Y}(x_i,y_1)$	$\cdots$	$p_{X,Y}(x_i,y_j)$	$\cdots$	$p_{X,Y}(x_i,y_k)$	$p_X(x_i)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_h	$p_{X,Y}(x_h,y_1)$	$\cdots$	$p_{X,Y}(x_h,y_j)$	$\cdots$	$p_{X,Y}(x_h,y_k)$	$p_X(x_h)$
$p_Y(y)$	$p_Y(y_1)$	$\cdots$	$p_Y(y_j)$	$\cdots$	$p_Y(y_k)$	1

Notiamo che, in generale

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) = \Pr(X \leq x, Y \leq +\infty) = F_{X,Y}(x,+\infty)\,, \\ F_Y(y) &= \Pr(Y \leq y) = \Pr(X \leq +\infty, Y \leq y) = F_{X,Y}(+\infty,y)\,. \end{aligned}$$

Esempio 22. *Si consideri il lancio di un dado e sia*

$$X = \begin{cases} 0, & \text{“\textit{esce un numero dispari}”} \\ 1, & \text{“\textit{esce un numero pari minore di 4}”} \\ 2, & \text{“\textit{esce un numero pari maggiore o uguale di 4}”} \end{cases}$$

e

$$Y = \begin{cases} 1, & \text{“\textit{esce un numero multiplo di 3}”} \\ 0, & \text{“\textit{non esce un numero multiplo di 3}”} \end{cases}$$

Calcoliamo la funzione di probabilità e di ripartizione della v.c. doppia $Z = (X, Y)$.

I possibili risultati dell'esperimento, le relative probabilità e i valori delle due variabili casuali sono riassunti nella seguente tabella:

evento	*1*	*2*	*3*	*4*	*5*	*6*
prob.	1/6	1/6	1/6	1/6	1/6	1/6
X	0	1	0	2	0	2
Y	0	0	1	0	0	1

È facile determinare la funzione di probabilità e di ripartizione di X ed Y considerate singolarmente; infatti abbiamo

x	0	1	2
$p_X(x)$	3/6	1/6	2/6

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 3/6, & 0 \leq x < 1 \\ 4/6, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

y	0	1
$p_Y(y)$	4/6	2/6

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ 4/6, & 0 \leq y < 1 \\ 1, & y \geq 1 \end{cases}$$

Congiuntamente invece abbiamo che la funzione di probabilità $p_{X,Y}(x, y)$ è

	Y		
X	*0*	*1*	$p_X(x)$
0	2/6	1/6	3/6
1	1/6	0	1/6
2	1/6	1/6	2/6
$p_Y(y)$	4/6	2/6	1

mentre la funzione di ripartizione congiunta $F_{X,Y}(x, y)$ è

	Y		
X	*0*	*1*	$F_X(x)$
0	2/6	3/6	3/6
1	3/6	4/6	4/6
2	4/6	1	1
$F_Y(y)$	4/6	1	

Indipendenza stocastica di due variabili casuali

L'indipendenza di due variabili casuali X e Y viene definita grazie alla probabilità per eventi indipendenti. Dati due eventi indipendenti A e B, la probabilità che questi due eventi si verifichino congiuntamente è data dal prodotto delle loro probabilità:

$$\Pr(A \cap B) = \Pr(A)\Pr(B)$$

Considerati i due eventi $A = \{X = x_i\}$ e $B = \{Y = y_j\}$ possiamo definire l'indipendenza di due variabili casuali discrete X e Y se **per ogni coppia di possibili risultati** (x_i, y_j) delle variabili casuali X e Y vale la relazione

$$\Pr(X = x_i, Y = y_j) = \Pr(X = x_i)\Pr(Y = y_j)$$

o equivalentemente

$$p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_j),$$

dove $p_X(x_i)$ e $p_Y(y_j)$ sono le relative distribuzioni di probabilità marginali.

Nell'esempio precedente si ha, ad esempio,

$$1/6 = p_{X,Y}(1,0) \neq p_X(1)p_Y(0) = 1/6 \cdot 4/6 = 1/9\,,$$

il che prova che X ed Y sono stocasticamente dipendenti.

Distribuzioni condizionate

Sia (X,Y) una v.a. doppia discreta con funzione di probabilità

$$p_{X,Y}(x,y) = \Pr(\{X = x\} \cap \{Y = y\})$$

allora in accordo con la definizione di probabilità condizionata

$$\begin{aligned} p_{Y|X}(y|x) &= \Pr(\{Y = y\}|\{X = x\}) \\ &= \frac{\Pr(\{X = x\} \cap \{Y = y\}}{\Pr(\{X = x\})} \\ &= \frac{p_{X,Y}(x,y)}{p_X(x)} \end{aligned}$$

con $p_X(x) > 0$.

Per ogni valore fissato di x la funzione $p_{Y|X}(y|x)$ prende il nome di **probabilità condizionata** di Y dato che X ha assunto il valore x.

Se X e Y sono indipendenti allora

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(x)} = p_Y(y)$$

se $p_X(x) > 0$, analogamente

$$p_{X|Y}(x|y) = p_X(x)$$

se $p_Y(y) > 0$.

Esempio 23 (Distribuzioni condizionate)**.** *Calcoliamo le funzioni di probabilità condizionate $p_{Y|X}(y|x)$ (una per ogni valore x).*

	Y		
X	*0*	*1*	$p_X(x)$
0	2/6	1/6	3/6
1	1/6	0	1/6
2	1/6	1/6	2/6
$p_Y(y)$	4/6	2/6	1

	Y	
X	$p_{Y\|X}(0\|x)$	$p_{Y\|X}(1\|x)$
0	2/3	1/3
1	1	0
2	1/2	1/2

dove, ad esempio,

$$p_{Y|X}(1|0) = \frac{p_{X,Y}(0,1)}{p_X(0)} = \frac{1/6}{3/6} = 1/3\,.$$

Calcoliamo anche le probabilità condizionate $p_{X|Y}(x|y)$

	Y	
X	*0*	*1*
$p_{X\|Y}(0\|y)$	1/2	1/2
$p_{X\|Y}(1\|y)$	1/4	0
$p_{X\|Y}(2\|y)$	1/4	1/2

Valori attesi condizionati

La funzione $Y|x$ (Per brevità indichiamo $Y|x$ per $Y|X = x$) è una v.c. con funzione di probabilità $p_{Y|X}(y|x)$.

Possiamo costruire la **funzione di ripartizione condizionata** applicando la definizione nel modo usuale

$$F_{Y|X}(y|x) = \sum_{y_i \le y} p_{Y|X}(y_i|x)\,.$$

Inoltre, ha senso definire il **valore atteso condizionato**

$$\mathbb{E}(Y|x) = \sum_{j=1}^{k} y_j p_{Y|X}(y_j|x)$$

e la **varianza condizionata**

$$\begin{aligned}
\mathbb{V}ar(Y|x) &= \sum_{j=1}^{k} [y_j - \mathbb{E}(Y|x)]^2 p_{Y|X}(y_j|x) \\
&= \mathbb{E}(Y^2|x) - \mathbb{E}(Y|x)^2\,.
\end{aligned}$$

Si noti che le due quantità ottenute sono, in generale, una funzione della v.c. condizionante X. E quindi, al variare dei valori di X, $\mathbb{E}(Y|X)$ e $\mathbb{V}ar(Y|X)$ **sono variabili casuali**.

Esempio 24 (Valori attesi condizionati)**.** *Continuando l'esempio si calcoli ad esempio* $\mathbb{E}(Y|x)$ *e* $\mathbb{V}ar(Y|x)$

$$\begin{aligned}
\mathbb{E}(Y|0) &= 0 \cdot 2/3 + 1 \cdot 1/3 = 1/3 \\
\mathbb{E}(Y|1) &= 0 \cdot 1 + 1 \cdot 0 = 0 \\
\mathbb{E}(Y|2) &= 0 \cdot 1/2 + 1 \cdot 1/2 = 1/2 \\
\mathbb{V}ar(Y|0) &= (0 - 1/3)^2 \cdot 2/3 + (1 - 1/3)^2 \cdot 1/3 = 2/9 \\
\mathbb{V}ar(Y|1) &= 0 \\
\mathbb{V}ar(Y|0) &= (0 - 1/2)^2 \cdot 1/2 + (1 - 1/2)^2 \cdot 1/2 = 1/4\,.
\end{aligned}$$

Le funzioni di ripartizione invece sono

$$F_{Y|X}(y|0) = \begin{cases} 0, & y < 0 \\ 2/3, & 0 \leq y < 1 \\ 1, & y \geq 1 \end{cases}$$

$$F_{Y|X}(y|1) = \begin{cases} 0, & y < 0 \\ 1, & y \geq 0 \end{cases}$$

$$F_{Y|X}(y|2) = \begin{cases} 0, & y < 0 \\ 1/2, & 0 \leq y < 1 \\ 1, & y \geq 1 \end{cases}$$

Variabili casuali bivariate continue

Diciamo che due v.c X e Y sono congiuntamente continue se esiste una funzione $f_{X,Y}(x,y)$ a valori non negativi, integrabile [3], tale che per ogni sottoinsieme C di $\mathbb{R}^2$ si abbia

$$\Pr((X,Y) \in C) = \iint_{(x,y)\in C} f_{X,Y}(x,y)dxdy$$

La funzione $f_{X,Y}(x,y)$ è detta **(funzione di) densità congiunta** di X e Y.

In particolare se $C = \{(t,u), -\infty < t < x, -\infty < u < y\}$

$$\begin{aligned} \Pr((X,Y) \in C) &= \Pr(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t,u)dtdu \\ &= F_{X,Y}(x,y) \end{aligned}$$

La $F_{X,Y}(x,y)$ è detta **funzione di ripartizione congiunta** di X e Y.

Se le derivate parziali di F hanno senso allora

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$$

Un'altra intepretazione della densità congiunta è la seguente, se Δ_x e Δ_y sono "piccoli"

$$\begin{aligned} \Pr(x < X \leq x + \Delta_x \quad &, \quad y < Y \leq y + \Delta_y) \\ &= \int_{x}^{x+\Delta_x} \int_{y}^{y+\Delta_y} f_{X,Y}(t,u)dtdu. \\ &\approx f_{X,Y}(x,y)\Delta_x\Delta_y. \end{aligned}$$

[3]Una funzione h definita su $\mathbb{R}^2$ si dice integrabile se $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |h(x,y)|dxdy < \infty$

Perciò $f_{X,Y}(x, y)$ può intendersi come la probabilità che la coppia (X, Y) sia vicina a (x, y).

Le **(funzioni di) densità marginali** sono

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx.$$

Due variabili casuali continue X e Y sono **indipendenti** se per ogni coppia di numeri reali (x, y) vale la relazione:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

La **densità condizionata** di Y dato $X = x$ è definita come

$$f_{Y|X}(y|x) \;=\; \frac{f_{X,Y}(x, y)}{f_X(x)} \qquad \text{per ogni } x : f_X(x) > 0$$

Per motivare questa definizione si consideri

$$\begin{aligned} f_{Y|X}(y|x)\Delta_y &= \frac{f_{X,Y}(x, y)\Delta_x\Delta_y}{f_X(x)\Delta_x} \\ &\approx \frac{\Pr(x < X \leq x + \Delta_x, y < Y \leq y + \Delta_y)}{\Pr(x < X \leq x + \Delta_x)} \\ &= \Pr(y < Y \leq y + \Delta_y | x < X \leq x + \Delta_x) \end{aligned}$$

Perciò $f_{Y|X}(y|x)$ può intendersi come la probabilità condizionata che Y appartenga all'intervallo infinitesimale $(y, y + \Delta_y]$ dato che X appartiene all'intervallo infinitesimale $(x, x + \Delta_x]$.

Possiamo definire la **funzione di ripartizione condizionata**

$$F_{Y|X}(y|x) = \Pr(Y \leq y | X = x) = \int_{-\infty}^{y} f_{Y|X}(t|x)dt$$

Nel caso di indipendenza

$$f_{Y|X}(y|x) = \frac{f_X(x)f_Y(y)}{f_x(x)} = f_Y(y)$$

Infine diamo la definizione di valore atteso condizionato e varianza condizionata per $f(y) > 0$

$$\begin{aligned} \mathbb{E}(Y|x) &= \int_{-\infty}^{\infty} t f_{Y|X}(t|x)dt \\ \mathbb{V}ar(Y|x) &= \int_{-\infty}^{\infty} [t - \mathbb{E}(Y|x)]^2 f_{Y|X}(t|x)dt \end{aligned}$$

Esempio 25 (Variabili casuali continue)**.** *Sia*

$$f_{X,Y}(x,y)=\begin{cases} x+y, & 0<x<1, 0<y<1\\ 0 & \textit{altrimenti}\end{cases}$$

Calcoliamo $f_X(x)$

$$\begin{aligned} f_X(x) &= \int_0^1 (x+y)dy = x\int_0^1 1dy + \int_0^1 ydy = \\ &= x\,[y]_0^1 + \left[y^2/2\right]_0^1 = x+1/2 \end{aligned}$$

Calcoliamo $\mathbb{E}(X)$

$$\begin{aligned} \mathbb{E}(X) &= \int_0^1 x(x+1/2)dx \\ &= \left[x^3/3\right]_0^1 + \left[x^2/4\right]_0^1 = 1/3+1/4 = 7/12 \end{aligned}$$

Si noti che $f_Y(y) = y+1/2$ *e che* $\mathbb{E}(Y) = 7/12$*. Da questo possiamo dedurre che*

$$f_{X,Y}(x,y) \neq (x+1/2)(y+1/2)$$

e quindi le v.c. non sono indipendenti. Ad esempio

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{x+y}{x+1/2}, \qquad \textit{per } 0<y<1$$

e

$$\begin{aligned} \mathbb{E}(Y|x) &= \int_0^1 y\frac{x+y}{(x+1/2)}dy = \int_0^1 \frac{xy}{(x+1/2)}dy + \int_0^1 \frac{y^2}{(x+1/2)}dy \\ &= \frac{x}{(x+1/2)}\left[y^2/2\right]_0^1 + \frac{1}{(x+1/2)}\left[y^3/3\right]_0^1 \\ &= \frac{1}{(x+1/2)}(x/2+1/3) \end{aligned}$$

Momenti di funzioni di variabili doppie

Proposizione 4. *Siano* X *e* Y *due v.c. e* g *una funzione di due variabili, allora valgono i seguenti risultati*

1. *se* X *e* Y *sono discrete*

$$\mathbb{E}(g(X,Y)) = \sum_{i=1}^{h}\sum_{j=1}^{k} g(x_i,y_j)p_{X,Y}(x_i,y_j)$$

2. se X e Y sono congiuntamente continue

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t,u) f_{X,Y}(t,u)dtdu$$

Ecco alcune interessanti risultati che si possono derivare dalla proposizione precedente. Le dimostrazioni sono mostrate supponendo X e Y discrete, ma rimangono invariate se a "$\sum$ si sostituisce opportunamente $\int$".

Proposizione 5.

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

Dimostrazione 2. *Infatti $g(X,Y) = aX + bY$ e*

$$\begin{aligned}
\mathbb{E}(aX + bY) &= \sum_{i=1}^{h}\sum_{j=1}^{k}(ax_i + by_j)p_{X,Y}(x_i, y_j) \\
&= a\sum_{i=1}^{h}\sum_{j=1}^{k} x_i p_{X,Y}(x_i, y_j) + b\sum_{i=1}^{h}\sum_{j=1}^{k} y_j p_{X,Y}(x_i, y_j) \\
&= a\sum_{i=1}^{h} x_i \sum_{j=1}^{k} p_{X,Y}(x_i, y_j) + b\sum_{j=1}^{k} y_j \sum_{i=1}^{h} p_{X,Y}(x_i, y_j) \\
&= a\sum_{i=1}^{h} x_i p_X(x_i) + b\sum_{j=1}^{k} y_j p_Y(y_j) = a\mathbb{E}(X) + b\mathbb{E}(Y)
\end{aligned}$$

Proposizione 6. *Se $X \leq Y$ allora*

$$\mathbb{E}(X) \leq \mathbb{E}(Y)$$

Dimostrazione 3. *$g(X,Y) = X - Y$ e*

$$\mathbb{E}(X - Y) = \sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - y_j)p_{X,Y}(x_i, y_j) \leq 0$$

poiché $x_i - y_j \leq 0$ per ogni i, j, e quindi

$$\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y) \leq 0$$

Proposizione 7. *Se X e Y sono **indipendenti** e u e v sono due funzioni a valori reali*

$$\mathbb{E}(u(X) \cdot v(Y)) = \mathbb{E}(u(X)) \cdot \mathbb{E}(v(Y))$$

Dimostrazione 4. *$g(X,Y) = u(X) \cdot v(Y)$ e*

$$\begin{aligned}
\mathbb{E}(u(X) \cdot v(Y)) &= \sum_{i=1}^{h}\sum_{j=1}^{k} u(x_i)v(y_j)p_{X,Y}(x_i, y_j) \\
&\quad \textit{grazie all'indipendenza} \\
&= \sum_{i=1}^{h}\sum_{j=1}^{k} u(x_i)v(y_j)p_X(x_i)p_Y p(y_j) \\
&= \left[\sum_{i=1}^{h} u(x_i)p_X(x_i)\right]\left[\sum_{j=1}^{k} v(y_j)p_Y p(y_j)\right] \\
&= \mathbb{E}(u(X)) \cdot \mathbb{E}(v(Y))
\end{aligned}$$

Facilmente si ottiene la seguente

Proposizione 8. *Se X e Y sono **indipendenti***

$$\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

Covarianza e correlazione

Un'altra applicazione del risultato precedente è costituita dal calcolo della covarianza

$$\mathbb{C}ov(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

e si può dimostrare che

$$\mathbb{C}ov(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

Si noti

Proposizione 9. *Se X e Y sono indipendenti, $\mathbb{C}ov(X, Y) = 0$*

Si noti che non è vero il viceversa. Infatti sia X una v.c. che assume i valori $\{-1, 0, 1\}$, ciascuno con probabilità $1/3$. Facilmente si verifica che $\mathbb{E}(X) = 0$. Sia Y la v.c. $Y = X^2$. Per definizione quindi Y dipende da X. Calcoliamo

$$\mathbb{E}(XY) = (-1 \cdot 1)1/3 + (0 \cdot 0)1/3 + (1 \cdot 1)1/3 = 0$$

per cui concludiamo che $\mathbb{C}ov(X, Y) = 0$

In forza della proposizione [9] la $\mathbb{C}ov(X, Y)$ assume il ruolo di un indice di dipendenza. La dipendenza che viene misurata è solo quella di tipo lineare (si veda più avanti).

Come indice di dipendenza lineare la $\mathbb{C}ov(X,Y)$ mal si presta ai confronti in quanto questa può assumere dei valori nell'intervallo $(-\infty,\infty)$. In realtà sappiamo dire qualcosa di più circa il suo intervallo di variazione, con **disuguaglianza di Cauchy-Schwarz**

$$[\mathbb{E}(XY)]^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

Grazie a questa, si ha

$$[\mathbb{C}ov(X,Y)]^2 \leq \mathbb{V}ar(X)\mathbb{V}ar(Y)$$

Ciò ha portato alla definizione di un indice detto **indice di correlazione**

$$\rho(X,Y) = \frac{\mathbb{C}ov(X,Y)}{\sqrt{\mathbb{V}ar(X)\cdot\mathbb{V}ar(Y)}}$$

che assume valori nell'intervallo $[-1,1]$.

Supponiamo che $Y = a + bX$, ovvero che tra X e Y ci sia una relazione lineare, allora

$$\rho(X,Y) = \frac{\mathbb{C}ov(X,a+bX)}{\sqrt{\mathbb{V}ar(X)\cdot\mathbb{V}ar(a+bX)}}$$

ma

$$\begin{aligned}
\mathbb{C}ov(X,a+bX) &= \mathbb{E}(X\cdot(a+bX)) - \mathbb{E}(X)\cdot\mathbb{E}(a+bX)) \\
&= b[\mathbb{E}(X^2) - (\mathbb{E}(X))^2] \\
&= b\mathbb{V}ar(X)
\end{aligned}$$

e quindi

$$\begin{aligned}
\rho(X,Y) &= \frac{b\mathbb{V}ar(X)}{\sqrt{\mathbb{V}ar(X)\cdot b^2\mathbb{V}ar(X)}} \\
&= \text{segno}(b)\cdot 1
\end{aligned}$$

Quindi $\rho = \pm 1$ a seconda che $b \gtrless 0$.

Proposizione 10. *Varianza di una combinazione lineare*

$$\mathbb{V}ar(aX+bY) = a^2\mathbb{V}ar(X) + b^2\mathbb{V}ar(Y) + 2ab\mathbb{C}ov(X,Y)$$

Dimostrazione 5.

$$\begin{aligned}
\mathbb{V}ar(aX+bY) &= \mathbb{E}\{[(aX+bY) - \mathbb{E}(aX+bY)]^2\} \\
&= \mathbb{E}\{[(a(X-\mathbb{E}(X)) + b(Y-\mathbb{E}(Y))]^2\} \\
&= a^2\mathbb{E}(X-\mathbb{E}(X))^2 + b^2\mathbb{E}(Y-\mathbb{E}(Y))^2 \\
&\quad +2ab\mathbb{E}(X-\mathbb{E}(X))(Y-\mathbb{E}(Y)) \\
&= a^2\mathbb{V}ar(X) + b^2\mathbb{V}ar(Y) + 2ab\mathbb{C}ov(X,Y)
\end{aligned}$$

In particolare se X e Y sono incorrelate ($\mathbb{C}ov(X,Y) = 0$) allora

$$\mathbb{V}ar(aX + bY) = a^2\mathbb{V}ar(X) + b^2\mathbb{V}ar(Y)$$

Esempio 26 (Covarianza e correlazione)**.** *Nel nostro esempio abbiamo*

$$\begin{aligned}
\mathbb{E}(X \cdot Y) &= 0 \cdot 0 \cdot 2/6 + 0 \cdot 1 \cdot 1/6 + 1 \cdot 0 \cdot 1/6 \\
&\quad +1 \cdot 1 \cdot 0 + 2 \cdot 0 \cdot 1/6 + 2 \cdot 1 \cdot 1/6 \\
&= 2/6 \\
\mathbb{E}(Y) &= 0 \cdot 4/6 + 1 \cdot 2/6 = 2/6 \\
\mathbb{E}(X) &= 0 \cdot 3/6 + 1 \cdot 1/6 + 2 \cdot 2/6 = 5/6 \\
\mathbb{C}ov(X,Y) &= \mathbb{E}(X \cdot Y) - \mathbb{E}(X)\mathbb{E}(Y) = 2/6 - 10/36 = 1/18\,.
\end{aligned}$$

Inoltre

$$\begin{aligned}
\mathbb{E}(Y^2) &= 0 \cdot 4/6 + 1 \cdot 2/6 = 2/6 \\
\mathbb{E}(X^2) &= 0 \cdot 3/6 + 1 \cdot 1/6 + 4 \cdot 2/6 = 9/6 \\
\mathbb{V}ar(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = 2/6 - 4/36 = 2/9 \\
\mathbb{V}ar(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 9/6 - 25/36 = 29/36\,,
\end{aligned}$$

e quindi

$$\rho(X,Y) = \frac{1/18}{\sqrt{2/9 \cdot 29/36}} = \frac{1}{\sqrt{58}} = 0.1313\,.$$

Valori attesi condizionati: proprietà

Proposizione 11.

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$$

Dimostrazione 6.

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(Y|X)) &= \sum_{i=1}^{h}\left[\sum_{j=1}^{k} y_j p_{Y|X}(y_j|x_i)\right] p_X(x_i) \\
&\quad \textit{poichè } p_{X,Y}(x_i,y_j) = p_{Y|X}(y_j|x_i)p_X(x_i) \\
&= \sum_{i=1}^{h}\sum_{j=1}^{k} y_j p_{X,Y}(x_i,y_j) \\
&= \sum_{j=1}^{k} y_j \left[\sum_{i=1}^{h} p_{X,Y}(x_i,y_j)\right] = \sum_{j=1}^{k} y_j p_Y(y_j) = \mathbb{E}(Y)
\end{aligned}$$

Proposizione 12.

$$\mathbb{E}(h(x)g(Y)|x)) = h(x)\mathbb{E}(g(Y)|x)$$

Dimostrazione 7.

$$\begin{aligned}
\mathbb{E}(h(x)g(Y)|x)) &= \sum_{j=1}^{k} h(x)g(y_j)p_{Y|X}(y_j|x) \\
&= h(x)\sum_{j=1}^{k} g(y_j)p_{Y|X}(y_j|x) = h(x)\mathbb{E}(g(Y)|x)
\end{aligned}$$

La seguente uguaglianza è detta **scomposizione della varianza**, ed è molto importante in Statistica.

Proposizione 13.

$$\mathbb{V}ar(Y) = \mathbb{E}(\mathbb{V}ar(Y|X)) + \mathbb{V}ar(\mathbb{E}(Y|X))$$

Dimostrazione 8.

$$\begin{aligned}
\mathbb{V}ar(Y) &= \sum_{j=1}^{k}[y_j - \mathbb{E}(Y)]^2 p_Y(y_j) \\
&= \sum_{j=1}^{k}[y_j - \mathbb{E}(Y)]^2 \sum_{i=1}^{h} p_{X,Y}(x_i, y_j) \\
&= \sum_{j=1}^{k}\sum_{i=1}^{h}[y_j - \mathbb{E}(Y|x_i) + \mathbb{E}(Y|x_i) - \mathbb{E}(Y)]^2 p_{X,Y}(x_i, y_j) \\
&= \sum_{j=1}^{k}\sum_{i=1}^{h}[y_j - \mathbb{E}(Y|x_i)]^2 p_{X,Y}(x_i, y_j) \\
&\quad + \sum_{j=1}^{k}\sum_{i=1}^{h}[\mathbb{E}(Y|x_i) - \mathbb{E}(Y)]^2 p_{X,Y}(x_i, y_j) \\
&\quad + 2\sum_{j=1}^{k}\sum_{i=1}^{h}[y_j - \mathbb{E}(Y|x_i)][\mathbb{E}(Y|x_i) - \mathbb{E}(Y)] p_{X,Y}(x_i, y_j)
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}ar(Y) &= \sum_{i=1}^{h}\left[\sum_{j=1}^{k}[y_j - \mathbb{E}(Y|x_j)]^2 p_{Y|X}(y_j|x_i)\right] p_X(x_i) \\
&\quad + \sum_{i=1}^{h}\left[\sum_{j=1}^{k}[\mathbb{E}(Y|x_i) - \mathbb{E}(\mathbb{E}(Y|X))]^2 p_{Y|X}(y_j|x_i)\right] p_X(x_i) \\
&\quad + 0 \\
&= \mathbb{E}(\mathbb{V}ar(Y|X)) + \mathbb{V}ar(\mathbb{E}(Y|X))
\end{aligned}$$

Rimane da mostrare

$$2\sum_{j=1}^{k}\sum_{i=1}^{h}[y_j - \mathbb{E}(Y|x_i)][\mathbb{E}(Y|x_i) - \mathbb{E}(Y)]p_{X,Y}(x_i, y_j) = 0$$

infatti

$$\begin{aligned}
&= 2\sum_{j=1}^{k}\sum_{i=1}^{h}[y_j - \mathbb{E}(Y|x_i)][\mathbb{E}(Y|x_i) - \mathbb{E}(Y)]p_{X,Y}(x_i, y_j) \\
&= \sum_{i=1}^{h}[\mathbb{E}(Y|x_i) - \mathbb{E}(Y)]\left\{\sum_{j=1}^{k}[y_j - \mathbb{E}(Y|x_i)]p_{Y|X}(y_j|x_i)\right\}p_X(x_i) \\
&= \sum_{i=1}^{h}[\mathbb{E}(Y|x_i) - \mathbb{E}(Y)]\left[\sum_{j=1}^{k}y_j p_{Y|X}(y_j|x_i) - \mathbb{E}(Y|x_i)\right]p_X(x_i) \\
&= \sum_{i=1}^{h}[\mathbb{E}(Y|x_i) - \mathbb{E}(X)]\,[\mathbb{E}(Y|x_i) - \mathbb{E}(Y|x_i)]\,p_X(x_i)
\end{aligned}$$

Esempio 27 (Valori attesi condizionati)**.** *Riprendendo l'esempio precedente abbiamo*

$$\begin{aligned}
\mathbb{E}(\mathbb{V}ar(Y|X)) &= 2/9 \cdot 3/6 + 0 \cdot 1/6 + 1/4 \cdot 2/6 \\
&= 1/9 + 0 + 1/12 \\
&= 7/36 \\
\mathbb{E}(\mathbb{E}(Y|X)) &= 1/3 \cdot 3/6 + 0 \cdot 1/6 + 1/2 \cdot 2/6 \\
&= 2/6 = \mathbb{E}(Y) \\
\mathbb{V}ar(\mathbb{E}(Y|X)) &= (1/3 - 2/6)^2 3/6 + (0 - 2/6)^2 1/6 \\
&\quad + (1/2 - 2/6)^2 2/6 \\
&= 0 \cdot 3/6 + 1/9 \cdot 1/6 + 1/36 \cdot 2/6 \\
&= 1/36
\end{aligned}$$

Quindi

$$\mathbb{V}ar(Y) = \mathbb{E}(\mathbb{V}ar(Y|X)) + \mathbb{V}ar(\mathbb{E}(Y|X)) = 1/36 + 7/36 = 2/9\,.$$

Considerando direttamente la distribuzione marginale si aveva

$$\begin{aligned}
\mathbb{E}(Y) &= 0 \cdot 4/6 + 1 \cdot 2/6 = 2/6 \\
\mathbb{V}ar(Y) &= (0 - 2/6)^2 4/6 + (1 - 2/6)^2 2/6 = 2/9\,.
\end{aligned}$$

Esempio 28 (Variabili casuali continue)**.** *Calcoliamo* $\mathbb{E}(XY)$

$$\begin{aligned}
\mathbb{E}(XY) &= \int_0^1 \int_0^1 xy(x+y)dxdy \\
&= \int_0^1 \left[\int_0^1 xy(x+y)dy\right] dx \\
&= \int_0^1 \left(x^2 \left[y^3/3\right]_0^1 + x\left[y^2/2\right]_0^1\right) dx \\
&= \int_0^1 (x^2/2 + x/3)dx \\
&= \left[x^3/6\right]_0^1 + \left[x^2/6\right]_0^1 = 1/6 + 1/6 = 1/3
\end{aligned}$$

e quindi

$$\mathbb{C}ov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 1/3 - 7/12 \cdot 7/12 = -1/144$$

Esercizio 14. *Calcolare l'indice di correlazione.*

Variabile casuale normale bivariata

Introduciamo ora l'estensione al caso bivariato della v.c. normale.

Diremo che (X,Y) sono v.c. congiuntamente distribuite come una v.c. normale bivariata se la sua funzione di densità è

$$\begin{aligned}
f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho(X,Y)^2)}} \cdot \\
&\cdot e^{-\frac{1}{2(1-\rho(X,Y)^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho(X,Y)\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}}
\end{aligned}$$

di medie $\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$, e varianze $\sigma_X^2 = \mathbb{V}ar(X)$, $\sigma_Y^2 = \mathbb{V}ar(Y)$.

Si noti che se $\rho(X,Y) = 0 \iff \mathbb{C}ov(X,Y) = 0$, allora

$$\begin{aligned}
f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{1}{2}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}} \\
& \quad \frac{1}{2\sqrt{\pi}\sigma_X} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2} \frac{1}{2\sqrt{\pi}\sigma_Y} e^{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2} \\
&= f_X(x) f_Y(y)
\end{aligned}$$

quindi X e Y sono indipendenti. Questo è l'unico caso in cui l'incorrelazione implica l'indipendenza.

2.5 Teoremi limite della Probabilità

Teoremi limite della probabilità

"Me so magna' i schei col 53"
(Anonimo veneziano)

- Legge dei grandi numeri
- Teorema del limite centrale

"Per la legge dei grandi numeri ..."

C'è stata in Italia quasi una psicosi (con effetti anche tragici) intorno all'uscita del numero 53 sulla ruota di Venezia nel gioco del lotto. Molti giocatori hanno puntato forti somme su questo numero in quanto la sua estrazione ritardava da molto tempo. Tali giocate erano motivate dall'affermazione che *"prima o poi per la legge dei grandi numeri il numero 53 doveva essere estratto"*.

Indichiamo con Y_i il risultato dell'estrazione i-esima (di 5 numeri in blocco da un'urna che ne contiene 90) sulla ruota di Venezia. Se $Y_i = 1$, il numero 53 stato estratto; se $Y_i = 0$, il numero 53 non stato estratto.

Se l'estrazione non è truccata (palline indistinguibili, urna ben mescolata, ...) ogni Y_i ha dunque una distribuzione di Bernoulli con parametro

$$p = \mathbb{E}(Y_i) = 1 \cdot \Pr(Y_i = 1) + 0 \cdot \Pr(Y_i = 0) = \Pr(Y_i = 1) = \frac{5}{90}$$

(perché?).

Inoltre possiamo assumere che ogni Y_i abbia la stessa distribuzione e che i risultati delle estrazioni (ovvero le Y_i) siano tra loro indipendenti.

La variabile $S_n = \sum_{i=1}^{n} Y_i$ conta il numero di successi (è uscito il 53) in n estrazioni indipendenti e si distribuisce dunque come una v.c. binomiale $\mathcal{B}(n, p)$.

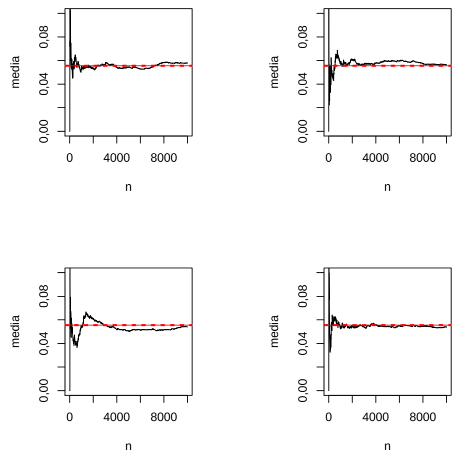
Intuitivamente, siamo portati a dire che per un numero n 'grande' di prove , $n \gg 0$, la frequenza relativa S_n/n si avvicina a p:

$$\frac{S_n}{n} \approx p$$

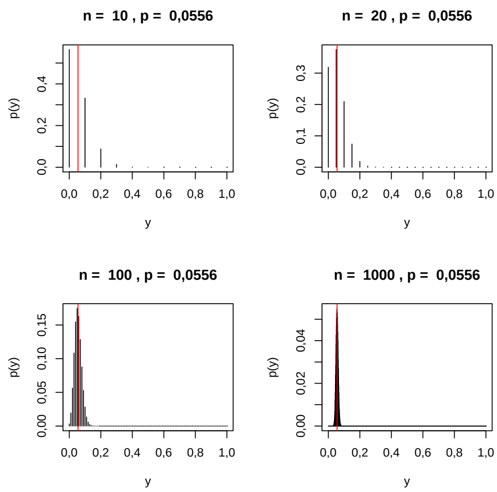
dove $p = 5/90$ la probabilità che il numero 53 sia estratto.

Ma cosa significa esattamente "si avvicina"?

Nel grafico rappresentiamo il risultato di 4 esperimenti con 10.000 estrazioni. Possiamo notare che S_n/n si avvicina a p (la linea rossa), in maniera erratica e in maniera differente per ogni esperimento.



Vediamo cosa accade alla distribuzione di S_n/n quando n aumenta (nei grafici, $p = 5/90$):



La distribuzione tende a concentrarsi intorno al valore $p = 5/90$.

Convergenza in probabilità (o debole)

Ci sono diversi modi per esprimere il fatto che S_n/n si avvicina a p.

Potremmo ad esempio scrivere che, per n grande e per ε piccolo a piacere

$$\Pr\{|S_n/n - p| \geq \varepsilon\} \approx 0,$$

o equivalentemente

$$\lim_{n\to\infty} \Pr\{|S_n/n - p| \geq \varepsilon\} = 0.$$

In simboli questo tipo di convergenza si denota con

$$S_n/n \xrightarrow{P} p$$

e $\xrightarrow{P}$ si legge "converge in probabilità (o in senso debole)".

Più in generale diremo che una successione $Y_1, Y_2 \ldots$ **converge in probabilità (o in senso debole)** ad una v.c. Y se, per ogni $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|Y_n - Y| \geq \varepsilon) = 0,$$

ovvero

$$\lim_{n\to\infty} \Pr(|Y_n - Y| < \varepsilon) = 1.$$

Convergenza in media quadratica

Un'altra formalizzazione del concetto di "vicinanza" potrebbe richiedere che in media gli scostamenti (al quadrato) di S_n/n da p siano piccoli, quando n è grande:

$$\mathbb{E}[(S_n/n - p)^2] \approx 0,$$

o equivalentemente

$$\lim_{n\to\infty} \mathbb{E}[(S_n/n - p)^2] = 0.$$

In simboli questo tipo di convergenza si denota con

$$S_n/n \xrightarrow{m.q.} p$$

e $\xrightarrow{m.q.}$ si legge "converge in media quadratica".

Più in generale diremo che una successione $Y_1, Y_2 \ldots$ **converge in media quadratica** ad una v.c. Y se

$$\lim_{n\to\infty} \mathbb{E}[(Y_n - Y)^2] = 0.$$

Proposizione 14. *La convergenza in media quadratica implica la convergenza in probabilità:*

$$Y_n \xrightarrow{m.q.} Y \Longrightarrow Y_n \xrightarrow{P} Y$$

Per dimostrare questa proposizione ci servono alcuni risultati preliminari.

Proposizione 15. *(Disuguaglianza di Markov) Sia Y una v.c. che assume valori non negativi, allora per ogni numero reale $a > 0$*

$$\Pr(Y \geq a) \leq \frac{\mathbb{E}(Y)}{a}.$$

Dimostrazione 9. *Sia*

$$X = \begin{cases} 1 & \textit{se } Y \geq a \\ 0 & \textit{altrimenti} \end{cases}$$

Si noti che, poiché $Y \geq 0$, $X \leq Y/a$ e quindi $\mathbb{E}(X) \leq \mathbb{E}(Y)/a$, con $\mathbb{E}(X) = 1 \cdot \Pr(X = 1) + 0 \cdot \Pr(X = 0) = \Pr(X = 1) = \Pr(Y \geq a)$.

Siamo ora in grado di dimostrare la Proposizione [14].

Dimostrazione 10. *Sia $\varepsilon > 0$. Grazie alla disuguaglianza di Markov applicata a $(Y_n - Y)^2$, si ha che*

$$\begin{aligned} \lim_{n\to\infty} \Pr(|Y_n - Y| \geq \varepsilon) &= \lim_{n\to\infty} \Pr((Y_n - Y)^2 \geq \varepsilon^2) \\ &\leq \lim_{n\to\infty} \frac{\mathbb{E}[(Y_n - Y)^2]}{\varepsilon^2}. \end{aligned}$$

Inoltre, per ipotesi, $\mathbb{E}[(Y_n - Y)^2] \longrightarrow 0$ e perciò $\lim_{n\to\infty} \Pr(|Y_n - Y| \geq \varepsilon) = 0$.

Dalla disuguaglianza di Markov discende un altro importante risultato:

Proposizione 16. *(Disuguaglianza di Chebychev) Sia Y una v.c. con valore atteso $\mathbb{E}(Y) = \mu$ e varianza $\mathbb{V}ar(Y) = \sigma^2$. Allora*

$$\Pr(|Y - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Dimostrazione 11. *Essendo $(Y - \mu)^2$ una v.c. non negativa, basta applicare la disuguaglianza di Markov con $a = \varepsilon^2$:*

$$\Pr((Y - \mu)^2 \geq \varepsilon^2) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Ma $(Y - \mu)^2 \geq \varepsilon^2$ se e solo se $|Y - \mu| \geq \varepsilon$ e quindi

$$\Pr(|Y - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

L'importanza delle due disuguaglianze di Markov e Chebychev è che ci danno dei limiti superiori per delle probabilità senza alcun riferimento alla distribuzione di probabilità della v.c. Y. Basta conoscerne il valore atteso o il valore atteso e la varianza.

Ad esempio, si supponga che il numero di clienti in un giorno ad uno sportello sia una v.c. Y con $\mathbb{E}(Y) = 130$. Cosa possiamo dire della probabilità che possano arrivare più di 250 clienti? Grazie alla disuguaglianza di Markov possiamo dire che $\Pr(Y \geq 250) \leq 130/250 = 0.52$.

Ovviamente questi limiti possono essere molto grossolani. Se $Y \sim \mathcal{N}(0, 1)$ sappiamo che $\Pr(|Y| \geq 1.96) = 0.05$, mentre la disuguaglianza di Chebychev afferma che $\Pr(|Y| \geq 1.96) \leq 1/(1.96^2) \approx 0.260$.

Richiami sulle somme variabili casuali

Proposizione 17. *Siano $Y_1, \ldots, Y_n$ v.c. con valore atteso rispettivamente $\mu_1, \ldots, \mu_n$. Allora*

$$\mathbb{E}(Y_1 + \ldots + Y_n) = \mu_1 + \ldots + \mu_n.$$

Proposizione 18. *Siano $Y_1, \ldots, Y_n$ v.c. indipendenti con varianza $\sigma_1^2, \ldots, \sigma_n^2$ rispettivamente. Allora*

$$\mathbb{V}ar(Y_1 + \ldots + Y_n) = \sigma_1^2 + \ldots + \sigma_n^2.$$

Una conseguenza importante di queste due proposizioni è la seguente:

Proposizione 19. *Siano $Y_1, \ldots, Y_n$ v.c. indipendenti, tutte con valore atteso μ e varianza σ^2 e sia $\overline{Y}_n = \sum_{i=1}^n Y_i/n$. Allora*

$$\mathbb{E}(\overline{Y}_n) = \sum_{i=1}^n \frac{\mathbb{E}(Y_i)}{n} = n\frac{\mu}{n} = \mu,$$

$$\mathbb{V}ar(\overline{Y}_n) = \sum_{i=1}^n \frac{\mathbb{V}ar(Y_i)}{n^2} = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

La legge debole dei grandi numeri

Proposizione 20 (Legge debole dei grandi numeri)**.** *Sia $Y_1, Y_2, \ldots$ una successione di v.c. indipendenti, ciascuna con valore atteso μ e varianza σ^2. Allora, per ogni $\varepsilon > 0$,*

$$\lim_{n \to \infty} \Pr\{|\overline{Y}_n - \mu| \geq \varepsilon\} = 0,$$

ovvero $\overline{Y}_n \xrightarrow{P} \mu$.

Dimostrazione 12. *Per la Proposizione [19], quando $n \longrightarrow \infty$,*

$$\mathbb{E}[(\overline{Y}_n - \mu)^2] = \mathbb{V}ar(\overline{Y}_n) = \frac{\sigma^2}{n} \longrightarrow 0,$$

che significa che $\overline{Y}_n \xrightarrow{m.q.} \mu$. La convergenza debole segue allora dalla Proposizione [14].

"Per la legge dei grandi numeri ...", un'amara conclusione

Tornando al gioco del lotto, la legge dei grandi numeri garantisce che $S_n/n \xrightarrow{P} p$ e quindi che la frequenza relativa del 53 tende a 5/90 quando il numero di estrazioni cresce.

Nella nostra discussione abbiamo supposto l'indipendenza e l'identica distribuzione delle v.c. Y_i. Nel gioco del lotto questo vuol dire che le estrazioni sono indipendenti e che la probabilità di estrarre il 53 è la stessa nelle varie estrazioni.

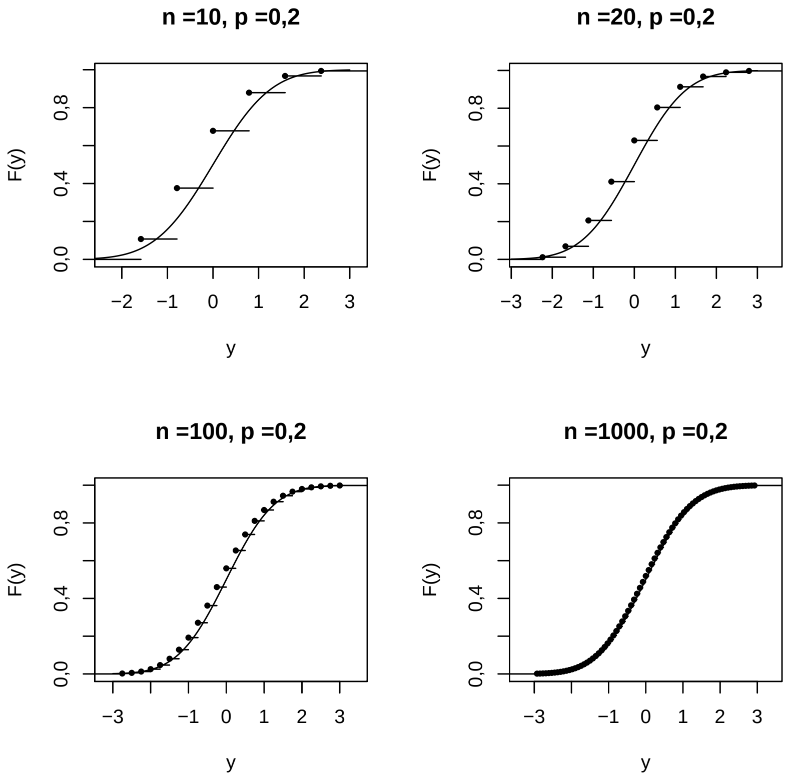
Anche se l'indipendenza non è un'ipotesi fondamentale per provare delle leggi dei grandi numeri (ma questo richiede nozioni più avanzate del calcolo delle probabilità), lo è nel gioco del lotto. E quindi anche se la frequenza dei successi tende in probabilità alla probabilità di successo, ogni volta che avviene una nuova estrazione il 53 ha **sempre la medesima probabilità** di uscire (sic!).

Il teorema del limite centrale

Ci occupiamo qui di un diverso tipo di convergenza. Abbiamo visto che, sotto determinate ipotesi, S_n/n converge debolmente alla probabilità di successo p. Consideriamo ora invece di S_n/n una sua espressione standardizzata

$$Z_n = \sqrt{n}\frac{S_n/n - p}{\sqrt{p(1-p)}}.$$

Nei grafici abbiamo riportato la funzione di ripartizione della variabile Z_n (con $p = 2$) e, con la linea continua, la funzione di ripartizione di una v.c. $\mathcal{N}(0, 1)$.



Proposizione 21 (Teorema del limite centrale)**.** *Sia* $Y_1, Y_2, \ldots$ *una successione di v.c. ciascuna con* $\mathbb{E}(Y_i) = \mu$ *e* $\mathbb{V}ar(Y_i) = \sigma^2$*. Allora, posto* $Z_n = \sqrt{n}(\overline{Y}_n - \mu)/\sigma$*, per ogni* z

$$\lim_{n \to \infty} \Pr(Z_n \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt \,.$$

In simboli il teorema del limite centrale si denota scrivendo

$$Z_n = \sqrt{n}\frac{(\overline{Y}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

dove $\xrightarrow{\mathcal{D}}$ si legge “converge in distribuzione”.

Una lettura grezza (ma pratica) del teorema del limite centrale la seguente:

$$\overline{Y}_n \overset{a}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

dove $\overset{a}{\sim}$ significa “distribuita approssimatamente”.

Somme di v.c. normali

Il teorema del limite centrale è un risultato valido solo asintoticamente, cioè quando $n \to \infty$.

Se le v.c. sono normali, la distribuzione esatta di $\overline{Y}_n$ può essere ottenuta dalla proposizione seguente.

Proposizione 22. *Se $Y_1, \ldots, Y_n$ sono variabili casuali normali* indipendenti *tra loro e se $a_0, \ldots, a_n$ sono delle costanti reali qualsiasi, allora*

$$a_0 + \sum_{i=1}^{n} a_i Y_i \sim \mathcal{N}\left(a_0 + \sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

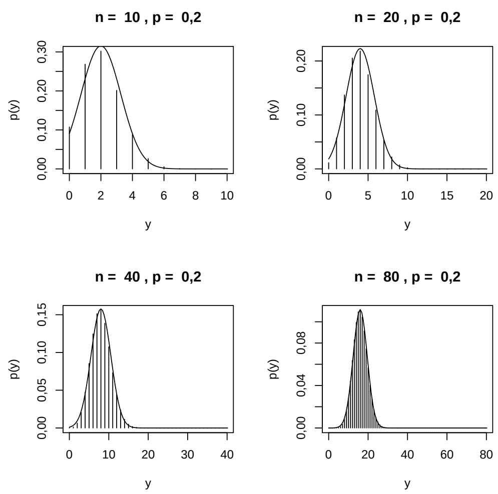
dove μ_i e σ_i^2 indicano rispettivamente la media e la varianza di Y_i.

Esempio 29. *Nel caso della media $\overline{Y}_n$ di n v.c. normali Y_i, indipendenti e con medesime media μ e varianza σ^2, si ha $a_0 = 0$ e $a_i = 1/n$, $i = 1, \ldots, n$. Quindi:*

$$\overline{Y}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) .$$

Approssimare la binomiale con la normale

All'aumentare del numero n di prove, la distribuzione di una variabile casuale binomiale di parametri n e p si "avvicina" sempre di più a quella di una normale con parametri $\mu = np$ e $\sigma^2 = np(1-p)$.



Alcune considerazioni:

- L'approssimazione è valida quando n è abbastanza grande e p e $1-p$ non sono vicini a zero.

- Regola pratica

  1. si calcola l'intervallo di estremi $np \pm 3\sqrt{np(1-p)}$
  2. se esso è contenuto nell'intervallo $[0, n]$ allora l'approssimazione può ritenersi valida.

Esempio 30. *Vogliamo calcolare* $\Pr(B \leq 25)$ *dove* $B \sim \mathcal{B}(30, 0.7)$*. L'intervallo* $[13.47, 28.53]$ *è contenuto nell'intervallo* $[0, 30]$ *quindi procediamo con l'approssimazione. Calcoliamo il valore standardizzato corretto corrispondente a 25:*

$$z_0 = \frac{(25 + \mathbf{0.5}) - 30 \cdot 0.7}{\sqrt{30 \cdot 0.7(1 - 0.7)}} \simeq 1.79$$

e utilizziamo le tavole della distribuzione normale:

$$\Pr(B \leq 25) \simeq \Pr(Z \leq 1.79) = 0.9633.$$

Si confronti questa probabilità con la probabilità esatta 0.9699.

Si osservi che nel calcolo di z_0 *abbiamo aggiunto 0.5 al valore dato 25. Si tratta della cosiddetta* ***correzione per continuità*** *che migliora l'approssimazione di una v.c. discreta con una continua.*

Capitolo 3

Inferenza statistica

3.1 Inferenza Statistica

"Statistica Descrittiva" vs "Inferenza Statistica"

- il punto di partenza di una indagine statistica è costituito da un insieme (che chiamiamo la **popolazione di riferimento**), disomogeneo all'interno (ovvero non tutti gli elementi sono uguali tra di loro) e che costituisce la "parte del mondo che ci interessa";
- gli elementi di questo insieme (che di volta in volta nei problemi concreti saranno persone, aziende, titoli azionari,...) vengono convenzionalmente indicati come **unità statistiche**;
- l'analisi statistica vuole, nelle sostanza, utilizzare i **dati** disponibili (misurazioni/rilevazioni di alcune delle caratteristiche delle unità statistiche condotte su alcune o tutte le unità statistiche che appartengono alla popolazione di riferimento) per fare delle affermazioni sulla popolazione.

Nel contesto brevemente schematizzato parliamo di

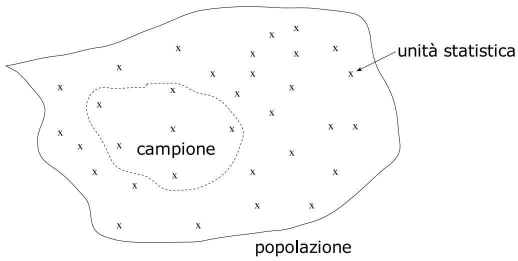
- **Statistica descrittiva**: ("quasi" sinonimi: esplorazione statistica dei dati, statistica senza modello probabilistico) se disponiamo di dati riferiti a tutta la popolazione di riferimento (o, come spesso accade, ci comportiamo come se l'affermazione precedente fosse vera!).
- **Inferenza statistica**: se, viceversa, 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 perché i problemi di inferenza vengono normalmente affrontati in accordo allo schema

$$\begin{array}{ccc} \text{descrizione} & & \text{affermazioni} \\ \text{caratteristiche} & \rightarrow & \text{sulle caratteristiche} \\ \text{campione} & & \text{della popolazione} \end{array}$$

Questo però non vuol dire che l'insieme delle idee e dei metodi riferibili ai due contesti non sia ben differenziato.

Lo schema qui sotto cerca di esemplificare la situazione. L'insieme delimitato dalla linea tratteggiata indica il campione. Le variabili di interesse sarebbero in questo caso rilevate solamente sulle sei unità statistiche che fanno parte del campione. Nonostante le informazioni sulla popolazione siano incomplete, in un problema di inferenza siamo però ambiziosi: con le informazioni rilevate sul solo campione vogliamo "produrre" affermazioni su tutta la popolazione.



Perché indagini di tipo campionario sono frequenti?

Esempio 31 (Tempo e/o costo)**.** *L'ISTAT fornisce informazioni sulla disoccupazione in Italia con cadenza trimestrale. Le informazioni provengono da una indagine campionaria piuttosto ampia (parecchie decine di migliaia di nuclei familiari). Non però esaustiva (non tutti sono infatti intervistati). Intervistare tutti i cittadini italiani ogni tre mesi è infatti organizzativamente troppo oneroso (richiederebbe una struttura organizzativa "immensa"). Il costo ovviamente diminuirebbe se ci accontentassimo di una rilevazione fatta non ogni trimestre. Ma in questo caso perderemmo la tempestività dell'informazione.*

Esempio 32 (Tempo e/o costo)**.** *Quanto tempo e denaro dovrebbe investire una azienda dolciaria per verificare, senza una rilevazione di tipo parziale, ovvero campionaria, se una nuova tortina potrebbe incontrare i gusti della clientela? Una rilevazione esaustiva richiederebbe di farla assaggiare a tutti i residenti in Italia o, perché no, se il piano è di vendere la tortina anche all'estero, in tutta Europa, in tutti i paesi occidentali,...*

Esempio 33 (La popolazione di interesse può essere infinita e virtuale)**.** *Una delle fasi dello studio di un nuovo farmaco è costituita dalla verifica che la tossicità del farmaco sia sufficientemente piccola rispetto alla gravità della malattia che vuole curare e alla tossicità di altri farmaci noti. Lasciando perdere i dettagli (anche se, in questo caso, sono importanti per ovvi aspetti etici), in pratica, questo si concretizza nel somministrare il farmaco ad alcuni pazienti e nel rilevare gli effetti secondari. La popolazione che ci interessa* *in questo caso è una popolazione teoricamente infinita e solamente virtuale: l'insieme di tutti i pazienti a cui potremmo voler somministrare il farmaco da oggi fino al giorno della fine del mondo.*

Non è ovviamente sensato somministrare il farmaco a tutta la popolazione prima di pronunciarci sulla tossicità del farmaco. Concludere con certezza, ovvero sulla base di una somministrazione esaustiva, che il farmaco è troppo "tossico" il giorno della fine del mondo è inutile. E per di più potrebbe essere non etico: magari qualche millenio prima lo potevamo già dire e allora perché abbiamo continuato a somministrarlo?

Esempio 34 (La rilevazione "distrugge" le unità statistiche e quindi, dopo una rilevazione esaustiva, la popolazione di partenza non interessa più perché non esiste più!)**.** *Una azienda farmaceutica produce tra le altre cose delle "pasticche" antibiotiche. Tra i controlli effettuati c'è la verifica a posteriori della titolazione delle "pasticche" prodotte in un determinato lotto di produzione. Un certo numero di "pasticche" vengono analizzate per verificare se la quantità di antibiotico che contengono è all'interno di un certo prescritto intervallo di tolleranza che include ovviamente il titolo nominale (che è quello indicato sulla confezione, ad esempio 5mg di sostanza attiva per "pasticca"). La misurazione della quantità di sostanza attiva richiede di norma la distruzione della "pillola" (la pillola viene triturata, mescolata a solventi,...). Se dovessimo farlo per tutte le "pillole" prodotte in un certo giorno non avremmo più pillole da dare ai pazienti!*

Esempio 35 (Precisione dei risultati)**.** *Può sembrare strano ma delle volte è stato dimostrato che rilevazioni campionarie (incomplete) portano a risultati più precisi di rilevazioni esaustive.*

È ad esempio il caso di rilevazioni semplici ma noiose fatte da operatori umani (non da macchine). La noia provoca cali di attenzione e quindi errori. Perciò...

Popolazione e campione: dobbiamo conoscere la relazione

- supponiamo che la popolazione di riferimento siate voi (gli studenti presenti alla prima lezione del corso di Statistica II presso la facoltà...)

- e che per qualche strano motivo io voglia conoscere la vostra altezza media ma che per qualche altro motivo ancora più misterioso possa misurare l'altezza solamente di 10 di voi.

- Il primo problema diventa come scegliere i dieci da misurare; due tra le molte possibilità "teoriche" sono:

A) scelgo completamente a caso 10 dei presenti (ad esempio, metto dei foglietti uguali con il vostro numero di matricola in un barattolo, mescolo bene, poi ne estraggo 10); misuro poi l'altezza dei 10 sorteggiati;

B) vi faccio allineare lungo il muro, vi ordino dal più alto al più piccolo (visivamente), scelgo i 10 più alti e misuro l'altezza di questi 10.

- In ambedue i casi, alla fine ci troviamo tra le mani 10 numeri (le altezze dei 10 studenti "misurati"). È però intuitivamente chiaro che per stimare l'altezza media di tutti i presenti nell'aula non posso utilizzare questi numeri (i nostri dati) nella stessa maniera.

- Ad esempio nel primo caso posso pensare di stimare l'altezza media utilizzando la media aritmetica delle 10 misurazioni fatte. Se non sono stato particolarmente sfortunato posso infatti pensare di non aver sorteggiato tutti studenti bassi o tutti studenti alti e quindi che la media delle dieci misure "cada vicino" alla altezza media di tutti.

- Nel secondo caso però non è sensato "stimare" l'altezza media nella stessa maniera: con certezza sappiamo che in questa maniera sovrastimeremo la quantità che vogliamo conoscere.

- È facile capire che quello che cambia nei due casi è la relazione tra il campione e la popolazione.

- In generale quindi non possiamo pensare di affrontare un problema di inferenza senza sapere (e saper descrivere appropriatamente) la relazione tra il campione e la popolazione (o almeno tra quello che abbiamo misurato sul campione e quello che della popolazione vogliamo conoscere).

Errare è l'unica certezza

"Produrre" affermazioni esatte sulla popolazione conoscendo solamente le caratteristiche di un sottoinsieme delle unità statistiche è impossibile (a meno che non supponiamo di avere ricevuto da Mago Merlino una sfera di cristallo!).

Quindi a priori sappiamo che commetteremo degli errori.

Per rendere utili le nostre affermazioni dovremo allora occuparci anche di capire di quanto sono sbagliate.

Esempio 36. *Supponiamo di sperimentare un nuovo farmaco su 20 pazienti e che solo uno di questi 20 pazienti mostri problemi gravi di tossicità (effetti secondari non voluti e non banali).*

Sembra naturale, sulla base di questi dati, "stimare" la probabilità che il farmaco induca effetti tossici rilevanti in 5% (ovvero un paziente ogni venti).

In questo caso la popolazione di riferimento è data da tutti i pazienti a cui potremmo pensare di somministrare il farmaco sotto analisi. È una popolazione virtuale e teoricamente infinita. È chiaro che non ci aspettiamo che la percentuale di tutti i possibili pazienti che potrebbero presentare problemi di tossicità sia esattamente uguale al 5%. Saremmo stati troppo fortunati.

Non è però irrilevante chiedere di quanto potrebbe essere differente (ovvero di quanto abbiamo sbagliato). Si considerino difatti le seguenti due ipotetiche alternative:

i) sulla base dei dati, procedendo in qualche maniera strana ancora da studiare, arriviamo a concludere che la percentuale incognita di pazienti della popolazione che potrebbero esibire problemi di tossicità è compresa tra il 2% e il 77%;

ii) oppure, seconda alternativa, è compresa tra il 4,8% e il 5,8%.

Le due alternative sono differenti tra di loro per il "differente errore" che attribuiscono alla "stima" di prima (5% di tossicità).

La differenza non è solo accademica.

Infatti, se fosse vera la prima alternativa la conclusione a cui saremmo arrivati è che, con i dati disponibili, della incognita probabilità di manifestare tossicità in realtà non sappiamo praticamente niente.

Viceversa, nel caso arrivassimo alla seconda alternativa potremmo concludere che "certo la vera probabilità di manifestare tossicità non la conosciamo esattamente ma che, sulla base dei dati possiamo dire che più o meno è uguale al 5%".

Inferenza Statistica e Probabilità

Il "trucco" alla base dell'inferenza statistica si concretizza nel descrivere la relazione tra la popolazione e il campione utilizzando il calcolo delle probabilità.

Ovvero, nella sostanza, interpreteremo i risultati sperimentali (ovvero i dati disponibili) come uno dei tanti risultati che un meccanismo probabilistico (un esperimento casuale) poteva fornirci.

Questa costruzione cercheremo di illustrarla nel seguito del corso. Inutile entrare quindi ora nei dettagli.

Una conseguenza importante sarà che potremo utilizzare in maniera naturale il calcolo delle probabilità "per misurare gli errori".

3.2 Controllo di qualità in un impianto che produce lastre di metallo

Un primo esempio di inferenza statistica

Stima del valore atteso, sua distribuzione campionaria, intervalli di confidenza e verifica d'ipotesi nel caso di un campione tratto da una v.c. normale di varianza nota.

Il problema ed i dati

Una industria metallurgica produce, tra l'altro, delle lastre di metallo con uno spessore nominale di $14mm$. In realtà esiste una tolleranza di $\pm 0.5mm$, ovvero, una lastra è considerata soddisfacente, per quello che riguarda lo spessore, se

$$13.5 \leq \text{spessore} \leq 14.5. \tag{3.1}$$

La produzione è organizzata in turni di 6 ore. All'inizio di ogni turno vengono estratte a caso da un operatore 5 lastre tra quelle prodotte nel turno precedente e ne viene misurato lo spessore. Queste 5 misure vengono utilizzate per decidere se le "macchine" stanno lavorando in maniera soddisfacente, ovvvero se il numero di lastre che non rispettano la (3.1) è sufficientemente piccolo. In particolare, se si decide per il sì la produzione del nuovo turno inizia immediatamente. Viceversa se si decide per il no, la produzione viene bloccata e le macchine vengono "ritarate".

I dati raccolti in un particolare turno (in mm) sono stati:

$$14.33 \;\; 14.19 \;\; 14.39 \;\; 14.43 \;\; 14.17.$$

Nel seguito considereremo il problema di utilizzare questi dati per decidere se bloccare o non bloccare temporaneamente la produzione.

Una possibile formulazione del problema

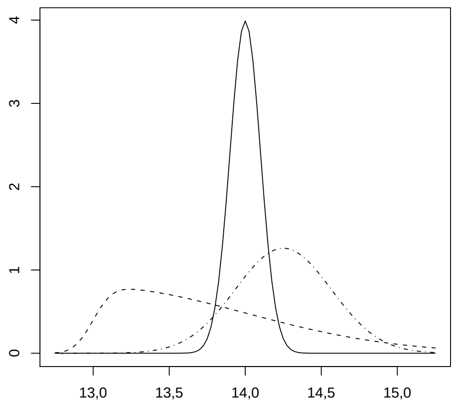
- Nessun processo produttivo è in grado di produrre lastre *esattamente* dello stesso spessore.

- All'inizio della lavorazione di una lastra (o all'inizio del turno,. . .) solo *Mago Merlino* sarebbe in grado di indovinarne esattamente lo spessore.

- Possiamo però pensare che lo spessore sia il risultato di un esperimento casuale e descriverne le caratteristiche utilizzando il calcolo delle probabilità.

- In particolare, potremmo guardare agli spessori che, in un determinato momento, il processo "potrebbe produrre" come ad una variabile casuale continua con funzione di densità $f(\cdot)$.
- Il problema diventa allora quello di utilizzare i dati disponibili per dire se la densità $f(\cdot)$ assegna una eccessiva probabilità all'evento "lastra difettosa" (= lastra il cui spessore non soddisfa la (3.1)). Si veda la pagina seguente, per alcuni esempi.
- Se questo accade e quindi se il processo sta, *almeno potenzialmente*, producendo "troppe" lastre difettose, si decide di sospendere la produzione.

Tre possibili situazioni



La densità disegnata con una linea continua indica una situazione soddisfacente: la probabilità di ottenere una lastra difettosa (spessore inferiore a $13.5mm$ o maggiore di $14.5mm$) è nulla (o quasi). Le altre due raccontano storie diverse: l'impianto sta producendo una frazione non piccola di lastre o troppo sottili o troppo spesse.

Informazioni aggiuntive sul processo

Cercare di stimare l'intera funzione di densità utilizzando solo le nostre 5 osservazioni sembra un'operazione eccessivamente avventurosa.

Fortunatamente nel caso in esame esistono delle informazioni aggiuntive. Infatti, precedentemente, le caratteristiche del processo sono state studiate raccogliendo alcune migliaia di misurazioni per alcune decine di turni.

Indicate con $Y_1, Y_2, \ldots$ le variabili casuali che descrivono lo spessore della prima lastra prodotta in un turno, della seconda e così via, le principali conclusioni delle analisi condotte sono:

- non esiste nessun tipo di dipendenza tra le Y_i;
- tutte le Y_i hanno la stessa distribuzione di probabilità;
- questa distribuzione comune è ben approssimata da una normale con valore atteso μ e varianza 0.01 dove μ è un *parametro* ignoto che può essere diverso da turno a turno.

Un modello è buono perché è utile non perché è vero

Nel seguito adotteremo come "esattamente" vere le conclusioni descritte nel lucido 162.

E' importante però rendersi conto che possono al più essere considerate una descrizione semplice ed operativamente utile di una realtà complessa.

Ad esempio la distribuzione dello spessore **non** può essere esattamente normale: una normale con varianza non nulla può assumere qualsiasi valore reale, lo spessore è però non negativo; dall'altra parte una normale può assegnare una probabilità così piccola a valori negativi che possiamo considerare quest'ultima trascurabile da un punto di vista pratico.

Analogo discorso può essere fatto per l'identica distribuzione e l'indipendenza.

Stima del valore atteso

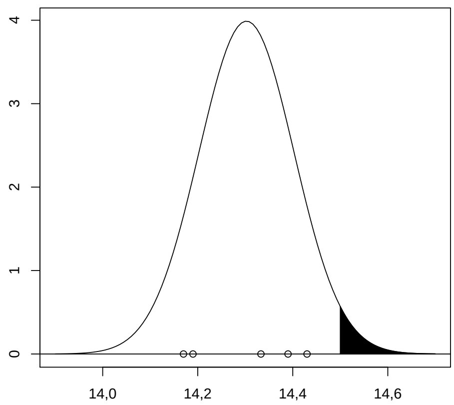
Le informazioni aggiuntive ci portano a considerare le 5 misure dello spessore come determinazioni di 5 v.c. indipendenti con la stessa distribuzione normale con valore atteso μ ignoto e varianza nota ed uguale a 0.01. Un'altra maniera di descrivere la situazione consiste nel dire che siamo in presenza di **determinazioni indipendenti ed identicamente distribuite (abbreviazione i.i.d.)** tratte da una variabile normale....

La funzione di densità dello spessore è quindi "quasi" nota. Ci manca solo il valore atteso. Sembra al proposito "ragionevole" utilizzare la media delle osservazioni come **"stima"** del vero valore atteso μ, ovvero porre

$$\text{stima del valore atteso} = \overline{y} = \frac{14.33 + \cdots + 14.17}{5} = 14.302.$$

Densità stimata

Il grafico mostra la densità di una normale con valore atteso 14.302 e varianza 0.01. L'area evidenziata rappresenta la probabilità (stimata) di produrre una lastra troppo spessa. La probabilità (stimata) di produrre una lastra troppo sottile è praticamente nulla. I "cerchietti" sull'asse delle x indicano le osservazioni.



Stima della "difettosità"

Due eventi particolarmente importanti nel presente contesto sono

$$\begin{array}{rcccl} A & = & \{\text{lastra troppo sottile}\} & = & \{Y < 13.5\} \\ B & = & \{\text{lastra troppo spessa}\} & = & \{Y > 14.5\} \end{array}$$

dove Y indica la variabile casuale che descrive lo spessore. Ovviamente sia $\Pr(A)$ che $\Pr(B)$ sono funzione di μ. In particolare, ricordando che[1]

$$\text{se } Y \sim \mathcal{N}(\mu, \sigma^2) \text{ allora } (Y - \mu)/\sigma \sim \mathcal{N}(0, 1)\,,$$

le probabilità di questi eventi possono essere agevolmente calcolate dalla funzione di ripartizione di una normale standard.

[1]Ci si ricordi che $\sim$ si legge "si distribuisce come"

In particolare,

$$\begin{aligned}
\Pr(A) &= \Pr(Y < 13.5) = \\
&= \Pr\left(\frac{Y-\mu}{0.1} < \frac{13.5-\mu}{0.1}\right) = \\
&= \Pr\left(\mathcal{N}(0,1) < \frac{13.5-\mu}{0.1}\right) = \\
&= \Pr\left(\mathcal{N}(0,1) \leq \frac{13.5-\mu}{0.1}\right).
\end{aligned}$$

Possiamo quindi scrivere

$$\Pr(A) = \Phi\left(\frac{13.5-\mu}{0.1}\right)$$

dove con $\Phi(\cdot)$ abbiamo indicato la funzione di ripartizione di una $\mathcal{N}(0,1)$. Si noti che abbiamo usato il fatto che, se Y è una variabile casuale continua, allora $\Pr(Y = y) = 0$ per qualsivoglia valore y.

Per l'altra probabilità troviamo

$$\begin{aligned}
\Pr(B) &= \Pr(Y > 14.5) = \\
&= 1 - \Pr(Y \leq 14.5) = \\
&= 1 - \Pr\left(\frac{Y-\mu}{0.1} \leq \frac{14.5-\mu}{0.1}\right) = \\
&= 1 - \Pr\left(\mathcal{N}(0,1) \leq \frac{14.5-\mu}{0.1}\right).
\end{aligned}$$

ovvero

$$\Pr(B) = 1 - \Phi\left(\frac{14.5-\mu}{0.1}\right).$$

Possiamo ottenere delle stime di queste due quantità sostituendo a μ, che è ignoto, la sua stima $\overline{y}$. Nel caso in esame

$$\widehat{\Pr(A)} = \Phi\left(\frac{13.5-14.302}{0.1}\right) = \Phi(-8.02) \approx 0$$

e

$$\widehat{\Pr(B)} = 1 - \Phi\left(\frac{14.5-14.302}{0.1}\right) = 1 - \Phi(1.98) \approx 0.024$$

ovvero, sulla base dei dati (e delle assunzioni fatte), stimiamo in 2.4% la probabilità di produrre una lastra troppo "alta" mentre valutiamo praticamente irrilevante la probabilità di produrre una lastra troppo sottile.

NOTA: In generale, se consideriamo una funzione del parametro θ, $\omega = g(\theta)$, e abbiamo una stima di θ, $\hat{\theta}$, allora una stima di ω è data da $\hat{\omega} = g(\hat{\theta})$.

Stima di qui, stima di là,..., ma se c'è una stima c'è un errore

• Abbiamo incontrato **due** quantit: una "vera" μ (il valore atteso) e una media **campionaria** $\overline{y}$; la prima la possiamo vedere come la media degli spessori di tutte le lastre che l'impianto potrebbe produrre se continuasse per un tempo infinito a produrre nelle condizioni attuali; la seconda è la media degli spessori delle 5 lastre effetivamente misurate. • Abbiamo incontrato **due** probabilità di produrre una lastra troppo "alta"; una che calcoleremmo se conoscessimo il "vero" valore atteso, l'altra che possiamo calcolare (e difatti abbiamo calcolato) utilizzando $\overline{y}$. • Ovvero abbiamo incontrato delle "vere" quantità (che hanno a che fare con la "vera" distribuzione di probabilità che ha generato i dati) e delle stime delle "vere" quantità. Ma se $\overline{y}$ è solo una "stima", ovvero una approssimazione, della "vera" media allora è spontaneo chiedere "quanto è buona?" ovvero "quanto è grande l'errore che commettiamo?"

Esercizio 15. *Si osservi che abbiamo sempre scritto vera tra virgolette. Lo studente ripensi a quanto detto nel lucido 163 e spieghi perché.*

Stima e Stimatore

- Definiamo **stima** di un parametro un **valore** (numerico) ottenuto dai dati osservati del campione $(y_1, \ldots, y_n)$.

  Nell'esempio precedente
  $$\overline{y} = \frac{14.33 + \cdots + 14.17}{5} = 14.302.$$

- La media campionaria, $\overline{y}$, può essere vista come una determinazione di una variabile casuale. Infatti se i dati da cui è calcolata sono il risultato di un esperimento casuale anche $\overline{y}$ ovviamente lo è. Definiamo **stimatore** di un parametro una **variabile casuale** funzione delle variabili casuali $Y_1, \ldots, Y_n$ che generano i valori osservati del campione $y_1, \ldots, y_n$.

  Nel nostro caso
  $$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n}$$
  è una variabile casuale.

- Giudicheremo la bontà della nostra stima valutando le proprietà dello stimatore (in realtà vale anche per le altre procedure inferenziali che vedremo più avanti...). In particolare, faremo riferimento al cosiddetto **principio del campionamento ripetuto**, secondo il quale se noi effettivamente replicassimo l'estrazione del campione per molte e molte volte, e per ogni campione calcolassimo ad esempio la media aritmetica, allora effettivamente ogni specifico valore ottenuto rappresenterebbe una determinazione della v.c. $\overline{Y}$. Nella grande maggioranza delle situazioni reali questa replica dell'esperimento non ha luogo, ma noi ragioniamo *come se* si verificasse.

- **Notazione:** In generale, se θ è il parametro di interesse, indicheremo con $\hat{\theta}$ sia la stima $\hat{\theta}(y_1, \ldots, y_n)$, sia lo stimatore $\hat{\theta}(Y_1, \ldots, Y_n)$ del parametro θ. Sar il contesto a rendere chiaro a quale quantit ci riferiamo.

- **Nota generale:** In generale, chiameremo **statistica** una qualsiasi funzione dei dati. Quindi, ad esempio, uno stimatore sarà una statistica (e la stima sarà il valore che tale statistica assume in un particolare campione osservato).

La distribuzione della media campionaria

Nelle ipotesi che stiamo facendo (normalità,...) la distribuzione di $\overline{Y}$ discende dal seguente risultato:

Proposizione 23. *Se $Y_1, \ldots, Y_n$ sono variabili casuali normali* indipendenti *tra loro e se $a_0, \ldots, a_n$ sono delle costanti reali qualsiasi, allora*

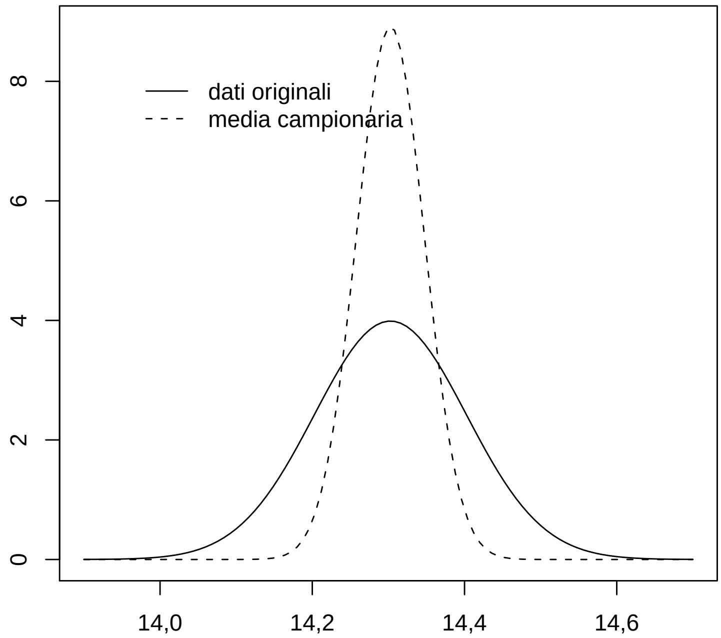
$$a_0 + \sum_{i=1}^{n} a_i Y_i \sim \mathcal{N}(a_0 + \sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$$

dove μ_i e σ_i^2 indicano rispettivamente il valore atteso e la varianza di Y_i.

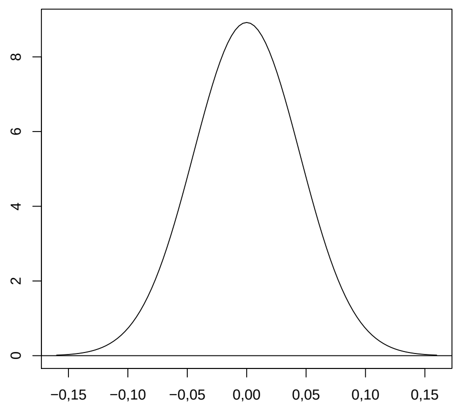
Quindi, se le n variabili casuali normali hanno tutte lo stesso valore atteso e varianza (diciamo μ e σ^2) allora (lo studente lo dimostri)

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) .$$

Si osservi che la distribuzione e la media sono quelle delle osservazioni originarie (ovvero le Y_i sono normali e $\overline{Y}$ è normale, le Y_i hanno valore atteso μ e $\overline{Y}$ ha valore atteso μ) e che la varianza della media campionaria è la varianza delle osservazioni originarie divisa per n (ovvero se il numero delle osservazioni è maggiore di 1 allora la media campionaria è meno variabile delle osservazioni originarie). Il grafico mostra le due funzioni di densità nel caso considerato nel nostro esempio, in cui $\mu = 14.3$, $\sigma = 0.1$ e $n = 5$.



La distribuzione dell'errore di stima



Il risultato precedente ci permette di calcolare anche la distribuzione dell'**errore di stima**, ovvero di $\overline{Y} - \mu$, che risulta (lo studente lo dimostri)

$$\overline{Y} - \mu \sim \mathcal{N}(0, \sigma^2/n).$$

Si noti che nel caso in esame, poiché σ^2 è noto, la distribuzione dell'errore di stima risulta anch'essa nota (è una normale di valore atteso 0 e varianza $0.01/5 = 0.002$). Inoltre il valore atteso dell'errore di stima è zero (non si commettono errori sistematici di stima).

Stimatore o stimatori?

Riprendiamo l'esempio precedente. L'idea di ricorrere alla media aritmetica un'idea piuttosto naturale ma consideriamo il caso di un operatore pigro che non vuole fare la media ma utilizza soltanto la prima misura y_1 realizzazione di Y_1. Per stimare μ egli impiega questa sola misura e quindi commette un errore di stima

$$Y_1 - \mu \sim \mathcal{N}(0, \sigma^2).$$

Alcune considerazioni:

- Esistono più modi di stimare un parametro (e conseguentemente più stimatori di un parametro).
- Per ognuno di questi stimatori si commette sempre un errore ma la distribuzione dell'errore diversa.
- Intuitivamente siamo portati a considerare $\overline{Y}$ come "migliore" dello stimatore Y_1 perché utilizza più informazione.

Proprietà degli stimatori

Indichiamo in generale con $\hat{\theta}$ uno stimatore per il parametro ignoto θ.

- Potendo scegliere, si vorrebbe che l'errore di stima $\hat{\theta}-\theta$ fosse nullo. Ma questa una richiesta insensata perch questo vorrebbe dire conoscere ci che si vuole stimare.

- Un requisito più ragionevole la **non distorsione**:

$$\mathbb{E}[(\hat{\theta}-\theta)] = \mathbb{E}(\hat{\theta}) - \theta = 0,$$

  ovvero il valore atteso dell'errore di stima nullo.

  Nell'esempio precedente $\overline{Y}$ e Y_1 sono due stimatori non distorti di μ. Invece $(Y_1+Y_2)/3$ non lo è, in quanto $\mathbb{E}[(Y_1+Y_2)/3] = 2\mu/3$

  Quest'ultimo esempio ci dice anche che la non distorsione può non essere valutabile, in quanto può dipendere da quantità non note (qui μ).

- La non distorsione ci dice se gli errori (in media) si compensano, ma nulla ci dice di quanto (in media) questi possono essere grandi. Una misura di ciò data dall'**errore quadratico medio**:

$$EQM(\hat{\theta}) = \mathbb{E}[(\hat{\theta}-\theta)^2].$$

  Ragionevolmente, preferiamo tra due stimatori quello con errore quadratico medio inferiore.

- l'errore quadratico medio può essere così decomposto

$$\begin{array}{rcl} EQM(\hat{\theta}) & = & \mathbb{E}[(\hat{\theta}-\theta)^2] \\ & = & \mathbb{E}[(\hat{\theta}-\mathbb{E}(\hat{\theta})+\mathbb{E}(\hat{\theta})-\theta)^2] \\ = \mathbb{E}[(\hat{\theta}-\mathbb{E}(\hat{\theta}))^2] & + & (\mathbb{E}(\hat{\theta})-\theta)^2 + 2\mathbb{E}[(\hat{\theta}-\mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta})-\theta)] \\ & = & \mathbb{V}ar(\hat{\theta}) + (\mathbb{E}(\hat{\theta})-\theta)^2 \\ & = & \text{varianza} + (\text{distorsione})^2, \end{array}$$

  essendo

$$\mathbb{E}[(\hat{\theta}-\mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta})-\theta)] = (\mathbb{E}(\hat{\theta})-\theta)\mathbb{E}[\hat{\theta}-\mathbb{E}(\hat{\theta})] = 0.$$

- Nell'esempio:

$$
\begin{aligned}
EQM(\overline{Y}) &= \mathbb{E}[(\overline{Y}-\mu)^2] = \sigma^2/n \\
EQM(Y_1) &= \mathbb{E}[(Y_1-\mu)^2] = \sigma^2 \\
EQM((Y_1+Y_2)/3) &= \mathbb{E}[((Y_1+Y_2)/3-\mu)^2] \\
&= \mathbb{E}[((Y_1+Y_2-3\mu)^2/9] \\
&= \mathbb{E}[((Y_1-\mu)+(Y_2-\mu)-\mu)^2/9] \\
&= (\mathbb{V}ar(Y_1)+\mathbb{V}ar(Y_2))/9+\mu^2/9 \\
&= 2\sigma^2/9+\mu^2/9
\end{aligned}
$$

  Alcune considerazioni:

  1. L'errore quadratico medio pu dipendere da quantità non note.
  2. Secondo questo criterio $\overline{Y}$ è preferibile a Y_1.
  3. Non sempre possiamo dire se uno stimatore preferibile ad un altro.

- Quando il numero di osservazioni n nel nostro campione aumenta, ci attendiamo che lo stimatore $\hat{\theta}_n$ si avvicini, ovvero tenda al parametro θ che intendiamo stimare. Abbiamo utilizzato qui la notazione $\hat{\theta}_n$ in luogo di $\hat{\theta}$ per sottolineare che lo stimatore dipende anche dal numero di osservazioni. Ma cosa vuol dire tendere?

  Ricordiamo che $\hat{\theta}_n$, in quanto funzione di $Y_1, \ldots, Y_n$, è una v.c.. Possiamo dunque dire che

  1. uno stimatore $\hat{\theta}$ **consistente in media quadratica** se
  $$\hat{\theta}_n \xrightarrow{m.q.} \theta,$$
  ovvero
  $$EQM(\hat{\theta}) = \mathbb{E}[(\hat{\theta}-\theta)^2] \longrightarrow 0.$$
  2. uno stimatore $\hat{\theta}$ **consistente in senso debole** se
  $$\hat{\theta}_n \xrightarrow{P} \theta.$$

Una condizione <u>sufficiente</u> per la consistenza debole

Sappiamo che la convergenza in media quadratica implica la convergenza in probabilità (Proposizione [14]). Perciò è possibile dimostrare la consistenza debole di uno stimatore $\hat{\theta}_n$ verificando la seguente condizione sufficiente:

$$\lim_{n\to\infty} EQM(\hat{\theta}_n) = \lim_{n\to\infty} \mathbb{E}\{(\hat{\theta}_n-\theta)^2\} = 0 \quad \Rightarrow \quad \hat{\theta}_n \xrightarrow{P} \theta\,.$$

Abbiamo visto che l'errore quadratico medio di uno stimatore è la somma della varianza dello stimatore e della sua distorsione al quadrato. Essendo i due addendi non negativi, la condizione appena vista può essere riformulata come:

$$\begin{cases} \lim\limits_{n\to\infty} \mathbb{E}(\hat{\theta}_n) = \theta \\ \lim\limits_{n\to\infty} \mathbb{V}ar(\hat{\theta}_n) = 0 \end{cases} \quad \Rightarrow \quad \hat{\theta}_n \xrightarrow{P} \theta$$

In pratica, se lo stimatore è (almeno quando $n \to \infty$) non distorto e la sua varianza diventa sempre più piccola (e cioè tende a 0), allora possiamo dire che lo stimatore è consistente in media quadratica e dunque anche in senso debole.

Nel nostro esempio, lo stimatore $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ è tale che $\mathbb{E}(\overline{Y}_n) = \mu$ e $\mathbb{V}ar(\overline{Y}_n) = 0.01/n$. Quindi, essendo non distorto (per qualsiasi valore di n) lo sarà anche al limite. Inoltre, è immediato vedere che $\lim\limits_{n\to\infty} \mathbb{V}ar(\overline{Y}_n) = 0$ e quindi lo stimatore è consistente in senso debole.

Invece lo stimatore Y_1 è tale che $\mathbb{E}(Y_1) = \mu$ e $\mathbb{V}ar(Y_1) = 0.01$. Quindi, pur essendo non distorto, la sua varianza non va a 0 al crescere di n e quindi non possiamo dire che lo stimatore sia consistente.

Esercizio 16. *Utilizzando la condizione sufficiente appena vista, verificate che lo stimatore* $\frac{1}{n-1}\sum_{i=1}^n Y_i$*, pur essendo distorto, è consistente in senso debole.*

Metodo dei momenti

Esistono vari metodi per 'costruire stimatori'. Nel nostro esempio, volendo stimare $\theta = \mathbb{E}(Y)$ abbiamo utilizzato $\sum_{i=1}^n Y_i/n$. Un tale stimatore ricavato secondo il **metodo dei momenti**.

In generale, poiché la distribuzione di Y dipende da θ, si ha $\mathbb{E}[g(Y)] = a(\theta)$. Lo stimatore dei momenti $\hat{\theta}$ si ottiene risolvendo rispetto a θ l'equazione (detta dei momenti)

$$a(\theta) = \sum_{i=1}^n \frac{g(Y_i)}{n}, \qquad \left[\theta = \sum_{i=1}^n \frac{Y_i}{n} \Rightarrow \hat{\theta} = \sum_{i=1}^n \frac{Y_i}{n}\right]$$

Esercizio 17. *Vogliamo stimare* $\theta = \mathbb{V}ar(Y)$*, ossia la varianza di una v.c.* Y*. Supponendo che la media di* Y *sia nulla (cioè che* $\mathbb{E}(Y) = 0$*), quale sarà lo stimatore di* θ *secondo il metodo dei momenti?*

Un intervallo di confidenza

Poichè la distribuzione dell'errore di stima è completamente nota ($\mathcal{N}(0, 0.01/5)$) possiamo "costruire" delle affermazioni del tipo: *"la probabilità che l'errore di stima sia in valore assoluto minore di* 0.1 *è uguale a* 0.975*"*. Infatti,

$$\begin{aligned}
\Pr(|\overline{Y} - \mu| < 0.1) &= \Pr(|\mathcal{N}(0, 0.002)| < 0.1) \\
&= \Pr\left(\left|\frac{\mathcal{N}(0, 0.002)}{\sqrt{0,002}}\right| < \frac{0.1}{\sqrt{0.002}}\right) \\
&= \Pr(|\mathcal{N}(0,1)| < 2.236) \\
&= \Phi(2.236) - \Phi(-2.236) \\
&= \Phi(2.236) - (1 - \Phi(2.236)) \\
&= 2 \cdot \Phi(2.236) - 1 \approx 0.975.
\end{aligned}$$

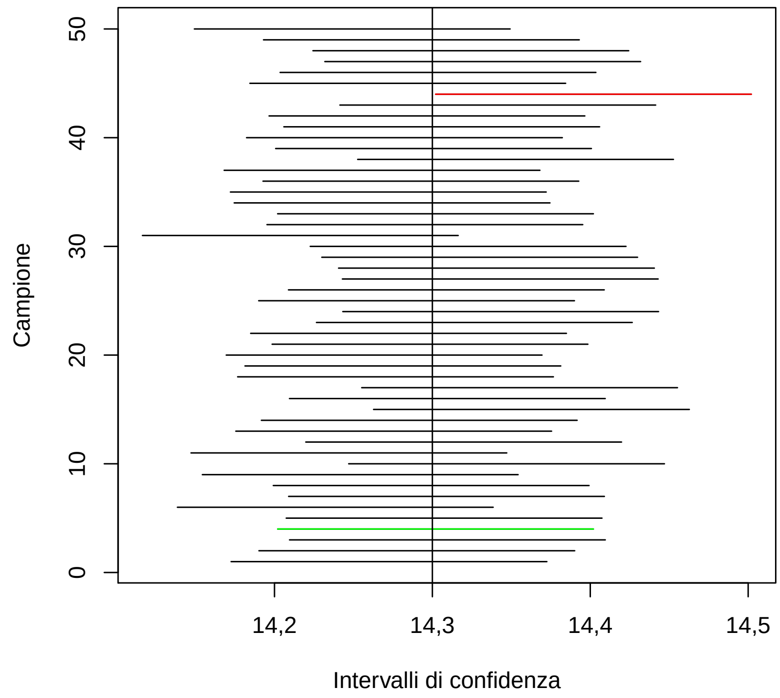
Si osservi ora che l'affermazione precedente può essere anche scritta come *"la probabilità che l'intervallo* $[\overline{Y} - 0.1, \overline{Y} + 0.1]$ *includa il vero valore atteso* μ *è* 0.975*"*. Infatti,

$$\begin{aligned}
\Pr(|\overline{Y} - \mu| < 0.1) &= \Pr(-0.1 < \mu - \overline{Y} < 0.1) \\
&= \Pr(\overline{Y} - 0.1 < \mu < \overline{Y} + 0.1).
\end{aligned}$$

In generale un intervallo che contiene il vero valore di un parametro ignoto con probabilità $1-\alpha$ viene chiamato un **intervallo di confidenza di livello** $1 - \alpha$.

L'intervallo $[\overline{Y} - 0.1, \overline{Y} + 0.1]$ è un intervallo casuale. Con i dati a nostra disposizione, $[\overline{y} - 0.1, \overline{y} + 0.1] = [14.202\ ,\ 14.402]$.

Se potessimo ripetere un numero infinito di volte l'esperimento casuale, intervalli di questo tipo conterrebbero il vero valore μ nel 97.5% dei casi.



Gli intervalli di confidenza costituiscono forse la maniera più semplice di comunicare la precisione (od imprecisione) di una stima. Si confrontino ad esempio le due affermazioni:

1. La stima della media è 14.302; la distribuzione dell'errore di stima è una normale con media nulla e varianza 0.002.

2. Con una "confidenza" molto alta, per la precisione 0.975, il "vero" valore di μ è compreso tra 14.202 e 14.402.

La prima affermazione è più generale ma la sua "decodifica" richiede nozioni non note a tutti (quale strana bestia è una distribuzione normale? E la varianza?). La seconda è molto più facile da interpretare.

Intervalli di confidenza di livello prefissato

Quasi sempre si calcolano intervalli di confidenza con un livello fissato a priori (le scelte più comuni sono 0.9 , 0.95 e 0.99).

In questo caso i passi da seguire sono i seguenti:

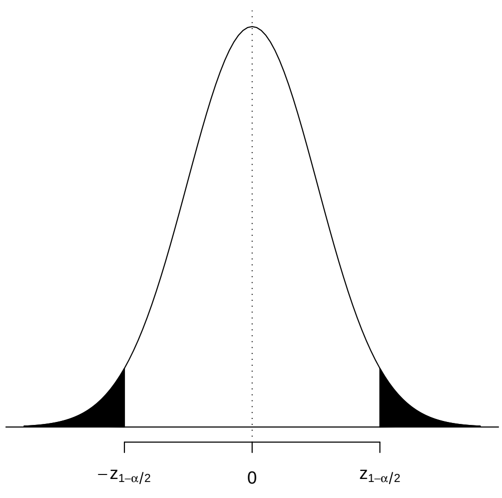
1. Fissiamo un valore per $1-\alpha$.

2. Determiniamo, utilizzando un programma o le tavole della normale standard, il percentile $1-\alpha/2$ di una normale standard, ovvero un punto $z_{1-\alpha/2}$ tale che $\Pr(\mathcal{N}(0,1) \leq z_{1-\alpha/2}) = 1-\alpha/2$. Per la simmetria della densità di una normale intorno alla sua media allora $\Pr(\mathcal{N}(0,1) \leq -z_{1-\alpha/2}) = \alpha/2$. E quindi $\Pr(|\mathcal{N}(0,1)| \leq z_{1-\alpha/2}) = 1-\alpha$. Si veda il grafico a pagina 175.

3. Ricordando che $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$, possiamo allora scrivere
$$\Pr\left(\left|(\overline{Y}-\mu)/\sqrt{\sigma^2/n}\right| \leq z_{1-\alpha/2}\right) = 1-\alpha.$$

4. Con semplici passaggi otteniamo
$$\Pr\left(\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

L'intervallo
$$\left[\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$
è quindi un intervallo di confidenza di livello $1-\alpha$ per μ.



Ambedue le aree "annerite" sono uguali ad $\alpha/2$. Quindi l'area "non annerita" è uguale a $1-\alpha$.

Esempio 37. *Supponiamo di volere un intervallo di confidenza di livello* 0.95*. Allora,* $\alpha = 0.05$ *e* $1-\alpha/2 = 0.975$*. Dalle tavole della distribuzione normale standard troviamo* $z_{0.975} = 1.96$*. Quindi l'intervallo risulta essere uguale a*

$$14.302 \pm 1.96 \frac{0.1}{\sqrt{5}} = [14.21, 14.39] \,.$$

Notiamo che se il livello di confidenza fosse $1-\alpha = 0.99$*, dovremmo utilizzare il quantile* $z_{0.995} = 2.575$*. L'intervallo sarebbe quindi pari a*

$$14.302 \pm 2.575 \frac{0.1}{\sqrt{5}} = [14.19, 14.42] \,.$$

Viceversa, se il livello di confidenza fosse $1-\alpha = 0.90$*, il quantile necessario sarebbe* $z_{0.95} = 1.645$*, con risultante intervallo di confidenza*

$$14.302 \pm 1.645 \frac{0.1}{\sqrt{5}} = [14.23, 14.38] \,.$$

Esercizio 18. *Commentate la relazione tra l'ampiezza dell'intervallo e il livello di confidenza. Cosa accade se* $1-\alpha = 0$*?*

Esercizio 19. *Per un fissato livello di confidenza* $1-\alpha$*, dite cosa accade all'ampiezza dell'intervallo all'aumentare della numerosità del campione* n*.*

Precisione nella stima della difettosità

Ricordiamoci che abbiamo ottenuto la formula

$$\Pr(\{\text{lastra troppo ``alta''}\}) = \pi(\mu) = 1 - \Phi\left(\frac{14.5-\mu}{0.1}\right)$$

dove con l'introduzione della nuova notazione $\pi(\mu)$ vogliamo enfatizzare il fatto che la probabilità di produrre una lastra troppo "alta" dipende dalla media μ.

E' facile verificare che $\pi(\mu)$ è una funzione monotona crescente (ci si ricordi che $\Phi(y)$ è crescente in y). Quindi, l'evento

$$\left\{\overline{y} : \pi\left(\overline{y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right) \le \pi(\mu) \le \pi\left(\overline{y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)\right\}$$

coincide con l'evento

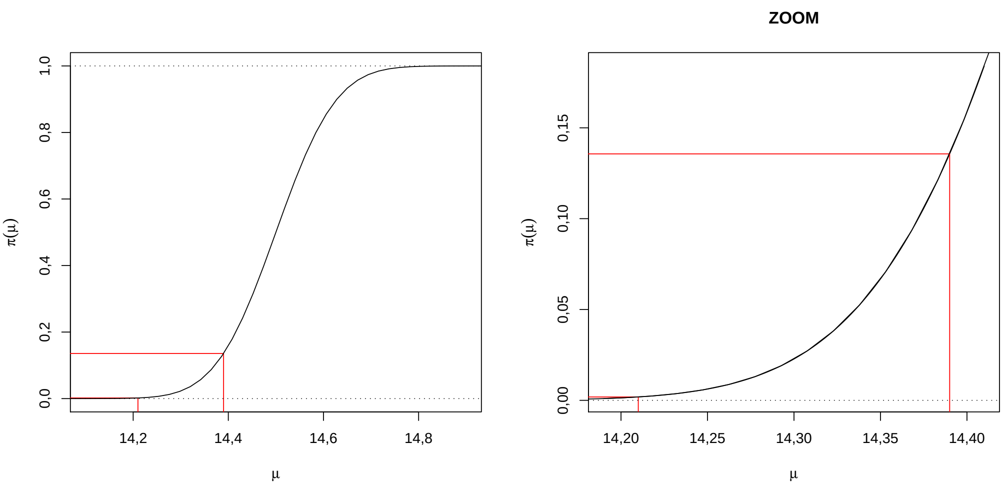
$$\left\{\overline{y} : \overline{y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \le \mu \le \overline{y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right\} .$$

Ma allora i due eventi hanno la stessa probabilità e quindi

$$\left[\pi\left(\overline{y}-\frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right), \pi\left(\overline{y}+\frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)\right]$$

è un intervallo di confidenza di livello $1-\alpha$ per $\pi(\mu)$. Si osservi che ci basta trasformare gli estremi di un intervallo di confidenza per μ.

Usando $\alpha = 0.05$, l'intervallo che otteniamo è $[0.002\ ,\ 0.135]$. Ovvero se teniamo conto dell'errore di stima i risultati ottenuti (gli spessori misurati) sono compatibili, anche senza fare riferimento ad eventi con probabilità particolarmente piccola, con una difettosità superiore al 10%. La conclusione è che sembra "prudente" bloccare la produzione, tenendo conto tra l'altro, che $\pi(14) \approx 2/10^6$, ovvero, che l'impianto, quando ben "tarato", può produrre un numero di lastre difettose realmente piccolo.



Un approccio diverso: la verifica di ipotesi

Fino ad ora ci siamo occupati di capire che cosa i dati ci potevano raccontare (e con quale affidabilità) sulla "vera" media e sulle "vere" probabilità di produrre lastre difettose. L'idea era di bloccare la produzione e ritarare le macchine quando i dati indicano che la "difettosità" dell'impianto è eccessiva.

Potremmo però anche ragionare lungo le seguenti linee:

- ad ogni manutenzione (ordinaria o straordinaria) l'impianto viene "tarato" in maniera tale che la media degli spessori prodotti risulti $14mm$;

- quindi un valore di μ diverso, anche di poco, da $14mm$ indica una qualche "sregolazione in corso";

- per questo motivo possiamo pensare di bloccare l'impianto appena i dati suggeriscono che la media è cambiata.

Uno dei possibili vantaggi di questo approccio è che potremmo riuscire a bloccare la produzione quando la "sregolazione" è iniziata ma la probabilità di produrre lastre difettose è ancora piccola.

Una maniera diversa di descrivere l'approccio appena suggerito consiste nel dire che all'inizio di ogni turno vogliamo utilizzare i dati per decidere tra le seguenti due ipotesi:

$$\mathrm{H}_0 : \mu = 14mm$$

e

$$\mathrm{H}_1 : \mu \neq 14mm.$$

L'interpretazione delle due ipotesi è (ovviamente):

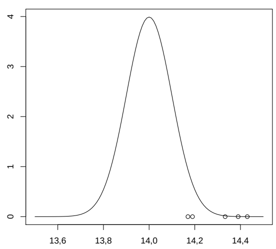
$$\mathrm{H}_0 : \text{l'impianto produce al meglio}$$

e

$$\mathrm{H}_1 : \text{l'impianto ha iniziato a "sregolarsi"}.$$

Problemi di scelta tra due (o più) alternative sono, in statistica, chiamati problemi di **verifica di ipotesi**. Le ipotesi (quando sono due) vengono spesso indicate come **ipotesi nulla** ed **ipotesi alternativa**. Lo "strumento" utilizzato per affrontare i problemi di verifica di ipotesi (ovvero la procedura che si segue per far "votare" i dati a favore o di H_0 o di H_1, ovvero per decidere quale ipotesi **accettare** o **rifiutare**) viene chiamato **test statistico**.

Analisi grafica



La figura mostra la densità con una normale di media 14 e varianza 0.01 (ovvero la distribuzione ipotizzata da H_0) con i dati osservati "marcati" sull'asse delle x. Sembra improbabile che i dati siano stati generati dalla distribuzione disegnata: sono troppo spostati a destra, anche in regioni a cui la distribuzione ipotizzata da

H_0 assegna probabilità quasi nulla. D'altra parte H_1 "prevede" alcune distribuzioni (ad es. si veda il grafico a pagina 164) che sembrano "più compatibili" con i dati. Quindi, i dati suggeriscono di rifiutare H_0. Sfortunatamente, una analisi grafica del tipo descritto è possibile solo nelle situazioni più semplici.

Un test statistico

Volendo definire una procedura "analitica" per scegliere tra le due ipotesi, sembra ragionevole basarsi sulla differenza tra la media stimata, $\overline{y}$, e la media ipotizzata da H_0, 14.

Ad esempio, potremmo pensare di usare una "regola" del tipo

$$-h \leq \frac{\sqrt{n}(\overline{y}-14)}{\sigma} \leq h$$

si → accettiamo H_0 no → rifiutiamo H_0

Si osservi che abbiamo diviso la differenza per lo scarto quadratico medio della media campionaria. Ovviamente, trattandosi nel nostro caso di una costante nota ($n = 5$ e $\sigma^2 = 0.01$) ciò non cambia l'interpretazione della "regola".

Per rendere operativa la "regola" dobbiamo decidere quale valore assegnare alla soglia h.

La soglia h sarà determinata utilizzando la distribuzione della variabile casuale, detta **test**,

$$T = \frac{\sqrt{n}(\overline{Y}-14)}{\sigma},$$

di cui il valore nel campione osservato, $t_{oss} = \sqrt{n}(\overline{y}-14)/\sigma$, è una determinazione.

Se H_0 è vera...

...vorremmo, ovviamente, rifiutare H_1. In altre parole non ci dispiacerebbe che

$$\Pr(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera}) = 1 \tag{3.2}$$

ovvero, che

$$\Pr(-h \leq \sqrt{n}(\overline{Y}-14)/\sigma \leq h \text{ quando } \mu = 14) = 1. \tag{3.3}$$

Ora, se $\mu = 14$, $T = \sqrt{n}(\overline{Y}-14)/\sigma$ ha distribuzione normale standard (lo studente spieghi perché). Quindi, la (3.3) è equivalente a

$$\Pr(-h \leq \mathcal{N}(0,1) \leq h) = 1 \tag{3.4}$$

e la (3.4) mostra che l'unico valore di h che garantisce la (3.2) è $h = +\infty$ (ci si ricordi che la densità di una normale è diversa da zero su tutta la retta reale).

L'utilizzo di una soglia infinita non è però molto sensato. Infatti se poniamo $h = +\infty$ non rifiutiamo mai H_0. In altre parole, se insistiamo sulla (3.2) finiamo con una "regola" per cui

$$\Pr(\text{accettare } \mathrm{H}_0 \text{ quando } \mathrm{H}_1 \text{ è vera}) = 1.$$

Un compromesso: test di livello fissato α

Chiedere che la (3.2) sia *esattamente* vera ci porta a determinare un valore di h inaccettabile. Sarebbe però inaccettabile anche una situazione in cui, ad esempio,

$$\Pr(\text{accettare } \mathrm{H}_0 \text{ quando } \mathrm{H}_0 \text{ è vera}) = 0.1$$

ovvero, una situazione in cui la (3.2) è pesantemente violata. Infatti, in questo caso, il test *sbaglierebbe* 9 volte su 10 quando l'ipotesi nulla è vera. E anche questo sembra poco sensato.

Non ci rimane quindi che considerare il caso in cui la (3.2) è approssimativamente (ma non esattamente) rispettata, ovvero

$$\Pr(\text{accettare } \mathrm{H}_0 \text{ quando } \mathrm{H}_0 \text{ è vera}) = 1 - \alpha\,, \tag{3.5}$$

per un valore "piccolo" di α, detto **livello di significatività** del test. La (3.5) può essere riscritta nella forma

$$\Pr(-h \leq \mathcal{N}(0,1) \leq h) = 1 - \alpha \tag{3.6}$$

ed è facile verificare (lo studente si aiuti con il grafico a pagina 175) che la soluzione in h della (3.6) è $h = z_{1-\alpha/2}$ dove con z_p abbiamo indicato il percentile p-simo di una normale con media zero e varianza uno, ovvero il numero per cui $\Phi(z_p) = p$.

Sintesi della procedura delineata

In definitiva, per verificare un sistema d'ipotesi del tipo

$$\begin{cases} \mathrm{H}_0 : \mu = \mu_0 \\ \mathrm{H}_1 : \mu \neq \mu_0 \end{cases}$$

siamo arrivati alla seguente procedura:

scegliere il livello di significatività α

↓

determinare $z_{1-\alpha/2}$

↓

calcolare $t_{oss} = \dfrac{\sqrt{n}(\overline{y} - \mu_0)}{\sigma}$

↓

verificare se

$-z_{1-\alpha/2} \leq t_{oss} \leq z_{1-\alpha/2}$

se la risposta è si → accettare H_0

se la risposta è no → rifiutare H_0

Nel caso in esame

$$\alpha = 0.01 \text{ (ad esempio)}$$

↓

$$z_{1-\alpha/2} = z_{0.995} = 2.58$$

↓

$$t_{oss} = \frac{\sqrt{5}(14.302 - 14)}{0.1} = 6.75$$

↓

$$-2.58 \leq 6.75 \leq 2.58\ ?$$

↓ no

rifiutiamo H_0

Struttura di un test

Quanto abbiamo fatto nel caso in esame illustra fedelmente la struttura di un test statistico. E' quindi conveniente "ricapitolare" la costruzione:

1. Abbiamo definito una **statistica**, ovvero una funzione dei dati, scelta in maniera tale che i valori che ci aspettiamo che la statistica assuma quando H_0 e H_1 sono vere siano "tendenzialmente" diversi. Nell'ambito della teoria dei test, la statistica scelta viene chiamata, guarda caso, **statistica test**. Nell'esempio considerato, la statistica utilizzata è

$$T(Y_1, \ldots, Y_5) = \frac{\sqrt{n}(\overline{Y} - \mu_0)}{\sigma}$$

e l'abbiamo scelta poichè ci aspettiamo che

ipotesi "vera"	valori assunti dalla statistica test
H_0	intorno allo zero
H_1	lontani dallo zero

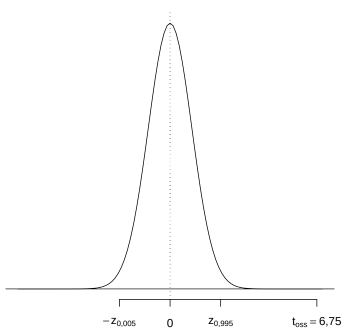
2. L'idea euristica di "la statistica test assume differenti valori sotto H_0 e H_1" si manifesta e concretizza da un punto di vista formale nell'osservare che T ha una diversa distribuzione di probabilità nei due casi. Ad esempio, nel caso in esame, se μ è la vera media degli spessori allora (lo studente lo dimostri utilizzando i risultati di pagina 167)

$$T \sim \mathcal{N}(\sqrt{n}(\mu - \mu_0)/\sigma, 1)$$

ovvero, solo sotto H_0, $T \sim \mathcal{N}(0,1)$. Invece, quando è vera H_1 la distribuzione è spostata o verso destra o verso sinistra (a seconda del segno di $\mu - \mu_0$).

3. A questo punto per decidere se H_0 debba essere accettata o meno possiamo confrontare il valore osservato della statistica, ovvero il valore di T calcolato dai dati, con la distribuzione sotto H_0 (si veda lucido seguente). Poiché nell'esempio il valore osservato della statistica è "troppo estremo" (ovvero, troppo poco probabile) abbiamo deciso di rifiutare H_0. In particolare si osservi che, desiderando una regola precisa, nella procedura operativa descritta dall'albero a pagina 188 abbiamo convenuto che "troppo estremo" significa $|T| > z_{1-\alpha/2}$ per qualche fissato (e non troppo grande) valore di α.

Distribuzione sotto H_0 e valore osservato della statistica test



Il valore osservato (6.75) non sembra essere stato generato dalla distribuzione disegnata. Quindi rifiutiamo H_0.

Si noti la somiglianza col grafico fatto a pagina 178. La differenza è che qui usiamo la statistica test e non le osservazioni direttamente per trarre le nostre conclusioni.

Esistono due tipi di errore

Si osservi che in un problema di verifica d'ipotesi esistono due possibili modi in cui possiamo sbagliare. Infatti può capitare di:

1. rifiutare H_0 quando H_0 è vera; questo è usualmente chiamato un **errore di primo tipo**.

2. accettare H_0 quando H_0 è falsa; questo è usualmente chiamato un **errore di secondo tipo**.

	H_0	H_1
Accettare H_0	OK	errore 2° tipo
Rifiutare H_0	errore 1° tipo	OK

Ovviamente

$$\Pr(\text{errore } 1^\circ \text{ tipo}) = 1 - \Pr\left(\begin{array}{c}\text{accettare } \mathrm{H}_0 \text{ quando} \\ \mathrm{H}_0 \text{ è vera}\end{array}\right) = \alpha.$$

Quindi, costruire un test che soddisfa la (3.5) equivale ad utilizzare un test in cui la probabilità di commettere un errore di 1° tipo sia α.

Si noti viceversa che nella costruzione delineata fino a questo punto la probabilità di commettere un errore di 2° tipo non è stata esplicitamente considerata (con la sola eccezione di pagina 179 il cui contenuto può essere parafrasato come "se vogliamo un test in cui la probabilità di errore di primo tipo sia nulla finiamo per costruire un test in cui la probabilità di errore di secondo tipo è uno".)

Il motivo per cui ci si preoccupa di più degli errori di 1° tipo è che spesso la domanda a cui si vuole rispondere con un test statistico è

A. *S*ono i dati sperimentali compatibili con H_0?

più che

B. *Q*uale tra H_0 e H_1 è vera?

Ovviamente esistono dei casi in cui B è la vera domanda. Diventa allora necessario considerare simultaneamente i due tipi di errore. Questo, all'interno della procedura delineata, può essere fatto scegliendo in maniera appropriata α e soprattutto, quando possibile, la numerosità compionaria (n). Infatti più n è grande più possiamo sperare di rendere piccoli ambedue i tipi di errore. Lasciamo a corsi più avanzati il mostrare come.

Ci limitiamo a menzionare che nel caso in esame il valore di n usato (ovvero 5) era stato scelto dall'impresa proprio sulla base di considerazioni di questo tipo.

Il livello di significatività osservato

Abbiamo visto che il test con livello di significatività fissato (α) si conclude con l'accettazione o il rifiuto dell'ipotesi nulla H_0, a seconda che il valore della statistica test appartenga o meno alla regione di accettazione (determinata in base al valore di α).

Un modo alternativo per verificare quanto i dati siano in accordo/disaccordo con l'ipotesi nulla H_0 passa attraverso il **livello di significatività osservato**:

$$\alpha_{oss} = \Pr(|T| > |t_{oss}|, \text{ quando } \mathrm{H}_0 \text{ è vera})$$

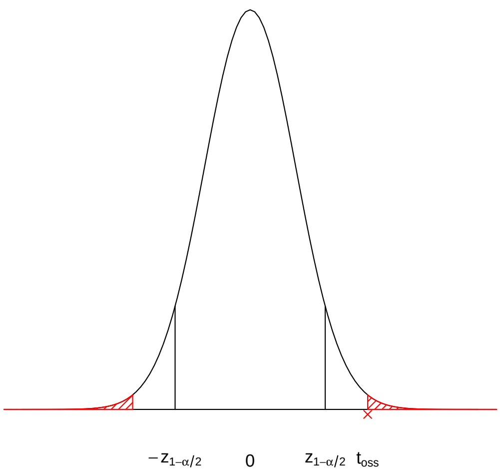
che è la probabilità, se H_0 è vera, di osservare un valore del test più lontano da H_0 del valore che abbiamo ottenuto nel campione osservato (ricordiamoci che T assume valori vicino allo zero se H_0 è vera e lontani dallo zero altrimenti...).

È chiaro quindi che se α_{oss} è piccolo vuol dire che i dati sono in disaccordo con l'ipotesi nulla. Questo perché è improbabile ottenere un valore del test ancora più lontano da H_0 del valore ottenuto nel campione osservato. Viceversa se α_{oss} non è piccolo, significa che i dati non sono in disaccordo con l'ipotesi nulla.

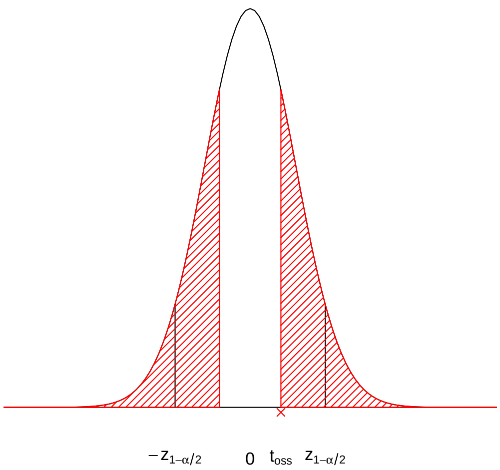
Nel nostro esempio, $t_{oss} = 6.75$ e $\alpha_{oss} \approx 0$ (calcolandolo, con un computer, sarebbe circa uguale a 0.0000000000147), confermando che i dati mostrano una forte evidenza contro H_0.

Relazione tra α_{oss} e α fissato

$\alpha_{oss} < \alpha \iff$ rifiuto H_0 con livello fissato α



$$\alpha_{oss} > \alpha \iff \text{accetto } \mathrm{H}_0 \text{ con livello fissato } \alpha$$



α_{oss} indica anche "quanto" siamo lontani/vicini da H_0.

Ancora sul livello di significatività osservato

La varietà del pur limitato insieme di test che abbiamo presentato dovrebbe aver chiarito l'utilità del livello di significatività osservato. Il suo merito principale consiste nel nascondere i dettagli dei vari test e nel, viceversa, presentare i risultati utilizzando una "scala" sempre uguale. Conoscendo il livello di significatività osservato non abbiamo bisogno di sapere, per trarre delle conclusioni, se sotto l'ipotesi nulla la statistica test si distribuisce come una normale, o come una t di Student (chi mai sarà?) o come ... Non abbiamo neanche bisogno di conoscere il valore della statistica test.

Sempre a proposito del livello di significatività osservato, ricordiamo che comunemente se è inferiore a 0,01 i risultati sono considerati *altamente significativi* contro H_0 mentre se risulta compreso tra 0,01 e 0,05 si parla di risultati *significativi*, sempre contro l'ipotesi nulla. Viceversa se risulta maggiore di 0,1 si conclude che i dati non contengono elementi tali da poter rifiutare H_0 e quindi si parla di *non significatività*. I valori che mancano, ovvero quelli compresi tra 0,05 e 0,1 sono i più difficili da interpretare. Siamo in una situazione di sostanziale indecisione, a volta indicata come risultato ai *margini della significatività* o *borderline*. Ovviamente, questi valori (0,01, 0,05 e 0,1) non hanno niente di *sacro*.

Intervalli di confidenza e test

Esiste una relazione tra un intervallo di confidenza per μ di livello $1-\alpha$ e il test di livello di significatività fissato α.

Infatti, l'intervallo di confidenza l'abbiamo costruito in modo tale che

$$\Pr\left(\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha\,,$$

ossia

$$\Pr\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\overline{Y}-\mu)}{\sigma} \leq z_{1-\alpha/2}\right) = 1-\alpha\,.$$

Se $\mathrm{H}_0 : \mu = \mu_0$ è vera, abbiamo che l'ultima equazione diventa

$$\Pr\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\overline{Y}-\mu_0)}{\sigma} \leq z_{1-\alpha/2}\right) = 1-\alpha\,.$$

che è la probabilità di stare nella regione di accettazione quando H_0 è vera.

Quindi, una volta osservato il campione, la condizione

$$-z_{1-\alpha/2} \leq t_{oss} \leq z_{1-\alpha/2}\,,$$

che altro non è che la condizione di accettazione, è equivalente alla condizione

$$\overline{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \overline{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

ossia a verificare se il valore di μ sotto l'ipotesi nulla (μ_0) è contenuto nell'intervallo di confidenza di livello $1-\alpha$.

Nel nostro esempio, l'intervallo di confidenza per μ di livello $1-\alpha = 0.99$ era pari a $[14.19, 14.42]$: $\mu_0 = 14$ non appartiene a questo intervallo e quindi rifiutiamo H_0 ad un livello di significatività fissato $\alpha = 0.01$.

Ipotesi unilaterali

In alcune situazioni si è interessati a verificare, sulla base dei dati, se il parametro μ eccede (ad esempio) una certa soglia. Il sistema di ipotesi indicato in questo caso è

$$\mathrm{H}_0 : \mu = \mu_0$$

contro l'alternativa

$$\mathrm{H}_1 : \mu > \mu_0.$$

H_1 viene chiamata allora un'**ipotesi alternativa unilaterale destra**.

Il sistema di ipotesi considerato è equivalente a:

$$\mathrm{H}_0 : \mu \leq \mu_0$$

e

$$\mathrm{H}_1 : \mu > \mu_0.$$

Esempio 38. *Supponiamo di voler verificare se l'impianto produttore di lastre è attualmente tarato su spessori superiori a $14mm$. Il sistema di ipotesi da verificare diventa allora*

$$H_0 : \mu = 14mm$$

contro l'alternativa

$$H_1 : \mu > 14mm.$$

E' ragionevole in questo caso rifiutare l'ipotesi nulla per valori osservati della statistica test

$$T = \frac{\sqrt{n}(\overline{Y} - 14)}{\sigma}$$

maggiori di una certa soglia h.

La soglia sarà determinata dalla condizione

$$\alpha = \Pr(T > h, \textit{ quando } H_0 \textit{ è vera}) = \Pr(\mathcal{N}(0,1) > h).$$

Si avrà dunque $h = z_{1-\alpha}$.

Sintesi della procedura (test unilaterale)

In definitiva, per verificare un sistema d'ipotesi del tipo

$$\begin{cases} \mathrm{H}_0 : \mu = \mu_0 \\ \mathrm{H}_1 : \mu > \mu_0 \end{cases}$$

siamo arrivati alla seguente procedura:

scegliere il livello di significatività α

↓

determinare $z_{1-\alpha}$

↓

calcolare $t_{oss} = \dfrac{\sqrt{n}(\overline{y} - \mu_0)}{\sigma}$

↓

verificare se $t_{oss} \leq z_{1-\alpha}$

se la risposta è si → accettare H_0

se la risposta è no → rifiutare H_0

Nel caso in esame

$\alpha = 0.01$ (ad esempio)

↓

$z_{1-\alpha} = z_{0.99} = 2.326$

↓

$$t_{oss} = \frac{\sqrt{5}(14.302 - 14)}{0.1} = 6.75$$

↓

$6.75 \leq 2.326$?

↓ no

rifiutiamo H_0

α_{oss} (test unilaterale)

Anche per test unilaterali possiamo definire il livello di significatività osservato, con alcuni accorgimenti.

Se l'alternativa è di tipo unilaterale destro, abbiamo visto che la regione di rifiuto per un test di livello α prende la forma $\{t_{oss} > z_{1-\alpha}\}$. Allora la probabilità di osservare sotto H_0 un valore della statistica test più lontano dall'ipotesi nulla del valore osservato in corrispondenza dei nostri dati, è

$$\alpha_{oss} = \Pr(T > t_{oss}, \text{ quando } H_0 \text{ è vera}) = \Pr(\mathcal{N}(0,1) > t_{oss}).$$

Nel nostro esempio, $t_{oss} = 6.75$ e $\alpha_{oss} \approx 0$ (calcolandolo, con un computer, sarebbe circa uguale a $7.39 \cdot 10^{-12}$), confermando che i dati mostrano una forte evidenza contro H_0.

Una situazione simmetrica si ottiene invece considerando un sistema di ipotesi con alternativa **unilaterale sinistra**:

$$H_0 : \mu = \mu_0$$

e

$$H_1 : \mu < \mu_0.$$

In questo caso la regione di rifiuto per un test di livello α prende la forma $\{t_{oss} < -z_{1-\alpha}\}$. Allora

$$\alpha_{oss} = \Pr(T < t_{oss}, \text{ quando } H_0 \text{ è vera}) = \Pr(\mathcal{N}(0,1) < t_{oss}).$$

3.3 Dove facciamo conoscenza con uno statistico birraio

Test t ad un campione.

Intervalli di confidenza per la media di una normale quando la varianza non è nota.

Saxon Plumbing company

La Saxon Plumbing company[2] è un'azienda di vendita all'ingrosso di forniture da bagno in una zona residenziale alla periferia di una città del Nord Est degli Stati Uniti. Allo scopo di mantenere un controllo interno delle vendite, l'azienda ha stabilito che alle fatture di vendita venga allegato uno scontrino di trasferimento del magazzino, senza il quale le merci non possono essere rimosse dal magazzino. Alla fine di ogni mese, viene estratto un campione di fatture per valutarne l'ammontare medio riportato. Nel corso degli ultimi 5 anni l'ammontare medio delle fatture è stato di 120 \$. Dal momento che i costi di trasporto sono influenzati dalla distanza di consegna, è importante mantenere un controllo dell'ammontare medio.

I dati seguenti si riferiscono agli importi di un campione di 12 fatture

108.98 152.22 111.45 110.59 127.46 107.26 93.32 91.97 111.56 75.71 128.58 135.11

- La media delle dodici fatture è 112.85\$. Quindi, se restringiamo l'attenzione alle dodici fatture considerate vi è una diminuzione .

- Viene però spontaneo porsi la domanda: "Sulla base di questi risultati ci aspettiamo che vi sia un calo dell'ammontare medio delle fatture *in generale*?".

Un possibile modello di riferimento

- Consideriamo l'insieme di tutte le fatture che avremmo potuto esaminare. Si tratta ovviamente di un insieme molto grande ovvero una popolazione molto grande.

- L'ammontare della fattura è il risultato di una miriade di fattori (il prodotto, il cliente che compra, ...). Ora se tutti questi fattori si

[2]Tratto da Levine, Krehbiel, Berenson (2002) Statistica, Apogeo.

"compongono" in maniera additiva possiamo pensare sulla base del teorema del limite centrale che la distribuzione dell'ammontare nella popolazione possa essere ben approssimata da una distribuzione normale di appropriata media e varianza, diciamo μ e σ^2.

- Supponiamo inoltre che le fatture scelte non abbiano caratteristiche particolari e quindi siano assimilabili ad individui *estratti casualmente dalla popolazione.*

- Allora, se tutto questo è vero, possiamo vedere i dati osservati, indichiamoli al solito con $y_1, \dots, y_{12}$, come delle determinazioni indipendenti ed identicamente distribuite di una $\mathcal{N}(\mu, \sigma^2)$.

Due precisazioni

- Nel seguito "lavoreremo" supponendo vera l'ipotesi di normalità. Nella realtà, questa ipotesi dovrebbe prima essere verificata con i dati disponibili. Ovvero, il primo stadio dell'analisi dovrebbe consistere nell'utilizzare i dati per rispondere alla domanda: "E' plausibile che i dati osservati siano stati generati da una normale?". Per rispondere a questa domanda esistono tecniche grafiche ed analitiche (di cui non parleremo). Può, comunque, interessare lo studente che utilizzando queste tecniche la risposta alla domanda precedente è: "Sì, è plausibile.".

- Il modello che stiamo utilizzando per interpretare i dati è simile a quello considerato nell'unità [3.2]. La differenza è che in quell'unità σ^2 era noto (od almeno assunto tale). Qui è un parametro ignoto.

Stima dei parametri del modello

- Nelle ipotesi fatte, la distribuzione dei dati (e soprattutto del fenomeno considerato nella popolazione) è nota con l'eccezione dei due parametri μ e σ^2. Sembra quindi ragionevole iniziare cercando di stimare questi due parametri dai dati.

- Le stime più usate per μ e σ^2 sono rispettivamente

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \approx 112.85$$

e

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2 \approx 432.56$$

dove, al solito, n indica il numero delle osservazioni (per l'esperimento considerato $n = 12$).

- Si noti che per stimare σ^2 si usa dividere la somma dei quadrati degli scarti dalla media campionaria per $n-1$, non per n. E' infatti possibile dimostrare che dividere per n porterebbe ad uno stimatore che tendenzialmente "sottostima" la vera varianza. Lo stesso non vale per s^2. Infatti, sarebbe possibile dimostrare (ma non lo facciamo...) che

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right) = \frac{n-1}{n}\sigma^2$$

e quindi, ponendo $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, si ha

$$\mathbb{E}(S^2) = \sigma^2\,,$$

e cioè S^2 è uno stimatore non distorto di σ^2.

Un problema di verifica d'ipotesi

- Un sistema d'ipotesi interessante in questo caso è

$$\begin{cases} \mathrm{H}_0 : \mu = \mu_0 \\ \mathrm{H}_1 : \mu \neq \mu_0 \end{cases}$$

con $\mu_0 = 120$. Accettare H_0, infatti, equivale a dire che, in media, non c'è stata variazione nell'ammontare delle fatture.

- Per verificare un sistema d'ipotesi analogo nell'unità 3.2 avevamo utilizzato la statistica test

$$T = \frac{\sqrt{n}(\overline{Y} - \mu_0)}{\sigma}.$$

Però in questa unità noi non conosciamo σ. Quindi, con i dati a disposizione, non possiamo calcolare il valore osservato di T, e cioè t_{oss}.

- D'altra parte, poiché abbiamo a disposizione una stima di σ, una statistica test simile è data da

$$T = \frac{\sqrt{n}(\overline{Y} - \mu_0)}{S}.$$

Se H_0 è vera ci aspettiamo che $t_{oss} = \sqrt{n}(\overline{y} - \mu_0)/s$ assuma valori vicini zero. Invece, se H_1 è vera ci aspettiamo che t_{oss} cada lontano da zero.

Quanto deve essere lontano da zero t_{oss} per concludere che H_0 è implausibile?

- Per rispondere alla domanda avremmo bisogno di sapere qual è la distribuzione di T quando H_0 è vera. Infatti, questa distribuzione ci "racconta" quali sono i valori di t_{oss} che ci aspettiamo sotto l'ipotesi nulla.

- Sappiamo dall'unità precedente che, quando σ è noto, la distribuzione di T è normale. Potremmo perciò pensare di approssimare la distribuzione di T con una $N(0,1)$ anche nel caso in esame (con σ ignoto). Ma la sostituzione del vero σ con s non può non essere "indolore" soprattutto nel caso di piccoli campioni in cui l'errore con cui s stima σ potrebbe anche essere grande.

- E' però possibile nelle nostre ipotesi (normalità delle osservazioni, indipendenza,...) determinare la distribuzione esatta di T. E' stato fatto da W.S.Gosset uno statistico che lavoravava alla birreria (nel senso di fabbrica di birra) Guinness. Poiché i suoi lavori furono pubblicati sotto lo pseudonimo di Student, e Gosset, usava la lettera t per indicare la statistica test, la distribuzione viene comunemente chiamata t di Student.

- Il test che stiamo descrivendo viene usualmente chiamato **test t a un campione**.

La distribuzione t di Student

- La distribuzione t di Student dipende da un solo parametro, chiamato i gradi di libertà. Nel caso in esame (verifica sulla media di una distribuzione normale) deve essere posto uguale a $n-1$, ovvero, quello che Student ha dimostrato è che

$$T \sim t_{n-1} \quad (t \text{ di Student con } n-1 \text{ gradi di libertà.})$$

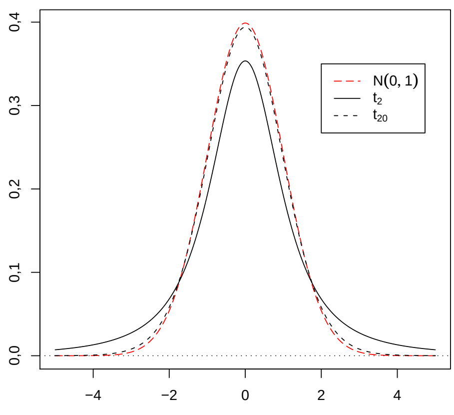
- Nel grafico che segue sono disegnate le densità di
  - una t di Student con 2 gradi di libertà;
  - una con 20 gradi di libertà;
  - una normale standard.

- Si osservi come:

1. le densità delle t siano simmetriche intorno allo zero;
2. abbiano delle "code" un po' più pesanti della normale e
3. la t con 20 gradi di libertà sia molto vicina alla $\mathcal{N}(0,1)$.

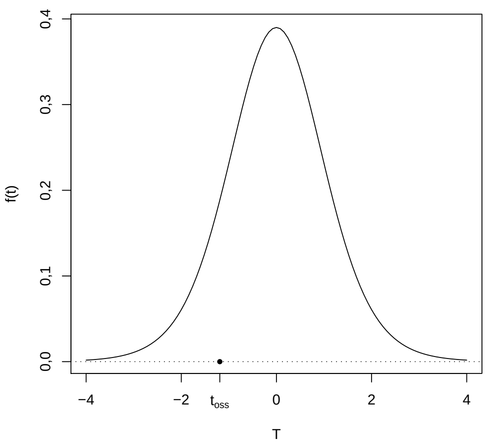
È possibile dimostrare che 1 e 2 valgono in generale (per qualsivoglia grado di libertà). L'osservazione 3 discende dal fatto che al divergere dei gradi di libertà la distribuzione di una t di Student converge ad una $\mathcal{N}(0,1)$. Il che tra l'altro è quello che ci aspettiamo. Infatti, più n è grande più s^2 dovrebbe avvicinarsi a σ^2 e quindi più T dovrebbe avvicinarsi all'analogo test calcolato con σ noto.

Grafico della densità della t di Student



Nota: I pedici indicano i gradi di libertà.

Analisi grafica del risultato



Il valore osservato di T sui nostri dati è pari a $t_{oss} = -1.19$. Nel grafico il valore è indicato dal punto sull'asse delle ascisse. La curva mostra la densità di una t di Student con 11 gradi di libertà. Non sembrano esserci elementi per dubitare che il valore osservato sia stato generato dalla distribuzione disegnata. Ovvero, non abbiamo elementi nei dati per rifiutare H_0.

Analisi mediante il livello di significatività osservato

- "Lontano da H_0" equivale a "lontano da 120 in ambedue le direzioni". Quindi, nel nostro caso,

$$\alpha_{oss} = \Pr(|t \text{ con 11 gradi di libertà}| \geq |t_{oss}|).$$

  che, per la simmetria della t di Student, possiamo anche calcolare come

$$\alpha_{oss} = 2 \times \Pr(t \text{ con 11 gradi di libertà} \geq |t_{oss}|).$$

- Disponendo solo di una tabella dei percentili, del tipo allegato al fondo di questa unità, possiamo determinare un intervallo che lo contiene.

- In particolare, dalla tabella vediamo che 1.19 è compreso tra i percentili di livello 0.75 e 0.90 di una t con 11 gradi di libertà. Quindi,

$$0.10 < \Pr(t \text{ con 11 gradi di libertà} \geq 1.19) < 0.25.$$

  Ma allora

$$0.2 < \alpha_{oss} < 0.5.$$

- Per quello che riguarda l'interpretazione, la prima disuguaglianza è la più importante. Ci racconta infatti che se l'ammontare medio non è cambiato allora noi ci aspetteremmo valori "più lontani da H_0 di quanto osservato" con una frequenza superiore al 20%. Questo, vuol dire che il valore osservato di t_{oss} non è "strano" quando H_0 è vera.

- Ad esempio, possiamo guardare al livello di significatività osservato in questa maniera. Supponiamo:

  - che sia vera l'ipotesi nulla,
  - di formare tutti i possibili campioni di numerosità 12 con gli individui che fanno parte della popolazione,
  - di calcolare per ciascun campione t_{oss}.

  Allora, il livello di significativà osservato è la percentuale di valori $|t_{oss}|$ con un valore maggiore di 1.19. Il calcolo precedente ci dice che questa percentuale è maggiore di 0.2 (e minore di 0.5). Ma allora -1.19 è un valore che "può capitare quando H_0 è vera". Del resto, non riteniamo sorprendente che il lancio di un dado equilibrato sia 3. Ma la percentuale di casi in cui un lancio ci dà come risultato 3 è inferiore al 20%.

- In conclusione, i dati ci dicono che non abbiamo elementi per rifiutare l'ipotesi nulla.

Una regola del tipo accetto/rifiuto

scegliere un valore α

$\downarrow$

determinare $t_{n-1,1-\alpha/2}$

$\downarrow$

calcolare $t_{oss} = \dfrac{\sqrt{n}(\overline{y} - \mu_0)}{s}$

$\downarrow$

se $|t_{oss}| \leq t_{n-1,1-\alpha/2}$

allora $\swarrow$ accettare H_0

altrimenti $\searrow$ rifiutare H_0

Nell'albero $t_{g,p}$ indica il percentile p-simo di una t di Student con g gradi di libertà. E' facile far vedere che l'albero fornisce una regola per accettare/rifiutare l'ipotesi nulla che garantisce che

$$\Pr(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera}) = 1 - \alpha$$

Con i dati sulle fatture

Supponiamo di porre $\alpha = 0.01$. Allora

$$\begin{array}{c} t_{n-1,1-\alpha/2} = t_{11,0.995} = 3.11 \\ \downarrow \\ t_{oss} = -1.19 \\ \downarrow \\ -3.11 \leq -1.19 \leq 3.11 \ ? \\ \Big\downarrow \text{si} \\ \text{accettiamo } \mathrm{H}_0 \end{array}$$

Un intervallo di confidenza

- Un intervallo di confidenza per μ può essere determinato, dai risultati precedenti utilizzando lo stesso ragionamento seguito nell'unità 3.2.
- Infatti quello che sappiamo è che se μ è il vero valore della media allora
$$\Pr(-t_{n-1,1-\alpha/2} \leq \sqrt{n}(\overline{Y} - \mu)/S \leq t_{n-1,1-\alpha/2}) = 1 - \alpha.$$
Ma allora, scrivendo le due disuguglianze in termini di μ, troviamo che
$$\Pr(\overline{Y} - S\ t_{n-1,1-\alpha/2}/\sqrt{n} \leq \mu \leq \overline{Y} + S\ t_{n-1,1-\alpha/2}/\sqrt{n}) = 1 - \alpha\,,$$
ovvero che
$$\left[\overline{y} - \frac{s\ t_{n-1,1-\alpha/2}}{\sqrt{n}}, \overline{y} + \frac{s\ t_{n-1,1-\alpha/2}}{\sqrt{n}}\right]$$
è un intervallo di confidenza di livello $1 - \alpha$ per la media.
- Supponiamo, ad esempio, di voler un intervallo di confidenza di livello 0.99. Allora $t_{n-1,1-\alpha/2} = t_{11,0.995} = 3.11$. Ricordando che $\overline{y} \approx 112.85$ e $s^2 \approx 432.56$ e quindi che $s \approx \sqrt{432.56} \approx 20.80$, la semi-ampiezza dell'intervallo richiesto è
$$18.67 = \frac{20.80 \times 3.11}{\sqrt{12}},$$
mentre l'intervallo stesso è
$$[112.85 - 18.67 \ ; \ 112.85 + 18.67] = [94.18 \ ; \ 131.52]$$

- Si osservi che l'intervallo include il valore $\mu = 120$. Questo era atteso dato che avevamo visto, con il test discusso precedentemente, che il valore 120 per μ era plausibile sulla base dei dati disponibili.

Quantili di una t di Student

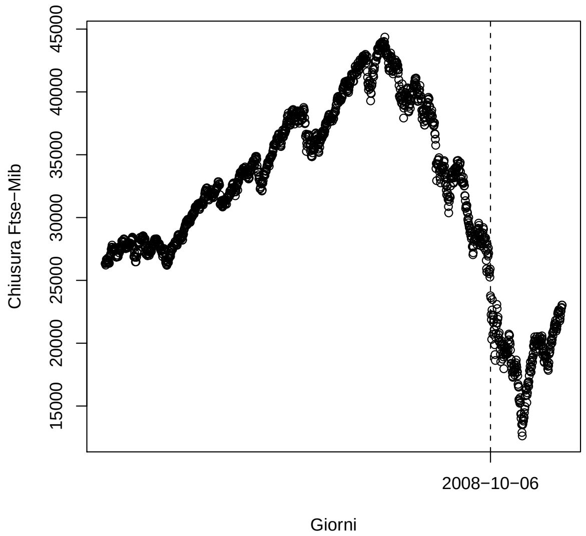
g indica i gradi di libertà. p la probabilità lasciata a "sinistra". Quindi, ad esempio, $\Pr(t \text{ con 2 gradi di libertà} \leq 6.96) = 0.99$. L'ultima riga ($g = \infty$) mostra i quantili di una $\mathcal{N}(0, 1)$. Possono essere usati come approssimazione, se $g > 30$.

	p							
g	0.75	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
1	1	3.08	6.31	12.71	31.82	63.66	318.31	636.62
2	0.82	1.89	2.92	4.3	6.96	9.92	22.33	31.6
3	0.76	1.64	2.35	3.18	4.54	5.84	10.21	12.92
4	0.74	1.53	2.13	2.78	3.75	4.6	7.17	8.61
5	0.73	1.48	2.02	2.57	3.36	4.03	5.89	6.87
6	0.72	1.44	1.94	2.45	3.14	3.71	5.21	5.96
7	0.71	1.41	1.89	2.36	3	3.5	4.79	5.41
8	0.71	1.4	1.86	2.31	2.9	3.36	4.5	5.04
9	0.7	1.38	1.83	2.26	2.82	3.25	4.3	4.78
10	0.7	1.37	1.81	2.23	2.76	3.17	4.14	4.59
11	0.7	1.36	1.8	2.2	2.72	3.11	4.02	4.44
12	0.7	1.36	1.78	2.18	2.68	3.05	3.93	4.32
13	0.69	1.35	1.77	2.16	2.65	3.01	3.85	4.22
14	0.69	1.35	1.76	2.14	2.62	2.98	3.79	4.14
15	0.69	1.34	1.75	2.13	2.6	2.95	3.73	4.07
16	0.69	1.34	1.75	2.12	2.58	2.92	3.69	4.01
17	0.69	1.33	1.74	2.11	2.57	2.9	3.65	3.97
18	0.69	1.33	1.73	2.1	2.55	2.88	3.61	3.92
19	0.69	1.33	1.73	2.09	2.54	2.86	3.58	3.88
20	0.69	1.33	1.72	2.09	2.53	2.85	3.55	3.85
21	0.69	1.32	1.72	2.08	2.52	2.83	3.53	3.82
22	0.69	1.32	1.72	2.07	2.51	2.82	3.5	3.79
23	0.69	1.32	1.71	2.07	2.5	2.81	3.48	3.77
24	0.68	1.32	1.71	2.06	2.49	2.8	3.47	3.75
25	0.68	1.32	1.71	2.06	2.49	2.79	3.45	3.73
26	0.68	1.31	1.71	2.06	2.48	2.78	3.43	3.71
27	0.68	1.31	1.7	2.05	2.47	2.77	3.42	3.69
28	0.68	1.31	1.7	2.05	2.47	2.76	3.41	3.67
29	0.68	1.31	1.7	2.05	2.46	2.76	3.4	3.66
30	0.68	1.31	1.7	2.04	2.46	2.75	3.39	3.65
∞	0.67	1.28	1.64	1.96	2.33	2.58	3.09	3.29

3.4 Crisi del mercato?

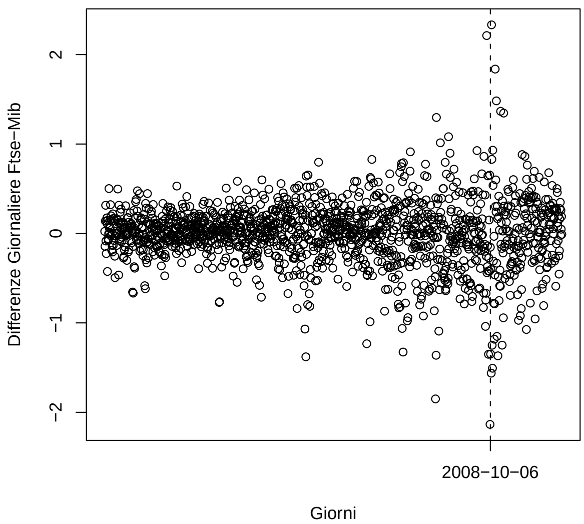
Intervalli di confidenza e verifica d'ipotesi per la varianza nel caso di un campione tratto da una v.c. normale.

FTSE MIB



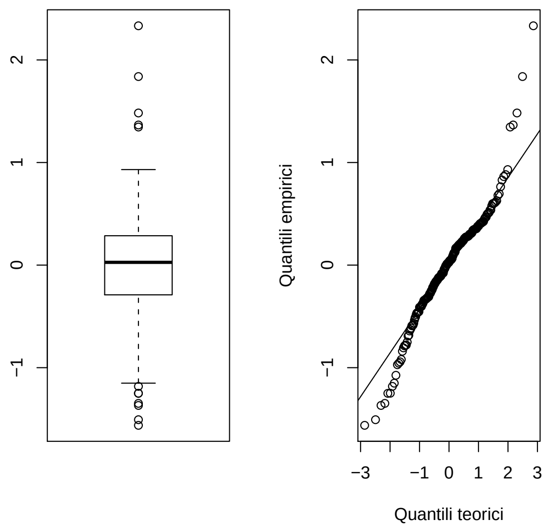
Valore giornaliero di chiusura aggiustata dell'indice FTSE MIB in migliaia di Euro. Fonte: it.finance.yahoo.com

Differenze FTSE MIB



Differenze dei valori giornalieri di chiusura aggiustata dell'indice FTSE MIB in migliaia di Euro. Fonte: it.finance.yahoo.com

Differenze FTSE MIB dopo il 6 Ottobre 2008



Differenze dei valori giornalieri di chiusura aggiustata dell'indice FTSE MIB, in migliaia di Euro, dal 6 Ottobre 2008 al 14 Settembre 2009. Fonte: it.finance.yahoo.com

Una possibile formulazione del problema

Sappiamo che la variabilità della differenze giornaliere dell'indice in un momento di stabilità dei mercati è di 0,105.

- Dai grafici precedenti possiamo concludere, in prima approssimazione, che le differenze giornaliere dell'indice, nel periodo 6 Ottobre 2008 – 14 Settembre 2009, possono essere assimilate a osservazioni da una variabile normale di media e varianza ignote e quindi
$$Y \sim N(\mu, \sigma^2)\,;$$
- Possiamo costruire un test per saggiare l'ipotesi:
$$\begin{cases} \mathrm{H}_0 : \sigma^2 \leq 0,105 \equiv \sigma_0^2 \\ \mathrm{H}_1 : \sigma^2 > \sigma_0^2 \end{cases}$$
- Possiamo pensare di costruire un intervallo di confidenza per la varianza e confrontarlo con il valore che ci attendiamo in caso di stabilità dei mercati;

Stima dei parametri ignoti

Sappiamo già che dato un campione $y_1, y_2, \cdots, y_n$ una stima dei parametri della distribuzione normale, media e varianza è ottenibile con il metodo dei momenti
$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = -0,003$$
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}) = 0,265$$

Per ottenere una statistica test possiamo utilizzare la distribuzione dello stimatore S^2.

Distribuzione di S^2

La distribuzione *esatta* di S^2 è facile da determinare. Infatti, abbiamo il seguente risultato

Proposizione 24. *Siano $Y_1, Y_2, \cdots, Y_n$, n variabili casuali indipendenti e identicamente distribuite secondo una normale $N(\mu, \sigma^2)$. Allora la variabile casuale*
$$W = \sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right)^2$$
ha distribuzione χ^2_n con n gradi di libertà.

Ora, ricordando che

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left[(Y_i - \mu) - (\bar{Y} - \mu)\right]^2 = \sum_{i=1}^{n}(Y_i - \mu)^2 - n(\bar{Y} - \mu)^2$$

da cui

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right)^2 - n\left(\frac{\bar{Y} - \mu}{\sigma}\right)^2$$

e quindi

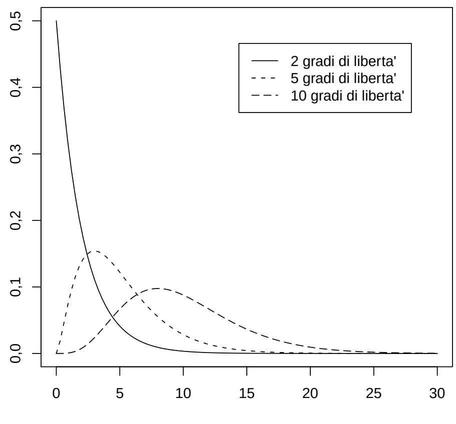
$$\sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right)^2 = (n-1)\frac{S^2}{\sigma^2} + \left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 .$$

Il termine di sinistra è distribuito come un χ^2_n, mentre il secondo termine di destra è un χ^2_1. Da cui si intuisce che

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} .$$

La variabile casuale χ^2

La variabile casuale χ^2 ha una distribuzione che dipende da un solo parametro, chiamato i gradi di libertà della distribuzione.



Si osservi che all'aumentare dei gradi di libertà la densità si sposta verso destra. Infatti, è possibile dimostrare che la media della variabile casuale coincide con i gradi di libertà.

Intervalli di confidenza

La distribuzione per S^2 può essere usata per costruire intervalli di confidenza di livello $1-\alpha$ prefissato.

$$\Pr(w_{n-1,\alpha/2} \leq (n-1)\frac{S^2}{\sigma^2} \leq w_{n-1,1-\alpha/2}) = 1-\alpha$$
$$\Pr\left(\frac{w_{n-1,\alpha/2}}{(n-1)S^2} \leq \frac{1}{\sigma^2} \leq \frac{w_{n-1,1-\alpha/2}}{(n-1)S^2}\right) = 1-\alpha$$
$$\Pr\left(\frac{(n-1)S^2}{w_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{w_{n-1,\alpha/2}}\right) = 1-\alpha$$

dove con $w_{n,p}$ indichiamo il quantile p-simo di un χ^2 con n gradi di libertà. ovvero, ci mostra che

$$\left[\frac{(n-1)s^2}{w_{n-1,1-\alpha/2}};\frac{(n-1)s^2}{w_{n-1,\alpha/2}}\right]$$

costituisce un intervallo di confidenza di livello $1-\alpha$ per σ^2.

Con i nostri dati

Supponiamo di voler calcolare un intervallo di confidenza di livello 0,9.

Allora $\alpha = 0{,}1$, $\alpha/2 = 0{,}05$ e $1-\alpha/2 = 0{,}95$. Da una tavola della distribuzione χ^2_{236} (o utilizzando un programma appropriato) troviamo che $w_{236,0,05} = 201,437$ e $w_{236,0,95} = 272,836$. Sappiamo già che $s^2 = 0,265$ e perciò l'intervallo è

$$\left[\frac{(237-1)0,265}{272,836}\ ;\ \frac{(237-1)0,265}{201,437}\right] = [0,229\ ;\ 0,311].$$

da cui si vede che il valore 0,105 non è incluso nell'intervallo. In maniera analoga si costruisco intervalli unilaterali.

È proprio crisi dei mercati?

Sulla base dell'intervallo di confidenza al 90% si evince che la variabilità nell'ultimo periodo è superiore alla variabilità del mercato in condizioni normali. Vogliamo verificare se la variabilità nell'ultimo periodo risulta doppia di quella usuale. Per rispondere a questo quesito dobbiamo considerare il sistema d'ipotesi seguente:

$$\begin{cases} \mathrm{H}_0 : \sigma^2 = \sigma_0^2 \\ \mathrm{H}_1 : \sigma^2 > \sigma_0^2 \end{cases}$$

con $\sigma_0^2 = 2 \times 0,105 = 0,211$. [3]

Una possibile *statistica test* è

$$T = (n-1)\frac{S^2}{\sigma^2}$$

Se l'ipotesi nulla è vera, per quanto ricordato a pagina 202, T ha una distribuzione χ^2_{n-1}.

Quindi possiamo confrontare il valore di T calcolato dai dati con questa distribuzione. Valori 'piccoli' di T ci indicheranno che H_0 è plausibile (rispetto ad H_1).

Allora la regione di rifiuto sarà data dai valori del test maggiori di $w_{n-1,1-\alpha}$, poiché fissiamo ad α la probabilità di commettere un errore di I tipo (e cioè di rifiutare H_0 quando è la vera ipotesi)

$$\Pr(T \geq w_{n-1,1-\alpha} \quad \text{quando } \mathrm{H}_0 \text{ è vera}) = \alpha\,.$$

Nel nostro caso, se $\alpha = 0{,}05$, $w_{236,0,95} = 272,836$ e, poiché

$$t_{oss} = (237-1)\frac{0,265}{0,211} = 296,855$$

accettiamo l'ipotesi nulla in favore dell'ipotesi alternativa.

Il livello di significatività osservato è pari a

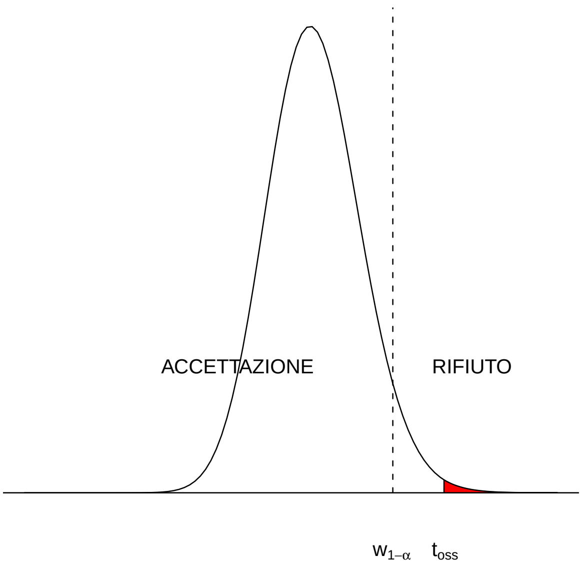
$$\begin{aligned} \alpha_{oss} = \Pr(T \geq t_{oss} \quad \text{quando } \mathrm{H}_0 \text{ è vera}) &= \Pr(\chi^2_{236} \geq 296,855) \\ &= 0,004 \end{aligned}$$

ed è, coerentemente, minore di 0,05.

---

[3] Il sistema è equivalente a

$$\begin{cases} \mathrm{H}_0 : \sigma^2 \leq \sigma_0^2 \\ \mathrm{H}_1 : \sigma^2 > \sigma_0^2 \end{cases}$$



E nel caso la media sia nota?

Se $Y_i \sim N(\mu, \sigma^2)$ e la media è nota allora il metodo dei momenti ci fornisce il seguente stimatore per la varianza (lo studente verifichi che è così)

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu)^2$$

che è uno stimatore non distorto e consistente di σ^2.

Dalla proposizione 24 abbiamo subito che

$$n \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_n^2 \ .$$

Utilizzando questa statistica si ottengono intervalli di confidenza e test statistici procedendo nella stessa maniera del caso in cui la media sia ignota.

3.5 Dove si incontrano un politico che dichiara: "Allora vinco" e uno statistico

Stima della probabilità di successo, intervalli di confidenza e verifica d'ipotesi nel caso di una binomiale.

Il caso

Un politico vuole candidarsi alle elezioni in una circoscrizione di 100000 elettori. Prima di candidarsi vuole però sapere se ha buona probabilità di successo. Per questo commissiona ad una società un sondaggio. La società contatta 2500 elettori di questi 1328 si dichiarano favorevoli al candidato, ovvero il $1328/2500 \times 100\% \simeq 53\%$. Il candidato (C) discute con uno statistico (S) il risultato e come prima affermazione dice: allora vinco!

Un dialogo

S. Piano con i proclami.

C. 1338 persone sono tante.

S. E' vero, ma si tratta di un campione. E come è stato selezionato il campione?

C. A caso. Ovvero sono state contattate delle persone per telefono.

S. Beh, allora proprio a caso il campione non è stato selezionato. Hanno risposto solo coloro che avevavo il telefono e intendevano rispondere.

C. Questo non lo so. E' grave?

S. Potrebbe esserlo.

C. Ma allora vinco?

S. Sicurezza non le darò, ma ora le spiego il mio modo di vedere la cosa.

Un possibile modello

- Indichiamo con y il numero di intervistati che sono favorevoli al candidato e con n in numero totale di intervistati. Nel caso dell'esperimento descritto $y = 1328$ e $n = 2500$. Vogliamo poter dire qualcosa su ϑ la percentuale incognita di elettori favorevoli.

- Facciamo una prima ipotesi (in mancanza di altra informazione) ovvero che tutti gli elettori siano raggiungibili telefonicamente e che tutti intendevano rispondere. Tutti saranno d'accordo che per la prima chiamata la probabilità di estrazione di 'Rossi Mario' (o di un altra qualsiasi persona nella circoscrizione) sia di 1/100000. Nella seconda intervista non richiamiamo più 'Rossi Mario' ma 'Verdi Giuseppe'. La probabilità di estrazione di 'Verdi Giuseppe' (o di un altra qualsiasi persona nella circosrizione ad l'esclusione di 'Rossi Mario') sia di 1/99999.

- Questo complica le cose. Tuttavia semplifichiamo ancora un po' (non sembra grave) e ammettiamo che tutte gli intervistati hanno la *stessa* probabilità di entrare nel campione. E ammettiamo che gli intervistati non usino il passaparola, ovvero che le risposte siano tra loro *indipendenti*.

Due formulazioni equivalenti

1. Possiamo assimilare il sondaggio all'estrazione casuale con reinserimento di n elettori da un'urna costituita da tutti gli elettori. Se accettiamo questo, allora
$$Y \sim \mathcal{B}(n, \vartheta) \tag{3.7}$$
ovvero, il numero di favorevoli tra gli n estratti può essere visto come determinazione, y, di una variabile casuale binomiale Y, con probabilità di successo ϑ e numero di prove n.

   Si osservi che la (3.7) è cruciale perché precisa la relazione tra quello che conosciamo (y e n) e quello che vogliamo conoscere (ϑ).

2. Notiamo inoltre che il parere sul candidato (favorevole o non favorevole) di ogni singolo elettore poteva essere rappresentato attraverso una variabile casuale $X_i \sim \mathcal{B}(1, \theta)$, per $i = 1, \ldots, n$, e che Y altro non è che la somma delle n v.c. X_i: $Y = \sum_{i=1}^{n} X_i$.

Stima di ϑ

Lo stimatore più "naturale" (forse l'unico "naturale" nel senso che qualsiasi altra scelta sembra "innaturale") per ϑ è

$$\hat{\vartheta} = \frac{Y}{n}$$

ovvero la proporzione di elettori favorevoli (il politico non era così sprovveduto).

Si ha che

$$\mathbb{E}(\hat{\theta}) = \theta$$

e

$$\mathbb{V}ar(\hat{\theta}) = \frac{\theta(1-\theta)}{n}.$$

$\hat{\theta}$ è perciò uno stimatore non distorto e consistente in media quadratica (e dunque anche debolmente).

Nel caso in esame, la stima è pari a $\hat{\vartheta} = y/n = 1328/2500 = 0{,}5312$.

Esercizio 20. *Verificate che lo stimatore $\hat{\vartheta}$ coincide con lo stimatore ottenuto con il metodo dei momenti.*

La distribuzione esatta di $\hat{\theta}$

Ovviamente, se Y è una variabile casuale anche $\hat{\vartheta}$ è una variabile casuale. Lo studio della sua distribuzione è importante perché permette di acquisire una idea sulla dimensione dell'errore di stima (come abbiamo già visto per la media della distribuzione normale nell'unità 3.2).

La distribuzione *esatta* di $\hat{\vartheta}$ è facile da determinare. Infatti, $\hat{\vartheta} \in \Theta_n = \{0/n, 1/n, \ldots, n/n\}$ e, per qualsivoglia $a \in \Theta_n$, risulta

$$\Pr(\hat{\vartheta} = a) = \Pr(Y = na) = \binom{n}{na} \vartheta^{na}(1-\vartheta)^{n-na}.$$

Approssimazione normale

Il fatto che la distribuzione esatta sia facile da determinare non implica che sia anche facile da maneggiare. La maniera più comoda per determinare intervalli di confidenza e test si basa sull'approssimazione normale alla binomiale.

Il risultato di partenza è costituito dal fatto che per n non troppo piccolo la distribuzione di

$$\frac{\hat{\vartheta} - \vartheta}{\sqrt{\vartheta(1-\vartheta)/n}}$$

è approssimabile con quella di una normale standard nel senso che per ogni intervallo della retta reale $[a, b]$

$$\Pr\left(a \leq \frac{\hat{\vartheta} - \vartheta}{\sqrt{\vartheta(1-\vartheta)/n}} \leq b\right) \approx \Pr(a \leq \mathcal{N}(0,1) \leq b)$$

Distribuzione approssimata dell'errore di stima

Se $(\hat{\vartheta} - \vartheta)/\sqrt{\vartheta(1-\vartheta)/n}$ è approssimatamente una normale standard allora

$$(\text{errore di stima}) = (\hat{\vartheta} - \vartheta) \overset{a}{\sim} \mathcal{N}(0; \vartheta(1-\vartheta)/n).$$

Si osservi che, a differenza di quanto accadeva all'errore di stima nell'unità 3.2, questa distribuzione oltre ad essere approssimata è anche parzialmente ignota. Infatti, la varianza della distribuzione dipende dal vero valore di ϑ.

Per acquisire delle informazioni sulla dimensione dell'errore di stima possiamo stimarne la varianza sostituendo $\hat{\vartheta}$ a ϑ. Nel caso in esame troviamo

$$\widehat{\mathbb{V}ar}(\hat{\vartheta} - \vartheta) = \frac{\hat{\vartheta}(1-\hat{\vartheta})}{n} \approx \frac{0{,}5312(1-0{,}5312)}{2500} \approx (0{,}01)^2.$$

Approssimazione dopo approssimazione, siamo arrivati alla conclusione che l'errore di stima è, grossomodo, normale di media zero e scarto quadratico medio 0,01.

Intervalli di confidenza

La distribuzione stimata per $\hat{\vartheta} - \vartheta$ può essere usata per costruire intervalli di confidenza (almeno approssimatamente) di livello $1 - \alpha$ prefissato.

Infatti se la distribuzione di $\hat{\vartheta} - \vartheta$ è approssimatamente una normale di media nulla e scarto quadratico medio 0,01 allora possiamo scrivere (perché?)

$$\Pr(z_{1-\alpha/2} \leq \frac{\hat{\vartheta} - \vartheta}{0{,}01} \leq z_{1-\alpha/2}) \approx 1 - \alpha \tag{3.8}$$

dove, al solito, con z_p indichiamo il quantile p-simo di una normale standard. La (3.8) può essere scritta come

$$\Pr(\hat{\vartheta} - 0.01 \times z_{1-\alpha/2} \leq \vartheta \leq \hat{\vartheta} + 0.01 \times z_{1-\alpha/2}) \approx 1 - \alpha$$

ovvero, ci mostra, ricordando come avevamo calcolato lo scarto quadradico medio dell'errore di stima, che

$$\left[\hat{\vartheta} - z_{1-\alpha/2}\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}}, \hat{\vartheta} + z_{1-\alpha/2}\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}}\right]$$

costituisce un intervallo di confidenza per ϑ di livello approssimato $1 - \alpha$.

Con i nostri dati

Supponiamo di voler calcolare un intervallo di confidenza di livello 0.9.

Allora $\alpha = 0.1$ e $1 - \alpha/2 = 0{,}95$. Da una tavola della distribuzione normale (o utilizzando un programma appropriato) troviamo che $z_{0,95} \approx 1{,}65$. Sappiamo già che $\hat{\vartheta} = 0{,}5312$ e che

$$\sqrt{\frac{0{,}5312 \times (1 - 0{,}5312)}{2500}} \approx 0{,}01.$$

Quindi, la semi-ampiezza dell'intervallo richiesto è $1{,}65 \times 0{,}01 = 0{,}0165$. Perciò l'intervallo stesso è

$$[0{,}5312 - 0{,}0165\ ;\ 0{,}5312 + 0{,}0165] = [0{,}5147\ ;\ 0{,}5477]\ .$$

Ma allora vinco! ovvero la verifica dell'ipotesi

Sulla base dell'intervallo di confidenza al 90% il politico conclude che, poiché 0.5 non è contenuto nell'intervallo, vi è una ragionevole aspettativa che possa vincere.

Per quanto detto, il sistema d'ipotesi che sottintende il politico, è il seguente:

$$\begin{cases} \mathrm{H}_0 : \vartheta = \vartheta_0 \\ \mathrm{H}_1 : \vartheta > \vartheta_0 \end{cases}$$

con $\vartheta_0 = 0.50$ Il sistema è equivalente a

$$\begin{cases} \mathrm{H}_0 : \vartheta \leq \vartheta_0 \\ \mathrm{H}_1 : \vartheta > \vartheta_0 \end{cases}$$

Si noti che a differenza dell'unità **??** il sistema d'ipotesi ha una alternativa di tipo **unilaterale**.

Volendo utilizzare un test statistico sembra ragionevole basare la decisione sulla distanza tra la stima di ϑ calcolata dai dati e il valore per il parametro previsto sotto H_0.

Una possibile *statistica test* è quindi [4]

$$T = \frac{\hat{\vartheta} - \vartheta_0}{\sqrt{\vartheta_0(1 - \vartheta_0)/n}}.$$

[4] Si osservi che, come nell'unità 3.2, preferiamo lavorare con una versione "standardizzata" della differenza; la cosa è però irrilevante poiché il tutto si concretizza nella divisione per una costante

Si osservi che qui, a differenza di quanto fatto per la costruzione di intervalli di confidenza, al denominatore di T compare la radice della varianza di $\hat{\theta}$ *sotto l'ipotesi nulla* e quindi per $\theta = \theta_0$. Non c'è dunque bisogno di stimare la varianza di $\hat{\theta}$.

Se l'ipotesi nulla è vera, per quanto ricordato a pagina 221, T ha una distribuzione approssimatamente normale di media zero e varianza 1. Quindi possiamo confrontare il valore di T calcolato dai dati con questa distribuzione. Valore positivi di T ‘grandi’ ci indicheranno che H_0 è implausibile (rispetto ad H_1).[5]

Quindi, la regione di rifiuto sarà data dai valori del test maggiori di $z_{1-\alpha}$, poiché fissiamo ad α la probabilità di commettere un errore di I tipo (e cioè di rifiutare H_0 quando è la vera ipotesi)

$$\Pr(T \geq z_{1-\alpha} \quad \text{quando } \mathrm{H}_0 \text{ è vera}) \approx \alpha\,.$$

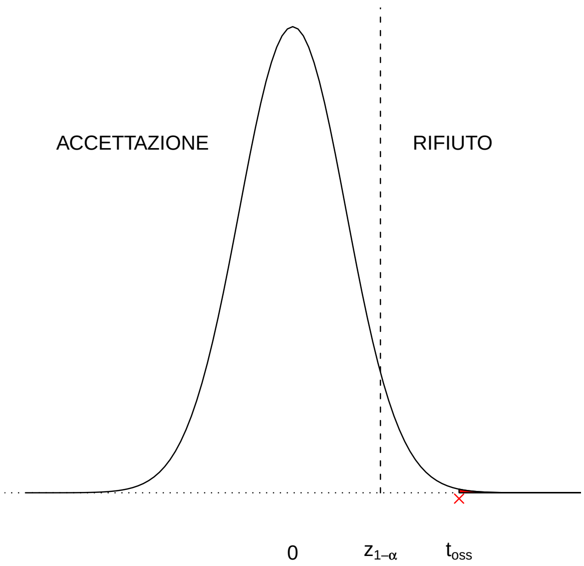
Nel nostro caso, se $\alpha = 0{,}05$, $z_{0{,}95} = 1{,}65$ e, poiché $t_{oss} = (0{,}5312 - 0{,}5)/\sqrt{0{,}5 \times 0{,}5/2500} = 3{,}12$, rifiutiamo l'ipotesi nulla in favore dell'ipotesi alternativa.

Il livello di significatività osservato è pari a

$$\begin{aligned} \alpha_{oss} = \Pr(T \geq t_{oss} \quad \text{quando } \mathrm{H}_0 \text{ è vera}) &\approx \Pr(\mathcal{N}(0;1) \geq 3{,}12) \\ &\approx 0{,}00088 \end{aligned}$$

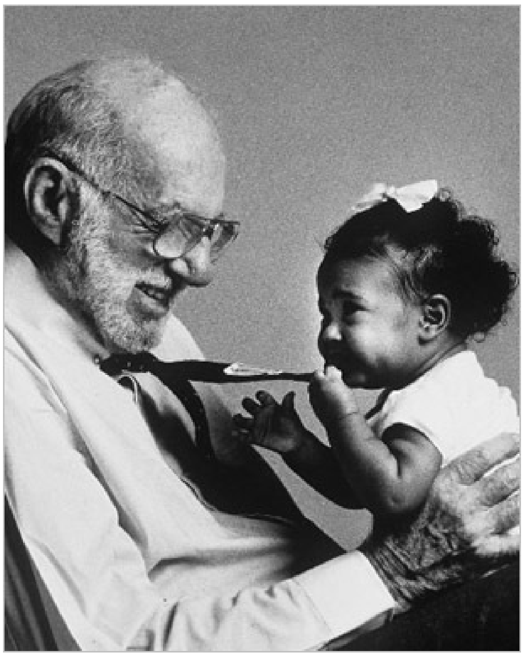
e quindi avremmo rifiutato H_0 a qualsiasi livello fissato “ragionevole”.

[5] Si osservi che questa è una differenza dalle unità 3.2 e 3.3


ACCETTAZIONE
RIFIUTO
0
$z_{1-\alpha}$
t_{oss}

3.6 Una giuria per il dottor Spock

Stima di una proporzione e verifica d'ipotesi nel caso di una distribuzione Binomiale.



Dott. Benjamin Spock: 2 Maggio 1903 – 15 Marzo 1998.

- Benjamin Spock è stato uno dei più famosi pediatri del secondo dopo guerra. In particolare i suoi libri ed articoli hanno contribuito notevolmente allo sviluppo di una pediatria e pedagogia meno autoritaria, più orientata verso i bisogni dei bambini che verso le "regole da rispettare".
- Nel 1968 il dott. Spock fu processato dal tribunale federale di Boston per cospirazione contro il *Military Service Act* (la legge sul servizio di leva). Il processo era la conseguenza della partecipazione di B. Spock al movimento contro la guerra nel Vietnam.
- La formazione delle giurie negli Stati Uniti era, ed è, un'operazione complicata.
- In particolare, nel caso in esame,
  - prima dovevano essere estratti 350 possibili giurati da una lista contenente centinaia di migliaia di *eleggibili*; la legge prevedeva che l'estrazione doveva essere casuale e fatta in maniera tale da garantire a ciascun eleggibile la stessa probabilità di estrazione;

- poi, sia l'accusa che la difesa potevano ricusare parte di questi potenziali giurati;
- infine, la giuria effettiva veniva estratta tra i giurati "non eliminati".

- Il processo fu affidato ad un giudice federale di nome Francis J.W. Ford, i cui compiti comprendevano l'estrazione dei 350 potenziali giurati.

- Era convinzione comune che giurati femminili avrebbero avvantaggiato la difesa, sia per un atteggiamento meno militarista delle donne, sia per il prestigio del dott. Spock tra il pubblico femminile.

- Ad esempio, quell'anno un avvocato scrisse sulla *Chicago Law Review*

  > *Of all defendants at such trials, Dr. Spock, who had given wise and welcome advice on child-bearing to millions of mothers, would haved liked women on his jury.*

- Il 53% della popolazione degli eleggibili era composto di donne. Destò sorpresa e polemica il fatto che solo 102 su 350 potenziali giurati risultarono donne.

- Il giudice Ford si difese affermando che il fatto che 102 donne erano state estratte dimostrava che non c'era stato nessun tentativo di escludere i possibili giurati di sesso femminile.

Il test statistico

- Possiamo inquadrare la questione "dare un giudizio sul comportamento del giudice Ford" come un problema di **verifica di ipotesi**.

- Problemi di scelta tra due (o più) alternative sono, in statistica, chiamati problemi di verifica di ipotesi.

- Le ipotesi (quando sono due) vengono spesso indicate come **ipotesi nulla** ed **ipotesi alternativa**. L'ipotesi nulla è solitamente indicata con la scrittura H_0, mentre quella alternativa come H_1.

- Lo "strumento" utilizzato per affrontare i problemi di verifica di ipotesi, ovvero, la procedura che si segue per decidere quale ipotesi accettare o rifiutare, viene chiamato **test statistico**.

- In altri termini, il test statistico valuta quanto i dati osservati siano compatibili o meno con l'ipotesi nulla.

Un possibile sistema di ipotesi

- Possiamo descrivere il problema del giudice Ford tramite il sistema di ipotesi

$$\begin{cases} \mathrm{H}_0 : \text{l'estrazione è stata fatta secondo la legge} \\ \mathrm{H}_1 : \text{l'estrazione è stata ``truccata''} \end{cases}$$

- I dati che possiamo utilizzare sono il risultato dell'estrazione (102 donne su 350 estratti).

- Per procedere abbiamo innanzitutto bisogno di specificare meglio l'ipotesi nulla. Ovvero, dobbiamo capire quale meccanismo probabilistico prevede la legge.

- Indichiamo con

  - N il numero degli eleggibili;

  - D il numero di donne tra gli eleggibili.

- La legge prevede che si debba

  - estrarre un primo individuo assegnando uguale probabilità a tutti gli eleggibili,

  - poi estrarre un secondo individuo tra i restanti $N-1$ assegnando anche questa volta uguale probabilità

  - e così via.

- La probabilità che il primo individuo sia donna è quindi D/N.

- Strettamente parlando, la probabilità che il secondo individuo sia donna dipende dal risultato della prima estrazione. Infatti la probabilità che il secondo estratto sia donna vale

$$\begin{cases} \frac{D-1}{N-1} & \text{se il } 1^\circ \text{ estratto è donna} \\ \frac{D}{N-1} & \text{se il } 1^\circ \text{ estratto è uomo} \end{cases}$$

- Nel nostro caso però N è molto grande (centinaia di migliaia) e quindi queste due probabilità sono "quasi" uguali tra di loro e "quasi" uguali a D/N. Ad esempio, se $N = 300.000$ e $D = 159.000$, allora $D/N = 0{,}53$, $(D-1)/(N-1) \approx 0{,}529998$ e $D/(N-1) \approx 0{,}530002$.

- Un discorso simile può essere fatto per le successive estrazioni.

- Quindi se il giudice Ford avesse seguito la legge, allora il numero di donne tra i potenziali giurati sarebbe quello ottenuto da una serie di 350 estrazioni tutte praticamente identiche, nel senso che in tutte le estrazioni la probabilità di estrarre un giurato femminile vale, approssimativamente, D/N.

- Ricordandoci che nel caso in esame $D/N = 0{,}53$, ovvero che il 53% degli eleggibili è donna, segue che

$$\left(\begin{array}{c} \text{numero donne} \\ \text{estratte} \end{array} \right) \sim \text{Binomiale}(350, 0{,}53)$$

- Descrivere in termini probabilistici l'ipotesi alternativa è viceversa complicato. Soprattutto perchè nessuno ci può garantire che, volendo "truccare" la giuria, si sia seguito un meccanismo in un qualsiasi senso assimilabile ad un esperimento casuale.

Un possibile test

- Siamo quindi davanti ad un problema di verifica d'ipotesi in cui H_0 è completamente specificata. Viceversa, H_1 è essenzialmente nebulosa.

- Ha senso in questo caso fare un test statistico? La risposta è sì. Con un test statistico cerchiamo di valutare se i dati potrebbero essere stati generati dal meccanismo previsto dall'ipotesi nulla. E questo è quello che vogliamo fare nel presente contesto visto che la domanda che ci stiamo ponendo è:

  > *È plausibile che il giudice Ford abbia seguito la legge ed estratto solo* 102 *donne?*

Il p-value

- In statistica per dare una valutazione di un test statistico si suole calcolare il cosiddetto **p-value** o **livello di significatività osservato** che corrisponde a calcolare

  > la probabilità di osservare qualcosa maggiormente in disaccordo con l'ipotesi nulla rispetto a quanto osservato *assumendo che l'ipotesi nulla sia vera.*

- Vediamo più in dettaglio il significato di questa affermazione

  – se il *p-value* è **alto**, allora è molto probabile osservare qualcosa di ancora più distante da H_0 rispetto a ciò che abbiamo osservato (assumendo che H_0 sia vera), quindi ciò che avevamo osservato non era così in disaccordo con H_0;

  – se, viceversa, il *p-value* è **basso**, allora è molto improbabile osservare qualcosa di più distante da H_0 di ciò che abbiamo osservato (assumendo che H_0 sia vera); in questo caso sembra ragionevole ritenere H_0 non compatibile con i dati osservati.

- Nel caso del nostro processo, il *p-value* è pari alla probabilità di osservare 102 o meno donne, ovvero (usando un computer!)

$$\Pr\left(D \leq 102 | D \sim \text{Binomiale}(350, 0{,}53)\right) =$$
$$\sum_{x=0}^{102} \binom{350}{x} 0{,}53^x (1 - 0{,}53)^{350-x} = 1{,}39 \times 10^{-19}$$

- cioè, se il giudice Ford avesse fatto le cose per bene, ci aspetteremmo di osservare 102 o meno donne in un campione di 350 potenziali giurati in meno di una volta ogni miliardo di miliardo di estrazioni!!!

- Ora, è chiaro che tutto può capitare, anche di estrarre solo 102 donne e anche di estrarne ancora di meno. Però il calcolo del *p-value* ci dice che un valore tanto o più estremo di quello ottenuto ce lo aspettiamo meno di una volta ogni miliardo di miliardo di estrazioni. Un po' troppo poco frequente per credere alle giustificazioni del giudice Ford!

Come andò a finire?

Il 10 Luglio 1968 la corte condannò il Dott. Spock a due anni di carcere e a 5000 dollari di multa. Il verdetto venne poi rovesciato dalla corte di Appello degli Stati Uniti nel 1969 per insufficienza di prove.

3.7 Ma quanti goal segna in media una squadra in trasferta?

Stima della media, intervalli di confidenza e verifica d'ipotesi nel caso di una distribuzione Poisson.

Risultati

La seguente tabella riporta i risultati della 32-esima giornata del campionato italiano di serie A 2006-2007.

Partita	Risultato	Goal fuori casa
Torino-Atalanta	1-2	2
Parma-Catania	1-1	1
Udinese-Chievo	2-1	1
Cagliari-Empoli	0-0	0
Ascoli-Lazio	2-2	2
Messina-Milan	1-3	3
Inter-Palermo	2-2	2
Livorno-Reggina	1-1	1
Roma-Sampdoria	4-0	0
Fiorentina-Siena	1-0	0

- In particolare, l'ultima colonna riporta i goal segnati dalla squadra che giocava in trasferta.

- Siamo interessati a stimare qual è il numero medio di goal segnati da una squadra in trasferta.

Un possibile modello

- Indichiamo con y_i il numero di goal segnati dalla squadra fuori casa nelle partita i-esima.

- Possiamo ipotizzare che il numero di goal segnati dalla squadra in trasferta nella i-esima partita sia una realizzazione di una variabile casuale Poisson con (ignota) media ϑ.

- La variabile casuale Poisson si presta a modellare situazioni in cui si conta quante volte si verifica un certo evento (il goal della squadra in trasferta) in un determinato intervallo di tempo (la durata di una partita di calcio):
$$Y_i \sim Poisson(\vartheta)\,,$$

con

$$\Pr(Y_i = y_i; \vartheta) = \frac{\vartheta^{y_i} e^{-\vartheta}}{y_i!}$$

e y_i che può essere un qualsiasi numero naturale (zero incluso).

- Inoltre, è abbastanza sensato assumere che il numero di goal segnati dalle squadre in trasferta in diverse partite siano rappresentati da variabili casuali indipendenti tra loro. Quindi possiamo pensare di avere un campione casuale semplice di numerosità $n = 10$ da una distribuzione Poisson con media ϑ e di voler usare questi dati per fare inferenza su ϑ.

Stima di ϑ

Essendo $\vartheta = \mathbb{E}(Y_i)$, uno stimatore "naturale" per ϑ è

$$\hat{\vartheta} = \frac{\sum_{i=1}^{n} Y_i}{n},$$

ovvero la media artimetica dei goal sengati fuori casa. $\hat{\vartheta}$ coincide con lo stimatore ottenuto con il metodo dei momenti.

Nel caso in esame, la stima è pari a $\hat{\vartheta} = \sum_{i=1}^{10} y_i/10 = 12/10 = 1{,}2$.

Ovviamente, se le Y_i sono variabili casuali, anche $\hat{\vartheta}$ è una variabile casuale. Lo studio della sua distribuzione è importante perché permette di acquisire una idea sulla dimensione dell'errore di stima (come abbiamo già visto nelle unità precendenti).

La distribuzione *esatta* di $\hat{\vartheta}$ è facile da determinare. Infatti, $\hat{\vartheta} \in \Theta_n = \{0/n, 1/n, 2/n, \ldots\}$ e, per qualsivoglia $a \in \Theta_n$, risulta

$$\Pr(\hat{\vartheta} = a) = \Pr\left(\sum_{i=1}^{n} Y_i = na\right) = \frac{(n\vartheta)^{na} e^{-n\vartheta}}{(na)!},$$

essendo la somma di v.c. Poisson (indipendenti) distribuita come una v.c. Poisson, con media uguale alla somma delle medie.

Approssimazione normale

Il fatto che la distribuzione esatta sia facile da determinare non implica che sia anche facile da maneggiare.

Anche in questo caso, la maniera più comoda per determinare intervalli di confidenza e test si basa sull'approssimazione normale ottenuta mediante il Teorema del Limite Centrale.

Infatti, essendo $\mathbb{E}(Y_i) = \mathbb{V}ar(Y_i) = \vartheta$, si ha

$$\begin{aligned} \mathbb{E}(\hat{\vartheta}) &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \vartheta\,, \\ \mathbb{V}ar(\hat{\vartheta}) &= \mathbb{V}ar\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \frac{\vartheta}{n}\,, \end{aligned}$$

e per n non troppo piccolo

$$\frac{\hat{\vartheta} - \vartheta}{\sqrt{\vartheta/n}} \overset{a}{\sim} N(0;1)\ .$$

L'approssimazione normale è tanto più accurata quanto più $n\vartheta$ è elevato. In questo caso, la bontà dell'approssimazione potrebbe essere motivo di preoccupazione ...

Distribuzione approssimata dell'errore di stima

Se $(\hat{\vartheta} - \vartheta)/\sqrt{\vartheta/n}$ è approssimativamente una normale standard allora

$$(\text{errore di stima}) = (\hat{\vartheta} - \vartheta) \overset{a}{\sim} N(0;\vartheta/n).$$

Si osservi che, a differenza di quanto accadeva all'errore di stima nell'unità 3.2, questa distribuzione, oltre ad essere approssimata è anche parzialmente ignota. Infatti, la varianza della distribuzione dipende dal vero valore di ϑ.

Per acquisire delle informazioni sulla dimensione dell'errore di stima possiamo stimarne la varianza sostituendo $\hat{\vartheta}$ a ϑ. Nel caso in esame troviamo

$$\widehat{\mathbb{V}ar}(\hat{\vartheta} - \vartheta) = \frac{\hat{\vartheta}}{n} \approx \frac{1{,}2}{10} \approx 0{,}12.$$

Approssimazione dopo approssimazione, siamo arrivati alla conclusione che l'errore di stima è, grossomodo, normale con media zero e varianza 0,12.

Intervalli di confidenza

La distribuzione stimata per $\hat{\vartheta} - \vartheta$ può essere usata per costruire intervalli di confidenza (almeno approssimatamente) di livello $1 - \alpha$ prefissato.

Infatti se la distribuzione di $\hat{\vartheta} - \vartheta$ è approssimatamente una normale di media nulla e varianza 0,12 allora possiamo scrivere (perché?)

$$\Pr(-\sqrt{0{,}12} \times z_{1-\alpha/2} \leq \hat{\vartheta} - \vartheta \leq \sqrt{0{,}12} \times z_{1-\alpha/2}) \approx 1 - \alpha \qquad (3.9)$$

dove, al solito, con z_p indichiamo il quantile p-simo di una normale standard. La (3.9) può essere scritta come

$$\Pr(\hat{\vartheta} - \sqrt{0{,}12} \times z_{1-\alpha/2} \leq \vartheta \leq \hat{\vartheta} + \sqrt{0{,}12} \times z_{1-\alpha/2}) \approx 1 - \alpha$$

ovvero, ci mostra, ricordando come avevamo calcolato lo scarto quadratico medio dell'errore di stima, che

$$\left[\hat{\vartheta} - z_{1-\alpha/2}\sqrt{\frac{\hat{\vartheta}}{n}}, \hat{\vartheta} + z_{1-\alpha/2}\sqrt{\frac{\hat{\vartheta}}{n}}\right]$$

costituisce un intervallo di confidenza di livello approssimato $1-\alpha$ per ϑ.

Con i nostri dati

Supponiamo di voler calcolare un intervallo di confidenza di livello 0,9.

Allora $\alpha = 0{,}1$ e $1-\alpha/2 = 0{,}95$. Da una tavola della distribuzione normale (o utilizzando un programma appropriato) troviamo che $z_{0,95} = 1{,}65$. Sappiamo già che $\hat{\vartheta} = 1{,}2$ e che

$$\sqrt{\frac{1{,}2}{10}} = 0{,}3464.$$

Quindi, la semi-ampiezza dell'intervallo richiesto è $1{,}65 \times 0{,}3464 \approx 0{,}5716$. Perciò l'intervallo stesso è

$$[1{,}2 - 0{,}5716\ ;\ 1{,}2 + 0{,}5716] = [0{,}6284\ ;\ 1{,}772].$$

Esercizio 21. *Sarebbe ragionevole affermare che una squadra in trasferta segna in media un goal a partita?*

Ma 2 goal a partita sono troppi?

Sulla base dell'intervallo di confidenza al 90% si evince che una squadra in trasferta segna in media meno di due goal a partita.

Ma qual è l'evidenza in favore dell'ipotesi che il vero numero medio di goal per una squadra in trasferta sia 2 (contro l'alternativa che sia inferiore a 2)?

Per rispondere a questo quesito dobbiamo considerare il sistema d'ipotesi seguente:

$$\begin{cases} \mathrm{H}_0 : \vartheta = \vartheta_0 \\ \mathrm{H}_1 : \vartheta < \vartheta_0 \end{cases}$$

con $\vartheta_0 = 2$. [6]

[6] Il sistema è equivalente a

$$\begin{cases} \mathrm{H}_0 : \vartheta \geq \vartheta_0 \\ \mathrm{H}_1 : \vartheta < \vartheta_0 \end{cases}$$

Si noti che come nell'unità 3.5 il sistema d'ipotesi ha una alternativa di tipo **unilaterale**.

Una possibile *statistica test* è

$$T = \frac{\hat{\vartheta} - \vartheta_0}{\sqrt{\vartheta_0/n}}$$

Se l'ipotesi nulla è vera, per quanto ricordato a pagina 221, T ha una distribuzione approssimativamente normale di media zero e varianza 1.

Quindi possiamo confrontare il valore di T calcolato dai dati con questa distribuzione. Valore negativi di T 'piccoli' ci indicheranno che H_0 è implausibile (rispetto ad H_1)

Allora la regione di rifiuto sarà data dai valori del test minori di z_α, poiché fissiamo ad α la probabilità di commettere un errore di I tipo (e cioè di rifiutare H_0 quando è la vera ipotesi)

$$\Pr(T \leq z_\alpha \quad \text{quando } \mathrm{H}_0 \text{ è vera}) \approx \alpha\,.$$

Nel nostro caso, se $\alpha = 0{,}05$, $z_{0{,}95} \approx 1{,}65$ e, poiché

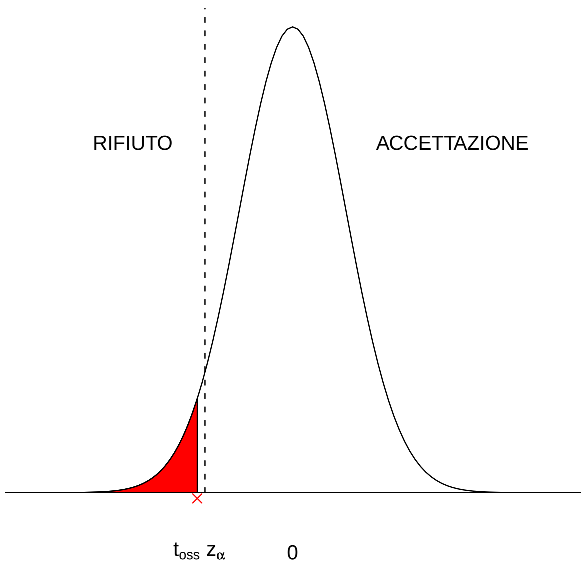
$$t_{oss} = (1{,}2 - 2)/\sqrt{2/10} = -1{,}789$$

rifiutiamo l'ipotesi nulla in favore dell'ipotesi alternativa.

Il livello di significatività osservato è pari a

$$\begin{aligned} \alpha_{oss} = \Pr(T \leq t_{oss} \quad \text{quando } \mathrm{H}_0 \text{ è vera}) &\approx \Pr(N(0;1) \leq -1{,}789) \\ &\approx 0{,}03682 \end{aligned}$$

ed è, coerentemente, minore di 0,05. Ma notiamo che se avessimo fissato $\alpha = 0{,}01$ avremmo accettato H_0.


RIFIUTO
ACCETTAZIONE
t_{oss} z_α
0

3.8 Sindacato o non sindacato?

Confronto delle medie di due popolazioni normali

Il sindacato fa guadagnare di più le donne?

"*Women who are union members earn* 2,50$ *per hour more than women who are not union members*".

(The Wall Street Journal, July 26, 1994).

Sembrerebbe quindi che per le donne statunitensi sia conveniente far parte di un sindacato.

Per verificare l'affermazione del *Wall Street Journal* abbiamo scelto due campioni indipendenti di lavoratrici del settore industriale.

Il primo campione è costituito da 15 lavoratrici iscritte ad un sindacato, mentre il secondo campione è costituito da 20 lavoratrici che non fanno parte di nessun sindacato.

Per ciascuna unità statistica (la lavoratrice) abbiamo misurato il salario orario (in $).

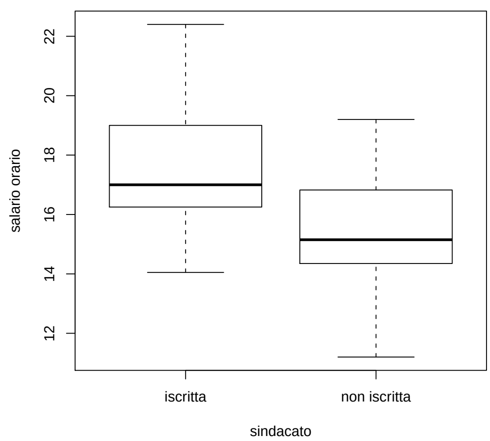
I dati

iscritte ad un sindacato:

22,40 18,90 16,70 14,05 16,20 20,00 16,10 16,30 19,10 16,50 18,50 19,80 17,00 14,30 17,20

non iscritte ad un sindacato:

17,60 14,40 16,60 15,00 17,65 15,00 17,55 13,30 11,20 15,90 19,20 11,85 16,65 15,20 15,30 17,00 15,10 14,30 13,90 14,50



Formalizzando...

Guardando il grafico precedente, sembrerebbe che, in media, il salario orario delle lavoratrici iscritte ad un sindacato sia in effetti maggiore di quello delle lavoratrici non iscritte.

Proviamo a vedere se i dati osservati ci permettono di generalizzare questa affermazione alle rispettive popolazioni di riferimento.

Supponiamo che:

- il salario orario di una lavoratrice appartenente ad un sindacato si possa rappresentare con una v.c.

$$Y_1 \sim N(\mu_1; \sigma_1^2)\,;$$

- il salario orario di una lavoratrice non iscritta al sindacato si possa rappresentare con una v.c.

$$Y_2 \sim N(\mu_2; \sigma_2^2)\,;$$

- le varianze di Y_1 e di Y_2 siano uguali: $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Questa assunzione sembra "sensata" alla luce dei dati osservati (perché?)

Stima delle medie

- Le medie delle due v.c. sono supposte diverse. Inoltre, i due campioni sono indipendenti.

- Sembra ragionevole stimare μ_1 utilizzando solo le $n_1 = 15$ osservazioni relative alle iscritte ad un sindacato e stimare μ_2 utilizzando le $n_2 = 20$ osservazioni provenienti dalle lavoratrici non appartenenti ad un sindacato:

$$\hat{\mu}_1 = \bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} = 17{,}54, \quad \hat{\mu}_2 = \bar{y}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i} = 15{,}36\,.$$

- Sappiamo inoltre che i relativi stimatori sono tali che

$$\begin{aligned} \bar{Y}_1 &\sim N\left(\mu_1; \frac{\sigma^2}{n_1}\right) \\ \bar{Y}_2 &\sim N\left(\mu_2; \frac{\sigma^2}{n_2}\right) \\ \bar{Y}_1 - \bar{Y}_2 &\sim N\left(\mu_1 - \mu_2; \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right) \end{aligned}$$

  L'ultima relazione deriva dall'indipendenza dei due campioni.

- Per fare inferenza su $\mu_1 - \mu_2$ sembra sensato lavorare con la differenza $\bar{Y}_1 - \bar{Y}_2$.

Stima della comune varianza

- Similmente a quanto visto nell'unità 3.3, dobbiamo comunque provvedere a stimare l'ignota varianza σ^2.

- Se utilizziamo solo il primo campione, una stima di σ^2 è data da

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 \,.$$

Analogamente, se utilizziamo solo il secondo campione, otteniamo

$$s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2 \,.$$

- Ma σ^2 è lo stesso nei due campioni e quindi conviene fornire un'unica stima che utilizzi entrambi i campioni (due campioni sono meglio di uno ...):

$$\begin{aligned} s^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2}{n_1 + n_2 - 2} = 4{,}41 \,. \end{aligned}$$

- Si potrebbe dimostrare che il relativo stimatore S^2 è uno stimatore non distorto di σ^2.

Verifica d'ipotesi

- Proviamo a verificare l'ipotesi (H_0) che i due gruppi di lavoratrici abbiano lo stesso salario orario medio (ossia che μ_1 si uguale a μ_2), contro l'ipotesi alternativa (H_1) che il salario orario medio delle iscritte al sindacato sia più elevato (ossia che μ_1 sia maggiore di μ_2)

$$\begin{cases} \mathrm{H}_0 : \mu_1 = \mu_2 \\ \mathrm{H}_1 : \mu_1 > \mu_2 \end{cases}$$

Questo sistema d'ipotesi è ovviamente equivalente al seguente

$$\begin{cases} \mathrm{H}_0 : \mu_1 - \mu_2 = 0 \\ \mathrm{H}_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

- Consideriamo la statistica test

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - 0}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  che, sotto H_0, ha distribuzione $t_{n_1+n_2-2}$ (t di Student con $n_1 + n_2 - 2$ gradi di libertà).

- Valori (positivi) grandi del test indicano evidenza contro H_0 in favore di H_1.

Regione di rifiuto

- La regione di rifiuto di livello α è determinata dalla condizione

$$\Pr(T > t_{n_1+n_2-2;1-\alpha}, \quad \text{quando } \mathrm{H}_0 \text{ è vera}) = \alpha$$

  Ad esempio, se $\alpha = 0{,}05$, $t_{33;0{,}95} = 1{,}692$ e quindi rifiutiamo per valori di t_{oss} maggiori di 1,692.

- Nel nostro caso, il valore osservato del test è pari a

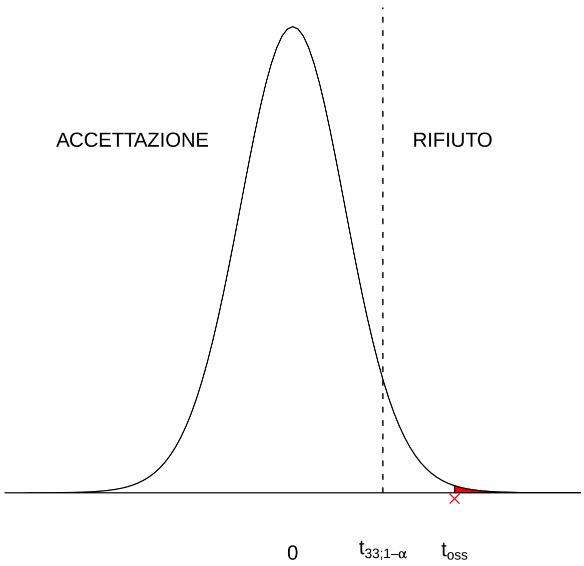
$$t_{oss} = \frac{17{,}54 - 15{,}36}{\sqrt{4{,}41\left(\frac{1}{15} + \frac{1}{20}\right)}} = 3{,}04$$

  e quindi rifiutiamo H_0 in favore di H_1 (a livello fissato 0,05).

- L'α_{oss} è dato da

$$\begin{aligned} \alpha_{oss} &= \Pr(T > t_{oss}, \quad \text{quando } \mathrm{H}_0 \text{ è vera}) \\ &= \Pr(t \text{ con 33 gradi di libertà} > 3{,}04) \\ &< 0{,}005 \quad \text{(dalle tavole\dots)} \end{aligned}$$

  quindi avrei rifiutato H_0 anche con un livello fissato pari a $\alpha = 0{,}01$.



Un intervallo di confidenza

Possiamo anche calcolare un intervallo di confidenza per la differenza delle medie $\mu_1 - \mu_2$ e cioè per la differenza di paga media oraria tra le iscritte e le non iscritte ad un sindacato.

Come fatto nelle unità precedenti otteniamo che (ponendo per brevità $g = n_1 + n_2 - 2$)

$$\Pr\left(-t_{g;1-\alpha/2} \leq \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \leq t_{g;1-\alpha/2}\right) = 1 - \alpha$$

e, con i soliti passaggi, l'intervallo di confidenza di livello $1 - \alpha$ risulta pari a

$$(\mu_1 - \mu_2) \in \left(\bar{y}_1 - \bar{y}_2 \mp t_{g;1-\alpha/2}\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Nel nostro caso, con $1 - \alpha = 0{,}95$, otteniamo l'intervallo $(0{,}718, 3{,}635)$. Quindi, una differenza di 2,50$ è plausibile.

3.9 Economia a Venezia non è così male ...

Confronto della probabilità di successo in due campioni indipendenti, intervalli di confidenza e verifica d'ipotesi nel caso di due binomiali.

I dati

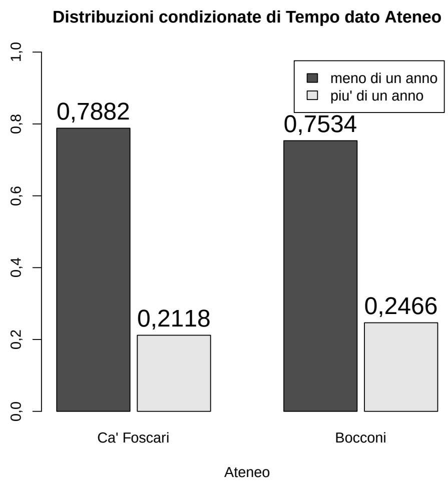
La seguente tabella classifica i laureati nel 2001 in Economia a Ca' Foscari e alla Bocconi rispetto al tempo intercorso (meno o più di un anno) tra la laurea e il primo lavoro continuativo iniziato dopo la laurea (fonte: ISTAT 2004):

Tempo	Ateneo		
	Ca' Foscari	*Bocconi*	Totale
meno di un anno	480	1338	**1818**
più di un anno	129	438	**567**
Totale	**609**	**1776**	**2385**

La tabella riporta la distribuzione congiunta di due variabili qualitative dicotomiche (Tempo: *meno di un anno*, *più di un anno*; Ateneo: *Ca' Foscari*, *Bocconi*) rilevate su 2385 unità statistiche (i laureati nel 2001 presso i due atenei).

Chi trova lavoro prima?

La distribuzione della variabile Tempo condizionata alle due modalità della variabile Ateneo mostra che la percentuale di laureati che trovano lavoro entro il primo anno dalla laurea è leggermente superiore per i laureati a Ca' Foscari rispetto a quelli laureati alla Bocconi. Ma allora i "veneziani" trovano lavoro prima?



Un possibile modello

Supponiamo, analogamente a quanto fatto nell'Unità 3.5, di rappresentare il numero di laureati a Ca' Foscari che trova lavoro entro un anno (sul totale degli n_1 laureati) con una variabile binomiale:

$$Y_1 \sim \mathcal{B}(n_1, \vartheta_1)$$

Analogamente, per i laureati alla Bocconi si può considerare una variabile casuale Y_2, indipendente da Y_1, tale che:

$$Y_2 \sim \mathcal{B}(n_2, \vartheta_2)\,.$$

Ricordiamo che questa formulazione è equivalente a rappresentare un singolo laureato a Ca' Foscari con una variabile casuale $X_{1i} \sim \mathcal{B}(1, \vartheta_1)$, $i = 1, \ldots, n_1$, che assume valore 1 se ha trovato lavoro entro l'anno e 0 altrimenti. Quindi $Y_1 = \sum_{i=1}^{n_1} X_{1i}$. Lo stesso discorso vale per i laureati alla Bocconi. Attraverso i dati a disposizione vogliamo fare delle affermazioni sulla diversità (o meno) delle probabilità dei laureati di trovare lavoro entro un anno nei due atenei, ossia sulla differenza $\vartheta_1 - \vartheta_2$.

Stimatori e stime dei parametri

Come già visto nell'Unità 3.5, lo stimatore più sensato di ϑ_1 è

$$\hat{\vartheta}_1 = \frac{Y_1}{n_1}\,.$$

ovvero la proporzione di laureati a Ca' Foscari che ha trovato lavoro entro un anno dalla laurea.

Analogamente, per ϑ_2, si ha

$$\hat{\vartheta}_2 = \frac{Y_2}{n_2}\,.$$

Le relative stime ottenute attraverso i dati a disposizione sono rispettivamente $\hat{\vartheta}_1 = y_1/n_1 = 480/609 = 0{,}7882$ e $\hat{\vartheta}_2 = y_2/n_2 = 1338/1776 = 0{,}7534$.

Dalle stime risulta quindi che esiste una probabilità maggiore per i laureati a Venezia di trovare lavoro entro il primo anno dalla laurea: $\hat{\vartheta}_1 - \hat{\vartheta}_2 = 0{,}7882 - 0{,}7534 = 0{,}0348$. Tuttavia, per vedere se questa differenza sia o meno significativa, bisogna considerare la distribuzione dello stimatore utilizzato, ossia di $\hat{\vartheta}_1 - \hat{\vartheta}_2 = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}$.

Distribuzione approssimata di $\hat{\vartheta}_1 - \hat{\vartheta}_2$

Essendo Y_j $(j = 1, 2)$ variabili casuali anche $\hat{\vartheta}_j$ sono variabili casuali, la cui distribuzione esatta è nota (si veda l'Unità 3.5).

Tuttavia per determinare intervalli di confidenza e test si preferisce basarsi sulla distribuzione approssimata degli stimatori, ottenuta usando il Teorema del Limite Centrale (si veda pagina 221):

$$\hat{\vartheta}_j \overset{a}{\sim} N\left(\vartheta_j; \frac{\vartheta_j(1-\vartheta_j)}{n_j}\right), j = 1, 2 \,.$$

Si osservi che $\hat{\vartheta}_j$ è uno stimatore non distorto e consistente di ϑ_j (perché?), e con distribuzione approssimata normale.

Inoltre, essendo Y_1 e Y_2 indipendenti, anche $\hat{\vartheta}_1$ e $\hat{\vartheta}_2$ risultano essere indipendenti. Quindi (perché?)

$$\hat{\vartheta}_1 - \hat{\vartheta}_2 \overset{a}{\sim} N\left(\vartheta_1 - \vartheta_2; \frac{\vartheta_1(1-\vartheta_1)}{n_1} + \frac{\vartheta_2(1-\vartheta_2)}{n_2}\right) \,.$$

Distribuzione approssimata di $\hat{\vartheta}_1 - \hat{\vartheta}_2$ (cont.)

La distribuzione di $\hat{\vartheta}_1 - \hat{\vartheta}_2$ oltre ad essere approssimata è anche parzialmente ignota. Infatti, la varianza della distribuzione dipende dal vero valore di ϑ_1 e ϑ_2 (in realtà anche la media, ma questo non è un problema. . .).

Possiamo stimare la varianza dello stimatore sostituendo $\hat{\vartheta}_j$ a ϑ_j (j=1,2). Nel nostro caso troviamo

$$\begin{aligned} \widehat{\mathbb{V}ar}(\hat{\vartheta}_1 - \hat{\vartheta}_2) &= \frac{\hat{\vartheta}_1(1-\hat{\vartheta}_1)}{n_1} + \frac{\hat{\vartheta}_2(1-\hat{\vartheta}_2)}{n_2} \\ &\approx \frac{0{,}7882(1-0{,}7882)}{609} + \frac{0{,}7534(1-0{,}7534)}{1776} \\ &\approx (0{,}0195)^2 \,. \end{aligned}$$

ovvero, approssimazione dopo approssimazione, siamo arrivati alla conclusione che l'errore di stima è, grossomodo, normale di media zero e scarto quadratico medio 0,0195.

Intervalli di confidenza

La distribuzione stimata per $\hat{\vartheta}_1 - \hat{\vartheta}_2$ può essere usata per costruire intervalli di confidenza (almeno approssimatamente) di livello $1 - \alpha$ prefissato.

Infatti possiamo scrivere (perché?)

$$\Pr\left(-z_{1-\alpha/2} \leq \frac{\hat{\vartheta}_1 - \hat{\vartheta}_2 - (\vartheta_1 - \vartheta_2)}{0.0195} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha$$

dove, al solito, con z_p indichiamo il quantile p-simo di una normale standard. La relazione precedente può essere scritta come

$$\Pr(\hat{\vartheta}_1 - \hat{\vartheta}_2 - 0{,}0195 \times z_{1-\alpha/2} \leq \vartheta_1 - \vartheta_2 \leq \hat{\vartheta}_1 - \hat{\vartheta}_2 + 0{,}0195 \times z_{1-\alpha/2}) \approx 1 - \alpha$$

ovvero, ci mostra, ricordando come avevamo calcolato lo scarto quadratico medio dell'errore di stima, che

$$\left[\hat{\vartheta}_1 - \hat{\vartheta}_2 \mp z_{1-\alpha/2} \sqrt{\frac{\hat{\vartheta}_1(1-\hat{\vartheta}_1)}{n_1} + \frac{\hat{\vartheta}_2(1-\hat{\vartheta}_2)}{n_2}} \right]$$

costituisce un intervallo di confidenza di dimensione approssimata $1 - \alpha$ per $\vartheta_1 - \vartheta_2$.

Con i nostri dati

Supponiamo di voler calcolare un intervallo di confidenza di livello 0,95.

Allora, $\alpha = 0{,}05$, $1 - \alpha/2 = 0{,}975$. Da una tavola della distribuzione normale (o utilizzando un programma appropriato) troviamo che $z_{0,975} \approx 1{,}96$. Sappiamo già che $\hat{\vartheta}_1 - \hat{\vartheta}_2 = 0{,}7882 - 0{,}7534 = 0{,}0348$ e che

$$\sqrt{\frac{0{,}7882 \times (1 - 0{,}7882)}{609} + \frac{0{,}7534(1 - 0{,}7534)}{1776}} \approx 0{,}0195 \ .$$

Quindi, la semi-ampiezza dell'intervallo richiesto è $1{,}96 \times 0{,}0195 = 0{,}0382$.

Perciò l'intervallo stesso è

$$[0{,}0348 - 0{,}0382 \ ; \ 0{,}0348 + 0{,}0382] = [-0{,}0034 \ ; \ 0{,}0730].$$

Cosa concludete sulla base di questo intervallo di confidenza?

Ma quindi Venezia non è meglio! Calma... una verifica d'ipotesi unilaterale

Sulla base dell'intervallo di confidenza al 95% (che include sia valori positivi che negativi) sembra non esserci una differenza significativa tra Ca' Foscari e la Bocconi (ricordo che stiamo guardando solo la percentuale di occupati ad un anno dalla laurea!).

Proviamo però a considerare una verifica d'ipotesi unilaterale

$$\left\{ \begin{array}{l} \mathrm{H}_0 : \vartheta_1 = \vartheta_2 \\ \mathrm{H}_1 : \vartheta_1 > \vartheta_2 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \mathrm{H}_0 : \vartheta_1 - \vartheta_2 = 0 \\ \mathrm{H}_1 : \vartheta_1 - \vartheta_2 > 0 \end{array} \right.$$

ossia

$$\left\{ \begin{array}{l} \mathrm{H}_0 : \text{Ca' Foscari} = \text{Bocconi} \\ \mathrm{H}_1 : \text{Ca' Foscari meglio della Bocconi} \end{array} \right.$$

Una statistica test

Se l'ipotesi nulla è vera, la distribuzione approssimata della seguente statistica test è normale standard

$$\frac{\hat{\vartheta}_1 - \hat{\vartheta}_2 - 0}{\sqrt{\widehat{\mathbb{V}ar}(\hat{\vartheta}_1 - \hat{\vartheta}_2)}}\,.$$

Notando però che, sotto H_0, Y_1 e Y_2 hanno la stessa distribuzione ($\vartheta_1 = \vartheta_2 = \vartheta$), si ha

$$\mathbb{V}ar_{\mathrm{H}_0}(\hat{\vartheta}_1 - \hat{\vartheta}_2) = \frac{\vartheta(1-\vartheta)}{n_1} + \frac{\vartheta(1-\vartheta)}{n_2} = \frac{(n_1+n_2)\vartheta(1-\vartheta)}{n_1 n_2}\,.$$

Inoltre, sempre sotto H_0, una stima di ϑ è data da

$$\hat{\vartheta} = \frac{y_1 + y_2}{n_1 + n_2} = \frac{n_1\hat{\vartheta}_1 + n_2\hat{\vartheta}_2}{n_1 + n_2} = 0{,}7623\,.$$

Quindi, una stima di $\mathbb{V}ar(\hat{\vartheta}_1 - \hat{\vartheta}_2)$, sotto H_0, è quindi pari a

$$\begin{aligned}\widehat{\mathbb{V}ar}_{\mathrm{H}_0}(\hat{\vartheta}_1 - \hat{\vartheta}_2) &= \frac{(n_1+n_2)\hat{\vartheta}(1-\hat{\vartheta})}{n_1 n_2} \\ &= \frac{2385 \cdot 0{,}7623 \cdot (1 - 0{,}7623)}{609 \cdot 1776} = (0{,}01999)^2\,.\end{aligned}$$

Una statistica test (cont.)

Per verificare l'uguaglianza di ϑ_1 e ϑ_2 utilizziamo quindi la statistica test

$$T = \frac{\hat{\vartheta}_1 - \hat{\vartheta}_2 - 0}{\sqrt{\widehat{\mathbb{V}ar}_{\mathrm{H}_0}(\hat{\vartheta}_1 - \hat{\vartheta}_2)}}\,.$$

che, sotto l'ipotesi nulla, ha distribuzione approssimata normale standard.

Un test di livello fissato α, rifiuterà H_0 in favore di H_1 per valori grandi della statistica test e precisamente la regione di rifiuto sarà data dai valori del test maggiori di $z_{1-\alpha}$, poichè fissiamo ad α la probabilità di commettere un errore di I tipo (e cioè di rifiutare H_0 quando è la vera ipotesi)

$$\Pr(T \geq z_{1-\alpha} \quad \text{quando } \mathrm{H}_0 \text{ è vera}) \approx \alpha\,.$$

Il livello di significatività osservato è invece

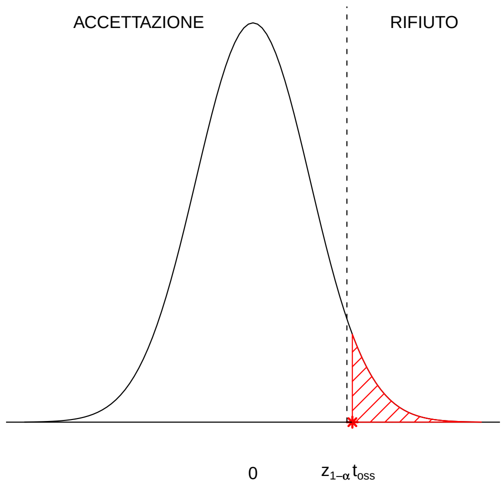
$$\alpha_{oss} = \Pr(T \geq t_{oss} \quad \text{quando } \mathrm{H}_0 \text{ è vera}) \approx \Pr(N(0;1) \geq t_{oss})\,.$$

Con i nostri dati

Nel nostro caso si ha $t_{oss} = 0{,}0348/0{,}01999 = 1{,}7409$. Quindi, se $\alpha = 0{,}05$, $t_{oss} = 1{,}7409 > z_{1-\alpha} = 1{,}645$, e quindi rifiutiamo l'ipotesi nulla in favore dell'ipotesi alernativa.

Il livello di significatività osservato è pari a

$$\begin{aligned} \alpha_{oss} &= \Pr(T \geq t_{oss} \quad \text{quando H}_0 \text{ è vera}) \\ &\approx \Pr(N(0;1) \geq 1{,}7409) = 0{,}0409. \end{aligned}$$



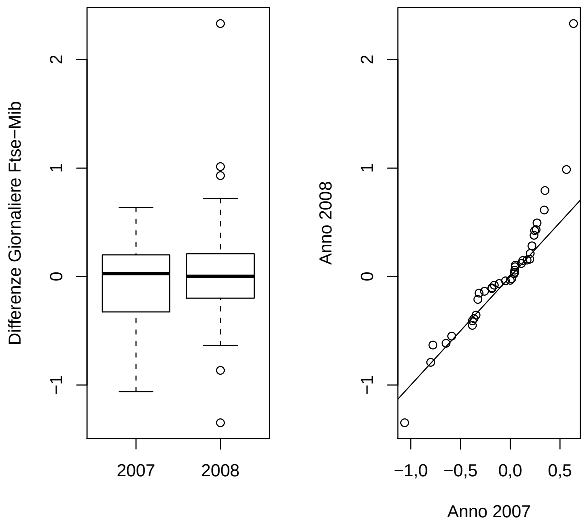
Ma allora Venezia è meglio!

Esercizio 22. *Avremmo rifiutato H_0 ad un livello fissato $\alpha = 0{,}01$?*

3.10 Ancora crisi nel mercato?

Intervalli di confidenza e verifica d'ipotesi per il rapporto di due varianze nel caso di due campioni tratti da una v.c. normale.

Differenze FTSE MIB



Campione ($n_{2007} = 38$ e $n_{2008} = 50$) di differenze dei valori giornalieri di chiusura aggiustata dell'indice FTSE MIB, in migliaia di Euro per gli anni 2007 e 2008. Fonte: it.finance.yahoo.com

Una possibile formulazione del problema

- Dai grafici precedenti possiamo concludere, in prima approssimazione, che i due campioni di differenze giornaliere dell'indice possono essere assimilati a osservazioni da due variabili normali di media e varianza ignote e quindi

$$Y_i \sim N(\mu_i, \sigma_i^2) \text{ con } i = 2007, 2008;$$

- Possiamo costruire un test per saggiare l'ipotesi:

$$\begin{cases} \mathrm{H}_0 : \sigma^2_{2007} = \sigma^2_{2008} \\ \mathrm{H}_1 : \sigma^2_{2007} \neq \sigma^2_{2008} \end{cases}$$

- Possiamo pensare di costruire un intervallo di confidenza per il rapporto delle varianze;

Stima dei parametri ignoti

Sappiamo già che dato un campione $y_1, y_2, \cdots, y_n$ una stima dei parametri della distribuzione normale, media e varianza è ottenibile con il metodo dei momenti

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})$$

Per i nostri due campioni abbiamo

	2007	2008
$\hat{\mu}$	-0,073	0,03
s^2	0,142	0,301

Per ottenere una statistica test possiamo utilizzare la distribuzione del rapporto tra i due stimatori, cioè S^2_{2007}/S^2_{2008}.

Distribuzione di S_x^2/S_y^2

Dati due campioni $x_1, x_2, \cdots, x_{n_x}$ e $y_1, y_2, \cdots, y_{n_y}$ con $X_i \sim N(\mu_x, \sigma_x^2)$ e $Y_i \sim N(\mu_y, \sigma_y^2)$ sappiamo che la distribuzione *esatta* di S_k^2 $(k = x, y)$ è

$$(n_k - 1)\frac{S_k^2}{\sigma_k^2} \sim \chi^2_{n_k - 1} \ .$$

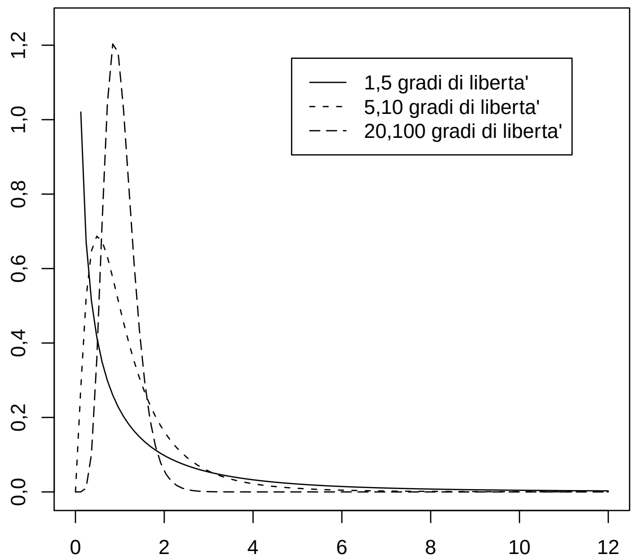
da cui abbiamo

$$\frac{\frac{S_x^2}{\sigma_x^2}}{\frac{S_y^2}{\sigma_y^2}} = \frac{S_x^2}{S_y^2}\,\frac{\sigma_y^2}{\sigma_x^2} \sim F_{n_x-1, n_y-1} \ .$$

dove F_{n_x-1,n_y-1} è la variabile casuale F di Fisher–Snedecor con $n_x - 1$ gradi di libertà al numeratore e $n_y - 1$ gradi di libertà al denominatore.

La variabile casuale F

La variabile casuale F ha una distribuzione che dipende da due parametri, chiamati, rispettivamente i gradi di libertà del numeratore e i gradi di libertà del denominatore.



Quantili di una variabile casuale F

	35	36	37	38	39	47	48	49	50	51
35	1,757	1,752	1,748	1,743	1,739	1,711	1,709	1,706	1,703	1,701
36	1,748	1,743	1,738	1,734	1,730	1,702	1,699	1,696	1,694	1,691
37	1,739	1,734	1,730	1,725	1,721	1,693	1,690	1,687	1,685	1,682
38	1,731	1,726	1,721	1,717	1,712	1,684	1,681	1,679	1,676	1,673
39	1,723	1,718	1,713	1,709	1,704	1,676	1,673	1,670	1,668	1,665
47	1,672	1,667	1,662	1,657	1,653	1,624	1,621	1,618	1,615	1,612
48	1,667	1,662	1,657	1,652	1,648	1,618	1,615	1,612	1,610	1,607
49	1,662	1,657	1,652	1,647	1,643	1,613	1,610	1,607	1,604	1,602
50	1,657	1,652	1,647	1,642	1,638	1,608	1,605	1,602	1,599	1,597
51	1,653	1,647	1,642	1,638	1,633	1,604	1,601	1,598	1,595	1,592

Tavola per $\alpha = 0{,}05$. Per ogni coppia di r_1 (colonna) e r_2 (riga), la tavola fornisce il quantile f_α di ordine α corrispondente. I quantili inferiori della distribuzione F di Fisher–Snedecor si possono determinare tramite la relazione $f_{r_1,r_2,1-\alpha} = 1/f_{r_2,r_1,\alpha}$. Ad esempio $f_{37,49,0{,}05} = 1/f_{49,37,0{,}95} = 1/1,687 = 0,593$.

Intervalli di confidenza

La distribuzione per S_x^2/S_y^2 può essere usata per costruire intervalli di confidenza di livello $1-\alpha$ prefissato per il rapporto tra le varianze.

$$\Pr\left(F_{n_x-1,n_y-1,\alpha/2} \leq \frac{\frac{S_x^2}{\sigma_x^2}}{\frac{S_y^2}{\sigma_y^2}} \leq F_{n_x-1,n_y-1,1-\alpha/2}\right) = 1-\alpha$$
$$\Pr\left(F_{n_x-1,n_y-1,\alpha/2}\frac{S_y^2}{S_x^2} \leq \frac{\sigma_y^2}{\sigma_x^2} \leq F_{n_x-1,n_y-1,1-\alpha/2}\frac{S_y^2}{S_x^2}\right) = 1-\alpha$$
$$\Pr\left(\frac{\frac{S_x^2}{S_y^2}}{F_{n_x-1,n_y-1,1-\alpha/2}} \leq \frac{\sigma_x^2}{\sigma_y^2} \leq \frac{\frac{S_x^2}{S_y^2}}{F_{n_x-1,n_y-1,\alpha/2}}\right) = 1-\alpha$$

dove con $F_{n,m,p}$ indichiamo il quantile p-esimo di una F con n, m gradi di libertà.

Ovvero, ci mostra che

$$\left[\frac{s_x^2/s_y^2}{F_{n_x-1,n_y-1,1-\alpha/2}}\ ;\ \frac{s_x^2/s_y^2}{F_{n_x-1,n_y-1,\alpha/2}}\right]$$

costituisce un intervallo di confidenza di livello $1-\alpha$ per σ_x^2/σ_y^2.

Con i nostri dati

Supponiamo di voler calcolare un intervallo di confidenza di livello 0,9.

Allora $\alpha = 0{,}1$, $\alpha/2 = 0{,}05$ e $1-\alpha/2 = 0{,}95$. Da una tavola della distribuzione $F_{37,49}$ (o utilizzando un programma appropriato) troviamo che $f_{37,49,0,05} = 1/f_{49,37,0,95} = 0,593$ e $f_{37,49,0,95} = 1,652$. Sappiamo che $s^2_{2007} = 0,142$, $s^2_{2008} = 0,301$ e perciò l'intervallo è

$$\left[\frac{0,142/0,301}{1,652}\ ;\ \frac{0,142/0,301}{0,593}\right] = [0,287\ ;\ 0,799].$$

da cui si vede che il valore 1 non è incluso nell'intervallo. In maniera analoga si costruisco intervalli unilaterali.

È proprio crisi dei mercati?

Sulla base dell'intervallo di confidenza al 90% si evince che la variabilità nel 2007 è stata diversa dalla variabilità del mercato nel 2008. Vogliamo verificare se la variabilità nel 2008 risulta superiore a quella del 2007. Per rispondere a questo quesito dobbiamo considerare il sistema d'ipotesi seguente:

$$\begin{cases} \mathrm{H}_0 : \sigma^2_{2007} = \sigma^2_{2008} \\ \mathrm{H}_1 : \sigma^2_{2007} > \sigma^2_{2008} \end{cases}$$

[7]

Una possibile *statistica test* è

$$T = \frac{S_x^2}{S_y^2}\,\frac{\sigma_x^2}{\sigma_y^2} \overset{H_0}{=} \frac{S_x^2}{S_y^2}$$

Se l'ipotesi nulla è vera, per quanto ricordato a pagina 238, T ha una distribuzione F_{n_x-1,n_y-1}.

Quindi possiamo confrontare il valore di T calcolato dai dati con questa distribuzione. Valori 'piccoli' di T ci indicheranno che H_0 è plausibile (rispetto ad H_1).

Allora la regione di rifiuto sarà data dai valori del test maggiori di $f_{n_x-1,n_y-1,1-\alpha}$, poiché fissiamo ad α la probabilità di commettere un errore di I tipo (e cioè di rifiutare H_0 quando è la vera ipotesi)

$$\Pr(T \geq f_{n_x-1,n_y-1,1-\alpha} \quad \text{quando } \mathrm{H}_0 \text{ è vera}) = \alpha\,.$$

Nel nostro caso, se $\alpha = 0{,}05$, $f_{37,49,0,95} = 1,652$ e, poiché

$$t_{oss} = \frac{0,142}{0,301}\,1 = 0,473$$

rifiutiamo l'ipotesi nulla in favore dell'ipotesi alternativa.

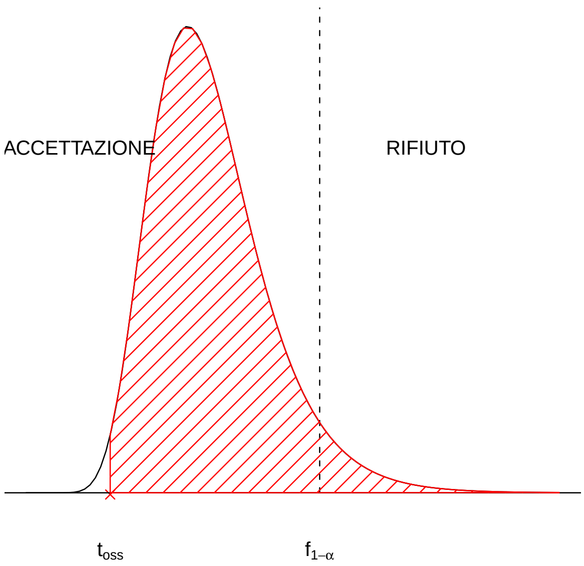
Il livello di significatività osservato è pari a

$$\begin{aligned}\alpha_{oss} = \Pr(T \geq t_{oss} \quad \text{quando } \mathrm{H}_0 \text{ è vera}) &= \Pr(F_{37,49} \geq 0,473)\\ &= 0,99\end{aligned}$$

ed è, coerentemente, maggiore di 0,05.

[7] Il sistema è equivalente a

$$\begin{cases} \mathrm{H}_0 : \sigma^2_{2007}/\sigma^2_{2008} \leq 1 \\ \mathrm{H}_1 : \sigma^2_{2007}/\sigma^2_{2008} > 1 \end{cases}$$



E nel caso la media sia nota?

Se $Y \sim N(\mu_y, \sigma_y^2)$ e la media è nota allora il metodo dei momenti ci fornisce il seguente stimatore per la varianza (lo studente verifichi che è così)

$$\hat{\sigma}_y^2 = \frac{1}{n_y} \sum_{i=1}^{n_y} (Y_i - \mu_y)^2$$

che è uno stimatore non distorto e consistente di σ_y^2. Allora

$$\frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \sim F^2_{n_x, n_y} \ .$$

Utilizzando questa statistica si ottengono intervalli di confidenza e test statistici procedendo nella stessa maniera del caso in cui la media sia ignota.

3.11 Telefonia mobile

Verifica dell'ipotesi di indipendenza in una tabella a doppia entrata

I dati campionari

Una nota azienda di telefonia mobile[8] vuole verificare se esiste una relazione tra la disattivazione del contratto da parte del cliente in un determinato mese e il piano tariffario che il cliente aveva sottoscritto.

In particolare, vuole confrontare tre particolari piani tariffari, nel seguito indicati con le lettere A, B e C rispetto alle disattivazioni da parte dei clienti nell'ultimo mese.

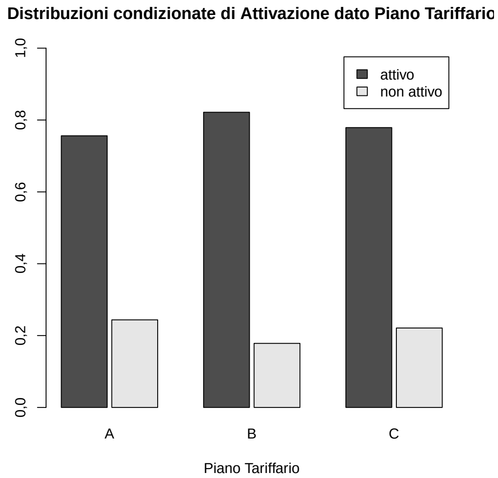
Cioè il problema che vogliamo affrontare è se esiste o no una qualche forma di dipendenza tra le variabili qualitative *Attivazione*, che assume le modalita *Attivo* e *Non attivo*, e *Piano Tariffario*, che assume modalità A, B e C.

La seguente tabella mostra le frequenze assolute (f_{ij}) per un campione (scelto a caso) di $n = \sum_{ij} f_{ij} = 2659$ unità statistiche (clienti) che sono state rilevate.

Attivazione	Piano Tariffario			
	A	B	C	Totale
Attivo	245	1677	229	**2151**
Non attivo	79	364	65	**508**
Totale	**324**	**2041**	**294**	**2659**

Grafico a barre

[8] Dati tratti da A. Azzalini e B. Scarpa, *Analisi dei Dati e Data Mining*, Springer-Verlag, 2004. I dati sono reali ma, per motivi di riservatezza, non è riportato il nome dell'azienda.



Distribuzione di *Attivazione* condizionata a *Piano Tariffario*. L'altezza delle barre è proporzionale alle frequenze relative.

Frequenze attese e X^2 di Pearson

La seguente tabella mostra le frequenze attese, $\hat{f}_{ij} = \frac{f_{i+}f_{+j}}{n}$, nell'ipotesi di perfetta indipendenza in distribuzione. Abbiamo indicato con $f_{i+} = \sum_j f_{ij}$ le frequenze marginali della variabile *Attivazione* e con $f_{+j} = \sum_i f_{ij}$ le frequenze marginali della variabile *Piano Tariffario*.

Attivazione	Piano Tariffario			
	A	*B*	*C*	Totale
Attivo	262,10	1651,07	237,83	**2151**
Non attivo	61,90	389,93	56,17	**508**
Totale	**324**	**2041**	**294**	**2659**

Il valore della statistica X^2 di Pearson è calcolato come segue

$$X^2 = \sum_{ij} \frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{\hat{f}_{ij}} = 9{,}69$$

Sappiamo che valori di X^2 vicini a zero indicano l'indipendenza tra le due variabili. Si pone adesso il problema di stabilire se il valore osservato 9,69 è sufficientemente lontano da 0 da indicare la presenza di una dipendenza tra le variabili.

Una tabella *fantasma*

- Consideriamo per un momento, l'insieme di tutti i clienti della nostra azienda, che costituiscono la nostra **popolazione di riferimento**.
- Se le due variabili fossere state rilevate su *tutte* queste unità statistiche avremmo potuto costruire una tabella, analoga a quella di pagina **??**, del tipo

Attivazione	Piano Tariffario			
	A	B	C	Totale
Attivo	F_{11}	F_{12}	F_{13}	F_{1+}
Non attivo	F_{21}	F_{22}	F_{23}	F_{2+}
Totale	F_{+1}	F_{+2}	F_{+3}	$N = F_{++}$

  dove (i) N indica il numero di tutti i clienti, (ii) F_{11} il numero di clienti ancora attivi e con il piano tariffario A, (iii)...

- La tabella precedente noi non la conosciamo. Per questo è una tabella *fantasma* e per questo le varie frequenze sono state indicate con lettere maiuscole.

Che relazione esiste tra la tabella osservata e quella *fantasma*?

- Il campione è stato formato: (1) estraendo un cliente tra gli N della popolazione; (2) estraendo un altro cliente tra gli $N-1$ clienti non estratti alla prima estrazione; ...; (2659) Estraendo un cliente tra gli $N-2658$ clienti non estratti nelle prime 2658 estrazioni. In tutte le estrazioni, è stata assegnata probabilità uguale a tutti i clienti non ancora estratti.
- Dividiamo tutte le frequenze della tabella *fantasma* per N ottenendo

Attivazione	Piano Tariffario			
	A	B	C	Totale
Attivo	π_{11}	π_{12}	π_{13}	π_{1+}
Non attivo	π_{21}	π_{22}	π_{23}	π_{2+}
Totale	π_{+1}	π_{+2}	π_{+3}	1

- Vista la maniera con cui è stato formato il campione, la probabilità che il primo cliente abbia, ad esempio, piano tariffario A e si sia disattivato nell'ultimo mese è π_{21}.

Le successive estrazioni non sono tra di loro indipendenti. Infatti, escludere un cliente già estratto altera ovviamente l'"urna" da cui stiamo estraendo.

- Nel caso in esame però N è molto grande e quindi la dipendenza è trascurabile da un punto di vista pratico.
- La probabilità (marginale) che un cliente abbia il piano tariffario A è $\pi_{+1} = \pi_{11} + \pi_{21}$. Analogamente la probabilità (marginale) che un cliente sia attivo è $\pi_{1+} = \pi_{11} + \pi_{12} + \pi_{13}$.
- Quindi, almeno approssimatamente, la tabella di pagina **??** mostra come si sono ripartiti nelle 6 "categorie" (*Attivo*, A), (*Attivo*, B), (*Attivo*, C), (*Non attivo*, A), (*Non attivo*, B), (*Non attivo*, C), i risultati di 2659 esperimenti casuali indipendenti tutti caratterizzati da

$$\begin{cases} \Pr(\text{estrarre un } (\textit{Attivo}, A)) = \pi_{11} \\ \Pr(\text{estrarre un } (\textit{Attivo}, B)) = \pi_{12} \\ \qquad\qquad\vdots \\ \Pr(\text{estrarre un } (\textit{Non attivo}, C)) = \pi_{23} \end{cases} .$$

- Una stima sensata con i dati del campione è data
  1. per le probabilità π_{ij} dalle frequenze relative osservate nel campione:
  $$\hat{\pi}_{ij} = f_{ij}/n, \qquad i = 1, 2; \;\; j = 1, 2, 3.$$
  2. per le probabilità (marginali) π_{i+} dalle frequenze relative marginali osservate nel campione
  $$\hat{\pi}_{i+} = f_{i+}/n, \qquad i = 1, 2.$$
  3. per le probabilit (marginali) π_{+j} dalle frequenze relative marginali osservate nel campione
  $$\hat{\pi}_{+j} = f_{+j}/n, \qquad j = 1, 2, 3.$$

Verifica dell'ipotesi di indipendenza

Una domanda interessante che possiamo fare ai dati è:

Nella tabella *fantasma* esiste indipendenza in distribuzione?

In altre parole l'indipendenza (o la dipendenza) che abbiamo rilevato nel campione è una peculiarità dei soli clienti estratti e quindi l'abbiamo osservata per puro caso oppure è la "manifestazione" della reale relazione tra i due fenomeni esistente nella popolazione.

Si tratta, ovviamente, di un problema di verifica d'ipotesi che può essere scritto nella forma

$$\begin{cases} \mathrm{H}_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \quad i = 1,2 \;\; j = 1,2,3 \quad \text{(indipendenza)} \\ \mathrm{H}_1 : \text{le } \pi_{ij} \text{ non rispettano i vincoli previsti da } \mathrm{H}_0 \quad \text{(dipendenza)} \end{cases}$$

Come può sembrare sensato, una stima dei π_{ij} sotto l'ipotesi di indipendenza è data dalle quantità $\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{f_{i+}}{n}\frac{f_{+j}}{n}$, $i = 1,2$ $j = 1,2,3$.

La statistica test più usata per verificare l'ipotesi d'indipendenza è l'X^2 di Pearson, che essenzialmente si basa sulla differenza tra le stime dei π_{ij} e le analoghe stime ottenute sotto l'ipotesi di indipendenza.

$$X^2 = \sum_{ij} \frac{\left(\hat{\pi}_{ij} - \hat{\pi}_{i+}\hat{\pi}_{+j}\right)^2}{\hat{\pi}_{i+}\hat{\pi}_{+j}} = \sum_{ij} \frac{\left(f_{ij} - \hat{f}_{ii}\right)^2}{\hat{f}_{ij}}$$

È certamente una statistica appropriata visto che ci aspettiamo che assuma valori

1. piccoli quando H_0 è vera e

2. grandi quando è falsa.

La distribuzione approssimata di X^2

E' possibile mostrare (rinviamo al solito la dimostrazione di questo risultato a corsi più avanzati) che se H_0 è vera e nessuna frequenza attesa è troppo piccola allora la distribuzione di X^2 può essere approssimata con la distribuzione di una variabile casuale chiamata χ^2 di Pearson.

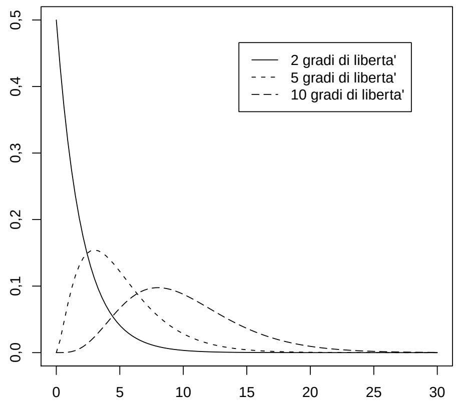
La distribuzione χ^2 dipende da un solo parametro, chiamato i gradi di libertà della distribuzione, che nel caso che stiamo trattando (verifica dell'ipotesi di indipendenza in una tabella di contingenza) deve essere posto uguale a

$$\left[\begin{pmatrix} \text{numero righe} \\ \text{tabella} \end{pmatrix} - 1\right] \times \left[\begin{pmatrix} \text{numero colonne} \\ \text{tabella} \end{pmatrix} - 1\right]$$

Ad esempio, per la tabella in esame, i gradi di libertà sono $2 = (2-1)\times(3-1)$.

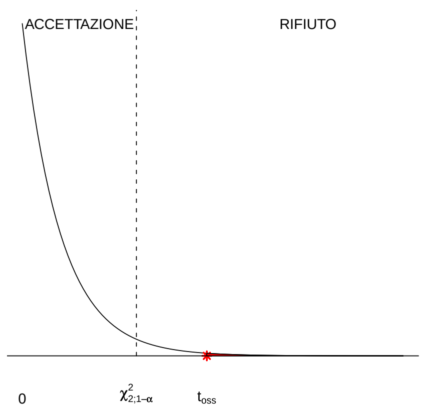
L'approssimazione è ritenuta "decorosa" se la più piccola delle frequenze attese (si noti, quelle attese, non quelle osservate) è più grande di 1 e migliora per valori grandi.

Densità di una variabile casuale χ^2 per tre valori dei gradi di libertà



Si osservi che all'aumentare dei gradi di libertà la densità si sposta verso destra. Infatti, è possibile dimostrare che la media della variabile casuale coincide con i gradi di libertà.

Test: analisi del risultato



Densità di una v.c. χ^2 con 2 gradi di libertà. L'asterisco sull'asse delle ascisse segna il valore osservato della statistica test. Il valore è nella regione di rifiuto e quindi rifiutiamo l'ipotesi nulla per $\alpha = 0{,}05$.

Livello di significatività osservato (e suo calcolo approssimato da una tavola dei percentili)

- "Lontano da H_0" vuol dire osservare una valore per la statistica test X^2 "grande". Quindi, in questo caso il livello di significatività osservato è la probabilità, di osservare quando è vera H_0 un valore uguale o maggiore di quello osservato. Quindi, per i dati presentati in questa unità,

$$\begin{pmatrix} \text{livello} \\ \text{significatività} \\ \text{osservato} \end{pmatrix} = P(\chi^2 \text{ con 2 gradi libertà} \geq 9{,}69)$$

- Supponiamo ora di voler determinare un intervallo che lo contenga conoscendo solo alcuni percentili della distribuzione. Ad esempio, supponiamo di conoscere solamente la seguente tabella in cui $\chi^2_{2,p}$ indica il percentile p-simo di un χ^2 con 2 gradi di libertà.

p	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995
$\chi^2_{2,p}$	0,0100	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2103	10,5966

- Il valore osservato del test ($X^2_{oss} = 9{,}69$) è compreso tra il 99-simo e il'99,5-simo percentile. Ora, per definizione la probabilità di assumere un valore più grande del 99-simo (99,5-simo) percentile è 1% (0,5%). Perciò

$$0{,}005 \leq (\text{livello significatività osservato}) \leq 0{,}01$$

- Quindi, ai fini pratici, sappiamo che il livello di significatività osservato è sufficientemente piccolo da farci rifiutare l'ipotesi di indipendenza anche a livello fissato $\alpha = 0{,}01$.

Quantili di un χ^2 di Pearson

g indica i gradi di libertà, p la probabilità lasciata a "sinistra", Quindi, ad esempio, $P(\chi^2 \text{ con 2 gradi di libertà} \leq 9{,}21) = 0{,}99$

	p							
g	0.1	0.25	0.5	0.75	0.90	0.95	0.99	0.999
1	0,02	0,1	0,45	1,32	2,71	3,84	6,63	10,83
2	0,21	0,58	1,39	2,77	4,61	5,99	9,21	13,82
3	0,58	1,21	2,37	4,11	6,25	7,81	11,34	16,27
4	1,06	1,92	3,36	5,39	7,78	9,49	13,28	18,47
5	1,61	2,67	4,35	6,63	9,24	11,07	15,09	20,52
6	2,2	3,45	5,35	7,84	10,64	12,59	16,81	22,46
7	2,83	4,25	6,35	9,04	12,02	14,07	18,48	24,32
8	3,49	5,07	7,34	10,22	13,36	15,51	20,09	26,12
9	4,17	5,9	8,34	11,39	14,68	16,92	21,67	27,88
10	4,87	6,74	9,34	12,55	15,99	18,31	23,21	29,59
11	5,58	7,58	10,34	13,7	17,28	19,68	24,72	31,26
12	6,3	8,44	11,34	14,85	18,55	21,03	26,22	32,91
13	7,04	9,3	12,34	15,98	19,81	22,36	27,69	34,53
14	7,79	10,17	13,34	17,12	21,06	23,68	29,14	36,12
15	8,55	11,04	14,34	18,25	22,31	25	30,58	37,7
16	9,31	11,91	15,34	19,37	23,54	26,3	32	39,25
17	10,09	12,79	16,34	20,49	24,77	27,59	33,41	40,79
18	10,86	13,68	17,34	21,6	25,99	28,87	34,81	42,31
19	11,65	14,56	18,34	22,72	27,2	30,14	36,19	43,82
20	12,44	15,45	19,34	23,83	28,41	31,41	37,57	45,31
21	13,24	16,34	20,34	24,93	29,62	32,67	38,93	46,8
22	14,04	17,24	21,34	26,04	30,81	33,92	40,29	48,27
23	14,85	18,14	22,34	27,14	32,01	35,17	41,64	49,73
24	15,66	19,04	23,34	28,24	33,2	36,42	42,98	51,18
25	16,47	19,94	24,34	29,34	34,38	37,65	44,31	52,62

Esercizio 23. *Si riconsideri la tabella dei laureati a Venezia e a alla Bocconi dell'Unità 3.9 e si verifichi l'ipotesi di indipendenza utilizzando il test* X^2.

Si controlli che il valore osservato del test, X^2_{oss}, *coincide con il quadrato del valore osservato,* t_{oss}, *del test* T *utilizzato per verificare il sistema d'ipotesi*

$$\begin{cases} H_0 : \vartheta_1 = \vartheta_2 \\ H_1 : \vartheta_1 \neq \vartheta_2 \end{cases}$$

Si verifichi inoltre che le due regioni di rifiuto sono equivalenti e anche l'equivalenza delle due ipotesi nulle

$H_0 : \vartheta_1 = \vartheta_2$ *e* $H_0' : \pi_{ij} = \pi_{i+}\pi_{+j}$ *(indipendenza).*

Insomma, in sostanza, si dimostri che si sta facendo esattamente la stessa cosa!

3.12 Hot–dog e calorie

- Rapporto tra medie e varianze condizionate e media e varianza marginali.
- Una misura della dipendenza in media.
- Analisi della varianza con un criterio di classificazione.

Hot–dog e calorie: i dati

Per cercare di capire se e di quanto la carne con cui vengono preparati gli *hot–dog* (wurstel) influenza il contenuto calorico degli stessi sono state misurate le calorie (per *hot–dog*) di 54 confezioni di diverse marche.

È inoltre noto se l'*hot–dog* è stato preparato con:

- solo carne bovina;
- carne mista (tipicamente a maggioranza maiale);
- pollame (pollo o tacchino).

Nel seguito mostriamo

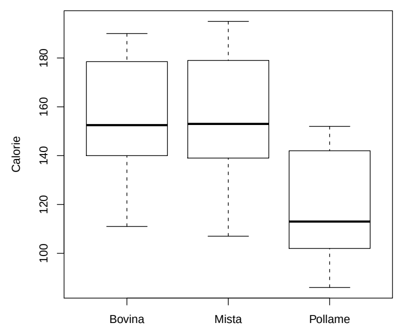
- i dati elementari;
- il diagramma scatola con baffi delle calorie classificate per tipo di carne e le numerosità, medie e scarti quadratici medi dei tre gruppi.

È evidente che, restringendo l'attenzione alle 54 misure disponibili, il tipo di carne influenza il contenuto calorico.

Tipo di carne e calorie (per pezzo) per 54 confezioni di _hot–dog_

Carne	Calorie	Carne	Calorie	Carne	Calorie
Bovina	186	Bovina	181	Bovina	176
Bovina	149	Bovina	184	Bovina	190
Bovina	158	Bovina	139	Bovina	175
Bovina	148	Bovina	152	Bovina	111
Bovina	141	Bovina	153	Bovina	190
Bovina	157	Bovina	131	Bovina	149
Bovina	135	Bovina	132	Mista	173
Mista	191	Mista	182	Mista	190
Mista	172	Mista	147	Mista	146
Mista	139	Mista	175	Mista	136
Mista	179	Mista	153	Mista	107
Mista	195	Mista	135	Mista	140
Mista	138	Pollame	129	Pollame	132
Pollame	102	Pollame	106	Pollame	94
Pollame	102	Pollame	87	Pollame	99
Pollame	107	Pollame	113	Pollame	135
Pollame	142	Pollame	86	Pollame	143
Pollame	152	Pollame	146	Pollame	144

Un primo sguardo ai dati



Carne	Numerosità	$\overline{y}$	s
Bovina	20	156,85	22,64
Mista	17	158,71	25,24
Pollame	17	118,76	22,55

Nota: s è la radice della stima della varianza ottenuta "dividendo per $n-1$"

Nel seguito ci concentremo sulla dipendenza in media rilevabile dalla tabella di pagina 252 ed in particolare affronteremo i seguenti punti:

- come misurare la forza della dipendenza in media e
- come verificare se è plausibile che le differenze osservate nelle medie siano *generalizzabili* a tutti gli *hot–dog* (o almeno a quelli prodotti con materie prime e tecnologia simili a quella usata per produrre le 54 confezioni).

Notazioni

- Per rendere il discorso generale indichiamo con k il numero dei gruppi (nel nostro caso $k = 3$) e con n_i, $i = 1, \ldots, k$ il numero di osservazioni per ogni gruppo (nel nostro caso, se 1 indica carne bovina, 2 carne mista e 3 pollame, $n_1 = 20$, $n_2 = 17$, $n_3 = 17$).
- L'insieme di tutte le osservazioni può poi essere indicato come y_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$. Si osservi che stiamo convenendo che il primo pedice indica il gruppo mentre il secondo l'osservazione dentro il gruppo.
- Per ogni gruppo possiamo calcolare la media
$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$
e la varianza
$$\mathrm{v}_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$$
Nel nostro caso, queste medie e varianze sono riferibili alle distribuzioni delle calorie condizionate ai vari tipi di carne.
- Possiamo anche calcolare la media e la varianza di *tutte* le osservazioni senza riferimento al gruppo di appartenenza
$$\overline{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$
$$\mathrm{v}^2 = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2$$
dove
$$n = \sum_{i=1}^{k} n_i$$

indica il numero totale di osservazioni disponibili. $\overline{y}$ e v^2 sono riferibili alla distribuzione marginale delle calorie.

- Si osservi che abbiamo definito le varianze dividendo la somma dei quadrati degli scarti dalla media per il "numero delle osservazioni" e non per "il numero delle osservazioni -1". Abbiamo quindi deciso, per il momento, di muoverci in un contesto descrittivo (le varie "v^2" sono la varianza delle osservazioni non la stima della varianza della popolazione da cui le osservazioni provengono). In realtà, le relazioni che vedremo sono facilmente estendibili anche se si segue l'altra strada.

La media della distribuzione marginale è la media delle medie delle distribuzioni condizionate

- Pensiamo ad una distribuzione di frequenza in cui le modalità sono le k medie condizionate e le frequenze (assolute) sono le numerosità delle osservazioni nei vari gruppi, ovvero, a

modalità	$\overline{y}_1$	$\overline{y}_2$	$\ldots$	$\overline{y}_k$
frequenze	n_1	n_2	$\ldots$	n_k

- La media (ponderata) di questa distribuzione è ovviamente

$$\frac{1}{n}\sum_{i=1}^{k} n_i \overline{y}_i$$

- E' immediato dimostrare che quest'ultima quantità coincide con la media marginale $\overline{y}$.

- Infatti
$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}.$$
Ma, per qualsivoglia i, dalla definizione di $\overline{y}_i$ segue che
$$\sum_{j=1}^{n_i} y_{ij} = n_i \overline{y}_i$$
e quindi,
$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{y}_i$$

La varianza della marginale è la media delle varianze condizionate + la varianza delle medie condizionate

- Ci si ricordi che v^2 indica la varianza di tutti i dati (= la varianza della distribuzione marginale), mentre le v_i^2 sono le varianze dentro il gruppo i-simo (= le varianze delle distribuzione condizionate).

- Dimostreremo che
$$v^2 = \frac{1}{n}\sum_{i=1}^{k} n_i v_i^2 + \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2. \tag{3.10}$$

- Si osservi che il primo addendo sul lato destro è la media (ponderata) di una distribuzione in cui le v_i^2 sono le modalità mentre le n_i sono le frequenze assolute, ovvero, è una media di varianze condizionate calcolata con pesi proporzionali alla numerosità dei vari gruppi.

- Viceversa, il secondo addendo è la varianza della distribuzione mostrata all'inizio di pagina 254, ovvero è la varianza delle medie condizionate (calcolata anche in questo caso con pesi proporzionali…).

- La verifica della (3.10) è agevole. Infatti
$$\begin{aligned}
v^2 &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \\
&= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}[(y_{ij} - \overline{y}_i) + (\overline{y}_i - \overline{y})]^2 = \\
&= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}[(y_{ij} - \overline{y}_i)^2 + (\overline{y}_i - \overline{y})^2 + 2(y_{ij} - \overline{y}_i)(\overline{y}_i - \overline{y})] = \\
&= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2 + \\
&\quad + \frac{2}{n}\sum_{i=1}^{k}(\overline{y}_i - \overline{y})\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i) = \\
&= \frac{1}{n}\sum_{i=1}^{k} n_i v_i^2 + \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2.
\end{aligned}$$

  Nell'ultima semplificazione abbiamo usato il fatto che, come sappiamo, la somma delle osservazioni del gruppo i-simo dalla media del gruppo i-simo è uguale a zero.

Una misura della dipendenza in media

- La (3.10) mostra come la varianza totale, v^2, sia scomponibile in due parti:

  1. la prima, il 1° addendo, dovuta alla variabilità *entro* i gruppi e
  2. la seconda, il 2° addendo, legata alle differenze (in media) *tra* i gruppi.

  Per questo motivo, i due addendi sono spesso indicati come **varianza entro i gruppi** e **varianza tra i gruppi**.

- Si osservi che se la varianza tra i gruppi è nulla, le medie condizionate sono tutte uguali a $\overline{y}$ e quindi esiste indipendenza in media.

- Viceversa, se la varianza tra i gruppi è molto grande rispetto alla varianza entro i gruppi, allora buona parte della variabilità totale mostrata dai dati diventa interpretabile in termini di differenze tra le medie condizionate. Siamo quindi in presenza di una situazione in cui la dipendenza in media esiste ed è importante (= le differenze tra le medie "spiegano" una larga frazione delle differenze che osserviamo nei dati).

- Sembra allora ragionevole usare

$$\begin{aligned} \eta^2 &= \frac{\text{varianza tra i gruppi}}{\text{varianza totale}} = \\ &= 1 - \frac{\text{varianza entro i gruppi}}{\text{varianza totale}} \end{aligned}$$

  per misurare la forza della dipendenza in media.

- In particolare si osservi che

  1. $0 \leq \eta^2 \leq 1$.
  2. $\eta^2 = 0$ implica indipendenza in media.
  3. $\eta^2 = 1$ implica che la varianza entro i gruppi è nulla. Siamo quindi in una situazione di dipendenza, in ogni senso e quindi anche in media, perfetta.
  4. η^2 non è ovviamente ne definito ne sensato quando $v = 0$. Questo non è un grande problema visto che v uguale a zero vuol dire che tutte le osservazioni sono uguali tra di loro e quindi che non esiste nessuna variabilità interessante da indagare.
  5. η^2 è chiamato indice **rapporto di correlazione** di Pearson.

- Nel caso degli *hot–dog*, η^2 è facilmente calcolabile dai risultati della tabella di pagina 252[9]. In particolare,

$$\begin{aligned} \text{varianza entro i gruppi} &\approx 519{,}76 \\ \text{varianza tra i gruppi} &\approx 327{,}63 \\ \text{varianza totale} &\approx 847{,}39 \end{aligned}$$

  e, quindi, $\eta^2 \approx 0{,}39$. Il valore trovato ci indica la presenza di una discreta ma non eccezionale dipendenza in media.

E se tutto fosse dovuto al caso

- Fino a questo punto abbiamo solo guardato ai dati disponibili. In realtà noi non compreremo mai nessuna delle 54 confezioni di *wurstel* analizzate e quindi siamo interessati a sapere quanto le differenze evidenziate siano estendibili ai *wurstel* che potremmo mangiare.

- Una maniera di vedere il problema consiste nel riconoscere che fino a questo punto abbiamo trascurato una ulteriore fonte di variabilità, quella **campionaria**. Ad esempio, almeno una parte delle differenze tra le medie presentate nella tabella di pagina 252 è specifica delle 54 confezioni utilizzate, nel senso che, replicando l'esperimento (ovvero, prendendo altre 54 confezioni,...) ci aspettiamo di trovare risultati diversi.

- La domanda è: "Di quanto diversi? Tanto diversi, ad esempio, da portarci a concludere che le minore calorie osservate per gli *hot–dog* di pollo e tacchino sono solamente una specificità del campione disponibile? Oppure, diversi si, ma non tanto da alterare le conclusioni suggerite dalla tabella?"

Un problema di verifica d'ipotesi

Pensiamo all'insieme[10] dei milioni e milioni di possibili *hot–dog* che potrebbero essere prodotti con gli ingredienti e la tecnologia attuale.

Questa *popolazione* ovviamente può essere divisa in tre gruppi:

- quelli prodotti con sola carne di bovino;

- quelli prodotti con carne mista;

[9] Ci si ricordi comunque che nella tabella, come spiegato nella legenda, s è la radice della stima della varianza calcolata dividendo per "il numero dei dati -1". Quindi, con notazioni ovvie, la varianza entro i gruppi deve essere calcolata con la formula $[\sum(n_i - 1)s_i^2]/n$.

[10] Un po' stomachevole, nella sua enormità.

- quelli prodotti con pollame.

Possiamo allora calcolare la media delle calorie per ciascuno di questi tre gruppi. Indichiamole rispettivamente con μ_1, μ_2 e μ_3.

Un sistema di ipotesi che può essere interessante verificare con i dati è

$$\begin{cases} \mathrm{H}_0 : \{\mu_1 = \mu_2 = \mu_3\} \\ \mathrm{H}_1 : \begin{Bmatrix} \text{almeno una delle uguaglianze previste} \\ \text{da } \mathrm{H}_0 \text{ è falsa} \end{Bmatrix} \end{cases} .$$

Infatti, se H_0 fosse vera allora nella popolazione, contrariamente a quanto osservato nel campione, non esisterebbe dipendenza in media.

Si osservi come il problema sia molto simile a quello che ci siamo posti nell'unità F. La differenza è che adesso sono coinvolte più di due medie.

Analisi della varianza con un criterio di classificazione

Al solito, per arrivare ad una soluzione abbiamo bisogno di descrivere la relazione che intercorre tra le osservazioni e la popolazione. In particolare, la relazione che interccorre tra le osservazioni e le tre medie μ_1, μ_2 e μ_3. Una relativamente "semplice" soluzione esiste nel caso in cui sia credibile assumere che:

1. La distribuzione all'interno dell'i-gruppo sia normale di media μ_i e varianza σ^2. Si osservi che stiamo supponendo che la varianza non dipenda da i, ovvero, che tutti i gruppi abbiano la stessa variabilità interna.

2. Le osservazioni sono tutte indipendenti tra di loro e, per qualsivoglia i e j, $y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$.

La statistica test comunemente usata è

$$T = \left(\frac{\text{varianza tra i gruppi}}{\text{varianza entro i gruppi}} \right) \left(\frac{n-k}{k-1} \right).$$

La statistica T è in stretta relazione con η^2. Infatti, come è facile verificare,

$$T = \left(\frac{\eta^2}{1-\eta^2} \right) \left(\frac{n-k}{k-1} \right).$$

Si noti inoltre che la funzione $f : x \to x/(1-x)$ è monotona crescente nell'intervallo $[0, 1]$ Quindi, più è grande η^2 più è grande t_{oss} e viceversa.

Ovviamente, poiché ci aspettiamo t_{oss} grande quando H_0 è falsa, consideriamo evidenza contro l'ipotesi nulla valori elevati della statistica. Il problema è al solito "quanto grande deve essere t_{oss} per farci dubitare di H_0?".

La risposta è facilitata dal fatto che è possibile dimostrare che, nelle ipotesi in cui ci siamo messi (normalità, indipendenza,...), T si distribuisce come una variabile casuale F di Fisher–Snedecor con $k-1$ gradi di libertà al numeratore e $n-k$ al denominatore.

In pratica

Per i dati sugli *hot–dog*,

$$t_{oss} \approx 16$$

. Questo valore deve essere confrontato con i quantile di una F di Fisher–Snedecor con 2 e 51 gradi di libertà.

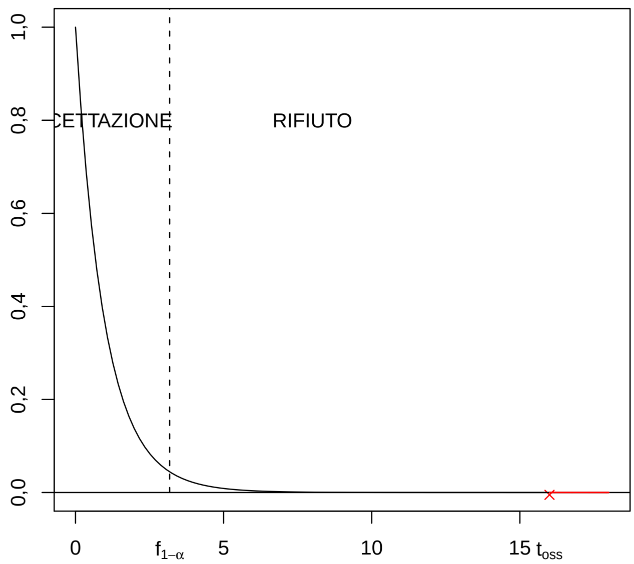
Se $\alpha = 0{,}05$, $f_{2,51,0{,}95} = 3,179$ e quindi rifiutiamo l'ipotesi nulla in favore dell'ipotesi alternativa.

Il livello di significatività osservato è pari a

$$\begin{aligned} \alpha_{oss} = \Pr(T \geq t_{oss} \quad \text{quando H}_0 \text{ è vera}) &= \Pr(F_{2,51} \geq 16) \\ &= 4e - 06 \ . \end{aligned}$$

il che indica, che un valore "uguale o più lontano da H_0" di quello osservato è molto improbabile quando l'ipotesi nulla è vera. In particolare, il livello di significatività osservato è inferiore a un millesimo.

In conclusione, i dati ci suggeriscono che non solo le medie nel campione ma anche quelle nella popolazione dovrebbero essere tra di loro diverse.



3.13 Cercar casa

- Diagramma di dispersione
- Dipendenza lineare
- Modello di regressione lineare semplice
- Minimi quadrati
- Inferenza

I dati

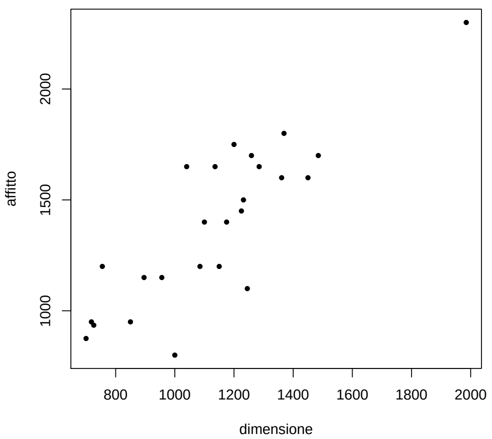
Un agente immobiliare [11] intende prevedere gli affitti mensili degli appartamenti sulla base della loro dimensione. Per questo conduce un'indagine e reperisce i dati su 25 appartamenti in una zona residenziale. La seguente tabella mostra i dati ottenuti per i 25 appartamenti, l'affitto l'affitto mensile in dollari e la dimensione è espressa in piedi quadrati.

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

Si vogliono utilizzare i dati per ottenere una equazione che permetta di prevedere l'affitto in base alla dimensione.

Diagramma di dispersione

[11] Dati tratti da Levine, Krehbiel, Berenson (2002) Statistica, Apogeo.



Abbiamo semplicemente disegnato i punti osservati sul piano. E' evidente una forte relazione, certamente crescente come ci si poteva attendere. Possiamo quantificare quanto forte è questa relazione?

Dipendenza lineare e indice di correlazione

Una possibilità per quantificare quanto forte è una relazione tra due variabili è quella di utilizzare l'indice di correlazione (che è una misura di dipendenza lineare). Dall' Unità ???? sappiamo che l'indice di correlazione tra due variabili casuali X e Y è

$$\rho(X,Y) = \frac{\mathbb{C}ov(X,Y)}{\sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y)}}$$

Quando, come nel nostro caso, abbiamo un campione $x = (x_1, x_2, \cdots, x_n)$ e $y = (y_1, y_2, \cdots, y_n)$, possiamo definire il **coefficiente di correlazione campionario** come

$$\text{corr}(x,y) = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\text{var}(y)}}$$

dove

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

indica la **covarianza campionaria** tra i due vettori x e y

e

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

sono le varianze.

Per il nostro insieme di dati

$$\mathrm{corr}(\text{dimensione}, \text{affitto}) \approx \frac{85200{,}75}{\sqrt{125589{,}04 \times 79989{,}90}} = 0{,}850 \ .$$

Calcolo della covarianza

Per il calcolo della covarianza è conveniente utilizzare la seguente relazione

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

ovvero

$$(\text{covarianza}) = \left(\begin{array}{c} \text{media dei} \\ \text{prodotti} \end{array} \right) - \left(\begin{array}{c} \text{prodotto delle} \\ \text{medie} \end{array} \right).$$

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 24. *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 sì, quale e in che senso?).*

Un possibile modello

- Come si determina il prezzo?
- Il prezzo è determinato da innumerevoli fattori, non tutti determinabili.
- Potremmo pensare che un fattore determinante sia dato dalla dimensione.
- Adottiamo per il momento l'ipotesi di una relazione lineare.

- Possiamo allora pensare ad un modello del tipo

$$(\text{affitto}) = \alpha + \beta(\text{dimensione}) + (\text{errore}) \tag{3.11}$$

dove la componente errore rappresenta quella parte di variabilità del prezzo di affitto determinata da variabii diverse dalla dimensione o comunque non spiegabile dalla relazione lineare (3.11).

Modelli di regressione lineare semplice: caso generale e terminologia

Come possiamo interpretare la variabilità dei dati a nostra disposizione ?

$$\begin{array}{cccc} (y_1, x_1) & (y_2, x_2) & \dots & (y_{25}, x_{25}) \\ (950, 850) & (1600, 1450) & \dots & (1750, 1200) \end{array}$$

Ipotizziamo che i valori y_i, $i = 1, \dots, 25$ siano altrettante osservazioni da v.c. Y_i tra loro **indipendenti**. Poiché pensiamo che la diversa dimensione dell'appartamento influisca sul prezzo dell'affitto le v.c. **non sono identicamente distribuite**. Per esempio potremmo ipotizzare che Y_i sia una v.c. con valore atteso che dipende dalla dimensione dell'appartamento (che denotiamo x_i).

$$\mathbb{E}(Y_i) = \alpha + \beta x_i$$

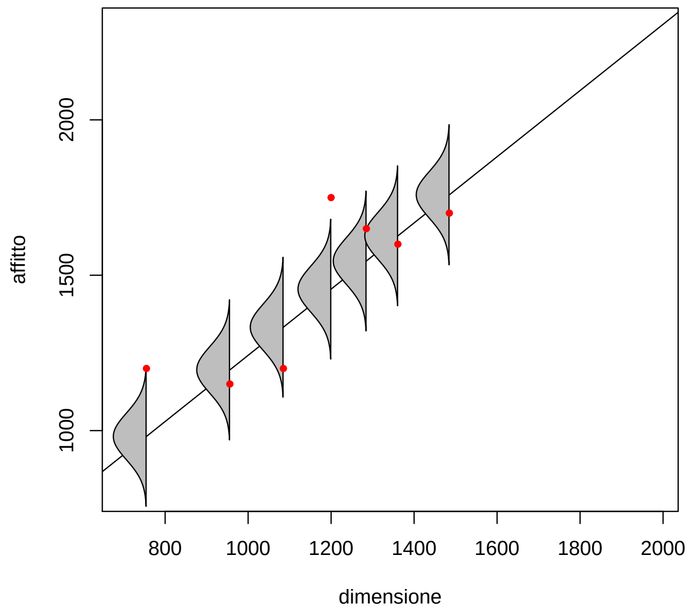
In questo modello ipotizziamo che il prezzo 'medio' dell'affitto sia determinato da dei fattori fissi (α) e dei fattori variabili (βx_i)

Una scrittura equivalente del modello che spiega la variabilità del prezzo d'affitto è la seguente

$$Y_i = \alpha + \beta x_i + \varepsilon_i. \tag{3.12}$$

dove ε_i è una v.c. detta **errore** tale che $\mathbb{E}(\varepsilon_i) = 0$.

Circa la variabilità dell'errore supponiamo che $\mathbb{V}ar(\varepsilon_i) = \sigma^2$ ovvero che $\mathbb{V}ar(Y_i) = \sigma^2$.



Un modello del tipo (3.12) viene usualmente chiamato **modello di regressione lineare semplice**. Nel caso generale, cerchiamo di **spiegare** una variabile, diciamo Y, utilizzando un'altra variabile, diciamo x.

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

Assunzioni del modello lineare

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

- Y viene usualmente indicata come **variabile risposta** o **variabile dipendente** e Y_i è la **variabile casuale** relativa alla $i-esima$ unità che fa parte del campione.

- $y = (y_1, y_2, \cdots, y_n)$ è il vettore che contiene le osservazioni relative alla variabile risposta.

- x è il **regressore**, detto anche **variabile esplicativa** o **variabile indipendente**, e x_i è il valore **predeterminato** assunto dalla $i-esima$ unità che fa parte del campione. Si noti il diverso ruolo di x e Y: i valori di x si considerano predeterminati. Spesso utilizzeremo lo stesso simbolo x anche per indicare il vettore $(x_1, x_2, \cdots, x_n)$ che contiene le osservazioni relative alla variabile esplicativa.

- ε rappresenta il termine d'**errore**. È una **variabile casuale** con $\mathbb{E}(\varepsilon_i) = 0$ e $\mathbb{V}ar(\varepsilon_i) = \sigma^2$
- le variabili Y_i sono supposte indipendenti [12]
- α, β e σ^2 sono i **parametri** del modello e sono **ignoti**.

Minimi quadrati: idea

Il problema è come stimare 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 prezzo d'affitto

$$\hat{\alpha} + \hat{\beta}(\text{dimensione}). \tag{3.13}$$

Sembra ragionevole cercare di "calcolare" $\hat{\alpha}$ e $\hat{\beta}$ in modo tale che la (3.13) fornisca buone "previsioni" sull'insieme di dati osservato. Indichiamo con n il numero delle osservazioni (in questo caso $n = 25$), e poniamo $y_i =$ (affitto della casa i-sima) e $x_i =$ (dimensione della casa i-sima).

Il nostro scopo è trovare dei valori per i parametri tali che

$$\begin{array}{rcl} y_1 & \approx & \hat{\alpha} + \hat{\beta} x_1 \\ y_2 & \approx & \hat{\alpha} + \hat{\beta} x_2 \\ & \vdots & \\ y_n & \approx & \hat{\alpha} + \hat{\beta} x_n \end{array} \tag{3.14}$$

Per rendere "operativa" la (3.14), dobbiamo decidere (i) in che senso interpretiamo gli $\approx$ che abbiamo scritto e (ii) come combiniamo tra di loro le varie linee della (3.14) stessa. La soluzione più usata si concretizza nello scegliere i due parametri minimizzando

$$s^2(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{3.15}$$

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 stimati utilizzando il **metodo dei minimi quadrati**".

[12] In realtà nell'usuale formulazione le variabili Y_i sono supposte tra loro incorrelate, ma per semplicit di trattazione introduciamo un'ipotesi più forte.

Minimi quadrati: stima dei parametri α e β

Per ottenere i valori di α e β che rendono minima $s^2(\alpha, \beta)$ definita in (3.15) possiamo procedere come seque. La funzione $s^2(\alpha, \beta)$ è derivabile e quindi i punti stazionari (di minimo, massimo o di sella) sono calcolabili andando a trovare i punti in cui le derivare (parziali) si annullano, si tratta quindi di risolvere il sistema

$$\begin{cases} \dfrac{\partial}{\partial \alpha} s^2(\alpha, \beta) & = & 0 \\ \dfrac{\partial}{\partial \beta} s^2(\alpha, \beta) & = & 0 \end{cases} \tag{3.16}$$

che porta alla seguente soluzione

$$\begin{cases} \hat{\alpha} & = & \overline{y} - \hat{\beta}\overline{x} \\ \hat{\beta} & = & \dfrac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} \end{cases} \tag{3.17}$$

dove $\overline{y}$, $\overline{x}$, $\operatorname{var}(x)$ e $\operatorname{cov}(x, y)$ sono rispettivamente la media della variabile risposta, la media e la varianza della variabile esplicativa e la covarianza tra le due variabili.

La (3.17) fornisce la soluzione del problema solamente se $\operatorname{var}(x) > 0$. Questo è molto ragionevole: β ci dice come varia la risposta al variare della esplicativa, ma se $\operatorname{var}(x) = 0$ l'esplicativa non è variata affatto nei dati disponibili.

Dettagli sulla soluzione del sistema di equationi dei minimi quadrati*

$$\begin{cases} \frac{\partial}{\partial \alpha} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 & = & 0 \\ \frac{\partial}{\partial \beta} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 & = & 0 \end{cases}$$

$$\begin{cases} -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) & = & 0 \\ -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) x_i & = & 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^n y_i - n\alpha - \beta \sum_{i=1}^n x_i & = & 0 \\ \sum_{i=1}^n y_i x_i - \alpha \sum_{i=1}^n x_i - \beta \sum_{i=1}^n x_i^2 & = & 0 \end{cases}$$

$$\begin{cases} \bar{y} - \alpha - \beta \bar{x} & = & 0 \\ \frac{1}{n} \sum_{i=1}^n y_i x_i - \alpha \bar{x} - \beta \frac{1}{n} \sum_{i=1}^n x_i^2 & = & 0 \end{cases}$$

$$\begin{cases} \bar{y} - \beta \bar{x} & = & \alpha \\ \frac{1}{n} \sum_{i=1}^n y_i x_i - (\bar{y} - \beta \bar{x}) \bar{x} - \beta \frac{1}{n} \sum_{i=1}^n x_i^2 & = & 0 \end{cases}$$

$$\begin{cases} \bar{y} - \beta\bar{x} & = & \alpha \\ \frac{1}{n}\sum_{i=1}^n y_i x_i - \bar{y}\bar{x} - \beta(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2) & = & 0 \end{cases}$$

$$\begin{cases} \bar{y} - \beta\bar{x} & = & \alpha \\ \mathrm{cov}(x,y) - \beta\mathrm{var}(x) & = & 0 \end{cases}$$

$$\begin{cases} \alpha & = & \bar{y} - \beta\bar{x} \\ \beta & = & \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} \end{cases}$$

Possiamo verificare che la soluzione del sistema è un punto di minimo della funzione $s^2(\alpha,\beta)$ verificando che il determinante della matrice delle derivate calcolato in tale soluzione sia positivo (visto che gli elementi diagonali sono positivi).

$$A = \begin{bmatrix} \frac{\partial}{\partial\alpha}\frac{\partial}{\partial\alpha}s^2(\alpha,\beta) & \frac{\partial}{\partial\beta}\frac{\partial}{\partial\alpha}s^2(\alpha,\beta) \\ \frac{\partial}{\partial\alpha}\frac{\partial}{\partial\beta}s^2(\alpha,\beta) & \frac{\partial}{\partial\beta}\frac{\partial}{\partial\beta}s^2(\alpha,\beta) \end{bmatrix} = \begin{bmatrix} 2n & 2\sum_{i=1}^n x_i \\ 2\sum_{i=1}^n x_i & 2\sum_{i=1}^n x_i^2 \end{bmatrix}$$

per cui il determinante è

$$|A| = 4n\sum_{i=1}^n x_i^2 - 4\sum_{i=1}^n x_i \sum_{i=1}^n x_i$$

e quindi

$$\begin{aligned} |A| &= 4n^2\left(\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2\right) \\ &= 4n^2\mathrm{var}(x) \ . \end{aligned}$$

La soluzione ottenuta è un punto di minimo cioè la soluzione del problema ai minimi quadrati se $\mathrm{var}(x) > 0$.

Stima dei parametri nel caso degli affitti

Da un punto di vista pratico, le stime di α e β possono essere ottenute conoscendo le seguenti quantità:

$$n \qquad \sum_{i=1}^n y_i \qquad \sum_{i=1}^n x_i \qquad \sum_{i=1}^n x_i^2 \qquad \sum_{i=1}^n x_i y_i \ .$$

da cui

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{3.18}$$

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \tag{3.19}$$

Nel nostro caso specifico,

$$\begin{array}{lll} n = 25 & \sum y_i = 34660 & \sum x_i = 28383 \\ & \sum x_i^2 = 34223535 & \sum x_i y_i = 41480210. \end{array}$$

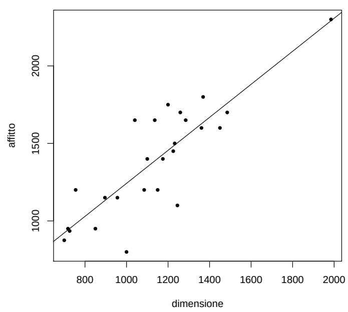
Perciò

$$\begin{array}{l} \overline{y} = 34660/25 = 1386{,}4 \\ \overline{x} = 28283/25 = 1135{,}32 \\ \mathrm{var}(x) = (34223535/25) - 1135{,}32^2 = 79989{,}9 \\ \mathrm{cov}(x, y) = (41480210/25) - 1135{,}32 \times 1386{,}4 = 85200{,}75. \end{array}$$

Quindi

$$\begin{array}{l} \hat{\beta} = 85200{,}75/79989{,}9 = 1{,}065144 \\ \hat{\alpha} = 1386{,}4 - 1{,}065144 \times 1135{,}32 = 177{,}1207 \end{array}$$

Modello Stimato



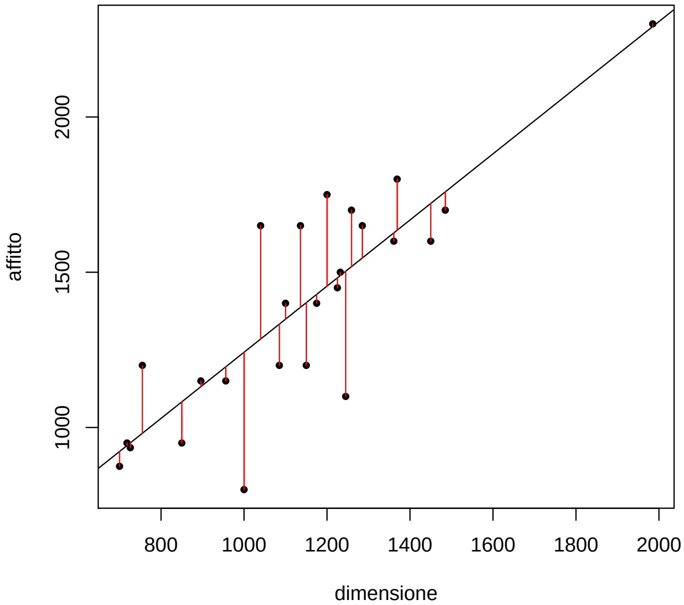
Il grafico mostra i dati osservati e la retta di regressione stimata. I due parametri hanno un ruolo geometrico, cioè $\hat{\alpha}$ è l'intercetta e $\hat{\beta}$ è il coefficiente angolare della retta. Quindi $\hat{\alpha}$ è il costo fisso che una persona deve sopportare per affittare un appartamento indipendentemente dalla sua dimensione, mentre $\hat{\beta}$ è quanto aumenta l'affitto se passiamo da un appartamento ad un altro di un piede quadrato più grande, cioè quanto aumenta Y all'aumentare di una unità di x.

I residui: stima di σ^2

Le differenze tra i valori osservati della variabile 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**. Essi si possono pensare come stime degli errori $\varepsilon_i = y_i - \alpha - \beta x_i$ (Si noti che in questo caso, non stimiamo parametri ma bensì il risultato di un esperimento casuale, i valori ottenuti dalla variabile casuale ε).



E' facile verificare che la media campionaria dei residui è nulla:

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

Questo è in linea con l'assunzione $\mathbb{E}(\varepsilon) = 0$ ma si noti che il risultato sui residui non è una evidenza in favore dell'assunzione.

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 "spiegherà" le variazioni della risposta.

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

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

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

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

Alla luce di quanto detto, la varianza campionaria dei residui può essere considerata una stima di $\mathbb{V}ar(\varepsilon) = \sigma^2$ ovvero della varianza attorno alla retta $\alpha + \beta x$.

Per motivi legati alla distribuzione delle variabili casuali che entrano in gioco per la stima di σ^2 si preferisce utilizzare

$$s^2 = \frac{\sum_{i=1}^{n} r_i^2}{n - 2}$$

perchè il relativo stimatore risulta essere non distorto, cioè $\mathbb{E}(S^2) = \sigma^2$.

Per il calcolo di s^2, teniamo conto che:

$$\begin{aligned} \mathrm{var}(r_1, \ldots, r_n) &= \mathrm{var}(y) - \mathrm{cov}^2(x, y)/\mathrm{var}(x) \\ &= \mathrm{var}(y) - \hat{\beta}^2 \mathrm{var}(x) \\ &= \frac{1}{n} \sum_{i=1}^{n} y_i^2 - \overline{y}^2 - \hat{\beta}^2 \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{x}^2 \right) \end{aligned}$$

e che $s^2 = \frac{n}{n-2}\mathrm{var}(r_1, \ldots, r_n)$; quindi

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

Notiamo che, oltre alle 5 quantità di pagina 267, ci basta calcolare $\sum_{i=1}^n y_i^2$. Nel nostro caso otteniamo $s^2 = 37867{,}37$.

Coefficiente di determinazione

È definito come

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

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 $\mathrm{var}(r_1, \ldots, r_n) = \mathrm{var}(y)$ 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 e ρ

Proposizione 25. *Nel caso di un modello di regressione lineare* ***semplice*** *abbiamo*

$$R^2 \equiv \rho^2$$

Dimostrazione 13.

$$\begin{aligned}
R^2 &= 1 - \frac{\mathrm{var}(y) - \mathrm{cov}^2(x, y)/\mathrm{var}(x)}{\mathrm{var}(y)} \\
&= \frac{\mathrm{var}(y) - \mathrm{var}(y) + \mathrm{cov}^2(x, y)/\mathrm{var}(x)}{\mathrm{var}(y)} \\
&= \frac{\mathrm{cov}^2(x, y)}{\mathrm{var}(x)\ \mathrm{var}(y)} \\
&= \rho^2
\end{aligned}$$

R^2 per le case

Abbiamo visto prima che

$\overline{y} = 34660/25 = 1386{,}4$
$\mathrm{var}(x) = (34223535/25) - 1135{,}32^2 = 79989{,}9$
$\mathrm{cov}(x, y) = (41480210/25) - 1135{,}32 \times 1386{,}4 = 85200{,}75.$

Inoltre

$$\sum y_i^2 = 51192350$$

Quindi

$$\mathrm{var}(y) = \frac{51192350}{25} - 1386,4^2 = 125589,04$$

e perciò

$$\mathrm{var}(r_1, \ldots, r_n) = 125589,04 - 85200,752^2/79989,898 = 34837,978.$$

Il coefficiente di determinazione vale

$$R^2 = 1 - \frac{34837,978}{125589,04} = 0,723 \ ,$$

ovvero il modello spiega il 72% della varianza della risposta.

La distribuzione degli stimatori $\hat{\alpha}$ e $\hat{\beta}$

Consideriamo lo stimatore di β

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{j=1}^n (x_j - \overline{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{\sum_{j=1}^n (x_j - \overline{x})^2} = \sum_{i=1}^n w_i Y_i \\ &\text{con } w_i = \frac{x_i - \overline{x}}{\sum_{j=1}^n (x_j - \overline{x})^2} \end{aligned}$$

e calcoliamo il valore atteso

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \mathbb{E}\left(\sum_{i=1}^n w_i Y_i\right) \\ &= \sum_{i=1}^n w_i \mathbb{E}(Y_i) \\ &= \frac{\sum_{i=1}^n (x_i - \overline{x})(\alpha + \beta x_i)}{\sum_{j=1}^n (x_j - \overline{x})^2} \\ &= \frac{\alpha \sum_{i=1}^n (x_i - \overline{x}) + \beta \sum_{i=1}^n (x_i - \overline{x}) x_i}{\sum_{j=1}^n (x_j - \overline{x})^2} \\ &= \frac{\beta \sum_{i=1}^n (x_i - \overline{x}) x_i}{\sum_{j=1}^n (x_j - \overline{x})^2} \\ &= \beta \end{aligned}$$

da cui deduciamo che lo stimatore è non distorto.

$$\begin{aligned}
\mathbb{V}ar(\hat{\beta}) &= \mathbb{V}ar\left(\sum_{i=1}^{n} w_i Y_i\right) \\
&= \sum_{i=1}^{n} w_i^2 \mathbb{V}ar(Y_i) \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\left(\sum_{j=1}^{n}(x_j - \overline{x})^2\right)^2}\sigma^2 \\
&= \frac{\sigma^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}
\end{aligned}$$

Se supponiamo che Y_i siano delle v.c. $N(\alpha+\beta x_i, \sigma^2)$ e indipendenti allora

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)$$

Per $\hat{\alpha}$ valgono considerazioni del tutto analoghe a quanto appena visto:

$$\begin{aligned}
\hat{\alpha} &= \overline{Y} - \hat{\beta}\overline{x} \\
&= \frac{\sum_{i=1}^{n} Y_i}{n} - \overline{x}\frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \\
&= \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{x}\frac{(x_i - \overline{x})}{\sum_{j=1}^{n}(x_j - \overline{x})^2}\right)Y_i \\
&= \sum_{i=1}^{n} \omega_i^* Y_i \\
&\quad \text{con} \quad \omega_i^* = \frac{1}{n} - \overline{x}\frac{(x_i - \overline{x})}{\sum_{j=1}^{n}(x_j - \overline{x})^2} .
\end{aligned}$$

e cioè anche $\hat{\alpha}$ è una combinazione lineare delle variabili Y_i.

Il valore atteso è

$$\begin{aligned}
\mathbb{E}(\hat{\alpha}) &= \sum_{i=1}^{n} \omega_i^* \mathbb{E}(Y_i) = \sum_{i=1}^{n} \omega_i^* (\alpha + \beta x_i) \\
&= \alpha \sum_{i=1}^{n} \left(\frac{1}{n} - \overline{x} \frac{(x_i - \overline{x})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) \\
&+ \beta \sum_{i=1}^{n} \left(\frac{x_i}{n} - \overline{x} \frac{(x_i - \overline{x}) x_i}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) \\
&= \alpha \left(\frac{n}{n} - \overline{x} \frac{\sum_{i=1}^{n} (x_i - \overline{x})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) \\
&+ \beta \left(\frac{\sum_{i=1}^{n} x_i}{n} - \overline{x} \frac{\sum_{i=1}^{n} (x_i - \overline{x}) x_i}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) \\
&= \alpha
\end{aligned}$$

Quindi $\hat{\alpha}$ è uno stimatore non distorto di α.
La varianza di $\hat{\alpha}$ si ottiene con alcuni semplici passaggi algebrici:

$$\begin{aligned}
&= \mathbb{V}ar(\hat{\alpha}) \\
&= \mathbb{V}ar \left(\sum_{i=1} \omega_i^* Y_i \right) \\
&= \sum_{i=1}^{n} (\omega_i^*)^2 Var(Y_i) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - \overline{x} \frac{(x_i - \overline{x})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right)^2 \sigma^2 \\
&= \sum_{i=1}^{n} \left(\frac{1}{n^2} + \overline{x}^2 \frac{(x_i - \overline{x})^2}{\left(\sum_{j=1}^{n} (x_j - \overline{x})^2 \right)^2} - \frac{2}{n} \overline{x} \frac{(x_i - \overline{x})}{\sum_{i=j}^{n} (x_j - \overline{x})^2} \right) \sigma^2 \\
&= \left(\frac{n}{n^2} + \overline{x}^2 \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\left(\sum_{j=1}^{n} (x_j - \overline{x})^2 \right)^2} - \frac{2}{n} \overline{x} \frac{\sum_{i=1}^{n} (x_i - \overline{x})}{\sum_{j=1}^{n} (x_j - \overline{x})^2} \right) \sigma^2 \\
&= \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=j}^{n} (x_j - \overline{x})^2} \right) \sigma^2
\end{aligned}$$

Se supponiamo che Y_i siano delle v.c. $N(\alpha + \beta x_i, \sigma^2)$ e indipendenti allora

$$\hat{\alpha} \sim N\left(\alpha, \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\sigma^2\right)$$

Infine si può dimostrare che se Y_i sono delle v.c. $N(\alpha + \beta x_i, \sigma^2)$ e indipendenti allora

$$\frac{(n-2)S^2}{\sigma^2} = \frac{\sum_{i=1}^n R_i^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 \sim \chi^2_{n-2}$$

e che lo stimatore S^2 è indipendente da $\hat{\alpha}$ e $\hat{\beta}$. Si ricordi che $\mathbb{E}(\chi^2_{n-2}) = n-2$.

Riassumendo possiamo concludere che, nel caso in cui le Y_i abbiano distribuzione normale, si ha

$$\frac{\hat{\alpha} - \alpha}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim N(0; 1)$$

$$\frac{\hat{\beta} - \beta}{\sigma\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \sim N(0; 1)$$

Un problema di verifica d'ipotesi

L'agente immobiliare aveva sempre applicato una semplice valutazione del prezzo d'affitto: il prezzo era proporzionale alla dimensione dell'appartamento, e affittava gli appartamenti ad un dollaro per piede quadro. Vediamo ora se questa valutazione errata in particolare sottoponiamo a verifica il sistema d'ipotesi

$$\begin{cases} \mathrm{H}_0 : \beta = \beta_0 \\ \mathrm{H}_1 : \beta \neq \beta_0 \end{cases}$$

con $\beta_0 = 1$. Accettare H_0, infatti, equivale a dire che,

$$\text{affitto medio} = \alpha + \text{dimensione}$$

Per verificare il sistema d'ipotesi potremmo utilizzare la statistica test

$$\frac{(\hat{\beta} - \beta_0)}{\sigma\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}}}.$$

Però anche in questo caso noi non conosciamo σ. Quindi, con i dati a disposizione, non posstemmo calcolare il valore osservato della statistica (pur conoscendone la sua distribuzione sotto l'ipotesi nulla).

D'altra parte, poichè abbiamo a disposizione una stima di σ, s, una statistica test analoga è data da

$$T = \frac{(\hat{\beta} - \beta_0)}{S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}}.$$

Se H_0 è vera ci aspettiamo che t_{oss} assuma valori vicini zero. Invece, se H_1 è vera ci aspettiamo che t_{oss} cada lontano da zero.

Si può dimostrare (e non lo faremo) che T ha una distribuzione già incontrata

$$T \sim t_{n-2} \quad (t \text{ di Student con } n-2 \text{ gradi di libertà.})$$

Si noti qui che i gradi di libertà sono pari a $n - 2$.

Esempio 39. *Supponiamo di porre $\alpha = 0{,}05$. Allora*

$$t_{n-2,1-\alpha/2} = t_{23,0,975} = 2{,}07$$

$$\downarrow$$

$$t_{oss} = \frac{1{,}065 - 1}{194{,}6\sqrt{\frac{1}{1999747}}} = 0{,}472$$

$$\downarrow$$

$$-2{,}07 \le 0{,}472 \le 2{,}07 \text{ ?}$$

$$\Big\downarrow \text{ si}$$

accettiamo H_0

E quindi concludiamo che la valutazione dell'agente era plausibile. Possiamo calcolare il livello di significatività osservato come segue:

$$\alpha_{oss} = \Pr(|T| > |t_{oss}|) = 2 \times \{1 - \Pr(t_{23} < 0{,}472)\} \approx 0{,}641$$

Il suo valore indica un supporto molto forte all'ipotesi nulla.

Intervalli di confidenza per $\hat{\beta}$

Un intervallo di confidenza per β può essere determinato, dai risultati precedenti utilizzando lo stesso ragionamento seguito nell'unità 3.2.

Infatti quello che sappiamo è che se β è il vero valore allora

$$\Pr(-t_{n-2,1-\alpha/2} \leq \frac{\hat{\beta}-\beta}{S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \leq t_{n-2,1-\alpha/2}) = 1-\alpha.$$

Ma allora, scrivendo le due disuguaglianze in termini di β, troviamo che

$$\Pr\left(\hat{\beta} - \sqrt{\hat{\text{var}}(\hat{\beta})}\; t_{n-2,1-\alpha/2} \leq \beta \leq \hat{\beta} + \sqrt{\hat{\text{var}}(\hat{\beta})}\; t_{n-2,1-\alpha/2}\right)$$
$$= 1-\alpha$$

dove

$$\sqrt{\hat{\text{var}}(\hat{\beta})} = S\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

ovvero che

$$\left[\hat{\beta} - s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\; t_{n-2,1-\alpha/2}, \hat{\beta} + s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\; t_{n-2,1-\alpha/2}\right]$$

è un intervallo di confidenza di livello $1-\alpha$ per β.

Esempio 40. *Supponiamo, ad esempio, di voler un intervallo di confidenza di livello* 0,95*. Allora,* $t_{n-2,1-\alpha/2} = t_{23,0,975} = 2{,}07$*. Ricordando che* $\hat{\beta} \approx 1{,}065$ *e* $s \approx 194{,}6$ *e quindi che* $\sqrt{\widehat{\mathbb{V}ar}(\hat{\beta})} \approx 194{,}6\sqrt{1/1999747} \approx 0{,}1376$*, la semi-ampiezza dell'intervallo richiesto è*

$$0{,}285 \approx 0{,}1376 \times 2{,}07$$

mentre l'intervallo stesso è

$$[1{,}065 - 0{,}285 \;\; ; \;\; 1{,}065 + 0{,}285] = [0{,}780 \;\; ; \;\; 1{,}350]$$

- Si osservi che l'intervallo include il valore $\beta = 1$. Questo era atteso visto che avevamo visto, con il test discusso precedentemente, che il valore 1 per β era plausibile sulla base dei dati disponibili.

- L'intervallo di confidenza non contiene $\beta = 0$ e quindi nel sistema d'ipotesi
$$\begin{cases} \text{H}_0 : \beta = 0 \\ \text{H}_1 : \beta \neq 0 \end{cases}$$
avremmo rifiutato l'ipotesi nulla in favore dell'ipotesi alternativa con un livello $\alpha = 0{,}05$, il che significa che la dimensione dell'appartamento gioca un ruolo nel determinare l'affitto.

3.14 Reddito e occupazione

- Diagramma di dispersione

- Modello di regressione lineare multipla

- Minimi quadrati

- Inferenza

3.14.1 Il problema

Reddito e occupazione

Abbiamo raccolto dal sito di EuroStat i dati relativi al reddito e all'occupazione di alcuni paesi per l'anno 2008. In particolare

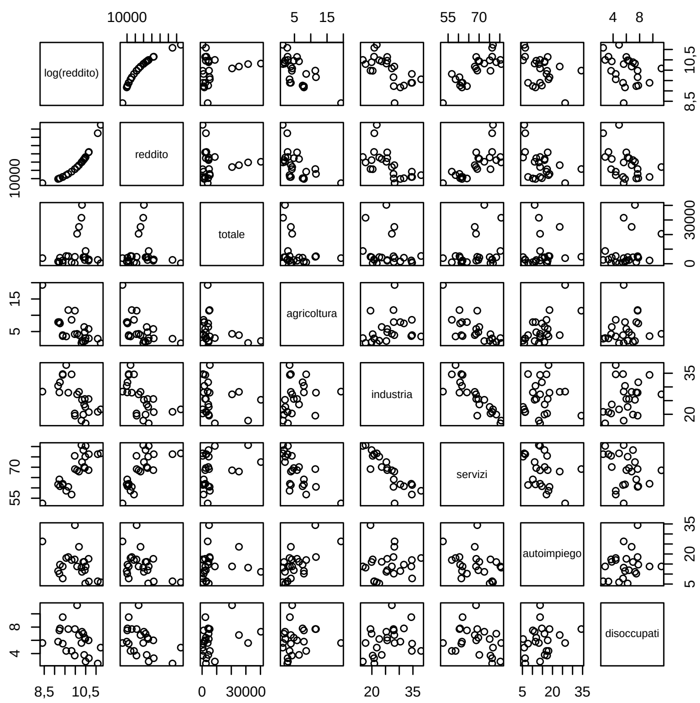
- Paesi: AT BE BG CY CZ DE DK EE ES FI GR HU IE IT LT LU LV NL NO PT SE SI SK UK;

- Reddito pro capite (*Income, saving and net lending / net borrowing - Current prices*, tavola *nama_inc_c*);

- Impiegati totali e percentuale per agricoltura, industria e servizi e percentuale di imprenditori e artigiani (*Employment growth and activity branches - Annual averages*, tavola *lfsi_grt_a*);

- Disoccupati (*Unemployment rate, annual average*, tavola *une_rt_a*);

paese	reddito	totale	agricoltura	industria	servizi	autoimpiego	disoccupati
AT	33800	4239,60	6,40	23,60	70,10	16,10	3,80
BE	32400	4436,30	1,80	19,90	78,30	16,00	7,00
BG	4500	3835,60	19,30	28,30	52,50	26,30	5,60
CY	21400	394,90	4,20	20,30	75,50	17,30	3,70
CZ	14200	5304,70	3,50	38,00	58,60	18,00	4,40
DE	30400	40331,00	2,10	25,30	72,50	11,10	7,30
DK	42300	2922,30	2,90	20,80	76,30	6,30	3,30
EE	12000	642,90	3,90	34,70	61,40	7,80	5,50
ES	24000	20532,20	4,30	27,30	68,50	13,80	11,30
FI	34800	2533,60	4,80	25,60	69,60	11,90	6,40
GR	21600	4758,70	11,40	19,50	69,10	34,40	7,70
HU	10500	4129,60	7,50	31,70	60,80	14,70	7,80
IE	41800	2097,80	5,80	25,60	68,60	17,60	6,00
IT	26300	25262,80	3,90	28,20	67,90	23,60	6,80
LT	9600	1521,50	7,90	30,40	61,70	11,50	5,80
LU	75100	348,80	1,40	21,80	76,70	5,90	4,90
LV	10200	1120,30	7,90	28,00	64,10	10,20	7,50
NL	36200	8743,00	3,00	16,70	80,30	13,80	2,80
NO	65000	2616,00	2,80	20,90	76,30	6,40	2,50
PT	15700	5147,00	11,60	27,90	60,50	18,50	7,70
SE	35600	4559,00	2,20	22,70	75,10	5,30	6,20
SI	18400	990,30	8,60	34,60	56,80	17,00	4,40
SK	12000	2237,10	3,60	34,40	62,00	13,80	9,50
UK	29600	31535,00	1,50	17,70	80,70	13,20	5,60



3.14.2 Introduzione

Definizione 14 (Modello lineare)**.** *Sia $\boldsymbol{Z} = (Y, \boldsymbol{X})$ un vettore casuale, Y una variabile casuale e $\boldsymbol{X}$ un vettore casuale di dimensione k. Un modello lineare è tale che*

1. $Y|\boldsymbol{X} \sim N(\mathbb{E}(Y|\boldsymbol{X}), \sigma^2)$ *(**Componente casuale**);*
2. $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^t\boldsymbol{\beta}$ *(**Componente sistematica**).*

dove $\boldsymbol{\beta}$ è un vettore di lunghezza k.

- La prima assunzione dice che:
  - La distribuzione condizionata $Y|\boldsymbol{X}$ è normale;
  - La varianza condizionata di $Y|\boldsymbol{X}$ non dipende dal valore di $\boldsymbol{X}$ (Ipotesi di omoschedasticità);
- La seconda assunzione ci dice che la media condizionata $Y|\boldsymbol{X}$ è una combinazione lineare di variabili casuali X_i attraverso il vettore di costanti $\boldsymbol{\beta}$ (i parametri del modello) ;
- Non facciamo assunzioni sulla forma della distribuzione del vettore $\boldsymbol{X}$.

Un altro modo per definire un modello lineare

Quando abbiamo un modello lineare, allora

$$\varepsilon = Y - \mathbb{E}(Y|\boldsymbol{X}) \sim N(0, \sigma^2)$$

che è chiamata la componente (stocastica) d'errore. Quindi

$$\begin{aligned} Y &= \mathbb{E}(Y|\boldsymbol{X}) + \varepsilon \\ &= \boldsymbol{x}^t\boldsymbol{\beta} + \varepsilon \end{aligned}$$

che è la formulazione classica del modello lineare. Da questa espressione possiamo vedere che Y può essere decomposto nella forma segnale+rumore.

Esplicitando tale relazione per le singole unità statistiche, esso equivale a

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \ , \qquad i = 1, \cdots, n$$

Modello lineare e osservazioni

Sia $Z = (\boldsymbol{y}, X)$ una matrice di dimensione $n \times (k+1)$ tale che le righe siano una determinazione del vettore casuale $\boldsymbol{Z} = (Y, \boldsymbol{X})$ indipendenti tra di loro. Cioè

$$Z = (\boldsymbol{y}, X) = \begin{bmatrix} y_1 & x_{11} & x_{12} & \cdots & x_{1k} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} = \begin{bmatrix} y_1 & \boldsymbol{x}_1{}^t \\ y_2 & \boldsymbol{x}_2{}^t \\ \vdots & \vdots \\ y_n & \boldsymbol{x}_n{}^t \end{bmatrix}$$

Questo porta a n equazioni osservazionali

$$y_i = \boldsymbol{x}_i{}^t \boldsymbol{\beta} + e_i \qquad 1 \leq i \leq n$$

che possono essere scritte nella seguente forma matriciale

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$$

dove $\boldsymbol{e}$ è un vettore di dimensione n contenente gli n valori non osservati dalla variabile casuale $\boldsymbol{\varepsilon}$. Gli elementi di $\boldsymbol{e}$ sono chiamati **errori**. Mantenendo fissati i valori di X, l'equazione precedente può essere pensata come una determinazione dal modello lineare

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

dove $\underset{n\times 1}{\boldsymbol{Y}}|X \sim N(\mathbb{E}(\boldsymbol{Y}|X), \sigma^2 I)$ and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I)$.

Tirando le somme

- $\boldsymbol{y}$ è un vettore di dimensione $n \times 1$ di osservazioni dal vettore casuale $\boldsymbol{Y}$;
- X è una matrice di dimensione $n \times k$;
- $\boldsymbol{\beta}$ è il vettore dei parametri di dimensione $k \times 1$;
- $\boldsymbol{e}$ è il vettore degli errori di dimensione $n \times 1$ (che è una osservazione dal vettore casuale $\boldsymbol{\varepsilon}$ che noi non osserviamo);
- $\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$;
- $\mathbb{V}ar(\boldsymbol{\varepsilon}) = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t) = \sigma^2 I$;
- $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I)$.

Terminologia

- Y è chiamata la variabile **dipendente**;
- Nei modelli lineari Y è sempre una variabile casuale quantitativa continua e distribuita normalmente;
- $\boldsymbol{X}$ è l'insieme delle variabili **indipendenti**;
- Siccome la distribuzione di $\boldsymbol{X}$ non è particolarmente rilevante nei modelli lineari allora X è considerata spesso come una matrice di dati non stocastica, cioè deterministica, in questo caso diciamo che X è fissata altrimenti che è stocastica;
- Le componenti di $\boldsymbol{X}$ possono essere variabili casuale quantitative o fattori.

Esempio 41. *Date le tre variabili che riguardano il prezzo di un'auto usata*

- *P: Prezzo dell'auto;*
- *A: Età dell'auto (in mesi dalla prima vendita);*
- *T: Tipo di auto (qui ci limitiamo a quattro tipi diversi che indichiamo con le etichette numeriche 0, 1, 2, 3).*

Allora ponendo P come variabile dipendente possiamo avere il seguente modello lineare

$$p_i = \beta_1 a_i + \beta_2 \mathrm{I}_{t_i=0} + \beta_3 \mathrm{I}_{t_i=1} + \beta_4 \mathrm{I}_{t_i=2} + \beta_5 \mathrm{I}_{t_i=3} + e_i$$

dove$\mathrm{I}_{t_i=0}$ *è la variabile indicatrice che vale* 1 *se* $t_i = 0$ *e* 0 *altrimenti. In forma matriciale abbiamo*

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} a_1 & 1 & 0 & 0 & 0 \\ a_2 & 1 & 0 & 0 & 0 \\ a_3 & 0 & 1 & 0 & 0 \\ \cdots & & \cdots & & \\ a_i & 0 & 0 & 1 & 0 \\ \cdots & & \cdots & & \\ a_n & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_5 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

*Uno stesso modello lineare può essere espresso in forme diverse (si parla di diverse **parametrizzazioni** del modello) ad esempio, il modello precedente può essere riscritto nella forma equivalente*

$$p_i = \alpha_0 + \alpha_1 a_i + \alpha_2 \mathrm{I}_{t_i=1} + \alpha_3 \mathrm{I}_{t_i=2} + \alpha_4 \mathrm{I}_{t_i=3} + e_i$$

Il modello è equivalente ma l'interpretazione dei parametri è diversa $\boldsymbol{\alpha}$ *rispetto ai* $\boldsymbol{\beta}$. *La forma matriciale di questo modello è*

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} a_1 & 1 & 0 & 0 & \\ a_2 & 1 & 0 & 0 & \\ a_3 & 1 & 1 & 0 & \\ \cdots & \cdots & & & \\ a_i & 1 & 0 & 1 & 0 \\ \cdots & \cdots & & & \\ a_n & 1 & 0 & 0 & \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Nel modello che abbiamo specificato, il prezzo dipende dall'età dell'auto nella stessa maniera, indipendentemente dal tipo di auto considerata. Possiamo costruire un modello che consenta relazioni diverse tra il prezzo e l'età a seconda del tipo di auto. Ad esempio

$$p_i = \gamma_0 + \gamma_1 a_i + \gamma_2 \,\mathrm{I}_{t_i=1} + \gamma_3 \,\mathrm{I}_{t_i=2} + \gamma_4 \,\mathrm{I}_{t_i=3} + \gamma_5 a_i \,\mathrm{I}_{t_i=3} + e_i$$

consente al modello 3 *di avere una diversa relazione rispetto agli altri tipi. Il termine* $a_i \,\mathrm{I}_{t_i=3}$ *è chiamata* ***interazione***.

3.14.3 Stima dei parametri attraverso il metodo dei minimi quadrati (ordinari)

Introduzione

Una procedura per la stima dei parametri $\boldsymbol{\beta}$ può essere basata sul minimizzare una ragionevole funzione degli errori:

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} M(e_i) = \mathrm{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} M(y_i - \boldsymbol{x}_i{}^t \boldsymbol{\beta}) \ .$$

Esempi sono

1. $M(x) = |x|$;
2. $M(x) = x^2$.

La prima è chiamata norma L1 o *Least Absolute Deviation* (LAD) regressione ed è stata introdotta da Pierre–Simon Laplace (1749–1827) (ma prima ancora da Galileo Galilei (1564–1642) e Ruggiero Giovanni Boscovich (1711–1787)). La seconda è conosciuta come la funzione dei Minimi Quadrati *Least Squares (LS) regression* (ed è la norma L2). È stata introdotta da Johann Carl Friedrich Gauss (1777–1855).

Minimi quadrati ordinari

Principio dei minimi quadrati (ordinari)

Sia $\mathcal{B}$ lo spazio parametrico per $\boldsymbol{\beta}$. In generale abbiamo $\mathcal{B} = \mathbb{R}^k$. L'obiettivo è quello di trovare un vettore $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k) \in \mathcal{B}$ che minimizzi la somma dei quadrati degli errori

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^t\boldsymbol{e} = (\boldsymbol{y} - X\boldsymbol{\beta})^t(\boldsymbol{y} - X\boldsymbol{\beta})$$

fissato il vettore di osservazioni $\boldsymbol{y}$ e la matrice di osservazioni X.

- La funzione $S(\boldsymbol{\beta})$ ha sembre (almeno) un minimo perchè è una funzione reale convessa e differenziabile;
- All'interno del principio dei minimi quadrati non abbiamo bisogno di assumere la normalità della variabile Y. È sufficiente assumere che $\boldsymbol{y}$ sia un campione di osservazioni indipendenti dalla variabile $Y|X = x \sim (E(Y|X = x), \sigma^2)$.

Possiamo manipolare algebricamente $S(\boldsymbol{\beta})$ e ottenere

$$\begin{aligned} S(\boldsymbol{\beta}) &= \boldsymbol{y}^t\boldsymbol{y} - \underbrace{\boldsymbol{\beta}^t X^t \boldsymbol{y}}_{1\times 1} - \underbrace{\boldsymbol{y}^t X \boldsymbol{\beta}}_{1\times 1} + \boldsymbol{\beta}^t X^t X \boldsymbol{\beta} \\ &= \boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{\beta}^t X^t \boldsymbol{y} + \boldsymbol{\beta}^t X^t X \boldsymbol{\beta} \end{aligned}$$

che andremo a differenziare per il vettore $\boldsymbol{\beta}$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta}) &= 2X^t X\boldsymbol{\beta} - 2X^t\boldsymbol{y} \\ \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} S(\boldsymbol{\beta}) &= 2X^t X \qquad \text{(non negative definite)} \end{aligned}$$

Uguagliando a zero la derivata prima otteniamo le **equazioni normali**.

Equazioni normali

$$X^t X\boldsymbol{\beta} = X^t\boldsymbol{y}$$

Ci sono due casi

1. X è una matrice di rango pieno k, allora $\underbrace{X^t X}_{k\times k}$ è non singolare e quindi

$$\hat{\boldsymbol{\beta}} = (X^t X)^{-1} X^t \boldsymbol{y} \ .$$

   Ricordiamo che $r\left((AA^t)\right) = r\left((A^t A)\right) = r\left(A\right) = r\left(A^t\right)$.

2. X non è una matrice di rango pieno allora la soluzione non è unica

$$\hat{\boldsymbol{\beta}} = (X^t X)^{-} X^t \boldsymbol{y} + \left(I - (X^t X)^{-} (X^t X)\right) \boldsymbol{w}$$

dove $(X^t X)^{-}$ è l'inversa generalizzata e $\boldsymbol{w}$ è un vettore arbitrario di dimensione k.

Non discuteremo questo secondo caso, anche se, la sua importanza dal punto di vista dell'analisi dei dati è immenso. Ne vedremo un esempio applicato ai nostri dati sul reddito e l'occupazione. Ricordiamo comunque che

- I valori previsti dal modello $\hat{\boldsymbol{y}}$ sono gli stessi per **tutte** le soluzioni $\hat{\boldsymbol{\beta}}$ delle equazioni normali;
- $S(\boldsymbol{\beta})$ attiene il minimo per ogni soluzione.

Definizione 15 (Valori previsti dal modello)**.** *Data una soluzione dell'equazione normale dei minimi quadrati* $\hat{\boldsymbol{\beta}}$ *abbiamo*

$$\begin{aligned} \hat{\boldsymbol{y}} &= X\hat{\boldsymbol{\beta}} \\ &= X(X^t X)^{-1} X^t \boldsymbol{y} && \text{se } X \text{ è di rango pieno} \\ &= H\boldsymbol{y} \end{aligned}$$

sono i ***valori previsti*** *dal modello, cioè è la previsione, basata sulle osservazioni, di* $\boldsymbol{y}$ *secondo il modello lineare utilizzato.*

Definizione 16 (Residui)**.** *I* ***residui*** $\boldsymbol{r}$ *sono definiti come*

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{y} - \hat{\boldsymbol{y}} \\ &= \boldsymbol{y} - X\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{y} - X(X^t X)^{-1} X^t \boldsymbol{y} && \text{se } X \text{ è di rango pieno} \\ &= (I - H)\boldsymbol{y} \\ &= \hat{\boldsymbol{e}}\ . \end{aligned}$$

Dalla definizione si evince che i residui sono delle stime degli errori $\boldsymbol{e}$ *non oservati.*

Teorema 1. *Le matrici* H *e* $(I - H)$ *sono matrici di proiezione ortogonali.*

Dimostrazione 14. *Dobbiamo mostrare che* H *è idempotente e simmetrica.*

$$H^2 = X\underbrace{(X^t X)^{-1} X^t X}_{I}(X^t X)^{-1} X^t = X(X^t X)^{-1} X^t = H$$

$$H^t = \left(X(X^t X)^{-1} X^t\right)^t = (X^t)^t (X^t X)^{-1^t} X^t = H$$

Valore atteso degli stimatori $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{Y}}$

Calcoliamo il valore atteso degli stimatori, nell'ambito del campionamento ripetuto. I calcoli sono effettuati assumento fissata la matrice X.

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}((X^tX)^{-1}X^t\boldsymbol{Y}) \\ &= (X^tX)^{-1}X^t\mathbb{E}(\boldsymbol{Y}) \\ &= (X^tX)^{-1}X^tX\boldsymbol{\beta} \\ &= \boldsymbol{\beta}\end{aligned}$$

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{Y}}) &= \mathbb{E}(X\hat{\boldsymbol{\beta}}) \\ &= X\mathbb{E}(\hat{\boldsymbol{\beta}}) \\ &= X\boldsymbol{\beta}\ .\end{aligned}$$

Varianza degli stimatori $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{Y}}$

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}) &= (X^tX)^{-1}X^t\mathrm{var}(\boldsymbol{Y})\big((X^tX)^{-1}X^t\big)^t \\ &= (X^tX)^{-1}X^t(\sigma^2 I)X(X^tX)^{-1} \\ &= \sigma^2(X^tX)^{-1}\end{aligned}$$

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{Y}}) &= X\,\mathrm{var}(\hat{\boldsymbol{\beta}})X^t \\ &= \sigma^2X(X^tX)^{-1}X^t \\ &= \sigma^2H\ .\end{aligned}$$

Stima di σ^2

Una stima non distorta di σ^2 è

$$\hat{\sigma^2} = \frac{\|\boldsymbol{r}\|^2}{n-k} = \frac{\sum_{i=1}^n r_i^2}{n-k}\ .$$

Infatti

$$\begin{aligned}
\mathbb{E}(\|\boldsymbol{r}\|^2) &= \mathbb{E}\left(\boldsymbol{Y}^t(I-H)\boldsymbol{Y}\right) \\
&= (X\boldsymbol{\beta})^t\underbrace{(I-H)(X\boldsymbol{\beta})}_{\mathbf{0}} + \mathbb{E}\left(\boldsymbol{\varepsilon}^t(I-H)\boldsymbol{\varepsilon}\right) \\
&= \mathbb{E}\left(\operatorname{tr}\left(\boldsymbol{\varepsilon}^t(I-H)\boldsymbol{\varepsilon}\right)\right) \\
&= \operatorname{tr}\left((I-H)\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t)\right) \\
&= \operatorname{tr}\left((I-H)\sigma^2 I\right) \\
&= \sigma^2(n-k)
\end{aligned}$$

siccome la traccia di una matrice idempotente è pari al suo rango e $r\,(I) = n$, $r\,(H) = k$.

Stimatori BLUE

Il prossimo teorema ci dice che gli stimatori trovati sono i migliori tra gli stimatori corretti lineari (*Best Unbiased Linear Estimator*)

Teorema 2 (Gauss–Markov)**.** *Se* $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *e* $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 I)$ *allora lo stimatore dei minimi quadrati* $\hat{\boldsymbol{\beta}}$ *è tale che*

$$\operatorname{var}(\hat{\boldsymbol{\beta}}) \leq \operatorname{var}(\tilde{\boldsymbol{\beta}})$$

rispetto a qualsiasi altro stimatore $\tilde{\boldsymbol{\beta}}$ *nella forma* $\tilde{\boldsymbol{\beta}} = C\boldsymbol{y}$*, dove* C *è una matrice di dimensione* $k \times n$ *di costanti tali che* $\mathbb{E}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$*.*

Dimostrazione 15. *Siccome* $\boldsymbol{\beta} = \mathbb{E}(C\boldsymbol{Y}) = CX\boldsymbol{\beta}$ *deve valere per tutti i vettori* $\boldsymbol{\beta}$ *allora* $CX = I = X^tC^t$*. Quindi*

$$\begin{aligned}
\underset{k\times k}{\operatorname{var}(\tilde{\boldsymbol{\beta}})} - \underset{k\times k}{\operatorname{var}(\hat{\boldsymbol{\beta}})} &= C\sigma_2 IC^t - \sigma^2(X^tX)^{-1} \\
&= \sigma^2\left(CIC^t - CX(X^tX)^{-1}X^tC^t\right) \\
&= \sigma^2 C(I-H)C^t \\
&= \sigma^2 C(I-H)(I-H)^t C^t\ .
\end{aligned}$$

e questa matrice $k \times k$ *è semidefinita positiva.*

Esempio 42 (Modello lineare semplice)**.** *Scriviamo il modello lineare semplice nella forma matriciale*

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Le stime dei minimi quadrati sono

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} = \begin{bmatrix} n & \sum(x_i-\bar{x}) \\ \sum(x_i-\bar{x}) & \sum(x_i-\bar{x})^2\end{bmatrix}^{-1}\begin{bmatrix}\sum y_i \\ \sum(x_i-\bar{x})y_i\end{bmatrix}$$
$$= \begin{bmatrix} n^{-1} & 0 \\ 0 & \left(\sum(x_i-\bar{x})^2\right)^{-1}\end{bmatrix}\begin{bmatrix}\sum y_i \\ \sum(x_i-\bar{x})y_i\end{bmatrix}$$
$$= \begin{bmatrix} \bar{y} \\ \frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}\end{bmatrix}$$

Se gli x_i sono tutti uguali allora X^tX non è invertibile e $\hat{\beta}_1$ è indeterminato e qualsiasi valore va bene.

La stima non distorta di σ^2 è

$$\hat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \bar{y} - (x_i-\bar{x})\frac{\sum_{j=1}^{n}(x_j-\bar{x})y_j}{\sum_{j=1}^{n}(x_j-\bar{x})^2}\right)^2$$

La matrice di varianza e covarianza dello stimatore $\hat{\boldsymbol{\beta}}$ è

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \sigma^2\begin{bmatrix} n^{-1} & 0 \\ 0 & \left(\sum(x_i-\bar{x})^2\right)^{-1}\end{bmatrix}$$

Esempio 43 (log(Reddito)∼Occupazione nel settore dei servizi)**.**

$$X = \begin{bmatrix} 1 & 70,1 \\ 1 & 78,3 \\ \cdots & \cdots \\ 1 & 62 \\ 1 & 80,7 \end{bmatrix} \quad y = \begin{bmatrix} 10,428 \\ 10,386 \\ \cdots \\ 9,393 \\ 10,296\end{bmatrix}$$

Il vettore delle stime dei coefficienti è

$$\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{bmatrix} = \left(X^tX\right)^{-1}X^ty$$

È necessario per prima cosa calcolare la matrice di dimensione 2×2 (X^tX):

$$\begin{aligned}\left(X^tX\right) &= \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 70,1 & 78,3 & \cdots & 62 & 80,7\end{bmatrix}\begin{bmatrix} 1 & 70,1 \\ 1 & 78,3 \\ \cdots & \cdots \\ 1 & 62 \\ 1 & 80,7 \end{bmatrix} \\ &= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2\end{bmatrix} = \begin{bmatrix} 24 & 1643,9 \\ 1643,9 & 114058,85\end{bmatrix}\end{aligned}$$

Calcoliamo ora $X^t y$:

$$
\begin{aligned}
X^t y &= \left[\begin{array}{ccccc} 1 & 1 & \cdots & 1 & 1 \\ 70,1 & 78,3 & \cdots & 62 & 80,7 \end{array} \right] \left[\begin{array}{c} 10,428 \\ 10,386 \\ \cdots \\ 9,393 \\ 10,296 \end{array} \right] \\
&= \left[\begin{array}{c} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i \; y_i \end{array} \right] = \left[\begin{array}{c} 240,581 \\ 16578,613 \end{array} \right]
\end{aligned}
$$

Ci rimane ora l'inversione di una matrice 2×2, posto:

$$
A = \left[\begin{array}{cc} a & b \\ c & d \end{array} \right]
$$

la matrice aggiunta è:

$$
\mathrm{agg} A = \left[\begin{array}{cc} d & -b \\ -c & a \end{array} \right]
$$

e il determinante è $\det(A) = a \; d - b \; c$, infine:

$$
A^{-1} = \frac{\mathrm{agg} A}{\det(A)}
$$

Nel nostro caso specifico abbiamo:

$$
\begin{aligned}
(X^t X)^{-1} &= \frac{1}{35005,19} \left[\begin{array}{cc} 114058,85 & -1643,9 \\ -1643,9 & 24 \end{array} \right] \\
&= \left[\begin{array}{cc} 24 & 1643,9 \\ 1643,9 & 114058,85 \end{array} \right]
\end{aligned}
$$

da cui:

$$
\left[\begin{array}{c} \hat{\beta}_0 \\ \hat{\beta}_1 \end{array} \right] = \left[\begin{array}{cc} 24 & 1643,9 \\ 1643,9 & 114058,85 \end{array} \right] \left[\begin{array}{c} 240,581 \\ 16578,613 \end{array} \right] = \left[\begin{array}{c} 5,337 \\ 0,068 \end{array} \right]
$$

I valori previsti dal modello possono essere calcolati semplicemente utilizzando la matrice di proiezione ($H = X(X^t X)^{-1} X^t$):

$$
\hat{y} = H y = X(X^t X)^{-1} X^t y
$$

e come differenza tra i valori osservati e i valori previsti otteniamo i residui della regressione: $r_i = y_i - \hat{y}_i$.

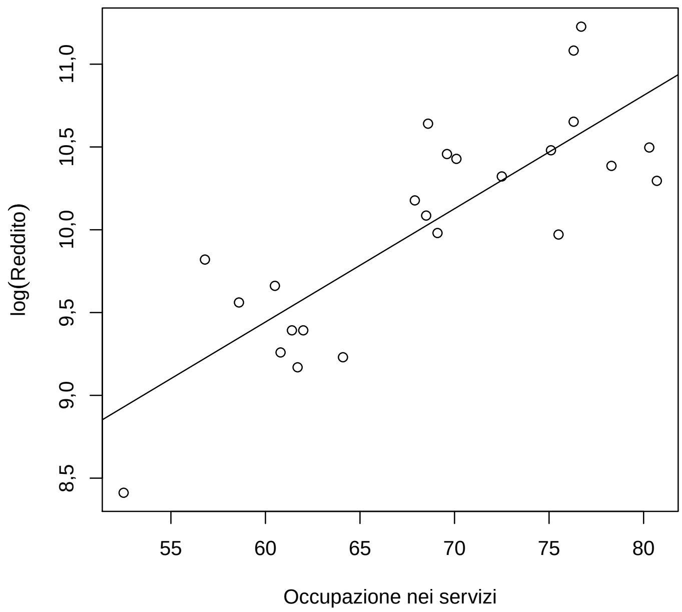


Grafico del modello $\log(Reddito) \sim Servizi$.

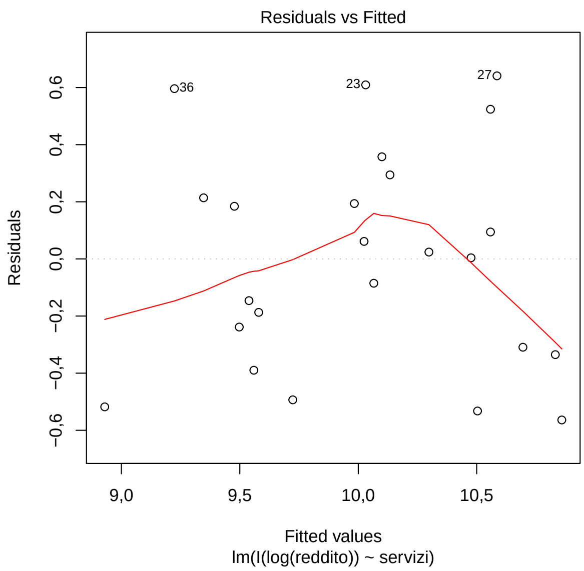


Grafico dei residui rispetto ai valori previsti per il modello $\log(Reddito) \sim Servizi$.

Calcoliamo infine il valore di R^2:

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ &= 1 - \frac{3,35}{10,179} \\ &= 0,671 \end{aligned}$$

3.14.4 Inferenza sui parametri

Primo esempio di inferenza sui parametri

Consideriamo il seguente sistema di ipotesi

$$\begin{cases} H_0 & : \boldsymbol{\beta} = \mathbf{0} \\ H_1 & : \boldsymbol{\beta} \neq \mathbf{0} \end{cases}$$

Una statistica test può essere costruita partendo dall'identità

$$\begin{aligned} \boldsymbol{Y}^t\boldsymbol{Y} &= \boldsymbol{Y}^t H\boldsymbol{Y} + \boldsymbol{Y}^t(I-H)\boldsymbol{Y} \\ &= \hat{\boldsymbol{Y}}^t\hat{\boldsymbol{Y}} + (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^t(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) \\ &= \hat{\boldsymbol{\beta}}^t X^t\boldsymbol{Y} + \boldsymbol{R}^t\boldsymbol{R} \\ SST_Y^* &= SSR_Y^* + SSE_Y^* \end{aligned}$$

dove $\boldsymbol{R}$ è la variabile casuale dei residui.

La statistica test è data da

$$\begin{aligned} T &= \frac{\frac{\boldsymbol{Y}^t H\boldsymbol{Y}}{k}}{\frac{\boldsymbol{Y}^t(I-H)\boldsymbol{Y}}{n-k}} \\ &= \frac{\frac{SSR_Y^*}{k}}{\frac{SSE_Y^*}{n-k}} \end{aligned}$$

la quale è distribuita come una F di Fisher–Snedecor con $k, n-k$ gradi di libertà sotto l'ipotesi nulla. Quindi la regione di rifiuto del test è data dall'intervallo

$$[F_{k,n-k,1-\alpha}, +\infty)\ .$$

La decomposizione nel caso del campione osservato è

$$\begin{aligned}\boldsymbol{y}^t\boldsymbol{y} &= \boldsymbol{y}^t H\boldsymbol{y} + \boldsymbol{y}^t(I-H)\boldsymbol{y}\\ &= \hat{\boldsymbol{y}}^t\hat{\boldsymbol{y}} + (\boldsymbol{y}-\hat{\boldsymbol{y}})^t(\boldsymbol{y}-\hat{\boldsymbol{y}})\\ \sum_{i=1}^n y_i^2 &= \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n r_i^2\\ SST_y^* &= SSR_y^* + SSE_y^*\end{aligned}$$

il valore della statistica osservato, sotto l'ipotesi nulla, è

$$\begin{aligned}t_{oss} &= \frac{\frac{\boldsymbol{y}^t H\boldsymbol{y}}{k}}{\frac{\boldsymbol{y}^t(I-H)\boldsymbol{y}}{n-k}}\\ &= \frac{\frac{SSR_y^*}{k}}{\frac{SSE_y^*}{n-k}}\\ &= \frac{\frac{\sum_{i=1}^n \hat{y}_i^2}{k}}{\frac{\sum_{i=1}^n r_i^2}{n-k}}\ .\end{aligned}$$

Significatività del modello rispetto al modello con la sola intercetta

Supponiamo di voler confrontare i due seguenti modelli

$$\begin{aligned}M_0 :& Y = \beta_0 + \varepsilon_0\\ M_1 :& Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon_1\end{aligned}$$

Il problema può essere riscritto in forma di sistema d'ipotesi

$$\begin{cases} H_0 & : \beta_1 = \beta_2 = \cdots = \beta_p = 0\\ H_1 & : \text{ almeno un coefficiente è diverso da zero}\end{cases}$$

e la stasistica test può essere costruita a partire dalla scomposizione

$$\begin{aligned}\boldsymbol{Y}^t(I-\frac{1}{n}J)\boldsymbol{Y} &= \boldsymbol{Y}^t\left(H-\frac{1}{n}J\right)\boldsymbol{Y} + \boldsymbol{Y}^t(I-H)\boldsymbol{Y}\\ SST_Y &= SSR_Y + SSE_Y\end{aligned}$$

dove J è la matrice in cui ogni elemento è pari a 1.

La statistica test è data da

$$\begin{aligned} T &= \frac{\frac{\boldsymbol{Y}^t(H-\frac{1}{n}J)\boldsymbol{Y}}{p}}{\frac{\boldsymbol{Y}^t(I-H)\boldsymbol{Y}}{n-p-1}} \\ &= \frac{\frac{SSR_Y}{p}}{\frac{SSE_Y}{n-p-1}} \end{aligned}$$

la quale è distribuita come una F di Fisher–Snedecor con $p, n-p-1$ gradi di libertà sotto l'ipotesi nulla. Quindi la regione di rifiuto del test è data dall'intervallo

$$[F_{p,n-p-1,1-\alpha}, +\infty) \ .$$

La decomposizione nel caso del campione osservato è

$$\begin{aligned} \boldsymbol{y}^t(I-\frac{1}{n}J)\boldsymbol{y} &= \boldsymbol{y}^t(H-\frac{1}{n}J)\boldsymbol{y} + \boldsymbol{y}^t(I-H)\boldsymbol{y} \\ \sum_{i=1}^n (y_i-\bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i-\bar{y})^2 + \sum_{i=1}^n r_i^2 \\ SST_y &= SSR_y + SSE_y \end{aligned}$$

il valore della statistica osservato, sotto l'ipotesi nulla, è

$$\begin{aligned} t_{oss} &= \frac{\frac{\boldsymbol{y}^t(H-\frac{1}{n}J)\boldsymbol{y}}{p}}{\frac{\boldsymbol{y}^t(I-H)\boldsymbol{y}}{n-p-1}} \\ &= \frac{\frac{SSR_y}{p}}{\frac{SSE_y}{n-p-1}} \\ &= \frac{\frac{\sum_{i=1}^n(\hat{y}_i-\bar{y})^2}{p}}{\frac{\sum_{i=1}^n r_i^2}{n-p-1}} \ . \end{aligned}$$

La procedura può essere riassunta in quella che viene chiamata Tavola di **Analisi della varianza** (ANOVA)

Fonte di variabilità	Somma dei quadrati	Gradi di libertà	Media della somma dei quadrati	t_{oss}
Regressione	SSR	p	$MSR = \frac{SSR}{p}$	MSR/MSE
Errore	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	
Totale	SST	$n-1$		

Caso generale

Consideriamo adesso il caso generale

$$\begin{cases} H_0 & : C\boldsymbol{\beta} = \boldsymbol{c} \\ H_1 & : C\boldsymbol{\beta} \neq \boldsymbol{c} \end{cases}$$

dove C è una matrice di costanti note di dimensione $m \times k$ per la quale $r(C) < k$ mentre $\boldsymbol{c}$ è un vettore di costanti note di dimensione $m \times 1$.

Casi particolari sono

- se C è un vettore riga con il solo elemento j–esimo pari a 1 e tutti gli altri pari a 0. mentre $\boldsymbol{c} = \mathbf{0}$, il test è equivalente alla verifica della significatività (rispetto a zero) del singolo parametro β_j, $j = 1, \cdots, k$;
- se occorre verificare la significatività dell'intera regressione (dove la prima variabile è l'intercetta del modello) basta porre $\boldsymbol{c} = \mathbf{0}$ e la matrice $C = [\mathbf{0}|I_{k-1}]$, cioè con prima colonna nulla e le seguenti identiche a quelle della matrice identità di ordine $k-1$;
- per verificare la significatività di un gruppo di parametri, basta organizzare la matrice C in modo che sia nullo quel gruppo, oltre a $\boldsymbol{c} = \mathbf{0}$;
- se si intende verificare l'ipotesi che un gruppo di parametri verifichi una relazione lineare (per esempio $c_1\beta_2 + c_2\beta_3 = c_3$), basta modificare C in modo che sia nulla tranne che in corrispondenza dei parametri indicati (ponendo pari a c_3 l'elemento corrispondente di $\boldsymbol{c}$).

$$\begin{bmatrix} 0 & c_1 & c_2 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \dots \\ \beta_k \end{bmatrix} = c_3$$

La statistica test è data da

$$T = \frac{\frac{(C\hat{\beta}-\boldsymbol{c})^t\left[C(X^tX)^{-1}C^t\right](C\hat{\beta}-\boldsymbol{c})}{m}}{\frac{\boldsymbol{Y}^t(I-H)\boldsymbol{Y}}{n-k}}$$

la quale è distribuita come una F di Fisher–Snedecor con $m, n-k$ gradi di libertà sotto l'ipotesi nulla. Quindi la regione di rifiuto del test è data dall'intervallo

$$\left[F_{m,n-k,1-\alpha}, +\infty\right) .$$

il valore della statistica osservato, sotto l'ipotesi nulla, è

$$t_{oss} = \frac{\frac{(C\hat{\beta}-\boldsymbol{c})^t\left[C(X^tX)^{-1}C^t\right](C\hat{\beta}-\boldsymbol{c})}{m}}{\frac{\boldsymbol{y}^t(I-H)\boldsymbol{y}}{n-k}}$$

Esempio 44. *Un caso importante è la verifica su di una componente soltanto, ad esempio la r–esima ($1 \leq r \leq k$), cioè*

$$\begin{cases} H_0 & : \beta_r = 0 \\ H_1 & : \beta_r \neq 0 \end{cases}$$

Per questo sistema di ipotesi abbiamo

$$\underset{1\times k}{C} = (0, 0, \cdots, 1, \cdots, 0, 0)$$

dove il valore 1 è in posizione r e $\mathbf{c} = \mathbf{0}$. Quindi abbiamo

$$\begin{aligned} (C\hat{\beta} - \mathbf{c}) &= \hat{\beta}_r - 0 \\ C(X^tX)^{-1}C^t &= v_{rr} \end{aligned}$$

dove v_{rr} è l'elemento di posizione (r, r) della matrice $V = (X^tX)^{-1}$. Allora

$$t_{oss} = \frac{\hat{\beta}_r^2}{\hat{\sigma}^2 v_{rr}}$$

e T ha distribuzione F di Fisher–Snedecor con $(1, n-k)$ gradi di libertà. Ricordatevi che $\hat{\sigma}^2$ è uno stimatore non distorto di σ^2.

La radice quadrata della statistica test è

$$\sqrt{t_{oss}} = \frac{\hat{\beta}_r}{\hat{\sigma}\sqrt{v_{rr}}}$$

che è la statistica analoga a quella che abbiamo visto nel caso del modello lineare semplice ed ha distribuzione t di Student con $n-k$ gradi di libertà.

Esempio 45 (Reddito e Occupazione)**.** *Consideriamo il seguente modello*

$$\log(Reddito) = \beta_0 + \beta_1 Servizi + \beta_2 Agricoltura + e$$

otteniamo la statistica test per verificare la significatività rispetto al modello

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6,3746	1,0871	5,86	0,0000
servizi	0,0560	0,0142	3,94	0,0007
agricoltura	-0,0339	0,0273	-1,24	0,2281

con la sola intercetta vale

$$t_{oss} = \frac{\frac{SSR_y}{p}}{\frac{SSE_y}{n-p-1}} = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{p}}{\frac{\sum_{i=1}^{n} r_i^2}{n-p-1}}$$

$$= \frac{\frac{7{,}058}{2}}{\frac{3{,}121}{24-2-1}} = 23,746$$

Dalle tavole della F con $2,21$ gradi di libertà si ottiene il quantile di ordine 0.95 *pari a* $3,467$ *e quindi si accetta l'ipotesi alternativa. E il p–valore è*

$$\alpha_{oss} = \Pr(T > t_{oss}) = \Pr(F_{2,21} > 23,746) \approx 0$$

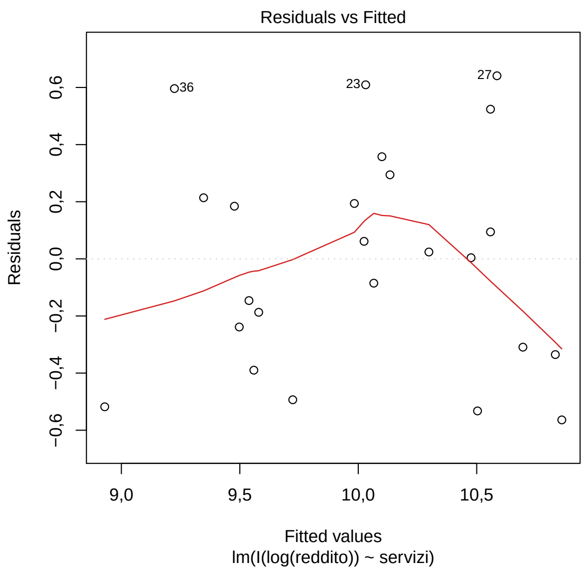


Grafico dei residui rispetto ai valori previsti per il modello $\log(\textit{Reddito}) \sim \textit{Servizi} + \textit{Agricoltura}$.

3.15 Il disastro del Challenger

Introduzione alla regressione logistica

Il disastro del Challenger

- La navicella spaziale Space Shuttle Challenger esplose il 28 gennaio 1986 dopo 73 secondi dal suo decollo da Capo Canaveral in Florida.
- Nell'esplosione morirono tutti e sette i membri dell'equipaggio, fra cui Sharon Christa McAuliffe, un'insegnante delle scuole superiori di scienze che avrebbe dovuto tenere la prima lezione di scienze dallo spazio...

L'esplosione

Lo Space Shuttle Challenger

Le immagini dell'esplosione mostrarono che l'incidente avvenne a causa di un cedimento del grande razzo a propellente solido (SRB, *Solid Rocket Boosters*) situato a destra dello Shuttle.

La commissione Rogers

- Vennero esaminate diverse tonnellate di rottami recuperati nell'Atlantico nei sette mesi successivi al disastro.

- L'operazione di recupero dei rottami coinvolse 22 navi, 6 sommergibili e 33 aerei in un'area di 241.000 km quadrati e fino a 370 metri di profondità.

- I rottami suggerirono che il cedimento fu causato dal guasto di una coppia di *o-ring* nel SRB destro (gli *o-ring* sono dei giunti in gomma non così diversi dalle guarnizioni delle moke di caffé ma di dimensioni molto più grandi).

- Nel Challenger c'erano 6 coppie di *o-ring* ognuna formata da un *o-ring* primario ed uno secondario. Il danneggiamento di entrambi gli *o-ring* di una certa coppia può provocare la fuoriuscita di fiamme dal SRB.

- Questo è proprio ciò che avvenne nel disastro del 1986: le fiamme fuoriuscite dal SRB fecero cedere il serbatoio esterno contenente idrogeno ed ossigeno liquido e quindi la navicella esplose.

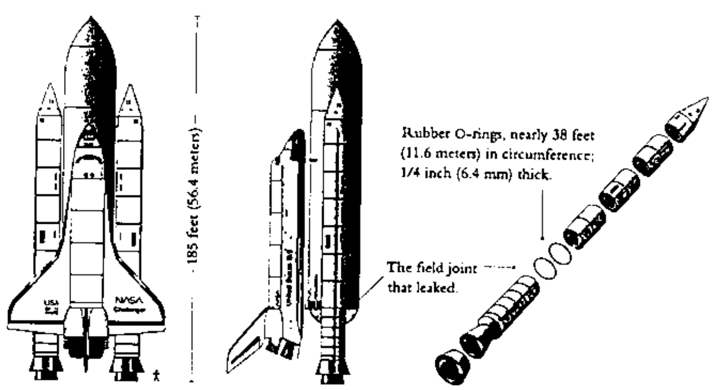
Faceva troppo freddo?

- Venne avanzata l'ipotesi che il danneggiamento degli *o-ring* fosse legato alla temperatura eccezionalmente bassa al momento del lancio pari a soli 31 gradi Fahrenheit, ovvero circa -0.5 gradi Celsius.

- Con una tale temperatura, gli *o-ring* perdono flessibilità e quindi si rompono più facilmente.

- Sebbene non vi fu completo accordo su questa spiegazione, o sul fatto che questa fosse l'unica spiegazione dell'esplosione, in seguito si decise di rafforzare queste guarnizioni e di aumentarle: ora, anziché coppie, vi sono gruppi di tre *o-ring*!

Challenger O-Rings



I dati

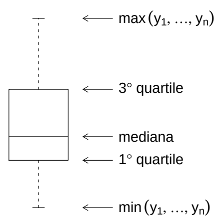
- Dati disponibili riguardo a 23 dei 24 precedenti lanci.
- In sette di questi vi furono danni all'*o-ring* primario, ma mai a quello secondario.
- Di seguito sono riportati il numero di danni, oltre alla temperatura e la pressione al momento del lancio.

Flight	Date	Number of O-rings with thermal distress, r	Temperature (°F) x_1	Pressure (psi) x_2
1	21/4/81	0	66	50
2	12/11/81	1	70	50
3	22/3/82	0	69	50
5	11/11/82	0	68	50
6	4/4/83	0	67	50
7	18/6/83	0	72	50
8	30/8/83	0	73	100
9	28/11/83	0	70	100
41-B	3/2/84	1	57	200
41-C	6/4/84	1	63	200
41-D	30/8/84	1	70	200
41-G	5/10/84	0	78	200
51-A	8/11/84	0	67	200
51-C	24/1/85	2	53	200
51-D	12/4/85	0	67	200
51-B	29/4/85	0	75	200
51-G	17/6/85	0	70	200
51-F	29/7/85	0	81	200
51-I	27/8/85	0	76	200
51-J	3/10/85	0	79	200
61-A	30/10/85	2	75	200
61-B	26/11/86	0	76	200
61-C	21/1/86	1	58	200
61-I	28/1/86	—	31	200

Tabella tratta da Davison (2005).

Boxplot

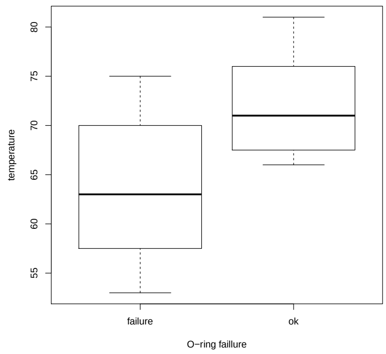
- Il nome è la forma contratta di **box and whiskers plot**.
- Il boxplot è costituito da una *scatola* e da due *baffi* costruiti in accordo al disegno sottostante.



- Il boxplot è costruito tramite le seguenti quantità: **mediana**: valore $y_{0.5}$ tale che il 50% delle osservazioni sono inferiori a $y_{0.5}$ e 50% sono superiori; **primo quartile**: valore $y_{0.25}$ tale che il 25% delle osservazioni sono inferiori a $y_{0.25}$ e 75% sono superiori; **terzo quartile**: valore $y_{0.75}$ tale che il 75% delle osservazioni sono inferiori a $y_{0.75}$ e 25% sono superiori.

O-ring e temperatura

Per partire osserviamo come varia la temperatura nei lanci in cui almeno un *o-ring* ha avuto un guasto rispetto ai lanci in cui non vi sono stati danni.



Cosa notate?

Un modello statistico

- Nei precedenti lanci la temperatura più bassa fu 53 gradi Fahrenheit.
- La domanda che ci poniamo è se data la temperatura eccezionalmente bassa per Cape Canaveral, 31 gradi Fahrenheit, non fosse stato il caso di annullare il lancio.
- Per rispondere a questa domanda abbiamo bisogno di un modello che, sulla base dell'informazione contenuta nei precedenti 23 lanci, ci dia una **previsione** di ciò che poi successe.
- In altri termini, abbiamo bisogno di costruire un modello che spieghi la variabile Y = "guasto di almeno un *o-ring*" sulla base della variabile X = temperatura.
- La variabile Y gioca il ruolo di variabile **risposta** e la variabile X quello di variabile **esplicativa**.
- Immagineremo che le quantità osservate (almeno un guasto e temperatura al lancio) siano delle realizzazioni di due variabili casuali.

Il modello logistico

- La variabile risposta è una variabile binaria che assume valore 0 se tutti gli *o-rings* sono intatti e 1 se almeno un *o-ring* subisce un guasto. Quindi, sembra ragionevole descrivere i 23 lanci precedenti come 23 variabili casuali Bernoulliane indipendenti con

$$p_i = \Pr(Y_i = 1) \ .$$

- Il nostro obiettivo è descrivere p_i in funzione della temperatura al lancio

$$p_i = f(x_i) \ ,$$

  dove $f(\cdot)$ è un'*opportuna* funzione.

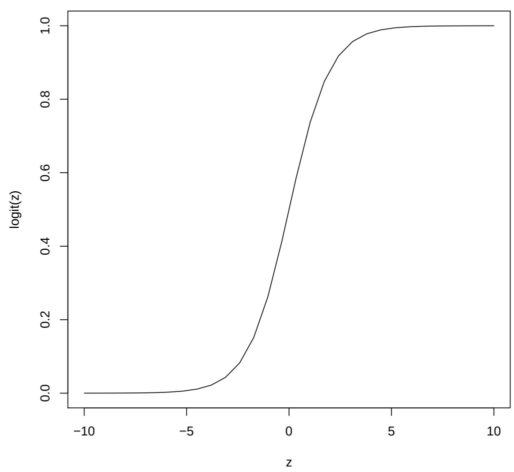
- Che caratteristica deve avere questa funzione per essere *opportuna*? Deve assumere valori fra 0 e 1!

- C'è letteralmente un'infinità di funzioni $f(\cdot)$ che potremmo scegliere, ma alcune sono privilegiate per motivi teorici.

- Fra queste funzioni privilegiate, la più importante (per motivi teorici che ci porterebbero troppo lontani nella discussione) è sicuramente la **funzione logistica**

$$f(z) = \text{logit}(z) = \frac{e^z}{1+e^z}.$$

La funzione logistica



Un modello logistico per il disastro del Challenger

- Ora dobbiamo decidere come strutturare il **modello logistico** che spiega il numero di *o-ring* danneggiati in funzione della temperatura.

- Una possibile scelta potrebbe essere

$$p_i = \frac{e^{\alpha+\beta \text{temp}_i}}{1+e^{\alpha+\beta \text{temp}_i}}$$

- Abbiamo espresso la probabilità sopra in termini di due **parametri** α e β. In particolare, siamo interessati al valore del parametro β che misura l'effetto differenziale dovuto alla temperatura.

- Questo è un esempio di **modello di regressione** o, più precisamente, di modello di regressione **non lineare**.

La funzione di verosimiglianza

- Possiamo scrivere la probabilità di osservare la sequenza degli esiti dei 23 lanci:

$$L(\alpha, \beta) = \underbrace{\left(\frac{1}{1 + e^{\alpha + 66\beta}} \right)}_{\text{primo lancio}} \times \underbrace{\left(\frac{e^{\alpha + 70\beta}}{1 + e^{\alpha + 70\beta}} \right)}_{\text{secondo lancio}} \times \cdots \times \underbrace{\left(\frac{e^{\alpha + 58\beta}}{1 + e^{\alpha + 58\beta}} \right)}_{\text{23-esimo lancio}}$$

- Al variare di α e β cambia il valore della probabilità della sequenza osservata.

- Per esempio, se
  - $\alpha = 11{,}3$ e $\beta = -0{,}5$ allora $\log\{L(\alpha, \beta)\} = -143{,}90$
  - $\alpha = 5$ e $\beta = 0$ allora $\log\{L(\alpha, \beta)\} = -80{,}15$
  - $\alpha = 9$ e $\beta = -0{,}1$ allora $\log\{L(\alpha, \beta)\} = -31{,}92$
  - ...

- Come scegliamo α e β? Sembra ragionevole prendere quella coppia di valori che massimizza la funzione $L(\alpha, \beta)$. Ovvero la coppia di valori che massimizza la probabilità della sequenza osservata.

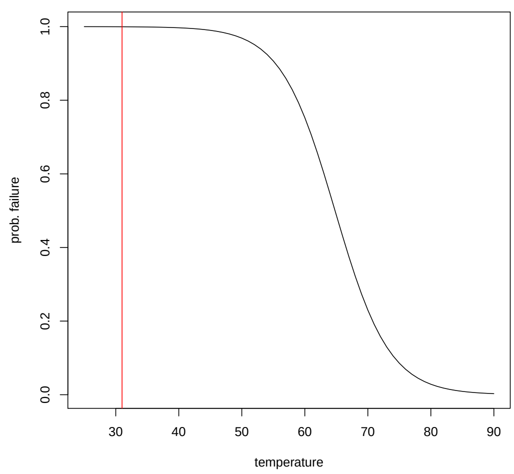
- I valori di questa coppia sono calcolabili con un computer e valgono $\hat{\alpha} = 15{,}043$ e $\hat{\beta} = -0{,}232$. Questi due valori sono i più *verosimili* alla luce dei 23 lanci osservati, ovvero quelli che ricevono maggior supporto sperimentale.

- I valori $\hat{\alpha}$ e $\hat{\beta}$ sono chiamati **stime di massima verosimiglianza** e la funzione $L(\alpha, \beta)$ è detta **funzione di verosimiglianza** (in inglese *Likelihood*).

Il modello stimato

- Il modello stimato è quindi

$$\Pr(\text{rottura } o\text{-}ring) = \frac{e^{15{,}043-0{,}232\text{temp}}}{1+e^{15{,}043-0{,}232\text{temp}}}$$



- Con questo modello la previsione di un danno con una temperatura di 31 gradi Fahrenheit è

$$\Pr(\text{rottura } o\text{-}ring) = \frac{e^{15{,}043-0{,}232\times 31}}{1+e^{15{,}043-0{,}232\times 31}} = 0{,}9996,$$

cioé il 99,96%!

Principio del campionamento ripetuto

- Abbiamo ottenuto una previsione per un danno di almeno un *o-ring* sulla base dei dati osservati.

- Possiamo pensare ai dati osservati come ad una delle possibili sequenze osservabili. Se potessimo tornare indietro nel tempo e “rivivere” i 23 lanci, allora potremmo osservare dei risultati diversi.

- Per esempio, il primo lancio avvenne alla temperatura di 66 gradi F e non avvenne alcuna rottura. Il nostro modello ci dice che la probabilità di rottura a quella temperatura è

$$\text{Pr(rottura } \textit{o-ring} \text{ — temp=66 F)} = \frac{e^{15{,}043-0{,}232\times 66}}{1+e^{15{,}043-0{,}232\times 66}} = 0{,}4305.$$

- Quindi, ci aspettiamo in una ipotetica replicazione dei lanci che nel 43% di questi avvenga almeno una rottura nel primo lancio.

- L'idea che abbiamo presentato viene chiamata **principio del campionamento ripetuto** e possiamo sfruttarla per valutare l'incertezza della nostra previsione.

- In altri termini, possiamo chiederci come cambierebbe la nostra previsione se potessimo tornare indietro nel tempo ed osservare nuovamente i 23 lanci un'infinità di volte.

- Se ripetiamo questo "salto indietro" per un'infinità di volte, otteniamo in corrispondenza ad ogni "nuova" sequenza di lanci una diversa previsione. Otteniamo, quindi, un'infinità di diverse previsioni.

Simulazioni

- Non possiamo tornare indietro nel tempo, però possiamo **simulare** con un computer tutte le sequenze di lanci che vogliamo. Nella tabella seguente sono riportati i dati osservati e tre sequenze simulate.

	y	temp	sim1	sim2	sim3	sim4
1	0	66	0	1	1	1
2	1	70	0	0	0	1
3	0	69	0	0	0	0
4	0	68	1	1	0	0
5	0	67	1	1	1	0
6	0	72	0	0	0	0
7	0	73	0	0	0	0
8	0	70	0	0	0	1
9	1	57	1	1	1	0
10	1	63	1	1	0	1
11	1	70	0	0	0	0
12	0	78	0	0	0	0
13	0	67	0	0	0	1
14	1	53	1	1	1	1
15	0	67	0	0	0	1
16	0	75	0	0	0	0
17	0	70	0	1	0	0
18	0	81	0	0	0	0
19	0	76	0	0	0	0
20	0	79	0	0	0	0
21	1	75	0	0	0	0
22	0	76	0	0	0	0
23	1	58	1	1	1	1

- Quindi, per ogni sequenza simulata possiamo ricalcolare la nostra previsione in corrispondenza a 31 gradi F.

- Per esempio, con la prima sequenza simulata la nostra previsione diventa 1 (100%), con la 45-esima 0.9797, ...

- Ripetendo queste simulazioni 1.000 volte, abbiamo ottenuto che solo il 2,5% delle volte la previsione di rottura a temperatura 31 era inferiore a 0,865 e che la previsione più piccola era 0,6745.

- Con questi risultati possiamo ritenere che il lancio quella mattina fosse ad altissimo rischio!

Bibliografia

A. Azzalini. *Inferenza Statistica.* Springer Italia, Milano, seconda edition, 2002. ISBN 88–470–0130–7.

G. Casella and R.C. Berger. *Statistical Inference.* Duxbury Press, seconda edition, 2002. ISBN 0–534–24312–6.

G. Cicchitelli. *Probabilità e Statistica.* Maggioli editore, Rimini, seconda edition, 2001. ISBN 88–387–1672.2.

G. Cicchitelli and M.A. Pannone. *Complementi ed esercizi di Statistica descrittiva e inferenziale.* Maggioli, Rimini, 1991.

M. Grigoletto and L. Ventura. *Statistica per le Scienze Economiche. Esercizi con richiami di teoria.* Giappichelli, Torino, 1998.

R. Guseo. *Istituzioni di Statistica – Lezioni.* CEDAM, 1997. ISBN 88–13-20425–6.

S.M. Iacus and G. Masarotto. *Laboratorio di Statistica con R.* McGraw–Hill, 2007. ISBN 88–386–6369–6.

Montanari, Agati, and Calò. *Statistica, con esercizi commentati e risolti.* Masson, Open, Milano, 1998.

A.M. Mood, F.A. Graybill, and D.C. Boes. *Introduzione alla Statistica.* McGraw–Hill, 1991. ISBN 88–386–0661–7.

L. Pace and A. Salvan. *Introduzione alla Statistica - I. Statistica descrittiva.* Cedam, Padova, 1996. ISBN 88–13–19939–2.

L. Pace and A. Salvan. *Introduzione alla Statistica - II. Inferenza, verosimiglianza, modelli.* Cedam, Padova, 2001.

F. Parpinel and S. Campostrini. *Introduzione all'inferenza statistica, teoria e esercizi.* Decibel, Zanichelli, 1996.

F. Parpinel and C. Provasi. *Elementi di probabilità e statistica per le scienze economiche*. Giappichelli, Torino, 2004.

D. Piccolo. *Statistica*. il Mulino, 1998. ISBN 88–15–07596–8.

D. Piccolo. *Statistica per le decisioni*. il Mulino, 2004. ISBN 88–15–09770–8.

Appendice A

Lettere dell'alfabeto greco antico

Lettere dell'alfabeto greco antico e loro pronuncia internazionale (e in italiano)

A	α	alpha
B	β	beta
Γ	γ	gamma
Δ	δ	delta
E	ϵ, ε	epsilon
Z	ζ	zeta
H	η	eta
Θ	θ, ϑ	theta
I	ι	iota
K	κ	kappa
Λ	λ	lambda
M	μ	mu (my, mi)
N	ν	nu (ny, ni)
Ξ	ξ	xi
O	o	omicron
Π	π, ϖ	pi
P	ρ, ϱ	rho
Σ	σ, ς	sigma
T	τ	tau
Υ	υ	upsilon (ypsilon)
Φ	ϕ, φ	phi
X	χ	chi
Ψ	ψ	psi
Ω	ω	omega

Appendice B

Strumenti di base

Sommatoria

Siano a_i e b_i, $i = 1, \ldots, n$, due serie di numeri e c una costante qualsiasi. Si definisce la **sommatoria** nel modo seguente:

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \ldots + a_n \,.$$

Alcune proprietà

1. $\sum_{i=1}^{n} c = c + c + \ldots + c = n\, c$.
2. $\sum_{i=1}^{n} c\, a_i = c\, a_1 + c\, a_2 + \ldots + c\, a_n = c\,(a_1 + a_2 + \ldots + a_n) = c \sum_{i=1}^{n} a_i$.
3. $\sum_{i=1}^{n} (a_i + b_i) = \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$.
4. $\left(\sum_{i=1}^{n} a_i\right)^c \neq \sum_{i=1}^{n} a_i^c$. Ad esempio, con $n = 2$ e $c = 2$, si ha $\left(\sum_{i=1}^{n} a_i\right)^2 = (a_1 + a_2)^2 = a_1^2 + 2\, a_1\, a_2 + a_2^2 \neq a_1^2 + a_2^2 = \sum_{i=1}^{n} a_i^2$.
5. $\sum_{i=1}^{n} (a_i\, b_i) \neq \sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i$. Ad esempio, con $n = 2$, si ha $\sum_{i=1}^{n} (a_i\, b_i) = a_1\, b_1 + a_2\, b_2 \neq (a_1 + a_2)\,(b_1 + b_2) = a_1\, b_1 + a_1\, b_2 + a_2\, b_1 + a_2\, b_2 = \sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i$.

La somma dei primi n interi[1]

Un caso famoso è quello della somma dei primi n numeri

$$\sum_{i=1}^{n} i = 1 + 2 + \cdots + n = \frac{n(n+1)}{2} \,.$$

[1] Si veda ad esempio Polymath

La dimostrazione più rapida è quella di Gauss, che secondo un aneddoto determinò in poco tempo, durante le elementari, la somma dei primi 100 numeri naturali, rispondendo alla richiesta del suo maestro che lo voleva tenere impegnato per un po' di tempo per potersi dedicare anche agli altri bambini, osservando che le somme del primo e dell'ultimo, del secondo e del penultimo, e così via, sono uguali. La dimostrazione per induzione della formula di Gauss, chiamata $P(n)$, è la seguente:

- (base) Per $n = 1$ si ha $1 = (1/2)1(1+1)$ e il secondo membro si riduce a 1.
- (passo induttivo) Ammesso per ipotesi induttiva $P(n)$ cioè che

$$P(n) = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \ .$$

si ha

$$\sum_{i=1}^{n+1} i = \sum_{i=1}^{n} i + n + 1 = \frac{n(n+1)}{2} + n + 1$$

e il secondo membro con facili calcoli si riduce a $(n+1)(n+2)/2$, e quindi si ha

$$P(n+1) = \sum_{i=1}^{n+1} i = \frac{(n+2)(n+1)}{2} \ .$$

Produttoria

Siano a_i e b_i, $i = 1, \ldots, n$, due serie di numeri e c una costante qualsiasi. Si definisce la **produttoria** nel modo seguente:

$$\prod_{i=1}^{n} a_i = a_1\, a_2\, \ldots a_n \ .$$

Alcune proprietà

1. $\prod_{i=1}^{n} c = c\, c\, \ldots c = c^n$.
2. $\prod_{i=1}^{n} c\, a_i = c\, a_1\, c\, a_2 \ldots c\, a_n = c^n\,(a_1\, a_2 \ldots a_n) = c^n \prod_{i=1}^{n} a_i$.
3. $\prod_{i=1}^{n}(a_i\, b_i) = \prod_{i=1}^{n} a_i \prod_{i=1}^{n} b_i$.
4. $\left(\prod_{i=1}^{n} a_i\right)^c = \prod_{i=1}^{n} a_i^c$.
5. $\prod_{i=1}^{n}(a_i + b_i) \neq \prod_{i=1}^{n} a_i + \prod_{i=1}^{n} b_i$.

Logaritmo

Il logaritmo in base b del numero reale positivo x, $l = \log_b x$, si definisce come l'esponente l da dare a b per ottenere x, ossia: $x = b^l$. Le basi più comunemente utilizzate sono quella *naturale*, ossia il numero neperiano $e = 2.7183\ldots$, e la base 10.

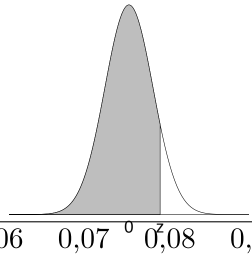
Alcune proprietà

1. $\log_b 1 = 0\,.$
2. $\log_b b = 1\,.$
3. $\lim_{x\to 0^+} \log_b x = -\infty$; $\lim_{x\to +\infty} \log_b x = +\infty$
4. $\log_b(x\,y) = \log_b x + \log_b y$
   (più in generale, $\log_b\left(\prod_{i=1}^n x_i\right) = \sum_{i=1}^n \log_b x_i$).
5. $\log_b(x^c) = c\log_b x$.
6. $\log_b(x/y) = \log_b(x\,y^{-1}) = \log_b x + \log_b y^{-1} = \log_b x - \log_b y$.
7. Cambio di base: $\log_b x = \log_b c \log_c x$.

Appendice C

Tavole di alcune variabili casuali

C.1 Funzione di ripartizione della distribuzione normale standard

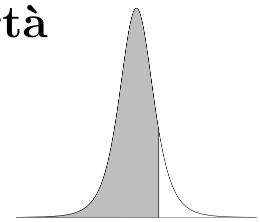


$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\ dx$$

	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

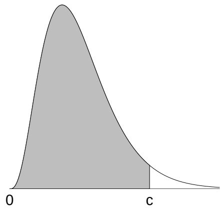
C.2 Alcuni quantili della distribuzione t di Student con r gradi di libertà

$$F_r(t) = \int_{-\infty}^{t} \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\Gamma[r/2](1+x^2/r)^{(r+1)/2}}\ dx$$



	0,6	0,75	0,9	0,95	0,975	0,99	0,995
1	0,3249	1,0000	3,0777	6,3138	12,7062	31,8205	63,6567
2	0,2887	0,8165	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,2767	0,7649	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,2707	0,7407	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,2672	0,7267	1,4759	2,0150	2,5706	3,3649	4,0321
6	0,2648	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,2632	0,7111	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,2619	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,2610	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,2602	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,2596	0,6974	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,2590	0,6955	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,2586	0,6938	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,2582	0,6924	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,2579	0,6912	1,3406	1,7531	2,1314	2,6025	2,9467
16	0,2576	0,6901	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,2573	0,6892	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,2571	0,6884	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,2569	0,6876	1,3277	1,7291	2,0930	2,5395	2,8609
20	0,2567	0,6870	1,3253	1,7247	2,0860	2,5280	2,8453
21	0,2566	0,6864	1,3232	1,7207	2,0796	2,5176	2,8314
22	0,2564	0,6858	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,2563	0,6853	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,2562	0,6848	1,3178	1,7109	2,0639	2,4922	2,7969
25	0,2561	0,6844	1,3163	1,7081	2,0595	2,4851	2,7874
26	0,2560	0,6840	1,3150	1,7056	2,0555	2,4786	2,7787
27	0,2559	0,6837	1,3137	1,7033	2,0518	2,4727	2,7707
28	0,2558	0,6834	1,3125	1,7011	2,0484	2,4671	2,7633
29	0,2557	0,6830	1,3114	1,6991	2,0452	2,4620	2,7564
30	0,2556	0,6828	1,3104	1,6973	2,0423	2,4573	2,7500
50	0,2547	0,6794	1,2987	1,6759	2,0086	2,4033	2,6778
75	0,2542	0,6778	1,2929	1,6654	1,9921	2,3771	2,6430
100	0,2540	0,6770	1,2901	1,6602	1,9840	2,3642	2,6259
∞	0,2533	0,6745	1,2816	1,6449	1,9600	2,3263	2,5758

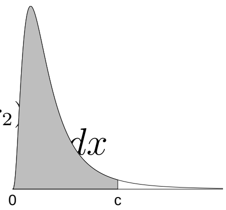
C.3 Alcuni quantili della distribuzione χ^2 con r gradi di libertà



$$F_r(c) = \int_0^c \frac{1}{\Gamma[r/2]2^{r/2}} x^{r/2-1} \exp\left[-\frac{x}{2}\right] dx$$

	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329	116,321

C.4 Alcuni quantili della distribuzione F con r_1 e r_2 gradi di libertà



$$F_{r_1,r_2}(c) = \int_0^c \frac{\Gamma((r_1+r_2)/2)}{\Gamma(r_1/2)\Gamma(r_2/2)} \left(\frac{r_1}{r_2}\right)^{r_1/2} x^{(r_1/2-1)} \left(1+\frac{r_1}{r_2}x\right)^{-(r_1+r_2)/2} dx$$

Per ogni coppia di r_1 (colonna) e r_2 (riga), la tavola fornisce il quantile f_α di ordine α corrispondente. I quantili inferiori della distribuzione F di Fisher–Snedecor si possono determinare tramite la relazione $f_{1-\alpha}(r_1, r_2) = 1/f_\alpha(r_2, r_1)$.

Tavole per l'ordine quantilico α= 0,95

	1	2	3	4	5	6	7	8	9
1	161,448	199,500	215,707	224,583	230,162	233,986	236,768	238,883	240,543
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124
50	4,034	3,183	2,790	2,557	2,400	2,286	2,199	2,130	2,073
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959
∞	3,841	2,996	2,605	2,372	2,214	2,099	2,010	1,938	1,880

Tavole per l'ordine quantilico α= 0,95

	10	15	20	30	40	50	60	120	∞
1	241,882	245,950	248,013	250,095	251,143	251,774	252,196	253,253	254,314
2	19,396	19,429	19,446	19,462	19,471	19,476	19,479	19,487	19,496
3	8,786	8,703	8,660	8,617	8,594	8,581	8,572	8,549	8,526
4	5,964	5,858	5,803	5,746	5,717	5,699	5,688	5,658	5,628
5	4,735	4,619	4,558	4,496	4,464	4,444	4,431	4,398	4,365
6	4,060	3,938	3,874	3,808	3,774	3,754	3,740	3,705	3,669
7	3,637	3,511	3,445	3,376	3,340	3,319	3,304	3,267	3,230
8	3,347	3,218	3,150	3,079	3,043	3,020	3,005	2,967	2,928
9	3,137	3,006	2,936	2,864	2,826	2,803	2,787	2,748	2,707
10	2,978	2,845	2,774	2,700	2,661	2,637	2,621	2,580	2,538
11	2,854	2,719	2,646	2,570	2,531	2,507	2,490	2,448	2,404
12	2,753	2,617	2,544	2,466	2,426	2,401	2,384	2,341	2,296
13	2,671	2,533	2,459	2,380	2,339	2,314	2,297	2,252	2,206
14	2,602	2,463	2,388	2,308	2,266	2,241	2,223	2,178	2,131
15	2,544	2,403	2,328	2,247	2,204	2,178	2,160	2,114	2,066
16	2,494	2,352	2,276	2,194	2,151	2,124	2,106	2,059	2,010
17	2,450	2,308	2,230	2,148	2,104	2,077	2,058	2,011	1,960
18	2,412	2,269	2,191	2,107	2,063	2,035	2,017	1,968	1,917
19	2,378	2,234	2,155	2,071	2,026	1,999	1,980	1,930	1,878
20	2,348	2,203	2,124	2,039	1,994	1,966	1,946	1,896	1,843
25	2,236	2,089	2,007	1,919	1,872	1,842	1,822	1,768	1,711
30	2,165	2,015	1,932	1,841	1,792	1,761	1,740	1,683	1,622
40	2,077	1,924	1,839	1,744	1,693	1,660	1,637	1,577	1,509
50	2,026	1,871	1,784	1,687	1,634	1,599	1,576	1,511	1,438
60	1,993	1,836	1,748	1,649	1,594	1,559	1,534	1,467	1,389
120	1,910	1,750	1,659	1,554	1,495	1,457	1,429	1,352	1,254
∞	1,831	1,666	1,571	1,459	1,394	1,350	1,318	1,221	1,000

Tavole per l'ordine quantilico α= 0,975

	1	2	3	4	5	6	7	8	9
1	647,789	799,500	864,163	899,583	921,848	937,111	948,217	956,656	963,285
2	38,506	39,000	39,165	39,248	39,298	39,331	39,355	39,373	39,387
3	17,443	16,044	15,439	15,101	14,885	14,735	14,624	14,540	14,473
4	12,218	10,649	9,979	9,605	9,364	9,197	9,074	8,980	8,905
5	10,007	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681
6	8,813	7,260	6,599	6,227	5,988	5,820	5,695	5,600	5,523
7	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823
8	7,571	6,059	5,416	5,053	4,817	4,652	4,529	4,433	4,357
9	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	4,026
10	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,779
11	6,724	5,256	4,630	4,275	4,044	3,881	3,759	3,664	3,588
12	6,554	5,096	4,474	4,121	3,891	3,728	3,607	3,512	3,436
13	6,414	4,965	4,347	3,996	3,767	3,604	3,483	3,388	3,312
14	6,298	4,857	4,242	3,892	3,663	3,501	3,380	3,285	3,209
15	6,200	4,765	4,153	3,804	3,576	3,415	3,293	3,199	3,123
16	6,115	4,687	4,077	3,729	3,502	3,341	3,219	3,125	3,049
17	6,042	4,619	4,011	3,665	3,438	3,277	3,156	3,061	2,985
18	5,978	4,560	3,954	3,608	3,382	3,221	3,100	3,005	2,929
19	5,922	4,508	3,903	3,559	3,333	3,172	3,051	2,956	2,880
20	5,871	4,461	3,859	3,515	3,289	3,128	3,007	2,913	2,837
25	5,686	4,291	3,694	3,353	3,129	2,969	2,848	2,753	2,677
30	5,568	4,182	3,589	3,250	3,026	2,867	2,746	2,651	2,575
40	5,424	4,051	3,463	3,126	2,904	2,744	2,624	2,529	2,452
50	5,340	3,975	3,390	3,054	2,833	2,674	2,553	2,458	2,381
60	5,286	3,925	3,343	3,008	2,786	2,627	2,507	2,412	2,334
120	5,152	3,805	3,227	2,894	2,674	2,515	2,395	2,299	2,222
∞	5,024	3,689	3,116	2,786	2,567	2,408	2,288	2,192	2,114

Tavole per l'ordine quantilico α= 0,975

	10	15	20	30	40	50	60	120	∞
1	968,627	984,867	993,103	1001,414	1005,598	1008,117	1009,800	1014,020	1018,258
2	39,398	39,431	39,448	39,465	39,473	39,478	39,481	39,490	39,498
3	14,419	14,253	14,167	14,081	14,037	14,010	13,992	13,947	13,902
4	8,844	8,657	8,560	8,461	8,411	8,381	8,360	8,309	8,257
5	6,619	6,428	6,329	6,227	6,175	6,144	6,123	6,069	6,015
6	5,461	5,269	5,168	5,065	5,012	4,980	4,959	4,904	4,849
7	4,761	4,568	4,467	4,362	4,309	4,276	4,254	4,199	4,142
8	4,295	4,101	3,999	3,894	3,840	3,807	3,784	3,728	3,670
9	3,964	3,769	3,667	3,560	3,505	3,472	3,449	3,392	3,333
10	3,717	3,522	3,419	3,311	3,255	3,221	3,198	3,140	3,080
11	3,526	3,330	3,226	3,118	3,061	3,027	3,004	2,944	2,883
12	3,374	3,177	3,073	2,963	2,906	2,871	2,848	2,787	2,725
13	3,250	3,053	2,948	2,837	2,780	2,744	2,720	2,659	2,595
14	3,147	2,949	2,844	2,732	2,674	2,638	2,614	2,552	2,487
15	3,060	2,862	2,756	2,644	2,585	2,549	2,524	2,461	2,395
16	2,986	2,788	2,681	2,568	2,509	2,472	2,447	2,383	2,316
17	2,922	2,723	2,616	2,502	2,442	2,405	2,380	2,315	2,247
18	2,866	2,667	2,559	2,445	2,384	2,347	2,321	2,256	2,187
19	2,817	2,617	2,509	2,394	2,333	2,295	2,270	2,203	2,133
20	2,774	2,573	2,464	2,349	2,287	2,249	2,223	2,156	2,085
25	2,613	2,411	2,300	2,182	2,118	2,079	2,052	1,981	1,906
30	2,511	2,307	2,195	2,074	2,009	1,968	1,940	1,866	1,787
40	2,388	2,182	2,068	1,943	1,875	1,832	1,803	1,724	1,637
50	2,317	2,109	1,993	1,866	1,796	1,752	1,721	1,639	1,545
60	2,270	2,061	1,944	1,815	1,744	1,699	1,667	1,581	1,482
120	2,157	1,945	1,825	1,690	1,614	1,565	1,530	1,433	1,310
∞	2,048	1,833	1,708	1,566	1,484	1,428	1,388	1,268	1,000

Tavole per l'ordine quantilico α= 0,99

	1	2	3	4	5	6	7	8	9
1	4052,181	4999,500	5403,352	5624,583	5763,650	5858,986	5928,356	5981,070	6022,473
2	98,503	99,000	99,166	99,249	99,299	99,333	99,356	99,374	99,388
3	34,116	30,817	29,457	28,710	28,237	27,911	27,672	27,489	27,345
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911
9	10,561	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780
17	8,400	6,112	5,185	4,669	4,336	4,102	3,927	3,791	3,682
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705	3,597
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,523
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457
25	7,770	5,568	4,675	4,177	3,855	3,627	3,457	3,324	3,217
30	7,562	5,390	4,510	4,018	3,699	3,473	3,304	3,173	3,067
40	7,314	5,179	4,313	3,828	3,514	3,291	3,124	2,993	2,888
50	7,171	5,057	4,199	3,720	3,408	3,186	3,020	2,890	2,785
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,559
∞	6,635	4,605	3,782	3,319	3,017	2,802	2,639	2,511	2,407

Tavole per l'ordine quantilico α= 0,99

	10	15	20	30	40	50	60	120	∞
1	6055,847	6157,285	6208,730	6260,649	6286,782	6302,517	6313,030	6339,391	6365,864
2	99,399	99,433	99,449	99,466	99,474	99,479	99,482	99,491	99,499
3	27,229	26,872	26,690	26,505	26,411	26,354	26,316	26,221	26,125
4	14,546	14,198	14,020	13,838	13,745	13,690	13,652	13,558	13,463
5	10,051	9,722	9,553	9,379	9,291	9,238	9,202	9,112	9,020
6	7,874	7,559	7,396	7,229	7,143	7,091	7,057	6,969	6,880
7	6,620	6,314	6,155	5,992	5,908	5,858	5,824	5,737	5,650
8	5,814	5,515	5,359	5,198	5,116	5,065	5,032	4,946	4,859
9	5,257	4,962	4,808	4,649	4,567	4,517	4,483	4,398	4,311
10	4,849	4,558	4,405	4,247	4,165	4,115	4,082	3,996	3,909
11	4,539	4,251	4,099	3,941	3,860	3,810	3,776	3,690	3,602
12	4,296	4,010	3,858	3,701	3,619	3,569	3,535	3,449	3,361
13	4,100	3,815	3,665	3,507	3,425	3,375	3,341	3,255	3,165
14	3,939	3,656	3,505	3,348	3,266	3,215	3,181	3,094	3,004
15	3,805	3,522	3,372	3,214	3,132	3,081	3,047	2,959	2,868
16	3,691	3,409	3,259	3,101	3,018	2,967	2,933	2,845	2,753
17	3,593	3,312	3,162	3,003	2,920	2,869	2,835	2,746	2,653
18	3,508	3,227	3,077	2,919	2,835	2,784	2,749	2,660	2,566
19	3,434	3,153	3,003	2,844	2,761	2,709	2,674	2,584	2,489
20	3,368	3,088	2,938	2,778	2,695	2,643	2,608	2,517	2,421
25	3,129	2,850	2,699	2,538	2,453	2,400	2,364	2,270	2,169
30	2,979	2,700	2,549	2,386	2,299	2,245	2,208	2,111	2,006
40	2,801	2,522	2,369	2,203	2,114	2,058	2,019	1,917	1,805
50	2,698	2,419	2,265	2,098	2,007	1,949	1,909	1,803	1,683
60	2,632	2,352	2,198	2,028	1,936	1,877	1,836	1,726	1,601
120	2,472	2,192	2,035	1,860	1,763	1,700	1,656	1,533	1,381
∞	2,321	2,039	1,878	1,696	1,592	1,523	1,473	1,325	1,000

Appendice D

Storia

Prima stesura del materiale in beamer e Sweave.

Aggiunte le lezioni: Spock e Challenger.

Appendice E

Versione software

Questo materiale è stato preparato con LaTeX, classe `beamer` e `book` e pacchetto `Sweave` in R . Sono stati generati in R ver. 2.9.2 nel sistema operativo (OS) linux-gnu e usando i seguenti pacchetti oltre a quelli standard:

Pacchetto	Versione
`e1071`	1.5-19
`labstatR`	1.0.5
`sn`	0.4-12
`xtable`	1.5-6

Appendice F

GNU Free Documentation License

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The **"Document"**, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as **"you"**. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A **"Modified Version"** of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A **"Secondary Section"** is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The **"Invariant Sections"** are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The **"Cover Texts"** are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A **"Transparent"** copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called **"Opaque"**.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The **"Title Page"** means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section **"Entitled XYZ"** means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as **"Acknowledgements"**, **"Dedications"**, **"Endorsements"**, or **"History"**.) To **"Preserve the Title"** of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties–for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright ©YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.