lucidi delle lezioni di inferenza statistica I (a.a. 2005/06)

guido masarotto

11 maggio 2006

guido masarotto
facoltà di scienze statistiche
università di padova
e-mail: guido.masarotto@unipd.it

Indice

Unità A

Introduzione al corso

Struttura del corso (e dell'esame)

Il corso è articolato in due parti che procedono in parallelo. Ha infatti due obbiettivi:

primo obbiettivo: presentare, soprattutto partendo da semplici problemi applicativi, le idee e alcune delle tecniche di base dell'inferenza statistica (6 ore di lezione alla settimana in aula "normale");

secondo obbiettivo: fornirvi una introduzione ad un primo ambiente per il calcolo statistico prendendo come pretesto le tecniche viste durante il corso di Descrittiva e quelle che via via vi presenterò durante questo corso (2 ore di esercitazione in laboratorio informatico alla settimana - divisi in due gruppi).

L'ambiente scelto per il laboratorio è **R** scaricabile gratuitamente da http://www.r-project.org e disponibile nel CD della Facoltà (disponibile sempre gratuitamente presso l'UID).

Anche l'esame (e il voto) è diviso in due parti:

prova pratica: una prova in laboratorio informatico (un ora, valutazione da 0 a 8, voto minimo per la sufficienza 4);

prova scritta: uno scritto in cui dovete risolvere alcuni esercizi in aula "normale" (un ora e mezza, valutazione da 0 a 24, voto minimo per la sufficienza 14).

Il voto complessivo è dato dalla somma dei voti delle due prove (e se la somma vale più di 30 c'è la lode).

“Statistica Descrittiva” vs “Inferenza Statistica”

Ricordiamoci, dal corso di “Descrittiva”, che:

- 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, animali, batteri, immagini raccolte da un satellite,...) vengono convenzionalmente indicate come **unità statistiche**;
- l’analisi statistica vuole, nella 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!).

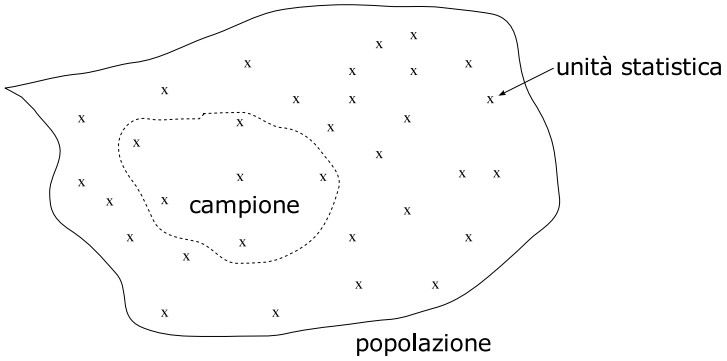
inferenza statistica: se, viceversa, i dati disponibili sono stati rilevati solamente su una parte delle unità statistiche (il *campione* da cui *indagini campionarie*). Vogliamo utilizzare le informazioni del campione per fare delle affermazioni sulle caratteristiche di tutta la popolazione.

Tra *Statistica Descrittiva* ed *Inferenza Statistica* esiste una ovvia “fratellanza” ed, in realtà, nelle applicazioni, non sono facilmente separabili anche perchè 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 sulle sei unità statistiche appartenenti al campione vogliamo "produrre" affermazioni su tutta la popolazione.



Perchè indagini di tipo campionario sono frequenti?

- **tempo e/o costo.**
  Esempi
  — 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.
  — 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, perchè no, se il piano è di vendere la tortina anche all'estero, in tutta Europa, in tutti i paesi occidentali,. . .

- **la popolazione di interesse può essere infinita e virtuale**
  Esempio: 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 perchè abbiamo continuato a somministrarlo?.

- **la rilevazione "distrugge" le unità statistiche** e quindi, dopo una rilevazione esaustiva, la popolazione di partenza non interessa più perchè non esiste più!
  Esempio: 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 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!
- **precisione dei risultati**: può sembrare strano ma delle volte è stato dimostrato che rilevazioni campionarie (incomplete) portano a risultati più precisi di rilevazioni esaustive.
  E' 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 conoscerne la relazione

— supponiamo che la polazione di riferimento siate voi (gli studenti presenti alla prima lezione del corso di inferenza statista I 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 (ad occhio), 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"). E' 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.

— E' 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 commetteremmo degli errori.
Per rendere utili le nostre affermazioni dovremmo allora occuparci anche di capire di quanto sono sbagliate.
Esempio. Supponiamo di sperimentare un nuovo farmaco su 20 pazienti e che solo 1 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. E' una popolazione virtuale e teoricamente infinita. E' 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 chiederci di quanto la nostra stima (5%) potrebbe essere differente dalla vera probabilità.
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, non siamo in grado di dire, in pratica, niente della incognita probabilità di manifestare tossicità.
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, interpreteremmo i risultati sperimentali (ovvero i dati disponibili) come uno dei tanti risultati che un meccanismo probabilistico (un esperimento casuale) poteva fornirci.
Questa costruzione cercherò di illustrarvela nel seguito del corso (già a partire dalla prossima lezione). Inutile entrare quindi ora nei dettagli.
Una conseguenza importante sarà che potremmo utilizzare in maniera naturale il calcolo delle probabilità “per misurare gli errori”.
Una seconda conseguenza importante, e il vero motivo di questo lucido, è il ricordarvi che i contenuti del corso di probabilità sono, almeno in parte, propedeutici a quelli di questo corso.

Unità B

Controllo di qualità in un impianto che produce lastre di metallo

Un primo esempio di inferenza statistica.
• Media e varianza campionaria.
• Inferenza sulla media (intervalli di confidenza e test) nel caso di un campione tratto da una v.c. normale di varianza nota.
• Inferenza sulla media quando la numerosità campionaria è grande.

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 $\pm 0{,}5$mm, ovvero, una lastra è considerata soddisfacente, per quello che riguarda lo spessore, se

$$13{,}5 \leq \text{spessore} \leq 14{,}5. \tag{B.1}$$

• La produzione è organizzata in turni di 6 ore.
• All'inizio di ogni turno vengono estratte a caso 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 (B.1) è sufficientemente piccolo.
• In particolare, se si decide per il si 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:

$$14{,}33 \quad 14{,}19 \quad 14{,}39 \quad 14{,}43 \quad 14{,}17.$$

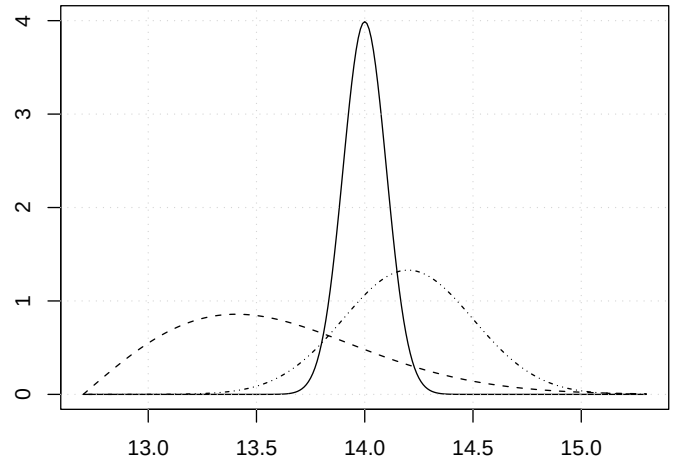
Nel seguito consideremo il problema di utilizzare questi dati per decidere se bloccare o non bloccare temporaneamente la produzione.

Una possibile formulazione del problema

• Nessun processo produttivo è in grado di produrre lastre *esattamente* dello stesso spessore.
• Possiamo però pensare che, durante un certo turno, il processo produttivo sia in un particolare "stato" dato dalle caratteristiche tecnologiche dell'impianto, dalla qualità delle materie prime,... e che le lastre prodotte durante il turno siano il risultato di un esperimento casuale le cui caratteristiche dipendono dallo "stato".
• Questo formalizza l'idea che, all'inizio di un turno, solo *Mago Merlino* sarebbe in grado di indovinare esattamente lo spessore delle lastre che saranno prodotte ma che, però, possiamo pensare di descrivere gli spessori delle lastre che saranno prodotte utilizzando il calcolo delle probabilità.
• In particolare, possiamo guardare agli spessori che, durante un certo turno, il processo produce come ad una variabile casuale continua con funzione di densità $f(\cdot)$.
• Il problema diventa allora quello di utilizzare[1] 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 (B.1)).
• Se questo accade, e quindi se il processo sta, *almeno potenzialmente*, producendo "troppe" lastre difettose decideremo di sospendere la produzione.

[1] si veda la pagina seguente, per alcuni esempi

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 avendo a disposizione solo le nostre 5 osservazioni sembra essere un'operazione eccessivamente avventurosa.
• Fortunamente esistono delle conoscenze aggiuntive sul processo.
• Infatti, precedentemente, le caratteristiche del processo sono state studiate raccogliendo alcune migliaia di misurazioni per alcune decine di turni.
• Le principali conclusioni delle analisi condotte su questi dati sono che, indicate con $Y_1, Y_2, \ldots$ le variabili casuali che descrivono lo spessore della prima lastra prodotto in un turno, della seconda e così via,

(a) non esiste nessun tipo di dipendenza tra le Y_i;

(b) tutte le Y_i hanno la stessa distribuzione di probabilità;

(c) questa distribuzione comune è ben approssimata da una normale di media μ e varianza $0{,}01$ dove μ è un *parametro* ignoto che può essere diverso da turno a turno.

Un modello è buono perchè è utile non perchè è vero

Nel seguito adotteremo come "esattamente" vere le (a)-(c) del lucido 19.
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 della media

Le informazioni aggiuntive ci portano a considerare le 5 misure dello spessore come 5 determinazioni indipendenti "estratte" da una stessa variabile casuale Gaussiana di media μ ignota 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 casuale normale....
La funzione di densità dello spessore è quindi "quasi" nota. Sappiamo infatti che è

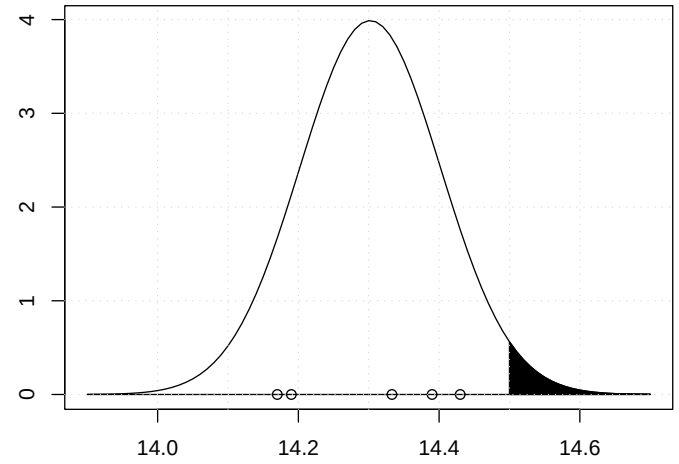
$$f(x) = f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

con $\sigma^2 = 0{,}01$ e per qualche ignoto "numero" μ.
Per conoscere completamente la distribuzione dei dati ci manca quindi solo la media μ. Possiamo però utilizzare le osservazioni disponibili (i "nostri" cinque spessori) per stimarla. Al proposito sembra "ragionevole" utilizzare la media delle osservazioni come **"stima"** della vera media μ, ovvero porre

$$\text{stima della media} = \overline{y} = \frac{14{,}33 + \cdots + 14{,}17}{5} = 14{,}302.$$

Poichè $\overline{y}$ è la media delle osservazioni nel campione viene usualmente chiamata *la media campionaria*.

Densità stimata



Il grafico mostra la densità di una normale di media 14,302 e varianza 0,01.
I "cerchietti" sull'asse delle x indicano le osservazioni. Si osservi come il "modello costruito" sia quantomeno "possibile": la distribuzione potrebbe realmente "generare" le 5 osservazioni.
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.

Stima della “difettosità”

Due eventi particolarmente importanti nel presente contesto sono

$$\begin{aligned} A &= \{\text{lastra troppo sottile}\} = \{Y < 13{,}5\} \\ B &= \{\text{lastra troppo spessa}\} = \{Y > 14{,}5\} \end{aligned}$$

dove Y indica la variabile casuale che descrive lo spessore.
Ovviamente sia $P(A)$ che $P(B)$ sono funzione di μ. In particolare risulta[2]

$$P(A) = P(N(\mu, 0{,}01) < 13{,}5) = P(N(\mu, 0{,}01) \leq 13{,}5) = \Phi\left(\frac{13{,}5 - \mu}{0{,}1}\right)$$

e

$$\begin{aligned} P(B) &= P(N(\mu, 0{,}01) > 14{,}5) = 1 - P(N(\mu, 0{,}01) \leq 14{,}5) = \\ &= 1 - \Phi\left(\frac{14{,}5 - \mu}{0{,}1}\right) \end{aligned}$$

dove $\Phi(\cdot)$ è la funzione di ripartizione di una variabile casuale normale standard[3].
Possiamo ottenere delle stime di queste due probabilità sostituendo a μ, che è ignoto, la sua stima $\overline{y}$.

$$\hat{P}(A) = \Phi\left(\frac{13{,}5 - 14{,}302}{0{,}1}\right) = \Phi(-8{,}02) \approx 0$$

e

$$\hat{P}(B) = 1 - \Phi\left(\frac{14{,}5 - 14{,}302}{0{,}1}\right) = 1 - \Phi(1{,}98) \approx 0{,}024$$

ovvero, sulla base dei dati (e delle assunzioni fatte), stimiamo in 2,4% la probabilità di produrre una lastra troppo “alta” mentre valutiamo praticamente nulla la probabilità di produrre una lastra troppo sottile.

[2][Probalità 7]. Usiamo anche il fatto che se X è una variabile casuale continua allora, $P(x < a) = P(x \leq a)$.

[3][Probalità 5]

Stima di qui, stima di là,. . . , ma se c’è una stima c’è un errore

• Abbiamo incontrato **due** medie: una “vera” μ e una **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 la “vera” media, 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 (e interessante da un punto di vista pratico) chiederci “quanto è buona?” ovvero “quanto è grande l’errore che commettiamo?”

Esercizio. Si osservi che abbiamo sempre scritto *vera* tra virgolette. Lo studente ripensi a quanto detto nel lucido 20 e spieghi perchè.

La distribuzione della media campionaria

• La media campionaria, $\overline{y}$, può essere vista come una determinazione di una variabile casuale e quindi ha una sua distribuzione di probabilità.
• Infatti se i dati da cui è calcolata, $y_1, \ldots, y_n$, sono il risultato di un esperimento casuale anche

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

ovviamente lo è[4].
• La distribuzione di probabilità di $\overline{y}$, che viene chiamata la *distribuzione campionaria* dello stimatore, ci racconta "dove ci aspettiamo di trovare" $\overline{y}$. Proviamo quindi a studiarla nel caso che stiamo considerando.
• *Non distorsione della media campionaria.* E' possibile far vedere[5] che, qualsiasi sia l'ignoto valore di μ,

$$\mathrm{E}\{\overline{y}\} = \mu$$

ovvero che

$$\mathrm{E}\{\text{stima di } \mu\} = \text{"vero" valore di } \mu.$$

• Si osservi che avremmo potuto anche scrivere

$$\mathrm{E}\{\overline{y} - \mu\} = 0 \text{ ovvero } \mathrm{E}\{\text{errori di stima}\} = 0.$$

• In generale, se la media di uno stimatore è uguale al valore che si vuole stimare si parla di stimatore corretto o non distorto. Le relazioni appena viste sono quindi equivalenti alla frase

> *"la media campionaria è uno stimatore non distorto della vera media"*

[4] se *ripetiamo* l'esperimento, nel caso delle lastre, ad esempio, estraendone altre 5, troveremo dei dati "differenti" e quindi una media campionaria "differente".
[5] [Probalità 40].

• La non distorsione ci garantisce che, qualsiasi sia μ, le determinazioni di $\overline{y}$, ovvero le stime della media, sono posizionate "intorno" al vero valore della media.
• Questa è ovviamente una proprietà fondamentale per uno stimatore. Si osservi comunque che perchè questo accada può, in generale, bastarci anche una non distorsione approssimata ovvero che

$$\mathrm{E}\{\overline{y}\} \approx \mu.$$

• *Varianza della media campionaria.* E' inoltre possibile far vedere[6] che

$$\mathrm{var}\{\overline{y}\} = \frac{\sigma^2}{n} \tag{B.2}$$

dove σ^2 è la varianza dei dati originali (nel nostro caso degli "spessori" e, quindi, $\sigma^2 = 0{,}01$);
La (B.2), che può anche essere scritta come

$$\mathrm{var}\{\text{errori di stima}\} = \mathrm{E}\left\{(\overline{y} - \mu)^2\right\} = \frac{\sigma^2}{n}$$

rende precisa l'idea che la media di n osservazioni è uno stimatore della vera media "più preciso" di ciascuna delle singole osservazioni. Potremmo infatti scriverla come

$$\mathrm{var}\{\text{media campionaria}\} = \frac{\mathrm{var}\{\text{singola osservazione}\}}{n}.$$

[6] [Probalità 40].

• *Consistenza della media campionaria.* La legge forte dei grandi numeri[7] ci assicura inoltre che, al tendere della numerosità campionaria ad infinito, $\overline{y}$ converge con probabilità uno verso la vera media μ[8].

• In generale, se uno stimatore converge [in probabilità, quasi certamente] verso il vero parametro si parla di stimatore *consistente* [in probabilità, quasi certamente] o in senso [debole,forte].

Equivalentemente quindi, la proprietà appena enunciata poteva essere raccontata dicendo

> *"la media campionaria è uno stimatore consistente (in senso forte) della vera media"*

• La consistenza è una proprietà di base di uno stimatore. Se la numerosità campionaria aumenta fino ad infinito la "quantità di informazione" contenuta nel campione diventa infinita. Quindi la stima deve diventare sempre più precisa e, almeno ad ∞, l'errore deve essere nullo.

• E' importante osservare che le tre proprietà di $\overline{y}$ appena viste (non distorsione, consistenza, formula per la varianza) *non dipendono* dalla normalità dei dati ma solo dal fatto che la media campionaria è stata calcolato a partire da n osservazioni indipendenti e identicamente distribuite come una variabile casuale di media μ e varianza σ^2.

[7] [Probalità 50]

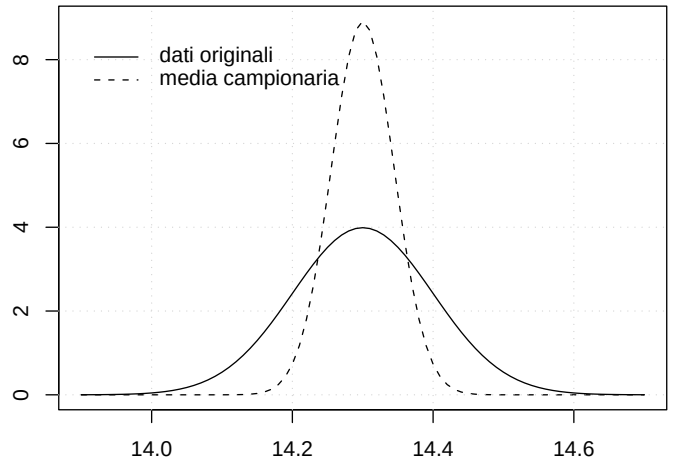
[8] ovviamente, questa proprietà non è particolarmente interessante nel caso degli "spessori" visto che abbiamo solo 5 osservazioni ovvero siamo molto lontani da infinito. Si tratta però di una proprietà in generale interessante della media campionaria.

• *Distribuzione della media campionaria nel caso di un campione tratto da una popolazione normale.* Nel caso in cui le osservazioni siano normali è però possibile mostrare anche che[9]

$$\overline{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

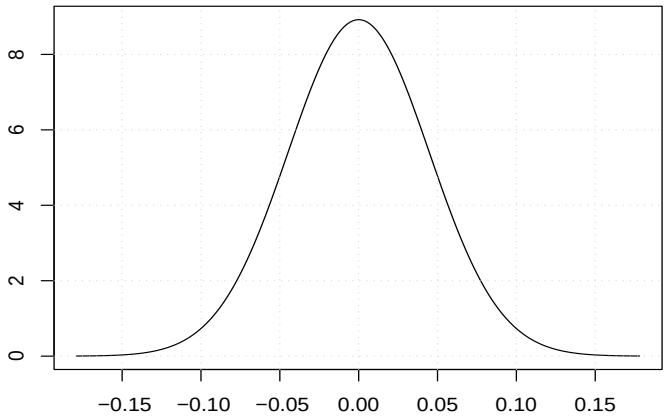
In questo caso conosciamo quindi "tutta" la distribuzione della media campionaria.

Il grafico mostra le funzioni di densità della media campionaria e delle osservazioni originali nel caso in cui $\mu = 14{,}3$ e $\sigma = 0{,}1$.



[9] [Probalità 41]

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)

$$\overline{y} - \mu \sim N(0, \sigma^2/n).$$

Si noti che nel caso in esame, poichè σ^2 è noto, la distribuzione dell'errore di stima risulta anche essa nota (è una normale di media 0 e varianza $0{,}01/5 = 0{,}002$).

Un intervallo di confidenza

• Poichè la distribuzione dell'errore di stima è completamente nota possiamo "costruire" delle affermazioni del tipo:

> *"la probabilità che l'errore di stima sia in valore assoluto minore di* 0,1 *è uguale a* 0,987*"*

Infatti,

$$\begin{aligned} P(|\overline{y} - \mu| < 0{,}1) &= P(|N(0, 0{,}002)| < 0{,}1) = \\ &= \Phi\left(\frac{0{,}1}{\sqrt{0{,}002}}\right) - \Phi\left(-\frac{0{,}1}{\sqrt{0{,}002}}\right) = \\ &= \Phi(2{,}236) - \Phi(-2{,}236) = 0{,}987. \end{aligned}$$

• L'affermazione precedente ci permette anche di dire che

> *"la probabilità che l'intervallo* [14,202 ; 14,402] *includa la vera media* μ *è* 0,987*"*

Infatti

$$\begin{aligned} P(|\overline{y} - \mu| < 0{,}1) &= P(-0{,}1 < \mu - \overline{y} < 0{,}1) = \\ &= P(\overline{y} - 0{,}1 < \mu < \overline{y} + 0{,}1) = \\ &= P(14{,}302 - 0{,}1 < \mu < 14{,}302 + 0{,}1) = \\ &= P(14{,}202 < \mu < 14{,}402) \end{aligned}$$

• 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** $1 - \alpha$.

• 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 di media nulla e varianza 0,002.
2. Con probabilità molto alta, per la precisione 0,987, il "vero" valore della media 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 0,5 , 0,9 , 0,95 e 0,99).
Nel caso che stiamo considerando, i passi da seguire sono i seguenti.
• Ovviamente, per prima cosa, dobbiamo fissare un valore per $1-\alpha$.
• Poi determiniamo o utilizzando un programma o le tavole della normale standard, il quantile $1-\alpha/2$ di una normale standard, ovvero un punto, indichiamolo con $z_{1-\alpha/2}$, tale che

$$P(N(0,1) \leq z_{1-\alpha/2}) = 1-\alpha/2.$$

Per la simmetria della densità di una normale intorno alla sua media allora

$$P(N(0,1) \leq -z_{1-\alpha/2}) = \alpha/2.$$

Quindi[10]

$$P(-z_{1-\alpha/2} \leq N(0,1) \leq z_{1-\alpha/2}) = 1-\alpha.$$

• Ricordando che[11]

$$\frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

possiamo allora scrivere

$$P\left(-z_{1-\alpha/2} \leq \frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

da cui, con semplici passaggi, otteniamo

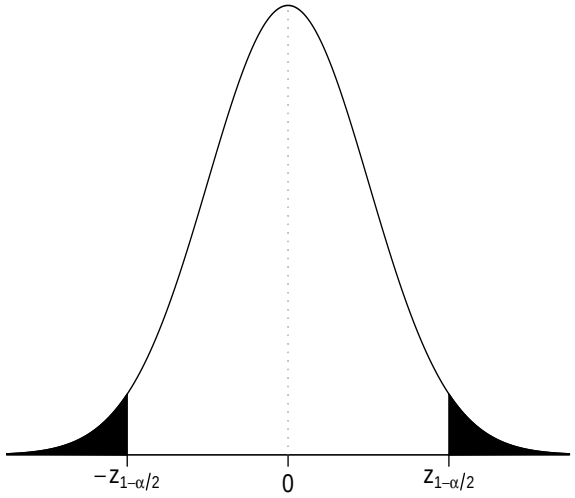
$$P\left(\overline{y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \overline{y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

• Quindi

$$\left[\overline{y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}; \overline{y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right]$$

è un intervallo di confidenza di livello $1-\alpha$ per μ.

[10] si veda il grafico a pagina 33.
[11] [Probalità 3] e [Probalità 41]



Ambedue le aree “annerite” sono uguali ad $\alpha/2$. Quindi l’area “non annerita” è uguale a $1-\alpha$.

Intervalli di confidenza per la difettosità

Ricordiamoci che abbiamo ottenuto le formule

$$\pi_B(\mu) = P(\text{“lastra troppo bassa”}) = \Phi\left(\frac{13{,}5-\mu}{0{,}01}\right)$$
$$\pi_A(\mu) = P(\text{“lastra troppo alta”}) = 1-\Phi\left(\frac{14{,}5-\mu}{0{,}01}\right)$$

dove con l’introduzione della nuova notazione $\pi_B(\cdot)$ e $\pi_A(\cdot)$ enfatizziamo il fatto che la probabilità di produrre una lastra difettosa dipende dalla media.
E’ facile verificare che $\pi_B(\mu)$ e $\pi_A(\mu)$ sono monotone in μ, la prima decrescente e la seconda crescente[12] Quindi, gli eventi

$$\left\{\overline{y} : \overline{y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}} \le \mu \le \overline{y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right\},$$
$$\left\{\overline{y} : \pi_B\left(\overline{y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right) \le \pi_B(\mu) \le \pi_B\left(\overline{y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)\right\}$$

e

$$\left\{\overline{y} : \pi_A\left(\overline{y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right) \le \pi_A(\mu) \le \pi_A\left(\overline{y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)\right\}$$

coincidono e perciò hanno la medesima probabilità.
Ricordando che il primo è vero con probabilità $1-\alpha$, questo ci permette di dire che

$$\left[\pi_B\left(\overline{y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right); \pi_B\left(\overline{y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)\right]$$

e

$$\left[\pi_A\left(\overline{y} - \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right); \pi_A\left(\overline{y} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}\right)\right]$$

sono intervalli di confidenza di dimensione $1-\alpha$, rispettivamente, per $\pi_B(\mu)$ e $\pi_B(\mu)$.

[12] ci si ricordi che $\Phi(y)$ è crescente in y.

Una prima conclusione

Supponiamo di volere un intervallo di confidenza di livello 0,95 per μ e $\pi(\mu)$. Allora,

$$\alpha = 0{,}05,\ \frac{\alpha}{2} = 0{,}025,\ 1 - \frac{\alpha}{2} = 0{,}975.$$

Utilizzando una funzione o consultando una tavola dei percentili della normale standard troviamo $z_{0{,}975} = 1{,}96$. Quindi l'intervallo di confidenza per μ è

$$14{,}302 \pm \frac{1{,}96 \times 0{,}1}{\sqrt{5}} = [14{,}21; 14{,}39].$$

L'intervallo di confidenza per $\pi_B(\mu)$ è quindi

$$[\pi_B(14{,}39)\ ;\ \pi_B(14{,}21)].$$

Ora, $\pi_B(14{,}39) < \pi_B(14{,}21) < 10^{-20}$. Quindi, per quanto non conosciamo esattamente la probabilità di produrre una lastra troppo "bassa", possiamo dire è, visti i dati, che è praticamente irrilevante. Viceversa, l'intervallo di confidenza per $\pi_A(\mu)$ è

$$[\pi_A(14{,}21)\ ;\ \pi_A(14{,}39)] = [0{,}002\ ;\ 0{,}135].$$

Quindi, sulla base dei dati sul processo produttivo (e delle ipotesi fatte), possiamo dire che la probabilità di produrre una lastra troppo "spessa" sta, con grande probabilità (esattamente 95%), tra il 2 per mille e il 13%.
La conclusione, se guardiamo all'estremo superiore, è che potrebbe essere "prudente" bloccare la produzione: una possibile difettosità superiore al 10% sarebbe disastrosa. Si tenga tra l'altro conto che

$$\pi_A(14) = \pi_B(14) \approx 2/10^6,$$

ovvero, che l'impianto, quando ben "tarato", può produrre un numero di lastre difettose veramente piccolo.

Un approccio diverso

• Fino ad adesso 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:
(i) ad ogni manutenzione (ordinaria o straordinaria) l'impianto viene "tarato" in maniera tale che la media degli spessori prodotti risulti 14mm;
(ii) quindi un valore di μ diverso, anche di poco, da 14mm indica una qualche "sregolazione in corso";
(iii) 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 per 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:

$$H_0 : \mu = 14\text{mm} \text{ verso } H_1 : \mu \neq 14\text{mm}.$$

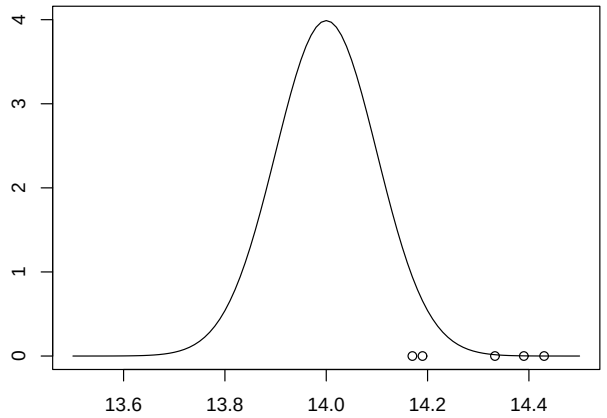
L'interpretazione delle due ipotesi è (ovviamente):

H_0 : l'impianto produce al meglio,
H_1 : l'impianto ha iniziato a "sregolarsi".

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**.
• 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, o meglio, come si usa dire, per decidere quale ipotesi **accettare** o **rifiutare**), viene chiamato **test statistico**.

Analisi grafica



La figura mostra la densità di 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. Dall'altra parte H_1 "prevede" alcune distribuzioni (ad es. si veda il grafico a pagina 22) 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 $\sigma = 0.1$) ciò non cambia l'interpretazione della "regola".

• Per rendere operativa la "regola" dobbiamo decidere quale valore assegnare alla soglia h.

Se H_0 è vera...

...vorremmo, ovviamente, rifiutare H_1. In altre parole non ci dispiacerebbe che

$$P(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera}) = 1 \tag{B.3}$$

ovvero, che

$$P\left(-h \leq \frac{\sqrt{n}(\overline{y} - 14)}{\sigma} \leq h \text{ quando } \mu = 14\right) = 1. \tag{B.4}$$

Ora,

$$\text{se } H_0 : \mu = 14 \text{ è vera allora } \frac{\sqrt{n}(\overline{y} - 14)}{\sigma} \sim N(0, 1)$$

e quindi la (B.4) è equivalente a

$$P(-h \leq N(0, 1) \leq h) = 1 \tag{B.5}$$

La (B.5) mostra che l'unico valore di h che garantisce la (B.3) 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 rifiuteremmo mai H_0. In altre parole, se insistiamo sulla (B.3) finiamo con una "regola" per cui

$$P(\text{accettare } H_0 \text{ quando } H_0 \text{ è falsa}) = 1.$$

Test con livello di significatività prefissato

• Chiedere che la (B.3) sia *esattamente* vera ci porta a determinare un valore di h inaccettabile.
• Sarebbe però inacettabile anche una situazione in cui, ad esempio,

$$P(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera}) = 0{,}1$$

ovvero, una situazione in cui la (B.3) 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 (B.3) è approssimativamente (ma non esattamente) rispettata, ovvero, in cui

$$P(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera}) = 1 - \alpha \qquad \text{(B.6)}$$

per un valore "piccolo" di α.
• La (B.6) può essere riscritta nella forma

$$P(-h \leq N(0,1) \leq h) = 1 - \alpha \qquad \text{(B.7)}$$

ed è facile verificare (lo studente si aiuti con il grafico a pagina 33) che la soluzione in h della (B.7) è

$$h = z_{1-\alpha/2}$$

dove con z_p abbiamo indicato il quantile p-simo di una normale di media zero e varianza uno, ovvero il numero per cui $\Phi(z_p) = p$.
• La probabilità α che compare nella (B.6) viene chiamata il *livello di significatività* del test.
• Per comunicare [l'accettazione,il rifiuto] di H_0 si utilizzano spesso frasi del tipo *"i risultati sono [non significativi,significativi] al* $100\alpha\%$*"*, o semplicemente, quando α è implicito, *"i risultati sono [non significativi,significativi]"* [13].

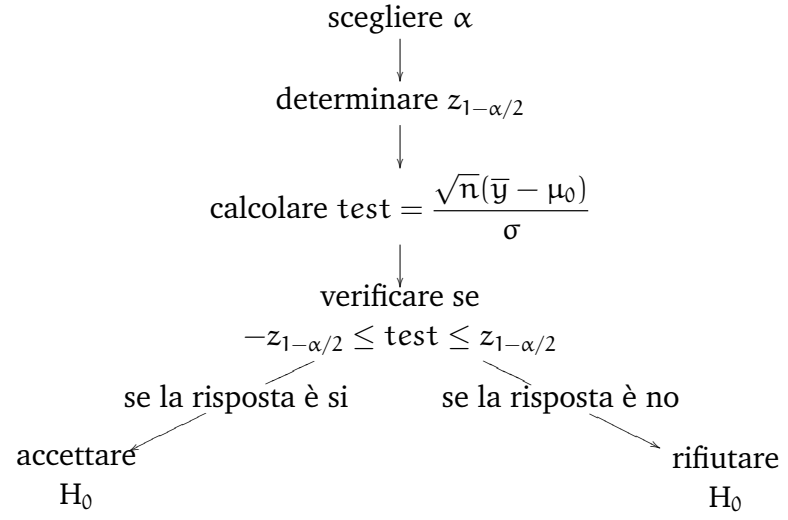
[13] la significatività è quindi da intendersi "contro" H_0

Sintesi della procedura delineata...

In definitiva, per verificare un sistema d'ipotesi del tipo

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

siamo arrivati alla seguente procedura:



... e applicazione al caso in esame

$$\alpha = 0{,}01 \text{ (ad es.)}$$
$$\downarrow$$
$$z_{1-\alpha/2} = z_{0{,}995} = 2{,}58$$
$$\downarrow$$
$$\text{test} = \frac{\sqrt{5}(14{,}302 - 14)}{0.1} = 6{,}75$$
$$\downarrow$$
$$-2{,}58 \leq 6{,}75 \leq 2{,}58 \text{ ?}$$
$$\downarrow$$
$$\text{no}$$
$$\downarrow$$
$$\text{rifiutiamo } H_0$$
$$\downarrow$$

Il risultato è significativo al 1%.

Inferenza sulla media quando la numerosità campionaria è grande

⋆ Gli intervalli di confidenza e il test sulla media che abbiamo costruito sono approssimativamente validi e quindi possono essere utilizzati anche se i dati disponibili, $y_1, \ldots, y_n$,

(a) sono n determinazioni indipendenti ed identicamente distribuite di una variabile casuale *non necessariamente normale* di media μ, incognita, e varianza σ^2 nota purchè

(b) la numerosità campionaria, n, sia "sufficentemente" grande.

⋆ Infatti, il risultato alla base degli intervalli di confidenza e del test sulla media che abbiamo costruito è che, se i dati, $y_1, \ldots, y_n$, sono determinazioni i.i.d. di una $N(\mu, \sigma)$ allora

$$\frac{\overline{y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

⋆ Ma, se sono vere le (a)-(b), per il teorema del *limite centrale*[14], se n tende ad infinito allora

$$\frac{\overline{y} - \mu}{\sigma/\sqrt{n}} \text{ converge in distribuzione verso una } N(0, 1),$$

ovvero, per qualsivoglia x

$$\lim_{n\to\infty} P\left(\frac{\overline{y} - \mu}{\sigma/\sqrt{n}} \leq x\right) = P(N(0, 1) \leq x) = \Phi(x).$$

[14] [Probalità 51].

⋆ Quindi, se n è sufficentemente grande,

$$P\left(\frac{\overline{y}-\mu}{\sigma/\sqrt{n}} \leq x\right) \approx P(N(0,1) \leq x) = \Phi(x).$$

e questo è sufficente, si ripercorra indietro quanto fatto fino ad ora, per mostrare che

- l'intervallo
$$\overline{y} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$
contiene l'incognito valore della media con una probabilità *approssimativamente* uguale a $1-\alpha$
- il livello di significatività del test descritto nel lucido 42 è *approssimativamente* pari ad α.

⋆ Se la varianza, σ^2, non è nota ma ne è disponibile una stima consistente[15], indichiamola con $\hat{\sigma}^2$, è possibile dimostrare che anche

$$\frac{\overline{y}-\mu}{\hat{\sigma}/\sqrt{n}} \text{ converge in distribuzione verso una } N(0,1).$$

Per questo motivo gli intervalli di confidenza e il test sulla media visti in questa unità rimangono *approssimativamente validi anche sostituendo alla vera varianza una sua stima consistente purchè la numerosità campionaria sia sufficentemente grande*.
Nota. Vedremo nelle prossime unità come trattare campioni "piccoli" provenienti da una popolazione normale quando la varianza non è nota.

[15] una possibilità è discussa nel lucido 47.

⋆ Una domanda spontanea è

quanto deve essere grande n perchè l'approssimazione sia decorosa?

⋆ Purtroppo, la domanda non ha una risposta precisa: la velocità di convergenza della distribuzione della media campionaria ad una normale dipende dalla distribuzione dei dati.
⋆ Una regola a spanne è

- n deve essere maggiore od uguale a 30 se la distribuzione dei dati è (almeno approssimativamente) simmetrica;
- n deve essere maggiore od uguale a 50 se la distribuzione dei dati è non simmetrica.

In ambedue i casi è inoltre importante verificare che non ci siano evidenti osservazioni anomale tra i dati.

La varianza campionaria

⋆ Lo stimatore usuale della varianza considerato in problemi di inferenza è

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

dove, al solito, con
- $y_1, \ldots, y_n$ abbiamo indicato i dati disponibili e
- con $\overline{y}$ la loro media.

⋆ s^2 è chiamato la *varianza campionaria*.

⋆ Si osservi che, in s^2, dividiamo la somma dei quadrati degli scarti dalla media per "$n-1$" non per n come è usuale fare in "Descrittiva".

⋆ Infatti è possibile far vedere che se i dati $y_1, \ldots, y_n$ sono determinazioni indipendenti e identicamente distribuiti di una variabile casuale[16] di varianza σ^2 allora

$$E\left\{s^2\right\} = \sigma^2$$

ovvero

> *"la varianza campionaria è uno stimatore non distorto della varianza della popolazione"*

⋆ Viceversa, visto che

$$E\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right\} = E\left\{\frac{n}{n-1}s^2\right\} = \frac{n}{n-1}\sigma^2 = \sigma^2 + \frac{1}{n-1}\sigma^2,$$

"dividendo per n" otteremmo uno stimatore distorto.

⋆ E' possibile anche dimostrare[17] che

> *"la varianza campionaria è uno stimatore consistente (in senso forte) della varianza della popolazione"*

[16] non necessariamente normale

[17] [Probalità 53]

Verifica d'ipotesi: struttura di un test statistico

Quanto abbiamo fatto per costruire il test sulla media illustra fedelmente la struttura di un test statistico. E' quindi conveniente "ricapitolarlo":

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[18] **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 "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[19]

$$T \sim N(\sqrt{n}(\mu - \mu_0)/\sigma, 1)$$

ovvero,
- se è vera H_0, $T \sim N(0,1)$ ma
- se è vera H_1, $T \sim N(\eta_n, 1)$ con $\eta_n \neq 0$.

[18] ma va!

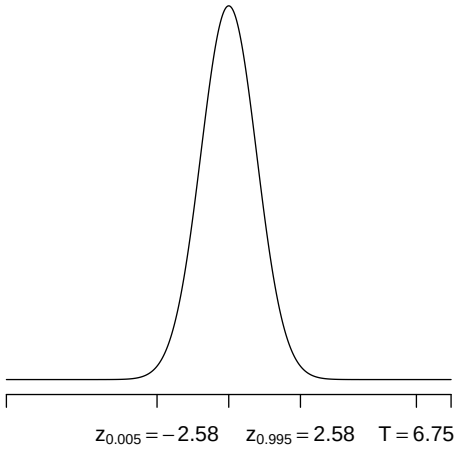
[19] lo studente lo dimostri utilizzando i risultati in appendice

3. A questo punto per decidere se H_0 doveva essere accettata o rifiutata abbiamo "confrontato" il valore osservato della statistica, ovvero il valore di T calcolato dai dati, con la distribuzione sotto H_0[20].
Poichè il valore osservato della statistica era "troppo estremo" (ovvero, troppo "poco probabile" per la distribuzione di T sotto H_0) abbiamo deciso di rifiutare H_0.
In particolare, si osservi che, desiderando una regola precisa, nella procedura operativa descritta dall'albero a pagina 42 abbiamo convenuto che "troppo estremo" significa $|T| > z_{1-\alpha/2}$ per qualche pre-scelto (e non troppo grande) valore di α.

Nota. Si osservi come in questo caso (ma in realtà accade sempre per i test che si "rispettano") per ogni prefissato $\mu \neq \mu_0$ la distribuzione della statistica test "scappi" verso o $+\infty$ o $-\infty$ all'aumentare di n.
Ovvero, come all'aumentare del numero di osservazioni (= della quantità di informazioni nel campione) le distribuzione di T sotto le due ipotesi si "separino" sempre più.

[20] si veda lucido seguente

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 con quanto fatto a pagina 38. Solamente qui usiamo la statistica test e non le osservazioni.

Verifica d'ipotesi: tipi di errore e funzione di potenza

• In un problema di verifica d'ipotesi esistono due possibili modi con cui sbagliare.
Infatti può capitare di:

1. rifiutare H_0 quando H_0 è vera; questo è usualmente chiamato[21] un *errore di primo tipo*.
2. accettare H_0 quando H_0 è falsa; questo è usualmente chiamato un *errore di secondo tipo*.

• Ovviamente

$$P(\text{errore } 1^\circ \text{ tipo}) = 1 - P\left(\begin{array}{c}\text{accettare } H_0 \\ \text{quando } H_0 \text{ è vera}\end{array}\right)$$

Quindi, costruire, come abbiamo fatto noi, un test per cui

$$P\left(\begin{array}{c}\text{accettare } H_0 \\ \text{quando } H_0 \text{ è vera}\end{array}\right) = 1 - \alpha$$

equivale ad utilizzare un test in cui la probabilità di commettere un errore di 1° tipo sia prefissata ed uguale ad α.
• O, in altre parole, il livello di significatività di un test è la probabilità che il test "commetta" un errore di 1° tipo.
• Si noti, viceversa, come, nella costruzione utilizzata, la probabilità di commettere un errore di 2° tipo non sia stata esplicitamente considerata[22].

[21] grande fantasia, giusto?

[22] con la sola eccezione di pagina 40 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. *I* dati sperimentali sono compatibili con H_0?

più che

B. *Q*uale tra H_0 e H_1 è vera?

Tra l'altro, come vedremo, a volte H_1 non è neanche specificabile.
• 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).
E' infatti intuitivamente chiaro che più n è grande più possiamo sperare di rendere piccoli ambedue i tipi di errore.
Incidentalmente, è proprio così che l'azienda ha scelto di "campionare" 5 lastre (e non di più o non di meno).
• Questo avviene, usualmente, utilizzando la *funzione di potenza* del test.
• Nel caso che stiamo considerando è definita come

$$\gamma(\mu) = P\left(\begin{array}{c}\text{rifiutare } H_0 \text{ quando } \mu \text{ è la} \\ \text{vera media}\end{array}\right)$$

• Si osservi che la funzione di potenza riassume le proprietà del test. Infatti

⋆ $\gamma(14)$, ovvero la funzione di potenza calcolata al valore della media previsto da H_0, è uguale alla probabilità di commettere un errore di I tipo e, nella costruzione di prima, $\gamma(14) = \alpha$;

⋆ $\gamma(\mu)$ con $\mu \neq 14$, ovvero i valori assunti dalla funzione di potenza per i valori di μ non previsti da H_0, forniscono le probabilità di *non* commettere un errore di II tipo.

• Proviamo a calcolarla. Ricordando che la probabilità di accettare H_1 è ovviamente uguale a 1 meno la probabilità di accettare H_0 scriviamo

$$\gamma(\mu) = 1 - P(-z_{1-\alpha/2} \leq \sqrt{n}(\overline{y} - \mu_0)/\sigma \leq z_{1-\alpha/2})$$

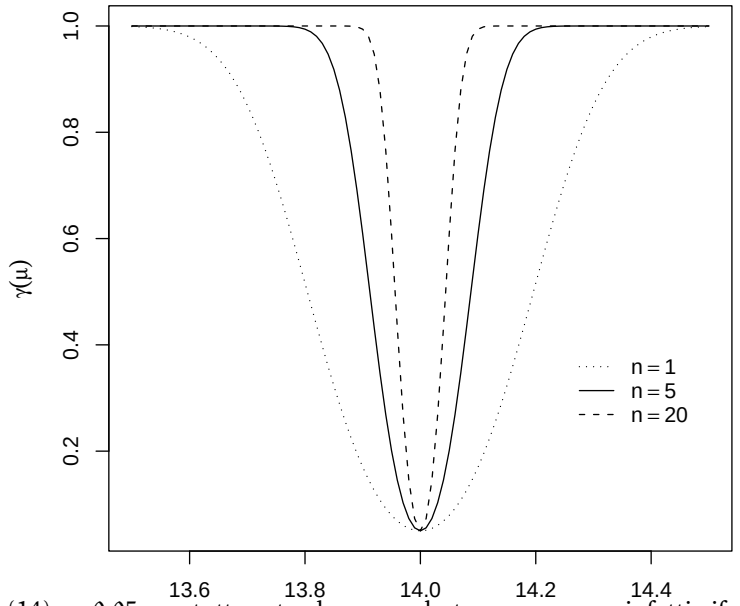
dove la probabilità deve essere calcolata supponendo che la media della normale che genera le osservazioni sia μ.
Sommando e sottraendo $\sqrt{n}\mu$ al numeratore della funzione test otteniamo

$$\begin{aligned}\gamma(\mu) &= 1 - P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\overline{y} - \mu_0 + \mu - \mu)}{\sigma} \leq z_{1-\frac{\alpha}{2}}\right) = \\ &= 1 - P\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\overline{y} - \mu)}{\sigma} + \delta_n(\mu) \leq z_{1-\frac{\alpha}{2}}\right)\end{aligned}$$

dove $\delta_n(\mu) = \sqrt{n}(\mu - \mu_0)/\sigma$.
Ricordando che, quando μ è la vera media, $\overline{y}$ si distribuisce come una normale di media μ e varianza σ^2/n e che, quindi, $\sqrt{n}(\overline{y} - \mu)/\sigma$ si distribuisce come una normale standard, otteniamo

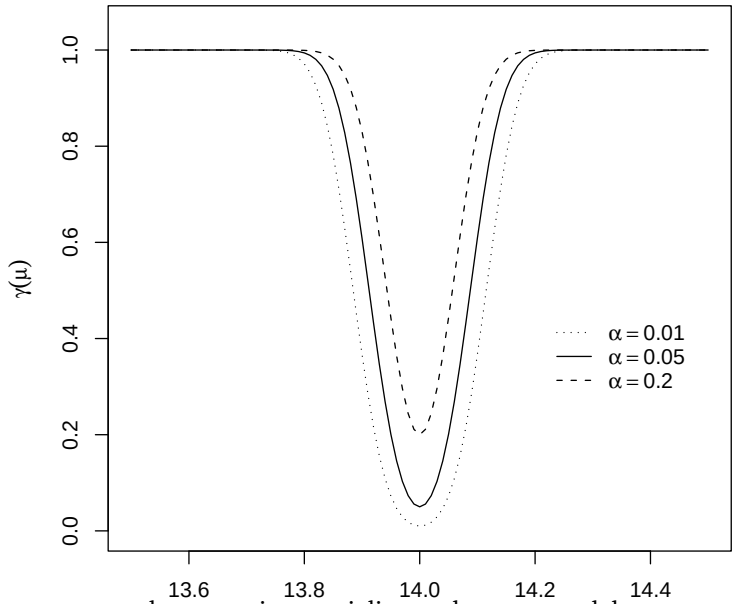
$$\begin{aligned}\gamma(\mu) &= 1 - P\left(-z_{1-\frac{\alpha}{2}} - \delta_n(\mu) \leq N(0,1) \leq z_{1-\frac{\alpha}{2}} - \delta_n(\mu)\right) = \\ &= 1 - \left[\Phi\left(z_{1-\frac{\alpha}{2}} - \delta_n(\mu)\right) - \Phi\left(-z_{1-\frac{\alpha}{2}} - \delta_n(\mu)\right)\right].\end{aligned}$$

Funzione di potenza del test considerato per tre valori della numerosità campionaria ($\alpha = 0{,}05$ in tutti e tre i casi)



(i) $\gamma(14) = 0{,}05$ per tutte e tre le curve; le tre curve sono infatti riferite a test costruiti per avere una probabilità di errore di I tipo uguale a 0,05.
(ii) A parità di n, la "potenza del test" aumenta man mano che μ si allontana da $\mu_0 = 14$, ovvero, più ci allontaniamo da H_0 più diventa probabile che il test ci segnali che H_0 è falsa. Questo sembra molto "logico". Comportamenti differenti sarebbero "sospetti".
(iii) Se considero un valore di $\mu \neq \mu_0$, la potenza aumenta all'aumentare della numerosità campionaria (n). Ovvero, più n è grande, più il test è in grado di mettere in luce differenze della vera media dal valore previsto da H_0.

Funzione di potenza del test considerato per tre valori di α ($n = 5$ in tutti e tre i casi)



Anche aumentando α possiamo migliorare la potenza del test, ovvero la sua capacità di rifiutare H_0 quando, effettivamente, H_0 non è vera.

In questo caso però diminuisce anche la capacità del test di dichiarare H_0 vera quando è effettivamente vera.

Unità C

Dove un prete ortolano incontra una binomiale che gli dice “Hai ragione. Io sono d’accordo con te”

Stima della probabilità di successo, intervalli di confidenza e verifica d’ipotesi nel caso di una binomiale.
Livello di significatività osservato (p-value).

Un esperimento

• Consideriamo in questa unità i risultati di uno dei primi esperimenti di **Mendel**, il grande genetista.
• Mendel aveva selezionato, tra gli altri, due gruppi di piante di piselli:
(i) il primo che presentava solo bacelli verdi
(ii) il secondo che presentava solo bacelli gialli.
• O, quanto meno, quello che Mendel sapeva era che impollinando piante del primo (secondo) gruppo con polline di piante dello stesso gruppo (procedura che aveva ripetuto per alcuni anni) nascevano sempre piante con bacello verde (giallo).
• A questo punto ha impollinato un certo numero di piante del gruppo “giallo” con polline prelevato da piante del gruppo “verde” ottenendo così una 1° generazione di piante incrociate.
Tutte le piante di questa generazione presentavano un bacello verde.
• Poi ha “auto-impollinato” le piante di 1° generazione ottenendo 56 piante di 2° generazione.
Di queste 39 avevano un bacello verde e 17 viceversa presentavano un bacello giallo.
• Quello di cui ci occuperemmo è di utilizzare le informazioni sperimentali per fare delle affermazioni su

$$\vartheta = \mathrm{P}\left(\begin{array}{c}\text{ottenere una pianta di } 2^\circ \\ \text{generazione con bacello verde}\end{array}\right)$$

• Abbiamo almeno due questioni da discutere preliminarmente:
1. esiste effettivamente un qualche spazio di probabilità in cui ϑ è definito?
2. quale relazione esiste tra ϑ ed i risultati sperimentali (39 bacelli verdi su 56 piante di 2° generazione)?

Si osservi in particolare che se non rispondiamo alla seconda domanda non possiamo pensare di utilizzare i dati per farci raccontare che cosa sanno sul parametro di interesse.

Un possibile modello

• Per quanto riguarda la prima domanda le risposte sono *probabilmente* tante quante le definizioni di probabilità.
• Una possibilità consiste nel pensare ad infinite ripetizioni dell'esperimento.
• Ad esempio, potremmo pensare di, per un numero infinito di generazioni,

(i) fare "auto-impollinare" metà dei "verdi" e metà dei "gialli" (la riproduzione separata ci serve per avere la materia prima per gli incroci)

(ii) incrociare le restanti metà e poi fare "auto-impollinare" le piante prodotte dall'incrocio.

• Oppure potremmo pensare ad un numero infinito di appassionati di genetica che vadano al mercato, comprano dei semi di pisello, selezionano due ceppi, uno "verde" e l'altro "giallo" e poi ripetano l'esperimento di Mendel.
• In ambedue i casi, tutto questo impollinare, far crescere, re-impollinare,... genera un numero infinito di piante di 2° generazione alcune delle quali con bacello verde, altre con bacello giallo.
• ϑ può essere identificato con la proporzione di piante "verdi" in questo insieme infinito di piante.
Stiamo, ovviamente, adottando una interpretazione *frequentista* dell'idea di probabilità.

• Indichiamo con

– y il numero di piante con bacello verde

– n in numero totale delle piante di 2° generazione.

Nel caso dell'esperimento descritto $y = 39$ e $n = 56$.
• La seconda questione è che relazione esiste tra (y, n) e ϑ.
Se accettiamo l'idea che Mendel non abbia fatto niente per influenzare i risultati ed abbia semplicemente lasciato lavorare il *"caso"*, possiamo assimilare l'esperimento all'estrazione casuale di n piante da un'urna costituita da tutte le piante di 2° generazione che abbiamo "evocato".
Ma allora[1]

$$y \sim \text{Bi}(n, \vartheta) \tag{C.1}$$

ovvero, il numero di piante "verdi" tra le n estratte può essere visto come una determinazione di una binomiale con *probabilità di successo* ϑ e *numero di prove* n.
• Si osservi che la (C.1) è cruciale perchè precisa la relazione tra quello che conosciamo (y e n) e quello che vogliamo conoscere (ϑ).

[1] [Probalità 22].

Stima di ϑ

• Uno stimatore "naturale"[2] di ϑ è

$$\hat{\vartheta} = \frac{y}{n}$$

ovvero la proporzione di piante "verdi" nei dati.
• Nel caso dell'esperimento di Mendel, $\vartheta = 39/56 \approx 0{,}70$.
• Ovviamente, se y è una variabile casuale anche $\hat{\vartheta}$ lo è.
• Lo studio della sua distribuzione è importante perchè permette di acquisire una idea sulla dimensione dell'errore di stima
• La media e la varianza di $\hat{\vartheta}$ sono facilmente calcolabili dai momenti primi e secondi di una binomiale[3]:

$$\mathrm{E}\left\{\hat{\vartheta}\right\} = \vartheta, \quad \mathrm{var}\left\{\hat{\vartheta}\right\} = \frac{\vartheta(1-\vartheta)}{n}.$$

Si osservi che $\hat{\vartheta}$ è uno stimatore non distorto della vera probabilità ϑ.
• E' inoltre possibile mostrare che $\hat{\vartheta}$ è uno stimatore consistente in senso forte di ϑ.
• Anche la distribuzione *esatta* di $\hat{\vartheta}$ può essere facilmente determinata.
Infatti, $\hat{\vartheta} \in \Theta_n = \{0/n, 1/n, \ldots, n/n\}$ e, per qualsivoglia $a \in \Theta_n$, risulta

$$\mathrm{P}(\hat{\vartheta} = a) = \mathrm{P}(y = na) = \binom{na}{n} \vartheta^{na}(1-\vartheta)^{n-na}.$$

• Da questa distribuzione è possibile ottenere intervalli di confidenza (e test) esatti per ϑ. I calcoli non sono però del tutto facili ed è necessario un calcolatore (in R è possibile utilizzare la funzione binom.test).
• Per questo motivo consideremo una procedura alternativa che, per quanto approssimata, è frequentemente utilizzata nelle applicazioni.

[2] forse l'unico "naturale" nel senso che qualsiasi altra scelta scelta sembra artefatta.
[3] [Probalità 24].

Approssimazione normale

• 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 $[a, b]$

$$\mathrm{P}\left(a \le \frac{\hat{\vartheta} - \vartheta}{\sqrt{\vartheta(1-\vartheta)/n}} \le b\right) \approx \mathrm{P}(a \le \mathrm{N}(0,1) \le b)$$

• Si ritiene generalmente che l'approssimazione normale "funzioni almeno decorosamente" quando sia $n\vartheta$ che $n(1-\vartheta)$ sono più grandi di 5.
• Se $(\hat{\vartheta} - \vartheta)/\sqrt{\vartheta(1-\vartheta)/n}$ è approssimativamente una normale standard allora, sempre approssimativamente,

$$(\text{errore di stima}) = (\hat{\vartheta} - \vartheta) \sim \mathrm{N}(0, \vartheta(1-\vartheta)/n).$$

• Si osservi che 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

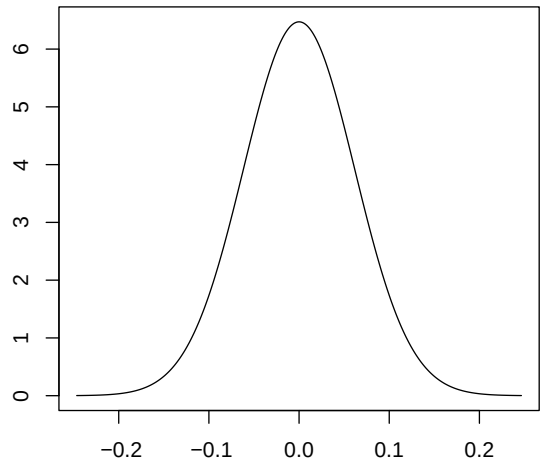
$$\widehat{\mathrm{var}}(\hat{\vartheta} - \vartheta) = \frac{\hat{\vartheta}(1-\hat{\vartheta})}{n} \approx \frac{0.70(1-0.70)}{56} \approx 0{,}0038$$

ovvero, approssimazione dopo approssimazione, siamo arrivati alla conclusione che

> l'errore di stima "subito" da Mendel è, grossomodo, normale di media zero e scarto quadratico medio 0,062.

La densità di questa distribuzione è mostrata nel lucido seguente.

Approssimazione della distribuzione dell'errore di stima



Si osservi che la densità è abbastanza "dispersa", ovvero che possiamo aspettare differenze tra il valore stimato ($\approx 0{,}7$) e il vero valore dell'ordine del $\pm 10\%$ senza fare riferimento ad eventi particolarmente poco probabili.

Intervalli di confidenza

- La distribuzione stimata per $\hat{\vartheta} - \vartheta$ può essere usata per costruire intervalli di confidenza (almeno approssimativamente) di livello $1 - \alpha$ prefissato.
- Infatti se la distribuzione di $\hat{\vartheta} - \vartheta$ è approssimativamente una normale di media nulla e scarto quadratico medio $0{,}062$ allora possiamo scrivere[4]

$$P(-0{,}062 \times z_{1-\alpha/2} \leq \hat{\vartheta} - \vartheta \leq 0{,}062 \times z_{1-\alpha/2}) \approx 1 - \alpha \qquad \text{(C.2)}$$

dove, al solito, con z_p indichiamo il quantile p-simo di una normale standard.
- La (C.2) può essere scritta come

$$P(\hat{\vartheta} - 0{,}062 \times z_{1-\alpha/2} \leq \vartheta \leq \hat{\vartheta} + 0{,}062 \times z_{1-\alpha/2}) \approx 1 - \alpha$$

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 (approssimativamente) un intervallo di confidenza di dimensione $1 - \alpha$ per ϑ.

[4] perchè?

Con i dati di Mendel

• Supponiamo di voler calcolare un intervallo di confidenza di livello

$$1 - \alpha = 0{,}9.$$

Allora,

$$\alpha = 0{,}1 \text{ e quindi } 1 - \alpha/2 = 0{,}95.$$

• Da una tavola della distribuzione normale (o utilizzando un programma appropriato) troviamo che

$$z_{0{,}95} \approx 1{,}65.$$

• Sappiamo già che $\hat{\vartheta} \approx 0{,}7$ e che

$$\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}} = \sqrt{\frac{0{,}7 \times 0{,}3}{56}} \approx 0{,}062.$$

Quindi, la semi-ampiezza dell'intervallo di confidenza è

$$z_{1-\alpha/2}\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}} = 1{,}65 \times 0{,}062 = 0{,}102.$$

• L'intervallo è quindi

$$[0{,}7 - 0{,}102\ ;\ 0{,}7 + 0{,}102] = [0{,}598\ ;\ 0{,}802].$$

Per Mendel ϑ vale $0{,}75$

• L'idea a cui stava lavorando Mendel è che ad ogni carattere osservabile (ad esempio, colore dei bacelli) corrisponda una coppia di geni.
• Questa coppia si divide al momento della riproduzione e la coppia di geni del "figlio" si forma combinando un gene del "padre" e un gene della "madre".
• Indichiamo con "V" un gene contenente l'informazione "bacello verde" e con "g" un gene associato a "bacello giallo".
• Il fatto che il gruppo "verde" per generazioni abbia dato solo piante con bacelli verdi viene da Mendel interpretato come indicazione del fatto che per tutte le piante del gruppo la coppia di geni è "VV".
• Simmetricamente, nel gruppo "giallo" la coppia di geni di tutte le piante deve essere "gg".
• Facendo incrociare piante del gruppo "giallo" con piante del gruppo "verde" dovremmo quindi ottenere una 1° generazione in cui tutte piante hanno la coppia di geni uguale a "Vg" (o se vogliamo anche "gV" ma l'ordine non è importante per Mendel).
• Il fatto che tutte le piante di questa generazione mostrino un bacello verde viene da Mendel interpretato come una manifestazione del fatto che "V domina su g". Maiuscole e minuscole sono state usate proprio per evidenziare questo aspetto.

• Arriviamo alla 2° generazione. Poichè tutte le piante di prima generazione sono "Vg" al momento della riproduzione metà dei geni forniti dal "papà" sono "V" e metà "g". Lo stesso vale per la "mamma".
• Quindi, le piante della 2° generazione possono essere o "VV" o "Vg" o "gg".
• Parte della teoria di Mendel è che le coppie si "ricompongono casualmente" (ovvero un gene "V" del "papà" ha probabilità 0,5 di "accasarsi" sia con un gene "V" che con un gene "g" della "mamma").
• Ma allora

$$\begin{aligned} P(\text{“VV”}) &= \frac{1}{4} \\ P(\text{“Vg”}) &= \frac{1}{2} \\ P(\text{“gg”}) &= \frac{1}{4} \end{aligned}$$

e quindi, ricordando che "V" domina su "g",

$$\vartheta = P(\text{“VV”}) + P(\text{“Vg”}) = \frac{3}{4}.$$

Verifica dell'ipotesi di Mendel

• Mendel aveva condotto l'esperimento essenzialmente per verificare il seguente sistema di ipotesi:

$$\begin{cases} H_0 : \vartheta = \vartheta_0 \\ H_1 : \vartheta \neq \vartheta_0 \end{cases}$$

con $\vartheta_0 = 0{,}75$.
• Volendo utilizzare un test statistico sembra ragionevole basare la decisione sulla distanza tra

- la stima di ϑ calcolata dai dati e
- il valore per ϑ previsto da H_0.

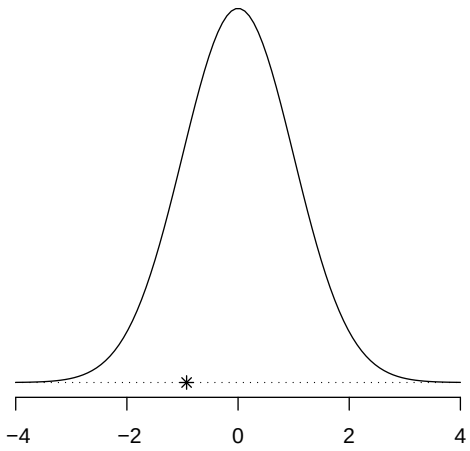
• Una possibile *statistica test* è[5]

$$T = \frac{\hat{\vartheta} - \vartheta_0}{\sqrt{\vartheta_0(1-\vartheta_0)/n}}$$

• Se l'ipotesi nulla è vera, ci aspettiamo che T assuma valori vicini allo zero (sia positivi che negativi).
• Viceversa se la vera probabilità di ottenere una pianta di 2° generazione è differente da ϑ_0 allora ci aspettiamo che T sia "più lontana" da zero.
• Quando è vera H_0, ricordando l'approssimazione normale alla binomiale, sappiamo che 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.

[5] Si osservi che come nell'unità precedente preferiamo lavorare con una versione "standardizzata" della differenza; la cosa è però irrilevante poichè il tutto si concretizza nella divisione per una costante

Confronto grafico



- Con i dati dell'esperimento che stiamo considerando $T \approx -0{,}93$.
- Il grafico mostra la densità di una normale standard con, sull'asse delle ascisse, indicato il valore osservato, della statistica test.
- Questo valore potrebbe benissimo essere stato generato dalla distribuzione disegnata ovvero lo scostamento tra la percentuale di piante con bacello verde nel campione ($\approx$ 70%) e quello previsto dalla teoria di Mendel (75%) potrebbe benissimo essere dovuto al caso.
- Non sembrano quindi esserci elementi per rifiutare l'ipotesi di Mendel che $\vartheta = 0{,}75$.

Un test di dimensione prefissata...

• Volendo una regola precisa per accettare del tipo

"se accade questo accetto H_0 altrimenti rifiuto"

possiamo procedere come nell'unità precedente.

• In particolare, non sembra irragionevole

(a) accettare l'ipotesi nulla se $|T|$ è sufficientemente piccolo, ovvero usare una regola del tipo

"accetto H_0 se $|T| \leq h$"

(b) e fissare h chiedendo che

$$P(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera}) = 1 - \alpha \qquad \text{(C.3)}$$

per qualche valore prefissato e non troppo grande di α.

• Ricordando che T è approssimativamente distribuito come una normale standard, possiamo concludere che ponendo

$$h = z_{1-\alpha/2}$$

otteniamo una regola che almeno approssimativamente soddisfa la (C.3).

• Quindi, a parte per la statistica test che è ovviamente differente, siamo arrivati ad una procedura "accetto/rifiuto" la cui meccanica è quella dell'unità B.

• Nel caso in esame, ad esempio, se scegliamo $\alpha = 0{,}1$ allora come già ricordato $z_{0,95} \approx 1{,}65$ e poichè $|T| \approx 0{,}93 \leq 1{,}65$ accettiamo H_0.

. . . [segue dal titolo precedente] è un pó troppo manicheo

• Nell'unità precedente (controllo spessore lastre di metallo) *dovevamo* per forza arrivare ad una regola del tipo "accetto/rifiuto". Infatti alle due alternative corrispondevano due azioni immediate. In un certo senso, eravamo ad un bivio e dovevamo decidere se andare verso destra o verso sinistra (= bloccare o continuare la produzione).

• Nel caso che stiamo considerando in questa unità questa urgenza non esiste. Ed allora, ridurre il tutto a "confrontiamo $|T|$ con una soglia h e se è minore accettiamo mentre se è maggiore rifiutiamo" è quantomeno inutilmente manicheo.

• Si pensi ad esempio al fatto che piccole differenze in T ci possono portare a conclusioni drammaticamente differenti. Ad esempio, nel caso in esame un valore di T pari a $1{,}649$ od a $1{,}651$ ci racconterebbero essenzialmente la stessa storia sulla teoria di Mendel. Però insistendo a fare un test con $\alpha = 0{,}1$ in un caso concluderemmo che Mendel ha ragione e nell'altro che ha torto.

Livello di significatività osservato

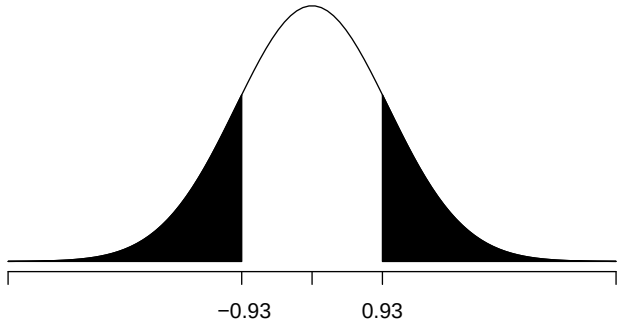
Se Mendel dovesse scrivere ai giorni nostri una memoria sulla sua teoria e sui risultati degli esperimenti da lui condotti probabilmente presenterebbe la parte di risultati che stiamo commentando con una frase del tipo

> . . . delle 56 piante della 2° generazione 39 (70%) mostravano un bacello verde ($p = 0{,}35$). . .

Quel "$p = \ldots$" tra parentesi indica che è stato fatto un test. Viene usualmente chiamato *livello di significatività osservato* o *p-value* o semplicemente *p* del test e costituisce la maniera più comune con cui vengono presentati i risultati di una verifica d'ipotesi.
In generale, la definizione è

$$\begin{pmatrix} \text{livello di} \\ \text{significatività} \\ \text{osservato} \end{pmatrix} = \begin{pmatrix} \text{probabilità di osservare} \\ \text{sotto } H_0 \text{ un valore di T più} \\ \text{o ugualmente lontano da} \\ H_0 \text{ di quanto} \\ \text{effettivamente osservato} \end{pmatrix}$$

Un grafico può aiutare



La curva mostra la densità di una normale standard. 0,93 è il valore della statistica test calcolata con i dati di Mendel. Poichè "lontano da 0 vuol dire lontano da H_0" l'area "annerita" fornisce una approssimazione della probabilità di osservare quando è vera H_0 un valore più lontano (o almeno ugualmente lontano) dall'ipotesi nulla di quanto osservato.

Esercizio 1. Perchè solo una "approssimazione della probabilità. . . "?

Esercizio 2. Si verifichi, utilizzando una tavole della normale, che l'area vale circa 0,35.

Interpretazione

• Il livello di significatività osservato costituisce una misura di quanto l'ipotesi nulla è plausibile sulla base dei dati.

• Varia tra 0 e 1[6] e più è grande più i dati "sono vicini ad H_0".

• In particolare si osservi che:

  – Se vale 0 vuol dire che sotto H_0 non è possibile osservare nessun altro valore più lontano da H_0, ovvero, il valore osservato per T è uno dei più lontani possibili.

  – Se vale 1 vuol dire che sotto H_0 tutti i possibili valori osservabili per T sono "più lontani" di quello osservato, ovvero, quello osservato è uno dei "più vicini possibili".

• Inoltre conoscendo il livello di significatività osservato possiamo facilmente dire se i dati sono significativi contro H_0 per qualsiasi valore di α.
Infatti,

$$\text{se (livello significatività osservato)} < \alpha$$

allora i risultati sono significativi al $100\alpha\%$ (il test con α prefissato rifiuta H_0) mentre

$$\text{se (livello significatività osservato)} \geq \alpha$$

allora i risultati sono non significativi al $100\alpha\%$ (il test con α prefissato accetta H_0)

[6] ovviamente, è una probabilità!

• Lo stretto legame esistente tra i test con livello di significatità prefissato e il livello di significatività osservato giustifica, tra le altre cose, il fatto che è abbastanza usuale parlare di risultati

- *non significativi* se il livello di significatività osservato è maggiore di 0,1 (10%);
- *significativi* se è compreso tra 0,01 e 0,05 (tra uno su 100 e uno su 20);
- *altamente significativi* se è minore di 0,01.

La "significatività" è da intendere contro H_0 e, difatti, negli ultimi due casi i dati ci stanno suggerendo di rifiutare l'ipotesi nulla.
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, le soglie utilizzate (0,01, 0,05 e 0,1) fanno parte della tradizione ma non per questo hanno qualcosa di *sacro*.

Unità D

Dove un pediatra anti-militarista incontra un giudice anti-femminista

Un esempio di verifica d'ipotesi in cui l'ipotesi alternative non è ben definita.

Un caso giudiziario

• 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 1969 il dott. Spock fu processato da un tribunale federale statunitense 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 da una lista contenente centinaia di migliaia di *elegibili* 350 possibili giurati; la legge prevedeva che l'estrazione doveva essere casuale e fatta in maniera tale da garantire a ciascun elegibile la stessa probabilità di estrazione
- poi, sia l'accusa che la difesa potevano ricusare parte di questi potenziali giurati
- e, infine, la giuria effettiva veniva estratta tra i giurati "non eliminati".

• Il processo fu affidato ad un giudice federale di nome 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, in media, 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 elegibili 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.

Un possibile sistema di ipotesi

• Possiamo inquadrare la questione di dare un giudizio sul comportamento del giudice Ford come un problema di verifica di ipotesi. In prima battuta il sistema di ipotesi è

$$\begin{cases} H_0 : \text{l'estrazione è stata fatta secondo la legge} \\ H_1 : \text{l'estrazione è stata "truccata"} \end{cases}$$

• 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 elegibili;
  – D il numero di donne tra gli elegibili.
• La legge prevede che si debba
— estrarre un primo individuo assegnando uguale probabilità a tutti gli elegibili;
— 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 probablità 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.
• La conclusione è quindi che, con una buona approssimazione, se si segue la legge,

> il numero di donne tra i potenziali giurati è il risultato del conteggio di quante donne vengono estratte in 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.

• Ma allora, ricordandoci che tra l'altro sappiamo che nel caso in esame $D/N = 0{,}53$, ovvero che il 53% degli eleggibili è donna

$$\begin{pmatrix} \text{numero donne} \\ \text{estratte} \end{pmatrix} \sim \text{Bi}(350, 0{,}53)$$

• 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.
• Siamo quindi davanti ad un problema di verifica d'ipotesi in cui H_0 è completamente specificata, ed in particolare, è esattamente del tipo che abbiamo considerato nella seconda parte dell'unità sui dati di Mendel. Viceversa, H_1 è essenzialmente nebulosa.

Ha senso lo stesso fare un test?

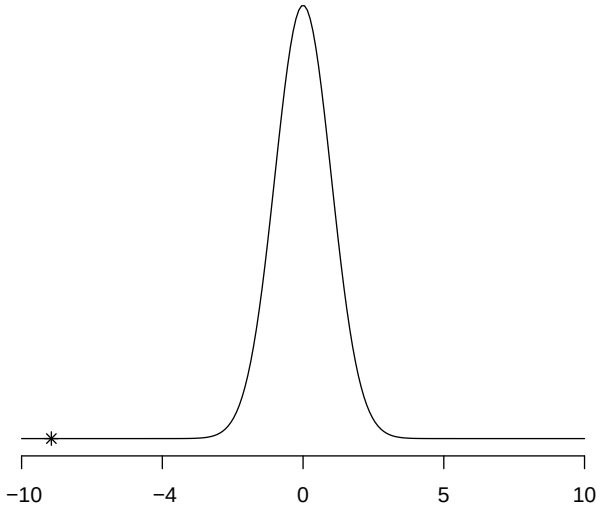
La risposta è si. Con un test statistico cerchiamo di valutare se i dati potrebbero essere stati generati dal meccanismo previsto dall'ipotesi nulla. E questo è quello che vogliamo fare nel presente contesto visto che la domanda che ci stiamo ponendo è:

> "E' plausibile che il giudice Ford abbia seguito la legge ed estratto solo 102 donne?".

In maniera analoga a quanto fatto nell'unità precedente possiamo "misurare la distanza" tra quanto osservato e quanto previsto dalla legge mediante la statistica test

$$T = \frac{\dfrac{\text{numero donne estratte}}{\text{numero potenziali giurati}} - 0{,}53}{\sqrt{0{,}53(1-0{,}53)/350}}.$$

Se H_0 è vera, T si distribuisce, almeno approssimativamente, come una normale stardard. Quindi, confrontando il valore osservato di T i valori "previsti" da questa distribuzione possiamo dare una risposta alla domanda.



- Il valore di T calcolato dai dati disponibili (102 donne tra 350 giurati potenziali) è $-8{,}94$.
- Il grafico mostra la densità di una normale standard. L'asterisco sull'asse delle ascisse indica il valore osservato di T.
- Il valore è troppo spostato verso destra. L'ipotesi nulla non sembra plausibile.

Il livello di significatività osservato

- Il livello di significatività osservato in questo caso potrebbe essere calcolato come (si veda il grafico a pagina 73)

$$P(N(0,1) \leq -8{,}94) + P(N(0,1) \geq 8{,}94)$$

- $8{,}94$ è "fuori" da tutte le usuali tavole della normale. Però possiamo calcolare la probabilità che ci interessa utilizzando un calcolatore ed una appropriata funzione.
- Procedendo in questa maniera il valore che troviamo è $\approx 3{,}8 \times 10^{-19}$.
- Ora, è chiaro che tutto può capitare. Anche di estrarre solo 102 donne. Però questo calcolo 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!

Unità E

Tonsille e *Streptococcus pyogenes*

Verifica dell'ipotesi di indipendenza in una tabella a doppia entrata

Il problema e i dati

• Nel corso di uno studio sulla determinazione di possibili fattori prognostici (predittivi) per alcune malattia otorino-laringoiatriche sono state rilevate le seguenti due variabili su 1398 bimbi o ragazzi:

(a) presenza (in un tampone nasale) di `Streptococcus pyogenes`; variabile dicotomica con modalità "portatore" e "non portatore";

(b) stato delle tonsille rilevato utilizzando la scala qualitativa ordinata:
   (i) normali (abbreviato in $+$),
   (ii) leggermente ingrossate ($++$)
   (iii) ingrossate ($+++$).

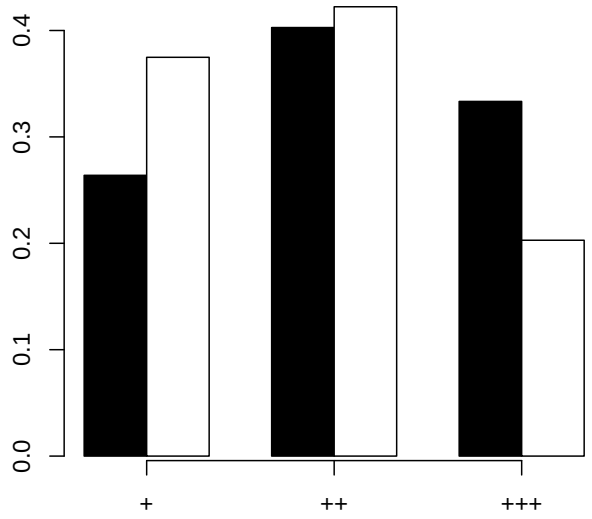
• I bimbi erano stati scelti casualmente, mediante sorteggio dalle liste anagrafiche, tra tutti gli individui tra i 3 e i 15 di età residenti in un ampia e popolosa regione inglese.

• La seguente tabella, che contiene le *frequenze osservate* nel campione, riassume i dati raccolti.

	tonsille			
streptococcus pyogenes	$+$	$++$	$+++$	totale
portatore	19	29	24	72
non portatore	497	560	269	1326
totale	516	589	293	1398

• Il problema che affrontiamo è se esiste o no una qualche forma di associazione tra le due variabili.

Diagramma a barre



- Il grafico mostra le distribuzione dello “stato delle tonsille” condizionato a
  - “portatore” (barre nere) e
  - “non portatore” (barre bianche).

  l’altezza della barre è proporzionale alle frequenze relative.
- I portatori sembrano avere le tonsille “più grosse”.

La popolazione di riferimento

- Il grafico a barre mostra chiaramente che la distribuzione di “stato delle tonsille” è diversa tra i portatori e i non portatori.
- Quindi, nei dati campionari c’è una qualche forma di dipendenza tra le due variabili.
- Una domanda che è spontaneo porsi è se e a chi è possibile estendere questi risultati.
- In realtà, se ci pensa questa è la vera domanda. Infatti, ci scusino i 1398 ragazzi, ma le tonsille di alcuni sconosciuti, probabilmente, non sono uno dei nostri principali problemi.
- I dati, viceversa, ci possono interessare per quello che ci possono raccontare sulla relazione intercorrente *in generale* tra `Streptococcos Pyogenes` e tonsille.
- Gli elementi del campione sono stati estratti casualmente tra i bimbi di una particolare regione geografica. Possiamo allora pensare che ci possano parlare direttamente della relazione esistente tra le due variabili in questo più grande gruppo di individui. Ovvero, l’insieme dei bimbi e ragazzini tra 3 e 15 abitanti nella regione inglese considerata costituisce quella che usualmente viene chiamata la *popolazione di riferimento*.
- Quello che vogliamo fare è “interrogare” i dati campionari per ottenere informazioni sulle caratteristiche di questa popolazione.
- Al solito, la prima cosa da discutere sarà la relazione che esiste tra il campione e la popolazione.

Breve digressione sui bimbi norvegesi, italiani, nigeriani,...

• Sarebbe interessante se i dati ci parlassero di tutti i bambini del mondo.
• Però questo richiede che non ci siano differenze, rispetto ai caratteri considerati, tra i bimbi inglesi (anzi di una particolare regione dell'Inghilterra) e, ad esempio, i bimbi nigeriani.
• Infatti nel campione non ci sono bimbi nigeriani. E quindi, tutto quello di particolare che riguarda quest'ultimi non può essere studiato con questi dati.
• Ovvero, un campione di bimbi inglesi è al più *rappresentativo* di tutti i bimbi inglesi[1].
• *Noi* possiamo anche decidere che le conclusioni che i dati ci suggeriscono valgono anche per i bimbi della Nigeria. Ma si tratta appunto di una *nostra* decisione.
• E, come è ovvio, estendere le conclusioni di una indagine su di una popolazione ad altre popolazioni è intrisincamente *pericoloso*. L'estensione può avvenire solo tramite nuovi studi (sulle altre popolazioni). Fino a che questi non sono condotti, le conclusioni su di una popolazione sono, al più, ipotesi da verificare per le altre.

[1] ovvero della popolazione in cui è stato estratto. E può anche non esserlo se l'estrazione è stata in qualche forma truccata (si pensi al giudice Ford!)

Ascensori, aspirine e la mutabilità dei comportamenti umani

• Quanto detto deve *sempre* essere tenuto presente.
• Ovvero, dobbiamo sempre chiederci di quale popolazione i dati sono rappresentativi. E dobbiamo stare attenti a non estendere in maniera arbitraria la validità delle storie che ci facciamo raccontare dai dati.
• Questo è importante, in modo particolare, nell'ambito delle scienze sociali[2]
• I meccanismi fisici, chimici e biologici sono piuttosto stabili nel tempo e nello spazio. Le leggi con cui si costruiscono gli ascensori a Oslo e a Sidney sono le stesse. E in tutte le farmacie del mondo contro il mal di testa si trovano prodotti che contengono acido acetilsalicilico (il prodotto commerciale più comune è l'aspirina). E, sempre senza differenza tra razze e ambienti, l'abuso di acido acetilsalicilico aumenta il rischio di gastrite.
• Lo stesso non si può dire per i fenomeni sociali. Due comunità separate da pochi chilometri possono avere comportamenti molto diversi. La stessa comunità a distanza di pochi anni può presentare comportamenti diversi,...

[2] che includono l'economia.

Una tabella *fantasma*

• Ritorniamo a considerare l'insieme dei bimbi tra i 3 e i 15 anni residenti nella regione considerata.
• Se le due variabili fossere state rilevate su *tutti* i bimbi avremmo potuto costruire una tabella, analoga a quella di pagina 86, del tipo

	tonsille			
streptococcus pyogenes	$+$	$++$	$+++$	totale
portatore	F_{11}	F_{12}	F_{13}	F_{1+}
non portatore	F_{21}	F_{22}	F_{23}	F_{2+}
totale	F_{+1}	F_{+2}	F_{+3}	N

dove
– N indica il numero di bimbi in quell'area dell'Inghilterra,
– F_{11} il numero di bimbi che sono portatori ma hanno le tonsille normali,
– F_{12} il numero di bimbi che sono portatori e hanno le tonsille leggermente ingrossate.
– ...
• La tabella non la conosciamo visto che, ad esempio, per conoscere F_{11} avremmo dovuto fare un tampone nasale a tutti i bimbi e ragazzini della regione. Per questo è una tabella *fantasma*.
• E' però la tabella di interesse: ci racconta, o meglio, lo farebbe se la conoscessimo, che cosa accade nella popolazione di riferimento.

Che relazione esiste tra la tabella osservata e quella *fantasma*?

• Dividiamo tutte le frequenze della tabella *fantasma* per N ottenendo

	tonsille			
streptococcus pyogenes	$+$	$++$	$+++$	totale
portatore	π_{11}	π_{12}	π_{13}	π_{1+}
non portatore	π_{21}	π_{22}	π_{23}	π_{2+}
totale	π_{+1}	π_{+2}	π_{+3}	1

• Il campione è stato formato:
– estraendo un bimbo tra gli N componenti della popolazione;
– estraendo un altro bimbo tra gli $N-1$ bimbi non estratti alla prima estrazione;
– ...;
– estraendo un bimbo tra gli $N-1397$ bimbi non estratti nelle prime 1397 estrazioni.
In tutte le estrazioni, è stata assegnata probabilità uguale a tutti i bimbi non ancora estratti.
• Vista la maniera con cui è stato formato il campione,

$$\begin{aligned} &P(1^\circ \text{ bimbo sia un (portatore,}+)) = \pi_{11} \\ &P(1^\circ \text{ bimbo sia un (non portatore,}+)) = \pi_{21} \\ &\qquad\qquad\vdots \\ &P(1^\circ \text{ bimbo sia un (non portatore,}+++)) = \pi_{23} \end{aligned}$$

• Le successive estrazioni non sono tra di loro indipendenti. Infatti, escludere i bimbi già estratti 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.

• Quindi, almeno approssimativamente, le frequenze osservate mostrano come si sono ripartiti nelle 6 "categorie" (portatori,+), (portatori,++), ..., (non portatori,+ + +) i risultati di 1398 esperimenti casuali indipendenti tutti caratterizzati da

$$\begin{aligned} &\text{P(estrarre un (portatore,+))} = \pi_{11} \\ &\text{P(estrarre un (portatore,++))} = \pi_{12} \\ &\qquad\vdots \\ &\text{P(estrarre un (non portatore,+++))} = \pi_{23} \end{aligned}$$

• Ma allora la tabella delle frequenze osservate, ovvero i nostri dati, è, approssimativamente, una determinazione di una variabile casuale Multinomiale$(n, (\pi_{11}, \pi_{12}, \ldots, \pi_{23}))$.

Verifica dell'ipotesi di indipendenza

• Una domanda interessante che possiamo fare ai dati è:

nella tabella fantasma esiste indipendenza in distribuzione? ovvero, nella popolazione di riferimento l'essere o non essere portatore è in qualche maniera associato con lo stato delle tonsille?

• In altre parole

la dipendenza che abbiamo rilevato nel campione è una peculiarità dei soli bimbi estratti e quindi l'abbiamo osservata per puro caso oppure è la manifestazione di una reale associazione tra i due fenomeni esistente nella popolazione?

• Si tratta, ovviamente, di un problema di verifica d'ipotesi che può essere scritto nella forma

$$\begin{cases} H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \;\; i = 1,2 \;\; j = 1,2,3 \\ H_1 : \text{le } \pi_{ij} \text{ non rispettano i vincoli previsti da } H_0 \end{cases}$$

• Infatti se H_0 è vera allora, per $x = +, ++, +++$,

$$\text{P(tonsille} = x|\text{portatore)} = \text{P(tonsille} = x|\text{non portatore)}$$

ovvero, la distribuzione dello stato delle tonsille è uguale tra portatori e non portatori.

• La statistica test più usata è l'X^2 di Pearson. E' certamente una statistica appropriata visto che assume valori, tendelzialmente,

– piccoli quando H_0 è vera e

– grandi quando è falsa.

Frequenze attese e X^2: richiami e applicazione

• *La tabella osservata: notazioni* Supponiamo che la seguente sia una generica tabella di frequenze (assolute) osservate.

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

(i) X e Y sono le due variabili considerate,

(ii) $\{x_1, \ldots, x_c\}$ e $\{y_1, \ldots, y_r\}$ indicano le modalità rispettivamente di X e di Y,

(iii) O_{ij} è il numero di unità statistiche nel campione che presentano simultaneamente la modalità x_j di X e la modalità y_i di Y,

(iv) O_{+j}, $j = 1, \ldots, c$, e O_{i+}, $i = 1, \ldots, r$ sono i totali rispettivamente delle colonne e delle righe, ovvero,

$$O_{+j} = \sum_{i=1}^{r} O_{ij} \text{ e } O_{i+} = \sum_{j=1}^{c} O_{ij}.$$

• *Frequenze attese sotto l'ipotesi di indipendenza.* Sono calcolabili come

$$A_{ij} = \frac{O_{+j}O_{i+}}{n} \quad (i = 1, \ldots, r;\ j = 1, \ldots, c).$$

Consideriamo, ad esempio, la tabella delle frequenze osservate su cui stiamo lavorando[3]. L'applicazione della formula alla prima cella, $i = j = 1$, da

$$A_{ij} = \frac{O_{+1}O_{1+}}{n} = \frac{516 \times 72}{1398} = 26{,}6.$$

La logica è semplice:

— in totale abbiamo trovato 516 bimbi su 1398 con tonsille normali;

— se non c'è differenza tra lo stato delle tonsille dei portatori e dei non portatori, la percentuale di portatori con tonsille normali dovrebbe essere circa uguale a 516/1398;

— ma il numero dei portatori nel campione è 72 e quindi, in ipotesi di indipendenza, ci aspettiamo di trovare circa

$$72 \times \frac{516}{1398}$$

portatori con tonsille normali nel campione;

— e così via per le altre celle della tabella.

[3] lucido 86.

• La tabella riporta le frequenze attese per tutte le celle.

	tonsille			
streptococcus pyogenes	+	++	+++	totale
portatore	26,6	30,3	15,1	72
non portatore	489,4	558,7	277,9	1326
totale	516	589	293	1298

Si osservi che, rispetto alla tabella attesa, nella tabella osservata ci sono troppi portatori con tonsille ingrossata e troppo pochi portatori con tonsille normali. E che il viceversa accade per i non portatori.
• X^2 misura, sostanzialmente, la distanza esistente tra le frequenze osservate e le frequenze attese. E' definito come

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

Nel caso in esame

$$X^2 = \frac{(19 - 26{,}6)^2}{26{,}6} + \cdots + \frac{(269 - 277{,}9)^2}{277{,}9} = 7{,}88.$$

La distribuzione approssimata di X^2

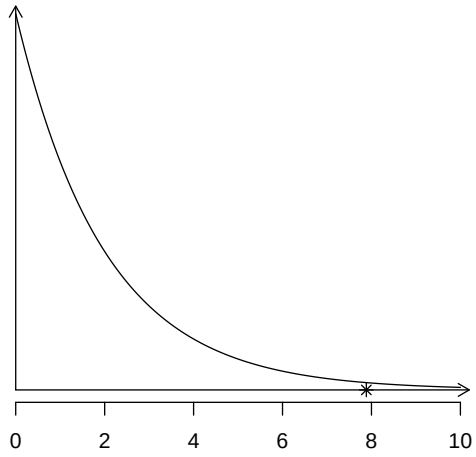
• E' possibile mostrare[4] che se l'ipotesi di indipendenza è vera e nessuna frequenza attesa è troppo piccola allora la distribuzione di X^2 può essere approssimata con la distribuzione di una variabile casuale[5] χ^2.
• La distribuzione χ^2 dipende da un solo parametro, chiamato i gradi di libertà 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[6] è più grande di 5 e migliora man mano che queste aumentano.

[4] rinviamo al solito la dimostrazione di questo risultato a corsi più avanzati
[5] si veda [Probalità 9] per la definizione e per alcune proprietà di una variabile casuale χ^2.
[6] si noti, quelle attese, non quelle osservate

Analisi grafica del risultato

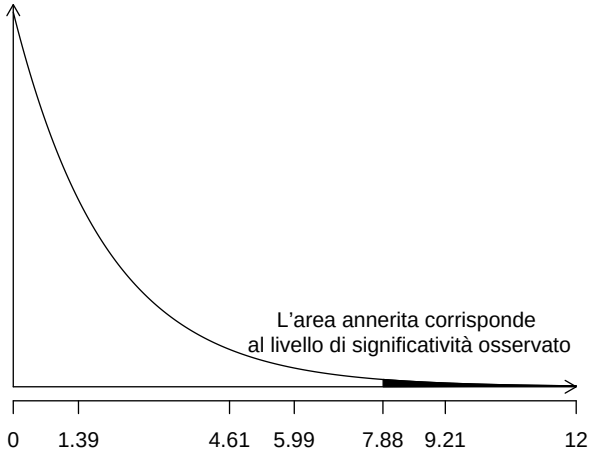


– Il grafico mostra la densità di una v.c. χ^2 con 2 gradi di libertà. L'asterisco sull'asse delle ascisse indica il valore osservato della statistica test.
– Il valore è "moderatamente" ma non "esageratamente" spostato verso destra, ovvero, verso H_1.
– La conclusione potrebbe essere una sorta di "dubbioso rifiuto di H_0".

Livello di significatività osservato (e suo calcolo approssimato da una tavola dei percentili)

• "Lontano da H_0" vuol dire per il test che stiamo considerando "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. Per i dati presentati in questa unità,

$$\begin{pmatrix}\text{livello}\\ \text{significatività}\\ \text{osservato}\end{pmatrix} = \mathrm{P}\begin{pmatrix}\chi^2 \text{ con } 2\\ \text{gradi libertà}\\ \geq 7{,}88\end{pmatrix}$$



• 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

p	0,5	0,90	0,95	0,99
$\chi^2_{2,p}$	1,39	4,61	5,99	9,21

in cui $\chi^2_p(2)$ indica il percentile p-simo di un χ^2 con 2 gradi di libertà.
• Il valore osservato (7,88) è compreso tra il 95-simo e il 99-simo percentile. Ora, per definizione, la probabilità di assumere un valore più grande del 95-simo (99-simo) percentile è 5% (1%). Perciò

$$0{,}01 \leq \text{(livello significatività osservato)} \leq 0{,}05 \qquad \text{(E.1)}$$

• I risultati sono quindi significativi al 5% ma non all'1%. I dati ci suggeriscono tendelzialmente di rifiutare l'ipotesi nulla ma non così chiaramente come ci è accaduto in altri casi.

Unità F

Dove parleremo di “rapporto” tra maschi e femmine e di demenza senile

- Ancora su X^2 e χ^2.
- Test di bontà dell’adattamento di un modello teorico completamente specificato per una multinomiale.
- Test di omogeneità (uguaglianza) tra più multinomiali.

Ancora sull’X^2

- La statistica

$$X^2 = \sum_i \frac{(O_i - A_i)^2}{A_i}$$

risulta utile per confrontare

  - un insieme di *frequenze osservate* O_i, $i = 1, \ldots, k$,
  - con delle *frequenze attese*, A_i, $i = 1, \ldots, k$ calcolate ipotizzando un particolare modello per il fenomeno di interesse.

- Nella unità su *“streptococchi e tonsille”*[1] abbiamo utilizzato X^2 come statistica test per verificare l’ipotesi di indipendenza tra due variabili.
- In questa unità, accenniamo ad un paio di di altre situazioni in viene usata.

[1] unità E

Speriamo che sia femmina!

• In un indagine, tra le altre cose, sono state raccolte informazioni su 1659 coppie con esattamente tre figli biologici.
• La tabella mostra la distribuzione di queste coppie per numero di figlie femmine.

figlie femmine	0	1	2	3
coppie	248	643	580	188

• Le assunzioni di un possibile "modello" sono
(i) il genere[2] di un nato è indipendente dal genere di altri nati siano essi figli della stessa coppia o no;
(ii) la probabilità di nascere femmina è $1/2$ per i figli di tutte le coppie e indipendentemente dall'ordine di nascita[3]
• Indichiamo con $y = (O_0, O_1, O_2, O_3) = (248, 643, 580, 188)$ il vettore delle frequenze osservate. Se sono vere le ipotesi (i)-(ii) allora y è una determinazione di una variabile casuale multinomiale con numero di prove uguale a $n = 1659$ e probabilità di "cadere" nelle varie celle pari a (p_0, p_1, p_2, p_3) dove[4]

$$\begin{aligned} p_i &= P((\text{numero di femmine}) = i) = \\ &= \binom{3}{i}\left(\frac{1}{2}\right)^i\left(1-\frac{1}{2}\right)^{3-i} = \\ &= \binom{3}{i}\left(\frac{1}{2}\right)^3 = \frac{1}{8}\binom{3}{i} \quad i = 0,1,2,3. \end{aligned}$$

• Con qualche semplice calcolo

$$p_0 = 0{,}125 \quad p_1 = 0{,}375 \quad p_2 = 0{,}375 \quad p_3 = 0{,}125.$$

[2] femmina/maschio.
[3] ovvero, per il primo nato, per il secondo,...
[4] infatti nelle ipotesi del modello il numero di figlie femmine è una binomiale con numero di prove par a tre e probabilità di successo uguale a $1/2$.

• Il numero di coppie con i figlie femmine che ci saremmo aspettati, sulla base modello, di osservare è

$$A_i = np_i = (\text{numero coppie}) \times (\text{probabilità } i \text{ figlie femmine})$$

• Le frequenze osservate e le frequenze attese non sono uguali.

figlie femmine (i)	0	1	2	3
frequenze osservate (O_i)	248	643	580	188
frequenze attese (A_i)	207,375	622,125	622,125	207,375

• E' chiaro però che una parte delle differenze è dovuta al caso, ovvero, al fatto che stiamo considerando quelle 1659 coppie e non altre.
• E' quindi spontaneo domandarsi se

la differenza tra frequenze osservate e frequenze attese potrebbe essere tutta dovuta al caso?

• La domanda può essere formalizzata come un problema di verifica di ipotesi:

$H_0 =$ la distribuzione di y è Multinomiale$(1659, (p_0, p_1, p_2, p_3))$

verso

$$H_1 : H_0 \text{ è falsa;}$$

che equivale a

$$H_0 : \text{ il modello formulato è vero}$$

verso

$$H_1 : \text{ il modello formulato è falso.}$$

Vogliamo cioè valutare la *bontà dell'adattamento* del modello ai dati.

• La statistica

$$X^2 = \sum_{i=0}^{3} \frac{(O_i - A_i)^2}{A_i} = \frac{(248 - 207{,}375)^2}{207{,}75} + \cdot + \frac{(188 - 207{,}375)^2}{207{,}375} \approx 13{,}32$$

misura la distanza tra le frequenze osservate e quelle attese, ovvero, tra quello che *conosciamo del mondo* (i dati) e il *modello*.
• Ovviamente più X^2 è grande più i dati "mettono in crisi" il modello.
• Se H_0 è vera[5], allora X^2 converge in distribuzione ad una variabile casuale χ^2 con $k - 1$ gradi di libertà dove k indica il "numero delle frequenze", ovvero quello delle "celle" della distribuzione multinomiale, ovvero, nel caso in esame 4.
• Il valore di X^2 deve quindi essere confrontato con i valori attesi da questa distribuzione.
• Un'occhiata ad una tavola dei quantili mostra che il valore di X^2 nel caso in esame è maggiore del quantile 0,99 di un χ^2 con tre gradi di libertà. Il livello di significatività osservato è quindi minore di 0,01. I dati sono altamente significativi contro H_0: il modello non sembrerebbe quindi adeguato a spiegare la realtà.

Esercizio (e spiegazione). Si ritiene che la proporzione di bimbe tra i nati sia, in tutto il mondo, pari al 48,6% ovvero, ogni 100 nuove bimbe nascono mediamente circa 106 bimbi maschi. Verificare che il modello formulato sembra fornire una più che adeguata descrizione dei dati osservati quando si utilizza questa probabilità di nascere femmina (e non 0,5, come precedentemente fatto).

[5] ovvero se y è realmente una multinomiale con le probabilità di "cadere" nelle varie classi suggerite dal modello e quindi tra l'altro completamente specificate

Demenza senile

• Per capire se una particolare alterazione neuronale può essere associata con la presenza di demenza senile[6] l'alterazione stessa è stata valutata su 100 persone affetti da demenza e su 100 persone non affette[7].
• Tutti gli individui coinvolti nello studio hanno almeno 70 anni e possono essere pensati come scelti a caso, rispettivamente, nelle due popolazioni:

A. persone con almeno 70 con demenza senile conclamata;

B. persone con almeno 70 senza segni di demenza senile.

• I dati racolti sono riassunti nella seguente tabella.

	alterazione neuronale			
	assente	leggera	importante	totale
demenza	2	41	57	100
non demenza	11	57	32	100
totale	13	98	89	200

• La domanda a cui vogliamo tentare di dare una risposta con questi dati è ovviamente

> *la presenza della demenza è legata all'alterazione neuronale considerata? ovvero, tra presenza della demenza e alterazione esiste una qualche forma di dipendenza?*

• La struttura dei dati e la domanda che ci poniamo è uguale a quanto incontrato quando abbiamo parlato di *"streptococchi e tonsille"*[8].
• Il dati però sono stati raccolti seguendo un disegno campionario differente.

[6] che costituisce una precisa patologia.
[7] o quantomeno senza segni clinici di demenza senile.
[8] unità E

• Nel caso degli *"streptococchi e tonsille"* la tabella di contingenza era stata ottenuta

(i) estraendo n individui dalla popolazione di riferimento

(ii) e poi rilevando su ciascun individuo le due variabili presenza di streptococco e stato delle tonsille.

Il risultato è che le frequenze congiunte della tabella possono essere, tutte insieme, pensate come una determinazione di una variabile casuale multinomiale.

• Nel caso in esame viceversa le unità statistiche sono state estratte separatamente da due differenti popolazioni: quella degli anziani con e senza demenza. Poi su ciascun individuo è stata rilevata la variabile alterazione neuronale.

• In questo caso quindi possiamo pensare che

– la prima riga della tabella sia la determinazione di una variabile casuale multinomiale che descrive il comportamento dell'alterazione neuronale nella popolazione degli anziani *con* demenza e

– la seconda riga sia la determinazione di un altra variabile casuale multinomiale che descrive il comportamento dell'alterazione neuronale nella popolazione degli anziani *senza* demenza.

• Potremmo dire che nel caso *"streptococco e tonsille"* i ricercatori avevano utilizzato "una sola urna" mentre in questo caso per ottenere i dati sono state utilizzate "due differenti urne".

• Questo fatto emerge anche dal fatto che, nel caso che stiamo considerando, la *distribuzione marginale* della variabile presenza di demenza non è il risultato di un esperimento causale ma è stata fissata a priori dai ricercatori prima dell'esperimento.

Viceversa nel caso *"streptococco e tonsille"* nessuna marginale era nota a priori.

• In una situazione del tipo considerato (campionamento separato da più popolazioni) quello che vogliamo verificare è se le "multinomiali coinvolte" sono tra di loro *omogenee* ovvero assegnano la stessa probabilità alle varie modalità.

• Sembra sensato anche in questo caso calcolare la tabella delle frequenze attese in maniera uguale a quanto fatto nel caso di indipendenza.

• Ad esempio, se non ci fossero differenze tra le distribuzioni della variabile di interesse (alterazione neuronale) nelle due popolazioni (persone con e senza demenza) quante persone con "alterazione importanti" ci aspetteremmo di osservare tra le persone con demenza?

Visto che le persone con "alterazione importante" sono 89 su un totale di 200 individui e che le persone con demenza sono 100 sembra sensato rispondere che, se non ci sono differenze tra le due popolazioni, circa

$$\frac{89}{200}100 = 44{,}5$$

persone con demenza dovrebbero presentare una "alterazione grave".

• La tabella mostra le frequenze attese per tutte le celle

	alterazione neuronale			
	assente	leggera	importante	totale
demenza	6,5	49	44,5	100
non demenza	6,5	49	44,5	100
totale	13	98	89	200

• Per misurare la "distanza" tra le frequenze osservate e le frequenze attese possiamo, al solito, usare X^2.

$$\begin{aligned} X^2 &= \frac{(2-6{,}5)^2}{6{,}5} + \frac{(41-49)^2}{49} + \frac{(57-44{,}5)^2}{44{,}5} + \\ &\quad + \frac{(11-6{,}5)^2}{6{,}5} + \frac{(57-49)^2}{49} + \frac{(32-44{,}5)^2}{44{,}5} = \\ &\approx 15{,}86 \end{aligned}$$

• Ovviamente, più X^2 è grande più i dati sono lontani da quanto ci aspettiamo nell'ipotesi di omogeneità.
• Nonostante il disegno campionario sia differente da quello considerato nel caso "streptococco e tonsille" è possibile dimostrare che la distribuzione asintotica della distribuzione di X^2 rimane, almeno sotto l'ipotesi nulla, la stessa[9].
• Quindi, per capire che cosa i dati ci raccontano sulla omegeità delle varie righe, possiamo confrontare il valore calcolato di X^2 con i valori plausibili per una distribuzione χ^2 con gradi di libertà uguali a

$$\left[\begin{pmatrix}\text{numero righe}\\ \text{tabella}\end{pmatrix} - 1\right] \times \left[\begin{pmatrix}\text{numero colonne}\\ \text{tabella}\end{pmatrix} - 1\right]$$

Ovvero, nel nostro caso $(2-1)\times(3-1) = 2$.
• 15,86 è più grande del quantile 0,999 di un $\chi^2(2)$. Quindi il livello di significatività osservato è in questo caso minore di 0,001: i dati ci stanno suggerendo che esistono delle differenze tra le due popolazioni e quindi che l'alterazione neuronale considerata è associata alla presenza o meno di demenza.

[9] rimane anche invariata la "regola a spanne" per utilizzarla: le frequenze attese devono tutte essere maggiori di 5.

Unità G

Dove facciamo conoscenza con uno statistico birraio

- Test t di Student ad un campione.
- Intervalli di confidenza per la media di una normale quando la varianza non è nota.
- *Normal probability plot*
- Test di Shapiro-Wilk

Un esperimento su un sonnifero

- Per verificare l'efficacia di una nuova sostanza "sonnifera"[1], su dieci individui, è stata misurata la variabile, denominata `ore di extra sonno`, definita come

$$\begin{pmatrix}\text{ore di sonno in una}\\ \text{notte in cui viene}\\ \text{somministrato il}\\ \text{sonnifero}\end{pmatrix} - \begin{pmatrix}\text{ore di sonno in una}\\ \text{notte in cui viene}\\ \text{somministrato un}\\ \text{placebo}\end{pmatrix}$$

- Le dieci osservazioni ottenute sono

$$0{,}7 \quad -1{,}6 \quad -0{,}2 \quad -1{,}2 \quad -0{,}1 \quad 3{,}4 \quad 3{,}7 \quad 0{,}8 \quad 0{,}0 \quad 2{,}0$$

- La media delle dieci misure disponibili per questa variabile è 0,75.
- Quindi, se restringiamo l'attenzione ai dieci individui considerati e alle notti in cui è stato condotto l'esperimento, il sonnifero ha avuto l'effetto atteso, ovvero gli individui hanno mediamente dormito di più[2].
- E' però spontaneo porsi la domanda:

  sulla base di questi risultati ci aspettiamo che la sostanza abbia effetto *in generale*, ovvero anche su altri individui a cui potremmo somministrarla?

[1] a cui ho già accennato nei lucidi di Descrittiva

[2] anzi, parecchio di più (circa 45 minuti) visto che gli individui a cui era stato somministrato il sonnifero non avevano particolari problemi di insonnia.

Un possibile modello di riferimento

• Consideriamo l'insieme di tutti gli individui a cui potremmo somministrare il farmaco. Si tratta ovviamente di un insieme molto grande.
• Le `ore di extra sonno` sono il risultato di un miriado di fattori (l'attitudine al sonno degli individui, la resistenza al farmaco, che cosa gli individui possono avere mangiato a cena, se una zanzara li ha punti durante la notte,...). Ora se tutti questi fattori si "compongono" in maniera additiva possiamo pensare sulla base del teorema del limite centrale che la distribuzione delle `ore di extra sonno` nella popolazione possa essere ben approssimata da una distribuzione normale di appropriata media e varianza, diciamo μ e σ^2.
• Supponiamo inoltre che gli individui scelti per l'esperimento non abbiano caratteristiche particolari e quindi siano assimilabili ad individui *estratti casualmente dalla popolazione*. Ed anche, come del resto era effettivamente accaduto, che siano stati tenuti separati durante l'esperimento in maniera tale che non si siano "condizionati" a vicenda.
• Allora, se tutto questo è vero, possiamo vedere i dati osservati, indichiamoli al solito con $y_1, \ldots, y_{10}$, come delle determinazioni indipendenti ed identicamente distribuiti di una $N(\mu, \sigma^2)$.

Due precisazioni

(i) In realtà la frase "tutti gli individui a cui potremmo somministrare il farmaco" è eccessivamente generica.
I risultati possono essere estesi propriamente solamente ad individui con le stesse caratteristiche di quelli che fanno parte del campione. Ad esempio se il campione fosse costituito solo da "donne sopra i 50 anni" l'insieme di queste donne costituirebbe la nostra *popolazione di riferimento*.
(ii) Il modello suggerito per interpretare i dati è simile a quello considerato nell'unità A. La differenza è che in quell'unità σ^2 era noto (od almeno assunto tale). Qui è un parametro ignoto.

Normal probability plot e test di Shapiro-Wilk

Domanda. E' plausibile il modello suggerito?
Risposta. Beh, quando gli statistici non sanno qualcosa cercano di interrogare i dati.
Come possiamo farlo? Per farlo useremo un procedimento grafico (*normal probability plot*) e uno analitico (test di normalità di Shapiro-Wilk).
Statistica ordinata. Siano $y_1, \ldots, y_n$ n osservazioni su di una variabile numerica. Una permutazione $y_{(1)}, \ldots, y_{(n)}$ di $y_1, \ldots, y_n$ 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.
Statistica ordinata e quantili. $y_{(j)}$ può essere visto[3] come una stima del quantile-p della distribuzione che ha generato i dati con $p \approx j/n$
Infatti, esattamente o approssimativamente, la frazione di osservazioni minori o uguali di $y_{(j)}$ è j/n.
Quantili di una distribuzione normale. E' facile verificare[4] che

$$\text{se } Y \sim N(\mu, \sigma^2) \text{ allora } (\text{quantile-p di } Y) = \mu + \sigma z_p$$

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

[3] 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).
[4] [Probalità 8]

Normal probability plot. Consideriamo un grafico ottenuto disegnando su di un piano cartesiano i punti

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

Il grafico (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. Per questo motivo è un esempio dei cosidetti grafici "quantile verso quantile".
Per quanto riguarda l'interpretazione si osservi che:
· se i dati sono normali ci aspettiamo di osservare un andamento, almeno approssimativamente, lineare; infatti, per quanto detto, ci aspettiamo, almeno se n non è piccolo[5], che

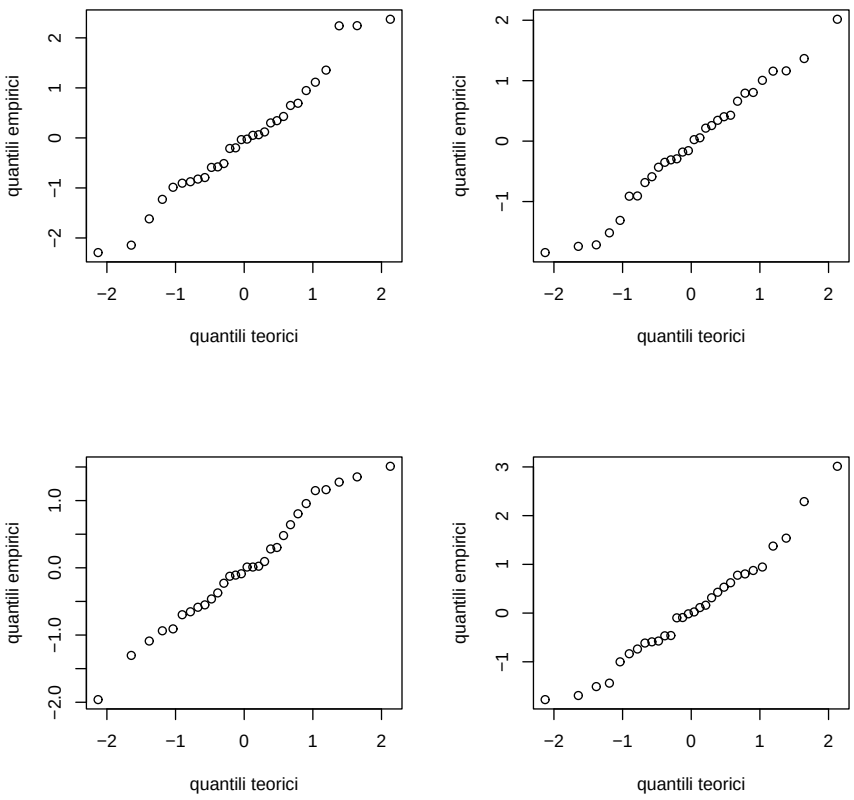
$$y_{(j)} \approx \mu + \sigma z_{\frac{j-0,5}{n}};$$

· viceversa se il grafico suggerisce un andamento non 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;
La *linearità* del grafico può quindi essere utilizzata per dare un giudizio sulla normalità della distribuzione che li ha generati.

Domanda: Perchè usiamo $z_{(j-0,5)/n}$ e non $z_{j/n}$?
Risposta: Perchè $z_1 = z_{n/n} = +\infty$ e quindi dovremmo disegnare l'osservazione più grande ad infinito.

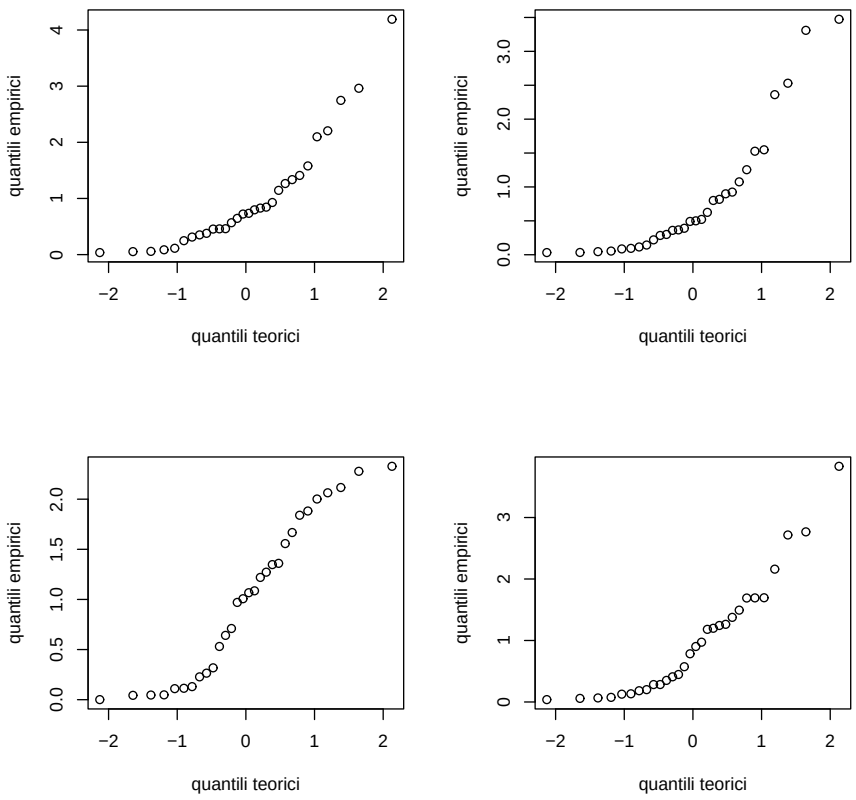
[5] ovvero, almeno quando i dati permettono una stima dei decorosa dei quantili

Esempio: campioni generati da una distribuzione normale



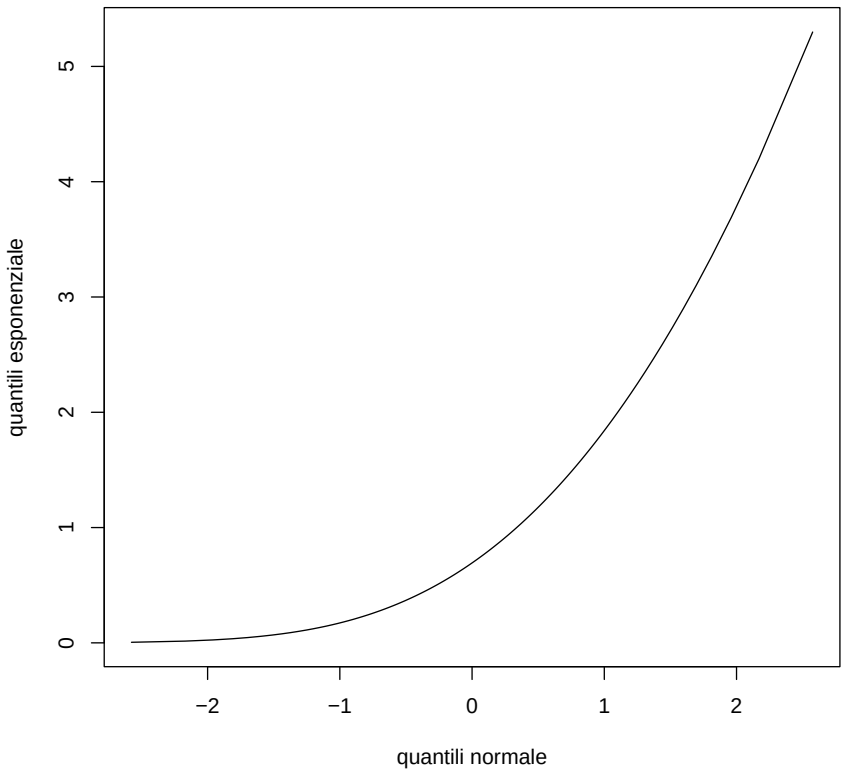
I grafici sono basati su 4 campioni di numerosità pari a 30 simulati da una distribuzione normale standard. In questo caso le considerazioni precedenti ci suggeriscono che i punti dovrebbero, come infatti accade, stare intorno alla bisettrice del 1° e 3° quadrante.
Si osservi inoltre come le maggiori deviazioni da una ipotetica retta si osservano agli estremi. Questa è una conseguenza della maggiore variabilità di $y_{(j)}$ quando j è "piccolo" (vicino a 1) e "grande" (vicino a n).

Esempio: campioni generati da una distribuzione esponenziale

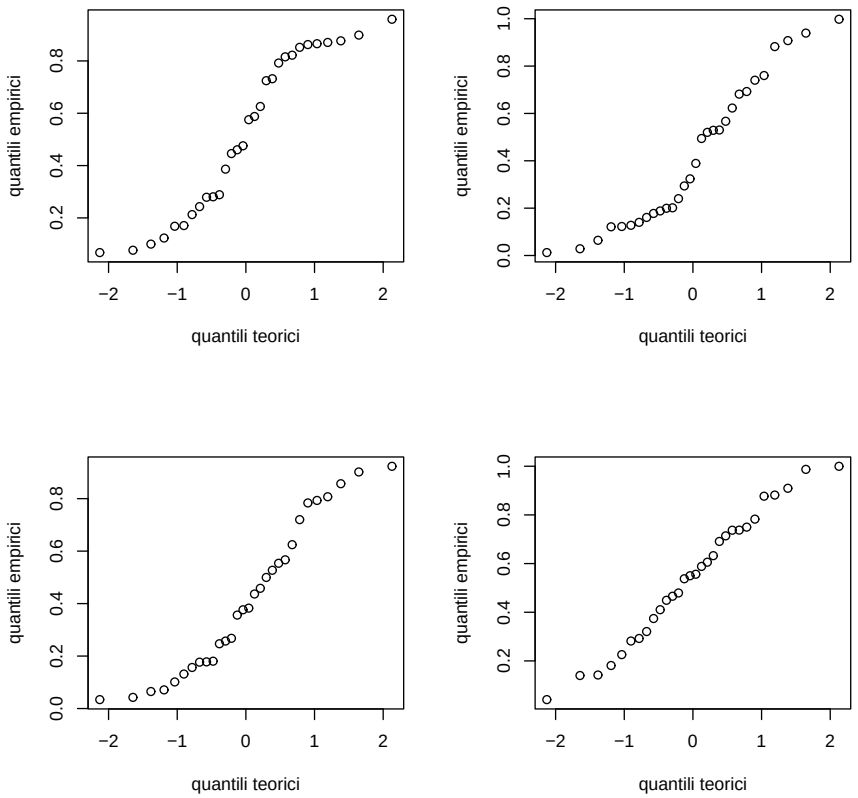


I grafici sono basati su 4 campioni di numerosità pari a 30 simulati da una distribuzione esponenziale di media 1. Si osservi che il quantile-p di questa distribuzione vale $-\log(1-p)$. Quindi, i punti li aspettiamo in questa caso intorno alla curva $(z_p, -\log(1-p))$, $0 < p < 1$, che è disegnata a pagina 121.

Quantili di una distribuzione esponenziale di media 1 verso quelli di una normale standard.

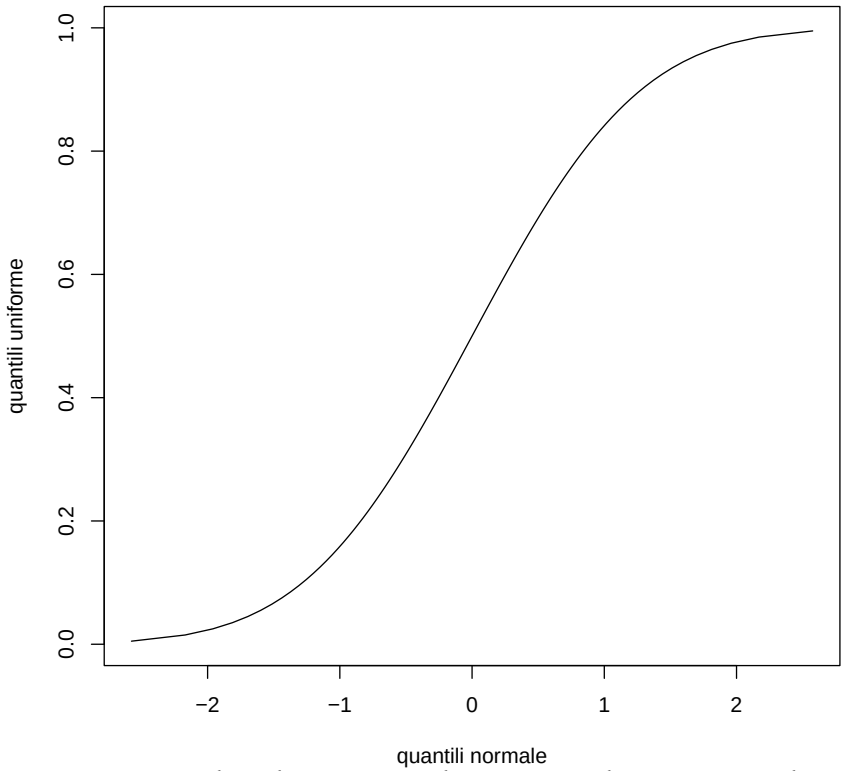


Esempio: campioni generati da una distribuzione uniforme



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), $0 < p < 1$, che è disegnata a pagina 123.

Quantili di una distribuzione uniforme tra 0 e 1 verso quelli di una normale standard.



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

test di Shapiro-Wilk Sul *normal probability plot* è basato uno dei molti test di normalità esistenti, ovvero, uno dei molti test che sono stati proposti per verificare il sistema d'ipotesi

$$\begin{cases} H_0 : \text{la distribuzione dei dati è normale} \\ H_1 : \text{la distribuzione dei dati non è normale} \end{cases}.$$

Il test, detto di Shapiro-Wilk dal nome degli autori, si basa su di una statistica che, nella sostanza, è il coefficiente di correlazione tra i punti disegnati nel *normal probability plot*.

Breve dialogo

Studente: cosa vuol dire "nella sostanza"? E' o non e' il coefficiente di correlazione?

Professore: la statistica test è una versione "appena appena" aggiustata del coefficiente di correlazione; l'aggiustamento apportato cerca di controbilanciare la correlazione che in ogni caso ci aspettiamo di trovare visto che i punti del *normal probability plot* sono in ogni caso posti su di una curva non decrescente.

S: ma come posso calcolare la statistica test?

P: solo un masochista la calcola a mano! per il calcolo è necessario un calcolatore con una funzione appropriata (in `R` la funzione si chiama `shapiro.test`); per questo motivo, vista la natura introduttiva del corso, non ti annoio con la formula precisa.

S: resta però inteso che rifiuto per valori troppo piccoli (lontani da uno) mentre più la statistica test è vicina ad uno più la interpreto come "i dati sostengono H_0"?

P: certo.

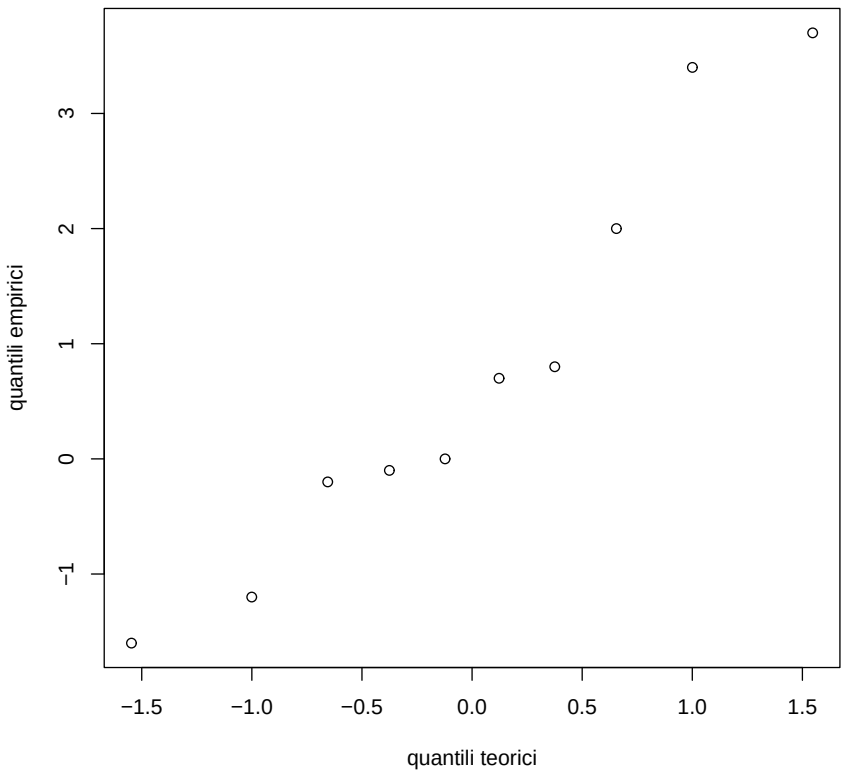
S: ma quanto lontana deve essere da uno questa benedetta statistica perchè io possa iniziare a dubitare di H_0?

P: se il programma che usi è ragionevole calcolerà per te il livello di significatività osservato; oramai hai imparato ad interpretarlo; quindi...

S: almeno un esempio svolto posso vederlo?

P: beh, se tu girassi pagina al posto di fare sempre domande!

Un esperimento su di un sonnifero



Il grafico mostra il *normal probability plot* dei dati sulle ore di extrasonno. La linearità sembra buona.
Il valore della statistica su cui è basato il test di Shapiro-Wilk vale in questo caso 0,926, il relativo livello di significatività osservato 0,408. Questo valore è elevato per dubitare della normalità dei dati.

Stima dei parametri del modello

• Viste anche le verifiche effettuate, proviamo a rispondere alla domanda sull'efficacia del sonnifero assumendo il modello suggerito prima, ovvero, ipotizzando che le osservazioni sulle ore di *extrasonno* siano determinazioni indipendenti di una variabile casuale $N(\mu, \sigma^2)$.
• La distribuzione del fenomeno considerato è nota con l'eccezione dei due parametri μ e σ^2.
• Sembra quindi ragionevole "iniziare" cercando di stimare questi due parametri dai dati.
• Gli stimatori più usati per μ e σ^2 sono rispettivamente la media e la varianza campionaria ovvero

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \approx 0{,}75$$

e

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \overline{y})^2 \approx 3{,}20$$

dove, al solito, n indica il numero delle osservazioni (per l'esperimento considerato $n = 10$).

Un problema di verifica d'ipotesi

• Un sistema d'ipotesi interessante in questo caso è

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

con $\mu_0 = 0$.
Accettare H_0, infatti, equivale a dire che, in media, prendendo il farmaco non si dorme né di più né di meno.
• Per verificare un sistema d'ipotesi analogo nell'unità B avevamo utilizzato come statistica test

$$z = \frac{\sqrt{n}(\overline{y} - \mu_0)}{\sigma}.$$

Però in questa unità noi non conosciamo σ. Quindi z non è direttamente utilizzabile.
• Dall'altra parte, poichè abbiamo a disposizione una stima di σ, una statistica test analoga a z è

$$t_{oss} = \frac{\sqrt{n}(\overline{y} - \mu_0)}{s}.$$

L'oss che abbiamo posto a denominatore è l'abbreviazione di "osservato".
• Se H_0 (H_1) è vera ci aspettiamo che t_{oss} assuma valori intorno allo (lontani dallo) zero.

Quanto deve essere lontana da zero t_{oss} per concludere che H_0 è implausibile?

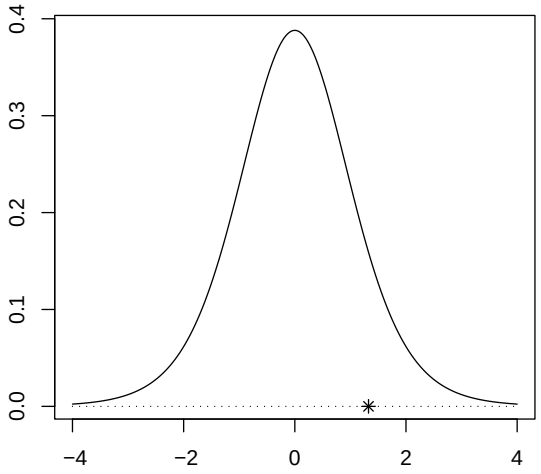
• Per rispondere alla domanda avremmo bisogno di sapere qual'è la distribuzione di t_{oss} quando H_0 è vera. Infatti, questa distribuzione ci "racconta" quali sono i valori di t_{oss} che ci aspettiamo sotto l'ipotesi nulla.
• Sappiamo che la distribuzione di z è normale. Potremmo perciò pensare di approssimare la distribuzione di t con quella di una $N(0, 1)$.
• La sostituzione del vero σ con s non può però essere "indolore" 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,...) mostrare che[6]

$$t_{oss} \sim t \text{ di Student con } n - 1 \text{ gradi di libertà.}$$

Il test che stiamo descrivendo viene usualmente chiamato test t a un campione.

[6] [Probalità 14] e [Probalità 43]

Analisi grafica del risultato



- Il valore di t_{oss} calcolato sui dati del primo sonnifero è 1,33. Nel grafico il valore è indicato dall'asterisco sull'asse delle ascisse.
- La curva mostra la densità di una t di Student con 9 gradi di libertà.
- Il valore osservato sembra "compatibile" con la distribuzione disegnata.
- Quindi, non abbiamo elementi nei dati per rifiutare H_0, ovvero, non possiamo affermare sulla base dei risultati sperimentali che il nuovo sonnifero ha una qualche effetto sulla media.

Analisi mediante il livello di significatività osservato

• "Lontano da H_0" equivale a "lontano da zero in ambedue le direzioni". Quindi, nel caso del sonnifero,

$$\begin{pmatrix}\text{livello di}\\ \text{significatività}\\ \text{osservato}\end{pmatrix} = P(|t \text{ con } 9 \text{ gradi di libertà}| \geq 1{,}33).$$

che, per la simmetria della t di Student, possiamo anche calcolare come

$$\begin{pmatrix}\text{livello di}\\ \text{significatività}\\ \text{osservato}\end{pmatrix} = 2 \times P(t \text{ con } 9 \text{ gradi di libertà} \geq 1{,}33).$$

• Disponendo solo di una tabella dei percentili, del tipo ad esempio contenuto in "Formule e Tavole", possiamo, come fatto nell'unità precedente, determinare un intervallo che lo contiene.

• In particolare, dalla tabella vediamo che 1,33 è compreso tra il 75% e il 90% percentile di una t con 9 gradi di libertà. Quindi,

$$0{,}10 < P(t \text{ con } 9 \text{ gradi di libertà} \geq 1{,}33) < 0{,}25.$$

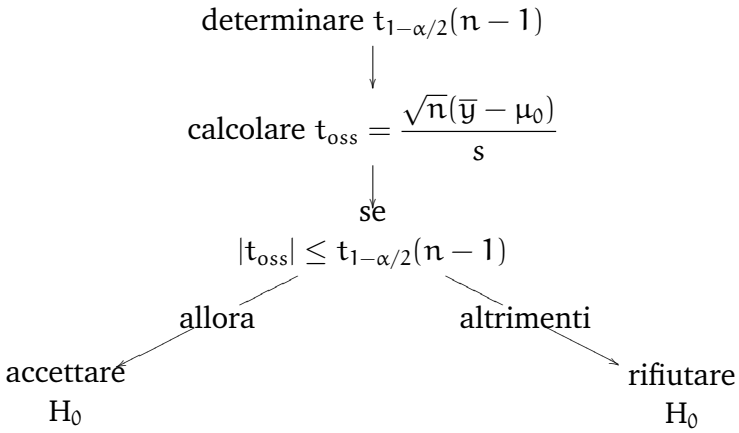
Ma allora

$$0{,}2 < \begin{pmatrix}\text{livello di}\\ \text{significatività}\\ \text{osservato}\end{pmatrix} < 0{,}5$$

• Per quello che riguarda l'interpretazione la prima disuguaglianza è la più importante. Ci racconta infatti che se il sonnifero non ha un effetto sulla media delle `ore di extra sonno` allora noi ci aspetteremmo valori "più lontani da H_0 di quanto osservato" con una frequenza superiore al 20% (ovvero, più di una volta ogni 5 replicazioni dell'esperimento). Questo, vuol dire che il valore osservato di t_{oss} non è "strano" quando H_0 è vera.

• In conclusione, i dati ci dicono che non abbiamo elementi per rifiutare l'ipotesi nulla.

Una regola del tipo accetto/rifiuto

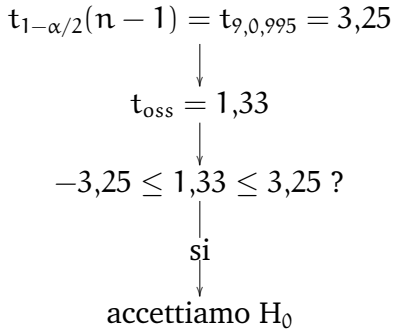


Nell'albero $t_p(g)$ indica il percentile p-simo di una t di Student con g gradi di libertà. E' facile far vedere che l'albero fornisce una regola per accettare/rifiutare l'ipotesi nulla che garantisce che

$$P(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera}) = 1 - \alpha$$

Con i dati

Supponiamo di porre $\alpha = 0{,}01$. Allora



Un intervallo di confidenza

· Un intervallo di confidenza per μ può essere determinato, dai risultati precedenti utilizzando lo stesso ragionamento seguito nell'unità B.
· Infatti quello che sappiamo è che se μ è il vero valore della media allora

$$P(-t_{1-\alpha/2}(n-1) \leq \sqrt{n}(\overline{y}-\mu)/s \leq t_{1-\alpha/2}(n-1)) = 1-\alpha.$$

Ma allora, scrivendo le due disuguglianze in termini di μ, troviamo che

$$P(\overline{y} - st_{1-\alpha/2}(n-1)/\sqrt{n} \leq \mu \leq \overline{y} + st_{1-\alpha/2}(n-1)/\sqrt{n}) = 1-\alpha$$

ovvero che

$$\left[\overline{y} - \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n}}\,;\, \overline{y} - \frac{st_{1-\alpha/2}(n-1)}{\sqrt{n}}\right]$$

è un intervallo di confidenza di livello $1-\alpha$ per la media.
· *Applicazione ai dati.* Supponiamo, ad esempio, di voler un intervallo che contenga con probabilità 90% il vero valore di μ. Allora, $t_{1-\alpha/2}(n-1) = t_{0,95}(9) = 1{,}83$. Ricordando che $\overline{y} = 0{,}75$ e $s^2 \approx 3{,}2$ e quindi che $s \approx \sqrt{3{,}2} \approx 1{,}79$, la semi-ampiezza dell'intervallo richiesto è

$$1{,}04 = \frac{1{,}79 \times 1{,}83}{\sqrt{10}}$$

mentre l'intervallo stesso è

$$[0{,}75 - 1{,}04 \;\;;\;\; 0{,}75 + 1{,}04] = [-0{,}29 \;\;;\;\; 1{,}79]$$

· Si osservi che l'intervallo include lo zero. Questo era atteso visto che avevamo visto, con il test discusso precedentemente, che un valore nullo per μ era plausibile sulla base dei dati disponibili.

Esercizio

Per una variante del sonnifero considerato si sono ottenute le seguenti ore di *extra-sonno*:

$$1{,}9 \;\; 0{,}8 \;\; 1{,}1 \;\; 0{,}1 \;\; -0{,}1 \;\; 4{,}4 \;\; 5{,}5 \;\; 1{,}6 \;\; 4{,}6 \;\; 3{,}4$$

Discutere l'efficacia della variante.

Unità H

Cuculi, scriccioli, pettirossi e Darwin

Test t a due campioni.

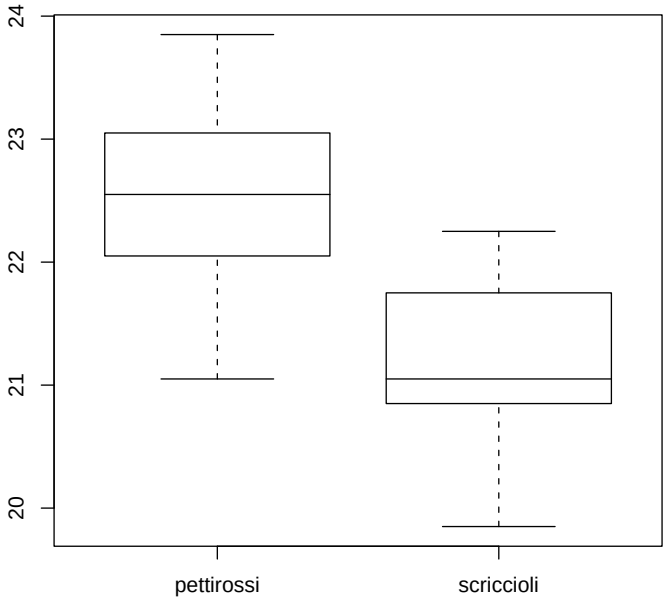
Il problema e i dati

• E' noto che i cuculi depongono le proprie uove nei nidi di altri uccelli a cui viene poi lasciato il compito della cova.
• E' possibile osservare una certa associazione tra territorio e uccello scelto come "ospite", ovvero, in certi territori i cuculi sembrano preferire una specie di uccello come "ospite", in altri un altra.
• Sulla base della teoria della selezione naturale, ci si aspetta quindi una qualche forma di adattamento dell'uovo del cuculo a quella dell'uccello "ospite".
Infatti, la probabilità di un uovo di essere covato (che viste le abitudini del cuculo influenza non poco la sopravvivenza del suo patrimonio genetico) dovrebbe essere tanto più alta quanta più le uova "abusive" sono simili a quelle dell'uccello "ospite".
• Per verificare questa idea sono state misurate le lunghezze (in mm) di alcune uova di cuculo trovate in nidi di pettirossi e di scriccioli in due territori, uno in cui i cuculi "preferiscono" i pettirossi, l'altro in cui "preferiscono" gli scriccioli.

I dati

Uova deposte in nidi di pettirosso:
21,05 21,85 22,05 22,05 22,05 22,25 22,45 22,45
22,65 23,05 23,05 23,05 23,05 23,05 23,25 23,85
Uova deposte in nidi di scricciolo:
19,85 20,05 20,25 20,85 20,85 20,85 21,05 21,05
21,05 21,25 21,45 22,05 22,05 22,05 22,25

ospite	media	mediana	sqm	mad
pettirosso	22,57	22,55	0,68	0,5
scricciolo	21,13	21,05	0,74	0,4



Primi commenti

- Gli scriccioli sono scriccioli e quindi le loro uova sono più piccole di quelle dei pettirossi! La differenza che ci si aspetta a priori tra i due gruppi ha quindi a che fare con la posizione della distribuzione.
- A livello puramente descrittivo, ovvero senza tenere conto di eventuali errori dovuti al fatto che conosciamo solamente le lunghezze di 31 uova (16 in un gruppo, 15 nell'altro), gli indici di posizione (media e mediana) e il diagramma a scatola con baffi suggeriscono che questo adeguamento all'"ospite" sia avvenuto.
- La breve analisi fatta e in particolare il grafico suggeriscono inoltre che la dispersione dei due insiemi di dati è praticamente la stessa.
- Una domanda interessante che ci possiamo fare è:

  "La differenza tra le lunghezze medie che abbiamo osservato sui dati disponibili può essere attribuita al caso? Ovvero, potrebbe essere dovuta al fatto che abbiamo considerato sola un piccolo numero di uova deposte? Oppure ci aspettiamo che valga più *in generale*?"

- Una possibile formulazione dell'ultima domanda è il seguente:

(a) La popolazione di riferimento è divisa in due gruppi Al primo (secondo) gruppo appartengono tutte le uova che i cuculi delle zone considerate depongono nei nidi di pettirosso (scricciolo).
(b) Indichiamo con μ e η la media delle lunghezze delle uova dei due gruppi. Utilizzando i dati disponibili siamo interessati a verificare il sistema di ipotesi

$$\begin{cases} H_0 : \mu = \eta \\ H_1 : \mu \neq \eta \end{cases}$$

Test t a due campioni: la situazione di riferimento

Una semplice procedura è disponibile nel caso in cui si accettino (o meglio, si verifichi con i dati che sono accettabili) le seguenti ipotesi:

1. La distribuzione della lunghezza delle uova in ambedue le popolazioni è normale.
2. Le due normali hanno una media μ, l'altra media η. La varianza è però la stessa, diciamo σ^2.
3. Le uova per cui abbiamo la misura delle lunghezze (i nostri dati) possono essere pensate come estratte a caso in maniera indipendente da una o dall'altra delle due popolazioni.

Ovvero, se, indicate con
– $y_1, \ldots, y_n$ le lunghezze delle uova trovate in nidi di pettirossi e
– $x_1, \ldots, x_m$ le lunghezze delle uova trovate in nidi di scricciolo
allora

$y_1, \ldots, y_n$ sono determinazioni i.i.d. distribuite come una $N(\mu, \sigma^2)$

$x_1, \ldots, x_m$ sono determinazioni i.i.d. distribuite come una $N(\eta, \sigma^2)$

e

le "y" e le "x" sono indipendenti tra di loro.

Test t a due campioni: la statistica test e la sua distribuzione

• La statistica test usualmente considerata per verificare l'ipotesi che le due medie sono uguali è[1]

$$t_{oss} = \frac{\overline{y} - \overline{x}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

dove $\overline{y}$ e $\overline{x}$ sone le medie dei due gruppi di osservazioni mentre

$$s^2 = \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(y_i - \overline{y})^2 + \sum_{i=1}^{m}(x_i - \overline{x})^2\right]$$

può essere vista come una stima di σ^2 basata su *tutti* i dati.
• t_{oss} è una versione standardizata della differenza tra le medie nei due gruppi.
• Il denominatore infatti è una stima di

$$\mathrm{var}\{\overline{y} - \overline{x}\} = \mathrm{var}\{\overline{y}\} + \mathrm{var}\{\overline{x}\} = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right).$$

Nel primo passaggio abbiamo usato[2] l'indipendenza tra le "y" e le "x"; nel secondo quello che sappiamo sulla varianza di una media campionaria di osservazioni i.i.d.
• Ovviamente, più è grande, in valore assoluto, il valore di t_{oss} più i dati ci suggeriscono di "dubitare" dell'ipotesi nulla.
• E' possibile far vedere che se H_0 è vera, ovvero se realmente $\mu = \eta$, allora t_{oss} si distribuisce come una t di Student con $n + m - 2$ gradi di libertà[3].
• I valore della statistica test può quindi essere analizzato in maniera analoga a quanto fatto nell'unità precedente.

[1] t_{oss} e s^2 indicano quantità diverse rispetto al test t a un campione.
[2] [Probalità 36].
[3] [Probalità 45]

• Si osservi che s^2 è facilmente calcolabile dalle varianza campionarie delle "y" e delle "x". Infatti , posto

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

e definito in maniera analoga s_x^2, risulta

$$s^2 = \frac{1}{n+m-2}[(n-1)s_y^2 + (m-1)s_x^2]$$

ovvero

> "s^2 è una media ponderata di s_y^2 e s_y^2 con pesi proporzionali ai gradi di libertà"

Applicazione alle lunghezze delle uove di cuculo

• In questo caso, abbiamo[4]

$$n = 16 \quad \overline{y} \approx 22{,}47 \quad s_y^2 \approx 0{,}46$$
$$m = 15 \quad \overline{x} \approx 21{,}13 \quad s_x^2 \approx 0{,}55$$

Quindi,

$$s \approx \sqrt{(15 \times 0{,}46 + 14 \times 0{,}55)/29} \approx 0{,}71$$

e

$$t_{oss} = \frac{22{,}47 - 21{,}13}{0{,}71\sqrt{\frac{1}{16} + \frac{1}{15}}} \approx 5{,}64$$

• La distribuzione sotto H_0 è una t di Student con 29 gradi di libertà.
• Dalla tabella dei quantili della t nell'unità precedente, vediamo che il valore calcolato di t_{oss} è più grande di $t_{29,0,9995}$.
Quindi, ci aspettiamo di osservare un valore più lontano da zero (in ambedue le direzioni) meno di una volta ogni 1000 replicazioni dell'esperimento o, in altre parole, il livello di significatività osservato è $\leq 0{,}001$.
• Un livello così basso del livello di significatività osservato è usualmente considerato altamente significativo contro H_0.
• La conclusione è quindi che, sulla base dei dati, sembra poco plausibile che la differenza osservata sia puramente dovuta al caso. Ci aspettiamo, viceversa, che la differenza osservata tra le due medie campionarie sia una manifestazione di una reale differenza tra le due popolazioni.

[4] Si ricordi che "y" vuol dire "pettirossi" e "x" scriccioli.

Esercizio. La distribuzione di t_{oss} data prima è un caso particolare di un risultato generale che dice che, nelle ipotesi con cui stiamo lavorando,

$$\frac{\overline{y} - \overline{x} - (\mu - \eta)}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2} \qquad \text{(H.1)}$$

Utilizzando questo risultato, mostrare che

$$\left[\overline{y} - \overline{x} - t_{1-\alpha/2} \frac{s}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \; ; \; \overline{y} - \overline{x} - t_{1-\alpha/2} \frac{s}{\sqrt{\frac{1}{n} + \frac{1}{m}}} \right]$$

è un intervallo di confidenza (con grado di copertura pari a 1α) per la differenza tra le due medie (ovvero per $\delta = \mu - \eta$).

La vera ipotesi è però unilaterale!

- Un sistema d'ipotesi unidirezionale, del tipo[5]

$$\begin{cases} H_0 : \eta = \mu \\ H_1 : \eta < \mu \end{cases} \text{ o, anche, } \begin{cases} H_0 : \eta \geq \mu \\ H_1 : \eta < \mu \end{cases}$$

sembra più appropriato di quello bilaterale considerato fino ad ora.
- Infatti, l'ipotesi "sul mondo" che stiamo esplorando prevede che le uove deposte nei nidi di scricciolo siano più piccole (almeno mediamente) di quelle deposte nei nidi di pettirosso.
- La statistica

$$t_{oss} = \frac{\overline{y} - \overline{x}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

sembra ancora appropriata.
- Cambiano però i "valori attesi" sotto le due ipotesi.

	valori attesi per t_{oss} sotto	
	H_0	H_1
bilaterale	vicino a zero	lontano in una delle due direzioni (negativa o positiva) da zero
unilaterale	negativo o vicino a zero	maggiore di zero

- Quindi, nel caso del sistema di ipotesi unilaterale definito sopra, "lontano da H_0" vuole dire "valori positivi di t_{oss}" e perciò il livello di significatività osservato è

$$P(t_{29} \geq 5{,}64)$$

e, non, come nel caso bilaterale, $P(t_{29} \leq -5{,}64) + P(t_{29} \geq 5{,}64)$. Nel caso in esame, il livello di significatività osservato risulta quindi minore di 0,005.

[5] si ricordi che 1° gruppo, media μ, pettirossi; 2° gruppo, media η, scriccioli]

• Volendo un test di tipo *accetto/rifiuto* con un livello di significatività prefissato α possiamo o utilizzare il livello di significatività osservato[6] oppure, in maniera analoga, utilizzare una "regola" del tipo

$$\begin{cases} \text{accettare} & \text{se } t_{oss} \leq h \\ \text{rifiutare} & \text{se } t_{oss} > h \end{cases}$$

• Per ottenere una "regola" che ci garantisca che, per ogni μ e η con $\mu \leq \eta$, la probabilità di accettare H_0 è maggiore o al più uguale a $1-\alpha$ dobbiamo porre

$$h = t_{1-\alpha}(n+m-2).$$

• Si osservi l'"$1-\alpha$" e non il solito "$1-\alpha/2$".
• *Esempio.* Supponiamo di porre $\alpha = 0.1$. Allora, nella tabella dei percentili di una t di Student troviamo $t_{29,0,9} = 1{,}31$. Il valore osservato della statistica test (5,64) è maggiore di questo livello di soglia e quindi... continuamo a concludere a favore di Darwin.

Attenzione. Tutte le considerazioni (grande, piccolo, a favore di H_0, a favore di H_1,...) dipendono, quando si ha a che fare con ipotesi unilaterali, da come si formulano le ipotesi e da come si scrive la statistica!

[6] si veda il lucido 74

E se le varianze nei due gruppi non sono uguali?

• La breve analisi preliminare condotta (vedi lucido 137) suggerisce che la dispersione all'interno dei due gruppi è sostanzialmente la stessa.
• E' però interessante, magari anche solo per assicurarsi che l'assunzione non "pesa", essere in grado di confrontare le medie di due gruppi anche quando le varianze non sono tra di loro uguali.
• Una possibilità approssimata in questo caso è offerta dalla cosidetta correzione di Welch.
• La statistica test da usare è

$$t^*_{oss} = \frac{\overline{y} - \overline{x}}{\sqrt{\frac{s_y^2}{n} + \frac{s_x^2}{m}}}.$$

• Se le due medie sono uguali, la distribuzione di t^*_{oss} può essere approssimata da una t di Student con gradi di libertà calcolati come

$$\frac{\left(\frac{s_y^2}{n} + \frac{s_x^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_y^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_x^2}{m}\right)^2}$$

Osservazione. I gradi di libertà calcolati con la formula precedente sono tipicamente non interi. Nell'utilizzo di una tabella dei quantili di una t di Student si può utilizzare l'intero più piccolo del valore ottenuto.

• **Esempio.** Nel caso dei dati che stiamo considerando in questa unità

$$t^*_{oss} = \frac{22{,}47 - 21{,}13}{\sqrt{\frac{0{,}46}{16} + \frac{0{,}55}{15}}} \approx 5{,}63$$

mentre applicando la formula per i gradi di libertà otteniamo

$$\frac{\left(\frac{0{,}46}{16} + \frac{0{,}55}{15}\right)^2}{\frac{1}{15}\left(\frac{0{,}46}{16}\right)^2 + \frac{1}{14}\left(\frac{0{,}55}{15}\right)^2} = 28{,}31$$

e quindi il valore di t^*_{oss} deve essere confrontato con i valori "previsti" da una t di Student con 28 gradi di libertà.
E' facile verificare che, anche procedendo in questa maniera, nulla cambia nelle conclusioni: il valore di t^*_{oss} rimane "troppo grande" perchè si possa pensare che i cuculi non si siano in qualche maniera adattati all'uccello ospite.
"Troppo grande" è ovviamente da intendersi rispetto ai valori previsti dalla t di Student.

Inferenza sulla differenza tra due medie: campioni di numerosità elevata

• E' possibile dimostrare che se sia n che m tendono ad infinito, allora t_{oss} se le varianza dei due gruppi sono uguali e $t^\star_{oss}$ in tutti i casi convergono in distribuzione ad una normale standard anche se la distribuzione dei due gruppi non è normale.
• Quindi, quanto visto in questa unità (test e intervalli di confidenza) può essere applicato per confrontare le medie di due gruppi di osservazioni purchè[7]

- ambedue le numerosità campionarie siano sufficentemente grandi[8];
- le osservazioni all'interno dei due gruppi siano indipendenti ed identicamente distribuite;
- ambedue le distribuzioni abbiano media e varianza finite;
- le osservazioni di un gruppo siano indipendenti dalle osservazioni dell'altro gruppo.

Ovviamente la validità delle procedure sarà solo approssimata se le distribuzioni dei dati all'interno di ogni gruppo non è esattamente normale.
• Strettamente parlando dovremmo utilizzare i quantili di una normale non quelli di una t. Però visto che stiamo pensando a situazioni in cui n e m sono grandi, utilizzare i quantili di una $N(0,1)$ o di una $t(n+m-2)$ è praticamente lo stesso.

[7] altrimenti non vale il risultato asintotico menzionato
[8] si veda il lucido 46 per alcune indicazioni a spanne.

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 o come ...
Non abbiamo neanche bisogno di conoscere il valore della statistica test.

Unità I

Un piccolo esperimento sulla coltivazione delle fragole

Test t per dati appaiati

Il problema e i dati

- Per confrontare l'efficacia di due differenti fertilizzanti[1],
  - 10 appezzamenti, di uguale estensione, sono stati divisi in due parti uguali;
  - tutti gli appezzamenti sono stati coltivati a fragole;
  - in una delle parti è stato però utilizzato il primo fertilizzante e nell'altra il secondo;
  - al momento della raccolta è stata poi "pesata" la quantità di fragole prodotte nelle varie parti.
- La tabella mostra i dati raccolti. I pesi delle fragole sono in kg.

appezzamento	1	2	3	4	5
fertilizzante A	216,7	149,9	136,9	211,2	171,4
fertilizzante B	196,8	108,9	134,7	195,6	160,2

appezzamento	6	7	8	9	10
fertilizzante A	138,0	127,6	160,3	153,8	150,4
fertilizzante B	141,9	114,1	130,6	116,2	150,2

- Il problema che ci poniamo è se i differenti fertilizzanti hanno un differente effetto sulla media delle fragole prodotte.

[1] in realtà si tratta di terreni arricchiti con sali minerali e altre sostanze da mescolare con il terreno prima della semina.

Perchè non utilizzare un test t a due campioni?

• Indichiamo con y_i e x_i, $i = 1, \ldots, 10$, le quantità di fragole raccolte nell'appezzamento i-simo. y_i è la quantità raccolta nel sotto-appezzamento coltivato con A. x_i l'analoga quantità riferita al sotto-appezzamento coltivato con B.

• In prima battuta potrebbe venire l'idea di utilizzare un test t a due campioni per vericare la significatività della differenze delle medie.

• L'assunzione su cui si basa questo test sono[2]:

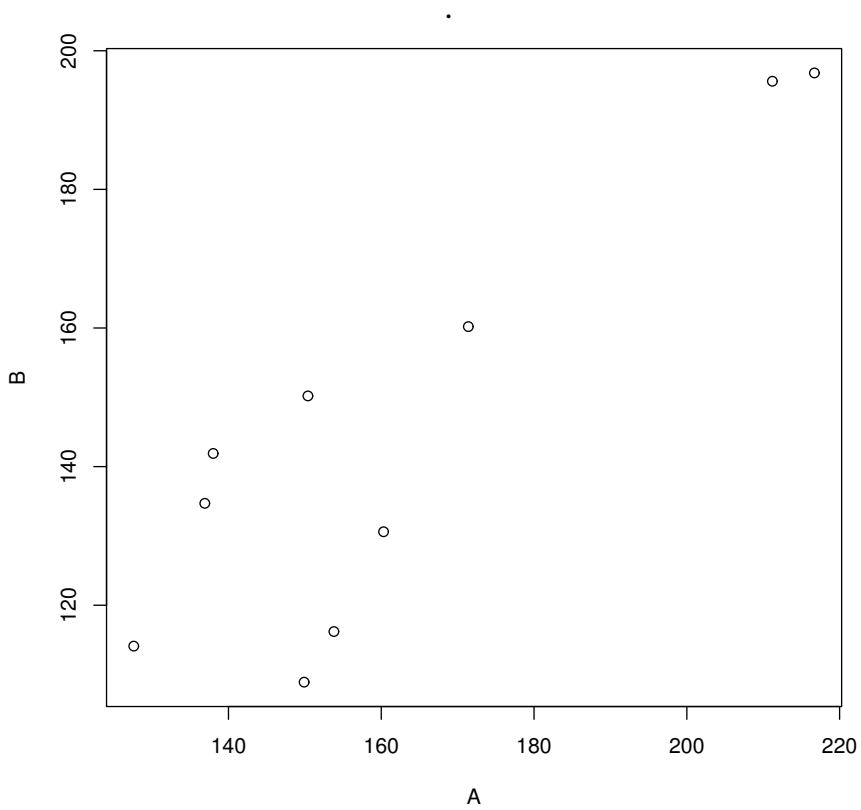
- indipendenza e normalità della distribuzione dentro i due gruppi (ovvero sia le "y" che le "x" devono essere determinazioni indipendenti di variabili casuali normali);
- indipendenza delle osservazioni nei due gruppi (ovvero le "y" devono essere indipendenti dalle "x").

• Trascurando per il momento l'ipotesi di normalità, si osservi come nel caso che stiamo considerando possa essere inappropriata la seconda assunzione.

• Ad esempio, se i vari appezzamenti hanno differenti livelli di fertilità, ci possiamo aspettare una dipendenza tra le quantità prodotte nei sotto-appezzamenti coltivati con A e B.

• Infatti, se l'appezzamento i-simo è particolarmente fertile (per la qualità del terreno, per il tipo di irrigazione, per l'esposizione al sole,...), potrebbe capitare che sia y_i che x_i siano grandi rispetto alle altre osservazioni.

[2] viste le numerisità campionarie

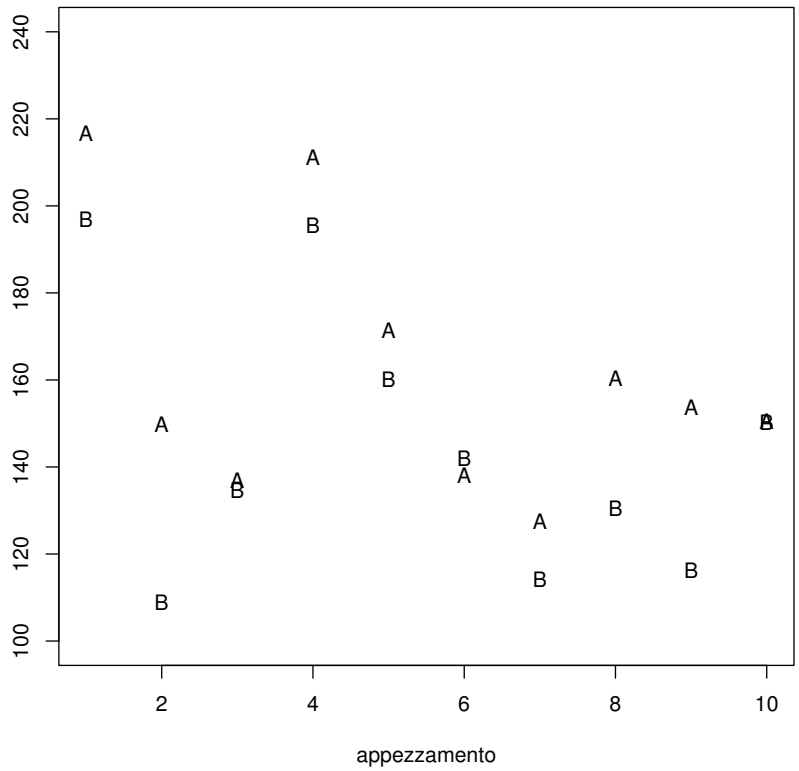


• Il diagramma di dispersione in cui abbiamo disegnato le coppie (y_i, x_i) mostra chiaramente che qualcosa del genere è accaduto. Abbiamo una discreta relazione tra la produzione nei due sotto-appezzamenti (il coefficente di correlazione vale $0{,}875$). Non possiamo quindi utilizzare il test t a due campioni per valutare la significatività delle differenze delle medie.

• Il problema nasce dal fatto che in questo caso abbiamo *misure ripetute* sulla stessa unità sperimentale (l'appezzamento). Si tratta di situazioni abbastanza comuni. Si pensi ad esempio rilevazioni fatte sugli stessi *prima* e *dopo* una terapia. O più in generale, a osservazioni fatte nel tempo sugli stessi soggetti.

Il test *t* per dati appaiati

• Il grafico mostra per ogni appezzamento la produzione ottenuta nel sotto-appezzamento coltivato con A (lettera A) e nel sotto-appezzamento coltivato con B (lettera B).



• Sembra evidente che oltre ad un effetto del fertilizzante sulle quantità prodotte (in 9 appezzamenti su 10 le A sono più grandi delle B) esiste anche un effetto dell'appezzamento. Ad esempio, ambedue le misure sul primo appezzamento sono superiori alle misure ottenute sugli appezzamenti 2 e 3. Quindi, il primo appezzamento sembra più fertile degli appezzamenti 2 e 3.

• In questa situazione, un possibile modello per le medie, potrebbe essere

$$E\{y_i\} = \mu_i, \quad E\{x_i\} = \mu_i + \delta \quad (i = 1, \ldots, 10)$$

ovvero chiedere che

- la media delle osservazioni dipenda sia dal fertilizzante (via δ) ma anche dall'appezzamento (visto che le "μ" dipendono da i, ovvero dall'appezzamento, le osservazioni in appezzamenti differenti hanno medie differenti)
- richiedendo però che la differenza legata ai fertilizzanti sia uguale in tutti gli appezzamenti (δ non dipende da i).

• Si osservi che in questo modello il problema di verificare se i due fertilizzanti hanno un effetto diverso diventa il problema di verificare

$$H_0 : \delta = 0 \text{ verso } H_1 : \delta \neq 0.$$

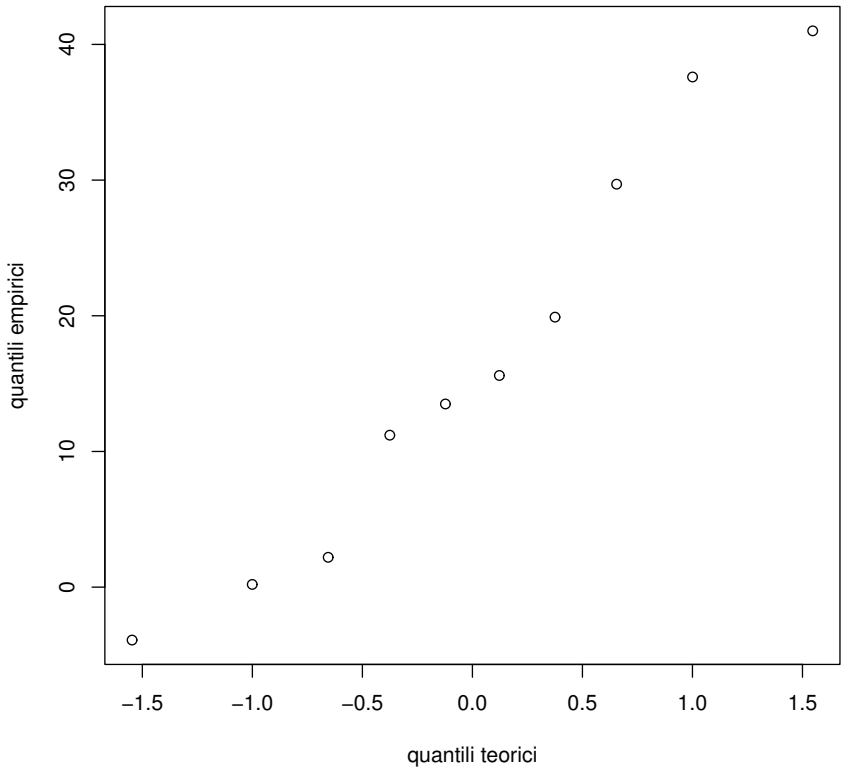
• Si ponga $z_i = x_i - y_i$, $i = 1, \ldots, 10$. Se vale il modello precedente

$$E\{z_i\} = E\{x_i\} - E\{y_i\} = \mu_i + \delta - \mu_i = \delta.$$

• Quindi lavorando con le "z" il problema di verifica di ipotesi precedente diventa un problema sulla media di un insieme di osservazioni univariate (non sulle differenze delle medie di più gruppi).

• Se le "z" sono normali può essere affrontato utilizzando un test t ad un campione.

• Il *normal probability plot*, confortato anche dal test di Shapiro-Wilks (livello di significatività osservato $\approx 0{,}6$, lascia pochi dubbi sulla normalità delle "z".



• Applicando[3] il test t ad un campione alle differenze (le "z") otteniamo un livello di significatività osservato inferiore a 0,01 e quindi, in definitiva, accettiamo l'ipotesi che ci siano delle differenze tra i due fertilizzanti (più precisamente che i due fertilizzanti abbiano effetti differenti sulle medie delle fragole prodotte).

[3] lo studente per esercizio lo verifichi.

• E' interessante osservare[4] che, se si fosse utilizzato un test t a due campioni per confrontare i due gruppi, il livello di significatività osservato sarebbe stato, utilizzando o no la correzione di Welch, $\approx 0{,}24$ ovvero saremmo arrivati ad una conclusione opposta.

[4] un altro esercizio da fare direi!

Unità J

Hot-dog e calorie

(a) Scomposizione della devianza totale.

(b) Misura della importanza delle differenze tra le medie

(c) Analisi della varianza con un criterio di classificazione.

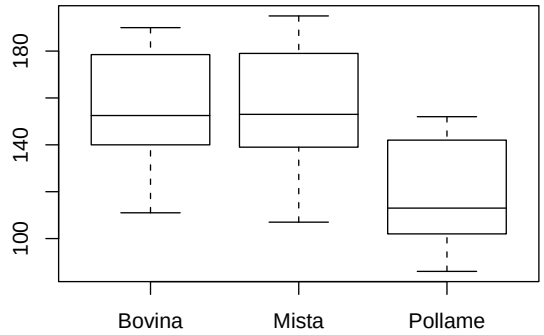
I dati

• 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 rilevando anche se l'*hot-dog* era stato preparato con:
— solo carne bovina;
— carne mista (tipicamente a maggioranza maiale);
— pollame (pollo o tacchino).
• I prossimi due lucidi mostrano:
(i) i dati elementari;
(ii) il diagramma scatola con baffi delle calorie classificate per tipo di carne e le numerosità, medie e scarti quadratici medi dei tre gruppi.
• E' evidente che, restringendo l'attenzione alle 54 misure disponibili, il tipo di carne influenza il contenuto calorico.
• Nel seguito dell'unità ci concentremo sulle differenze tra le medie rilevabili dalla tabella di pagina 162 ed in particolare cercheremo di dare una risposta alle seguenti domande:
– come possiamo "misurare" l'importanza di queste differenze?
– come verificare se è plausibile che le differenze osservate 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)?

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	$\bar{y}$	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 $n - 1$"

Notazioni

• Per rendere il discorso generale indichiamo con
– k il numero dei gruppi;
– n_i, $i = 1, \ldots, k$ il numero di osservazioni per ogni gruppo.
Nel nostro caso, ovviamente, $k = 3$ e, convenendo che, 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 entro il gruppo.
• Per ogni gruppo possiamo calcolare la media e la *devianza* campionaria

$$\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad d_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$$

Nel nostro caso, queste quantità sono riferibili alla posizione e alla dispersione delle distribuzioni delle calorie *condizionate* ai vari tipi di carne.
• Possiamo inoltre anche calcolare la media e la devianza *totali* ovvero di *tutte* le osservazioni senza riferimento al gruppo di appartenenza

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} \text{ e } d^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$$

dove

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

indica il numero totale di osservazioni disponibili.
$\overline{y}$ e d^2 sono riferibili alla distribuzione *marginale* delle calorie.

La media totale è uguale alla media delle medie dei gruppi

• Pensiamo alla distribuzione di frequenza in cui le modalità sono le medie dei k gruppi e le frequenze (assolute) sono le numerosità delle osservazioni nei vari gruppi, ovvero, a

modalità	$\overline{y}_1$	$\overline{y}_2$	...	$\overline{y}_k$
frequenze	n_1	n_2	...	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 $\overline{y}$. 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, sostituendo, troviamo

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k} n_i\overline{y}_i$$

• Si osservi che questa relazione non vale solo nel campione ma anche nella popolazione. E' infatti, in generale, possibile dimostrare, e spesso molto utile da ricordare, che la media di una distribuzione marginale può essere calcolata come media delle medie condizionate.

La devianza totale è la somma delle devianze dei gruppi + la devianza delle medie dei gruppi

• Ci si ricordi che d^2 indica la devianza di tutti i dati (= la devianza della "distribuzione marginale"), mentre d_i^2 è la devianza dentro il gruppo i-simo (= le devianze delle "distribuzione condizionate").
• Dimostreremo che

$$d^2 = \sum_{i=1}^{k} d_i^2 + \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2. \qquad \text{(J.1)}$$

• Si osservi che il primo addendo sul lato destro è la somma delle devianze interne ai vari gruppi.
• Viceversa, il secondo addendo è la devianza della distribuzione mostrata all'inizio di pagina 164, ovvero è la "devianza delle medie dei gruppi".
• La verifica della (J.1) è agevole. Infatti[1]

$$\begin{aligned}
d^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 = \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(y_{ij} - \overline{y}_i) + (\overline{y}_i - \overline{y})]^2 = \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(y_{ij} - \overline{y}_i)^2 + (\overline{y}_i - \overline{y})^2 + 2(y_{ij} - \overline{y}_i)(\overline{y}_i - \overline{y})] = \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2 + 2\sum_{i=1}^{k}(\overline{y}_i - \overline{y})\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i) = \\
&= \sum_{i=1}^{k} d_i^2 + \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2.
\end{aligned}$$

[1] nell'ultimo passaggio si ricordi che la somma delle osservazioni del gruppo i-simo dalla media del gruppo i-simo vale zero.

Una misura dell'importanza delle differenze tra le medie dei vari gruppi

• La (J.1) mostra come la devianza totale, d^2, sia scomponibile in due parti:

(i) la prima, il 1° addendo, legata alla dispersione all'interno dei vari gruppi e

(ii) la seconda, il 2° addendo, legata le differenze (in media) *tra* i gruppi.

Per questo motivo, i due addendi sono spesso indicati come *devianza entro i gruppi* e *devianza tra i gruppi*.
• Si osservi che se la devianza tra i gruppi è nulla, allora le medie di tutti i gruppi sono tutte uguali a $\overline{y}$ e quindi tutte uguali tra di loro.
• Viceversa, se la varianza tra i gruppi è molto grande rispetto alla varianza entro i gruppi, allora buona parte della variabilità totale dei dati è interpretabile in termini di differenze tra le medie dei gruppi. Siamo quindi in presenza di una situazione in cui la differenza tra le medie è importante (= "spiega" una larga frazione della variabilità che osserviamo nei dati).
• Sembra allora ragionevole usare

$$\begin{aligned}
\eta^2 &= \frac{\text{devianza tra i gruppi}}{\text{devianza totale}} = \\
&= 1 - \frac{\text{devianza entro i gruppi}}{\text{devianza totale}}
\end{aligned}$$

per misurare l'importanza delle differenze tra le medie dei gruppi.

• In particolare si osservi che

(a) $0 \leq \eta^2 \leq 1$.

(b) $\eta^2 = 0$ implica che le medie dei gruppi sono tutte uguali tra di loro (*indipendenza in media* almeno nel campione).

(c) $\eta^2 = 1$ implica che la devianza entro i gruppi è nulla. Siamo quindi in una situazione di dipendenza perfetta.

(d) η^2 non è ovviamente definito quando $d^2 = 0$. Questo non è un grande problema visto che d^2 uguale a zero vuol dire che tutte le osservazioni sono uguali tra di loro e quindi che non esiste nessuna variabilità interessante da indagare.

• Nel caso degli *hot-dog*, η^2 è facilmente calcolabile dai risultati della tabella di pagina 162.[2].

$$\begin{aligned} \text{devianza entro i gruppi} &\approx 28067{,}78 \\ \text{devianza tra i gruppi} &\approx 17698{,}32 \\ \text{devianza totale} &\approx 45766{,}11 \end{aligned}$$

e, quindi, $\eta^2 \approx 0{,}39$. Il valore trovato ci indica la presenza di una discreta ma non eccezionale dipendenza in media.

[2] la devianza entro i gruppi può essere calcolata come $\sum(n_i - 1)s_i^2$.

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.

• Viceversa, potremmo essere 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 fonte di variabilità, quella *campionaria*: almeno una parte delle differenze tra le medie delle osservazioni dei vari gruppi è specifica alle 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[3] 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;
– 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

$H_0 : \mu_1 = \mu_2 = \mu_3$
H_1 : almeno una delle uguaglianze previste da H_0 è falsa

• Infatti, se H_0 fosse vera, allora nella popolazione, contrariamente a quanto osservato nel campione, il tipo di carne utilizzato non influenzerebbe il contenuto degli *hot-dog*. Ovvero, quello che abbiamo osservato nei dati sarebbe un artefatto legato al campionamento.
• 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.

[3] un po' stomachevole?

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 intercorre tra le osservazioni e le tre medie μ_1, μ_2 e μ_3.
• Una soluzione relativamente "semplice" esiste quando sia credibile assumere che:

1. la distribuzione all'interno dell'i-gruppo è normale di media μ_i e varianza σ^2, ovvero,

$$y_{ij} \sim N(\mu_i, \sigma^2) \quad (i = 1, \ldots, k; j = 1, \ldots, n_i);$$

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.

• La statistica test comunemente usata è

$$F_{oss} = \frac{(\text{devianza tra i gruppi})/(k-1)}{(\text{devianza entro i gruppi})/(n-k)}$$

• La statistica F_{oss} è in stretta relazione con η^2. Infatti, come è facile verificare,

$$F_{oss} = \left(\frac{\eta^2}{1-\eta^2}\right)\left(\frac{n-k}{k-1}\right).$$

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 F_{oss} e viceversa.

• Ovviamente, poichè ci aspettiamo F_{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 F_{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,...), F_{oss} si distribuisce come una variabile casuale F di Snedecor con $k-1$ gradi di libertà al numeratore e $n-k$ al denominatore[4].
Possiamo quindi confrontare il valore osservato di F_{oss} con i valori "possibili" per questa variabile casuale.
• *Applicazione ai dati.* Per i dati sugli *hot-dog*, $F_{oss} \approx 16$. Questo valore deve essere confrontato con i quantile di una F di Snedecor con 2 e 51 gradi di libertà. Consultando una tabella dei quantili di una distribuzione F possiamo vedere che il valore osservato è molto più grande del quantile 0,999 di questa distribuzione e, quindi, 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.

[4] per la definizione di questa variabile casuale si veda [Probalità 19].

Unità K

Dove facciamo la conoscenza con delle statistiche di alto rango

Cenno ai test basati sui ranghi.

Trasformazione rango

Definizione. Sia $z = (z_1, \ldots, z_N)$ un vettore di N numeri. Allora la trasformazione rango di z è il vettore di interi $r = (r_1, \ldots, r_N)$ tale che

$$r_i = \text{numero di “z” minori od uguali a } z_i = \sum_{j=1}^{N} I(z_j \leq z_i)$$

dove

$$I(A) = \begin{cases} 0 & \text{se A è falsa} \\ 1 & \text{se A è vera} \end{cases} .$$

In altre parole, r_j, ovvero il rango di z_j, è la posizione di z_j nella seguenza ordinata dei numeri. Ad esempio se $r_5 = 2$ allora solo un'altra osservazione è più piccola o al più uguale a z_5, tutte le altre “z” sono più grandi.

Esempio. Supponiamo

$$z = (3{,}1\ ;\ 0{,}4\ ;\ 4{,}3\ ;\ -1{,}6\ ;\ 0{,}4).$$

Allora il vettore dei ranghi di z è

$$r = (4, 3, 5, 1, 3).$$

Osservazione. Esistono altre *“versioni”* della trasformata rango di un insieme di osservazioni. Tutte coincidono nei casi in cui non ci siano valori ripetuti tra le “z” Trattano però in maniera diversa osservazioni uguali (nella definizione di prima viene assegnato il “rango più elevato”, in altre il “rango medio”, in altre ancora un “rango casuale”,...).

Trasformata rango e variabili casuali i.i.d.

Siano $z_1, \ldots, z_N$ delle determinazioni indipendenti ed identicamente distribuite di una variabile casuale assolutamente continua con valori in $\mathcal{R}$. Si indichi con $r = (r_1, \ldots, r_N)$ il vettore dei ranghi di $z_1, \ldots, z_N$.

I ranghi sono tutti distinti e quindi Il vettore dei ranghi è una delle $N!$ permutazioni di $(1, \ldots, N)$. Infatti, con probabilità uno, le osservazioni sono distinte (la probabilità che due determinazioni di una variabile casuale continua siano uguali è nulla).

Tutti i valori che r può assumere sono equiprobabili. Ovvero, è possibile dimostrare è che per qualsivoglia $s = (s_1, \ldots, s_N)$, permutazione di $(1, \ldots, N)$, allora

$$\Pr(r = s) = \frac{1}{N!}.$$

Importanza del risultato enunciato. Si osservi che la distribuzione del vettore dei ranghi non dipende dalla distribuzione dei dati; le $z_1, \ldots, z_N$ potrebbero essere normali, esponenziali, beta,... ma la distribuzione dei ranghi delle osservazioni rimane costante e completamente nota (se le osservazioni sono determinazioni i.i.d. di una v.c. continua).

Test di Wilcoxon per due campioni

I dati. I dati sono del tipo di quelli considerati per il test t a due campioni:

- $(y_1, \ldots, y_n)$ determinazioni indipendenti di una variabile casuale continua con funzione di ripartizione $F(\cdot)$;
- $(x_1, \ldots, x_m)$ determinazioni indipendenti di una variabile casuale continua con funzione di ripartizione $G(\cdot)$;
- le "y" sono indipendenti dalle "x".

Nessuna assunzione su F e G Tolta l'assoluta continuità, supporremmo però $F(\cdot)$ e $G(\cdot)$ completamente ignote: sono due qualsiasi funzioni di ripartizione.

Nelle unità precedenti, la distribuzione di probabilità dei dati osservati era nota a meno di un certo numero di parametri reali[1] Nella situazione che stiamo considerando questo non è più vero. Per questo motivo quello che stiamo per affrontare è un problema di *inferenza statistica non parametrica*.

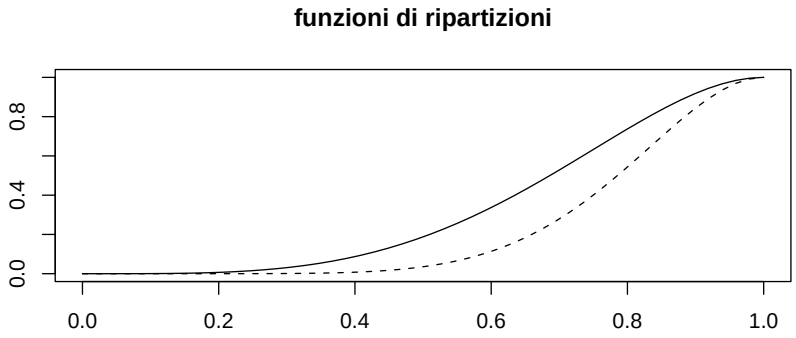
Ipotesi L'ipotesi nulla prevede che i due gruppi abbiano la stessa distribuzione:

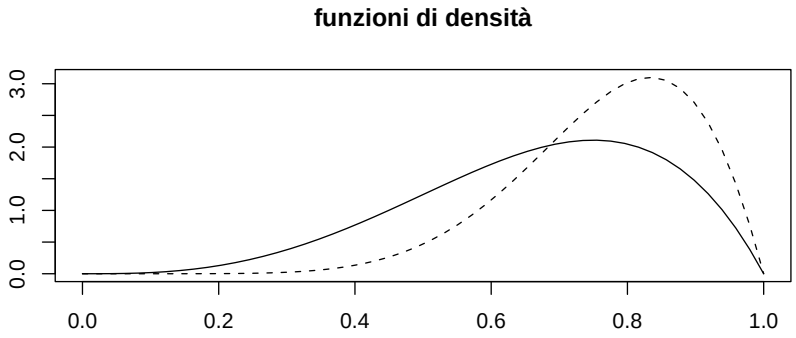
$$H_0 : F(x) = G(x), \quad \forall x \in \mathcal{R}.$$

L'ipotesi alternativa che consideriamo è *unilatelare* e, lasciamola espressa informalmente, prevede che la distribuzione delle x sia *"spostata verso destra"* rispetto alla distribuzione delle y[2]. Ovvero, l'ipotesi alternativa prevede che, tendelzialmente, le "x" siano *più grandi* delle "y".

[1] ad esempio la distribuzione era normale di (i "parametri" della distribuzione) media e varianza ignota

[2] come esercizio, lo studente può provare a formulare la versione "bilaterale" del test.





La figura mostra un esempio di una delle situazioni “previste” da H_1: due distribuzioni di probabilità differenti con quella a cui corrispondono le curve tratteggiate che “genera” valori tendelzialmente più verso destra dell’altra.

Statistica test Formiamo il vettore (di dimensione $N = n + m$) di tutte le osservazioni disponibili[3]

$$z = (x_1, \ldots, x_m, y_1, \ldots, y_n)$$

e poi calcoliamone il vettore dei ranghi

$$r = (r_1, \ldots, r_m, \ldots, r_N).$$

I primi m valori sono i ranghi delle “x” nel campione combinato. I restanti valori sono i ranghi delle “y”.
Si osservi che

- quando è vera H_0 ci aspettiamo che i ranghi delle “x” siano “mescolati” con i ranghi delle “y”;
- viceversa, quando è vera H_1, ci aspettiamo che i ranghi delle “x” siano “più grandi” dei ranghi delle “y” visto che sotto H_1 le “x” sono tendelzialmente “più grandi” delle “y”.

Poniamo

$$W = \sum_{i=1}^{m} r_i - \frac{m(m+1)}{2} = \\ = \left(\begin{array}{c}\text{somma dei ranghi delle} \\ \text{“x”}\end{array}\right) - \left(\begin{array}{c}\text{costante che dipende solo} \\ \text{dal numero delle “x”}\end{array}\right)$$

Per il discorso fatto ci aspettiamo valori di W più grandi sotto H_1 che sotto H_0. Possiamo quindi utilizzare W come statistica test.

[3] che per ipotesi sono tutte distinte; ma vedi anche dopo...

Interpretazione alternativa della statistica test Si osservi che

$$W = \sum_{i=1}^{m}\sum_{j=1}^{N} I(z_j \le x_i) - m(m+1)/2 =$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{m} I(x_j \le x_i) + \sum_{i=1}^{m}\sum_{j=1}^{n} I(y_j \le x_i) - m(m+1)/2.$$

E' facile far vedere che, quando le osservazioni sono tutte distinte,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} I(x_j \le x_i) = 1 + 2 + \cdots + m = \frac{m(m+1)}{2}.$$

Quindi

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} I(y_j \le x_i) =$$
$$= \text{numero coppie } (x_i, y_j) \text{ con } y_j \le x_i.$$

Anche scritta in questa maniera è evidente che più i dati sono a favore di H_1, ovvero, più le "x" sono a destra delle "y", più W assume valori grandi.
La scrittura mostra inoltre immediatamente che W è un numero intero che assume valori tra 0 e $n \times m$.
Distribuzione di W sotto l'ipotesi nulla La distribuzione di r, il vettore dei ranghi, è nota quanto è vera l'ipotesi nulla. Infatti, sotto H_0, il "campione combinato" z è un vettore di $N = n+m$ determinazioni indipendenti ed identicamente distribuite di variabile casuale continua (l'ipotesi nulla prevede che la distribuzione delle "x" sia uguale a quella delle "y").
La statistica test W è semplicemente una trasformata di r e quindi se conosciamo la distribuzione di r, possiamo calcolare anche la distribuzione di W.
Il punto importante è che riusciamo a determinare la distribuzione sotto H_0 di W anche se non conosciamo la funzione di ripartizione delle osservazioni.

Livello di significatività osservato Più W è grande più è "contro H_0". Quindi il livello di significatività osservato può essere calcolato come

$$P(W \ge W_{oss})$$

dove con W_{oss} è stato indicato il valore di W calcolato dai dati.
Ipotesi alternativa bilaterale. Il discorso fatto è facilmente estendibile al caso di ipotesi alternativa bilaterali, ovvero, quando, sotto H_1, la distribuzione delle "x" può essere o a destra o a sinistra della distribuzione delle "y".
La statistica W continua ad essere appropriata. Sotto H_1, ci aspettiamo valori di W o più grandi o più piccoli di quelli attesi sotto H_0.
Visto che è possibile far vedere che la distribuzione[4] sotto H_0 è simmetrica intorno a

$$\frac{nm}{2}$$

il livello di significatività osservato in questo caso è

$$P(|W - nm/2| \ge |W_{oss} - nm/2|).$$

E se ci sono dati uguali? Per il modello, dati uguali non possono capitare. Nella realtà può capitare di trovare due o più dati uguali. Al proposito, e' necessario considerare separatamente due casi:
- la variabile considerata è fondamentalmente continua; i dati uguali sono pochi e semplicemente il frutto di arrotondamenti; in questo caso, possiamo nella sostanza ignorarli utilizzando una qualsiasi conveniente definizione di rango.
- la variabile considerata è realmente discreta e può assumere pochi valori; in questo caso non ci sono le condizioni per applicare il test che stiamo considerando.

[4] ovviamente è la stessa sia se l'ipotesi alternativa è unilaterale sia se è bilaterale.

Un esempio

• In una ricerca sono state utilizzate due modalità differenti di coltivazione di una certa pianta (officinale):

- 10 piante sono state coltivate con la tecnica “classica”; i pesi delle piante raccolte ed essiccate sono risultati (le “y”):

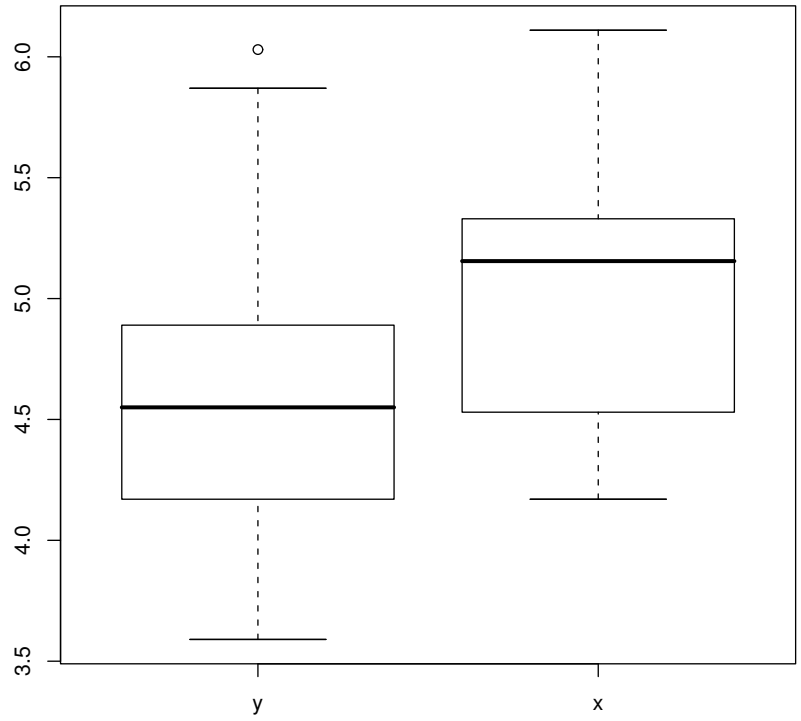
  ```
  4,81 4,17 4,41 3,59 5,87
  3,83 6,03 4,89 4,32 4,69
  ```

- altre 10 piante sono state coltivate con una tecnica “nuova”; i pesi in questo caso sono risultati (le “x”):

  ```
  4,17 5,58 5,18 6,11 4,50
  4,61 5,17 4,53 5,33 5,14
  ```

• Vogliamo verificare se la nuova tecnica è migliore di quella classica ovvero se la distribuzione da cui provengono le “x” è in un qualche senso a destra della distribuzione da cui provengono le “y”.

• Il *boxplot* indica che nel campione la nuova tecnica si è “comportata meglio”.



• Proviamo ad utilizzare il test di Wilcoxon per verificare se la differenza è significativa, ovvero se possiamo aspettarci che sia un risultato di una reale differenza tra le due tecniche e non un semplice artefatto legato al campione.

• Si osservi che due delle osservazioni sono uguali. In questo caso, possiamo attribuire l’uguaglianza ad un semplice effetto di arrotondamento e quindi procedere lo stesso.

• Osservazioni ordinate e ranghi. Per la coppia di osservazioni uguali sono utilizzate due definizioni alternative di rango (rango massimo e rango medio).

dati	x o y?	rango massimo	rango medio
3,59	y	1	1
3,83	y	2	2
4,17	y	4	3,5
4,17	x	4	3,5
4,32	y	5	5
4,41	y	6	6
4,5	x	7	7
4,53	x	8	8
4,61	x	9	9
4,69	y	10	10
4,81	y	11	11
4,89	y	12	12
5,14	x	13	13
5,17	x	14	14
5,18	x	15	15
5,33	x	16	16
5,58	x	17	17
5,87	y	18	18
6,03	y	19	19
6,11	x	20	20
m=(numero di “x”)=		10	10
$m(m+1)/2 =$		55	55
somma ranghi “x”=		123,0	122,5
$W =$		67	67,5

• La statistica test vale 67 o 67,5 a seconda della definizione di rango che si adotta.

• Utilizzando o delle tavole o una funzione appropriata[5] troviamo che il livello di significatività osservato è all’incirca del 10%.

• Il valore non è molto grande ma è ancora compatibile con H_0. Siamo quindi in una situazione di accettazione, con qualche dubbio, dell’ipotesi che le due tecniche di coltivazione non abbiano differente efficienza.

[5] in R, la funzione che calcola la funzione di ripartizione della statistica test di Wilcoxon a due campioni si chiama `pwilcox`.

Wilcoxon o Student? Una guerra non ci serve!

Vantaggio del test di Wilcoxon E' utilizzabile anche per piccoli campioni senza che sia necessario assumere la normalità dei dati.
Vantaggio del test *t* a due campioni Se i dati sono normali, il test basato sulla t è più *potente*, ovvero, a parità di errore di primo tipo, permette di ottenere una probabilità di errore di secondo tipo più bassa (= dichiara più spesso che H_1 è vera quando H_1 è realmente vera).
Nelle applicazioni... ...è comunque conveniente utilizzarli in maniera combinata.

> *"...Per verificare l'ipotesi che la nuova tecnica sia migliore abbiamo utilizzato il test* t *di Student (*$p = 0{,}125$*) e il test di Wilcoxon (*$p \approx 0{,}109$*)..."*

Risultati simili (come nel caso illustrato qui sopra) si confermano a vicenda. La discussione di risultati contrastanti è spesso illuminante.

Altri test di "alto rango"

Esistono test basati sui ranghi, e quindi utilizzabili anche per piccoli campioni senza assunzione parametriche, per svariati problemi di verifica di ipotesi.
Mi limito a menzionarne due.
Wilcoxon a un campione. E' un test sulla mediana di un singolo campione e quindi "fratello" del test t ad un campione. Richiede la simmetria della distribuzione dei dati ma non la normalità.
Può anche essere utilizzato confrontare due gruppi nel caso di dati appaiati.
Kruskal-Wallis. E' l'analogo basato sui ranghi dell'analisi della varianza ad un criterio di classificazione. Confronta quindi k gruppi. L'ipotesi nulla è che abbiano la stessa distribuzione. L'ipotesi alternativa è che almeno un gruppo abbia una distribuzione che genera valori o più piccole o più grandi delle altre.

Appendice

Richiami e complementi di probabilità

Per facilitarmi i richiami a lezione riporto in questa appendice alcuni "flash informali" di probabilità.

La distribuzione normale

Probabilità 1 Una variabile casuale continua, chiamiamola Y, si dice normale di media μ e varianza σ^2 se la sua funzione di densità è

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \quad (-\infty < x < +\infty).$$

Scriveremo in questo caso $Y \sim N(\mu,\sigma)$ dove il simbolo $\sim$ si legge *"si distribuisce come"*. Sinonimo di distribuzione normale è distribuzione gaussiana o di Gauss. Nel caso in cui $\mu = 0$ e $\sigma = 1$ diremo che Y è una normale standard.

Probabilità 2 La densità è simmetrica intorno a μ. Il supporto, se $\sigma > 0$, è tutta la retta reale (ovvero una variabile casuale normale può assumere valori da $-\infty$ a $+\infty$). Però quasi tutta la sua massa è concentrata nell'intervallo $[\mu - 3\sigma; \mu + 3\sigma]$ visto che

$$\text{se } Y \sim N(\mu,\sigma^2) \text{ allora } P(\mu - 3\sigma \leq Y \leq \mu + 3\sigma) \approx 0{,}9973.$$

Probabilità 3 Se $Y \sim N(\mu,\sigma^2)$ e ν_0 e ν_1 sono due costanti reali, allora $\nu_0 + \nu_1 Y \sim N(\nu_0 + \nu_1\mu, \nu_1^2\sigma^2)$, ovvero, trasformate lineari di una variabile casuale normale sono normali con media e varianza appropriate.
Quindi, ad esempio,

$$Y \sim N(\mu,\sigma^2) \;\Rightarrow\; \frac{Y-\mu}{\sigma} \sim N(0,1).$$

Probabilità 4 Se Y_1 e Y_2 sono variabili casuali normali indipendenti tra loro allora anche le loro combinazioni lineari, ovvero le variabili casuali del tipo $Y = \nu_1 Y_1 + \nu_2 Y_2$ dove ν_1 e ν_2 sono delle costanti reali, hanno distribuzione normale con media e varianza appropriate[1].
Quindi, ad esempio, somme ($Y = Y_1 + Y_2$) e differenze ($Y = Y_1 - Y_2$) di variabili casuali normali indipendenti sono normali.

[1] per il calcolo della media e della varianza si veda [Probalità 33] e [Probalità 35].

Probabilità 5 Seguendo un uso abbastanza comune, nei lucidi vengono indicati con:

- $\Phi(\cdot)$ la funzione di ripartizione di una variabile casuale normale standard; quindi

$$\Phi(x) = P(N(0,1) \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx$$

- z_p il quantile p-esimo della stessa distribuzione

$$P(N(0,1) \leq z_p) = \Phi(z_p) = p.$$

Probabilità 6 Per il calcolo di $\Phi(\cdot)$ e dei relativi quantili z_p è necessario utilizzare delle funzioni o delle tabelle appropriate. In R le funzioni sono pnorm e qnorm. Una tabella dei quantili di una normale standard è contenuta in "Formulario e tavole" scaricabile dalla pagina del corso.

Probabilità 7 E' importante notare che riuscendo a calcolare $\Phi(\cdot)$ riusciamo a calcolare la funzione di ripartizione di una normale di media e varianza qualsiasi. Infatti se $Y \sim N(\mu, \sigma^2)$ allora

$$P(Y \leq x) = P\left(\frac{Y-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(N(0,1) \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

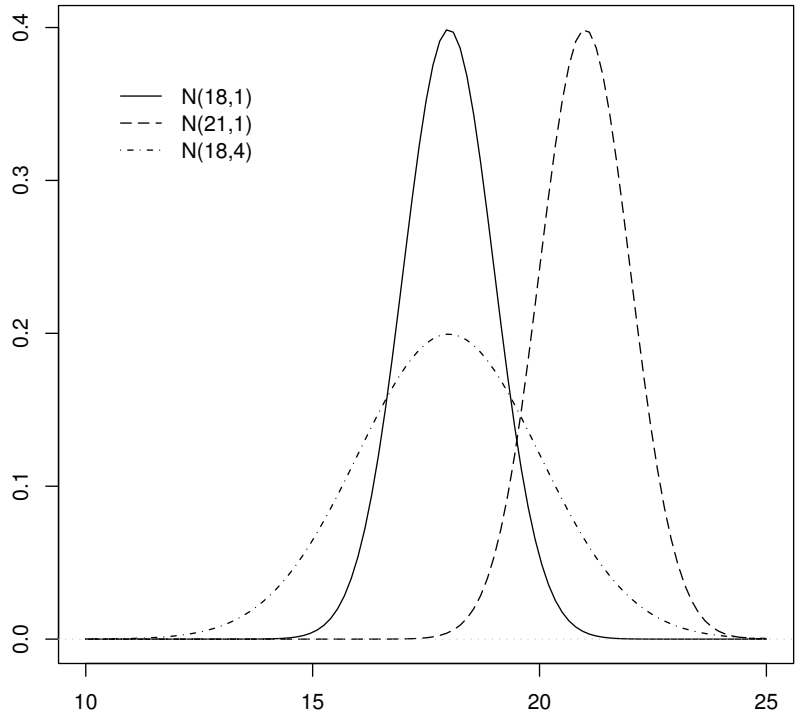
Probabilità 8 Analogamente si osservi che riuscendo a calcolare i quantili di una normale standard riusciamo anche a calcolare i quantili di una normale qualsiasi. Infatti, se $Y \sim N(\mu, \sigma^2)$ allora

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

ovvero

(quantile-p di una $N(\mu, \sigma^2)$) $= \mu + \sigma$(quantile-p di una $N(0,1)$).

Densità di una variabile casuale normale per tre differenti valori di μ e σ^2



Tre distribuzioni di probabilità legate alla distribuzione normale: χ^2

Probabilità 9 Siano $Y_1,\ldots,Y_k$ k variabili casuali indipendenti tra loro e tutte distribuite come una normale standard ($Y_i \sim N(0,1)$, $i = 1,\ldots,k$). Allora diremo che

$$X^2 = Y_1^2 + \cdots + Y_k^2 = \sum_{i=1}^{k} Y_i^2$$

è una variabile casuale χ^2 con k gradi di libertà. Scriveremo in questi casi $X^2 \sim \chi^2(k)$.

Probabilità 10 Per costruzione, una variabile casuale χ^2 è continua e assume solamente valori non negativi.

Probabilità 11 La media e la varianza di un χ^2 con k gradi di libertà valgono rispettivamente k e $2k$, ovvero

$$E\left\{X^2\right\} = k \text{ e } \text{var}\left\{X^2\right\} = 2k.$$

Probabilità 12 Siano X_1^2 e X_2^2 due variabili casuali indipendenti tali che

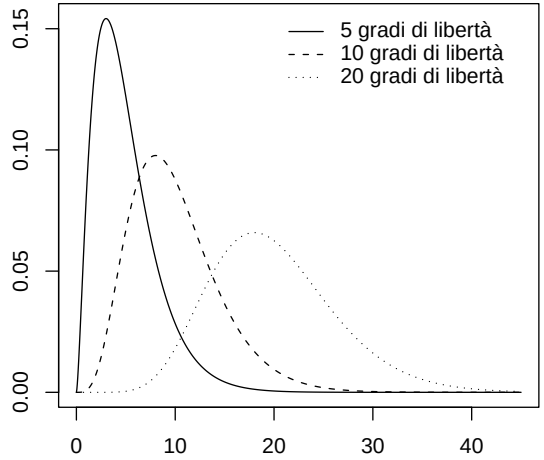
$$X_1^2 \sim \chi^2(k) \text{ e } X_2^2 \sim \chi^2(h).$$

Allora

$$X^2 = X_1^2 + X_2^2 \sim \chi^2(h+k).$$

Probabilità 13 Per il calcolo della funzione di ripartizione e dei quantili è necessario utilizzare delle funzioni o delle tabelle appropriate. In R le funzioni sono pchisq e qchisq. Una tabella dei quantili è contenuta in "Formulario e tavole" scaricabile dalla pagina del corso.

Densità di una variabile casuale χ^2 per tre valori dei gradi di libertà



Si noti inoltre dal grafico come all'aumentare dei gradi di libertà la densità si sposta verso destra (= un χ^2 tende ad assumere valori sempre più grandi più aumentano i gradi di libertà).
Si osservi anche l'asimmetria positiva delle distribuzioni.

Tre distribuzioni di probabilità legate alla distribuzione normale: *t* di Student

Probabilità 14 Siano Y e X^2 due variabili casuali indipendenti tali che

$$Y \sim N(0,1) \text{ e } X^2 \sim \chi^2(k).$$

Allora diremo che

$$t = \frac{Y}{\sqrt{X^2/k}}$$

è una variabile casuale t di Student con k gradi di libertà e scriveremo $t \sim t(k)$.
La distribuzione prende il nome (e il simbolo) da W.S.Gosset, uno statistico che lavoravava alla birreria (nel senso di fabbrica di birra) Guiness. I lavori di Gosset furono pubblicati sotto lo pseudonimo di Student, e Gosset, come anche noi abbiamo fatto, usava la lettera t per indicare la distribuzione, da cui, appunto, t di Student.

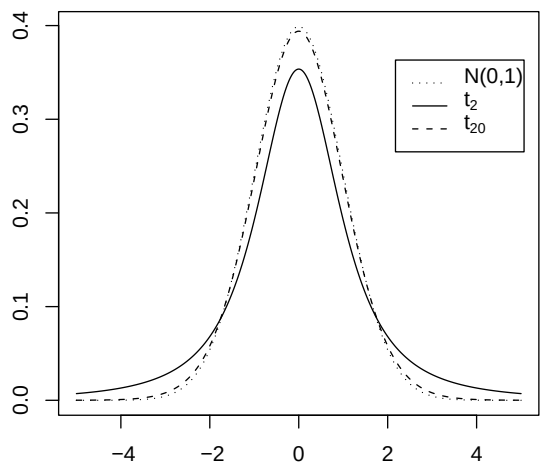
Probabilità 15 La distribuzione è simmetrica intorno allo zero. Il supporto coincide con la retta reale (= una t può assumere valori da $-\infty$ a $+\infty$).

Probabilità 16 Per qualsiasi numero finito dei gradi di libertà k, una t ha code "più pesanti" di quelle di una normale standard (= può assumere con probabilità più grande valori "lontani" da zero);

Probabilità 17 Per $k \to \infty$ la distribuzione converge in distribuzione ad una normale standard. Quindi, una variabile casuale t di Student può essere approssimata con una $N(0,1)$ se k è abbastanza grande.

Probabilità 18 Per il calcolo della funzione di ripartizione e dei quantili è necessario utilizzare delle funzioni o delle tabelle appropriate. In R le funzioni sono pt e qt. Una tabella dei quantili è contenuta in "Formulario e tavole" scaricabile dalla pagina del corso.

Grafico della densità della *t* di Student



Nota: I pedici indicano i gradi di libertà.

Si osservi come già per $k = 20$ non ci siano più grandi differenze tra la densità di una t di Student e quella di una normale standard.

Tre distribuzioni di probabilità legate alla distribuzione normale: *F* di Snedecor

Probabilità 19 Siano X_1^2 e X_2^2 due variabili casuali indipendenti tali che

$$X_1^2 \sim \chi^2(k) \text{ e } X_2^2 \sim \chi^2(h).$$

Allora diremo che

$$F = \frac{X_1^2/k}{X_2^2/h}$$

è una variabile casuale F di Snedecor con (k, h) gradi di libertà (o con k gradi di libertà al numeratore e h al denominatore) e scriveremo $F \sim F(k, h)$.

Probabilità 20 Per costruzione, una variabile casuale F di Snedecor è continua e assume solamente valori non negativi.

Probabilità 21 Per il calcolo della funzione di ripartizione e dei quantili è necessario utilizzare delle funzioni o delle tabelle appropriate. In R le funzioni sono pf e qf. Una tabella dei quantili è contenuta in "Formulario e tavole" scaricabile dalla pagina del corso.

La distribuzione binomiale

Probabilità 22 Una variabile casuale Y, discreta e con supporto $\{0, 1, \ldots, n\}$, viene chiamata binomiale con numero di prove pari ad n e probabilità di successo ϑ se

$$P(Y = y) = \begin{cases} \binom{n}{y} \vartheta^y (1-\vartheta)^{n-y} & \text{se } y = 0, \ldots, n \\ 0 & \text{altrimenti} \end{cases}.$$

Scriveremo in questo caso $Y \sim Bi(n, \vartheta)$.

Probabilità 23 Una variabile casuale binomiale descrive il numero di "successi" ottenuto in n esperimenti casuali che possono risultare o in un "successo" o in un "insuccesso" quando

(i) gli n esperimenti sono completamente indipendenti tra di loro;
(ii) la probabilità di ottenere un successo è uguale a ϑ in ciascuno degli esperimenti.

Il racconto in termini di palline colorate e di urne è il seguente:

(i) esiste un urna contenente palline di 2 colori diversi: "arancione" e "azzurro";
(ii) tutte le palline possono essere estratte con la stessa probabilità;
(iii) la frazione di palline di colore arancione è ϑ (ad esempio, se il 12% delle palline dell'urna è "arancione" allora $\vartheta = 0{,}12$);
(iv) n palline sono estratte dall'urna in maniera indipendente e con *reintroduzione* (quindi la composizione dell'urna è la stessa in ogni estrazione)

allora la variabile casuale Y che descrive il numero di palline estratte di colore arancione è una $Bi(n, \vartheta)$.

Probabilità 24 É possibile far vedere che

$$E\{Y\} = n\vartheta \text{ e } \text{var}\{Y\} = n\vartheta(1-\vartheta).$$

Probabilità 25 *[Approssimazione normale]*. Se n è sufficentemente grande e ϑ è differente da 0 e da 1, la distribuzione binomiale può essere approssimata con una distribuzione normale[2]. In particolare è possibile far vedere che se $Y \sim Bi(n, \vartheta)$ allora, per qualsivoglia reale x,[3]

$$\lim_{n\to\infty} P\left(\frac{Y - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}} \leq x\right) = P(N(0,1) \leq x) = \Phi(x).$$

Quindi, se n è grande, risulta

$$P(Y \leq y) = P\left(\frac{Y - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}} \leq \frac{y - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right) \approx \Phi\left(\frac{y - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right).$$

L'approssimazione è considerata ragionevolmente buona se

$$n\vartheta \geq 5 \text{ e } n(1-\vartheta) \geq 5.$$

[2]si tratta di una delle tante versioni del teorema del limite centrale, in particolare, probabilmente della prima dimostrata

[3]si veda [Probalità 5] per la definizione della funzione $\Phi(\cdot)$.

La distribuzione multinomiale

Probabilità 26 Costituisce la generalizzazione della distribuzione binomiale al caso di più classi/categorie.
Il racconto, in termini di palline colorate e di urne, è:

(i) esiste un urna contenente palline di k colori diversi;
(ii) tutte le palline possono essere estratte con la stessa probabilità;
(iii) la frazione di palline del colore i-simo è π_i (ad esempio, se l'i-simo colore è "viola" allora $\pi_i = 0.12$ indica che il 12% delle palline dell'urna è "viola");
(iv) n palline sono estratte dall'urna con *reintroduzione* (ovvero la composizione dell'urna non cambia)

allora la variabile casuale k-dimensionale $Y = (Y_1, \ldots, Y_k)$ che descrive il numero di palline estratte del primo colore,del secondo colore,..., è una Multinomiale$(n, (\pi_1, \ldots, \pi_k))$.

Probabilità 27 Un esperimento su una variabile casuale multinomiale ci fornisce un vettore di k interi.
Ad esempio, se $k = 3$, i colori sono {blu, viola, arancione} e $n = 10$ un possibile risultato sperimentale potrebbe essere $y = (3, 1, 6)$ e ci indicherebbe che nelle 10 estrazioni dall'urna abbiamo ottenuto, in ordine qualsiasi, 3 palline blu, 1 pallina viola e 6 palline arancione.

Probabilità 28 Si osservi che, in generale,

$$Y_i \in \{0, \ldots, n\} \ \ (i = 1, \ldots, k) \text{ e } Y_1 + \cdots + Y_k = n.$$

Probabilità 29 Si osservi inoltre che, "per costruzione",

$$Y_i \sim Bi(n, \pi_i) \ \ (i = 1, \ldots, k)$$

o che più in generale se $i_1, \ldots, i_h$ sono h interi, maggiori di zero, minori o uguali a k e distinti tra loro ($h \in \{1, \ldots, k\}$), allora

$$Y = Y_{i_1} + \cdots + Y_{i_h} \sim Bi(n, \pi_{i_1} + \cdots + \pi_{i_h}).$$

Media e varianza di "combinazioni lineari" di variabili casuali

Probabilità 30 Una proprietà di base del valore atteso, conseguenza della definizione, è la sua *linearità*:

(i) se Y è una variabile casuale con valore atteso finito e ν una costante reale allora, $\mathrm{E}\{\nu Y\} = \nu \mathrm{E}\{Y\}$;

(ii) se Y_1 e Y_2 sono due variabili casuali ed esistono $\mathrm{E}\{Y_1\}$ e $\mathrm{E}\{Y_2\}$ allora $\mathrm{E}\{Y_1 + Y_2\} = \mathrm{E}\{Y_1\} + \mathrm{E}\{Y_2\}$.

Probabilità 31 Sia Y una variabile casuale e si supponga che esista $\mathrm{E}\{Y\}$. Allora, se ν_0 e ν_1 sono due costanti reali

$$\mathrm{E}\{\nu_0 + \nu_1 Y\} = \nu_0 + \nu_1 \mathrm{E}\{Y\}.$$

Per dimostrarla si usi [Probalità 30] ponendo Y_1 uguale ad una variabile casuale degenere tale che $\mathrm{P}(Y_1 = \nu_0) = 1$ e $Y_2 = \nu_1 Y$.

Probabilità 32 Sia Y una variabile casuale e si supponga che esista $\mathrm{var}\{Y\}$. Allora, se ν_0 e ν_1 sono due costanti reali

$$\mathrm{var}\{\nu_0 + \nu_1 Y\} = \nu_1^2 \mathrm{var}\{Y\}.$$

Infatti,

$$\begin{aligned} var\{\nu_0 + \nu_1 Y\} &= E\left\{[\nu_0 + \nu_1 Y - E\{\nu_0 + \nu_1 Y\}]^2\right\} = \\ &= E\left\{[\nu_0 + \nu_1 Y - (\nu_0 + \nu_1 E\{Y\})]^2\right\} = \\ &= E\left\{\nu_1^2 (Y - E\{Y\})^2\right\} = \\ &= \nu_1^2 E\left\{(Y - E\{Y\})^2\right\} = \\ &= \nu_1^2 var\{Y\}) \end{aligned}$$

Probabilità 33 Siano Y_1 e Y_2 due variabili casuali ambedue con valore atteso finito. Allora, se ν_1 e ν_2 sono due costanti reali

$$\mathrm{E}\{\nu_1 Y_1 + \nu_2 Y_2\} = \nu_1 \mathrm{E}\{Y_1\} + \nu_2 \mathrm{E}\{Y_2\}.$$

E' nient'altro che una formulazione alternativa di [Probalità 30].

Probabilità 34 Siano Y_1 e Y_2 due variabili casuali tali che $\mathrm{var}\{Y_1\}$, $\mathrm{var}\{Y_2\}$ e $\mathrm{cov}\{Y_1, Y_2\}$ esistono finiti[4]. Allora, se ν_1 e ν_2 sono due costanti reali

$$\mathrm{var}\{\nu_1 Y_1 + \nu_2 Y_2\} = \nu_1^2 \mathrm{var}\{Y_1\} + \nu_2^2 \mathrm{var}\{Y_2\} + 2\nu_1 \nu_2 \mathrm{cov}\{Y_1, Y_2\}.$$

Infatti

$$\begin{aligned} var\{\nu_1 Y_1 + \nu_2 Y_2\} &= E\left\{[\nu_1 Y_1 + \nu_2 Y_2 - E\{\nu_1 Y_1 + \nu_2 Y_2\}]^2\right\} = \\ &= E\left\{[\nu_1 Y_1 + \nu_2 Y_2 - (\nu_1 E\{Y_1\} + \nu_2 E\{Y_2\})]^2\right\} = \\ &= E\left\{[\nu_1 (Y_1 - E\{Y_1\}) + \nu_2 (Y_2 - E\{Y_2\}))]^2\right\} = \\ &= \nu_1^2 E\left\{(Y_1 - E\{Y_1\})^2\right\} + \nu_2^2 E\left\{(Y_2 - E\{Y_2\})^2\right\} + \\ &\quad + 2\nu_1 \nu_2 E\{(Y_1 - E\{Y_1\})(Y_2 - E\{Y_2\})\} = \\ &= \nu_1^2 var\{Y\} + \nu_2^2 var\{Y_2\} + 2\nu_1 \nu_2 cov\{Y_1, Y_2\}. \end{aligned}$$

Probabilità 35 Siano Y_1 e Y_2 due variabili casuali con medie e varianze finite e incorrelate tra di loro ($\mathrm{cov}\{Y_1, Y_2\} = 0$). Allora

$$\mathrm{var}\{Y_1 + Y_2\} = \mathrm{var}\{Y_1 - Y_2\} = \mathrm{var}\{Y_1\} + \mathrm{var}\{Y_2\}.$$

Si tratta di due casi particolari di [Probalità 34].

Probabilità 36 L'indipendenza implica l'incorrelazione. Quindi [Probalità 35] vale anche quando Y_1 e Y_2 sono indipendenti (purchè ovviamente $\mathrm{var}\{Y_1\}$ e $\mathrm{var}\{Y_2\}$ esistano).

Probabilità 37 *A proposito di [Probalità 35].* Capita di trovare utilizzata nei compiti d'esame la seguente "versione" di [Probalità 35]:

$$\mathrm{var}\{Y_1 - Y_2\} = \mathrm{var}\{Y_1\} - \mathrm{var}\{Y_2\}.$$

La conseguenza è un compito non sufficente qualsiasi altra cosa lo studente faccia. Nei casi in cui $\mathrm{var}\{Y_1\} < \mathrm{var}\{Y_2\}$ si possono anche sentire a Santa Caterina delle urla "poco divertite" del docente che sta correggendo il compito.

[4] in realtà sarebbe possibile dimostrare che l'esistenza delle varianze implica l'esistenza della covarianza.

Probabilità 38 Siano $Y_1, \ldots, Y_n$ n variabili casuali tutte di media finita e si ponga ν_0,

$$Y_L = \nu_0 + \nu_1 Y_1 + \cdots + \nu_n Y_n = \nu_0 + \sum_{i=1}^{n} \nu_i Y_i$$

dove $\nu_0, \ldots, \nu_n$ sono $n+1$ costanti reali. Allora,

$$E\{Y_L\} = \nu_0 + \nu_1 E\{Y_1\} + \cdots + \nu_n E\{Y_n\} = \nu_0 + \sum_{i=1}^{n} \nu_i E\{Y_i\}.$$

Può essere ottenuta utilizzando [Probalità 31] e, iterativamente, [Probalità 33] di cui la formula appena data costituisce una generalizzazione.

Probabilità 39 Sia Y_L definito come in [Probalità 38]. Allora, se esistono anche $\mathrm{var}\{Y_1\}, \ldots, \mathrm{var}\{Y_n\}$,

$$\begin{aligned} \mathrm{var}\{Y_L\} &= \nu_1^2 \mathrm{var}\{Y_1\} + \cdots + \nu_n^2 \mathrm{var}\{Y_n\} + \\ &\quad + \nu_1\nu_2 \mathrm{cov}\{Y_1, Y_2\} + \cdots + \nu_{n-1}\nu_n \mathrm{cov}\{Y_{n-1}, Y_n\} = \\ &= \sum_{i}^{n} \nu_i^2 \mathrm{var}\{Y_i\} + \sum_{i \neq j} \nu_i \nu_j \mathrm{cov}\{Y_i, Y_j\} = \\ &= \sum_{i}^{n} \nu_i^2 \mathrm{var}\{Y_i\} + 2 \sum_{i<j} \nu_i \nu_j \mathrm{cov}\{Y_i, Y_j\}. \end{aligned}$$

dove $\sum_{i \neq j}$ indica la somma estesa a tutte le coppie di indici

$$(i, j) \in \{(i, j) \in N^2 : 1 \leq i \leq n, 1 \leq j \leq n, i \neq j\}$$

e in maniera analoga $\sum_{i<j}$ indica la somma estesa a tutte coppie di indici

$$(i, j) \in \{(i, j) \in N^2 : 1 \leq i \leq n, 1 \leq j \leq n, i < j\}.$$

La dimostrazione può essere ottenuta applicando [Probalità 32] e, iterativamente, [Probalità 34].

Media e varianza della media campionaria

Probabilità 40 Siano Y_1, Y_2,$\ldots$,Y_n n variabili casuali indipendenti e identicamente distribuite.
Si indichino con μ e σ^2 la media e la varianza comune (che supponiamo esistere). Quindi

$$\mu = E\{Y_1\} = \cdots = E\{Y_n\}$$

e

$$\sigma^2 = \mathrm{var}\{Y_1\} = \cdots = \mathrm{var}\{Y_n\}.$$

Sia

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

$\overline{Y}$ è la variabile casuale "media campionaria".
Allora

$$E(\overline{Y}) = \mu \text{ e } \mathrm{var}\left\{\overline{Y}\right\} = \frac{\sigma^2}{n}.$$

La dimostrazione è immediata utilizzando [Probalità 38] e [Probalità 39] se si tiene conto che l'indipendenza implica l'incorrelazione, ovvero che

$$Y_i \textit{ indipendente da } Y_j \;\Rightarrow\; \textit{cov}\{Y_i, Y_j\} = 0.$$

Distribuzione della media e della varianza campionaria nel caso di un campione estratto da una popolazione normale

Probabilità 41 Si supponga che $(Y_1, \ldots, Y_n)$ siano delle variabili casuali indipendenti e identicamente distribuite come una normale di media μ e varianza σ^2.
Si ponga

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ e } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Allora è possibile dimostrare che:

(i) la distribuzione di $\overline{Y}$ è normale di media μ e varianza σ^2/n, ovvero,

$$\overline{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right);$$

(ii) la distribuzione di $(n-1)S^2/\sigma^2$ è un χ^2 con $n-1$ gradi di libertà, ovvero,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1);$$

(iii) $\overline{Y}$ e S^2 sono stocasticamente indipendenti.

Probabilità 42 Utilizzando [Probalità 3], la parte riguardante la media campionaria dell'enunciato [Probalità 41] può anche essere scritta come

$$\frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma} \sim N(0,1).$$

Probabilità 43 Ricordando la definizione della t di Student [Probalità 14], è immediato far vedere che [Probalità 41] implica anche che

$$\frac{\sqrt{n}(\overline{Y} - \mu)}{S} \sim t(n-1).$$

Infatti

$$\begin{aligned}\frac{\sqrt{n}(\overline{Y} - \mu)}{S} &= \frac{\sqrt{n}(\overline{Y} - \mu)/\sigma}{\sqrt{((n-1)S^2/\sigma^2)/(n-1)}} = \\ &= \frac{N(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} \sim t(n-1)\end{aligned}$$

dove, nell'ultimo passaggio, oltre alla definizione della t *di Student, abbiamo utilizzato il fatto che* $\overline{Y}$ *e* S^2 *sono tra di loro indipendenti.*

Distribuzione delle medie e delle varianze campionarie e di alcune loro funzioni notevoli nel caso di due campioni estratti da popolazioni normali

Probabilità 44 Siano $Y_1, \ldots, Y_n$ delle variabili casuali indipendenti tra di loro e identicamente distribuite come una normale di media μ_y e varianza σ_y^2 e $X_1, \ldots, X_n$ delle variabili casuali indipendenti tra di loro e dalle "Y" e identicamente distribuite come una normale di media μ_x e varianza σ_x^2. Definiamo

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ e } S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

e in maniera analoga $\overline{X}$ e S_x^2.
Allora, ricordando che *"trasformate separate di variabili casuali indipendenti sono indipendenti"*[5], da [Probalità 41] segue che $\overline{Y}$, $\overline{X}$, S_y^2 e S_x^2 sono variabili casuali indipendenti tra loro tali che

$$\overline{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n}\right), \quad \frac{(n-1)S_y^2}{\sigma_y^2} \sim \chi(n-1),$$
$$\overline{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{m}\right), \quad \frac{(m-1)S_x^2}{\sigma_x^2} \sim \chi(m-1).$$

Quindi da [Probalità 4] e [Probalità 19] segue che

$$\overline{Y} - \overline{X} \sim N\left(\mu_y - \mu_x, \frac{\sigma_y^2}{n} + \frac{\sigma_x^2}{m}\right)$$

e

$$\frac{S_y^2/\sigma_y^2}{S_x^2/\sigma_x^2} \sim F(n-1, m-1).$$

[5] ovvero che se Z è una variabile casuale, eventualmente multidimensionale, e W è un'altra variabile casuale, eventualmente multidimensionale, indipendente da Z allora $f(Z)$ è indipendente da $g(W)$ per qualsiasi $f(\cdot)$ e $g(\cdot)$ per cui $f(Z)$ e $g(W)$ sono variabili casuali.

Probabilità 45 Nelle stessa situazione di [Probalità 44] si ipotizzi che

$$\sigma_y^2 = \sigma_x^2 = \sigma^2$$

ovvero che le "Y" e le "X" abbiano la stessa dispersione. Si ponga

$$S^2 = \frac{(n-1)S_Y^2 + (m-1)S_X^2}{n+m-2}.$$

Allora, da [Probalità 12], [Probalità 14] e [Probalità 44] segue che S^2 è una variabile casuale indipendente da $\overline{Y}$ e $\overline{X}$ e tale che

$$\frac{(n+m-2)S^2}{\sigma^2} \sim \chi^2(n+m-2).$$

Inoltre,

$$\frac{\overline{Y} - \overline{X} - (\mu - \eta)}{S\left(\frac{1}{n} + \frac{1}{m}\right)} \sim t(n+m-2).$$

Alcuni risultati asintotici

Probabilità 46 *Modi di convergenza* Sia $Y_1, Y_2, \ldots$, una successioni di variabili casuali, l una costante e Y_∞ una variabile casuale. Si dice che

1. la successione $\{Y_n\}$ converge *quasi certamente* o *con probabilità uno* a l se
$$P(\lim_{n\to\infty} Y_n = l) = 1;$$
2. la successione $\{Y_n\}$ converge in probabilità a l se, $\forall \epsilon > 0$
$$\lim_{n\to\infty} P(|Y_n - l| \leq \epsilon) = 1;$$
3. la successione $\{Y_n\}$ converge *in distribuzione* a Y_∞ se per ogni intervallo $[a, b]$
$$\lim_{n\to\infty} P(a \leq Y_n \leq b) = P(a \leq Y_\infty \leq b).$$

La convergenza quasi certa implica la convergenza in probabilità. Per questo motivo la prima è volte chiamata convergenza *forte* e la seconda *debole*.

Probabilità 47 Siano $Y_1, Y_2, \ldots$ una successione di variabili casuali e $f(\cdot)$ una funzione da R in R.
(i) Se Y_n converge in probabilità/quasi certamente alla costante l e $f(\cdot)$ è continua in l allora $f(Y_n)$ converge in probabilità/quasi certamente a $f(l)$.
(ii) Se Y_n converge in distribuzione alla variabile casuale Y_∞ e $f(\cdot)$ è continua, $f(Y_n)$ converge in distribuzione a $f(Y_\infty)$.
Esempio 1. Se Y_n converge in probabilità a 25, allora $\sqrt{Y_n}$ converge in probabilità a 5.
Esempio 2. Se Y_n converge ad una $N(0, 1)$, allora Y_n^2 converge ad un $\chi^2(1)$.

Probabilità 48 Siano $Y_1, Y_2, \ldots$ e $X_1, X_2, \ldots$ due successioni di variabili casuali convergenti in probabilità/quasi certamente rispettivamente a l e m. Sia inoltre $f(\cdot, \cdot)$ una funzione da R^2 a R continua in (l, m). Allora, $f(Y_n, X_n)$ converge in probabilità/quasi certamente a $f(l, m)$.
Quindi, ad esempio, le successioni $Y_n + X_n$, $Y_n - X_n$, $Y_n X_n$ e, se $m \neq 0$ $Y_n + X_n$ convergono a $l + m$, $l - m$, lm e l/m.

Probabilità 49 Siano $Y_1, Y_2, \ldots$ e $X_1, X_2, \ldots$ due successioni di variabili casuali la prima convergente in distribuzione a Y_∞ e la seconda in probabilità a m. Sia inoltre $f(\cdot, \cdot)$ una funzione da R^2 a R continua. Allora, $f(Y_n, X_n)$ converge in distribuzione a $f(Y_\infty, m)$.
Quindi, ad esempio, se Y_n converge in distribuzione ad una normale standard, allora $Y_n + X_n$, $Y_n - X_n$, $Y_n X_n$ e, se $m \neq 0$ $Y_n + X_n$ convergono in distribuzione rispettivamente ad una $N(m, 1)$, $N(-m, 1)$, $N(0, m^2)$ e $N(0, m^{-2})$.

Probabilità 50 *Legge forte dei grandi numeri.* Sia $Y_1, Y_2, \ldots$ una successione di variabili casuali indipendenti e identicamente distribuite tali che $E\{Y_1\}$, chiamiamolo μ, esista[6]. Allora

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

converge quasi certamente (e quindi anche in probabilità) a μ.

Probabilità 51 *Teorema del limite centrale.* Nella stessa situazione di [Probalità 50] se esiste anche $\sigma^2 = \mathrm{var}\{Y_1\}$ allora

$$\frac{\overline{Y}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

converge in distribuzione ad una normale standard.

Probabilità 52 *Limite centrale con varianza "stimata".* Nelle ipotesi del teorema del limite centrale [Probalità 51], si supponga di conoscere una successione $\hat{\sigma}_n$ convergente (almeno) in probabilità a σ.
Allora anche

$$\frac{\overline{Y} - \mu}{\frac{\hat{\sigma}_n}{\sqrt{n}}} \text{ converge in distribuzione ad una } N(0,1).$$

Infatti,

$$\frac{\overline{Y} - \mu}{\frac{\hat{\sigma}_n}{\sqrt{n}}} = \frac{\overline{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} \times \frac{\sigma}{\hat{\sigma}_n}$$

e quindi il risultato segue da [Probalità 49].

[6] essendo le "Y" identicamente distribuite ovviamente l'esistenza del valore atteso di Y_1 implica l'esistenza del valore atteso di tutte le "Y".

Probabilità 53 *Applicazione alla varianza campionaria.* Sia $Y_1, Y_2, \ldots$ una successione di variabili casuali indipendenti e identicamente distribuite con media μ e varianza σ^2 (che supponiamo esistere).
Poniamo

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ e } S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Per la legge forte dei grandi numeri [Probalità 50], $\overline{Y}_n$ converge con probabilità uno a μ.
S^2, vista la presenza in tutti gli addendi di $\overline{Y}$, non è però una somma di variabili casuali indipendenti. Quindi non possiamo applicare direttamente la legge forte dei grandi numeri. Però possiamo scrivere

$$S_n^2 = \frac{n}{n-1}(V_n^2 - D_n^2)$$

dove

$$V_n^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu)^2 \text{ e } D_n = \overline{Y}_n - \mu.$$

Osserviamo che

- $n/(n-1)$ è una successione numerica convergente a uno;
- $(Y_1 - \mu)^2, (Y_2 - \mu)^2, \ldots$ è una successione di variabili casuali indipendenti e identicamente distribuite di media σ^2; la legge forte dei grandi numeri ci garantisce quindi che V_n^2 converge quasi certamente a σ^2;
- D_n converge a zero con probabilità uno e per [Probalità 47] lo stesso quindi accade a D_n^2;
- quindi, applicando [Probalità 48], troviamo che S_n^2 converge con probabilità uno a σ^2;
- per [Probalità 47] anche che

$$S_n = \sqrt{S_n^2} \text{ converge con probabilità uno a } \sigma$$

Probabilità 54 *Applicazione alla binomiale.* Sia $X_1, X_n, \ldots$ una successione di variabili casuali indipendenti e identicamente distribuite come una $\mathrm{Bi}(1, \vartheta)$, $0 < \vartheta < 1$. Sappiamo che[7]

$$\mathrm{E}\{X_i\} = \vartheta \text{ e } \mathrm{var}\{X_i\} = \vartheta(1-\vartheta), \quad (i = 1, 2, \ldots)$$

ed inoltre che, per la stessa definizione di binomiale,

$$Y_n = \sum_{i=1}^{n} X_i \sim \mathrm{Bi}(n, \vartheta).$$

Poniamo

$$\hat{\vartheta}_n = \frac{Y_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

(i) Per la legge forte dei grandi numeri [Probalità 50], $\hat{\vartheta}_n$ converge con probabilità uno a ϑ.
(ii) Per il teorema del limite centrale [Probalità 51],

$$\frac{\hat{\vartheta}_n - \vartheta}{\sqrt{\frac{\vartheta(1-\vartheta)}{n}}}$$

converge in distribuzione ad una normale standard.
(iii) Combinando le due affermazioni appena viste e utilizzando [Probalità 47] e [Probalità 49] anche

$$\frac{\hat{\vartheta}_n - \vartheta}{\sqrt{\frac{\hat{\vartheta}(1-\hat{\vartheta})}{n}}}$$

converge in distribuzione ad una normale standard.

[7] [Probalità 24]

Indice analitico