Inferenza Statistica I
(lucidi a.a. 2001/2002)

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

6 giugno 2002

Indice

Unità A

Controllo di qualità in un impianto che produce lastre di metallo

Un primo esempio di inferenza statistica.

Stima della media, sua distribuzione campionaria, intervalli di confidenza e verifica d'ipotesi nel caso di un campione tratto da una v.c. normale di varianza nota.

Il problema ed i dati

Una industria metallurgica produce, tra l'altro, delle lastre di metallo con uno spessore nominale di 14mm. In realtà esiste una tolleranza di ±0,5mm, ovvero, una lastra è considerata soddisfacente, per quello che riguarda lo spessore, se

$$13{,}5 \leq \text{spessore} \leq 14{,}5. \qquad \text{(A.1)}$$

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 (A.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:

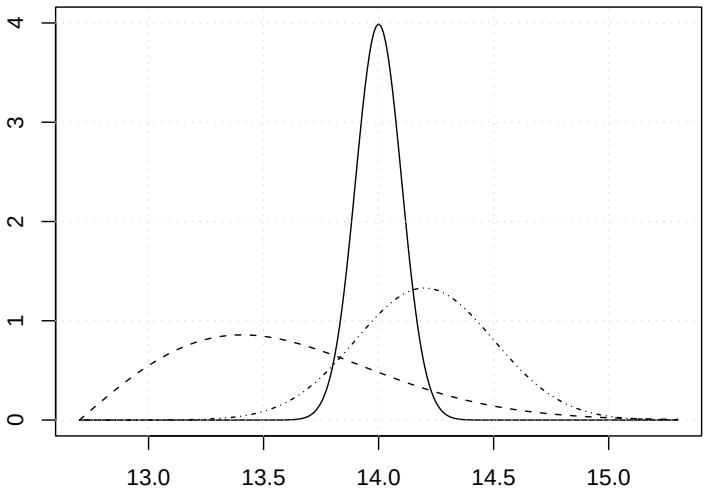
14,33 14,19 14,39 14,43 14,17.

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.
• All'inizio della lavorazione di una lastra (o all'inizio del turno,. . .) solo *Mago Merlino* sarebbe in grado di indovinarne esattamente lo spessore.
• Possiamo però pensare che lo spessore sia il risultato di un esperimento casuale e descriverne le caratteristiche utilizzando il calcolo della probabilità.
• In particolare, potremmo guardare agli spessori che, in un determinato momento, il processo "potrebbe produrre" come ad una variabile casuale continua con funzione di densità $f(\cdot)$.
• Il problema diventa allora quello di utilizzare i dati disponibili per dire se la densità $f(\cdot)$ assegna una eccessiva probabilità all'evento "lastra difettosa" (= lastra il cui spessore non soddisfa la (A.1)). Si veda la pagina seguente, per alcuni esempi.
• Se questo accade, e quindi se il processo sta, *almeno potenzialmente*, producendo "troppe" lastre difettose decidere di sospendere la produzione.

Tre possibili situazioni



La densità disegnata con una linea continua indica una situazione soddisfacente: la probabilità di ottenere una lastra difettosa (spessore inferiore a 13,5mm o maggiore di 14,5mm) è nulla (o quasi). Le altre due raccontano storie diverse: l'impianto sta producendo una frazione non piccola di lastre o troppo sottili o troppo spesse.

Informazioni aggiuntive sul processo

Cercare di stimare l'intera funzione di densità utilizzando solo le nostre 5 osservazioni sembra un'operazione eccessivamente avventurosa.

Fortunamente nel caso in esame esistono delle informazioni aggiuntive. Infatti, precedentemente, le caratteristiche del processo sono state studiate raccogliendo alcune migliaia di misurazioni per alcune decine di turni.

Indicato 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, le principali conclusioni delle analisi condotte, sono:
• non esiste nessun tipo di dipendenza tra le Y_i;
• tutte le Y_i hanno la stessa distribuzione di probabilità;
• questa distribuzione comune è ben approssimata da una normale 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 conclusioni descritte nel lucido 6.

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

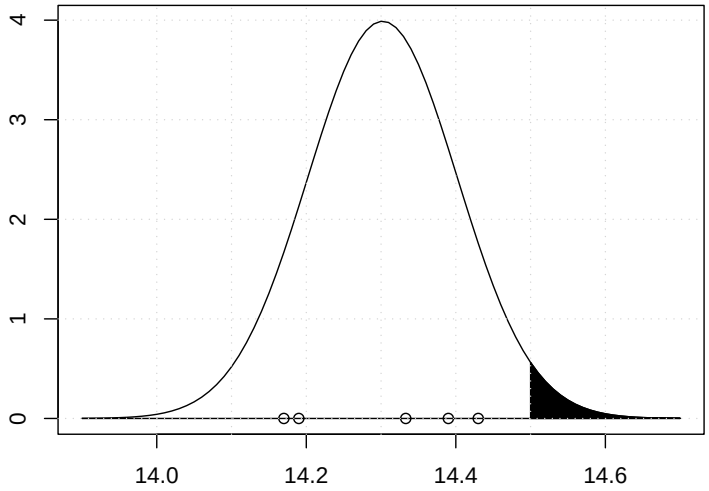
Le informazioni aggiuntive ci portano a considerare le 5 misure dello spessore come 5 determinazioni indipendenti "estratte" da una stessa variabile casuale Gaussiana di media μ ignota e varianza nota ed uguale a 0,01. Un altra maniera di descrivere la situazione consiste nel dire che siamo in presenza di **determinazioni indipendenti ed identicamente distribuite (abbreviazione i.i.d.)** tratte da una variabile normale. . . .

La funzione di densità dello spessore è quindi "quasi" nota. Ci manca solo la media. Sembra al proposito "ragionevole" utilizzare la media delle osservazioni come **"stima"** della vera media μ, ovvero porre

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

Densità stimata

Il grafico mostra la densità di una normale di media 14,302 e varianza 0,1. L'area evidenziata rappresenta la probabilità (stimata) di produrre una lastra troppo spessa. La probabilità (stimata) di produrre una lastra troppo sottile è praticamente nulla. I "cerchietti" sull'asse delle x indicano le osservazioni.



Stima della "difettosità"

Due eventi particolarmente importanti nel presente contesto sono

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

dove Y indica la variabile casuale che descrive lo spessore. Ovviamente sia $P(A)$ che $P(B)$ sono funzione di μ. In particolare, ricordando che[1]

$$\text{se } Y \sim \mathrm{N}(\mu, \sigma^2) \text{ allora } (Y-\mu)/\sigma \sim N(0,1)$$

le probabilità di questi eventi possono agevolmente essere calcolate dalla funzione di ripartizione di una normale standard. In particolare,

$$\begin{aligned} P(A) &= P(Y < 13{,}5) = \\ &= P\left(\frac{Y-\mu}{0{,}1} < \frac{13.5-\mu}{0{,}1}\right) = \\ &= P\left(\mathrm{N}(0,1) < \frac{13.5-\mu}{0{,}1}\right) = \\ &= P\left(\mathrm{N}(0,1) \le \frac{13.5-\mu}{0{,}1}\right). \end{aligned}$$

[1] Ci si ricordi che $\sim$ si legge "si distribuisce come"

Possiamo quindi scrivere

$$P(A) = \Phi\left(\frac{13.5 - \mu}{0{,}1}\right)$$

dove con $\Phi(\cdot)$ abbiamo indicato la funzione di ripartizione di una $\mathrm{N}(0,1)$. Si noti che abbiamo usato il fatto che, se Y è una variabile casuale continua, allora $P(Y = y) = 0$ per qualsivoglia valore y. Per l'altra probabilità troviamo

$$\begin{aligned} P(B) &= P(Y > 14{,}5) = \\ &= 1 - P(Y \leq 14{,}5) = \\ &= 1 - P\left(\frac{Y - \mu}{0{,}1} \leq \frac{14.5 - \mu}{0{,}1}\right) = \\ &= 1 - P\left(\mathrm{N}(0,1) \leq \frac{14.5 - \mu}{0{,}1}\right). \end{aligned}$$

ovvero

$$P(B) = 1 - \Phi\left(\frac{14.5 - \mu}{0{,}1}\right).$$

Possiamo ottenere delle stime di queste due quantità sostituendo a μ, che è ignoto, la sua stima $\overline{y}$. Nel caso in esame

$$\hat{P}(A) = \Phi\left(\frac{13{,}5 - 14{,}302}{0{,}01}\right) = \Phi(-8{,}02) \approx 0$$

e

$$\hat{P}(B) = 1 - \Phi\left(\frac{14{,}5 - 14{,}302}{0{,}01}\right) = 1 - \Phi(1{,}98) \approx 0{,}024$$

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

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 soprattutto interessante da un punto di vista pratico) chiedere "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 7 e spieghi perchè.

La distribuzione della media campionaria

• La media campionaria, $\overline{y}$, può essere vista come una determinazione di una variabile casuale. Infatti se i dati da cui è calcolata sono il risultato di un esperimento casuale anche $\overline{y}$ ovviamente lo è.
• Indichiamo con $\overline{Y}$ la variabile casuale. Nelle ipotesi che stiamo facendo (normalità,. . .) la distribuzione di $\overline{Y}$ discende dal seguente risultato:

se $Y_1, \ldots, Y_n$ sono variabili casuali normali indipendenti *tra loro e se $a_0, \ldots, a_n$ sono delle costanti reali qualsiasi, allora*

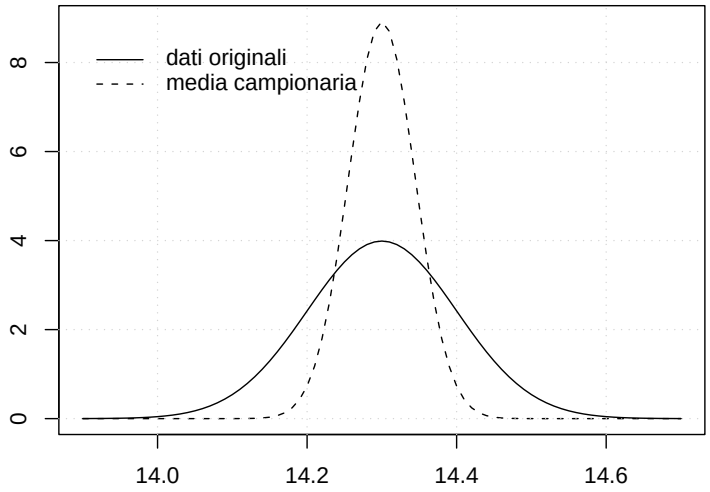
$$a_0 + \sum_{i=1}^{n} a_i Y_i \sim N(a_0 + \sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$$

dove μ_i e σ_i^2 indicano rispettivamente la media e la varianza di Y_i.

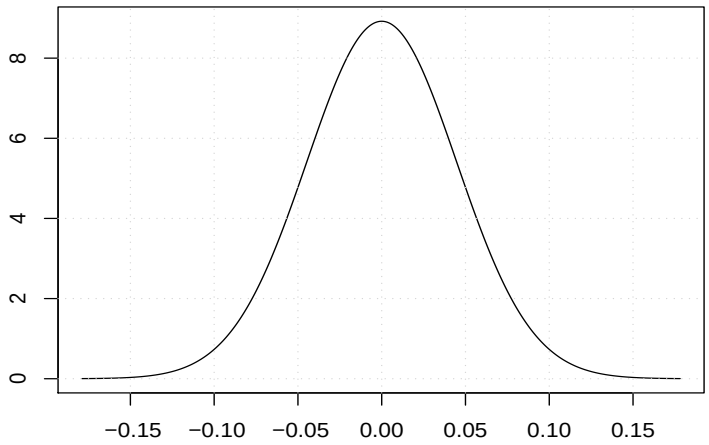
• Quindi, se le n variabili casuali normali hanno tutte la stessa media e varianza (diciamo μ e σ^2) allora (lo studente lo dimostri)

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

• Si osservi che la distribuzione e la media sono quelle delle osservazioni originarie (ovvero le Y_i sono normali e $\overline{Y}$ è normale, le Y_i hanno media μ e $\overline{Y}$ ha media μ) e che la varianza della media campionaria è la varianza delle osservazioni originarie divisa per n (ovvero se il numero delle osservazioni è maggiore di 1 allora la media campionaria è meno variabile delle osservazioni originarie). Il grafico mostra le due funzioni di densità nel caso in cui $\mu = 14{,}3$ e $\sigma = 0{,}1$.



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)

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

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,

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

Si osservi ora che l'affermazione precedente può essere anche scritta come
"la probabilità che l'intervallo $[\overline{y} - 0{,}1, \overline{y} + 0{,}1]$*, ovvero, l'intervallo* [14,202 , 14,402]*, includa la vera media* μ *è* 0,987".
Infatti

$$\begin{aligned} &P(\overline{Y} - 0{,}1 < \mu < \overline{Y} + 0{,}1) = \\ &= P(-0{,}1 < \mu - \overline{Y} < 0{,}1) = P(|\overline{Y} - \mu| < 0{,}1) \end{aligned}$$

In generale un intervallo che contiene il vero valore di un parametro ignoto con probabilità $1-\alpha$ viene chiamato un **intervallo di confidenza di livello** $1-\alpha$.

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).
In questo caso i passi da seguire sono i seguenti:
• Ovviamente fissiamo un valore per $1-\alpha$.
• Determiniamo o utilizzando un programma o le tavole della normale standard, il percentile $1-\alpha/2$ 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 $P(N(0,1) \leq -z_{1-\alpha/2}) = \alpha/2$. E quindi $P(|N(0,1)| \leq z_{1-\alpha/2}) = 1-\alpha$. Si veda il grafico a pagina 22.
• Ricordando che $\overline{Y} \sim \mathrm{N}(\mu, \sigma^2/n)$, possiamo allora scrivere

$$P\left(\left|(\overline{Y}-\mu)/\sqrt{\sigma^2/n}\right| \leq z_{1-\alpha/2}\right) = 1-\alpha.$$

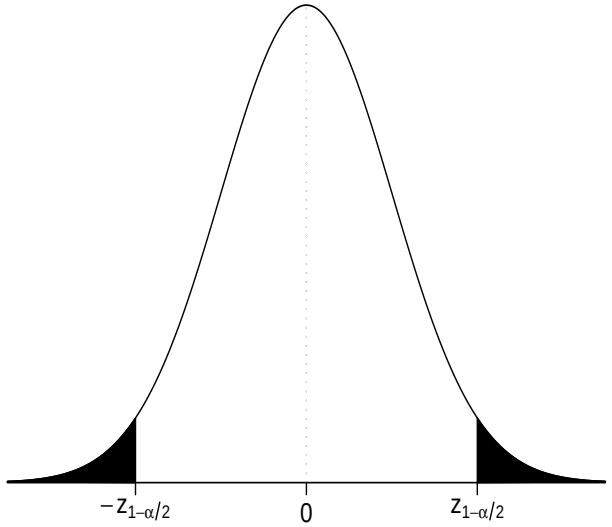
Con semplici passaggi otteniamo

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

L'intervallo

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

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



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

Esempio

Supponiamo di volere un intervallo di confidenza di livello 0,95. Allora, $\alpha = 0{,}05$ e $z_{0,975} = 1{,}96$. Quindi l'intervallo risulta essere uguale a

$$14{,}302 \pm \frac{1{,}96 \times 0{,}1}{\sqrt{5}}$$

ovvero [14,21, 14,39].

Precisione nella stima della difettosità

Ricordiamoci che abbiamo ottenuto la formula

$$P(\{\text{lastra troppo "alta"}\}) = \pi(\mu) = 1 - \Phi\left(\frac{14{,}5 - \mu}{0.01}\right)$$

dove con l'introduzione della nuova notazione $\pi(\mu)$ vogliamo enfatizzare il fatto che abbiamo un valore della probabilità di produrre una lastra troppo "alta" per ogni valore della media.

E' facile verificare che $\pi(\mu)$ è una funzione monotona crescente (ci si ricordi che $\Phi(y)$ è crescente in y). Quindi, l'evento

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

coincide con l'evento

$$\left\{\overline{y} : \overline{y} - \frac{z_{1-\alpha/2}\sigma}{n} \leq \mu \leq \overline{y} + \frac{z_{1-\alpha/2}\sigma}{n}\right\}.$$

Ma allora i due eventi hanno la stessa probabilità e quindi

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

è un intervallo di confidenza di dimensione $1 - \alpha$ per $\pi(\mu)$. Si osservi che ci basta trasformare gli estremi di un intervallo di confidenza per μ.

Usando $\alpha = 0.05$, l'intervallo che otteniamo è [0,002 , 0,135]. Ovvero se teniamo conto dell'errore di stima i risultati ottenuti (gli spessori misurati) sono compatibili anche senza fare riferimento ad eventi con probabilità particolarmente piccola con una difettosità superiore al 10%. La conclusione è che sembra "prudente" bloccare la produzione[2].

[2]Si tenga tra l'altro conto che $\pi(14) \approx 2/10^6$, ovvero, che l'impianto, quando ben "tarato", può produrre un numero di lastre difettose realmente piccolo

Un approccio diverso

• 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 di descrivere l'approccio appena suggerito consiste nel dire che all'inizio di ogni turno vogliamo utilizzare i dati per decidere tra le seguenti due ipotesi:

$$H_0 : \mu = 14mm$$

e

$$H_1 : \mu \neq 14mm.$$

L'interpretazione delle due ipotesi è (ovviamente):

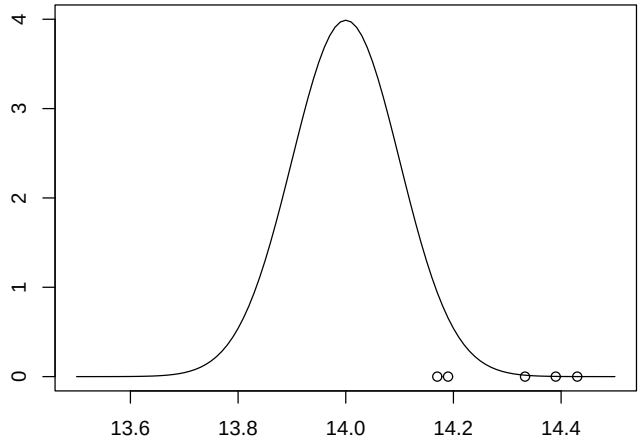
H_0 : l'impianto produce al meglio

e

H_1 : l'impianto ha iniziato a "sregolarsi".

Problemi di scelta tra due (o più) alternative sono, in statistica, chiamati problemi di **verifica di ipotesi**. Le ipotesi (quando sono due) vengono spesso indicate come **ipotesi nulla** ed **ipotesi alternativa**. Lo "strumento" utilizzato per affrontare i problemi di verifica di ipotesi (ovvero la procedura che si segue per far "votare" i dati a favore o di H_0 o di H_1, ovvero per decidere quale ipotesi **accettare** o **rifiutare**) viene chiamato **test statistico**.

Analisi grafica



La figura mostra la densità 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 9) che sembrano "più compatibili" con i dati. Quindi, i dati suggeriscono di rifiutare H_0. Sfortunatamente, una analisi grafica del tipo descritto è possibile solo nelle situazioni più semplici.

Un test statistico

• Volendo definire una procedura "analitica" per scegliere tra le due ipotesi, sembra ragionevole basarsi sulla differenza tra la media stimata, $\overline{y}$, e la media ipotizzata da H_0, 14.

• Ad esempio, potremmo pensare di usare una "regola" del tipo

$$-h \leq \frac{\sqrt{n}(\overline{y} - 14)}{\sigma} \leq h$$

si → accettiamo H_0

no → rifiutiamo H_0

Si osservi che abbiamo diviso la differenza per lo scarto quadratico medio della media campionaria. Ovviamente, trattandosi nel nostro caso di una costante nota ($n = 5$ e $\sigma^2 = 0.1$) 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

$$P(\text{accettare } H_0 \text{ quando } H_0 \text{ è vera}) = 1 \qquad \text{(A.2)}$$

ovvero, che

$$P(-h \leq \sqrt{n}(\overline{y} - 14)/\sigma \leq h \text{ quando } \mu = 14) = 1. \qquad \text{(A.3)}$$

Ora, se $\mu = 14$, $\sqrt{n}(\overline{y} - 14)/\sigma$ è una determinazione di una normale standard (lo studente spieghi perchè). Quindi, la (A.3) è equivalente a

$$P(-h \leq \mathrm{N}(0,1) \leq h) = 1 \qquad \text{(A.4)}$$

e la (A.4) mostra che l'unico valore di h che garantisce la (A.2) è $h = +\infty$ (ci si ricordi che la densità di una normale è diversa da zero su tutta la retta reale).

L'utilizzo di una soglia infinita non è però molto sensato. Infatti se poniamo $h = +\infty$ non rifiuteremmo mai H_0. In altre parole, se insistiamo sulla (A.2) finiamo con una "regola" per cui

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

Un compromesso

Chiedere che la (A.2) sia *esattamente* vera ci porta a determinare un valore di h inaccettabile. Sarebbe però inacettabile anche una situazione in cui, ad esempio,

$$P(\text{accettare } \mathrm{H}_0 \text{ quando } \mathrm{H}_0 \text{ è vera}) = 0{,}1$$

ovvero, una situazione in cui la (A.2) è pesantemente violata. Infatti, in questo caso, il test *sbaglierebbe* 9 volte su 10 quando l'ipotesi nulla è vera. E anche questo sembra poco sensato.

Non ci rimane quindi che considerare il caso in cui la (A.2) è approssimativamente (ma non esattamente) rispettata, ovvero, in cui

$$P(\text{accettare } \mathrm{H}_0 \text{ quando } \mathrm{H}_0 \text{ è vera}) = 1 - \alpha \quad \text{(A.5)}$$

per un valore "piccolo" di α. La (A.5) può essere riscritta nella forma

$$P(-h \leq \mathrm{N}(0,1) \leq h) = 1 - \alpha \quad \text{(A.6)}$$

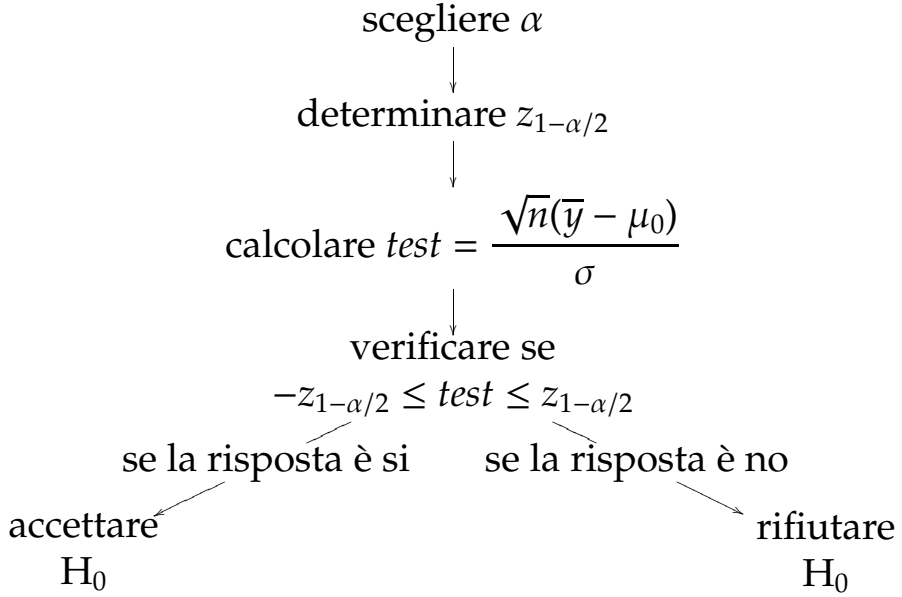
ed è facile verificare (lo studente si aiuti con il grafico a pagina 22) che la soluzione in h della (A.6) è $h = z_{1-\alpha/2}$ dove con z_p abbiamo indicato il percentile p-simo di una normale di media zero e varianza uno, ovvero il numero per cui $\Phi(z_p) = p$.

Sintesi della procedura delineata

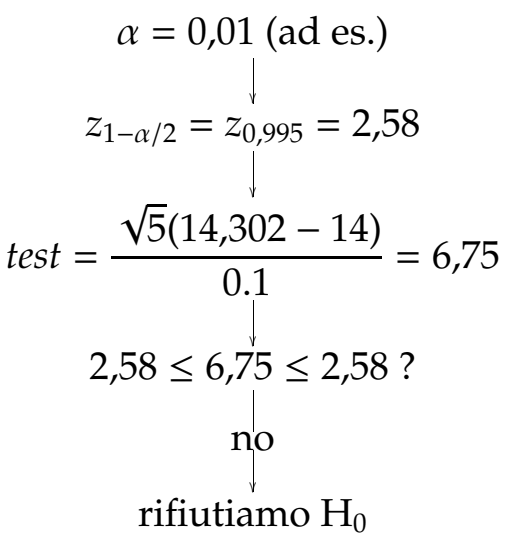
In definitiva, per verificare un sistema d'ipotesi del tipo

$$\begin{cases} \mathrm{H}_0 : \mu = \mu_0 \\ \mathrm{H}_1 : \mu \neq \mu_0 \end{cases}$$

siamo arrivati alla seguente procedura:



Nel caso in esame



Struttura di un test

Quanto abbiamo fatto nel caso in esame illustra fedelmente la struttura di un test statistico. E' quindi conveniente "ricapitolare" la costruzione:

1. Abbiamo definito una **statistica**, ovvero una funzione dei dati, scelta in maniera tale che i valori che ci aspettiamo che la statistica assuma quando H_0 e H_1 sono vere siano "tendenzialmente" diversi. Nell'ambito della teoria dei test, la statistica scelta viene chiamata, guarda caso, **statistica test**. Nell'esempio considerato, la statistica utilizzata è

$$T(y_1, \ldots, y_5) = \frac{\sqrt{n}(\overline{y} - \mu_0)}{\sigma}$$

   e l'abbiamo scelta poichè ci aspettiamo che

ipotesi "vera"	valori assunti dalla statistica test
H_0	intorno allo zero
H_1	lontani dallo zero

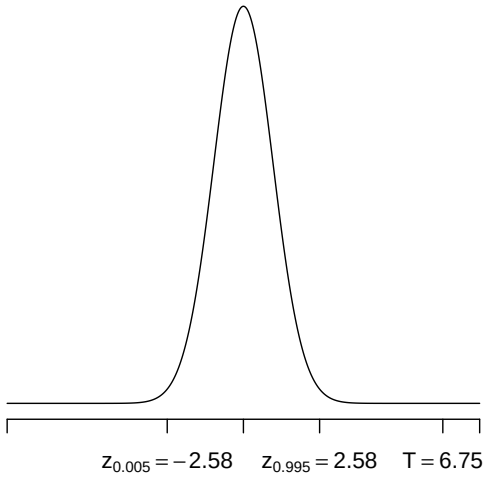
2. L'idea euristica di "la statistica test assume differenti valori sotto H_0 e H_1" si manifesta e concretizza da un punto di vista formale nell'osservare che T ha una diversa distribuzione di probabilità nei due casi. Ad esempio, nel caso in esame, se μ è la vera media degli spessori allora (lo studente lo dimostri utilizzando i risultati di pagina 14)

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

   ovvero, solo sotto H_0, $T \sim \mathrm{N}(0, 1)$ mentre, quando è vera H_1 la distribuzione è spostata o verso destra o verso sinistra (a seconda del segno di $\mu - \mu_0$).

3. A questo punto per decidere se H_0 doveva essere accettata o rifiutata abbiamo essenzialmente "confrontato" il valore osservato della statistica, ovvero il valore di T calcolato dai dati, con la distribuzione sotto H_0 (si veda lucido seguente). Poichè il valore osservato della statistica era "troppo estremo" (ovvero, troppo poco probabile) abbiamo deciso di rifiutare H_0. In particolare, si osservi che, desiderando una regola precisa, nella procedura operativa descritta dall'albero a pagina 33 abbiamo convenuto che "troppo estremo" significa $|T| > z_{1-\alpha/2}$ per qualche pre-scelto (e non troppo grande) valore di α.

Distribuzione sotto H_0 e valore osservato della statistica test



Il valore osservato (6,75) non sembra essere stato generato dalla distribuzione disegnata. Quindi rifiutiamo H_0.

Si noti la somiglianza con quanto fatto a pagina 28. Solamente qui usiamo la statistica test e non le osservazioni.

Esistono due tipi di errore

- Si osservi che in un problema di verifica d'ipotesi esistono due possibili modi con cui possiamo sbagliare. Infatti può capitare di:

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

- Ovviamente

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

  Quindi, costruire un test che soddisfa la (A.5) equivale ad utilizzare un test in cui la probabilità di commettere un errore di 1° tipo sia α.

- Si noti viceversa come nella costruzione delineata fino a questo punto la probabilità di commettere un errore di 2° tipo non è stata esplicitamente considerata (con la sola eccezione di pagina 30 il cui contenuto può essere parafrasato come "se vogliamo un test in cui la probabilità di errore di primo tipo sia nulla finiamo per costruire un test in cui la probabilità di errore di secondo tipo è uno".)

- Il motivo per cui ci si preoccupa di più degli errori di 1° tipo è che spesso la domanda a cui si vuole rispondere con un test statistico è

  A. *S*ono i dati sperimentali compatibili con H_0?

  più che

  B. *Q*uale tra H_0 e H_1 è vera?

  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. Lasciamo a corsi più avanzati il mostrare come. Ci limitiamo a menzionare che nel caso in esame il valore di n usato (ovvero 5) era stato scelto dall'impresa proprio sulla base di considerazioni di questo tipo.

Unità B

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 e (ii) il secondo che presentava solo bacelli gialli. 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 queste informazioni per fare delle affermazioni su

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

Un possibile modello

• Indichiamo con y il numero di piante con bacello verde e con n in numero totale delle piante di 2° generazione. Nel caso dell'esperimento descritto $y = 39$ e $n = 56$.

• Abbiamo almeno due questioni da discutere. La prima è se esiste effettivamente un qualche spazio di probabilità in cui ϑ è definito. La seconda è che la relazione esiste tra ϑ ed i risultati sperimentali (y, n) (altrimenti, non possiamo pensare di utilizzare i secondi per fare delle affermazioni su ϑ)

• 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) e (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.

• 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". Se accettiamo questo, allora

$$y \sim \text{Bi}(n, \vartheta) \qquad \text{(B.1)}$$

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 (B.1) è cruciale perchè precisa la relazione tra quello che conosciamo (y e n) e quello che vogliamo conoscere (ϑ).

Stima di ϑ

Lo stimatore più "naturale" (forse l'unico "naturale" nel senso che qualsiasi altra scelta sembra "innaturale") per ϑ è

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

ovvero la proporzione di piante "verdi" nei dati. Nel caso dell'esperimento di Mendel, $\vartheta = 39/56 \approx 0{,}70$.

Ovviamente, se y è una variabile casuale anche $\hat{\vartheta}$ è una variabile casuale. Lo studio della sua distribuzione è importante perchè permette di acquisire una idea sulla dimensione dell'errore di stima (come abbiamo già visto per la media della distribuzione normale nell'unità A).

La distribuzione *esatta* di $\hat{\vartheta}$ è facile da determinare. Infatti, $\hat{\vartheta} \in \Theta_n = \{0/n, 1/n, \ldots, n/n\}$ e, per qualsivoglia $a \in \Theta_n$, risulta

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

Approssimazione normale

Il fatto che la distribuzione esatta sia facile da determinare non implica che sia anche facile da maneggiare.

La maniera più "rapida" per determinare intervalli di confidenza e test si basa sull'approssimazione normale alla binomiale.

Il risultato di partenza è costituito dal fatto che per n non troppo piccolo la distribuzione di

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

è approssimabile con quella di una normale standard nel senso che per ogni intervallo della retta reale $[a, b]$

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

Si ritiene generalmente che l'approssimazione normale "funzioni almeno decorosamente" quando sia $n\vartheta$ che $n(1-\vartheta)$ sono più grandi di 5.

Distribuzione (approssimata) dell'errore di stima

Se $(\hat{\vartheta} - \vartheta)/\sqrt{\vartheta(1-\vartheta)/n}$ è approssimativamente una normale standard allora, sempre approssimativamente,

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

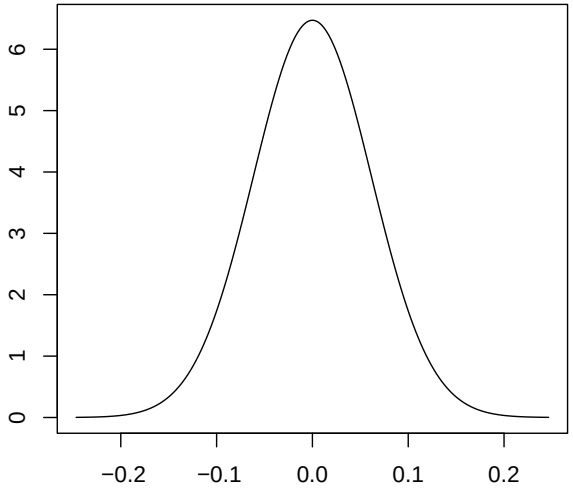
Si osservi che, a differenza di quanto accadeva all'errore di stima nell'unità A, questa distribuzione, oltre ad essere approssimata è anche parzialmente ignota. Infatti, la varianza della distribuzione dipende dal vero valore di ϑ.

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 commesso/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 ±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 (perchè?)

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

dove, al solito, con z_p indichiamo il quantile p-simo di una normale standard. La (B.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.

Con i dati di Mendel

Supponiamo di voler calcolare un intervallo di confidenza di livello 0,9.

Allora, $\alpha = 0{,}1$, $1 - \alpha/2 = 0{,}95$. 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

$$\sqrt{\frac{0{,}7 \times 0{,}3}{56}} \approx 0{,}062.$$

Quindi, la semi-ampiezza dell'intervallo richiesto è $1{,}65 \times 0{,}062 = 0{,}102$. Perciò l'intervallo stesso è

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

Per Mendel ϑ vale 0,75

L'idea di base su 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

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

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

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

Verifica dell'ipotesi di Mendel

Per quanto detto, 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 $\vartheta_0 = 0{,}75$.

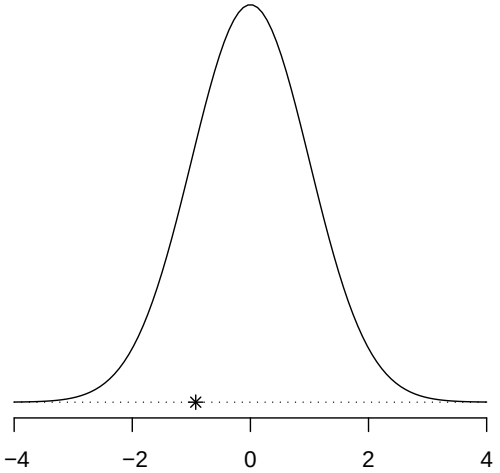
Volendo utilizzare un test statistico sembra ragionevole basare la decisione sulla distanza tra la stima di ϑ calcolata dai dati e il valore per il parametro previsto sotto H_0. Una possibile *statistica test* è quindi [1]

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

Se l'ipotesi nulla è vera, per quanto ricordato a pagina γ-50, 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.

[1] 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 se $|T|$ è sufficientemente piccolo, ovvero usare una regola del tipo "accetto se $|T| \leq h$" e (b) fissare h chiedendo che

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

per qualche valore prefissato e non troppo grande di α. Ricordando che T è approssimativamente distribuito come una normale standard, possiamo concludere che scegliendo $h = z_{1-\alpha/2}$ otteniamo una regola che almeno approssimativamente soddisfa la (B.3). Quindi, a parte per la statistica test che è diversa arriviamo ad una procedura "accetto/rifiuto" la cui meccanica è quella dell'unità A.

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

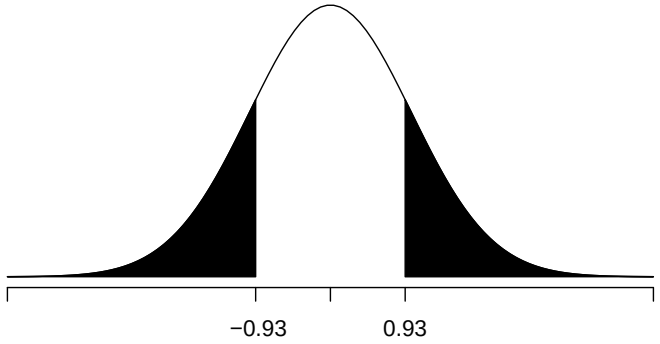
> . . . delle 56 piante della 2° generazione 39 (70%) mostravano un bacello verde (p = 0,35). . .

Quel "p = . . ." 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

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

Rappresentazione grafica nel caso in esame



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 (ovviamente, è una probabilità!) 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 "non più vicini" di quello osservato, ovvero, quello osservato è uno dei "più vicini possibili".

Unità C

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 la giuria effettiva veniva poi 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 costituiti dal risultato dell'estrazione (102 donne su 350 estratti).

Cerchiamo di specificare meglio l'ipotesi nulla. Ovvero cerchiamo di capire quale meccanismo probabilistico prevede la legge.

Sia N il numero degli elegibili. La legge prevede che si debba estrarre un primo individuo assegnando uguale probabilità a tutti gli elegibili. Poi che si debba estrarre un secondo individuo tra i restanti $N-1$ assegnando anche questa volta uguale probabilità. E così via.

Indichiamo con D il numero di donne tra gli elegibili. Quindi la probabilità che il primo individuo sia donna è 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

$$\left(\begin{array}{c} \text{numero donne} \\ \text{estratte} \end{array} \right) \sim \text{Bi}(350, 0{,}53)$$

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?

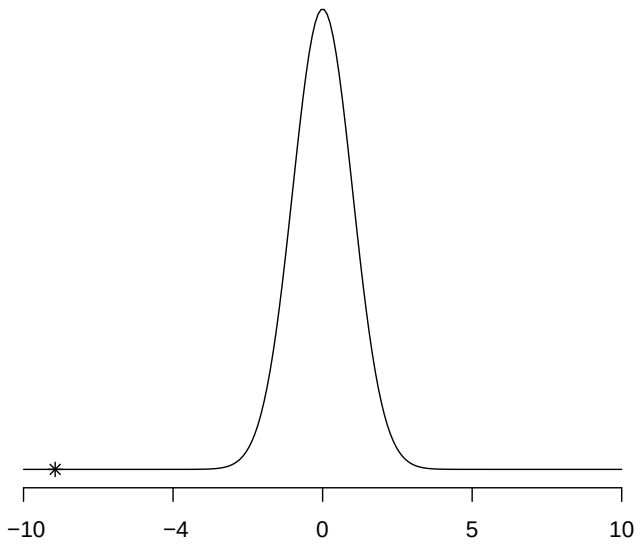
La risposta è si. Con un test statistico cerchiamo di valutare se i dati potrebbero essere stati generati dal meccanismo previsto dall'ipotesi nulla. E questo è quello che vogliamo fare nel presente contesto visto che la domanda che ci stiamo ponendo è:

> "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

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

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 62)

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

Ora, 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 D

Tonsille e *Streptococcus pyogenes*

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

I dati campionari

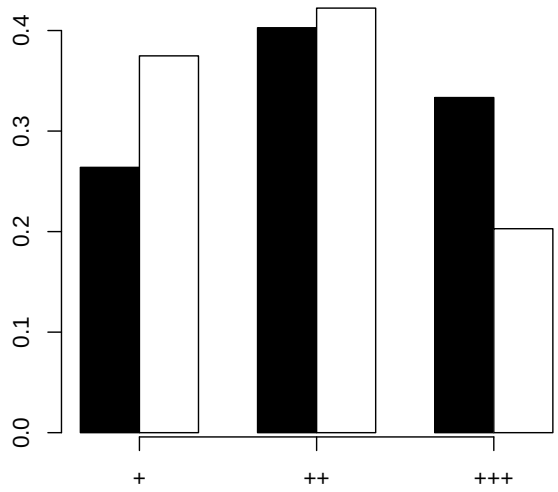
Nel corso di uno studio sulla determinazione di possibili fattori prognostici (predittivi) per alcune malattia otorino-laringoiatriche su 1398 bimbi o ragazzi sono state rilevate le seguenti due variabili:
(a) Presenza (in un tampone nasale) di `Streptococcus pyogenes`; variabile dicotomica con modalità "portatore" e "non portatore".
(b) Stato delle tonsille rilevato utilizzando la scala qualitativa ordinata: (i) normali (abbreviato in +), (ii) leggermente ingrossate (++) e (iii) ingrossate (+ + +).
I bimbi erano stati scelti casualmente tra tutti gli individui tra i 3 e i 15 di età residenti in un ampia e popolosa regione inglese.

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

La seguente tabella mostra le frequenze assolute.

Streptococcus	Tonsille			
pyogenes	+	++	+ + +	Totale
portatore	19	29	24	72
non portatore	497	560	269	1326
Totale	516	589	293	1298

Grafico a barre



Distribuzione di "stato delle tonsille" condizionata 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".

Frequenze attese e X^2 di Pearson

La seguente tabella mostra le frequenze attese nell'ipotesi di perfetta indipendenza in distribuzione.

Streptococcus pyogenes	Tonsille +	++	+ + +	Totale
portatore	26,6	30,3	15,1	72
non portatore	489,4	558,7	277,9	1326
Totale	516	589	293	1298

Il valore della statistica X^2 di Pearson è 7,88.

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.

La popolazione di riferimento

⋆ L'analisi precedente mostra che la distribuzione di "stato delle tonsille" è diversa tra i portatori e i non portatori. Quindi, nella tabella di pagina 76 non esiste indipendenza in distribuzione.

⋆ 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.

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 (ovvero della popolazione da cui è stato estratto)[1].

[1] E può anche non esserlo se l'estrazione è stata in qualche forma truccata (si pensi al giudice Ford!)

• *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.

Ascensori, aspirine e la mutabilità dei comportamenti umani

• Quanto detto deve *sempre* essere tenuto presente. In modo particolare, per gli studi nell'ambito delle scienze sociali.

• 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,. . .

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 76, del tipo

Streptococcus pyogenes	Tonsille +	++	+ + +	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 (i) N indica il numero di bimbi in quell'area dell'Inghilterra, (ii) F_{11} il numero di bimbi che sono portatori ma hanno le tonsille normali, (iii). . .

• La tabella precedente noi non la conosciamo. Per questo è una tabella *fantasma* e per questo le varie frequenze sono state indicate con lettere.

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

• Il campione è stato formato: (1) estraendo un bimbo tra gli N della popolazione; (2) estraendo un altro bimbo tra gli $N - 1$ bimbi non estratti alla prima estrazione; . . . ; (1398) 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.

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

Streptococcus pyogenes	Tonsille +	++	+ + +	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

• Vista la maniera con cui è stato formato il campione, la probabilità che il primo bimbo sia, ad esempio, un non portatore di *Streptococcus pyogenes* con le tonsille normali è π_{21}. 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, la tabella di pagina 76 mostra come si sono ripartiti nelle 6 "categorie" (portatori,+), (portatori,++), . . . , (non portatori,+ + +) i risultati di 1398 esperimenti casuali indipendenti tutti caratterizzati da

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

• Ma allora

$$(F_{11}, F_{12}, \ldots, F_{23}) \sim \text{Multinomiale}(n, (\pi_{11}, \pi_{12}, \ldots, \pi_{23}))$$

dove $(F_{11}, F_{12}, \ldots, F_{23})$ indica la variabile casuale che descrive il numero di (portatori,+), (portatori,++) e così via estratti[2].

[2]Si ricordi che il racconto di una variabile casuale multinomiale in termini di palline colorate e di urne è: se (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 $(N_1, \ldots, N_k)$ che descrive il numero di palline estratte del primo colore,del secondo colore,. . . , è una Multinomiale$(n, (\pi_1, \ldots, \pi_k))$.

Verifica dell'ipotesi di indipendenza

Una domanda interessante che possiamo fare ai dati è: nella tabella *fantasma* esiste indipendenza in distribuzione? 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

$$\begin{cases} \text{H}_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \quad i = 1, 2 \quad j = 1, 2, 3 \\ \text{H}_1 : \text{le } \pi_{ij} \text{ non rispettano i vincoli previsti da H}_0 \end{cases}$$

La statistica test più usata è l'X^2 di Pearson. E' certamente una statistica appropriata visto che ci aspettiamo che assuma valori (i) piccoli quando H_0 è vera e (ii) grandi quando è falsa.

La distribuzione approssimata di X^2

E' possibile mostrare (rinviamo al solito la dimostrazione di questo risultato a corsi più avanzati) che se H_0 è vera e nessuna frequenza attesa è troppo piccola allora la distribuzione di X^2 può essere approssimata con la distribuzione di una variabile casuale chiamata χ^2 di Pearson.

La distribuzione χ^2 dipende da un solo parametro, chiamato i gradi di libertà della distribuzione, che nel caso che stiamo trattando (verifica dell'ipotesi di indipendenza in una tabella di contingenza) deve essere posto uguale a

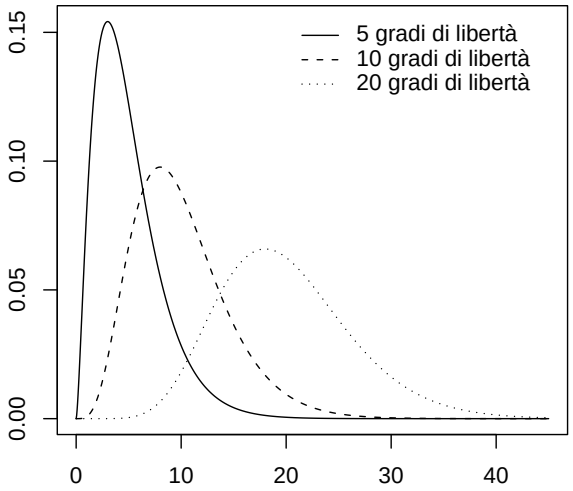
$$\left[\binom{\text{numero}}{\text{righe tabella}} - 1\right] \times \left[\binom{\text{numero}}{\text{colonne}}{\text{tabella}} - 1\right]$$

Ad esempio, per la tabella in esame, i gradi di libertà sono $2 = (2-1) \times (3-1)$.

L'approssimazione è ritenuta "decorosa" se la più piccola delle frequenze attese (si noti, quelle attese, non quelle osservate) è più grande di 1 e migliora man mano che queste aumentano.

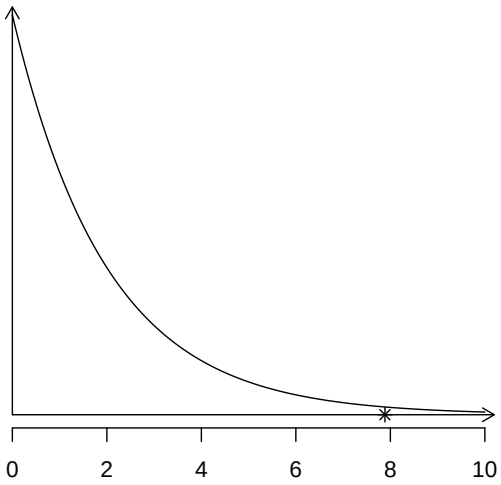
Alla fine di questa unità è allegata una tabella con pre-calcolati alcuni percentili di un χ^2.

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



Si osservi che all'aumentare dei gradi di libertà la densità si sposta verso destra. Infatti, è possibile dimostrare che la media della variabile casuale coincide con i gradi di libertà.

Test: analisi grafica del risultato



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

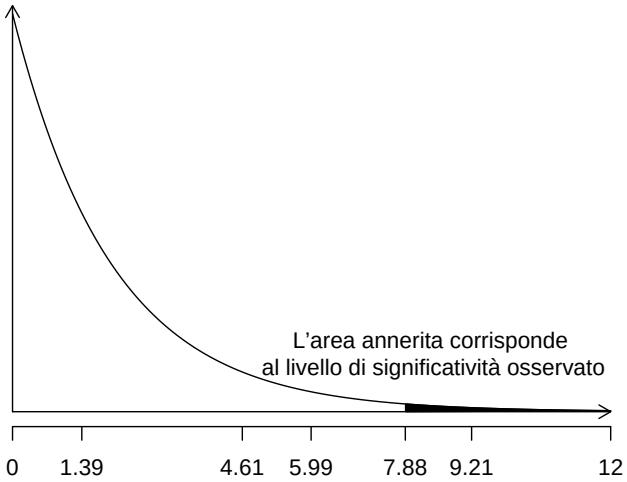
Livello di significatività osservato (e suo calcolo approssimato da una tavola dei percentili)

• "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. Quindi, per i dati presentati in questa unità,

$$\begin{pmatrix}\text{livello}\\ \text{significati-}\\ \text{vità}\\ \text{osservato}\end{pmatrix} = P(\chi^2 \text{ con 2 gradi libertà} \geq 7{,}88)$$

• Supponiamo ora di voler determinare un intervallo che lo contenga conoscendo solo alcuni percentili della distribuzione. Ad esempio, supponiamo di conoscere solamente la seguente tabella in cui $\chi^2_{2,p}$ indica il percentile p-simo di un χ^2 con 2 gradi di libertà.

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



• Il valore del test (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

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

Quantili di un χ^2 di Pearson

g indica i gradi di libertà, p la probabilità lasciata a "sinistra", Quindi, ad esempio, $P(\chi^2$ con 2 gradi di libertà $\leq 9{,}21) = 0{,}99$

	p							
g	0,1	0,25	0,5	0,75	0,90	0,95	0,99	0,999
1	0,02	0,1	0,45	1,32	2,71	3,84	6,63	10,83
2	0,21	0,58	1,39	2,77	4,61	5,99	9,21	13,82
3	0,58	1,21	2,37	4,11	6,25	7,81	11,34	16,27
4	1,06	1,92	3,36	5,39	7,78	9,49	13,28	18,47
5	1,61	2,67	4,35	6,63	9,24	11,07	15,09	20,52
6	2,2	3,45	5,35	7,84	10,64	12,59	16,81	22,46
7	2,83	4,25	6,35	9,04	12,02	14,07	18,48	24,32
8	3,49	5,07	7,34	10,22	13,36	15,51	20,09	26,12
9	4,17	5,9	8,34	11,39	14,68	16,92	21,67	27,88
10	4,87	6,74	9,34	12,55	15,99	18,31	23,21	29,59
11	5,58	7,58	10,34	13,7	17,28	19,68	24,72	31,26
12	6,3	8,44	11,34	14,85	18,55	21,03	26,22	32,91
13	7,04	9,3	12,34	15,98	19,81	22,36	27,69	34,53
14	7,79	10,17	13,34	17,12	21,06	23,68	29,14	36,12
15	8,55	11,04	14,34	18,25	22,31	25	30,58	37,7
16	9,31	11,91	15,34	19,37	23,54	26,3	32	39,25
17	10,09	12,79	16,34	20,49	24,77	27,59	33,41	40,79
18	10,86	13,68	17,34	21,6	25,99	28,87	34,81	42,31
19	11,65	14,56	18,34	22,72	27,2	30,14	36,19	43,82
20	12,44	15,45	19,34	23,83	28,41	31,41	37,57	45,31
21	13,24	16,34	20,34	24,93	29,62	32,67	38,93	46,8
22	14,04	17,24	21,34	26,04	30,81	33,92	40,29	48,27
23	14,85	18,14	22,34	27,14	32,01	35,17	41,64	49,73
24	15,66	19,04	23,34	28,24	33,2	36,42	42,98	51,18
25	16,47	19,94	24,34	29,34	34,38	37,65	44,31	52,62

Esercizi

1. Si costruisca una regola del tipo (accetto H_0)/(rifiuto H_0) che accetti H_0 quando H_0 è vera con una probabilità prefissata, diciamo $1 - \alpha$. La si applichi ai dati considerati usando sia $\alpha = 0{,}05$ che $\alpha = 0{,}01$.

2. Si supponga di voler calcolare un intervallo di confidenza per la proporzione di portatori dello `Streptococcus pyogenes` (diciamo che deve contenere il vero valore con probabilità 0,9). Come potremmo fare?.

3. Lo studente si convinca che la (D.1) e il grafico di pagina 89 forniscono essenzialmente la stessa informazione.

Unità E

Dove facciamo conoscenza con uno statistico birraio

Test t ad un campione.

Intervalli di confidenza per la media di una normale quando la varianza non è nota.

Ancora su di un esperimento su due sonniferi

• Nella parte di descrittiva abbiamo considerato un piccolo insieme di dati concernenti un esperimento su due sonniferi. Limitiamoci a considerare i risultati per la prima sostanza.

• Per dieci individui, era stata misurata la variabile, denominata `ore di extra sonno`, definita come

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

• 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.

• 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?"

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 $N(\mu, \sigma^2)$.

Tre precisazioni

1. 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 con problemi di insonnia" l'insieme di queste donne costituirebbe la nostra *popolazione di riferimento*.

2. Nel seguito "lavoreremo" supponendo vera l'ipotesi di normalità. Nella realtà, questa ipotesi dovrebbe prima essere verificata con i dati disponibili. Ovvero, il primo stadio dell'analisi dovrebbe consistere nel rispondere alla domanda: "E' plausibile che i dati osservati siano stati generati da una normale?". Per rispondere a questa domanda esistono tecniche grafiche ed analitiche che però fuoriescono dall'interesse di questo corso. Può, comunque, interessare lo studente che utilizzando queste tecniche la risposta alla domanda è: "Si. E' plausibile.".

3. Il modello che stiamo utilizzando 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.

Stima dei parametri del modello

• Nelle ipotesi fatte, la distribuzione dei dati (e soprattutto del fenomeno considerato nella popolazione) è nota con l'eccezione dei due parametri μ e σ^2. Sembra quindi ragionevole "iniziare" cercando di stimare questi due parametri dai dati.

• Gli stimatori più usati per μ e σ^2 sono rispettivamente

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

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 $n = 10$).

• Si noti che per stimare σ^2 si usa dividere la somma dei quadrati degli scarti dalla media campionaria per $n-1$ non per n. E' infatti possibile dimostrare che dividere per n porterebbe ad uno stimatore che tenzialmente "sottostima" la vera varianza. Lo stesso viceversa non vale per s^2.

Un problema di verifica d'ipotesi

• Un sistema d'ipotesi interessante in questo caso è

$$\begin{cases} \mathrm{H}_0 : \mu = \mu_0 \\ \mathrm{H}_1 : \mu \neq \mu_0 \end{cases}$$

con $\mu_0 = 0$. Accettare H_0, infatti, equivale a dire che, in media, prendendo il farmaco non si dorme ne di più ne di meno.

• Per verificare un sistema d'ipotesi analogo nell'unità A 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 $N(0,1)$. Ma la sostituzione del vero σ con s non può non essere "indolore" soprattutto nel caso di piccoli campioni in cui l'errore con cui s stima σ potrebbe anche essere grande.

• E' però possibile nelle nostre ipotesi (normalità delle osservazioni, indipendenza,...) determinare la distribuzione esatta di t_{oss}. E' stato fatto da W.S.Gosset uno statistico che lavoravava alla birreria (nel senso di fabbrica di birra) Guiness. Poichè i suoi lavori furono pubblicati sotto lo pseudonimo di Student, e Gosset, come anche noi abbiamo fatto, usava la lettera t per indicare la statistica test, la distribuzione viene comunemente chiamata t di Student.

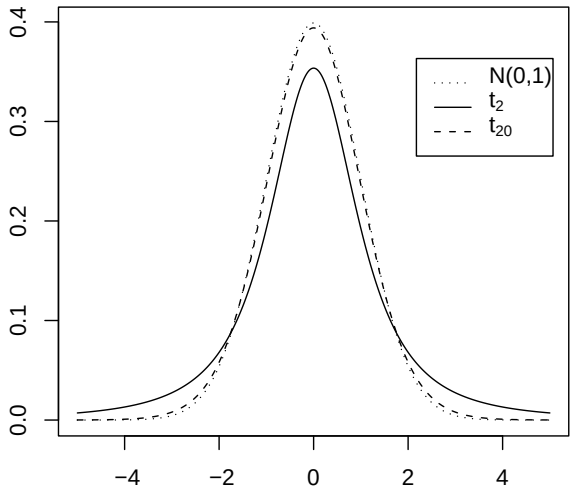
• La distribuzione t di Student dipende da un solo parametro, chiamato i gradi di libertà. Nel caso in esame (verifica sulla media di una distribuzione normale) deve essere posto uguale a $n-1$, ovvero, quello che Student ha dimostrato è che

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

• Nel grafico della pagina seguente sono disegnate le densità di (i) una t di Student con 2 gradi di libertà; (ii) una con 20 gradi di libertà e (iii) una normale standard. Si osservi come (a) le densità delle t siano simmetriche intorno allo zero; (b) abbiano delle "code" un po' più pesanti della normale e (c) la t con venti gradi di libertà sia molto vicina alla $N(0,1)$. E' possibile dimostrare che (a) e (b) valgono in generale (per qualsivoglia grado di libertà). L'osservazione (c) discende dal fatto che al divergere dei gradi di libertà la distribuzione di una t di Student converge ad una $N(0,1)$. Il che tra l'altro è quello che ci aspettiamo. Infatti, più n è grande più s^2 dovrebbe avvicinarsi a σ e quindi più t_{oss} dovrebbe avvicinarsi a z.

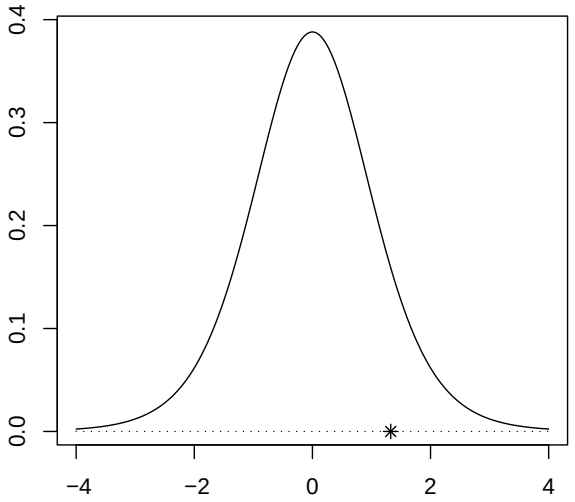
• Il test che stiamo descrivendo viene usualmente chiamato test **t a un campione**.

Grafico della densità della *t* di Student



Nota: I pedici indicano i gradi di libertà.

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. Non sembrano esserci elementi per dubitare che il valore osservato sia stato generato dalla distribuzione disegnata. Ovvero, non abbiamo elementi nei dati per rifiutare H_0.

Analisi mediante il livello di significatività osservato

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

$$\left(\begin{array}{c}\text{livello di}\\ \text{significati-}\\ \text{vità}\\ \text{osservato}\end{array}\right) = P(|t \text{ con 9 gradi di libertà}| \geq 1{,}33).$$

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

$$\left(\begin{array}{c}\text{livello di}\\ \text{significati-}\\ \text{vità}\\ \text{osservato}\end{array}\right) = 2\times P(t \text{ con 9 gradi di libertà} \geq 1{,}33).$$

• Disponendo solo di una tabella dei percentili, del tipo allegato al fondo di questa unità, 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,

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

Ma allora

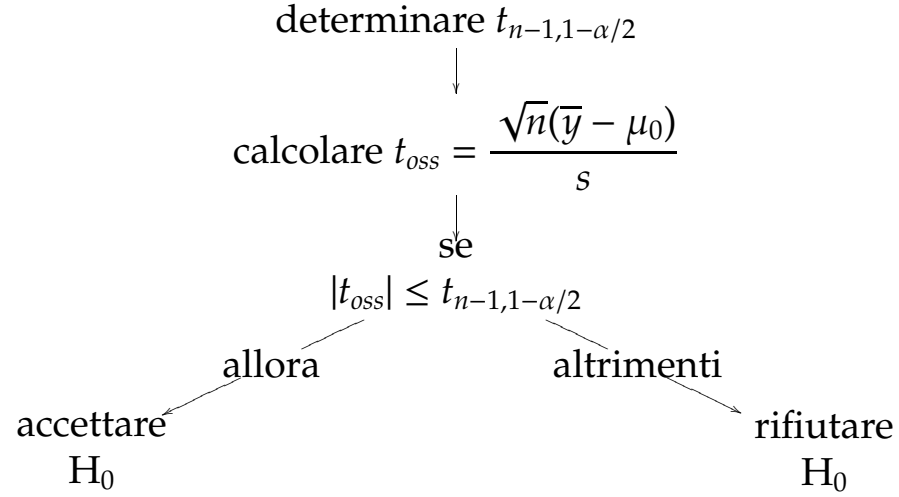
$$0{,}2 < \left(\begin{array}{c}\text{livello di}\\ \text{significatività}\\ \text{osservato}\end{array}\right) < 0{,}5$$

• 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.

• Ad esempio, possiamo guardare al livello di significatività osservato in questa maniera. Supponiamo: (i) che sia vera l'ipotesi nulla, (ii) di formare tutti i possibili campioni di numerosità 10 con gli individui che fanno parte della popolazione e (iii) di calcolare per ciascun campione t_{oss}. Allora, il livello di significativà osservato è la percentuale di valori di t_{oss} con un valore maggiore di 1,33. Il calcolo precedente ci dice che questa percentuale è maggiore di 0,2. Ma allora 1,33 è un valore che "può capitare quando H_0 è vera". Del resto, non riteniamo sorprendente che il lancio di un dado equilibrato sia 3. Ma la percentuale di casi in cui un lancio ci da come risultato 3 è inferiore al 20%.

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

Una regola del tipo accetto/rifiuto



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

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

Con i dati sul primo sonnifero

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

$$t_{n-1,1-\alpha/2} = t_{9,0{,}995} = 3{,}25$$
$$\downarrow$$
$$t_{oss} = 1{,}33$$
$$\downarrow$$
$$-3{,}25 \leq 1{,}33 \leq 3{,}25\ ?$$
$$\downarrow$$
$$\text{si}$$
$$\downarrow$$
$$\text{accettiamo } H_0$$

Un intervallo di confidenza

• Un intervallo di confidenza per μ può essere determinato, dai risultati precedenti utilizzando lo stesso ragionamento seguito nell'unità A.

• Infatti quello che sappiamo è che se μ è il vero valore della media allora

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

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

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

ovvero che

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

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

• Supponiamo, ad esempio, di voler un intervallo che contenga con probabilità 90% il vero valore di μ. Allora, $t_{n-1,1-\alpha/2} = t_{9,0,95} = 1{,}83$. 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

Si ripeta l'analisi (test ed intervallo di confidenza) con i dati del secondo sonnifero.

Quantili di una t di Student

g indica i gradi di libertà. p la probabilità lasciata a "sinistra". Quindi, ad esempio, $P(t \text{ con 2 gradi di libertà} \leq 6{,}96) = 0.99$. L'ultima riga ($g = \infty$) mostra i quantili di una $N(0, 1)$. Possono essere usati come approssimazione, se $g > 30$.

	p							
g	0,75	0,90	0,95	0,975	0,99	0,995	0,999	0,9995
1	1	3,08	6,31	12,71	31,82	63,66	318,31	636,62
2	0,82	1,89	2,92	4,3	6,96	9,92	22,33	31,6
3	0,76	1,64	2,35	3,18	4,54	5,84	10,21	12,92
4	0,74	1,53	2,13	2,78	3,75	4,6	7,17	8,61
5	0,73	1,48	2,02	2,57	3,36	4,03	5,89	6,87
6	0,72	1,44	1,94	2,45	3,14	3,71	5,21	5,96
7	0,71	1,41	1,89	2,36	3	3,5	4,79	5,41
8	0,71	1,4	1,86	2,31	2,9	3,36	4,5	5,04
9	0,7	1,38	1,83	2,26	2,82	3,25	4,3	4,78
10	0,7	1,37	1,81	2,23	2,76	3,17	4,14	4,59
11	0,7	1,36	1,8	2,2	2,72	3,11	4,02	4,44
12	0,7	1,36	1,78	2,18	2,68	3,05	3,93	4,32
13	0,69	1,35	1,77	2,16	2,65	3,01	3,85	4,22
14	0,69	1,35	1,76	2,14	2,62	2,98	3,79	4,14
15	0,69	1,34	1,75	2,13	2,6	2,95	3,73	4,07
16	0,69	1,34	1,75	2,12	2,58	2,92	3,69	4,01
17	0,69	1,33	1,74	2,11	2,57	2,9	3,65	3,97
18	0,69	1,33	1,73	2,1	2,55	2,88	3,61	3,92
19	0,69	1,33	1,73	2,09	2,54	2,86	3,58	3,88
20	0,69	1,33	1,72	2,09	2,53	2,85	3,55	3,85
21	0,69	1,32	1,72	2,08	2,52	2,83	3,53	3,82
22	0,69	1,32	1,72	2,07	2,51	2,82	3,5	3,79
23	0,69	1,32	1,71	2,07	2,5	2,81	3,48	3,77
24	0,68	1,32	1,71	2,06	2,49	2,8	3,47	3,75
25	0,68	1,32	1,71	2,06	2,49	2,79	3,45	3,73
26	0,68	1,31	1,71	2,06	2,48	2,78	3,43	3,71
27	0,68	1,31	1,7	2,05	2,47	2,77	3,42	3,69
28	0,68	1,31	1,7	2,05	2,47	2,76	3,41	3,67
29	0,68	1,31	1,7	2,05	2,46	2,76	3,4	3,66
30	0,68	1,31	1,7	2,04	2,46	2,75	3,39	3,65
∞	0,67	1,28	1,64	1,96	2,33	2,58	3,09	3,29

Unità F

Ancora su cuculi e Darwin

Cenno al test *t* a due campioni.

Il problema

• I dati sono stati presentati nella parte di descrittiva. Da un punto di vista descrittivo avevamo visto che la media delle lunghezze di uovo di cuculo deposte in nidi di pettirosso era diversa dalla media di uova di cuculo deposte in nidi di scricciolo. Avevamo anche notato che la dispersione dei due insiemi di dati era 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 maniera di guardare al problema è la seguente:
(a) Esistono due popolazioni di riferimento (o, equivalentemente, una popolazione divisa in due gruppi). Alla prima (seconda) popolazione 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 nelle due popolazioni. 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 – ma per imparare come si fa dovete attendere corsi più avanzati) 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. In particolare, se $y_1, \ldots, y_n$ sono le lunghezze rilevate dalle uove trovate in nidi di pettirossi e $x_1, \ldots, x_m$ le lunghezze delle uove trovate in nidi di scricciolo allora $y_1, \ldots, y_n$ sono determinazioni indipendenti di una $N(\mu, \sigma^2)$ mentre $x_1, \ldots, x_m$ sono determinazioni indipendenti, tra di loro e dalle "y", di una $N(\eta, \sigma^2)$.

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

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

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

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

può essere vista come una stima di σ^2 basata su *tutti* i dati.

• Si osservi che t_{oss} e s^2 indicano quantità diverse rispetto alla prima parte all'unità E.

• Ovviamente, più è grande, in valore assoluto, il valore di t_{oss} più i dati ci stanno suggerendo 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. I valore calcolato dai dati della statistica test può quindi essere analizzato in maniera analoga a quanto fatto nell'unità precedente.

• Si osservi che s^2 è facilmente calcolabile dalle stime di σ^2 costruite utilizzando solo le "y" e solo le "x". In particolare, 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

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

Applicazione alle lunghezze delle uove di cuculo

• In questo caso, abbiamo[1]

$$n = 16 \quad \overline{y} \approx 22{,}47 \quad s_y^2 \approx 0{,}46$$
$$m = 15 \quad \overline{x} \approx 21{,}13 \quad s_x^2 \approx 0{,}57$$

Quindi,

$$s \approx \sqrt{(15 \times 0{,}46 + 14 \times 0{,}57)/29} \approx 0{,}72$$

e

$$t_{oss} = \frac{22{,}47 - 21{,}13}{0{,}72\sqrt{\frac{1}{16} + \frac{1}{15}}} \approx 5{,}58$$

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

• La distribuzione sotto H_0 è una t di Student con 29 gradi di libertà.

• Dalla tabella dei quantili della t nell'unità E, 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.

Unità G

Hot-dog e calorie

(a) Rapporto tra medie e varianze condizionate e media e varianza marginali.

(b) Una misura della dipendenza in media.

(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.

E' inoltre noto se l'*hot-dog* era stato preparato con: (i) solo carne bovina; (ii) carne mista (tipicamente a maggioranza maiale); (iii) pollame (pollo o tacchino).

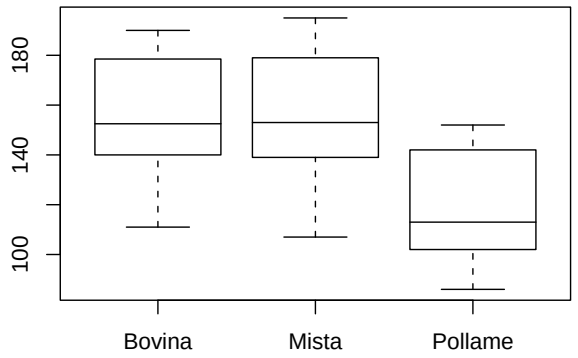
Le prossime due pagine mostrano: (i) i dati elementari; (ii) il diagramma scatola con baffi delle calorie classificate per tipo di carne e le numerosità, medie e scarti quadratici medi dei tre gruppi.

E' evidente che, restringendo l'attenzione alle 54 misure disponibili, il tipo di carne influenza il contenuto calorico.

Tipo di carne e calorie (per pezzo) per 54 confezioni di *hot-dog*

Carne	Calorie	Carne	Calorie	Carne	Calorie
Bovina	186	Bovina	181	Bovina	176
Bovina	149	Bovina	184	Bovina	190
Bovina	158	Bovina	139	Bovina	175
Bovina	148	Bovina	152	Bovina	111
Bovina	141	Bovina	153	Bovina	190
Bovina	157	Bovina	131	Bovina	149
Bovina	135	Bovina	132	Mista	173
Mista	191	Mista	182	Mista	190
Mista	172	Mista	147	Mista	146
Mista	139	Mista	175	Mista	136
Mista	179	Mista	153	Mista	107
Mista	195	Mista	135	Mista	140
Mista	138	Pollame	129	Pollame	132
Pollame	102	Pollame	106	Pollame	94
Pollame	102	Pollame	87	Pollame	99
Pollame	107	Pollame	113	Pollame	135
Pollame	142	Pollame	86	Pollame	143
Pollame	152	Pollame	146	Pollame	144

Un primo sguardo ai dati



Carne	Numerosità	$\overline{y}$	s
Bovina	20	156,85	22,64
Mista	17	158,71	25,24
Pollame	17	118,76	22,55

Nota: s è la radice della stima della varianza ottenuta "dividendo per $n-1$"

Nel seguito dell'unità ci concentremo sulla dipendenza in media rilevabile dalla tabella di pagina 128 ed in particolare affronteremo i seguenti punti:

- come misurare la forza della dipendenza in media e
- come verificare se è plausibile che le differenze osservate nelle medie siano *generalizzabili* a tutti gli *hot-dog* (o almeno a quelli prodotti con materie prime e tecnologia simili a quella usata per produrre le 54 confezioni).

Notazioni

- Per rendere il discorso generale indichiamo con k il numero dei gruppi (nel nostro caso $k = 3$) e con n_i, $i = 1, \ldots, k$ il numero di osservazioni per ogni gruppo (nel nostro caso, se 1 indica carne bovina, 2 carne mista e 3 pollame, $n_1 = 20$, $n_2 = 17$, $n_3 = 17$).

- L'insieme di tutte le osservazioni può poi essere indicato come y_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$. Si osservi che stiamo convenendo che il primo pedice indica il gruppo mentre il secondo l'osservazione dentro il gruppo.

- Per ogni gruppo possiamo calcolare la media

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

e la varianza

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

Nel nostro caso, queste medie e varianze sono riferibili alle distribuzioni delle calorie condizionate ai vari tipi di carne.

• Possiamo anche calcolare la media e la varianza di *tutte* le osservazioni senza riferimento al gruppo di appartenenza

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

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

dove

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

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

• Si osservi che abbiamo definito le varianze dividendo la somma dei quadrati degli scarti dalla media per il "numero delle osservazioni" e non per "il numero delle osservazioni -1". Abbiamo quindi deciso, per il momento, di muoverci in un contesto descrittivo (le varie "v^2" sono la varianza delle osservazioni non la stima della varianza della popolazione da cui le osservazioni provengono). In realtà, le relazioni che vedremo sono facilmente estendibili anche se si segue l'altra strada.

La media della distribuzione marginale è la media delle medie delle distribuzioni condizionate

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

modalità	$\overline{y}_1$	$\overline{y}_2$	...	$\overline{y}_k$
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 marginale $\overline{y}$.

- Infatti

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}.$$

Ma, per qualsivoglia i, dalla definizione di $\overline{y}_i$ segue che

$$\sum_{j=1}^{n_i} y_{ij} = n_i \overline{y}_i$$

e quindi,

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

La varianza della marginale è la media delle varianze condizionate + la varianza delle medie condizionate

- Ci si ricordi che v^2 indica la varianza di tutti i dati (= la varianza della distribuzione marginale), mentre le v_i^2 sono le varianze dentro il gruppo i-simo (= le varianze delle distribuzione condizionate).

- Dimostreremo che

$$v^2 = \frac{1}{n} \sum_{i=1}^{k} n_i v_i^2 + \frac{1}{n} \sum_{i=1}^{k} (\overline{y}_i - \overline{y})^2. \qquad \text{(G.1)}$$

- Si osservi che il primo addendo sul lato destro è la media (ponderata) di una distribuzione in cui le v_i^2 sono le modalità mentre le n_i sono le frequenze assolute, ovvero, è una media di varianze condizionate calcolata con pesi proporzionali alla numerosità dei vari gruppi.

- Viceversa, il secondo addendo è la varianza della distribuzione mostrata all'inizio di pagina 132, ovvero è la varianza delle medie condizionate (calcolata anche in questi caso con pesi proporzionali. . .).

- La verifica della (G.1) è agevole. Infatti

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

Nell'ultima semplificazione abbiamo usato il fatto che, come sappiamo, la somma delle osservazioni del gruppo i-simo dalla media del gruppo i-simo è uguale a zero.

Una misura della dipendenza in media

★ La (G.1) mostra come la varianza totale, v^2, sia scomponibile in due parti:

(i) la prima, il 1° addendo, dovuta alla variabilità *entro* i 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 **varianza entro i gruppi** e **varianza tra i gruppi**.

★ Si osservi che se la varianza tra i gruppi è nulla, le medie condizionate sono tutte uguali a $\overline{y}$ e quindi esiste indipendenza in media.

★ Viceversa, se la varianza tra i gruppi è molto grande rispetto alla varianza entro i gruppi, allora buona parte della variabilità totale mostrata dai dati diventa interpretabile in termini di differenze tra le medie condizionate. Siamo quindi in presenza di una situazione in cui la dipendenza in media esiste ed è importante (= le differenze tra le medie "spiegano" una larga frazione delle differenze che osserviamo nei dati).

★ Sembra allora ragionevole usare

$$\eta^2 = \frac{\text{varianza tra i gruppi}}{\text{varianza totale}} = \\ = 1 - \frac{\text{varianza entro i gruppi}}{\text{varianza totale}}$$

per misurare la forza della dipendenza in media.

★ In particolare si osservi che

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

(b) $\eta^2 = 0$ implica indipendenza in media.

(c) $\eta^2 = 1$ implica che la varianza entro i gruppi è nulla. Siamo quindi in una situazione di dipendenza, in ogni senso e quindi anche in media, perfetta.

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

★ Nel caso degli *hot-dog*, η^2 è facilmente calcolabile dai risultati della tabella di pagina 128[1]. In particolare,

$$\begin{aligned} \text{varianza entro i gruppi} &\approx 519{,}76 \\ \text{varianza tra i gruppi} &\approx 327{,}63 \\ \text{varianza totale} &\approx 847{,}39 \end{aligned}$$

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

[1]Ci si ricordi comunque che nella tabella, come spiegato nella legenda, s è la radice della stima della varianza calcolata dividendo per "il numero dei dati -1". Quindi, con notazioni ovvie, la varianza entro i gruppi deve essere calcolata con la formula $[\sum(n_i - 1)s_i^2]/n$.

E se tutto fosse dovuto al caso

• Fino a questo punto abbiamo solo guardato ai dati disponibili. In realtà noi non compreremo mai nessuna delle 54 confezioni di *wurstel* analizzate e quindi siamo interessati a sapere quanto le differenze evidenziate siano estendibili ai *wurstel* che potremmo mangiare.

• Una maniera di vedere il problema consiste nel riconoscere che fino a questo punto abbiamo trascurato una ulteriore fonte di variabilità, quella **campionaria**. Ad esempio, almeno una parte delle differenze tra le medie presentate nella tabella di pagina 128 è specifica delle 54 confezioni utilizzate, nel senso che, replicando l'esperimento (ovvero, prendendo altre 54 confezioni,...) ci aspettiamo di trovare risultati diversi.

• La domanda è: "Di quanto diversi? Tanto diversi, ad esempio, da portarci a concludere che le minore calorie osservate per gli *hot-dog* di pollo e tacchino sono solamente una specificità del campione disponibile? Oppure, diversi si, ma non tanto da alterare le conclusioni suggerite dalla tabella?"

Un problema di verifica d'ipotesi

• Pensiamo all'insieme[2] 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: (i) quelli prodotti con sola carne di bovino; (i) quelli prodotti con carne mista; (i) quelli prodotti con pollame. Possiamo allora calcolare la media delle calorie per ciascuno di questi tre gruppi. Indichiamole rispettivamente con μ_1, μ_2 e μ_3.

• Un sistema di ipotesi che può essere interessante verificare con i dati è

$$\begin{cases} \mathrm{H}_0 : \{\mu_1 = \mu_2 = \mu_3\} \\ \mathrm{H}_1 : \left\{ \begin{array}{c} \text{almeno una delle} \\ \text{uguaglianze previste da} \\ \mathrm{H}_0 \text{ è falsa} \end{array} \right\} \end{cases}.$$

Infatti, se H_0 fosse vera allora nella popolazione, contrariamente a quanto osservato nel campione, non esisterebbe dipendenza in media.

• Si osservi come il problema sia molto simile a quello che ci siamo posti nell'unità F. La differenza è che adesso sono coinvolte più di due medie.

[2] Un po' stomachevole, nella sua enormità.

Analisi della varianza con un criterio di classificazione

Al solito, per arrivare ad una soluzione abbiamo bisogno di descrivere la relazione che intercorre tra le osservazioni e la popolazione. In particolare, la relazione che interccorre tra le osservazioni e le tre medie μ_1, μ_2 e μ_3. Una relativamente "semplice" soluzione esiste nel caso in cui sia credibile assumere che:

1. La distribuzione all'interno dell'i-gruppo sia normale di media μ_i e varianza σ^2. Si osservi che stiamo supponendo che la varianza non dipenda da i, ovvero, che tutti i gruppi abbiano la stessa variabilità interna.

2. Le osservazioni sono tutte indipendenti tra di loro e, per qualsivoglia i e j, $y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2)$.

La statistica test comunemente usata è

$$F_{oss} = \left(\frac{\text{varianza tra i gruppi}}{\text{varianza entro i gruppi}}\right)\left(\frac{n-k}{k-1}\right).$$

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.

Per quello che ci riguarda una F di Snedecor è una ulteriore variabile casuale, dipendente da due parametri, i gradi di libertà menzionati precedentemente. Alcuni percentili, per alcune combinazioni dei gradi di libertà sono riportati, alla fine dell'unità.

In pratica

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. Dalla tabella alla fine dell'unità vediamo 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.

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.

Sempre a proposito del livello di significatività osservato, ricordiamo che comunemente se è inferiore a 0,01 i risultati sono considerati *altamente significativi* contro H_0 mentre se risulta compreso tra 0,01 e 0,05 si parla di risultati *significativi*, sempre contro l'ipotesi nulla. Viceversa se risulta maggiore di 0,1 si conclude che i dati non contengono elementi tali da poter rifiutare H_0 e quindi si parla di *non significatività*. I valori che mancano, ovvero quelli compresi tra 0,05 e 0,1 sono i più difficili da interpretare. Siamo in una situazione di sostanziale indecisione, a volta indicata come risultato ai *margini della significatività* o *borderline*. Ovviamente, questi valori (0,01, 0,05 e 0,1) non hanno niente di *sacro*.

Quantili di una F di Snedecor

La tabella è deliberatamente limitata a quello che può essere utile per esercizi ed esami. $g1$ e $g2$ indicano rispettivamente i gradi di libertà del numeratore e del denominatore, p la probabilità lasciata a "sinistra". Quindi, ad esempio, $P(F \text{ con } 2 \text{ e } 10 \text{ gradi di libertà} \leq 4{,}1) = 0{,}95$.

		p					
$g1$	$g2$	0,5	0,75	0,90	0,95	0,99	0,999
1	10	0,49	1,49	3,29	4,96	10,04	21,04
1	15	0,48	1,43	3,07	4,54	8,68	16,59
1	20	0,47	1,4	2,97	4,35	8,1	14,82
1	30	0,47	1,38	2,88	4,17	7,56	13,29
1	50	0,46	1,35	2,81	4,03	7,17	12,22
1	50	0,46	1,35	2,81	4,03	7,17	12,22
1	51	0,46	1,35	2,81	4,03	7,16	12,19
2	10	0,74	1,6	2,92	4,1	7,56	14,91
2	15	0,73	1,52	2,7	3,68	6,36	11,34
2	20	0,72	1,49	2,59	3,49	5,85	9,95
2	30	0,71	1,45	2,49	3,32	5,39	8,77
2	50	0,7	1,43	2,41	3,18	5,06	7,96
2	50	0,7	1,43	2,41	3,18	5,06	7,96
2	51	0,7	1,42	2,41	3,18	5,05	7,93

		p					
$g1$	$g2$	0,5	0,75	0,90	0,95	0,99	0,999
3	10	0,85	1,6	2,73	3,71	6,55	12,55
3	15	0,83	1,52	2,49	3,29	5,42	9,34
3	20	0,82	1,48	2,38	3,1	4,94	8,1
3	30	0,81	1,44	2,28	2,92	4,51	7,05
3	50	0,8	1,41	2,2	2,79	4,2	6,34
3	50	0,8	1,41	2,2	2,79	4,2	6,34
3	51	0,8	1,41	2,19	2,79	4,19	6,32
4	10	0,9	1,59	2,61	3,48	5,99	11,28
4	15	0,88	1,51	2,36	3,06	4,89	8,25
4	20	0,87	1,47	2,25	2,87	4,43	7,1
4	30	0,86	1,42	2,14	2,69	4,02	6,12
4	50	0,85	1,39	2,06	2,56	3,72	5,46
4	50	0,85	1,39	2,06	2,56	3,72	5,46
4	51	0,85	1,39	2,06	2,55	3,71	5,44
5	10	0,93	1,59	2,52	3,33	5,64	10,48
5	15	0,91	1,49	2,27	2,9	4,56	7,57
5	20	0,9	1,45	2,16	2,71	4,1	6,46
5	30	0,89	1,41	2,05	2,53	3,7	5,53
5	50	0,88	1,37	1,97	2,4	3,41	4,9
5	50	0,88	1,37	1,97	2,4	3,41	4,9
5	51	0,88	1,37	1,96	2,4	3,4	4,88

Unità H

Veleni e antidoti

Un esempio di analisi della varianza a due criteri di classificazione

I dati

- ⋆ Durante, una ricerca sono stati considerati tre differenti veleni e quattro possibili antidoti.
- ⋆ Ogni combinazione formata da un particolare veleno e un particolare antidoto è stata iniettata a quattro cavie.
- ⋆ Per ogni cavia è stato poi misurato il tempo di sopravvivenza (in ore) dopo l'iniezione.
- ⋆ I dati sono raccolti nella seguente tabella

	Veleno		
Antidoto	1	2	3
1	0,31 0,45 0,46 0,43	0,36 0,29 0,40 0,23	0,22 0,21 0,18 0,23
2	0,82 1,1 0,88 0,72	0,92 0,61 0,49 1,24	0,30 0,37 0,38 0,29
3	0,43 0,45 0,63 0,76	0,44 0,35 0,31 0,4	0,23 0,25 0,24 0,22
4	0,45 0,71 0,66 0,62	0,56 1,02 0,71 0,38	0,3 0,36 0,31 0,33

- ⋆ Nel seguito lavoreremo non con i tempi di sopravvivenza ma con i loro reciprochi. Useremo infatti delle tecniche basate su assunzioni che sembrano soddisfatte per i reciprochi ma sospette per i dati originali.
- ⋆ Ovviamente, reciprochi "piccoli" equivale a "sopravvivenze" alte e viceversa.

Domande

1. I differenti veleni hanno un impatto differente sulla sopravvivenza?
2. I differenti antidoti hanno un effetto differente?
3. Se si, l'effetto dell'antidoto dipende dal veleno?

Il modello di riferimento

Si indichi con y_{ijk} il reciproco del tempo di sopravvivenza

- della k-sima cavia ($k = 1, \ldots, 4$) a cui è stato iniettato
- il veleno i-simo ($i = 1, 2$ o 3)
- e l'antidoto j-simo ($j = 1, \ldots, 4$).

Quando ci servirà per rendere il discorso generale indicheremo i massimi dei 3 pedici rispettivamente con I, J, K. Nel nostro caso ovviamente $I = 3$, $J = 4$ e $K = 4$.

Supporremmo nel seguito che sia accettabile assumere che, almeno approssimativamente,

1. tutte le y_{ijk} sono indipendenti tra di loro;
2. ogni y_{ijk} può essere vista come una determinazione di una variabile casuale normale di media μ_{ij} e varianza σ^2.

Le ipotesi precedenti definiscono quello che è usualmente chiamato il **modello di analisi della varianza a due criteri di classificazione**.

Riparametrizzazione delle medie: formule

Poniamo

$$\begin{aligned}
\alpha &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij} \\
\beta_i &= \frac{1}{J}\sum_{j=1}^{J}(\mu_{ij}-\alpha) \quad i=1,\ldots,I \\
\gamma_j &= \frac{1}{I}\sum_{i=1}^{I}(\mu_{ij}-\alpha) \quad j=1,\ldots,J \\
\delta_{ij} &= \mu_{ij}-\alpha-\beta_i-\gamma_j
\end{aligned}$$

Ovviamente possiamo scrivere

$$\mu_{ij} = \alpha + \beta_i + \gamma_i + \delta_{ij} \qquad \text{(H.1)}$$

Riparametrizzazione delle medie: interpretazione

- ⋆ α ci fornisce un livello medio prescindendo sia dai veleni che dagli antidoti;
- ⋆ β_i misura un effetto (medio tra gli antidoti) del veleno i-simo; infatti $\mu_{ij}-\alpha$ può essere preso come "effetto" della combinazione (veleno i,antidoto j); β_i è la media tra tutti gli antidoti di questi effetti; questi coefficienti vengono chiamati **l'effetto principale del veleno**;
- ⋆ l'interpretazione di γ_j è simmetrica; γ_j può essere visto come una misura (media tra i veleni) dell'effetto del antidoto j-simo; questi coefficienti vengono chiamati **l'effetto principale degli antidoti**;
- ⋆ le δ_{ij} costituiscono quello che usualmente si chiama il **fattore di interazione** tra veleno e antidoto; misurano se e come varia l'effetto dell'antidoto al variare del veleno.

Sull'interazione

Quello che capita se si passa dall'antidoto 3 all'antidoto 2 (ovviamente solo per fare un esempio) può essere misurato dalle differenze

$$\begin{aligned} \text{primo veleno} &= \mu_{12} - \mu_{13} = (\gamma_2 - \gamma_3) + (\delta_{12} - \delta_{13}) \\ \text{secondo veleno} &= \mu_{22} - \mu_{23} = (\gamma_2 - \gamma_3) + (\delta_{22} - \delta_{23}) \\ \text{terzo veleno} &= \mu_{32} - \mu_{33} = (\gamma_2 - \gamma_3) + (\delta_{32} - \delta_{33}) \end{aligned}$$

Quindi se $\delta_{i2} = \delta_{i3}$ per $i = 1, 2$ o 3, l'effetto sulla media del passaggio dall'antidoto 2 all'antidoto 3 non dipende dal veleno (ovvero per qualsiasi veleno, l'effetto del passaggio sarebbe $\gamma_2 - \gamma_3$).

Viceversa se $\delta_{i2} \neq \delta_{i3}$ per qualche i l'effetto del passaggio dipende anche dal veleno e potrebbe anche essere vantaggioso per un veleno e svantaggioso per l'altro.

Un esempio numerico

Supponiamo (a) di lavorare con i tempi di sopravvivenza (non con i loro reciproci) e (b) che la vera tabella delle medie sia la seguente (ci si ricordi comunque che noi non la conosciamo!):

μ_{ij}

	Antidoto			
Veleno	1	2	3	4
1	0,45	0,81	0,53	0,67
2	0,38	0,74	0,46	0,60
3	0,11	0,47	0,19	0,33

Allora

$$\alpha \approx 0{,}48.$$

β_1	β_2	β_3
0,14	0,07	-0,21

γ_1	γ_2	γ_3	γ_4
-0,16	0,20	-0,09	0,05

Le δ_{ij} sono tutte praticamente nulle.

Pezzi dell'interpretazione sarebbero:

1. mediamente le cavie a cui viene somministrata una qualsiasi combinazione dei veleni e degli antidoti considerati soppravvivono per mezzora;

2. le cavie che hanno ricevuto il terzo veleno sopravvivono in media 12 minuti di meno della media complessiva;

3. le cavie che ricevono il secondo antidoto sopravvivono circa 12 minuti di più della media di tutte le cavie;

4. visto che l'interazione è assente passare dall'antidoto 2 all'antidoto 3 comporta, qualsiasi sia il veleno somministrato, un diminuzione di circa 17 minuti nella sopravvivenza media; infatti, visto che le δ_{ij} sono nulle

$$\mu_{12} - \mu_{13} = \mu_{22} - \mu_{23} = \mu_{32} - \mu_{33} = \gamma_2 - \gamma_3 = 0{,}29$$

Un altro esempio

Supponiamo che α, β_i e le γ_j siano come nell'esempio di prima ma che

δ_{ij}

	Antidoto			
Veleno	1	2	3	4
1	0,0	0,5	-0,5	0,0
2	0,0	0,0	0,0	0,0
3	0,0	-0,5	0,5	0,0

Consideriamo, tanto per cambiare l'esempio, il passaggio dall'antidoto 1 all'antidoto 2. Possiamo "misurarne" l'effetto con le differenze

veleno	variazione media	valore
1	$\mu_{11} - \mu 12$	-0,54
2	$\mu_{21} - \mu 22$	-0,04
3	$\mu_{21} - \mu 22$	0,46

ovvero, l'effetto del passaggio dipende dal veleno, ovvero, l'effetto dell'antidoto dipende dal veleno. In questo caso, ad esempio, passare dall'antidoto 1 all'antidoto 2, peggiora la sopravvivenza delle cavie che hanno ricevuto il veleno 1 di circa mezzora, la migliora di altrettato per quelle che hanno ricevuto il terzo veleno, la lascia più o meno invariata per quelle che hanno preso il veleno 2.

La riparametrizzazione non è unica

Esistono molte riparametrizzazioni che permettono di scomporre le medie in maniera analoga alle (H.1).

Si osservi, ad esempio, che la (H.1) continua a valere se sommiamo una qualsiasi costante a tutte le β_i e sottraiamo la medesima costante alle δ_{ij}.

La parametrizzazione presentata risolve queste ambiguità imponendo, come è facile verificare, che

$$\sum_{i=1}^{I} \beta_i = 0$$
$$\sum_{j=1}^{J} \gamma_j = 0$$
$$\sum_{i=1}^{I} \delta_{ij} = 0 \quad j = 1, \ldots, J$$
$$\sum_{j=1}^{J} \delta_{ij} = 0 \quad i = 1, \ldots, I$$

Alcune ipotesi di interesse

Le domande di pagina 153 possono essere ricondotte alla verifica delle seguenti ipotesi (ognuna contro la sua negazione):

$$_1H_0 : \{\beta_1 = \cdots = \beta_I = 0\},$$

$$_2H_0 : \{\gamma_1 = \cdots = \gamma_J = 0\},$$

e

$$_3H_0 : \{\delta_{ij} = 0 \text{ per ogni } i = 1, \ldots, I \text{ e } j = 1, \ldots, J\}.$$

Ad esempio, accettare $_3H_0$ equivale ad accettare l'ipotesi che l'effetto degli antidoti non dipende dal veleno. Oppure, accettare, sia $_2H_0$ che $_3H_0$, equivale ad affermare sulla base dei dati, che non esistono differenze tra gli antidoti.

Stima dei parametri

Gli stimatori usuali per μ_{ij} e σ^2 sono rispettivamente

$$\overline{y}_{ij} = \frac{1}{K}\sum_{k=1}^{K} y_{ijk}$$

e

$$s^2 = \frac{1}{IJK - IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \overline{y}_{ij})^2$$

Si osservi che, come nei casi precedenti, la stima di σ^2 è il rapporto tra

- la somma dei quadrati delle differenze delle osservazioni dalla stima della loro media e
- il numero delle osservazioni meno il numero di medie stimate.

Utilizzando le formule di pagina 155, ma sostituendo alle vere medie dei gruppi le loro stime, è poi possibile ottenere stime della media generale (α), degli effetti principali dei veleni e degli antidoti (le "β_i" e le "γ_j") e dei coefficienti di interazione (le "δ_{ij}"). Ad esempio, lo stimatore usuale di α è

$$\hat{\alpha} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\overline{y}_{ij} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}$$

ovvero, nient'altro che la media di tutte le osservazioni, mentre lo stimatore di β_2 è

$$\hat{\beta}_2 = \frac{1}{J}\sum_{j=1}^{J}(\overline{y}_{2j} - \hat{\alpha}) = \left(\frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{2jk}\right) - \hat{\alpha}$$

ovvero la differenza tra

- la media delle osservazioni delle cavie che hanno ricevuto il secondo veleno e
- la media di tutte le osservazioni.

Scomposizione dei dati...

Alle stime dei parametri trasformati ovviamente corrisponde la seguente scomposizione delle singole osservazioni

$$y_{ijk} = \hat{\alpha} + \hat{\beta}_i + \hat{\gamma}_j + \hat{\delta}_{ij} + r_{ijk}$$

dove

$$r_{ij} = y_{ijk} - \overline{y}_{ij}$$

costituisce la "parte dell'osservazione (i, j, k)" *non spiegata* dalla media del gruppo. Tra l'altro si osservi che

$$s^2 = \frac{1}{IJK - IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} r_{ijk}^2$$

...e relativa scomposizione della devianza

E' piuttosto agevole dimostrare[1] che la "devianza totale" ovvero la somma dei quadrati di tutte le osservazioni dalla loro media può essere scomposta come

$$\begin{aligned}\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{ijk} - \hat{\alpha})^2 = \quad & JK \sum_{i=1}^{I} \hat{\beta}_i^2 + IK \sum_{j=1}^{J} \hat{\gamma}_j^2 + \\ & + K \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{\delta}_{ij}^2 + \\ & + \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} r_{ijk}^2\end{aligned}$$

ovvero in quattro addendi che possono essere visti rispettivamente come una misura della importanza degli effetti principali dei veleni e degli antidoti, della forza della interazione e della componente residua.

[1] Per lo studente che ci vuole tentare: una possibilità semplice consiste nell'applicare in maniera iterativa la formula per la decomposizione della varianza vista nell'unità precedente

Tabella di analisi della varianza

I risultati vengono tipicamente presentati in una tabella del tipo

fattore	somma dei quadrati	gradi di libertà	errore quadratico medio	F	p
veleni	d^2_{v}	$I-1$	$s^2_{\mathrm{v}} = \frac{d^2_{\mathrm{v}}}{I-1}$	$\mathrm{F}_v = \frac{s^2_{\mathrm{v}}}{s^2_{\mathrm{r}}}$	p_v
antidoti	d^2_{a}	$J-1$	$s^2_{\mathrm{a}} = \frac{d^2_{\mathrm{a}}}{J-1}$	$\mathrm{F}_{\mathrm{a}} = \frac{s^2_{\mathrm{a}}}{s^2_{\mathrm{r}}}$	p_a
interazione	d^2_{i}	$(I-1) \times (J-1)$	$s^2_{\mathrm{i}} = \frac{d^2_{\mathrm{i}}}{(I-1)(J-1)}$	$\mathrm{F}_{\mathrm{i}} = \frac{s^2_{\mathrm{i}}}{s^2_{\mathrm{r}}}$	p_i
residuo	d^2_{r}	$IJ(K-1)$	$s^2_{\mathrm{r}} = \frac{d^2_{\mathrm{r}}}{IJ(K-1)}$		
totale	d^2_{T}	$IJK-1$	$s^2_{\mathrm{T}} = \frac{d^2_{\mathrm{T}}}{IJK-1}$		

dove

$$d_v^2 = JK\sum_i \hat{\beta}_i^2 \quad d_a^2 = IK\sum_j \hat{\gamma}_j^2$$
$$d_i^2 = K\sum_{ij} \hat{\delta}_{ij}^2 \quad d_r^2 = \sum_{ijk} r_{ijk}^2$$
$$d_T^2 = \sum_{ijk}(y_{ijk} - \hat{\alpha})^2 = d_v^2 + d_a^2 + d_i^2 + d_r^2$$

Le ultime due colonne, quelle etichettate F e p, contengono le statistiche test e i relativi livelli di significatività dei test usuali per verificare le ipotesi descritte a pagina 162.

L'interpretazione delle statistiche è immediata: come nell'unità precedente sono, nella sostanza, il rapporto tra la parte della somma dei quadrati "spiegata" dal fattore considerato e quella attribuile al residuo. Nel caso una delle ipotesi nulla sia falsa ci aspettiamo valori grandi per la statistica relativa. I livelli di significatività osservati possono essere calcolati come la probabilità (sotto l'ipotesi nulla) di osservare un valore più grande di quello osservato.

Questo è possibile poichè si dimostra che

$$\begin{aligned} F_v &\sim F \text{ con } I-1 \text{ e } IJ(K-1) \text{ gradi di libertà} \\ F_a &\sim F \text{ con } J-1 \text{ e } IJ(K-1) \text{ gradi di libertà} \\ F_i &\sim F \text{ con } (I-1)(J-1) \text{ e } IJ(K-1) \text{ gradi di libertà} \end{aligned}$$

Quindi, ad esempio,

$$p_v = \mathrm{pr}\left\{\begin{pmatrix}\text{F di Snedecor con } I-1 \text{ e} \\ JK(K-1) \text{ gradi di libertà}\end{pmatrix} \geq F_v\right\}.$$

Con i dati

fattore	somma dei quadrati	gradi di libertà	errore quadratico medio	F	p
veleno	34,877	2	17,439	72,6347	< 0,0001
antidoto	20,414	3	6,805	28,3431	< 0,0001
interazione	1,571	6	0,262	1,0904	0,3867
residuo	8,643	36	0,240		
totale	65,505	47	1,394		

I dati suggeriscono quindi che

- ⋆ gli effetti principali dei veleni e degli antidoti non sono nulli; ovvero, ci sono differenze sia tra i veleni che tra antidoti;
- ⋆ l'interazione non è importante ovvero, l'effetto dei vari antidoti è lo stesso per tutti i veleni

Stime degli effetti principali

veleni		
1	2	3
-0,8217	-0,3530	1,1747

antidoti			
1	2	3	4
0,8970	-0,7604	0,3248	-0,4614

L'ordinamento di pericolosità dei veleni sembra quindi essere

$$1 < 2 < 3$$

mentre l'ordinamento di efficienza tra gli antidoti suggerito è

$$1 < 3 < 4 < 2$$

[Ci si ricordi che stiamo lavorando con il reciproco del tempo di sopravvivenza. Quindi, un effetto positivo e magari grande è associato ad una media alta del reciproco, ovvero ad un basso tempo di sopravvivenza]