Introduzione alla verifica d'ipotesi

Alessandra Nardi

Dipartimento di Matematica

Università di Roma "Tor Vergata"

28 marzo 2023

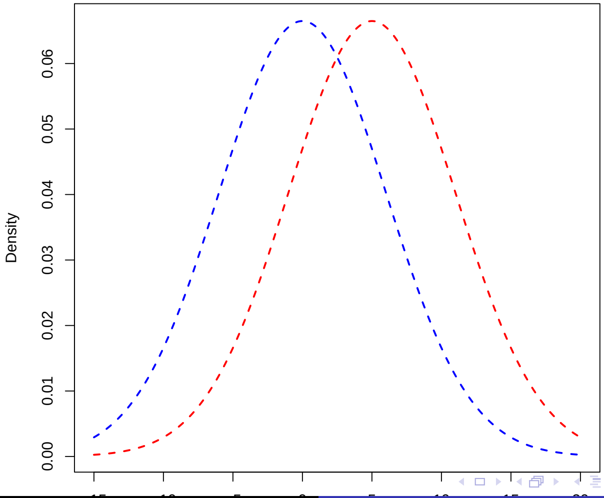
Consideriamo il caso in cui la nostra variabile risposta sia continua
Immaginiamo che obiettivo del nostro studio sia valutare l'efficacia di un farmaco sperimentale per la riduzione del livello di colesterolo totale in pazienti adulti, di entrambi i sessi, affetti da ipercolesterolemia.
Ipotizziamo di misurare l'efficacia del farmaco sulla base della variazione del livello di colesterolo dopo 30 giorni di trattamento (Delta:=valore basale - valore post-trattamento).
Assumiamo che tale variazione segua nella popolazione un **modello Normale** di valore atteso μ e varianza σ^2
Ricordiamo che μ rappresenta la riduzione media che osserveremmo nella popolazione se trattata con il nuovo farmaco e σ^2 la variabilitÃ della risposta al farmaco.

Vogliamo valutare il **sistema d'ipotesi**

$$\begin{array}{lcll} H_0 & : & \mu = 0\, mg/dl & \text{Ipotesi nulla: nessuna efficacia in media} \\ H_1 & : & \mu = 5\, mg/dl & \text{Ipotesi alternativa: 5 mg/dl di riduzione media} \end{array}$$

Ipotizziamo di conoscere il valore di $\sigma = 6mg/dl$ a priori
L'unico elemento incognito resta l'efficacia media nella popolazione in termini di variazione di colesterolo anche se la nostra incertezza e' limitata alle due possibili ipotesi, 0 mg/dl o 5 mg/dl
Questa incertezza residua dovrà essere risolta sulla base dei dati sperimentali

Di fatto stiamo ipotizzando che, se tutta la popolazione venisse trattata con il farmaco sperimentale, la distribuzione della variazione di colosterolo sarebbe approssimabile con una delle due densità in figura.

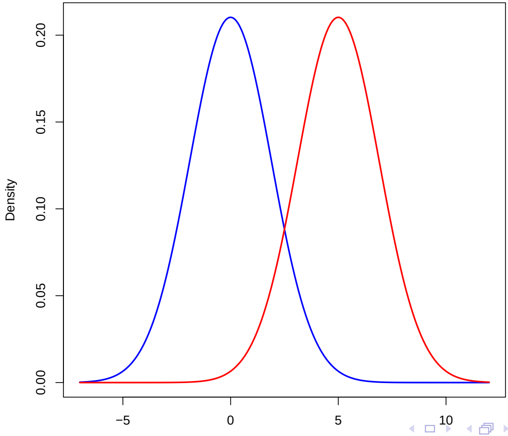


Ipotizziamo di aver estratto casualmente un **campione** di 10 pazienti su ciascuno dei quali verrà osservata la variazione del livello di colesterolo dopo 30 giorni di terapia
In simboli
$(X_1, \ldots, X_{10})$ dove X_i i.i.d. con $X_i \sim N(\mu, \sigma^2)$
Sintetizziamo le nostre esservazioni attraverso una **statistica test** , ovvero una funzione delle (sole) osservazioni campionarie
Scegliamo come **statistica test** la media campionaria, stimatore naturale di μ
Dall'assunzione di un modello normale segue che la **distribuzione campionaria** della nostra statistica test sarà ancora normale, in simboli

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

dove $\sigma = 6mg/dl$

Confrontiamo la distribuzione della media campionaria delle nostre 10 osservazioni sotto l'ipotesi nulla (in blu) e sotto quella alternativa (in rosso) Ricordiamo che sull'asse X abbiamo tutti i valori che la nostra statistica test può assumere al variare del campione nello spazio dei campioni



Le densità campionarie della statistica test sotto l'ipotesi nulla e sotto l'ipotesi alternativa ci consentono di capire quali intervalli sono più probabili sotto le due ipotesi

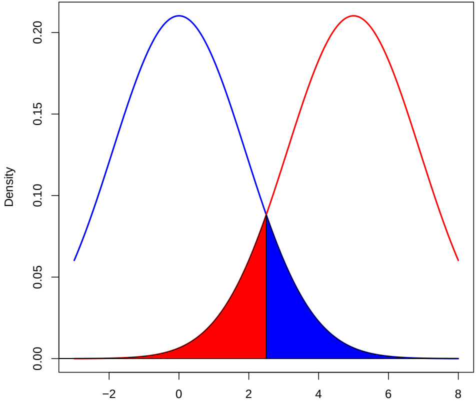
Sulla base di queste probabilità definiremo la regione di rifiuto del nostro test

Un test d'ipotesi infatti è formalmente una regola di decisione che associa ad ogni possibile risultato la scelta di una delle due ipotesi.

Dovremo pertanto individuare la **regione di rifiuto del test**, $\mathcal{R}$, definita come l'insieme dei valori osservabili della media campionaria per i quali andremo a rifiutare H_0

Il complemento della regione di rifiuto sarà la regione di accettazione del test , $\mathcal{A}$

Nel nostro esempio scegliamo come regione di rifiuto $\mathcal{R} = (\overline{x} : \overline{x} > 2.5)$
Quale regola di decisione abbiamo utilizzato?



Una volta definita $\mathcal{R}$ non resta che osservare i nostri dati e decidere di conseguenza.

Come sempre, tuttavia, al risultato finale è necessario affiancare una misura d'errore (anzi due).

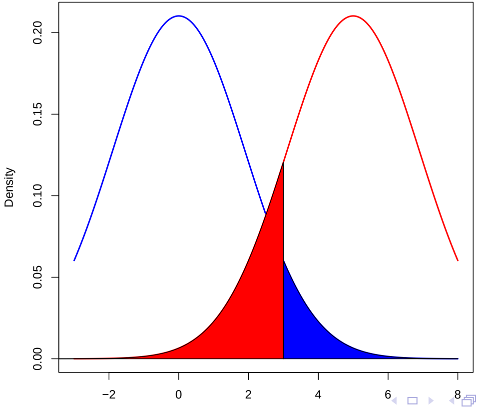
	H_0	H_1
Non rifiutiamo H_0	OK	Errore di seconda specie
Rifiutiamo H_0	Errore di prima specie	OK

Ad ogni tipo di errore resta associata una probabilità di essere commesso che andiamo a valutare

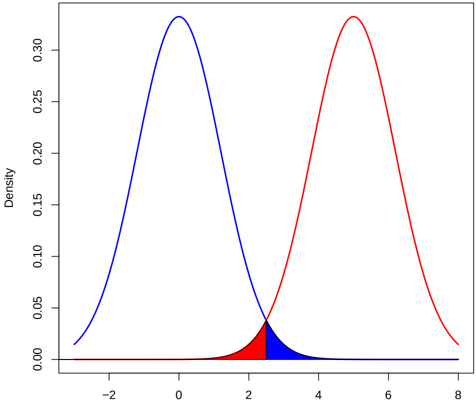
- probabilità di un errore di prima specie o dimensione del test
  $\alpha = Prob\{\overline{X} \, \epsilon \, \mathcal{R} \mid H_0\}$
- probabilità di un errore di seconda specie
  è $1 - \beta = Prob\{\overline{X} \, \epsilon \, \mathcal{A} \mid H_1\}$

dove $\mathcal{A}$ è la regione di accettazione del test (complemento di $\mathcal{R}$).
NB la probabilità $\beta = Prob\{\overline{X} \, \epsilon \, \mathcal{R} \mid H_1\}$ è nota come potenza del test
Nel nostro esempio abbiamo $\alpha = 0.094$ e $1 - \beta = 0.094$

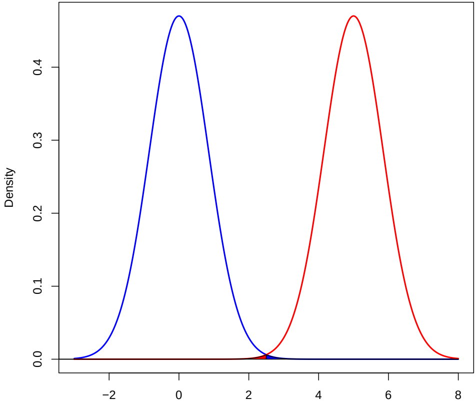
Sarebbe possibile ridurre queste probablità di errore scegliendo diversamente la regione di rifiuto? Potremmo facilmente ridurre α scegliendo $\mathcal{R} = (\overline{x} : \overline{x} > 3)$, cioè diventando più conservativi La diretta conseguenza sarebbe tuttavia un aumento della probabilità di errore di seconda specie; diventando molto conservativi rischieremmo di non riconoscere H_1 anche quando è vera



L'unico modo effettvo di ridurre l'errore è aumentare la dimensione campionaria, di fatto seprando le due distribuzioni campionarie della nostra statistica test Se aumentiamo il numero di osservazioni passando da 10 a 25 avremo $\alpha = 0.018$ e $1 - \beta = 0.018$



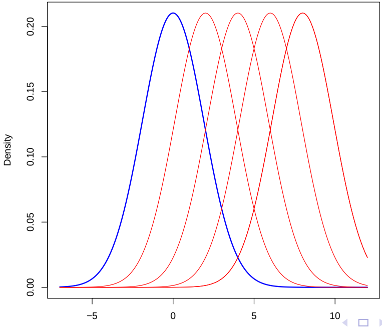
Se aumentiamo ancora il numero di osservazioni passando da 25 a 50 avremo le probabilità di errore saranno vicine a 0, $\alpha = 0.002$ e $1 - \beta = 0.002$



Nella realtà la situazione si complica in primo luogo perché l'ipotesi alternativa è spesso composta

$$\begin{array}{lcll} H_0 & : & \mu = 0\, mg/dl & \text{Ipotesi nulla: nessuna efficacia in media} \\ H_1 & : & \mu > 0\, mg/dl & \text{Ipotesi alternativa: efficacia attesa sempre in media} \end{array}$$

Questo implica che avremo un insieme di distribuzioni campionarie per la nostra statistica test sotto l'ipotesi alternativa, una per ogni possibile valore di $\mu > 0$ e la potenza del test (e la probablità di errore di seconda specie) diventerà una funzione di μ



Inoltre generalmente σ^2 non è noto e sarà necessario ricorrere al suo stimatore

$$S^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$$

La costruzione della nostra statistica test partirà dalla media campionaria che andremo a standardizzare assumendo vera l'ipotesi nulla

$$Z = \frac{\overline{X} - 0}{\sigma/\sqrt{n}}$$

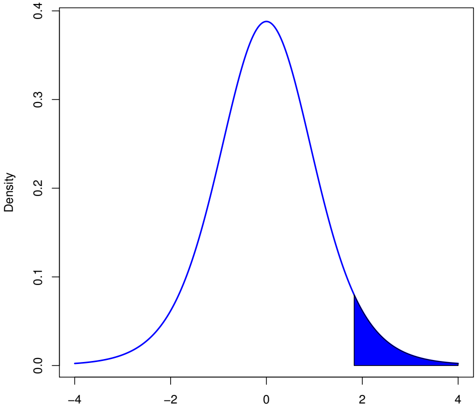
per poi sotituire σ con il suo stimatore

$$T = \frac{\overline{X} - 0}{S/\sqrt{n}}$$

T dipende adesso da due quantità aleatorie e la sua distribuzione campionaria sarà la distribuzione t di Student con n-1 gradi di libertà

Il test ottimale sarà caratterizzato dalla regione di rifiuto $R = \{t \ tali \ che \ t > t_\alpha\}$ dove t_α si ricava dall'uguaglianza $\int_{t_\alpha}^{+\infty} f_T(t)dt = \alpha$

Nel nostro caso $t_\alpha = 1.83$..



Osservati i dati del nostro campione non resterà che calcolare il valore della media camponaria e verificare se si trova o meno in regione di rifiuto Nel primo caso rifiuteremo l'ipotesi nulla per accettare quella alternativa $\mu > 0$ di una effettiva efficacia del farmaco nel ridurre i livelli di colesterolo Ovviamente resterà da stimare il valore di μ

Nel secondo caso non potremo rifiutare l'ipotesi nulla Questo tuttavia non dimostra definitivamente che il farmaco non sia efficace Resta aperta infatti anche la possibilità che il farmaco abbia un'efficacia ma la numerosità del nostro campione sia troppo bassa per riconoscerla