Modulo 3: Inferenza

Prof. Davide Ferrari

Faculty of Economics & Management, UniBz

ARC Centre of Excellence in Maths & Stats

Schema concettuale

$X, \mathcal{P}_N, \theta \longrightarrow \{x_1, \ldots, x_n\} \longrightarrow \widehat{\theta} \longrightarrow$ valido per $\mathcal{P}_N$?

$\Downarrow$ $\Downarrow$

Statistica descrittiva (sintesi)

Inferenza statistica (induzione inversa)

Esempio

Il gestore di un impianto industriale afferma che tale impianto scarica meno di 90 litri all'ora di acqua residua in un lago. Un gruppo locale per la protezione ambientale decide di monitorare l'impianto.

- Si sono raccolte 75 osservazioni:

$$75, 95, 89, 92, \ldots, 91$$

  - Media campionaria: $\overline{x} = 92.1$ l/h
  - Deviazione standard campionaria: $s = 7.1$ l/h

- Possiamo affermare che l'impianto scarichi effettivamente piu' di 90 l/h?

Exampio

In uno dei suoi famosi esperimenti di ibridazione con le piante di pisello, Gregor Mendel ottenne 428 piante con baccello verde e 152 piante con baccello giallo. Secondo la sua teoria, 1/4 delle piante avrebbe dovuto avere il baccello giallo.

- n = 428+152 = 580
  - Frequenza di baccelli gialli: $\widehat{p} = 152/580 = 0.262$
- La teoria di Mendel e' giusta?

I fondamenti dell'inferenza statistica

- Le variabili $X_1, X_2, \ldots, X_n$ sono i.i.d. con una funzione di probabilità (se le X_i sono discrete) o di densità (se le X_i sono continue) che indicheremo in modo generico con $f(x)$.

- La funzione $f(x)$ dipende da uno o più parametri ignoti che rappresentano caratteristiche di interesse della popolazione di riferimento.

$\Longrightarrow$ Obiettivo dell'Inferenza Statistica è utilizzare le osservazioni campionarie per risalire al vero valore dei parametri ignoti e da qui alla distribuzione di probabilità del fenomeno di interesse.

  - Si vuole stimare il vero valore dei parametri ignoti in base alle osservazioni del campione e capire quanto accurata è la stima proposta (**stima puntuale**).
  - Si vuole identificare un insieme di valori ragionevoli per i parametri ignoti (**stima intervallare**).
  - Si formula un'ipotesi sul vero valore dei parametri ignoti e si vuole verificare se tale ipotesi è vera oppure no, in base alle osservazioni campionarie. (**verifica di ipotesi**).

Il campionamento

- Il campione deve essere rappresentativo della popolazione di riferimento.

- Esistono vari metodi di campionamento ma noi considereremo solo il campionamento casuale:

  $\Longrightarrow$ le unità statistiche del campione sono estratte in modo *casuale* dalla popolazione di riferimento

- Se le estrazioni delle unità statistiche che entrano nel campione sono indipendenti si parla di **campione casuale semplice** (c.c.s)
  - se la popolazione è finita le estrazioni devono essere con reinserimento (schema bernoulliano)
  - se la popolazione è infinita oppure è molto più ampia della numerosità campionaria, le estrazioni possono essere con (schema bernoulliano) o senza reinserimento (in blocco)

  In seguito ci limiteremo a considerare campioni casuali semplici (c.c.s.).

I fondamenti dell'inferenza statistica

- $\mathcal{P}_N, N, X$
- $\mathcal{C}_n, n, \{x_1, x_2, \ldots, x_n\}$
- X_i è la generica v.a. che descrive l'insieme dei valori che possono avverarsi quando si assume l'i-esimo elemento campionario e x_i è il valore effettivamente osservato chiamato **realizzazione** di X_i.

- x_1 è la realizzazione della v.a. X_1 che descrive il risultato che osserveremo sulla prima unità estratta, x_2 è la realizzazione della variabile casuale X_2 che descrive il risultato che osserveremo sulla seconda unità estratta e così via.
- X_i coincide con X nella popolazione quindi abbiamo n repliche della variabile X che sono n variabili identicamente distribuite ed, essendo le estrazioni indipendenti parliamo di variabili **i.i.d.**.
- Dopo aver effettuato l'osservazione, un c.c.s. è rappresentabile da una n-pla di valori, $(x_1, x_2, \ldots, x_n)$ ciascuno realizzazione di una v.a.

Campionamento casuale

Un campione di 260 observazioni da una popolazione normale, $N(\mu = 67, \sigma^2 = 12^2)$, e' formata da variabili aleatore indipendenti

$$X_1 \sim N(67, 12^2),\ X_2 \sim N(67, 12^2),\ \ldots,\ X_{260} \sim N(67, 12^2).$$

"PRIMA"

Il *campione osservato* contiene realizzazioni di tali variabili:

$$x_1 = 74.23, \quad x_2 = 61.07, \cdots, \quad x_{260} = 58.93.$$

"DOPO"

Quando si fa inferenza, μ non e' direttamente osservabile ed utilizziamo il campione per stimare μ.

Le statistiche campionarie

- Cerchiamo un criterio per utilizzare i dati del campione per fare inferenza sui parametri della popolazione.
- Le statistiche campionarie sono degli indicatori sintetici da calcolare nel campione che possono darci informazioni sui parametri.

- Una **statistica campionaria** $T = h(X_1, \ldots, X_n)$ è una funzione che dipende solo dai dati del campione e non da quantità incognite. Dato un c.c.s.
  - la media campionaria: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - la varianza campionaria: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
  - la semisomma dei valori estremi: $(X_{\max} - X_{\min})/2$
  - ...
- Prima dell'estrazione del campione T è una v.a. e descrive tutti i possibili valori che T può assumere al variare dei possibili campioni estraibili.
- Dopo l'estrazione del campione, $t = h(x_1, \ldots, x_n)$ non è una v.a., ma è il valore t che la statistica campionaria assume nel campione estratto.

Si consideri un campione i.i.d. tale che

$$X_1 \sim N(67, 12^2),\ X_2 \sim N(67, 12^2),\ \ldots,\ X_{260} \sim N(67, 12^2).$$

Si osserva

$$x_1 = 74.23, \quad x_2 = 61.07, \cdots, \quad x_{260} = 58.93.$$

e quindi

$$\bar{x} = \frac{1}{260}(74.23 + 61.07 + \cdots + 58.93) = 67.31$$

Questa e' una realizzazione della v.a.

$$\bar{X} = \frac{1}{260}(X_1 + X_2 + \cdots + X_{260})$$

Si noti che, $X_1,\ X_2,\ \ldots,\ X_{260}$ sono v.a. e quindi anche $\bar{X}$.

Conoscere la distribuzione di una statistica e' importante per capire come varia e quindi per fare inferenza.

$\bar{X} = \frac{1}{260}(X_1 + X_2 + \cdots + X_{260})$ ha una distribuzione precisa con valore atteso e varianza

$$
\begin{aligned}
E(\bar{X}) &= \frac{1}{260} E(X_1 + X_2 + \cdots + X_{260}) \\
&= \frac{1}{260} \big(\mathrm{E}(X_1) + E(X_2) + \cdots + E(X_{260})\big) \\
&= \frac{1}{260} \big(67 + 67 + \cdots + 67\big) = 67 \\
&= \frac{1}{n} \big(\mu + \mu + \cdots + \mu\big) = \mu
\end{aligned}
$$

$$
Var(\bar{X}) = \left(\frac{1}{260}\right)^2 Var(X_1 + X_2 + \cdots + X_{260})
$$

$$
= \left(\frac{1}{260}\right)^2 (Var(X_1) + Var(X_2) + \cdots + Var(X_{260}))
$$

$$
= \left(\frac{1}{260}\right)^2 Var(12^2 + 12^2 + \cdots + 12^2) = \frac{12^2}{260}
$$

$$
= \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}
$$

But wait, there's more!

Il Theorema del Limite Centrale ci dice che:

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Se la distribuzione di X non e' troppo strana e n abbastanza grande.

La v.a. media campionaria

- La v.a. descritta dai valori medi di X, v.a. con media μ e varianza σ^2, la v.a. media campionaria:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad E(\bar{X}) = \mu \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$$

- Si noti che la dispersione della media campionaria intorno al baricentro è tanto ridotta quanto maggiore è la dimensione del campione.

- Se $X \sim N(\mu, \sigma^2)$ la distribuzione della media campionaria è

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

- Se X non proviene da $N(\mu, \sigma)$ la distribuzione della media campionaria è

$$\bar{X} \approx N\left(\mu, \sigma^2/n\right)$$

La v.a. varianza campionaria

- La v.a. descritta dalle varianze di X, v.a. con media μ e varianza σ^2, in $\mathcal{U}_n$ è la v.a. varianza campionaria. La varianza campinaria "non corretta" e'

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad E(\widehat{\sigma}^2) = \sigma^2\left(\frac{n-1}{n}\right)$$

- $\frac{n-1}{n}$ è il fattore usato per "correggere" la v.a. varianza campionaria. La varianza campionaria corretta diventa:

$$S^2 = \widehat{\sigma}^2\frac{n}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

che ha

$$E(S^2) = \sigma^2$$

- Considerata la seguente trasformata di S^2, $\frac{(n-1)S^2}{\sigma^2}$ si ha che

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

La v.a. proporzione campionaria

- La v.a. descritta dalle frequenze relative di $X \sim Ber(p)$ è la v.a. proporzione campionaria:

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

che ha

$$E(\widehat{p}) = p \qquad Var(\widehat{p}) = \frac{p(1-p)}{n}$$

- Si noti che la dispersione della media campionaria intorno al baricentro è tanto ridotta quanto maggiore è la dimensione del campione.

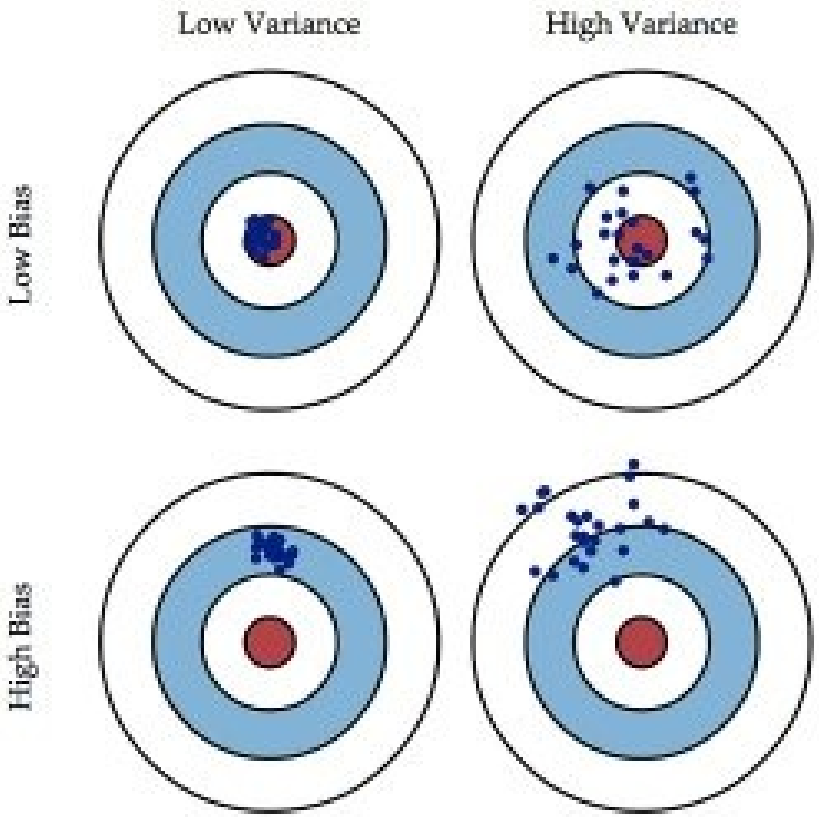
- La distribuzione approssimata della proporzione campionaria è

$$\widehat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

Stima parametrica: teoria fisheriana-neymaniana

- Stima parametrica: stima uno o più parametri incogniti di popolazione note le osservazioni campionarie e un modello funzionale, e.g. una funzione di densità di probabilità.
- Stima ogni parametro incognito di popolazione identificando un valore puntuale oppure un range di valori che con una certa fiducia contiene il parametro.
- Il problema della stima puntuale può ricondursi alla scelta di una statistica campionaria T opportuna per stimare il valore del parametro ignoto.
- La statistica campionaria T usata per stimare il parametro ignoto θ è chiamata **stimatore**. Lo stimatore è quindi una v.a. con una sua distribuzione campionaria che varia nell'universo dei campioni.
- La realizzazione $t = h(x_1, \ldots, x_n)$ dello stimatore sul campione effettivamente estratto viene chiamata **stima** puntuale.
- L'accuratezza della stima puntuale dipende dalla **deviazione standard** dello stimatore $SD(T) = \sqrt{Var(T)}$.

Proprietà degli stimatori: Bias e varianza



Proprietà degli stimatori: la correttezza

- Uno stimatore T per θ si dice **corretto** se il suo valore atteso coincide con il parametro di popolazione, cioè

$$E(T) = \theta$$

- La distribuzione campionaria di uno stimatore corretto è centrata attorno a θ.

- La correttezza di uno stimatore garantisce che la media delle stime ottenute su tutti i c.c.s. di dimensione n che possiamo estrarre dalla popolazione sia uguale al parametro ignoto.

- Seppure la correttezza è una proprietà auspicabile, essa fornisce solo delle garanzie in media e non ci assicura che la stima ottenuta sul c.c.s. effettivamente estratto sia uguale a θ (e anche solo vicina).

- Esempi: $T = \bar{X}$, $T = \frac{1}{2}(X_1 + X_n)$, $T = S^2$

Proprietà degli stimatori: la correttezza

- Uno stimatore T per θ si dice **distorto** se

$$Bias(T) = E(T) - \theta \neq 0$$

- Uno stimatore T per θ è **asintoticamente corretto** se al limite il suo valore atteso coincide con il parametro di popolazione, cioè se

$$E(T) \to \theta, \quad n \to \infty$$

- Se utilizzassimo uno stimatore distorto, ossia tale che $E(T) < \theta$ o $E(T) > \theta$, allora in media le stime ottenute su tutti i possibili c.c.s. sottostimerebbero o sovrastimerebbero, rispettivamente, il vero valore del parametro.

- Esempio: $\widehat{\sigma}^2 = \frac{1}{n} \sum_i (X_i - \bar{X})^2$.

Proprietà degli stimatori: efficienza

- Si definisce **errore quadratico medio** (mean squared error) di uno stimatore T per θ

$$\begin{aligned} MSE(T) &= E(T-\theta)^2 \\ &= Var(T) + Bias(T)^2 \end{aligned}$$

  che può essere interpretato come la distanza media di T da θ poichè misura di quanto in media le realizzazioni di T distano da θ.
- E' auspicabile, pertanto, scegliere uno stimatore con MSE piccolo.
- Lo stimatore T_1 per θ si dice **più efficiente** di T_2 se il suo errore quadratico medio è non superiore a quello di T_2, cioè

$$MSE(T_1) \leq MSE(T_2) \qquad \forall\, \theta \in \Theta$$

- L'MSE è pertanto dato da una componente legata alla posizione della distribuzione di T, la distorsione di T, e da una componente legata alla variabilità della distribuzione di T.
- Esempi: $MSE(\bar{X})$, $MSE\left(\frac{1}{2}(X_1 + X_n)\right)$

Proprietà degli stimatori: consistenza

- Uno stimatore T per θ si dice **consistente** se

$$MSE(T) = \left[Var(T) + Bias(T)^2\right] \to 0, \quad n \to \infty$$

  cioè se la sua accuratezza migliora all'aumentare della numerosità campionaria.
- La consistenza di uno stimatore è una proprietà asintotica, si analizza cioè cosa succede alla distribuzione campionaria di T per $n \to \infty$.
- L'idea di base è che una proprietà asintotica sarà almeno approssimativamente soddisfatta in campioni finiti, purchè sufficientemente grandi.
- Se T è uno stimatore corretto per θ, allora $MSE(T) = Var(T)$ e lo stimatore è consistente se

$$Var(T) \to 0, \quad n \to \infty$$

  ossia la distribuzione di T, all'aumentare di n, diventa sempre più concentrata attorno al suo valore atteso che, per la correttezza, coincide con θ.

- Esempi: $\bar{X}$, S^2, frequenza, ma anche $\widehat{\sigma}^2$

Stima puntuale della media

- Sia estratto e osservato il campione $(x_1, \ldots, x_n)$, la stima puntuale della media di popolazione μ incognita è

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

e l'accuratezza della stima di μ è data dalla deviazione standard dello stimatore

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Dato che σ^2 di solito non è nota, sostituiamo $\sigma^2 \approx s^2$ ottendendo l'errore standard

$$SE = \frac{s}{\sqrt{n}} \approx SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- $\bar{X}$ è uno stimatore corretto e consistente per μ. Se $X_1, \ldots, X_n$ sono i.i.d. $N(\mu, \sigma^2)$, allora $\bar{X}$ è anche lo stimatore migliore (con MSE più piccolo) tra tutti gli stimatori corretti di μ.

Stima puntuale della varianza

- Sia estratto e osservato il campione $(x_1, \ldots, x_n)$, la stima puntuale della varianza di popolazione σ^2 incognita è

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

  Un calcolo mostra che

$$SD(S^2) = \sigma^2 \sqrt{\frac{2}{n-1}}$$

  e quindi sostituendo $\sigma^2 \approx s^2$ l'accuratezza della stima di σ^2 è data dall'errore standard

$$SE = s^2 \sqrt{\frac{2}{n-1}} \approx SD(S^2)$$

- S^2 è uno stimatore corretto e consistente per σ^2. Se $X_1, \ldots, X_n$ sono i.i.d. $N\left(\mu, \sigma^2\right)$, allora S^2 è anche lo stimatore con MSE più piccolo.

Stima puntuale della proporzione

- Sia $X \sim Ber(p)$ con p incognito. Sia estratto e osservato il campione $(x_1, \ldots, x_n)$, la stima puntuale di p è

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

  l'errore standard di $\widehat{p}$ è

$$SE = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

- $\widehat{p}$ è uno stimatore corretto e consistente di una probabilità di successo. E' anche lo stimatore con il più piccolo MSE nella classe degli stimatori corretti di p.

Intervalli di confidenza

Uno stimatore e' una statistica usata per stimare un parametro. E' una v.a., ad esempio $\bar{X} \sim (55.4, 15.2^2)$.

Una stima e' la realizzazione dello stimatore. E' un numero: ad esempio $\bar{x} = 56.82$

Una stima puntuale e' utile ma sarebbe piu' utile capire esattamente quanto vicini siamo al parametro. Quindi si costruisce una stima intervallare (= intervallo di confidenza).

La precisione e' data dall'errore standard. $SE \approx SD(\bar{X}) = \frac{s}{\sqrt{n}} = 4.81$

Potremmo usare questa informazione per definire un intervallo di "valori plausibili" per μ. Ovvero quali valori di μ potrebbero avere portato ad osservare observed value $\bar{x} = 56.82$? Per esempio un intervallo ptrebbe essere

$$\bar{x} \pm 2 \times SE,$$

ovvero (47.2, 66.44). Ma perche' "$2 \times SE$" e non "$1 \times SE$"?

Consideriamo $n = 20$ osservazioni da $X \sim N(\mu, 12^2)$. Allora, $\bar{X} \sim (\mu, \frac{12^2}{20})$.

$$P(\mu - 1.96\frac{12}{\sqrt{20}} < \bar{X} < \mu + 1.96\frac{12}{\sqrt{20}}) = 0.95$$

$$P(\bar{X} - 1.96\frac{12}{\sqrt{20}} < \mu < \bar{X} + 1.96\frac{12}{\sqrt{20}}) = 0.95$$

$$P(\bar{X} - 5.26 < \mu < \bar{X} + 5.26) = 0.95$$

Quindi l'intervallo aleatorio $\bar{X} \pm 5.26$ contiente μ con probabilità 0.95. Una stima intervallare e' data da:

$$\bar{x} \pm 5.26$$

Intervallo di confidenza (IC)

$n = 20$ observations on $X \sim N(\mu, 12^2)$.

$\bar{X} \sim N(\mu, \frac{12^2}{20})$ $\qquad$ $P(\bar{X} - 5.26 < \mu < \bar{X} + 5.26) = 0.95$

realizzazione osservazione

realizzazione osservazione

stima $= \bar{x} = 72.06$ $\qquad$ CI $= (72.06 \pm 5.26) = (66.80, 77.32)$

stima puntuale

stima intervallare

Questo e' l'IC al 95% per μ

Livello di confidenza $(1-\alpha)\%$

level	quantile	confidence interval
99.9%	$z_{0.9995} = 3.2905$	$\bar{x} \pm 3.2905 \frac{12}{\sqrt{20}} = (63.2, 80.9)$
99%	$z_{0.995} = 2.5758$	$\bar{x} \pm 2.5758 \frac{12}{\sqrt{20}} = (65.1, 79.0)$
95%	$z_{0.975} = 1.9600$	$\bar{x} \pm 1.9600 \frac{12}{\sqrt{20}} = (66.8, 77.3)$
90%	$z_{0.95} = 1.6449$	$\bar{x} \pm 1.6449 \frac{12}{\sqrt{20}} = (67.6, 76.5)$
80%	$z_{0.90} = 1.2816$	$\bar{x} \pm 1.2816 \frac{12}{\sqrt{20}} = (68.6, 75.5)$
50%	$z_{0.75} = 0.6745$	$\bar{x} \pm 0.6745 \frac{12}{\sqrt{20}} = (70.3, 73.9)$
0%	$z_{0.5} = 0.0000$	$\bar{x} \pm 0.0000 \frac{12}{\sqrt{20}} = (72.1, 72.1)$

Un intervallo di confidenza rappresenta un insieme di valori per μ che sono "compatibili" con la media campionaria osservata $\bar{x}$.

$$\mu \quad \longrightarrow \quad (\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}})$$

"valori plausibili" $\bar{x}$

[intervallo di probabilita' al 95%]

Un intervallo di confidenza rappresenta un insieme di valori per μ che sono "compatibili" con la media campionaria osservata $\bar{x}$.

$$\mu \quad \longrightarrow \quad (\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}})$$

"valori plausibili" $\bar{x}$
[intervallo di probabilita' al 95%]

Se osservo $\bar{x}$, quali sono valori plausibili di μ?

Ovvero valori μ che potrebbero avere indotto l'osservare $\bar{x}$

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}) \quad \longleftarrow \quad \bar{x}$$

"valori plausibili" di μ
[Intervallo di confidenza al 95%]

Un'industria salinara produce sacchi di sale che hanno un peso distribuito normalmente con media μ incognita e varianza nota 0.6. Si estrae un campione di 20 sacchi e se ne registra il peso. Il peso medio risulta pari a 7 Kg. Qual è l'intervallo di confidenza per μ di livello 95%? E al 99%? Come si potrebbe ridurre il margine di errore commesso? Si utilizzi il seguente output

```
> x <- c(0.9, 0.95, 0.975, 0.99, 0.995)
> qnorm(x)
[1] 1.281552 1.644854 1.959964 2.326348 2.575829
```

Intervallo al 95% di confidenza per μ :

$$\bar{x} \pm z_{0.975} \frac{\sigma}{\sqrt{n}}, \Rightarrow 7 \pm 1.96 \sqrt{\frac{0.6}{20}}$$

Ovvero $(6.66, 7.34)$. Intervallo al 99% di confidenza per μ :

$$\bar{x} \pm z_{0.995} \frac{\sigma}{\sqrt{n}} \Rightarrow 7 \pm 2.58 \sqrt{\frac{0.6}{20}}$$

Ovvero $(6.55, 7.46)$.

Intervalli di confidenza per p

Se n e' abbastanza grande abbiamo:

$\widehat{p} \approx N\big(p, \frac{p(1-p)}{n}\big)$, [come $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$].

$SD(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}$, e quindi $SE = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$.
Questo implica che

$$0.95 = P\left(\widehat{p} \pm 1.96\frac{p(1-p)}{n} \leq p \leq \widehat{p} \pm 1.96\frac{p(1-p)}{n}\right)$$

ed il corrispondente intervallo di confidenza al 95% e'

$$\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Exampio

Gregor Mendel ottenne 428 piante con baccello verde e 152 piante con baccello giallo. Secondo la sua teoria, 1/4 delle piante avrebbe dovuto avere il baccello giallo.

- n = 428+152 = 580.
- Frequenza di baccelli gialli: $\widehat{p} = 152/580 = 0.262$
- Intervallo di confidenza al 95%:

$$0.262 \pm 1.96\sqrt{\frac{0.262(1-0.262)}{580}}$$

  Ovvero $(0.226, 0.298)$.
- La teoria di Mendel e' giusta? I dati non ci forniscono sufficiente evidenza per poter rigettare tale teoria.

IC approssimati per stimatori normalmente distribuiti

Un costruttore decide di acquistare una partita di listelli di alluminio se il numero medio di difetti in una partita di 35 listelli e' inferiore a 2.1 difetti per listello. Si assuma che $X =$"il numero di difetti per listello" segue una distribuzione Poisson $X \sim Poi(\lambda)$. Dal seguente campione di grandezza $n = 35$.

7	4	5	5	5	9	5	2	3	4	3	5	5	7	4	9	4	7
4	5	8	6	3	2	5	6	8	4	7	4	3	4	9	6	10	

si calcola $\bar{x} = 5.34$.

(a) Utilizzando il teorema del limite centrale, si derivi un intervallo di confidenza per λ.

(b) Si calcoli l'intervallo di confidenza utilizzando i dati ed un livello di confidenza pari a 95%.

(c) Al costruttore conviene comprare i listelli? Si giustifichi la propria risposta.

Dato che $E(X) = Var(X) = \lambda$, il TLC ci dice che

$$\bar{X} \approx N(\lambda, \lambda/n)$$

Si noti che uno stimatore non distorto per λ e' dato da $\bar{X}$. Quindi la stima di λ e'

$$\bar{x} = 5.34$$

Dato che l'errore standard e' $SE = \bar{x}/n$, l'IC al 95% e' quindi dato da

$$\bar{x} \pm 1.96\sqrt{\frac{\bar{x}}{n}}$$

Quindi abbiamo $5.34 \pm 1.96\sqrt{5.34/35}$, ovvero $(4.58, 6.10)$. Sulla base di questo campione siamo confidenti al 95% che il numero di difetti medio λ e' maggiore di 2.1. Quindi al produttore non conviene comprare la partita di listelli.

IC per μ se σ non e' nota

Euando σ e' noto, possiamo usare il fatto che

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \text{ oppure } \approx N(0,1)$$

e calcolare ricavare quantili dalla seguente equazione

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

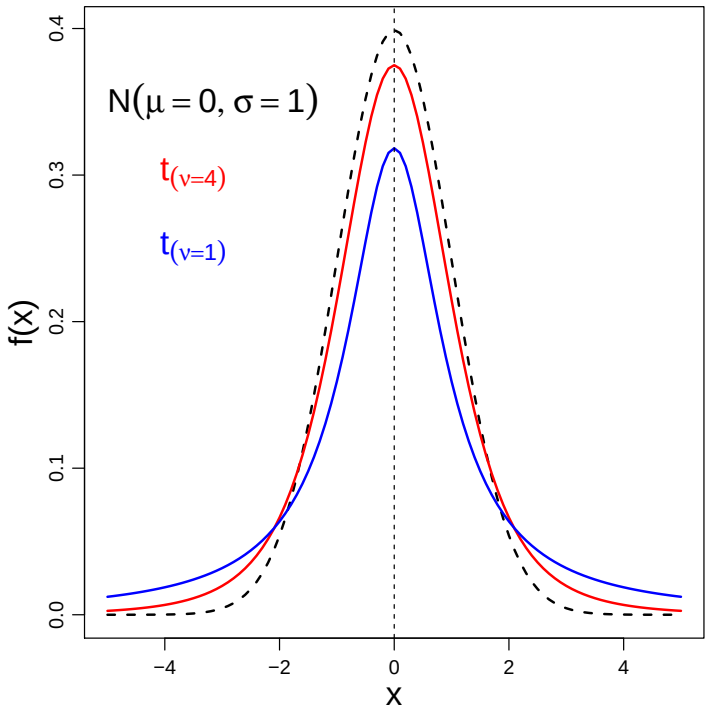
da cui l'intervallo $\bar{x} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$. Cosa facciamo se σ non e' noto?

Utilizziamo S al posto di σ ottenendo la quantita'

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim ????$$

Tuttavia, questa statistica non segue piu' $N(0,1)$. Segue invece una distribuzione $t-$student con $\nu = n - 1$ gradi di liberta'

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$



Basandoci sul fatto che

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

i quantili nella seguente equazione vengono tratti dalla t-student

$$1 - \alpha = P\left(t_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq t_{1-\alpha/2}\right),$$

L'intervallo di confidenza con livello di confidenza $(1-\alpha)\%$ sara' quindi dato da

$$\bar{x} \pm t_{1-\alpha/2}\sqrt{\frac{s}{n}}.$$

Esempio

Il gestore di un impianto industriale afferma che tale impianto scarica meno di 90 litri all'ora di acqua residua in un lago. Un gruppo locale per la protezione ambientale decide di monitorare l'impianto.

- Si sono raccolte 10 osservazioni:

  95 94 90 95 102 110 93 93 111 93

  - Media campionaria: $\overline{x} = 97.6$ l/h
  - Deviazione standard campionaria: $s = 7.4$ l/h
- Possiamo affermare che l'impianto scarichi effettivamente piu' di 90 l/h?

Rispondiamo utilizzando un intervallo di confidenza al 95%:

```
> t <- c(0.90, 0.95, 0.975, 0.995)
> qt(t, df =9)
[1] 1.383029 1.833113 2.262157 3.249836
```

$$\bar{x} \pm t_{1-\alpha/2} \frac{s}{\sqrt{n}}, \quad \Rightarrow 97.6 \pm 2.26 \frac{7.4}{\sqrt{10}}$$

Ovvero $(92.3, 102.8)$. Questi dati ci forniscono evidenza sul fatto che l'impianto scarica piu' di 90 l/h nel lago.

Attenzione!

- Se n e' piccolo ($n \leq 25$), si richede che $X \sim N(0,1)$, affinche' T segua una t-student con n gradi di liberta'.
- Se n e' grande e $X \sim N$, possiamo usare i quantili della t-student che sono circa uguali a quelli della $N(0,1)$
- Se n e' grande e X non segue una normale, possiamo usare i quantili della della $N(0,1)$.

Verifica di ipotesi: esempi

Il direttore di produzione di una certa azienda ha proposto ad un consulente esterno di valutare un nuovo processo produttivo per produrre un certo prodotto. L'attuale processo ha una media di 70 unità orarie e il direttore sostiene di non voler passare al nuovo processo a meno che non ci sia una forte evidenza empirica che questo determini un aumento della produzione media oraria.

- La decisione si basa sull'osservazione di un campione di n ore di produzione con il nuovo processo per le quali si calcola la media $\bar{x}$ del numero di unità prodotte;
- dato che non si conosce l'intera popolazione, la decisione deve tenere conto dell'incertezza dovuta alla stima campionaria $\bar{x}$;
- per decidere se adottare o meno il nuovo processo produttivo sulla base del campione, è necessario definire una regola che tenga conto dell'errore campionario $\bar{x} - \mu$.

Una società di ricerche di mercato vuole verificare se i clienti dei supermercati siano consapevoli dei prezzi degli articoli comprati. La società ritiene che almeno la metà dei clienti debba essere consapevole dei prezzi degli articoli per evitare di avvviare una campagna di informazione.

- La decisione si basa sull'osservazione di un campione di n clienti su cui si osserva una proporzione $\widehat{p}$ di clienti consapevoli dei prezzi degli articoli e deve tener conto dell'errore campionario $\widehat{p} - p$;
- si vuole verificare, sfruttando l'informazione campionaria, se almeno la metà dei clienti sia in grado di ricordare i prezzi degli articoli acquistati, cioè l'ipotesi $H_0 : p < 0.5$ contro la sua alternativa $H_1 : p \geq 0.5$.

Online shopping

Una catena di supermercati che vuole espandersi nell'area di Bolzano progetta il lancio di un nuovo servizio di ordine on-line (tramite App) e recapito a domicilio.

L'azienda si aspetta che il nuovo servizio porti profitto se la spesa media per ordine e' di almeno 90 euro.

Online shopping

- A 75 clienti campionati causalmente viene offerto un incentivo per accedere al servizio. Si registrano le seguenti :

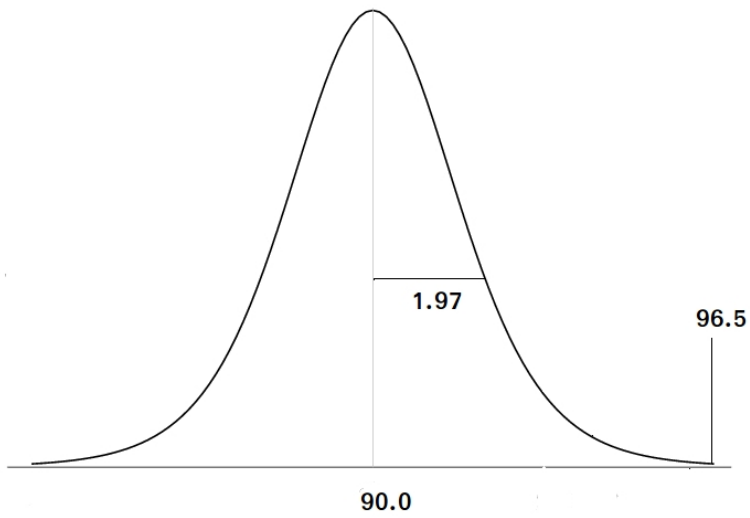
$$125.4, 51.1, 89.6, 92.1, \ldots, 65.5$$

  - Media campionaria e deviazione st campionaria: $\overline{x} = 96.5$ and $s = 17.1$
  - Possiamo affermare che i clienti spendono in media piu' di 90 euros per ordine?

How likely is 96.5?

- Assumiamo che la media per l'intera popolazione sia $\mu = 90$
- Al crescere di n sappiamo che

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \approx N\left(90, \frac{17.1^2}{75}\right) = N\left(90, 1.97^2\right).$$



- La probabilita' di osservare un valore della statistica di 96.5 o maggiore di 96.5 e' molto piccola (circa 0).
- Di conseguenza, abbiamo molta evidenza contro l'assunzione iniziale.

Ingredienti per un test d'ipotesi formale:

1. Sistema di ipotesi
2. Statistica test
3. P-value
4. Significatività e conclusione
5. (Ingredente segreto)

1 Ipotesi

- Ipotesi nulla (si scrive H_0). Affermazione teorica circa i **vero valore** di un parametro (μ, p or σ). Rappresenta il caso dove nulla di interessante prende luogo.

$$H_0 : \mu = 90$$

  Nota: questo equivale a dire $H_0 : \mu \leq 90$.

- Ipotesi alternativa (si scrive H_1). Afferma che il parametro cade in range alternativo di valori values. Rappresenta il fenomeno che siamo interessati a dimostrare.

$$H_1 : \mu > 90$$

1 Ipotesi

In generale:

Ipotesi nulla
$H_0 : \mu = \mu_0$

Ipotesi alternative	Tipo di test
$H_1 : \mu > \mu_0$	Destro
$H_1 : \mu < \mu_0$	Sinistro
$H_1 : \mu \neq \mu_0$	Bilaterale

2 Test statistic

Diversi parametri corrispondono a stime puntuali:

$$\{\mu, \overline{x}\},\ \{p, \widehat{p}\},\ \{\sigma, s\}$$

La statistica test misura la distanza tra il la stima puntuale del parametro ed il valore del parametro assumendo che H_0 sia vera. Normalmente tale distanza viene misurata in termini di errore standard (ovvero si standardizza).

$$\frac{\text{Stima del parametro} - \text{Valore parametro sotto } H_0}{\text{Errore standard della stima}}$$

Per la media μ, possiamo utilizzare la statistica t

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} = \frac{96.5 - 90}{17.1/\sqrt{75}} = 3.29$$

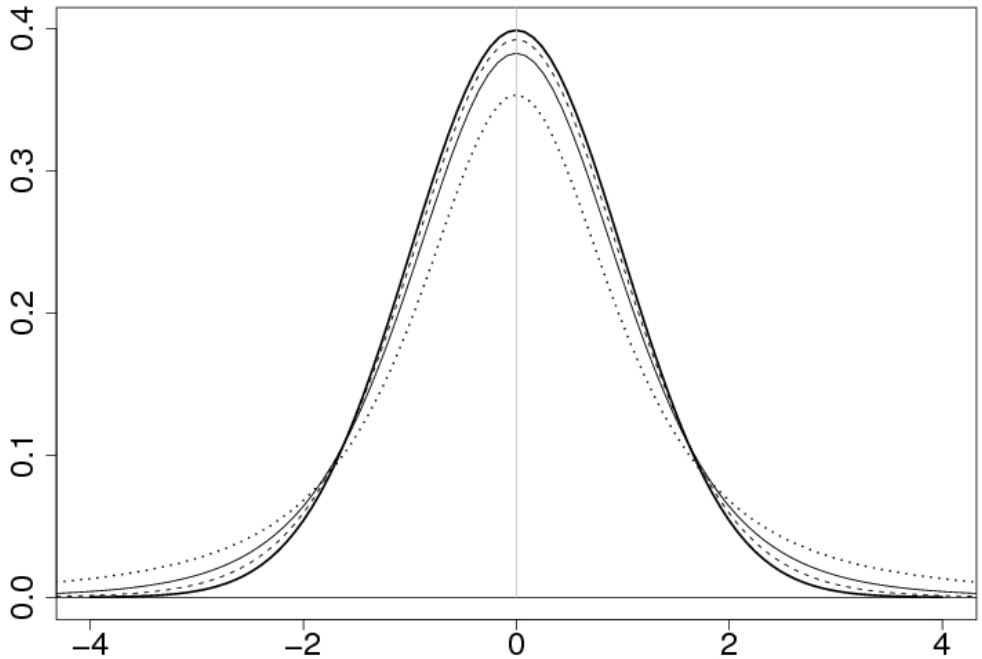
3 P-value

- Probabilita' riassuntiva dell'evidenza contro H_0, che utilizziamo per interpretare il risultato della statistica.
- Il p-value e' la probabilità di ottenere una statistica tanto estrema o piu' estrema di quella osservata nei dati.
- Per la media della popolazione usiamo

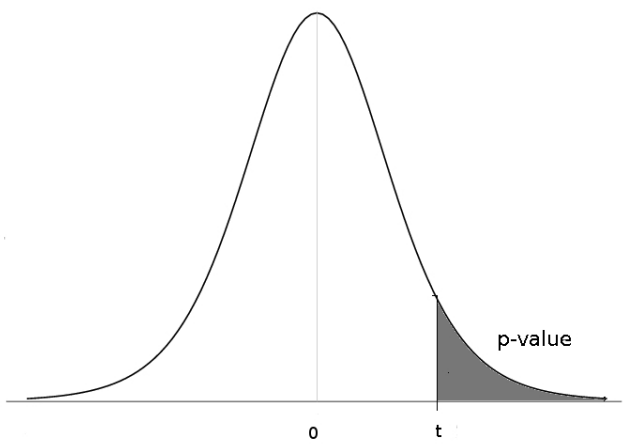
$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

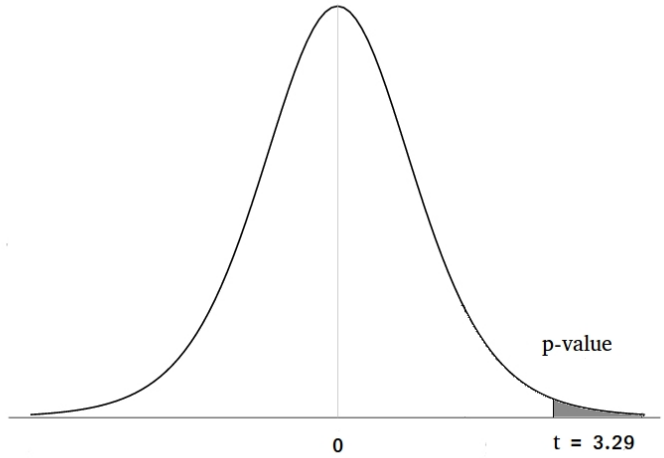
ovvero una distribuzione t-student con $n - 1$ gradi di liberta'.

La distribuzione t-student



3 P-value





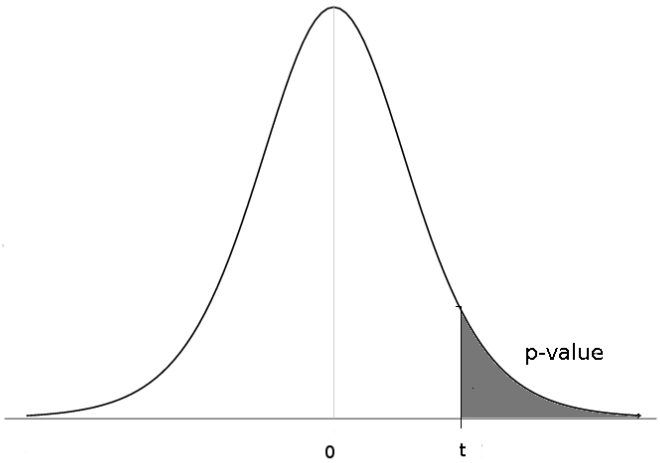
$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

$$\frac{96.5 - 90}{17.1/\sqrt{75}} = 3.29$$

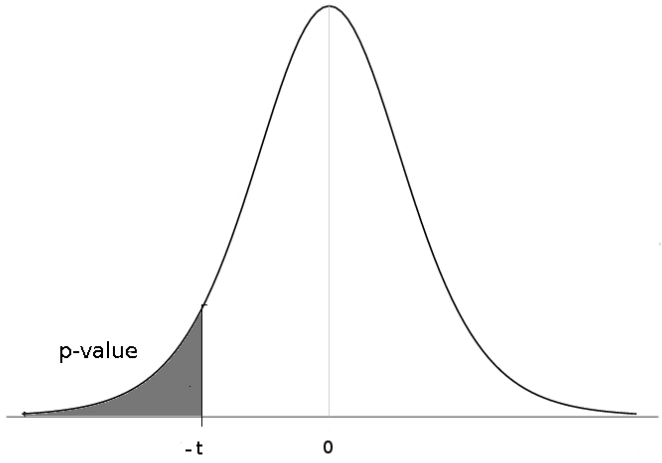
p-value $= P(T_{n-1} > t)$

p-value $= P(T_{74} > 3.29) = 0.00077$

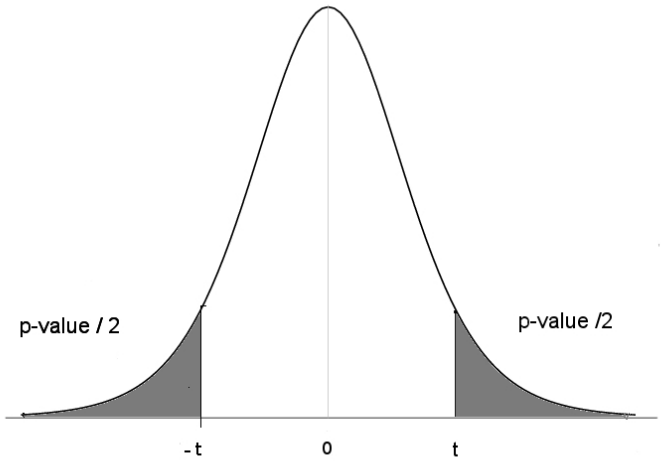
Test destro. $H_1 : \mu > \mu_0$.



Test sinistro. $H_1 : \mu < \mu_0$.



Test bilaterale. $H_1 : \mu \neq \mu_0$



4 Decisione/conclusione

Significatività. Il livello di Significatività, α, e' una soglia decisionale definita come

$$\alpha = P(\text{Errore di tipo I}) = P(\text{ Rifiutare } H_0 | H_0 \text{e' vera })$$

Per decidere, si paragona α al p-value. Typically,

$$\alpha = 0.05 \text{ or } \alpha = 0.01.$$

Possiamo dormire sonni tranquilli con 1 errore su 20 ($\alpha = 0.05$)?

4 Decisione/conclusione

Significatività. Il livello di Significatività, α, e' una soglia decisionale definita come

$$\alpha = P(\text{Type I Error}) = P(\text{ Reject } H_0 | H_0 \text{e' vera })$$

Per decidere, si paragona α al p-value. Tipicamente,

$$\alpha = 0.05 \text{ or } \alpha = 0.01.$$

Possiamo dormire sonni tranquilli con 1 errore su 20 ($\alpha = 0.05$)?
Decisione su H_0:

1. p-value $< \alpha$. I dati mostrano forte evidenza contro H_0.
   Quindi rigettiamo H_0.
2. p-value $> \alpha$. I dati NON mostrano forte evidenza contro H_0.
   Quindi non possiamo rigettare H_0.

4 Decisione/conclusione

Esempio (E-shopping). Esegui un test al 5% di significatività. Se $\alpha = 0.05$

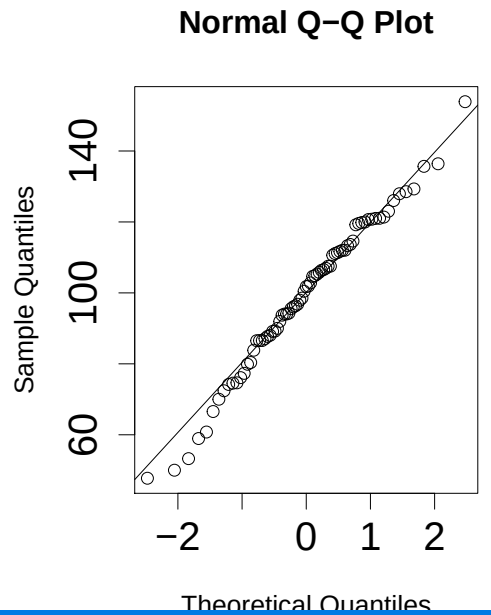
$$\text{p-value} = 0.00077 < \alpha = 0.05.$$

Conclusione: rigettiamo H_0, ovvero che l'ordine medio non supera 90 euro. Quindi concludiamo che probabilmente l'ordine medio supera 90 euro.

5 Ingrediente segreto

Assunzioni! Ogni test si basa su assunzioni sul processo generatore dei dati.
Per il t-test su μ abbiamo bisogno delle seguenti assunzioni:

1. Variabile quantitativa
2. Dati campionati casualmente
3. La popolazione da cui campioniamo e' approssimativamente normale (oppure disponiamo di campione sufficientemente grande)



Test per la media con varianza nota

- Sia $(x_1, x_2, \ldots, x_n)$ un campione casuale da $X \sim N\left(\mu; \sigma^2\right)$ con σ^2 nota. Si vuole verificare al livello di significatività α il sistema di ipotesi

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

- Sappiamo che, in generale, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

- Se H_0 è vera, si ha che

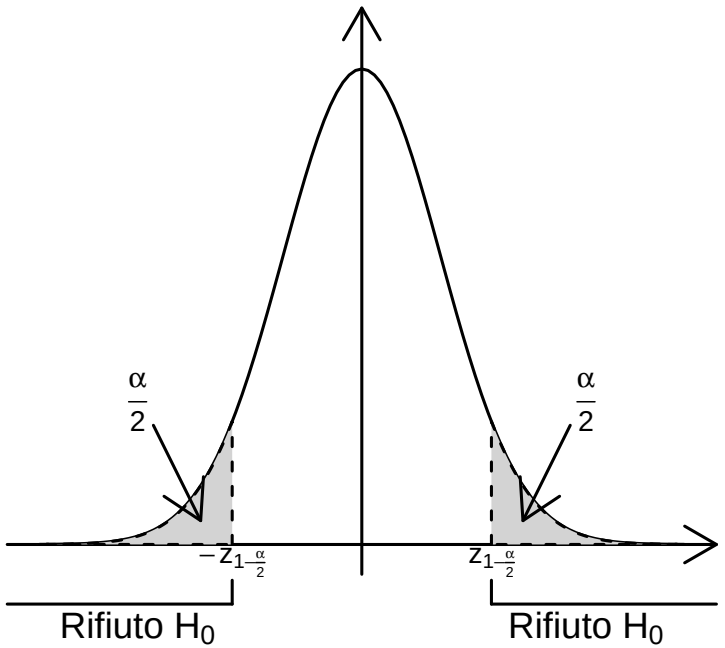
$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

  che è la statistica test per μ.

- Sia $\bar{x}$ il valore di μ nel campione estratto e $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ la sua versione standardizzata sotto H_0.

Test per la media e zone di rifiuto e accettazione

- I valori critici sulla $\mathcal{N}(0;1)$ sono: $\pm z_{1-\frac{\alpha}{2}}$. Quindi:
  - $A = \left(-z_{1-\frac{\alpha}{2}}, +z_{1-\frac{\alpha}{2}}\right)$
  - $R = \left(-\infty, -z_{1-\frac{\alpha}{2}}\right) \cup \left(z_{1-\frac{\alpha}{2}}, +\infty\right)$.
- Si rifiuta H_0 se $z < -z_{1-\frac{\alpha}{2}}$ oppure $z > z_{1-\frac{\alpha}{2}}$.



Test per la media e zone di rifiuto e accettazione

- Se il sistema di ipotesi da saggiare è il seguente

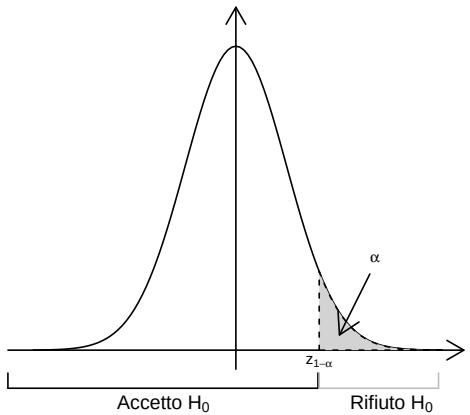
$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

allora il valore critico sulla $N(0;1)$ è $z_{1-\alpha}$.

Quindi:

  - $A = (-\infty, z_{1-\alpha})$
  - $R = (z_{1-\alpha}, +\infty)$

e si rifiuta H_0 se $z > z_{1-\alpha}$.



Test per la media e zone di rifiuto e accettazione

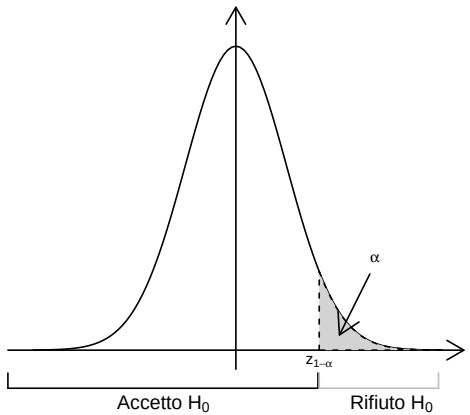
- Se il sistema di ipotesi da saggiare è il seguente

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

allora il valore critico sulla $N(0;1)$ è $z_{1-\alpha}$.
Quindi:

  - $A = (-\infty, z_{1-\alpha})$
  - $R = (z_{1-\alpha}, +\infty)$

e si rifiuta H_0 se $z > z_{1-\alpha}$.



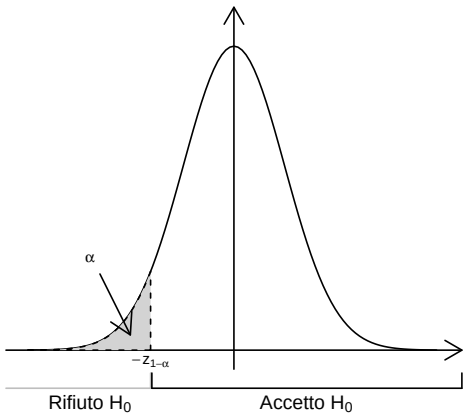
- Se il sistema di ipotesi da saggiare è il seguente

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

allora il valore critico sulla $N(0;1)$ è $-z_{1-\alpha}$. Quindi:

  - $A = (-z_{1-\alpha}, +\infty)$
  - $R = (-\infty, -z_{1-\alpha})$

e si rifiuta H_0 se $z < -z_{1-\alpha}$.



Esempio

Una catena di fast-food controlla ogni giorno che il suo panino gigante pesi in media 700 gr. L'ipotesi alternativa è che il peso medio sia inferiore a 700 gr. Si seleziona in maniera casuale un campione di $n = 20$ panini giganti che risultano avere un peso medio pari a 665 gr.

- Si verifichi il sistema di ipotesi a cui la catena è interessata sapendo che il peso dei panini ha distribuzione normale con deviazione standard pari a 60 gr. e scegliendo $\alpha = 0.05$.
- Si calcoli il p-value.
- Quale sarebbe la regola di decisione se $n = 10$? E se $n = 40$? Che cosa accade all'aumentare della dimensione campionaria?

Test per la media con varianza ignota

- Sia $(x_1, x_2, \ldots, x_n)$ un campione casuale da $X \sim N\left(\mu; \sigma^2\right)$ con σ^2 non nota.
- Sia $H_0 : \mu = \mu_0$; S^2 lo stimatore per σ^2 e $\bar{X}$ quello per μ.

- Se H_0 è vera, si ha che

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \sim t_{n-1}$$

  che è la statistica test per μ.

- Sia $\bar{x}$ il valore di μ nel campione estratto e s^2 la varianza campionaria corretta, $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ è la versione campionaria standardizzata di T sotto H_0.

- Se $H_1 : \mu \neq \mu_0$ allora si rifiuta H_0 se $t < -t_{n-1;1-\alpha/2}$ o $t_c > t_{n-1;1-\alpha/2}$.

- Se $H_1 : \mu > \mu_0$ allora si rifiuta H_0 se $t > t_{n-1;1-\alpha}$.

- Se $H_1 : \mu < \mu_0$ allora si rifiuta H_0 se $t < -t_{n-1;1-\alpha}$.

Esempio

- Una farmacia accetta una fornitura di cerotti solo dopo averne verificato un campione per valutare la superiorità del loro tempo medio di resistenza/tenuta rispetto a quello dei cerotti che attualmente ha in vendita che è pari a 9 ore.

- Un campione di $n = 140$ cerotti viene estratto casualmente da cui abbiamo: $\bar{x} = 5.5$ e $s^2 = 3.9$.

- Verificare al livello di significatività $\alpha = 0.01$ se la farmacia accetterà la fornitura o no.

Esempio t-test in R

Il gestore di un impianto industriale afferma che tale impianto scarica meno di 90 litri all'ora di acqua residua in un lago. Un gruppo locale per la protezione ambientale decide di monitorare l'impianto. Si consideri il seguente campione di 10 osservazioni.

```
> x <- c(95,  94,  90,  95, 102, 110,  93,  93, 111,  93)
> t.test(x, mu=90, conf.level=0.95, alternative="greater")
        One Sample t-test

data:  x
t = 3.2231, df = 9, p-value = 0.005219
alternative hypothesis: true mean is greater than 90
95 percent confidence interval:
 93.27758      Inf
sample estimates:
mean of x
     97.6
```

Test per la proporzione (per grandi campioni)

Sia $(x_1, x_2, \ldots, x_n)$ un campione casuale da una popolazione $X \sim Ber(p)$. Si vuole verificare al livello di significatività α il seguente sistema di ipotesi

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

Per il teorema del limite centrale, sotto H_0, si ha che

$$Z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \approx N(0,1)$$

Sia $z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ il suo standardizzato sotto H_0.

- I valori critici sulla $N(0;1)$ sono: $\pm z_{1-\alpha/2}$. Quindi: $A = (-z_{1-\alpha/2}, +z_{1-\alpha/2})$ e $R = (-\infty, -z_{1-\alpha/2}) \cup (z_{1-\alpha/2}, +\infty)$ e si rifiuta H_0 se $z < -z_{1-\alpha/2}$ oppure $z > z_{1-\alpha/2}$.
- Se $H_1 : p > p_0$, invece, il valore critico è $z_{1-\alpha}$ e si rifiuta H_0 se $z > z_{1-\alpha}$.
- Se $H_1 : p < p_0$, allora il valore critico è $z_{1-\alpha}$ e si rifiuta H_0 se $z_< - z_{1-\alpha}$.

Esempio

Mendel in un esperimento ottenne 428 baccelli verdi e 152 gialli. Secondo la sua teoria genetica 1/4 di essi avrebbero dovuto avere baccelli gialli.

- $n = 428 + 152 = 580$
  - $\widehat{p} = 152/580 = 0.262$
- La teoria di mendel e' giusta?

Utilizziamo un test al 5% di significatività ($\alpha = 0.05$)

$$H_0 : p = 1/4$$
$$H_1 : p \neq 1/4$$

Quindi

$$z = \frac{0.262 - 0.25}{\sqrt{\frac{0.25(1-0.25)}{580}}} = 0.667$$

Si noti che z non cade nell'area di rifiuto dato che $z < z_{0.975} = 1.96$.

$$\text{p-value} = 2P(Z > 0.667) \approx 0.5 > \alpha = 0.05.$$

Conclusione: non possiamo rifiutare H_0, quindi non abbiamo sufficiente evidenza in questi dati per rifiutare la teoria di Mendel (probabilmente la teoria e' vera).

Esempio in R:

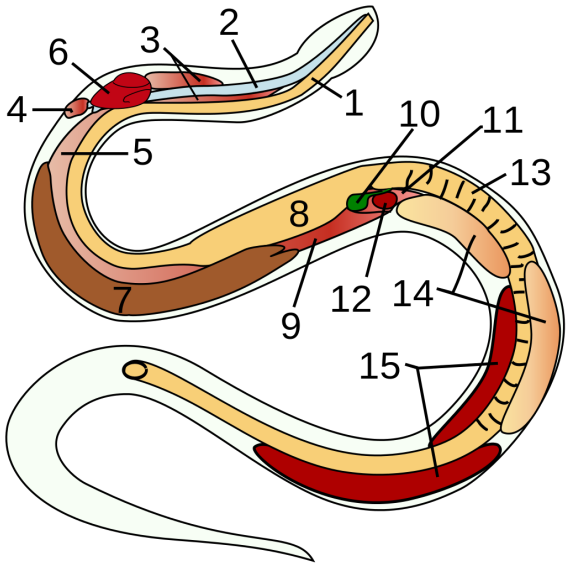
```
> prop.test(x=152, n =580, p=0.25, conf.level=0.95, alternative="two.sided",
correct=FALSE)

        1-sample proportions test without continuity correction

data:  152 out of 580, null probability 0.25
X-squared = 0.45057, df = 1, p-value = 0.5021
alternative hypothesis: true p is not equal to 0.25
95 percent confidence interval:
 0.2279290 0.2993399
sample estimates:
       p
0.262069
```

Inferenza comparativa

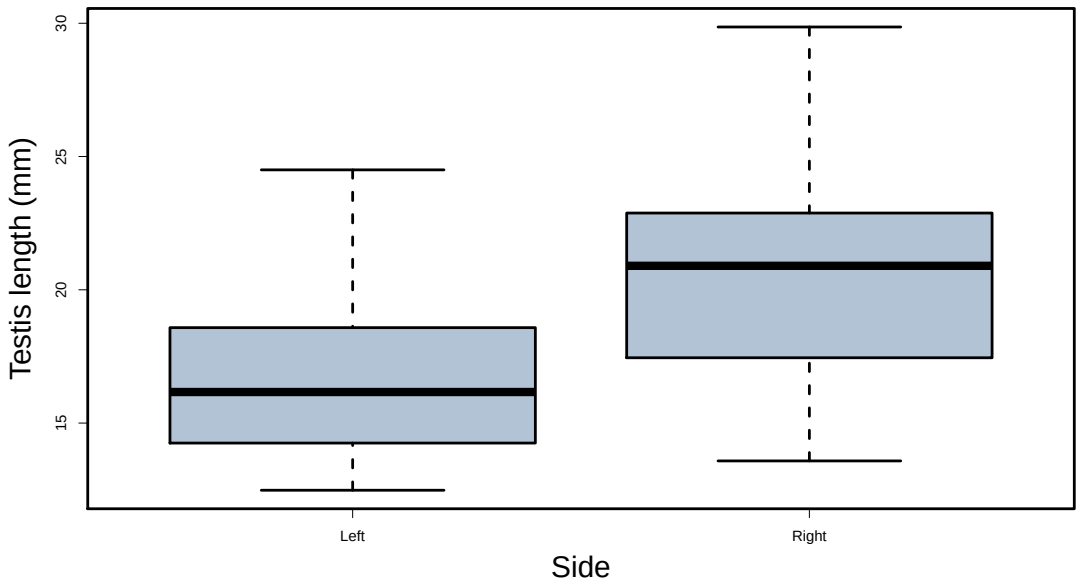
È comune considerare il confronto degli effetti di
due *trattamenti* o *attributi qualitativi*.

Quindi le popolazioni da confrontare sono la popolazione ipotetica con il primo trattamento (o attributo) e il popolazione ipotetica con l'altro.

I serpenti sono mancini o destri?

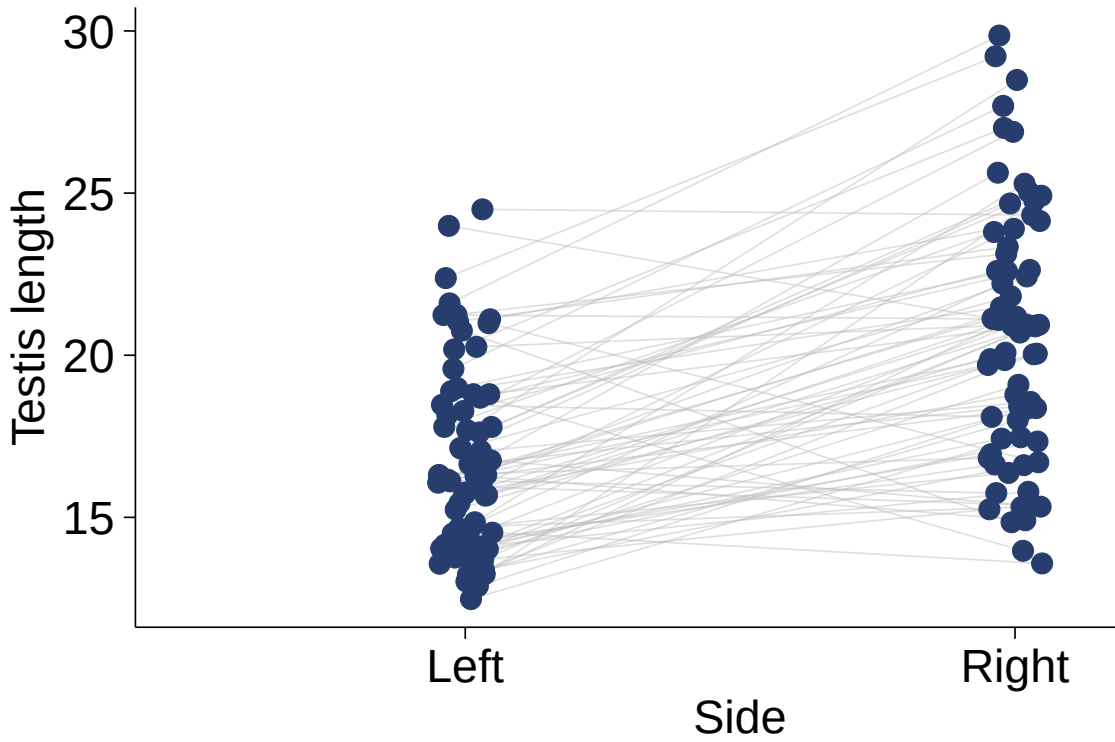


1. Variable numerica (testis length)
2. Variabile esplicativa categorica (side of body)



I dati sono appaiati

1. Variable numerica (testis length)
2. Variabile esplicativa categorica (side of body)



Creiamo una nuova variabile !

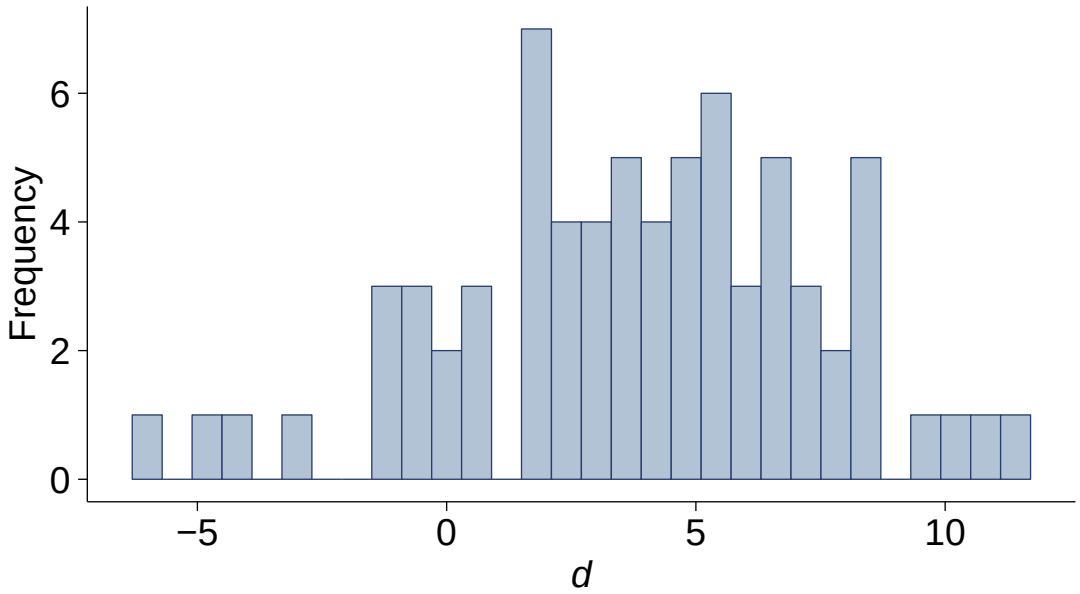
Siano X_{sx} e X_{dx} le misure a dx e sx. Definiamo

$$D = X_{dx} - X_{sx}$$

E quindi prendiamo le differenze osservate:

$$d_1, \ldots, d_n$$

dove d_i rappresenta la differenza per $i$simo serpente snake.



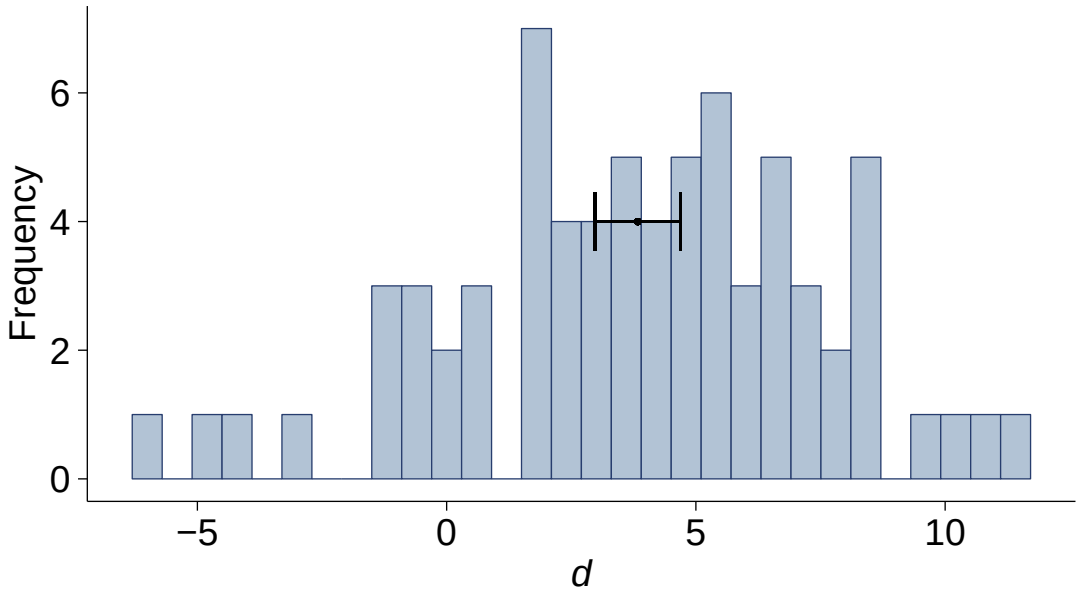
Se le variabili di partenza sono normali, allora

$$D \sim \mathrm{N}(\mu_D, \sigma_D^2)$$

Ma a questo punto sappiamo some procedere. Per esempio un intervallo di confidenza al 95% per μ_D e':

$$\bar{d} \pm t_{0.975;n-1} s_D / \sqrt{n},$$

Distribuzione di D



Test d'ipotesi

Procediamo come nel caso di un singolo campione: $H_0 : \mu_D = 0$ vs $H_1 : \mu_D \neq 0$.La statistica test e'

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$$

con $n - 1 = 71$ gradi di liberta'. Dai dati otteniamo $t = 8.93$ che corrisponde a

$$\text{p-value} = 2P(T \geq 8.93) \approx 0.$$

.

Campioni indipendenti

Assumiamo di avere due campioni da $X_1 \sim N(\mu_1, \sigma_1^2)$ e $X_2 \sim (\mu_2, \sigma_2^2)$. I due campioni sono indipendenti.

		media	dev. st.
campione 1 (da X_1)	n_1	$\bar{x}_1$	s_1
campione 2 (da X_2)	n_2	$\bar{x}_2$	s_2

Le ipotesi sono di solito $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$. Siamo anche interessati a stimare la differenza $\mu_1 - \mu_2$.

Infernza per $\mu_1 - \mu_2$ [varianze note]

$n_1 = 25 \quad \bar{x}_1 = 11.43 \quad (\sigma_1 = 2.0)$
$n_2 = 10 \quad \bar{x}_2 = 9.74 \quad (\sigma_2 = 2.0)$

Si noti che $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, 0.56)$; dato che $Var(\bar{X}_1 - \bar{X}_2) = \frac{2.0^2}{25} + \frac{2.0^2}{10}$.

e quindi un CI al 95% per $\mu_1 - \mu_2$ e' dato da $1.69 \pm 1.96\sqrt{0.56}$, ovvero $(0.22, 3.16)$.

Per testare $H_0 : \mu_1 = \mu_2$ (ovvero $\mu_1 - \mu_2 = 0$), usiamo

$$z = \frac{11.43 - 9.74}{\sqrt{0.56}} = 2.258$$

Con P-value $= 2P(Z > 2.258) = 0.024$. Quindi rifiutiamo H_0 al livello $\alpha = 0.05$.

Infernza per $\mu_1 - \mu_2$ [varianze non note ma uguali]

$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N$ e $\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$ Un IC al 95% CI per $(\mu_1-\mu_2)$ e' dato da

$$(\bar{x}_1-\bar{x}_2) \pm t_{0.975;n_1+n_2-2}\, s \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Nota che quando $\sigma_1^2 = \sigma_2^2 = \sigma^2$ il modo migliore di stimare σ^2 e' quello di combinare $s_1^2 = \frac{\sum(x_1-\bar{x}_1)^2}{n_1-1}$ e $s_2^2 = \frac{\sum(x_2-\bar{x}_2)^2}{n_2-1}$ utilizzando

$$s^2 = \frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{(n_1-1)+(n_2-1)} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

ovvero una media pesata di s_1^2 and s_2^2 con pesi proporzionali ai gradi di liberta'.

Esempio

Un gruppo di quattro pazienti non correlati con retinite pigmentosa (RP) hanno subito una mutazione in un particolare. Un fattore di riduzione campo visivo fattore è stato misurato per ciascuno di questi pazienti. Da questo campione abbiamo ottenuto una media di 24.6 e varianza 4.5. I risultati sono stati confrontati con sedici pazienti con RP non affetti da mutazione. Per questi pazienti la media è risultata essere 21.4 con varianza 5.1.

$$\begin{array}{llll} n_1 = 4 & \bar{x}_1 = 24.6 & s_1^2 = 4.5 & (s_1 = 2.12) \\ n_2 = 16 & \bar{x}_2 = 21.4 & s_2^2 = 5.1 & (s_2 = 2.26) \end{array}$$

$$s^2 = \frac{3\times 4.5 + 15\times 5.1}{18} = \frac{4.5 + 5\times 5.1}{6} = \frac{30.0}{6} = 5.0$$

$\bar{x}_1 - \bar{x}_2 = 3.2$. Quindi

$$\mathrm{se}(\bar{x}_1 - \bar{x}_2) = \sqrt{5(\frac{1}{4} + \frac{1}{16})} = 1.25$$

Dunque

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{se}(\bar{x}_1 - \bar{x}_2)} = \frac{3.2}{1.25} = 2.48.$$

Quindi il p-value e' p-value $= 2P(T_{18} > 2.48) \approx 0.02$. Quindi rifiutiamo $(\mu_1 = \mu_2)$. Un IC al 95% per $\mu_1 - \mu_2$ e' $(3.2 \pm 2.101 \times 1.25) = (0.6, 5.8)$.