Fondamenti di Statistica. a.a. 2008-2009

Prof. V. Simoncini

Orario di Lezione: Mer 9-11

Orario di ricevimento: per appuntamento `valeria@dm.unibo.it`

Sito del corso:
`www.dm.unibo.it/~simoncin/Fondamenti.html`

Testi di Riferimento

Appunti di Lezione

[CC] *A basic course in Statistics*, G. M. Clarke e D. Cooke, IV edizione, Arnold ed., 1998.

[JB] *Statistics. Principles and Methods*, R. A. Johnson e G. K. Bhattacharyya, III edizione, Wiley & sons, 1996.

[SP] *Teoria ed applicazioni della Statistica*, M. R. Spiegel, Collana Schaum, ETAS Libri, II edizione, 1976.

Programma

- Elementi di probabilità

  Analisi Combinatoria, Probabilità condizionata

  Distribuzione di Probabilità e Valore atteso, Standardizzazione

- Distribuzioni discrete di dati casuali

  Distribuzione uniforme, binomiale, di Poisson

- Distribuzioni continue di variabili casuali

  Distribuzione esponenziale, Distribuzione normale (o Gaussiana)

- Teoria dei Campioni. Stime
- Test di ipotesi statistiche ed Intervalli di confidenza
- Test di ipotesi e distribuzione normale
- Test χ^2 ed adattamento a distribuzioni teoriche
- Confronto di coppie

Elementi di probabilità

Prime definizioni: Sia E un evento.

La probabilità che tale evento si verifichi è

$$P(E) = \frac{\#\text{ casi favorevoli}}{\#\text{ casi possibili}} = \frac{N(E)}{N(T)}$$

$(Pr\{E\})$

$\star$ $P(E) \in [0, 1]$

La probabilità che **non** si verifichi l'evento vale $1 - P(E)$

Esempio: Qual'è la probabilità che lanciando un dado si ottenga il numero 5?

Sol. evento E=\{esce la faccia con il numero 5\}
Casi favorevoli=1. Casi possibili=6. Quindi $P(E) = \frac{1}{6}$

N.B. La probabilità è la stessa per ogni altro numero del dado (non truccato).

Esempio: Qual'è la probabilità che lanciando un dado si ottenga un numero dispari ?

Sol. evento E={esce la faccia con un numero dispari}
Casi favorevoli=3. Casi possibili=6. Quindi $P(E) = \frac{3}{6} = \frac{1}{2}$

Osservazione: Probabilità è il limite della frequenza quando il numero delle osservazioni è molto grande.

Esempio: Supponiamo di lanciare 1000 volte una moneta e di ottenere 529 teste. La frequenza è 529/1000. Su altri 1000 lanci si ottiene testa 493 volte, cioè con una frequenza 493/1000. Sul totale dei 2000 lanci si ha

$$\frac{529 + 493}{2000} = \frac{1022}{2000} = 0.511$$

All'aumentare del numero di lanci la frequenza si avvicina alla probabilità ($= 0.5$)

Analisi Combinatoria

Per studiare la probabilità di eventi complessi, occorre un metodo "pratico" per gestire grandi # di prove.

Permutazioni.

Sia n il numero di oggetti da considerare. Il numero di permutazioni di n oggetti è $n!$ (n fattoriale).

Esempio. Supponiamo di avere tre palline colorate, una rossa (R), una blu (B) ed una gialla (G). Le possibili permutazioni sono

R B G R G B B G R B R G G B R G R B

cioè $3! = 6$.

$\star$ Se abbiamo due insiemi distinti di n e k oggetti ciascuno, allora il numero totale di permutazioni dei primi n oggetti seguiti dagli altri k oggetti è

$$n! \times k!$$

Permutazioni in gruppi di r.

Dati n oggetti, si vuole valutare il numero totale di raggruppamenti possibili di r degli n oggetti, dove due raggruppamenti differiscono se considerano almeno un elemento diverso od almeno una posizione diversa.

(Permutazioni di n oggetti presi r alla volta)

$$P_{n,r} = \frac{n!}{(n-r)!} = n(n-1)(n-2)\cdots(n-r+1)$$

Esempio. Siano date 4 palline: blu (B), rossa (R), gialla (G), verde (V). Determinare tutte le possibili permutazioni prendendo solo 3 palline.

Sol. Si ha:

B R G	V R G	G R V	R B V
B G R	V G R	G V R	R V B
B R V	V B G	G B V	R G V
B V R	V G B	G V B	R V G
B V G	V R B	G R B	R G B
B G V	V B R	G B B	R B G

che equivale a $P_{4,3} = \frac{4!}{(4-3)!} = 24$.

Combinazioni.

Siano dati n oggetti. Se non si vuole distinguere tra due gruppi di r oggetti se differiscono per la loro posizione (permutazione degli oggetti), ma solo se differiscono almeno di un elemento, allora si parla di combinazione degli oggetti. In tal caso, dalla formula precedente, bisogna *eliminare* $r!$, che equivale alla permutazione di r oggetti. Si ha quindi

$$C_{n,r} = \left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n!}{r!(n-r)!}$$

Esempio: In una lega di calcio ci sono 8 squadre. Quante partite dovranno essere giocate perchè ogni squadra giochi contro tutte le altre?

Sol. Si ha $n = 8$ ed ad ogni partita ci sono 2 squadre. Quindi si ha

$$C_{8,2} = \frac{8!}{2!(8-2)!} = \frac{8!}{2!6!} = \frac{8 \cdot 7}{2} = 28,$$

cioè 28 partite.

Esempio: Una classe è costituita da 9 maschi e 11 femmine. In quanti modi distinti è possibile formare un gruppo di rappresentanti con 2 maschi e 2 femmine?

Sol. La scelta delle femmine è indipendente da quella dei maschi.

Femmine: $C_{11,2} = \frac{11!}{2!9!} = 55$

Maschi, $C_{9,2} = \frac{9!}{2!7!} = 36$

$\Rightarrow 55 \cdot 36 = 1980$ modi per formare un gruppo di rappresentanti

Probabilità composta

E^c evento complementare di E $\qquad P(E^c) = 1 - P(E)$

Intersezione: Siano E_1, E_2 due eventi.

$$P(E_1 \cap E_2) = \frac{N(E_1 \cap E_2)}{N(Tot)}$$

probabilità che i due eventi abbiano luogo contemporaneamente:

$N(E_1 \cap E_2)$ numero di casi favorevoli

$N(Tot)$ numero di casi possibili

Esempio: 3 lanci successivi di una moneta ed i seguenti eventi:

E_1 : la seconda volta viene testa;

E_2: solo una volta viene testa.

Qual'è la probabilità dell'intersezione dei due eventi?

Sol. Indichiamo con (T) testa e con (C) croce.

Tutti i casi possibili sono $8 (= 2 \cdot 2 \cdot 2)$:

CCC, CTC, CCT, CTT, TCC, TCT, TTC, TTT.

Eventi favorevoli per E_1: CTC, CTT, TTC, TTT.

Eventi favorevoli per E_2: CTC, CCT, TCC.

Quindi $N(E_1 \cap E_2) = 1$ (CTC) $\Rightarrow \quad P(E_1 \cap E_2) = \frac{1}{8}$

unione:

$$P(E_1 \cup E_2) = \frac{N(E_1 \cup E_2)}{N(Tot)}$$

probabilità che almeno uno dei due eventi abbia luogo

Nell'esempio precedente,

casi favorevoli per $E_1 \cup E_2$:

CTC, CCT, TCC, CTT, TTC, TTT

$\Rightarrow \quad N(E_1 \cup E_2) = 6 \quad \Rightarrow \quad P(E_1 \cup E_2) = \frac{6}{8} = \frac{3}{4}$

Nota. Unione e intersezione di eventi possono essere facilmente generalizzati ad un numero di eventi superiore a due,
$E_1, E_2, \ldots, E_k$.

Eventi mutuamente esclusivi

Se due eventi E_1, E_2 si escludono a vicenda, allora

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Esempio: Determinare la probabilità di ottenere **almeno** 2 assi in una mano a briscola (3 carte in mano su 40).

Sol. $N(Tot) = C_{40,3}$.

E_1 : avere esattamente 2 assi $\qquad$ E_2 : avere esattamente 3 assi

$N(E_1) = C_{4,2} \cdot C_{36,1} \qquad N(E_2) = C_{4,3} = 4$

E_1, E_2 mutuamente esclusivi $\Rightarrow P(E_1 \cup E_2) = P(E_1) + P(E_2)$

$$P(E_1) = \frac{3 \cdot 18}{130 \cdot 19}, \quad P(E_2) = \frac{1}{130 \cdot 19} \quad \Rightarrow \quad P(E_1 \cup E_2) = 0.022$$

Esempio: Trovare la probabilità che si presenti almeno due volte la somma 7 in quattro lanci di due dadi.

Sol. La somma 7 si ottiene con le 6 coppie

$$(3,4),(4,3),(2,5),(5,2),(1,6),(6,1)$$

E_1: prob. di avere 2 volte la somma 7
E_2: prob. di avere 3 volte la somma 7
E_3: prob. di avere 4 volte la somma 7
$\Rightarrow P = P(E_1) + P(E_2) + P(E_3)$

(mutuamente esclusivi)

$$P = C_{4,2}\left(\frac{6}{36}\right)^2\left(\frac{30}{36}\right)^2 + C_{4,1}\left(\frac{6}{36}\right)^3\left(\frac{30}{36}\right) + \left(\frac{6}{36}\right)^4 = \frac{171}{6^4}.$$

Eventi **non** sono mutuamente esclusivi, allora (v. insiemistica)

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Esempio: Un'urna contiene 3 palline: rossa (R), blu (B), verde (V).
Due estrazioni (con riposizione).
E_1 : almeno una delle palline sia rossa
E_2 : le palline estratte siano dello stesso colore.

Determinare $P(E_1)$, $P(E_2)$, $P(E_1 \cup E_2)$ e $P(E_1 \cap E_2)$

Sol. Si ha (direttamente)

ToT={RR, RB, RV, BB, BR, BV, VV, VB, VR}

$E_1 = \{RR, RB, RV, BR, VR\}$ e $E_2 = \{RR, BB, VV\}$

$\Rightarrow \quad P(E_1) = \frac{5}{9} \quad P(E_2) = \frac{3}{9} = \frac{1}{3}$

$E_1 \cap E_2 = \{RR\} \qquad E_1 \cup E_2 = \{RR, RB, RV, BR, VR, BB, VV\}$

$\Rightarrow P(E_1 \cap E_2) = \frac{1}{9}, \quad P(E_1 \cup E_2) = \frac{5}{9} + \frac{1}{3} - \frac{1}{9} = \frac{7}{9}$

Eventi indipendenti

Se E_1 non influenza E_2 si dice che i due eventi sono indipendenti. In tal caso

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

Esempio delle targhe

Qual'è la probabilità che una targa (di tre lettere e tre cifre) contenga esattamente una volta la lettera A ed esattamente il numero 99?

Sol. $E_1 = \{$ appare la lettera A$\}$, $\quad E_2 = \{$appare il numero 99$\}$

eventi indipendenti. Quindi $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$

$P(E_1) = C_{3,1} \left(\frac{1}{26}\right) \left(\frac{25}{26}\right)^2 = 3 \left(\frac{25^2}{26^3}\right) \approx 0.107$

$P(E_2) = C_{2,1} \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right) \approx 0.018$

Quindi $P(E_1 \cap E_2) \approx 0.107 \cdot 0.018 = 0.0019$

Probabilità condizionata

Supponiamo che l'evento E_2 dipenda dal verificarsi dell'evento E_1.

Probabilità di E_2 sapendo che E_1 si è verificato:

$$P(E_2|E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)}$$

Osservazione. Riscrivendo la formula sopra si ottiene

$$P(E_1 \cap E_2) = P(E_2|E_1)P(E_1) \tag{1}$$

la probabilità di due eventi si ottiene moltiplicando la probabilità di un evento con la probabilità del secondo evento, dato il primo.

Legge Moltiplicativa

Si ha anche $P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$ da cui si ottiene

$$P(E_2|E_1)P(E_1) = P(E_1|E_2)P(E_2) = P(E_1 \cap E_2)$$

Esempio: Una carta viene estratta da un mazzo di 52 carte. Ci viene detto che è un asso. Qual'è la probabilità che sia un asso di quadri?

Sol. Via breve: la probabilità è 1/4.

Come probabilità condizionata:

E_1 : l'evento "estrazione di un asso"

E_2 : l'evento "estrazione dell'asso di quadri" $\qquad P(E_2|E_1)$?

$$P(E_1) = \frac{4}{52}, \quad P(E_1 \cap E_2) = P(E_2) = \frac{1}{52} \Rightarrow P(E_2|E_1) = \frac{1}{52} \cdot \frac{52}{4} = \frac{1}{4}$$

Esempio: Si estraggono 2 carte da un mazzo di 52 carte. Trovare la probabilità che si tratti di due assi, nel caso che la prima carta (a) sia (b) non sia rimessa nel mazzo.

Sol. E_1= asso alla prima estrazione
E_2= asso alla seconda estrazione.
Allora:

a) Eventi indipendenti.
$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2) = \frac{4}{52} \cdot \frac{4}{52} = \frac{1}{169}$.

b) Eventi dipendenti.

b.1) Alla prima estrazione: $P(E_1) = \frac{4}{52}$. Alla seconda estrazione: $P(E_2|E_1) = \frac{3}{51}$. Quindi $P(E_1 \cap E_2) = \frac{4}{52} \cdot \frac{3}{51}$.

b.2) Casi possibili: $C_{52,2} = \frac{52 \cdot 51}{2}$ Casi favorevoli $C_{4,2} = \frac{4 \cdot 3}{2}$
$\Rightarrow P = \frac{3 \cdot 4}{52 \cdot 51}$

es. (23/6/2008). Un monitoraggio di una serie di siti di interesse storico naturalistico riporta la seguente situazione di degrado di due dei materiali strutturali:

Materiale	Buono Stato	Stato Critico	Ottimo Stato
Arenaria	35	30	10
Marmo	25	20	30

Preso a caso un sito tra quelli esaminati, determinare:

(a) La probabilità che il materiale sia in stato critico; *Sol.* $P = 50/150 = 1/3$.

(b) La probabilità che il materiale sia di tipo arenario e sia in buono stato;

Sol. $P(\text{'aren'} \cap \text{'buono stato'}) = 35/150$. *Oppure*

$P = P(\text{arenario}|\text{buono stato})\, P(\text{buono stato}) = 35/60 * 60/150$.

(c) La probabilità che il materiale sia in stato critico, sapendo che è di tipo marmoreo;

Sol. $P(\text{'critico'} \mid \text{'marmo'}) = (20/75) / (75/150) = 20 * 150 / (75 * 75)$.

(d) La probabilità che il materiale non sia in stato critico.

Sol. $P(\text{'non critico'}) = 1 - P(\text{'critico'}) = 1 - 50/150 = 2/3$. *Oppure,* $P = 100/150$.

Probabilità binomiale

In caso di evento con successo-insuccesso netto

Consideriamo n prove indipendenti

p : probabilità di successo ad ogni prova
$(1-p)$: probabilità di insuccesso
Qualè la probabilità di ottenere r successi in n prove?

Sol. Ogni prova è indipendente (con prob.di successo p)

Se i primi r sono r successi, dovrà essere

$$\underbrace{p \cdot p \cdots p}_{r}\underbrace{(1-p)\cdots(1-p)}_{(n-r)}$$

Tutte le possibili combinazioni di questa sequenza ci vanno bene:

$$C_{n,r}p^r(1-p)^{n-r}$$

es. (19/2/08) In un contenitore di un laboratorio sono tenuti sotto osservazione 20 individui di una specie, e tra questi ci sono 8 individui malati. Vengono presi a caso tre individui.

a) Supponendo che ogni individuo sia rimesso nel contenitore dopo la sua estrazione, determinare la probabilità che tutti gli individui siano sani o malati.

Sol. $P = (12/20)(12/20)(12/20) + (8/20)(8/20)(8/20)$.

b) Come in a), ma senza riposizione

Sol. $P = (12/20)(11/19)(10/18) + (8/20)(7/19)(6/18)$.

c) Supponendo che ogni individuo sia rimesso nel contenitore dopo la sua estrazione, determinare la probabilità che almeno uno degli individui sia malato.

Sol. $P = 3\ (8/20)(12/20)(12/20) + 3\ (8/20)\ (8/20)\ (12/20) + (8/20)(8/20)(8/20)$.

Distribuzione di Probabilità

Sia X variabile con valori discreti $X_1, X_2, \ldots, X_N$ aventi probabilità $p_1, p_2, \ldots, p_N$ $(\sum_i p_i = 1)$

(X variabile discreta *aleatoria*, o *stocastica*, o *casuale*, random)

Funzione di probabilità di X: $P(X)$ tale che $P(X = X_i) = p_i$

Esempio: Lanciamo due dadi e denotiamo con X la somma di punti ottenuta. Allora si ha

X	2	3	4	5	6	7	8	9	10	11	12
$P(X = X_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Valore atteso (*expected value*)

Valore atteso (Speranza Matematica) di una variabile X:

$$E[X] = p_1 X_1 + p_2 X_2 + \ldots, + p_N X_N$$

$\mu = E[X]$ si dice anche *media* della variabile X

Il valore atteso è una funzione (lineare)

Esempio: Ogni settimana un padre dà al proprio figlio dei soldi come segue: lancia 3 monete da 2 Euro, e consegna al figlio quelle che cadono con il "numero" rivolto verso l'alto. Quante monete il figlio si aspetta di avere?

Sol. Si può definire X_i= $\{i$ monete da 2 Euro ottenute $\}$. Quindi $E[X] = X_0 p_0 + X_1 p_1 + X_2 p_2 + X_3 p_3$.
Casi possibili $(8 = 2^3)$:

HTT, HTH, HHT, HHH, TTT, THT, TTH, THH

Casi favorevoli: che venga 1 volta il "numero" sono $C_{3,1} = 3$ e i casi per cui vengano 2 volte il numero sono $C_{3,2} = 3$

$$p_0 = P(0\,\text{monete}) = \frac{1}{8}, \qquad p_1 = P(1\,\text{moneta}) = \frac{3}{8},$$
$$p_2 = P(2\,\text{monete}) = \frac{3}{8}, \qquad p_3 = P(3\,\text{monete}) = \frac{1}{8}$$

$$\Rightarrow E[X] = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{6}{8} = \frac{3}{4}$$

Il ragazzo spera di ottenere $\frac{3}{4}$ di moneta da due Euro (1.50 Euro)

Varianza e Valore atteso

La varianza è :

$$\sigma^2 = Var[X] = \sum_i p_i (x_i - \mu)^2$$

Osservazioni:

- $Var[X] = E[(X - \mu)^2]$
- $Var[X] = E[(X - E[X])^2)$
- $Var[X] = E[X^2] - \mu^2$

σ: deviazione standard

Standardizzazione

Trasformiamo una variabile casuale con media μ e varianza σ^2 in una variabile U con media 0 e varianza 1:

$$U = \frac{X - \mu}{\sigma}$$

Infatti si ottiene

$$\begin{aligned} E[U] &= \frac{1}{\sigma} E[X - \mu] = \frac{1}{\sigma}(E[X] - E[\mu]) = 0 = \mu_U \\ Var[U] &= E[(U - \mu_U)^2] = E[U^2] = \frac{1}{\sigma^2} E[(X - \mu)^2] \\ &= \frac{1}{\sigma^2} Var[X] = 1 \end{aligned}$$

Distribuzioni discrete di variabili casuali

Distribuzioni: curve ipotetiche rappresentanti (approssimanti) il fenomeno

Variabile: dati, probabilità, ...

X : variabile casuale

$P(X = r)$: probabilità che l'evento abbia valore r

Distribuzione cumulativa: $F(b) = P(X \leq b) = \sum_{r=1}^{b} P(X = r)$

Mediana:

M tale che $P(X \leq M) \geq \frac{1}{2}$ e $P(X \geq M) \geq \frac{1}{2}$

Distribuzione uniforme

Contesto: tutti i valori $X_1, \ldots, X_N$ hanno la stessa probabilità.

Def. Una variabile casuale discreta X che prende i valori $1, 2, \ldots, k$ e tale che

$$P(X = r) = \begin{cases} \frac{1}{k} & r = 1, \ldots, k \\ 0 & \end{cases}$$

segue una distribuzione discreta uniforme.

Distribuzione uniforme

Media (Valore atteso):

$$E[X] = \sum_{r=1}^{k} r \frac{1}{k} = \frac{1}{k} \sum_{r=1}^{k} r = \frac{1}{k} \frac{k(k+1)}{2} = \frac{k+1}{2}.$$

Esempio: Determinare la mediana della distribuzione discreta uniforme, per $k = 3$ e per $k = 6$.

Sol. Per $k = 3$, un possibile candidato è 2. Si ha

$$P(X \leq 2) = \frac{2}{3} > \frac{1}{2} \quad e \quad P(X \geq 2) = \frac{2}{3} > \frac{1}{2}$$

per cui la mediana è effettivamente 2.

Per $k = 6$, possibili candidati sono 3 e 4. Si ha

$$P(X \leq 3) = \frac{1}{2}, \quad P(X \geq 3) = \frac{2}{3} \quad P(X \leq 4) = \frac{2}{3} \quad P(X \geq 4) = \frac{1}{2}$$

Prendendo $M = 3.5$ (media di 3 e 4):

$P(X \leq M) = \frac{1}{2}$ e $P(X \geq M) = \frac{1}{2}$

Distribuzione binomiale

Sia X l'evento con probabilità p

La probabilità che X si verifichi esattamente r volte su N prove è

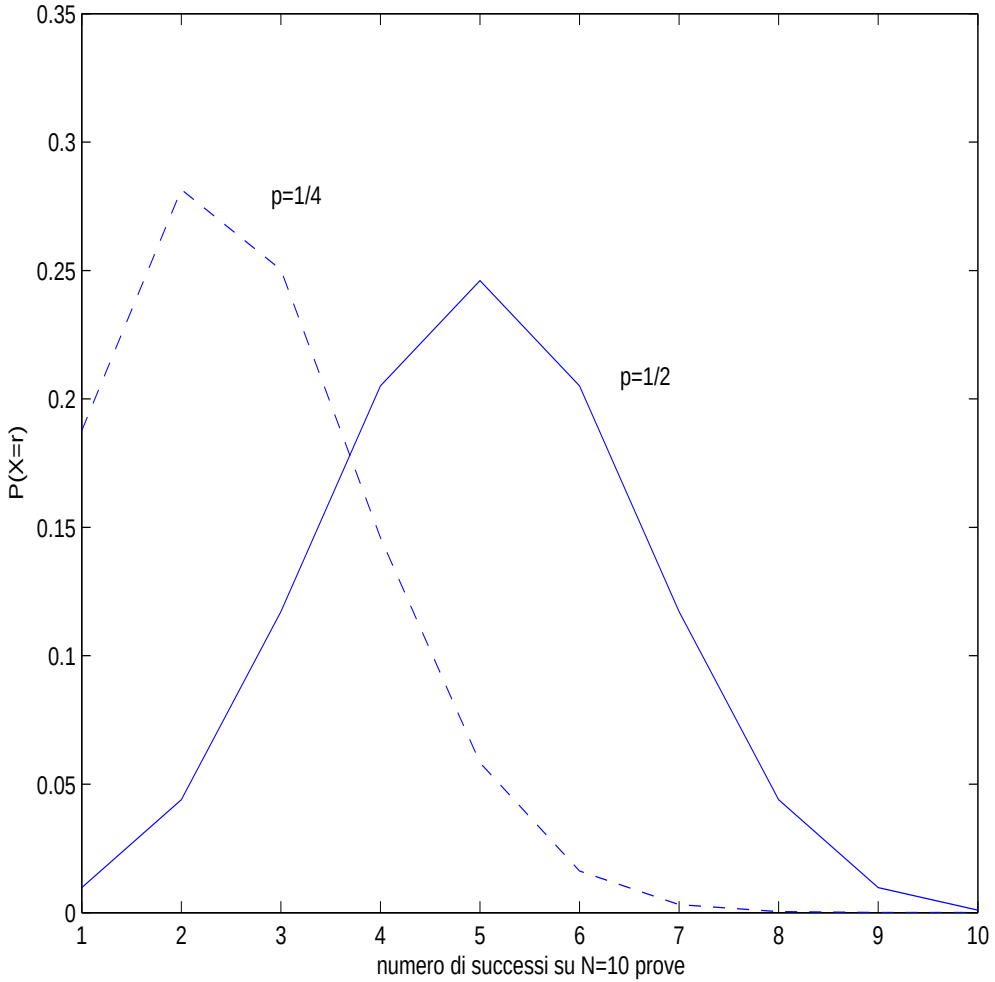
$$P(r) = \binom{N}{r} p^r (1-p)^{N-r} \tag{2}$$

Def. Una variabile discreta casuale X si dice che segue una distribuzione binomiale se vale (2)

Condizioni:

1. Un numero fisso N di prove
2. Due probabilità: p (successo) e $(1-p)$ (insuccesso)
3. Prove indipendenti
4. La probabilità rimane costante durante tutte le prove
5. La variabile è il numero totale di successi in N prove.

Curve di distribuzioni binomiali per $p = \frac{1}{2}$ e $p = \frac{1}{4}$



Distribuzione binomiale cumulativa

Esempio: Qual'è la probabilità di avere *almeno* 4 teste in 6 lanci di una moneta?

Sol. Si ottiene come
$P(X \geq 4) = P(teste = 4) + P(teste = 5) + P(teste = 6) = \cdots = \frac{1}{32}$.

In generale:

$$F(b) = P(X \leq b) = \sum_{r=0}^{b} P(X = r)$$

binomiale perchè per $r = 0, 1, \ldots, N$ ogni $P(X = r)$ corrisponde ai coefficienti dello sviluppo del binomio:

$$(p+q)^N = q^N + C_{N,1} p q^{N-1} + C_{N-1,2} p^2 q^{N-2} + \cdots + p^N$$

dove nel nostro caso $q = 1 - p$

Esempio: In un processo produttivo, vengono ispezionati campioni di 10 elementi per ogni pacchetto, selezionati in modo casuale. Se il numero di elementi danneggiati è inferiore a 2, allora il il pacchetto è accettato, altrimenti il processo viene fermato per registrare la macchina. Determinare la probabilità che il pacchetto sia accettato quando la proporzione effettiva p di trovare elementi danneggiati sull'intera produzione è (i) $p = 0.04$; (ii) $p = 0.07$.

Sol. La probabilità di trovare pacchetti con un numero di elementi danneggiati inferiore a 2 è dato da $(1-p)^{10} + 10(1-p)^9 p$, quindi nel caso (i) si ottiene 0.9418 e per (ii) si ottiene 0.8482.

Alcune proprietà della distribuzione binomiale

Media:

$$\begin{aligned} E[X] &= \sum_{r=0}^{N} rP(X=r) = \sum_{r=0}^{N} r\frac{N!}{r!(N-r)!}p^r(1-p)^{N-r} \\ &= pN\sum_{r=1}^{N}\frac{(N-1)!}{(r-1)!(N-r)!}p^{r-1}(1-p)^{N-r} \\ &= pN(p+(1-p))^{N-1} = Np. \end{aligned}$$

Varianza:

$$\sigma^2 \equiv Var[X] = Np(1-p)$$

Esempio: Determinare la probabilità che in una famiglia di 4 figli ci siano almeno 0, 1, 2, 3, 4 maschi, sapendo che la probabilità media della nascita di un maschio è $p = 0.52$.

Sol. Sia X=figli maschi in famiglia. La variabile X può assumere valori $0, 1, 2, 3, 4$. Le relative probabilità sono

$$P(X = r) = \binom{4}{r} p^r (1-p)^{4-r}, \quad r = 0, \ldots, 4$$

per cui $P(X = 0) = 0.05$, $P(X = 1) = 0.23$, $P(X = 2) = 0.37$, $P(X = 3) = 0.28$ e $P(X = 4) = 0.07$.

Nota: La somma delle probabilità deve dare 1 se tutti i casi possibili sono considerati.

Esempio: Un pacchetto di 20 caramelle viene confezionato scegliendo casualmente da un grande contenitore contenente il 40% di caramelle gommose ed il resto di caramelle dure. Determinare la media (μ) e la deviazione standard (σ) del numero di caramelle gommose nel pacchetto. Determinare anche la probabilità che tale numero sia minore di (i) $\mu - \sigma$, (ii) $\mu - 2\sigma$.

Sol. Si ha $\mu = Np = 20 \cdot 0.4 = 8$ e
$\sigma = \sqrt{Np(1-p)} = \sqrt{20 \cdot 0.4(1-0.4)} = 2.19$. Per ottenere la probabilità che il pacchetto contenga non più di $\mu - \sigma = 5.81$ caramelle gommose, si calcola la probabilità cumulativa

$$\begin{aligned} P(X < \mu - \sigma) &= P(X \leq 5) \\ &= P(X = 0) + P(X = 1) + \cdots + P(X = 5) \\ &= \sum_{r=0}^{5} \binom{20}{r} (0.40)^r (0.60)^{N-r} = 0.1256. \end{aligned}$$

Per (ii) si ottiene $P(X < \mu - 2\sigma) = P(X \leq 3) = 0.01596 \approx 0.016$.

Distribuzione geometrica

Una variabile casuale discreta X segue la distribuzione geometrica se

$$P(X = r) = (1 - p)^{r-1}p, \qquad r = 1, 2, \ldots$$

Condizioni:

1. C'è una successione di prove;
2. Due possibili risultati (successo/insuccesso);
3. Le prove sono indipendenti;
4. La probabilità ad ogni prova rimane costante;
5. La variabile è il numero di prove necessarie per avere il primo successo

Non analizzata ulteriormente (vedi Appunti)

Distribuzione di Poisson

Si dice che una variabile aleatoria discreta X segue la distribuzione di Poisson se, fissato $\lambda > 0$, vale

$$P(X = r) = \frac{\lambda^r}{r!} e^{-\lambda} \qquad r = 0, 1, \ldots,$$

Condizioni:

1. Gli eventi sono *casuali* nello spazio (tempo) continuo $\star$;
2. Gli eventi hanno luogo singolarmente e sono esclusivi;
3. Il numero di eventi che ha luogo in un dato intervallo è uniforme (proporzionale alla lunghezza dell'intervallo) $\star$;
4. Gli eventi sono indipendenti $\star$;
5. La variabile è *il numero di eventi* aventi luogo nell'intervallo considerato.

Alcune proprietà

Media:

$$\begin{aligned} E[X] &= \sum_{r=0}^{\infty} rP(X=r) = \sum_{r=1}^{\infty} \frac{r}{r!}\lambda^r e^{-\lambda} \\ &= \lambda e^{-\lambda} \sum_{r=1}^{\infty} \frac{1}{(r-1)!}\lambda^{r-1} = \lambda e^{-\lambda} e^{\lambda} = \lambda \end{aligned}$$

Varianza:

$$Var[X] = \lambda$$

Eventi rari

La distribuzione di Poisson si può derivare come limite della distribuzione binomiale per

$N \to \infty$ e $p \to 0$, ponendo $\lambda = N \cdot p$.

Esercizio: Analizzare graficamente la distribuzione binomiale e Poissoniana al variare del parametro λ.

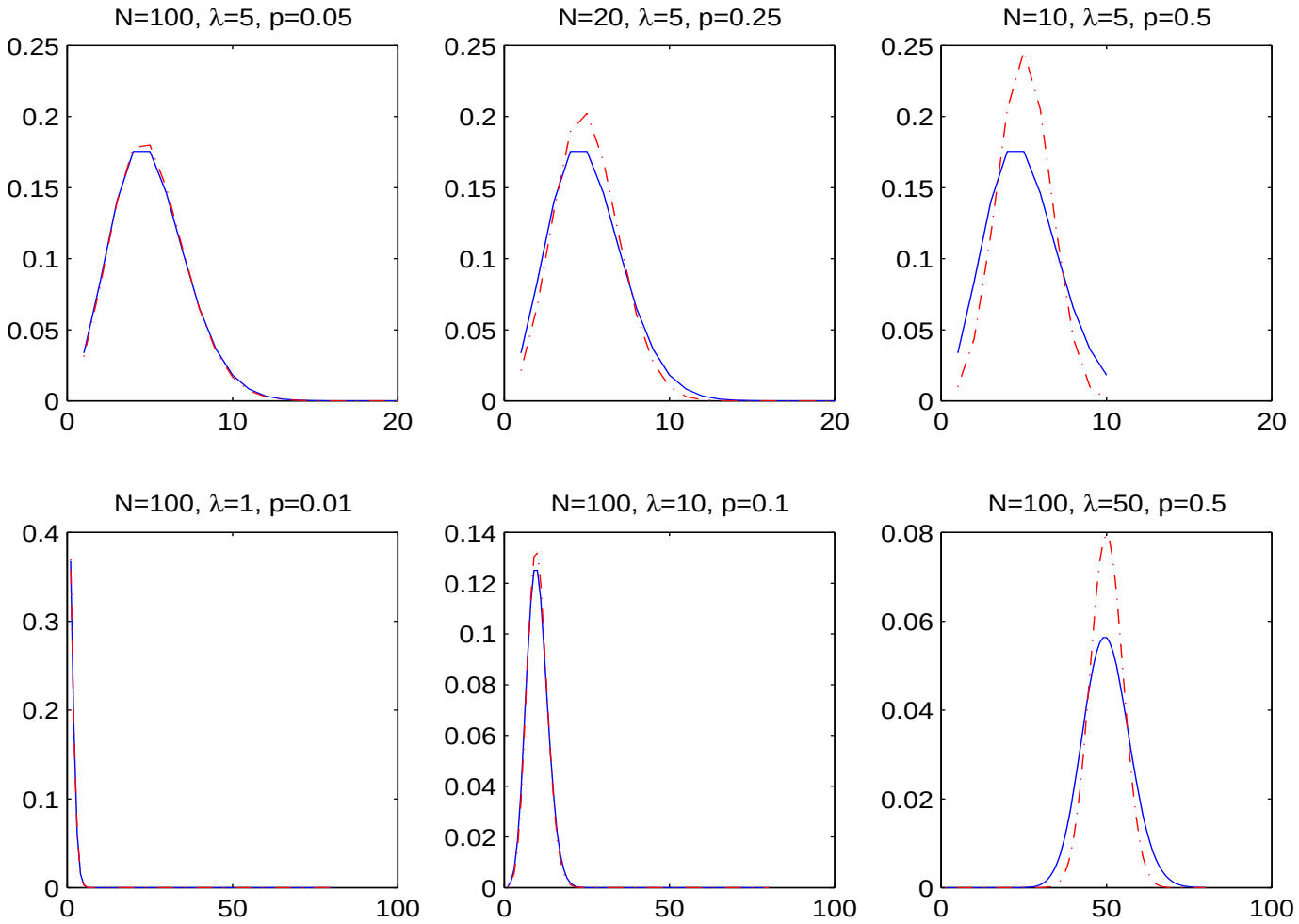


Figura 1: Poligoni per le distribuzioni binomiale e di Poisson per vari valori dei parametri di riferimento, con $\lambda = N \cdot p$. Poisson. Binomiale

Esempio: Il numero medio di errori di battitura per pagina in un libro è 1.2. Qual'é la probabilità di trovare in una particolare pagina (di 2000 lettere): (a) nessun errore; (b) tre o più errori?

Sol. La variabile aleatoria X è il numero di errori.

La probabilità di trovare un errore in una pagina di $N = 2000$ lettere è $p = 1.2/2000 = 0.0006$. Con la distribuzione di Poisson per $\lambda = Np = 1.2$:

(a) $P(X = 0) = 1e^{-1.2} \approx 0.30$

(b) $P(X \geq 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2) =$
$1 - 0.30 - 0.36 - 0.22 = 0.12$

Esempio: Una malattia rara viene presa in media dallo 0.5% dei neonati. Cento bambini nascono in un ospedale in una settimana. Qual'è la probabilità che esattamente tre di loro abbiano la malattia?

Sol. La probabilità è bassa. Ponendo $\lambda = 100 \cdot 0.005 = 0.5$, si ottiene usando la distribuzione di Poisson,

$$P(X = 3) = \frac{\lambda^3}{3!} e^{-0.5} = 0.01263$$

Distribuzioni continue di variabili casuali

X può assumere tutti i valori in un intervallo $[a, b]$

Definizione. Densità di probabilità di X in $[a, b]$:

la funzione f integrabile a cui è associata la probabilità

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Nota: $P(X = a) = 0$

Probabilità di ottenere un valore $c \in [a, b]$:

$$P(X = c) = \int_{c-\delta}^{c+\delta} f(x)dx \approx f(c) \cdot (2\delta)$$

Condizioni per f:

- f continua
- $f \geq 0, \forall x$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

La terza condizione: Per X definita in $[a, b]$:

$$\int_a^b f(x)dx = 1$$

Funzione di distribuzione cumulativa:

$$F(b) = P(X \leq b) = \int_{-\infty}^{b} f(x)dx$$

Mediana:

$$M \text{ tale che } \int_{-\infty}^{M} f(x)dx = \frac{1}{2}.$$

Moda:

$$M \text{ tale che } \max_{x \in [a,b]} f(x) = f(M).$$

Valore atteso (**media**) nel continuo:

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

(momento di ordine 1)

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

Varianza nel continuo:

$$Var[X] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = E[X^2] - \mu^2$$

(momento di ordine 2 dalla media)

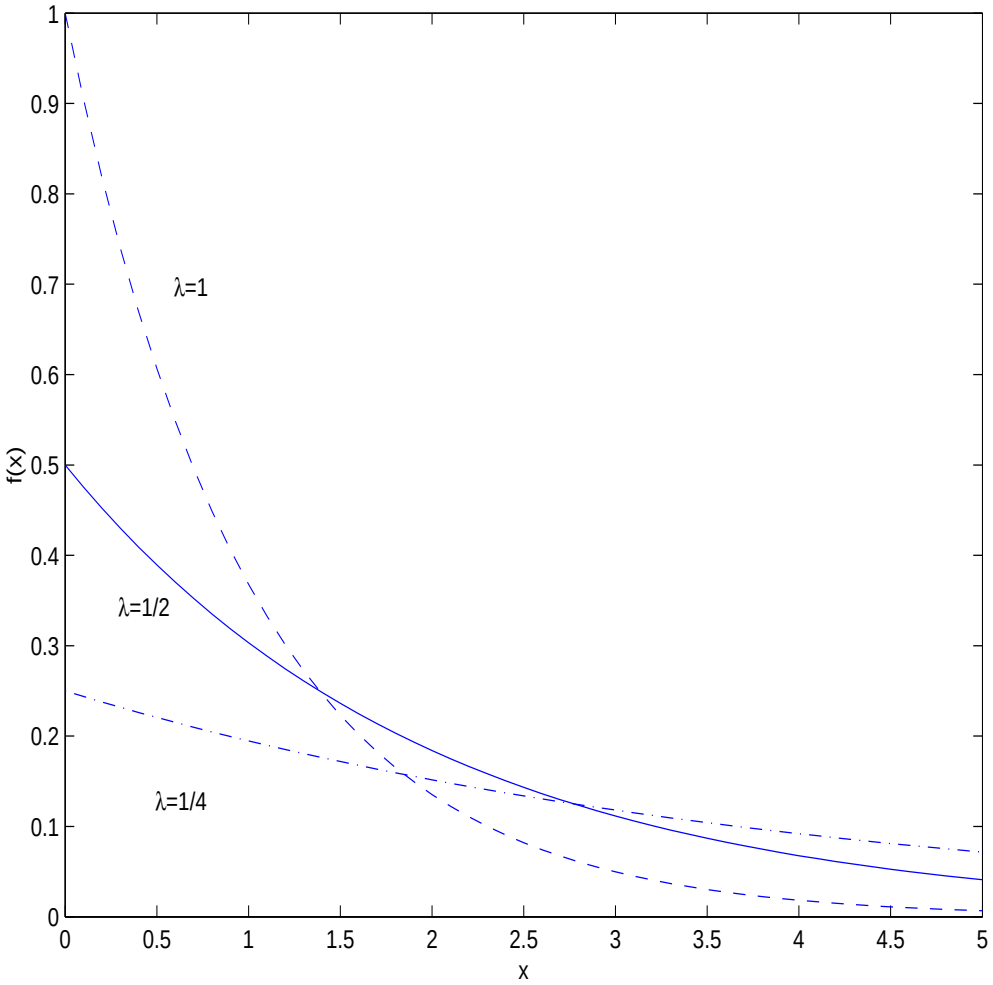
Distribuzione esponenziale

Funzione densità

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Funzione parametrica (parametro λ)

Funzione di densità della distribuzione esponenziale



$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- f è una funzione di densità di distribuzione
- Media: $\mu \equiv E[X] = \frac{1}{\lambda}$ (integrazione per parti)
- Varianza: $Var[X] = \frac{1}{\lambda^2}$.
- La funzione distribuzione cumulativa soddisfa

$$P(X \leq b) = \int_0^b \lambda e^{-\lambda x} dx = 1 - e^{-\lambda b}.$$

Esempio. I tempi di durata di 500 batterie elettriche sono stati raggruppati con le seguenti frequenze

Ore	0-50	50-100	100-150	150-200	200-250	250-300	300-350	350-400
Freq.	208	112	75	40	30	18	11	6

Stimare la media e fare un istogramma. Suggerire una distribuzione che possa essere usata per modellare i dati, e determinare le frequenze per ogni intervallo-tempo sul tale modello di distribuzione (la frequenza per ogni intervallo si determina moltiplicando la probabilità in tale intervallo per il numero totale delle frequenze, cioè $N = 500$). Riportare tali valori sull'istogramma fatto in precedenza e commentare sulla bontà del modello.

Sol. Usando il punto medio x_i^c di ogni classe ed $N = 500$,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{8} f_i x_i^c = 95.$$

Se il modello di distribuzione è quello esponenziale, dev'essere
$E(X) = \frac{1}{\lambda} = \bar{x}$

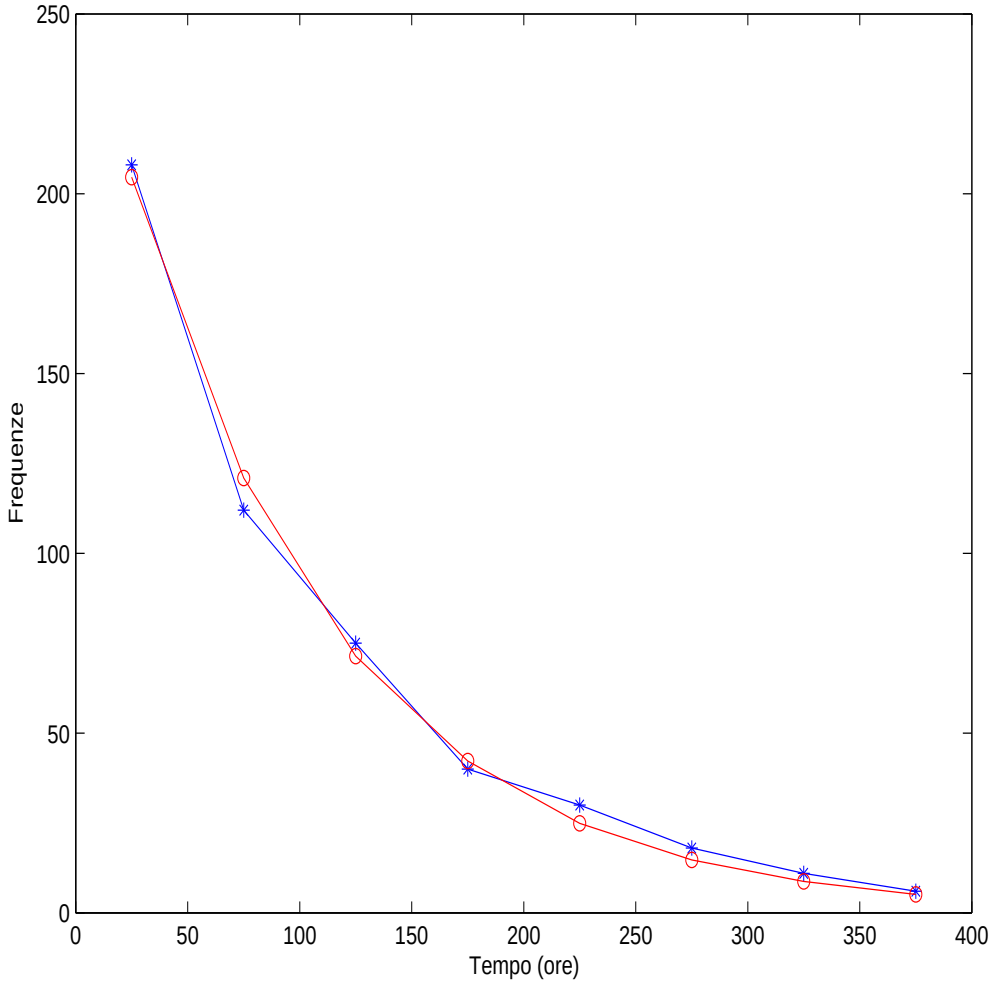
$$\lambda = \frac{1}{95}$$

Usando $F(b) = 1 - e^{-\lambda b}$:

$P(x_i \leq X \leq x_{i+1}) = F(x_{i+1}) - F(x_i)$ $\qquad$ $\widetilde{F}_i = P(x_i \leq X \leq x_{i+1}) \cdot N$

(x_i, x_{i+1} estremi della classe)

Ore	0-50	50-100	100-150	150-200	200-250	250-300	300-350	350-400
$\widetilde{F}$	204.6	120.9	71.4	42.2	24.9	14.7	8.7	5.1


250
200
150
100
50
0
Frequenze
0
50
100
150
200
250
300
350
400
Tempo (ore)

Distribuzione normale (o Gaussiana)

Si deriva dalla distribuzione binomiale per $N \to \infty$

X variabile casuale (media μ e varianza σ^2)

Funzione densità

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$X \in \mathcal{N}(\mu, \sigma^2)$

$\star$ Dati μ e σ, se X è distribuita seguendo la funzione f sopra, allora

$$E[X] = \mu, \qquad Var[X] = \sigma^2$$

Valgono: $f \geq 0, \int_{-\infty}^{\infty} f(x)dx = 1$

Distribuzione normale standardizzata

$$Z = \frac{X - \mu}{\sigma}$$

Funzione di densità *normalizzata*

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

Z ha media zero e varianza uno $\qquad Z \in \mathcal{N}(0,1)$

Probabilità cumulata

$$P(Z \leq b) = \Phi(b) = \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

$P(Z \leq b) =$ Area sotto la curva

- 65% della probabilità in $[\mu - \sigma, \mu + \sigma]$,
- 95% in $[\mu - 2\sigma, \mu + 2\sigma]$

Esempio: Supponiamo $X \in \mathcal{N}(5,4)$. Determinare

(a) $P(4 \leq X \leq 6)$ (b) $P(X \geq 9)$.

Sol. Definiamo Z in $\mathcal{N}(0,1)$, $Z = (X-5)/2$.

$4 \leq X \leq 6 \Rightarrow -0.5 \leq Z \leq 0.5$ Si sfrutta

$$P(-0.5 \leq Z \leq 0.5) = P(Z \leq 0.5) - P(Z \leq -0.5)$$

Usando le tabelle: $P(Z \leq 0.5) = .6915$
$P(Z \leq -0.5) = P(Z \geq 0.5) = 1 - P(Z \leq 0.5) = 0.3085.$

$\Rightarrow P(-0.5 \leq Z \leq 0.5) = P(Z \leq 0.5) - P(Z \leq -0.5) = 0.3830.$

La probabilità $P(X \geq 9)$ si trasforma in $P(Z \geq 2)$ e si ha
$P(Z \geq 2) = 1 - P(Z \leq 2) = 1 - 0.9772 = 0.0228.$

Esempio: Le uova di gallina hanno peso medio di 60 gr. con deviazione standard di 15 gr., e si può pensare ad una loro distribuzione normale. Uova di peso inferiore a 45 gr. sono classificate come *piccole*. Il resto viene diviso tra standard e grandi, ed è sperabile che queste si verifichino con uguale frequenza. A quale peso bisognerebbe distinguere tra standard e grandi? (arrotondare al grammo)

Sol. Determiniamo x_0 tale che $P(45 \leq X \leq x_0) = P(X \geq x_0)$. Si ha
$P(X \geq x_0) = 1 - P(X \leq x_0)$ e
$P(45 \leq X \leq x_0) = P(X \leq x_0) - P(X \leq 45)$, quindi deve valere

$$2P(X \leq x_0) - P(X \leq 45) = 1 \quad \Leftrightarrow \quad P(X \leq x_0) = \frac{1 + P(X \leq 45)}{2}.$$

In forma standard, $Z = (X - 60)/15$, per cui $P(X \leq 45)$ si trasforma in
$P(Z \leq -1) = P(Z \geq 1) = 1 - P(Z \leq 1) = 0.1587$ (dalla Tabella).
Quindi, $P(Z \leq z_0) = (1 + 0.1587)/2 = 0.57935$ e sulla Tabella $z_0 = 0.205$.
Nell'unità di misura dei dati, $x_0 = 15z_0 + 60 = 63.075 \approx 63$ gr.

Esempio: In una specie di pesci adulti, maschi e femmine sono distinguibili dalle dimensioni medie: misurate in cm., le femmine hanno $\mu = 37.5$ e $\sigma = 3.8$, mentre i maschi hanno $\mu = 34.5$ e $\sigma = 3.2$. Qual'è la lunghezza minima del 5% delle femmine con dimensioni maggiori? Quale percentuale di maschi avranno le stesse dimensioni del 30% delle femmine di dimensioni maggiori?

Sol. In una distribuzione normale, nella variabile standardizzata, 95% dei valori sono contenuti per $z \leq 1.645$ (si veda la Tabella).

Il valore soglia è quindi $z_0 = 1.645$, al quale corrisponde $x_0 = z_0\sigma + \mu = 43.7510$. La percentuale di maschi è 5.94%.

Addattamento della distr. normale

In un ospedale viene valutata la dipendenza tra il peso (in grammi) dei neonati e la caratteristica che la neo mamma non abbia mai fumato. I dati risultanti sono riportati nella seguente tabella.

Peso	1-	501-	1001-	1501-	2001-	2501-	3001-	3501-	4001-	4501-	5001-	Tot.
Freq.	4	20	25	73	236	1089	2530	1927	551	101	7	6563

Studiare l'adattamento della distribuzione normale ai dati.

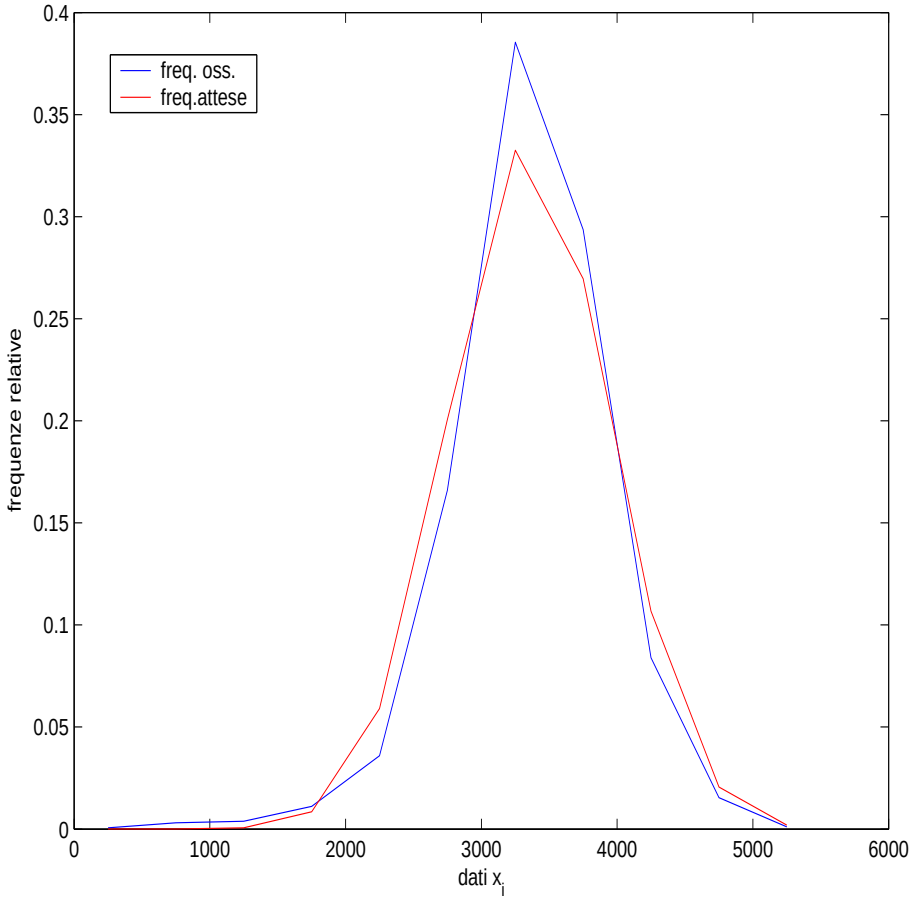
Sol. Considerando come osservazioni i punti centrali di ogni classe, si ha che $\bar{x} = 3353.30$, $s^2 = 3.2782 \cdot 10^5$ e $s = 572.55$.

In tabella:

- dati (valore destro dell'intervallo)
- frequenze relative
- valori standardizzati $z = (x - \bar{x})/s$
- probabilità della distribuzione normale associate ai valori di z
- Frequenze attese (teoriche) relative, $\widetilde{F}(z_i) = P(z_i) - P(z_{i-1})$

Valori ottenuti con il calcolatore

x	f_i/N	z	$P(z)$	$\widetilde{F}(z)$
500	0.00060948	-4.9835	3.1225e-007	3.1225e-007
1000	0.0030474	-4.1102	1.9766e-005	1.9453e-005
1500	0.0038092	-3.2369	0.00060414	0.00058437
2000	0.011123	-2.3636	0.0090483	0.0084441
2500	0.035959	-1.4904	0.068065	0.059017
3000	0.16593	-0.61707	0.26859	0.20053
3500	0.38549	0.25621	0.60111	0.33251
4000	0.29362	1.1295	0.87065	0.26955
4500	0.083956	2.0028	0.9774	0.10674
5000	0.015389	2.8761	0.99799	0.020587
5500	0.0010666	3.7493	0.99991	0.0019247


freq. oss.
freq.attese
frequenze relative
dati x_i
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1000
2000
3000
4000
5000
6000

Valutazione della normalità: Q-Q plot

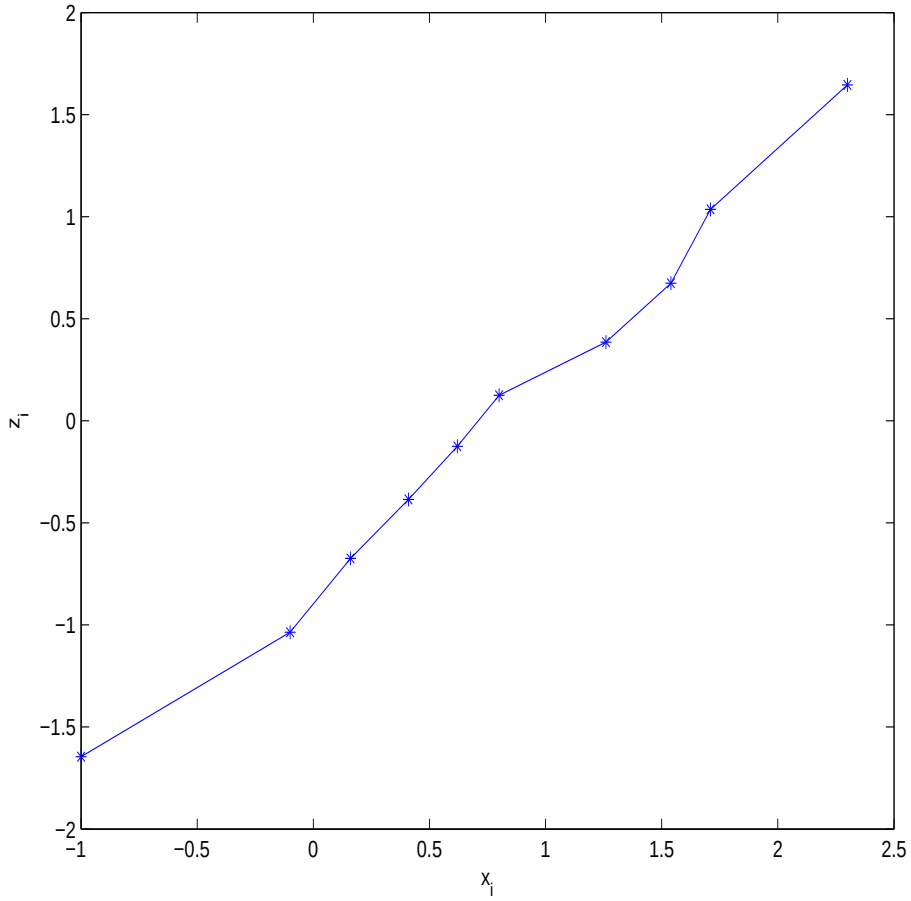
x_i $i = 1, \ldots, N$ ordinati crescenti

z_i tali che $P(z \leq z_i) = \frac{i - \frac{1}{2}}{N}$

Q-Q plot: grafico di (x_i, z_i) (quantili)

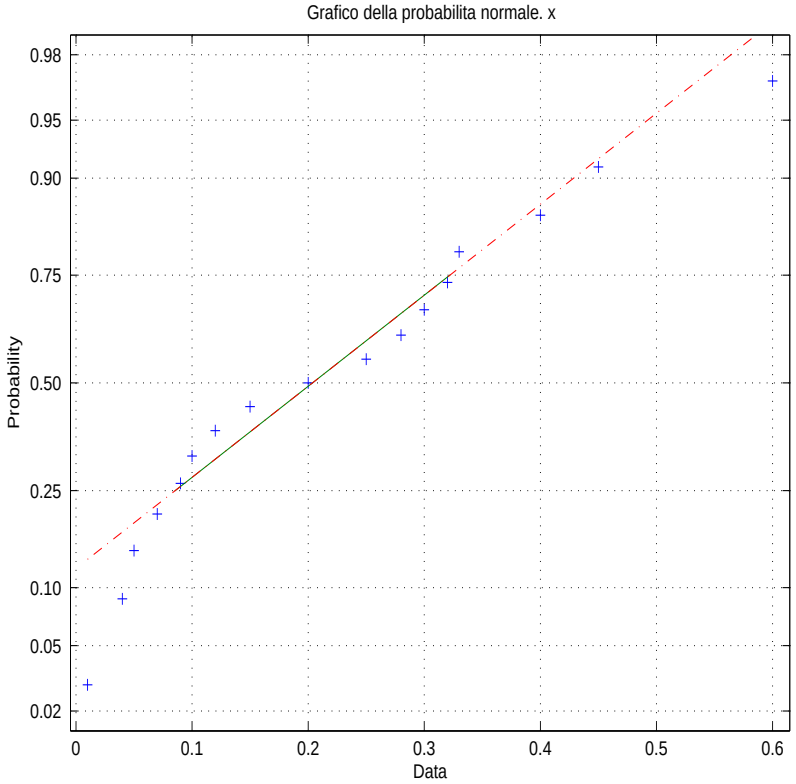
Osserv. x_i	Prob $(i-\frac{1}{2})/n$	quantili (std. normal) z_i
-1.00	0.05	-1.645
-0.10	0.15	-1.036
0.16	0.25	-0.674
0.41	0.35	-0.385
0.62	0.45	-0.125
0.80	0.55	0.125
1.26	0.65	0.385
1.54	0.75	0.674
1.71	0.85	1.036
2.30	0.95	1.645

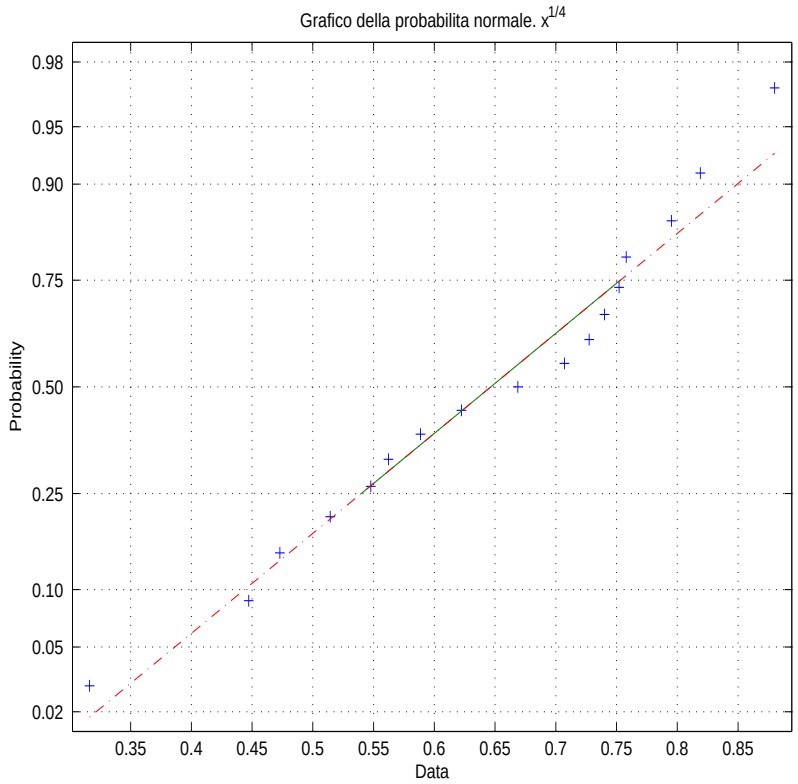
es. $P(z \leq z_7) = (7-0.5)/10 = 0.65$ e si ottiene $z_7 = 0.385$.


2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−1
−0.5
0
0.5
1
1.5
2
2.5
z_i
x_i

Trasformazioni verso la normalità

$x^2, x^3, x^4, \sqrt{x}, \sqrt[4]{x}, \log x, \ln x, e^x, \ldots$

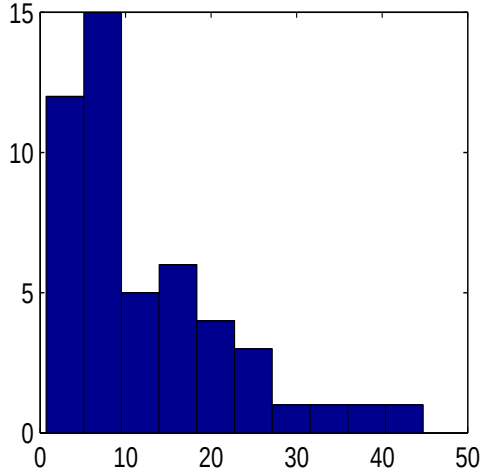


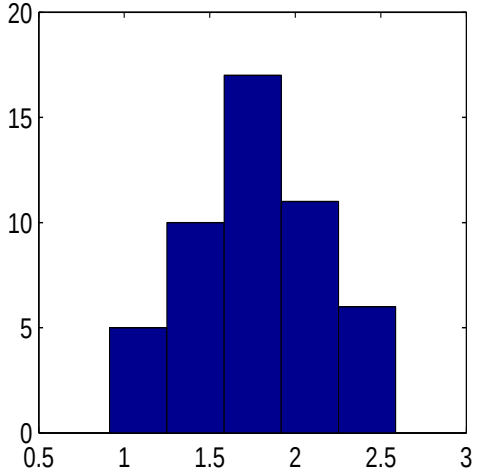


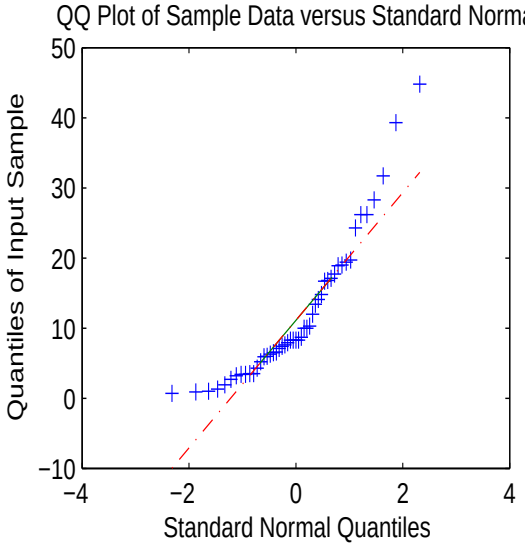
Trasformazioni verso la normalità

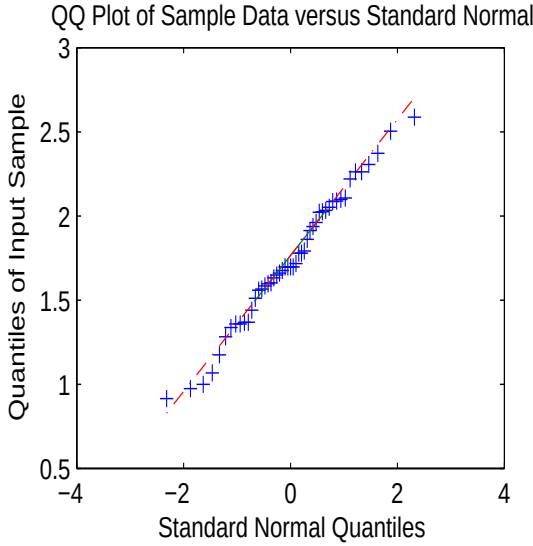
Volume di legno da segheria (stimato dal campionamento del numero di alberi in zone casuali della foresta)

$$\ell = \left\{ \begin{array}{rrrrrrr} 39.3 & 3.5 & 6.0 & 2.7 & 7.4 & 3.5 & 19.4 \\ 19.7 & 1.0 & 8.7 & 14.8 & 8.3 & 17.1 & 26.2 \\ 6.6 & 8.3 & 19.0 & 10.3 & 7.6 & 18.9 & 6.3 \\ 10.0 & 16.8 & 24.3 & 5.2 & 44.8 & 14.1 & 3.4 \\ 28.3 & 3.4 & 0.9 & 1.3 & 0.7 & 17.7 & 8.3 \\ 8.3 & 1.9 & 16.7 & 26.2 & 10.0 & 6.5 & 7.1 \\ 7.9 & 3.2 & 5.9 & 13.4 & 12.0 & 4.3 & 31.7 \end{array} \right\}$$









Volume

$\sqrt[4]{\text{Volume}}$

Outliers

⋆ Criterio di Chauvenet per la distribuzione normale

⋆ Criterio di Peirce

⋆ Test di Grubbs (per distr. normali)

Ma prima ricordare che:

- Arbitrarietà nella decisione sui dati (info a priori sono cruciali);
- Prima ripetere la misurazione varie volte;
- Criteri possono avere ipotesi non sempre realistiche;
- Campioni piccoli sono preziosi.
- Spesso diagrammi sono sufficienti (istogrammi, qqplot, box-plot, ecc.)

Stime

⋆ Distribuzione campionaria

(in generale sono discrete)

Domanda:

è possibile prendere la statistica della distribuzione campionaria come statistica della popolazione?

(*stima* della statistica della popolazione)

Distribuzione campionaria della Media

X variabile casuale, media μ, varianza σ^2 (non note)

$X_1, \ldots, X_n$ campioni casuali $(E[X_i] = \mu) \quad \Rightarrow \bar{X}$ media

Valore atteso: $E[\bar{X}] = \frac{1}{n} E[X_1 + \cdots + X_n] = \mu$

Varianza: $Var[\bar{X}] = \frac{\sigma^2}{n}$

La varianza differisce per un fattore n

Errore standard: La deviazione standard della distribuzione campionaria di una statistica

Esempio: l'errore standard della media campionaria $\bar{x}$ è $\frac{\sigma}{\sqrt{n}}$

Stimatore unbiased e biased

stimatore: operatore (di media, per es.) associato ai campioni

stima: il valore assunto da tale operatore per campioni specifici.

Uno stimatore si dice corretto (*unbiased*) se il valore atteso della sua distribuzione campionaria eguaglia il parametro che vuole stimare.

Es. $\bar{X}$ è uno stimatore corretto di μ nell'esempio precedente.

- Stimatore non corretto è uno stimatore distorto (o *biased*).
- Correttezza non implica accuratezza
- Tra stimatori corretti si sceglie la minore dev.standard (media vs. mediana)

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad \sigma_{mediana} = \sigma\sqrt{\frac{\pi}{2n}} \approx \frac{1.2533\sigma}{\sqrt{n}}.$$

Stime corrette della varianza

X variabile aleatoria (σ, μ incognite)

$x_1, \ldots, x_n$ sono valori campionari

Proviamo a stimare σ^2 sfruttando la varianza del campione,

$$s_0^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$$

Vogliamo valutare se tale varianza è una stima corretta.

Nota: μ non è noto! usiamo $\bar{x}$ invece di μ

Deve valere:

$$E[s_0^2] = \sigma^2$$

Ma si ottiene:

$$E\left[\frac{1}{n}\sum(x_i-\bar{x})^2\right]=\frac{(n-1)}{n}\sigma^2.$$

$\Rightarrow$ s_0^2 stimatore distorto di σ^2

Stimatore corretto:

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

$(E[s^2]=\sigma^2)$

Stime (unbiased) di media e varianza

(riassunto) $x_1, \ldots, x_N$ campione

Parametro media μ:

$$\bar{x} = \frac{1}{N} \sum (x_i)^2$$

Stimatore con distribuzione con media μ e varianza $\frac{\sigma^2}{N}$

Nota: per grandi campioni, $\bar{x}$ ha minore dispersione della popolazione, rispetto alla media

Parametro varianza σ^2:

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

(Stimatore corretto, $E[s^2] = \sigma^2$)

Distribuzioni di medie campionarie

TEOREMA: (Limite centrale). Sia X una variabile casuale con media μ e varianza σ^2. Se $\bar{x}_N$ è la media di un campione casuale di dimensione N, allora la distribuzione di

$$\frac{\bar{x}_N - \mu}{\sigma/\sqrt{N}}$$

tende alla distribuzione normale *standardizzata* per $N \to \infty$.

Considerazioni:

- Nessuna ipotesi sul tipo di distribuzione
- Richieste informazioni su media e varianza
- Per N grande, distribuzioni di medie campionarie sono normali

Esempio: Qual'è la probabilità che il punteggio medio dopo 100 lanci di un dado ecceda 4?

Sol. X : variabile casuale associata ad un singolo lancio, e può assumere valori $1, 2, \ldots, 6$, ognuno con probabilità $p = \frac{1}{6}$

Si ha $\mu = Np = 3.5$ e $\sigma^2 = Np(1-p) = 2.917$

Per $N = 100$:

$$z = \frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} = \frac{\bar{X}_N - 3.5}{0.171},$$

z è circa $\mathcal{N}(0,1)$, cioè $\bar{X}_N$ è $\mathcal{N}(3.5, 2.917)$

Per $\bar{X} = 4$, si ha $z = 2.92$. usando la Tabella:

$$P(z \geq 2.92) = 1 - P(z \leq 2.92) = 1 - 0.9982 = 0.0018$$

Nota:

La probabilità in un singolo lancio di ottenere un valore superiore a 4 è $2/6 = 1/3 \approx 0.333$. Quindi la probabilità che un singolo lancio si discosti di molto dalla media è più alta di quella di un campione.

Proposizione. Se è noto che X ha una distribuzione normale, allora la distribuzione della media $\bar{X}_N$ è esattamente $\mathcal{N}(\mu, \frac{\sigma^2}{N})$ per ogni N.

Esempio: Il contenuto di umidità per pound di una concentrato di proteina disidratata è distribuita normalmente con media 3.5 e deviazione standard 0.5. Viene testato un campione casuale di 16 esemplari, ognuno dei quali corrisponde ad un pound di tale concentrato. a) Qual'è la distribuzione della media campionaria $\bar{x}$? b) Qual'è la probabilità che $\bar{x}$ sia superiore a 3.7?

Sol. a) Dato che la distribuzione della popolazione è normale, anche il campione si distribuisce esattamente in maniera normale, con stessa media e deviazione standard $\sigma/\sqrt{N}$, dove N è il numero di esemplari del campione.

b) Si vuole determinare $P(\bar{X} \geq 3.7)$. Si ha

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} = \frac{\bar{x} - 3.5}{0.5/4}$$

In termini di Z, si vuole quindi determinare $P(Z \geq 1.6)$, dove $1.6 = \frac{3.7-3.5}{0.5/4}$.
Dalle tabelle si ottiene $P(Z \geq 1.6) = 1 - P(Z \leq 1.6) = 1 - 0.9452 = 0.0548$.

Approssimazione normale alla distribuzione binomiale

- $P_b(X \leq r)$ costoso
- $P_b(X \leq r) \approx P(X \leq r)$ per N grande

Teorema: Se la variabile casuale X ha una distribuzione binomiale con parametri N e p, allora, per N grande, la distribuzione di X può essere approssimata da una distribuzione normale con media Np e varianza $Np(1-p)$.

$(Np > 5,\ N(1-p) > 5)$

Correzione di continuità

Attenzione: $P_b(X = r) \geq 0$, mentre $P(X = r) = 0$.

Il valore r è allargato all'intervallo $(r - \frac{1}{2}, r + \frac{1}{2})$:

$$P_b(X = r) \qquad \leftrightarrow \qquad P\left(r - \frac{1}{2} \leq X \leq r + \frac{1}{2}\right)$$

$$P_b(X \leq r) \approx P\left(X \leq r + \frac{1}{2}\right)$$

$$P_b(X \geq r) \approx P\left(X \geq r - \frac{1}{2}\right)$$

Ipotesi statistiche

Ipotesi poste sulla (distribuzione di) popolazione per raggiungere una decisione sulla popolazione stessa

L'ipotesi che si vuole testare: H_0 (ipotesi nulla)

L'ipotesi che si prende in caso si scarti H_0: H_1 (ipotesi alternativa)

es. Moneta truccata? p=prob. testa. $H_0 =$ "$p = 0.5$",
$H_1 =$ "$p > 0.5$", oppure $H_1 =$ "$p = 0.7$", $H_1 =$ "$p \neq 0.5$".

Test di significatività

1. Si suppone H_0 vera
2. Si considera il campione con le statistiche di H_0
3. Si verifica se il campione differisce *significativamente* dalla popolazione

- Errore di I tipo: rifiutiamo H_0 che invece era buona
- Errore di II tipo: accettiamo H_0 che invece era da scartare

Per limitare l'errore, allargare il campione

Livello di significatività

In H_0, distribuzione nota

Campione: $\Rightarrow$ Evento con probabilità π

$$\Downarrow$$

π inferiore a soglia $\alpha \Rightarrow$ Scartiamo H_0

α=livello di significatività (scelto a priori)

α=5%, 1%, ...

$\pi < \alpha$: Evento improbabile (unlikely)

$z : |z| > z_0$, Regione critica

La distribuzione normale standard

$$P(-1.96 \leq z \leq 1.96) = 0.95 \qquad P(-2.576 \leq z \leq 2.576) = 0.99$$

$$P(-3 \leq z \leq 3) = 0.9973 \qquad P(-3.291 \leq z \leq 3.291) = 0.999.$$

Nota: L'evento di essere al di fuori di (-1.96, 1.96) è *improbabile*, se è stato scelto il 5% di livello di significatività.

X in $\mathcal{N}(\mu, \sigma^2)$

$$P(\mu - 1.96\sigma \leq x \leq \mu + 1.96\sigma) = 0.95$$

Test ad una e a due code

L'ipotesi alternativa H_1 influenza l'ampiezza della regione critica

Test a due code. Zona critica comprende entrambi gli estremi

Test ad una coda. Zona critica comprende un solo estremo

Esempio: Sia H_0="x è $\mathcal{N}(0,1)$".

- H_1="x è $\mathcal{N}(\mu,1)$ con $\mu \neq 0$". Rifiutiamo H_0 se x cade sotto -1.96 o sopra 1.96. (5% **livello di sign.**)

  Per tali valori, le distribuzioni $\mathcal{N}(\mu_1,1)$, $\mu_1 < 0$ o $\mathcal{N}(\mu_2,1)$, $\mu_2 > 0$ possono meglio rappresentare l'osservazione nell'ipotesi alternativa.

- H_1="x è $\mathcal{N}(\mu,1)$ con $\mu > 0$", Solo le osservazioni in $z > 1.96$ sono meglio rappresentate da H_1 (una sola coda) (2.5% livello sign.)

  $z > 1.64$ per 5% liv. sign.

Esempio In una certo laghetto un ecologo ha trovato parecchi pesci appartenenti ad una certa specie. Nel laghetto vicino, dopo alcune ore di ricerca ha recuperato solo un pesce che assomiglia alla specie del primo laghetto. I pesci del primo laghetto hanno lunghezza media 12 cm e varianza 0.64 cm^2, con distribuzione apparentemente normale. La lunghezza del nuovo pesce è 13.3 cm. Il nuovo pesce è della stessa razza degli altri?

Sol. H_0="il pesce è della stessa razza dei pesci dell'altro laghetto, con media 12cm." H_1=" Il pesce è dello stesso tipo, ma di razza diversa, con media diversa da 12cm". Test a due code. Sotto H_0, la lunghezza standardizzata è $z = (13.3 - 12)/0.8 = 1.63$. Usando un livello di significatività 5%, la regione critica è data dai valori di z tali che $|z| > 1.96$, e quindi il nostro valore di z non rientra in tale intervallo.

Sulla base di questo test possiamo confermare che il pesce trovato appartiene alla razza dell'altro laghetto.

Esempio: L'altezza di una certa varietà di pianta in una serra, durante la sua prima stagione di vita è stata per lungo tempo distribuita normalmente, con media 53 cm e varianza 12 cm^2. Un anno, per colpa di un errore, una di tale piante ha ricevuto tre applicazioni di fertilizzante invece di due, ed ha raggiunto i 60 cm. Valutare se il fertilizzante ha avuto un effetto benefico.

Sol. Altezza in $\mathcal{N}(\mu, 12)$. H_0="$\mu = 53$". H_1="$\mu > 53$". Test ad una coda. 5% livello di significatività. $z = (x - \mu)/\sigma = 2.02$. La regione critica è $z > 1.64$. Viene rifiutata l'ipotesi nulla in favore di H_1.

Test di ipotesi con la distribuzione normale

Campione casuale di N osservazioni da $\mathcal{N}(\mu, \sigma^2)$

$$\Downarrow$$

$\bar{X}$ in $\mathcal{N}(\mu, \frac{\sigma^2}{N})$

Quindi

$[\mu - \frac{\sigma}{\sqrt{N}}, \mu + \frac{\sigma}{\sqrt{N}}]$ corrisponde al 65% della probabilità

$[\mu - 2\frac{\sigma}{\sqrt{N}}, \mu + 2\frac{\sigma}{\sqrt{N}}]$ corrisponde al 95% della probabilità

Esempio: I punteggi di un test del quoziente di intelligenza (IQ) sono distribuiti normalmente con deviazione standard 15. Sull'intera popolazione la media è 100. Una compagnia ha un proprio test di selezione, che sembra garantire la selezione di persone con un IQ migliore, in media. 20 degli impiegati di questa compagnia sono stati assunti usando questo test, ed il loro IQ standard ha media 104.2. Esaminare l'asserto della compagnia.

Sol. $H_0 =$ "$\mu = 100$", $H_1 =$ "$\mu > 100$". Test ad una coda, con livello di significatività 2.5%. Secondo l'ipotesi nulla, $z = (\bar{x}_N - 100)/(15/\sqrt{20})$ è $\mathcal{N}(0,1)$. Usando i dati, si ha $z = (104.2 - 100)\sqrt{20}/15 = 1.25$. La regione critica è per $z > 1.96$, quindi per il valore di z trovato, $z = 1.25$, l'ipotesi nulla non deve essere scartata. L'evidenza è che il test della compagnia non differisca da quello standard.

Test su grandi campioni (Teorema del Limite centrale)

Esempio: I tempi di volo di un velivolo su una certa tratta hanno avuto sempre media 16 1/4 ore, sin dal primo volo con quel velivolo su quella tratta. La distribuzione dei tempi di volo ha avuto deviazione standard 1 1/2 ore, anche se non è normale. Negli ultimi 120 voli, il tempo di volo è stato di 15 ore e 56 minuti. Questo tempo è diverso dalla media?

Sol. $N = 120$ grande. Assumiamo che la distribuzione delle ore di volo non sia troppo asimmetrica, e che quindi la media delle ore di volo $\bar{X}$ abbia distribuzione quasi normale: $\bar{X} \approx \mathcal{N}(\mu, \frac{(1.5)^2}{120})$

$$H_0 = \text{“}\mu = 16.25\text{”}, \qquad H_1 = \text{“}\mu \neq 16.25\text{”}$$

Test a due code: $z = \dfrac{(15.933 - 16.250)}{1.5/\sqrt{120}} = -2.31$

Regione critica: $|z| > 1.96$ (5% di livello di sign.)

Anche se la nostra distribuzione non è esattamente normale, il valore -2.31 è sufficientemente dentro la regione critica da poter senza dubbio rifiutare l'ipotesi H_0 in favore di H_1, cioè il velivolo ha cambiato tempo di volo medio.

Probabilità di significatività: i p-valori

Esempio. Supp. livello di sign. $\alpha = 2.5\%$ (piccolo!)

Supp. di rifiutare l'ipotesi H_0 perchè il valore osservato $z_0 = -2.52$ ricade nella regione critica (ad una coda) $z \leq -1.96$.

Domanda: quant'è dipesa la decisione dal valore di α?

Valore minimo di α che dà ancora rifiuto?

- $z_0 = -2.52$ valore soglia (valore critico) per la variabile standardizzata
- Si valuta $P(z \leq -2.52) = 0.0059$.

$\Rightarrow \quad \alpha_{\min} = 0.0059 \qquad$ (rifiuto forte di H_0)

Test sulla media campionaria e t-distribuzione

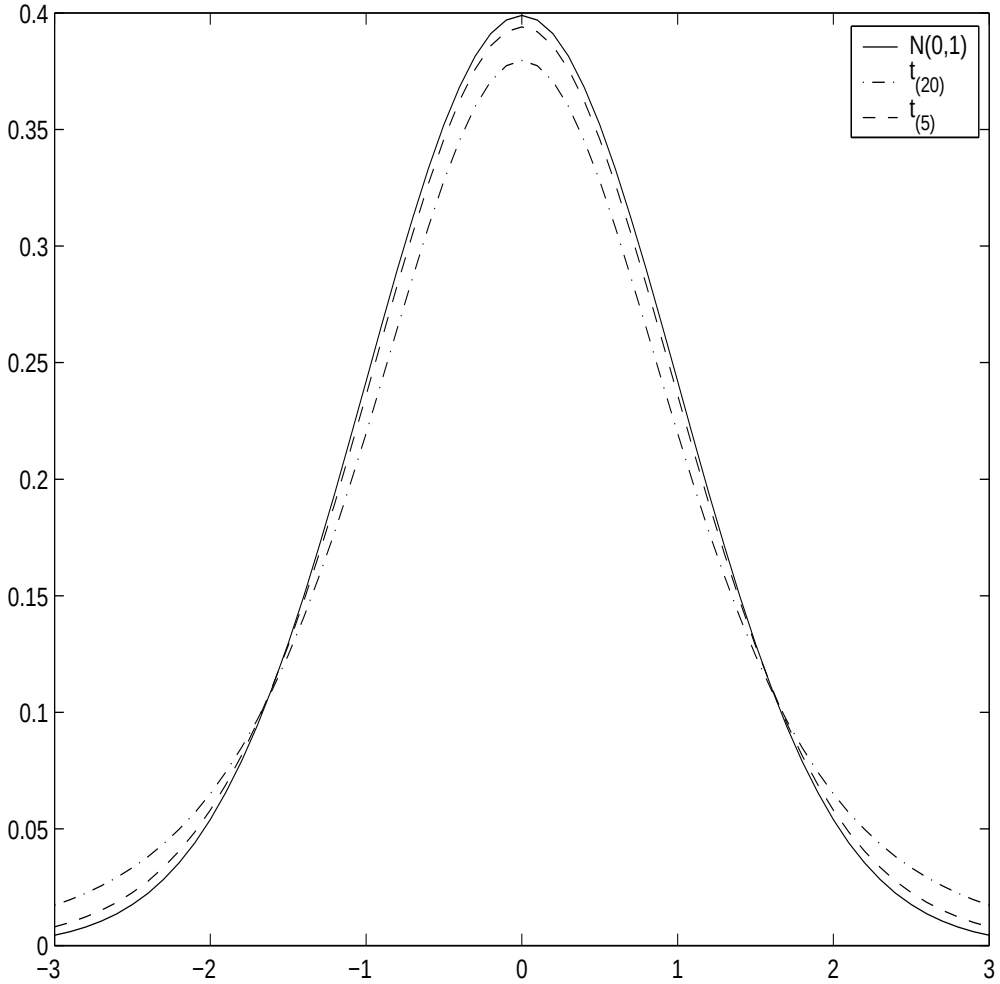
Test del significato di una media di variabili distribuite normalmente, se σ incognita (N piccolo):

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \Rightarrow t = \frac{\bar{X} - \mu}{s/\sqrt{N}}. \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

t distribuzione di Student (Gosset)

- t non $\mathcal{N}(0,1)$
- t ha due variabili $\bar{X}$, s
- t varia al variare di N $\quad \Rightarrow$ distribuzione
- t ha media zero
- t ha $N-1$ gradi di libertà: $t = t_{(N-1)}$

Funzione di densità della t di Student



Tabelle

Esempio: In un campione di coccinelle, di una particolare località, la larghezza (in mm.) di ogni coccinella è 28, 21, 26, 16, 18, 13, 15, 22, 19, 22 mm. Precedenti misurazioni avevano stabilito che le coccinelle della stessa specie hanno media 23 mm. Testare se le coccinelle della località prescelta hanno una larghezza media diversa da quella della popolazione.

Sol. Modello: la larghezza X è $\mathcal{N}(\mu, \sigma^2)$.

$$H_0 = \text{“}\mu = 23'' \qquad H_1 = \text{“}\mu \neq 23''$$

Test a due code. Dal campione, si trova $\bar{x} = 20$ e $s^2 = \frac{68}{3} = 22.67$,

$$t = \frac{(\bar{x} - \mu)}{s/\sqrt{N}} = -1.99$$

Nell'ipotesi H_0, t è distribuita come la t di Student con 9 gradi di libertà.

Regione critica: $|t| > 2.26$ (5% livello sign.)

$\Rightarrow$ H_0 non viene scartata (non ci sono indizi perchè le coccinelle della località abbiano media diversa dalla popolazione)

Un'applicazione: valutazione della correlazione

Supp. (x_i, y_i), $i = 1, \ldots n$ coppie di osservazioni.

Coefficiente di correlazione: $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$

$0 \leq r^2 \leq 1$. Come valutare la bontà della correlazione (lineare) ?

Supp. X e Y variabili con distribuzione normale. La correlazione ρ delle due variabili può essere testata dal campione come:

$$H_0 \ : \ \rho = 0 \qquad \text{Testata con } t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{(n-2)}$$

solitamente $H_1 \ : \ \rho \neq 0$ (ma anche $\rho > 0$ può avere un senso)

Esempio

Test per determinare la relazione tra il contenuto chimico di un particolare costituente (x) in soluzione, e la temperatura (y) di cristallizzazione.

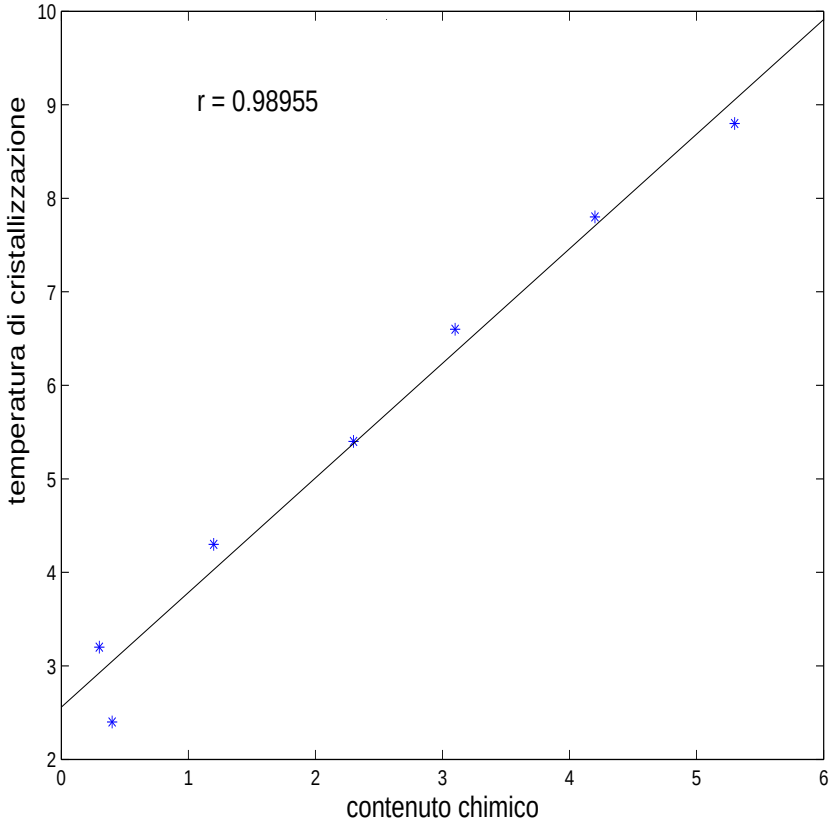
Primo test:

x	0.3	0.4	1.2	2.3	3.1	4.2	5.3
y	3.2	2.4	4.3	5.4	6.6	7.8	8.8

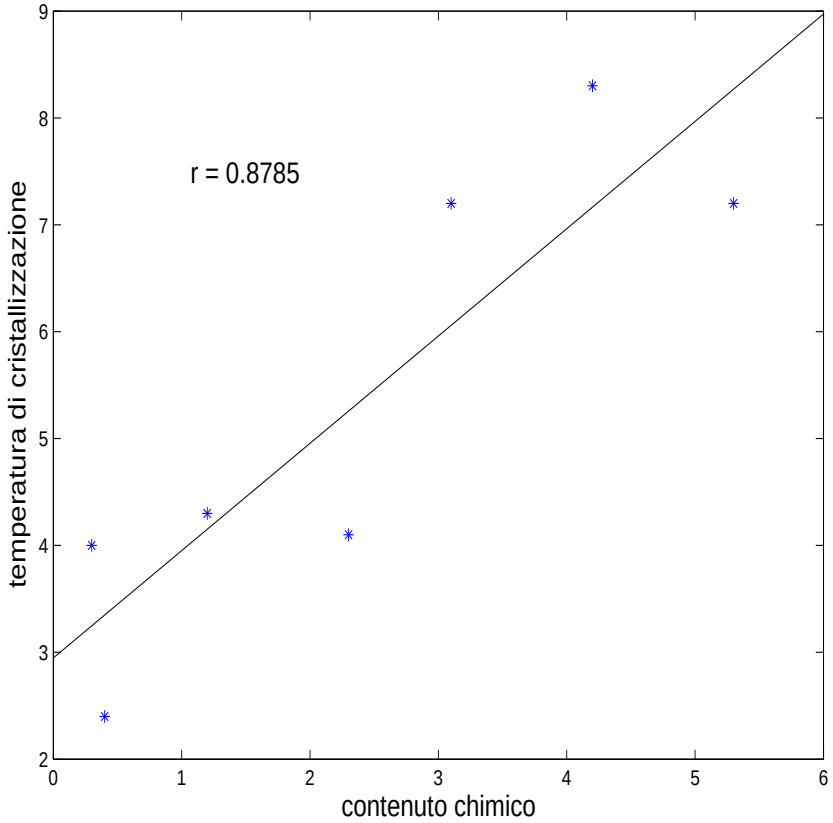
Secondo test:

x	0.3	0.4	1.2	2.3	3.1	4.2	5.3
$\hat{y}$	4.0	2.4	4.3	4.1	7.2	8.3	7.2

Livello di significatività $\alpha = 5\%$



Primo test. Diagramma di dispersione e retta di regressione.



Secondo test. Diagramma di dispersione e retta di regressione.

$$n = 7,\ r = 0.8785, \quad t = 0.8785\sqrt{\frac{5}{1 - 0.8785^2}} = 4.1127 > t_{(5)} = 2.571$$

Riassunto...

Test di Ipotesi statistiche:

- Distribuzione normale *standard*
  1. Analisi di una nuova osservazione in $\mathcal{N}(0,1)$
  2. Analisi di N osservazioni in $\mathcal{N}(\mu,\sigma^2)$ (in $\mathcal{N}(\mu,\sigma^2/N)$)
- Test su grandi campioni
- Test su piccoli campioni (da distr. normale) con σ non nota (t di Student)

Intervalli di confidenza

- X variabile casuale $\mathcal{N}(\mu, \sigma^2)$
- Un solo campione disponibile

Stime per μ incognita: intervallo (già viste stime puntuali)

Campione $\{x_1, \ldots, x_N\}$ è $\mathcal{N}(\mu, \frac{\sigma^2}{N})$

$$\Rightarrow \qquad \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \quad \text{è} \quad \mathcal{N}(0,1)$$

$z \in \mathcal{N}(0,1) \quad \Rightarrow \quad P(-1.96 \leq z \leq -1.96) = 0.95$

($P(-1.96 \leq z) = 0.025$)

Livello di significatività: 95%

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \leq 1.96\right) = 0.95$$

Da cui

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{N}}$$

disuguaglianza vera con probabilità 95%

Intervalli basati su grandi campioni

Campione ($N \geq 30$) Statistiche campionarie $\approx \mathcal{N}(\mu, \sigma^2/N)$

$$\bar{X} \pm 2\frac{\sigma}{\sqrt{N}} \qquad \text{(int.confidenza 95\%)}$$

Esempio. È stato riscontrato che un campione casuale tra 100 cestini di mele con un peso segnalato di 1 kg l'uno, ha in effetti peso 1020 gr. La varianza stimata dal campione è 144 gr.^{2}. Determinare un intervallo di confidenza del 95% per la media dei cestini.

Sol. Usiamo $\bar{X} - 2\frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + 2\frac{\sigma}{\sqrt{N}}$.

Si ottiene $\bar{X} \pm -2\frac{\sigma}{\sqrt{N}} = 1020 \pm 2 \cdot \frac{12}{10}$

$\Rightarrow \quad 1017.6 \leq \mu \leq 1022.4$

Intervalli per piccoli campioni

Campione ($N \leq 30$) $\quad \Rightarrow \quad \sigma^2 \to s^2$

$$\frac{\bar{X} - \mu}{s/\sqrt{N}}$$

ha distribuzione t di Student ($N - 1$ gradi di libertà)

Con una confidenza del 95%, intervallo di stima per la media μ:

$$P\left(-t_{(N-1)} \leq \frac{\bar{X} - \mu}{s/\sqrt{N}} \leq t_{(N-1)}\right) = 0.95$$

Estremi: $\quad \bar{X} \pm \frac{s}{\sqrt{N}} t_{(N-1)}$

Esempio. Quindici pomodori in una serra hanno altezza media 83 cm e deviazione standard 5.8 cm. Determinare l'intervallo di confidenza al 95% per l'altezza media (si assume una distribuzione normale della popolazione).

Sol. $N = 15$, per $P = 0.05$ si ottiene $t_{(14)} = 2.145$. Gli estremi sono

$$83 \pm \frac{5.8}{\sqrt{15}} 2.145 \approx 83 \pm 3.21$$

Osservazione: Nella maggior parte dei casi, nè μ nè σ sono note in una popolazione con distribuzione normale. Stime della varianza: distribuzione χ^2

Osservazione: Metodi usati basati su stima da campione

$\Rightarrow$ Metodo dei momenti

Più generali: **Metodi di massima verosimiglianza** e Minimi quadrati

Distribuzione χ^2

X variabile aleatoria $\mathcal{N}(\mu, \sigma^2)$

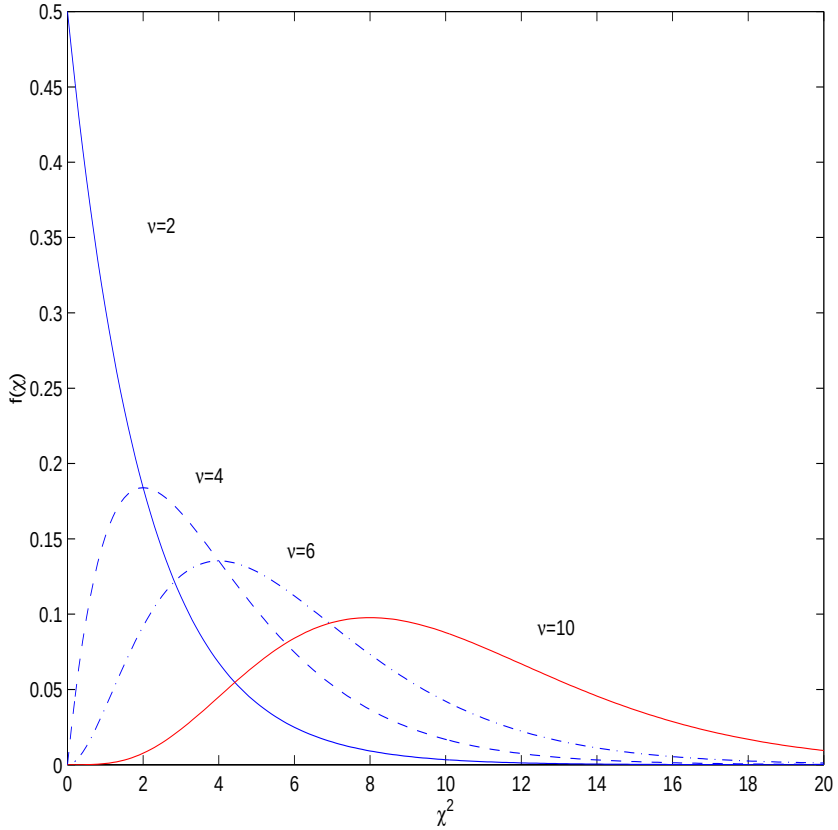
Campione $x_1, \ldots, x_N$

Introduciamo la statistica:

$$\chi^2 = \frac{(N-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{\sigma^2} \qquad \nu = N - 1$$

La Distribuzione χ^2 è la distribuzione campionaria di questa statistica

Funzione di densità di probabilità: $f(\chi) = y_0 \chi^{\nu-2} e^{-\frac{1}{2}\chi^2}$



media: ν varianza: 2ν Asimmetria a destra (media $>$ moda)

Tabella: per regioni critiche $\chi^2 \geq \chi^2_{(\nu)}$ ed intervalli di confidenza con la distribuzione χ^2.

Solitamente usata ad una coda, a destra o a sinistra.

Intervalli di confidenza per la varianza di una distribuzione normale

X variabile aleatoria in $\mathcal{N}(\mu, \sigma^2)$

Intervallo di confidenza di σ^2 da un campione di grandezza N:

$\bar{x}, s^2$ media e varianza campionarie

$$\frac{(N-1)s^2}{\sigma^2} \quad \text{distribuzione } \chi^2$$

$$P\left(\chi^2_L \leq \frac{(N-1)s^2}{\sigma^2} \leq \chi^2_U\right) = 0.95$$

$$P\left(\frac{(N-1)s^2}{\chi^2_U} \leq \sigma^2 \leq \frac{(N-1)s^2}{\chi^2_L}\right) = 0.95$$

Esempio. L'altezza media di un campione di quindici pomodori in serra è di 83 cm con deviazione standard 5.8 cm. Determinare limiti di confidenza del 95% per la varianza della popolazione, assumendo che l'altezza della popolazione sia distribuita in modo normale

Sol.

Nella Tabella $\chi^2_{(14)}$ con entrambe le code:

Per 2.5% $\chi^2_U = 26.12$

Per 97.5% $\chi^2_L = 5.63$

$$\frac{(N-1)s^2}{\chi^2_L} = \frac{14 \cdot 5.8^2}{5.63} = 83.36 \qquad \frac{(N-1)s^2}{\chi^2_U} = \frac{14 \cdot 5.8^2}{26.12} = 18.03$$

$$18.03 \leq \sigma^2 \leq 83.36$$

Test χ^2

Distribuzione χ^2 usata per valutare *fitting* di frequenze teoriche attese a frequenze osservate

$\{o_i\}$ frequenze osservate $\{e_i\}$ frequenze attese

Statistica: $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} \approx$ distr. $\chi^2_{(\nu)}$

Livello di significatività $\chi^2 \geq \chi^2_{(\nu)}$

ν gradi di libertà: $\nu =$ # frequenze - 1 - # parametri stimati

Nota: $\{e_i\}$, $i = 0, 1, \ldots, k$, $e_i \geq 5$

Fitting di dati a distribuzione discrete

Distribuzione binomiale:

Esempio. Tra 320 famiglie con 5 figli. Numero r di figli maschi

r:# maschi	0	1	2	3	4	5	Totale
freq. osserv.	8	40	88	110	56	18	320=N

Evento R="nascita di un figlio maschio "

Test: H_0 = "R ha distribuzione binomiale, $n = 5$ e $p = 0.5$"

Livello di significatività: 5%

Frequenze attese:

$$e_r = N \cdot P(R = r) = N \frac{n!}{r!(n-r)!}(0.5)^n$$

r:# maschi	0	1	2	3	4	5	Totale
freq. osserv.	8	40	88	110	56	18	320=N
freq. attese	10	50	100	100	50	10	320

$$\chi^2 = \sum_{r=0}^{5} \frac{(o_r - e_r)^2}{e_r} = \frac{(8-10)^2}{10} + \frac{(40-50)^2}{50}$$
$$+ \frac{(88-100)^2}{100} + \frac{(110-100)^2}{100} + \frac{(56-50)^2}{50} + \frac{(18-10)^2}{10} = 11.96$$

Gradi di libertà: $6 - 1 = 5 \quad \Rightarrow \chi^2_{(5)} = 11.07 \quad \chi^2 = 11.96 > \chi^2_{(5)}$

L'ipotesi nulla deve essere scartata

Nota: con livello di sign. all' 1% l'ipotesi non verrebbe scartata

Alternativa

Evento R=“nascita di un figlio maschio ”

Test: H_0 = “R ha distribuzione binomiale, $n = 5$ e $p = \tilde{p}, \tilde{p} > 0.5$”

Stima di $\tilde{p}$ dai dati: $\bar{o} \approx \mu = E[R] = np$

$$\bar{o} = \frac{1}{N}\sum_r (r \cdot o_r) = \frac{860}{320} = \frac{43}{16}, \qquad \tilde{p} = \frac{\bar{o}}{n} = 0.5375.$$

$$\tilde{e}_r = NP(X = r) = N \binom{5}{r} \tilde{p}^r (1 - \tilde{p})^{n-r}$$

r:# maschi	0	1	2	3	4	5	Totale
# osserv.	8	40	88	110	56	18	320=N
# approx. attese	6.8	39.3	91.5	106.3	61.8	14.4	320.1

$\chi^2 = 1.93$. Gradi di libertà: $6 - 1 - 1 = 4$

$\chi^2_{(4)} = 9.49$ $\chi^2 < \chi^2_{(4)}$

Accettiamo l'ipotesi di distribuzione binomiale, con $p > 0.5$

Fitting di dati a distribuzione di Poisson

Esempio. Il numero r di lettere ricevute ogni giorno in un arco di 100 giorni è distribuito con la frequenza riportata nella tabella sotto.

Si puó dire che la distribuzione è di tipo Poissoniana? (sign. 5%)

r: # lettere	0	1	2	3	4	5	Totale
freq.oss.	48	32	17	2	0	1	100

Sol. Stimiamo λ con $\bar{r} = 0.77 \quad \Rightarrow \quad P(X = r) = e^{-\bar{r}} \frac{(\bar{r})^r}{r!}$

r: # lettere	0	1	2	≥ 3
freq.att.	46.30	35.65	13.73	4.32
freq.oss.	48	32	17	3

Gradi di libertà: $4 - 2 = 2$.

$\chi^2_{(2)} = 5.99 \qquad \chi^2 = 1.618$

Non rifiutiamo l'ipotesi

Fitting di dati a distribuzioni continue. Distribuzione esponenziale.

Esempio. Tempi di durata batterie elettriche.

$$\lambda = \frac{1}{\bar{x}}$$

Quindi otteniamo la seguente tabella

Tempo	-50	-100	-150	-200	-250	-300	-350	-400	> 400
freq. oss.	208	112	75	40	30	18	11	6	0
freq. attese	204.6	120.9	71.4	42.2	24.9	14.7	8.7	5.1	7.4

$\chi^2 = 10.96$

Gradi di libertà: $9 - 2 = 7$

(livello di sign. 5%) $\chi_{(7)} = 14.07 \Rightarrow$ buon adattamento

Fitting di dati a distribuzioni continue. Distribuzione normale

Esempio. Peso, in gr., di un gruppo di 100 animali.

Peso (gr.)	≤ 89	-109	-129	-149	-169	-189	≥ 190
Freq. oss.	3	7	34	43	10	2	1
Freq. att.	2	12	32	36	15	3	0

media e deviazione standard stimate dai dati: $\bar{o} = 131.5, \bar{s} = 20$

Raggruppamento di frequenze attese

Peso (gr.)	≤ 109	-129	-149	-169	≥ 170
Freq. oss.	10	34	43	10	3
Freq. att.	14	32	36	15	3

Gradi di libertà $5 - 3 = 2$, $\chi_{(2)} = 5.99$

$$\chi^2 = \sum_r \frac{(o_r - e_r)^2}{e_r} = 4.2956$$

Accettiamo l'ipotesi di distribuzione normale

Indice di dispersione per il Fitting della distr. di Poisson

Distr. Poisson $\Rightarrow \quad \lambda = \mu = \sigma^2$

Statistiche campionarie: $\frac{\bar{r}}{s^2} \approx 1$ se da popolazione di Poisson

$$\text{Indice di dispersione : } \quad \mathcal{I} = \sum_{i=1}^{n} \frac{f_i (r_i - \bar{r})^2}{\bar{r}}$$

$\star$ Per n, λ grandi, $\mathcal{I}$ distribuita come $\chi^2_{(n-1)}$

Esempio: (test precedente 5% liv.sign.) $\bar{r} = 0.77,$

r: # lettere	0	1	2	3	4	5	Totale
freq.oss.	48	32	17	2	0	1	100 $=n$

$$\mathcal{I} = \sum_{i=1}^{n} \frac{f_i (r_i - \bar{r})^2}{\bar{r}} = \frac{83.71}{0.77} \approx 108.71 \leftarrow \chi^2_{(99)}$$

$\chi^2_{(100)} = 124.34, \chi^2_{(90)} = 113.15$ $\quad \mathcal{I}$ non significativo

Confronto di Trattamenti

Test sui risultati di un nuovo prodotto, ecc.

Disegno di esperimenti:

- Coppie di campioni

  Elementi di ogni coppia sono simili

  Coppie diverse sono dissimili

- 2 insiemi di campioni indipendenti

Disegno sperimentale con confronto di coppie

Coppia	1	2	$\cdots$	N
Tratt. 1	x_1	x_2	$\cdots$	x_N
Tratt. 2	y_1	y_2	$\cdots$	y_N

La risposta è influenzata da

- le caratteristiche della coppia
- l'effetto del trattamento

Con la differenza si prevede di eliminare l'influenza della coppia

Campione casuale: $d_i = x_i - y_i, \quad i = 1, \ldots, N$

$\bar{d}$ media camp. $\quad s_D^2 = \dfrac{1}{N-1} \sum_{i=1}^{N} (d_i - \bar{d})^2$ varianza camp.

Confronto di coppie

Campione casuale: $d_i = x_i - y_i, \quad i = 1, \ldots, N$

$$E(d_i) = \delta \qquad Var(d_i) = \sigma_D^2 \qquad i = 1, \ldots, N$$

$\delta = 0$ trattamenti equivalenti.

Per N grande, $\dfrac{\bar{d} - \delta}{s_D/\sqrt{N}} \quad \sim \quad \mathcal{N}(0,1)$

Per N piccolo, assumendo distr. normale $\mathcal{N}(\delta, \sigma_D^2)$

$t = \dfrac{\bar{d} - \delta}{s_D/\sqrt{N}}$ è t-distribuzione con $N-1$ g.d.l.

Esempio. Studio degli effetti di un medicinale sulla pressione sanguigna del paziente. Dati x: pressione di 15 pazienti prima della cura. Dati y: pressione degli stessi 15 pazienti dopo 6 mesi di cura

dati $x_i - y_i$: 2, 8, 10, 6, 18, 10, 4, 26, 18, -8, 0, 32, 0, -4, 10

a) L'uso del medicinale riduce la pressione? (1% liv.sign.) b) Calcolare intervalli di conf. (95%) per la media della variazione della pressione.

Sol. Si ha $\bar{d} = 8.80 \quad s_D = 10.98$ con $N = 15$ piccolo.

a) $H_0 : \delta = 0, \quad H_1 : \delta > 0 \qquad t = \frac{\bar{d}}{s_D/\sqrt{N}}$, g.d.l. 14

regione critica: $t \geq t_0 = 2.624$. Ma $t = \frac{8.80}{2.84} = 3.10$ rifiutiamo l'ipotesi

b) Assumiamo che i dati provengano da distribuzione normale.

$$\bar{d} \pm t_0 \frac{s_D}{\sqrt{15}} = 8.80 \pm 6.08 \qquad g.d.l. = 14, t_0 = 2.145$$

(intervallo positivo)

Insiemi di campioni indipendenti

Confronto di due campioni indipendenti da stessa popolazione

- Inferenze su grandi campioni (test, intervalli)
- Inferenze su piccoli campioni
  (*Pooling o no.*)

Inferenze su grandi campioni

$x_1, \ldots, x_{N_1}$ campione da popolazione 1

$y_1, \ldots, y_{N_2}$ campione da popolazione 2

(in effetti, stessa popolazione ...)

campioni indipendenti

Popolazione 1: μ_1, σ_1 Popolazione 2: μ_2, σ_2

Statistiche campionarie:

$$\bar{x}, \quad s_1^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (x_i - \bar{x})^2 \qquad \bar{y}, \quad s_2^2 = \frac{1}{N_2 - 1} \sum_{i=1}^{N_2} (y_i - \bar{y})^2$$

Differenze tra i due trattamenti: parametro $\mu_1 - \mu_2$

Inferenze sulla media

Ipotesi: N_1, N_2 grandi (≥ 30)

$\Rightarrow$ $\bar{x} - \bar{y} \sim \mathcal{N}$ con

$$E[\bar{x} - \bar{y}] = \mu_1 - \mu_2 \qquad Var[\bar{x} - \bar{y}] = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

$$ErroreSt.[\bar{x} - \bar{y}] = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

Inferenze sulla media. Test di ipotesi

$H_0 : \mu_1 = \mu_2$ (od anche $\mu_1 - \mu_2 = 0$)

ipotesi alternativa: $H_1 : \mu_1 \neq \mu_2$, oppure per es. $H_1 : \mu_1 - \mu_2 > 0$

usando la statistica

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim \mathcal{N}(0, 1)$$

Nota: analogo per $H_0 : \mu_1 - \mu_2 = \delta_0$

Esempio. Nel giugno 1993, analisi chimiche su 85 campioni d'acqua (di una unità di volume l'uno) del lago di una città. Studio del contenuto di cloro: media 18.3 e deviazione standard 1.2. Nei successivi due inverni, si ha una riduzione di uso di sale nelle strade nei dintorni del lago. Nel giugno 1995, 110 campioni d'acqua. Studio del contenuto di cloro: media 17.8 e deviazione standard 1.8. Valutare l'affermazione (liv.sign. 5%) che la diminuizione di uso di sale ha influito sulla quantità di cloro nel lago.

Sol. $N_1 = 85, \bar{x} = 18.3, s_1 = 1.2, \quad N_2 = 110, \bar{y} = 17.8, s_2 = 1.8$

$H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 > 0$

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \approx \frac{0.5}{0.2154} \approx 2.32$$

Test ad una coda. $z > 1.645$, scartiamo l'ipotesi. I p-valori:

$P(z \geq 2.32) = 1 - P(z \leq 2.32) = 1 - 0.9898 = 0.0102$ molto basso.

Forte rifiuto di H_0 in favore di H_1

Inferenze sulla media per piccoli campioni

Ipotesi aggiuntive per fare inferenza:

- Entrambe le popolazioni sono normali
- $\sigma_1 = \sigma_2$

la bontà dell'inferenza dipende dalla veridicità di queste ipotesi

Quindi:

$x_1, \ldots, x_{N_1}$ campione casuale da $\mathcal{N}(\mu_1, \sigma)$

$y_1, \ldots, y_{N_2}$ campione casuale da $\mathcal{N}(\mu_2, \sigma)$

$x_1, \ldots, x_{N_1}$ e $y_1, \ldots, y_{N_2}$ indipendenti

$$E[\bar{x} - \bar{y}] = \mu_1 - \mu_2 \qquad Var[\bar{x} - \bar{y}] = \frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2} = \sigma^2 \left(\frac{1}{N_1} + \frac{1}{N_2} \right)$$

Stime per σ?

Stima della varianza per inferenze

Stima di σ^2 usando *entrambi* i campioni:

$$\sigma^2 \approx \frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}$$

$$= \frac{\sum_{i=1}^{N_1}(x_i-\bar{x})^2 + \sum_{i=1}^{N_2}(y_i-\bar{y})^2}{N_1+N_2-2} = S^2_{pooled}$$

Inferenze sulla media con piccoli campioni

La variabile

$$T = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S_{pooled}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

è distribuita come la t di Student con $N_1 + N_2 - 2$ gradi di libertà.

Intervallo di confidenza per $\mu_1 - \mu_2$:

$$(\bar{x} - \bar{y}) \pm t_0 S_{pooled}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

(t_0 valore di riferimento con $N_1 + N_2 - 2$ gradi di libertà)

Esempio. Test sull'alimentazione di 25 mucche da latte. Campione di 13 mucche alimentate con prodotto A. Le rimanenti 12 mucche alimentate con prodotto B. Per 3 settimane, per ogni mucca viene segnata la produzione giornaliera di latte. Il totale viene riportato in tabella.

A	44	44	56	46	47	38	58	53	49	35	46	30	41
B	35	47	55	29	40	39	32	41	42	57	51	39	

Verificare se il prod. A induce una maggiore produzione rispetto al prod. B

Sol. $N_1 = 13, \bar{x} = 45.15, s_1 = 7.998 \qquad N_2 = 12, \bar{x} = 42.25, s_1 = 8.740$

$$S_{pooled} = \sqrt{\frac{12(7.998)^2 + 11(8.740)^2}{23}} = 8.36 \quad \text{gradi di libertà: } N_1 + N_2 - 2 = 23$$

Test: $H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 > 0$

regione critica (una coda!): $t \geq t_0 = 1.714$ (5% liv.sign.)

$$T = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S_{pooled}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} = \frac{45.15 - 42.25}{8.36\sqrt{\frac{1}{13} + \frac{1}{12}}} \approx 0.87 \Rightarrow \text{Ipotesi } H_0 \text{ non rifiutata}$$

Grandi vs. piccoli campioni

- ⋆ Per grandi campioni no pooling
- ⋆ Per N_1, N_2 piccoli,

$$\frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

  non segue una t-distribuzione (e dipende da $\frac{\sigma_1}{\sigma_2}$)

- ⋆ Ipotesi $\sigma_1 = \sigma_2$ e uso S_{pooled} permettono l'uso della t-distribuzione per N_1, N_2 piccoli

Quand'è che $\sigma_1 = \sigma_2$ è un'ipotesi ragionevole?

$$\frac{1}{2} \leq \frac{s_1}{s_2} \leq 2$$

altrimenti (procedura prudente):

$$\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}} \quad \text{con} \quad \min(N_1-1, N_2-1) \quad \text{g.d.l.}$$

prudente: livello di confidenza è *almeno* $(1-\alpha)$

Confronto tra i due disegni di esperimenti

	Campioni indip $(N_1 = N_2 = N)$	N Coppie $d_i = x_i - y_i$
Errore Standard stimato	$S_{pooled}\sqrt{\frac{1}{N} + \frac{1}{N}}$	$\frac{s_D}{\sqrt{N}}$
g.d.l. in t	$2N - 2$	$N - 1$

Difetti nel confronto delle due procedure:

- Coppie: perde gradi di libertà (t_0 più grande)
- Camp. Indip.: risente di caratteristiche intrinseche del campione. maggiore variabilità

  (nella coppia, la differenza le elimina)

$\Rightarrow$ "Coppie" da preferire se porta a riduzione di variabilità (nota a priori). Se campioni tutti molto simili, allora "Campioni indipendenti".