Modulo I

—

January 20, 2021

Contents

1 matlab

Insiemistica

- $A \cap B = A \cdot B$
- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(AB)$
- $A \cap B = \emptyset \rightarrow disgiunti$
- $P(A \cup B) = P(A) + P(B) \rightarrow disgiunti$
- $P(A)$ e $P(B)$ sono indipendenti allora:

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

Probabilità condizionata

Esprime la probabilità che si verifichi A nell'ipotesi in cui sia verificato B

$$P(A|B) = \frac{P(AB)}{P(B)}$$
$$P(AB) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Probabilità totali

$$B_1, B_2, ..., B_\mathrm{k} \rightarrow \textit{disgiunti, partizioni di } \Omega$$
$$P(A) = P(A) \cdot P(B_1) + P(A) \cdot P(B_2) + ... + P(A) \cdot P(B_\mathrm{k})$$
$$P(A) = P(AB_1) \cdot P(AB_2) \cdot P(AB_\mathrm{k})$$

$$P(A) = P(A|B_1) \cdot P(B_1) + ... + P(A|B_\mathrm{k}) \cdot P(B_\mathrm{k})$$
$$\rightarrow P(A) = P(A|B) \cdot P(B) + ... + P(A|\bar{B}) \cdot P(\bar{B})$$

Bayes

Probabilità inverse, o a posteriori.

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(BA)}{P(A)} = \frac{P(A|B) \cdot P(B)}{P(A)}$$

NOTA BENE:

$$\rightarrow P(A|B) + P(\bar{A}|B) = 1$$

$$P(R|F) = 1 - P(B|F)$$
$$P(F) = 1 - P(M)$$
$$P(B|M) = 1 - P(R|M)$$

Esempio:

$$P(D) = 0.20$$
$$P(E|D) = 0.995$$
$$P(E|\bar{D}) = 0.001$$

a) Supponendo che la produzione sia di 1000 pezzi, qual'è la probabilità che vengano eliminati 2 pezzi non difettosi.

$$n = 1000$$

$$P(\bar{D}) = 1 - P(D)$$
$$P(E\bar{D}) = P(E|\bar{D}) \cdot P(\bar{D})$$
$$sol = binopdf(2, n, P(E\bar{D}))$$

(b) Si calcoli la probabilità che un pezzo che non sia stato eliminato al controllo di qualità sia difettoso.

$$P(E) = P(E|D) \cdot P(D) + P(E|\bar{D}) \cdot P(\bar{D})$$

$$P(\bar{E}) = 1 - P(E)$$

$$P(\bar{E}|D) = 1 - P(E|D)$$

$$P(D|\bar{E}) = \frac{P(\bar{E}|D) \cdot P(D)}{P(\bar{E})}$$

2 matlab

Insiemistica es

dadi =[
11 , 12 , 13 , 14 , 15 , 16 ;
21 , 22 , 23 , 24 , 25 , 26 ;
31 , 32 , 33 , 34 , 35 , 36 ;
41 , 42 , 43 , 44 , 45 , 46 ;
51 , 52 , 53 , 54 , 55 , 56 ;
61 , 62 , 63 , 64 , 65 , 66]

l'evento E: la somma dei numeri usciti pari → E = [dadi(1,1) dadi(1,2) .....];
l'evento F: Almeno uno dei due numeri è 1 → E = [dadi(1,1) dadi(1,2) .....];

- $E \cup F$ = numel(intersect(E, F)) / numel(dadi)
- $E \cap F$ = pE + pF - $(E \cup F)$
- $\bar{F}$ = nonF = setdiff(dadi, F)'

Estrazione palline urna

- estrazione delle palline da due urne indipendenti → A · B
- la probabilità totale = → A · B + C · D
- + → "oppure"
- · → "e"

3 Calcolo Combinatorio

- Il Fattoriale è un caso particolare di disposizioni in cui la dimensione del gruppo da formare (n) è pari alla dimensione totale del insieme (N)
- Disposizione $\rightarrow$ diverso per elemento e posizione dell'elemento (r = gruppi, n = oggetti)

$$D_{(\mathrm{n},\ \mathrm{r})} = \frac{n!}{(n-r)!}$$

- Combinazione $\rightarrow$ diverso per elemento

$$C_{(\mathrm{n},\ \mathrm{r})} = \frac{n!}{r! \cdot (n-r)!}$$

$$nchoosek(n, r)$$

- Con ripetizione / reinserimento $\rightarrow$ diverso per elemento e posizione dell'elemento

$$D*_{\mathrm{n},\ \mathrm{r}} = n^r$$

4 Binomiale

Contatore di successi, fornisce una percentuale.

$$p(x) = C_{(\mathrm{n},\ \mathrm{x})} \cdot p^{\mathrm{x}} \cdot (1-p)^{(n-x)}$$
$$E(X) = n \cdot p$$
$$Var(X) = n \cdot p \cdot (1-p)$$

Probabilità nel punto

$$binopdf(x, n, p)$$

- x = successi
- n = prove
- p = probabilità

Probabilità cumulata

$$binocdf(x, n, p)$$

- $P(X < a) = P(X <= a-1)$
- almeno 30, probabilità compresa la 30-esima $\rightarrow 1 - binopdf(29, n, p)$

Esempio:
Y = v.c. che descrive il numero di giocatori che effettua 1 solo canestro

$$P(Y) = binopdf(1, 10, 0.25) = 0.1877$$
$$E(Y) = n \cdot p = round(70 * 0.1877, 0)$$

5 Poisson

Distribuzione di eventi rari (probabilità molto bassa che accada qualcosa).

$$p(x) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$$

$$E(X) = \lambda$$
$$Var(X) = \lambda$$

In matlab:
X = n = variabile di conteggio
X(t) = k = nel tempo

$$Poissonpdf(n, lambda \cdot t)$$
$$Poissoncdf \rightarrow sulla\ variabile\ di\ conteggio$$

Se n grande e p piccolo:

$$Bin(n, p) \cong \mathbb{P}(\lambda = np)$$

Statistica modulo II

Silviu Filote

December 7, 2020

Contents

Variabili casuali continue

Una variabile casuale **VC** è continua se può assumere un qualunque valore in un intervallo continuo e la probabilità delle singole realizzazioni è pari a *0*.

- Ω è continuo [-∞, ∞]
- $P(X = x) = 0$
- $F(x) = P(X \leq x)$

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

La variabile casuale è continua se e solo se la funzione di ripartizione è continua

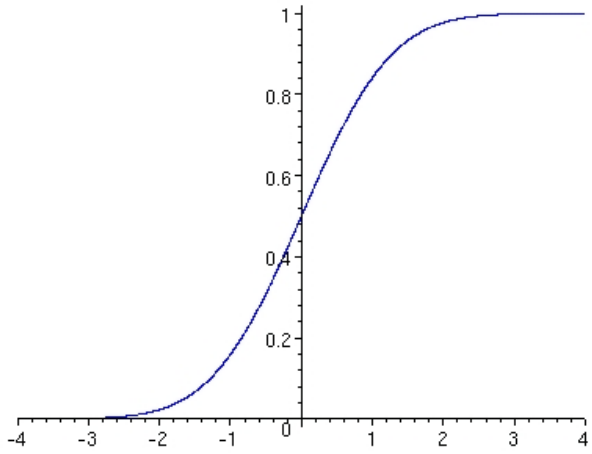


Figure 1: variabile casuale continua

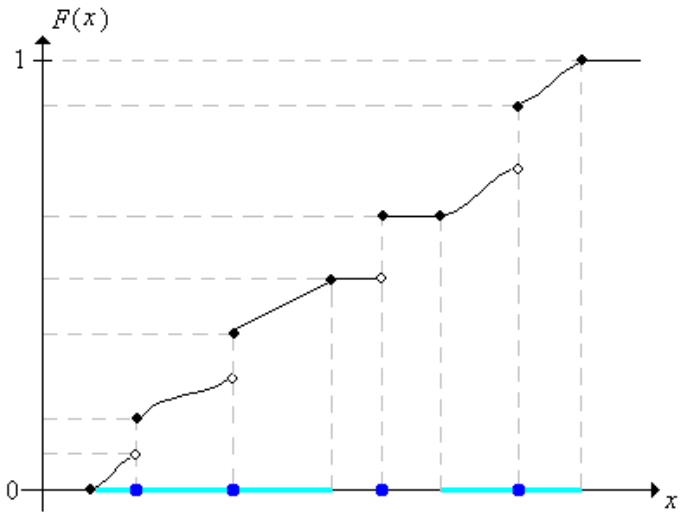
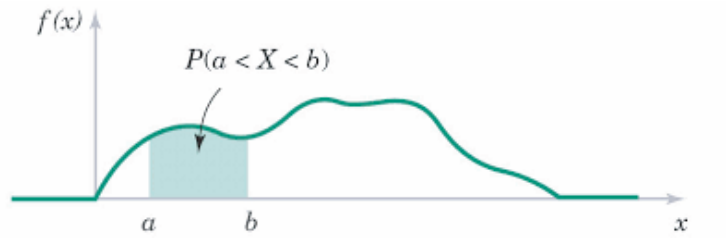


Figure 2: variabile casuale non continua

1 Densità di probabilità

Chiameremo funzione di densità la funzione matematica f(x) per cui l'area sottesa alla funzione, corrispondente ad un certo intervallo, è uguale alla probabilità che X assuma un valore in quell'intervallo



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

- La f(x) è sempre positiva
- L'area totale sottesa alla funzione è pari a 1, ossia:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

La f(X) non dà la probabilità di X, ma è proporzionale alla probabilità che X ricada in un intervallo infinitesimale centrato su X. Quindi se vogliamo intepretare questa densità non è una probabilità, ma mi dice come varia la probabilità vicino a x.

- quando la densità è alta: probabilità intorno a quel punto alta
- quando la densità è bassa: probabilità intorno a quel punto bassa

Lo stesso problema può essere risolto attrarveso l'utilizzo della funzione di ripartizione $F(x)$, esempio:

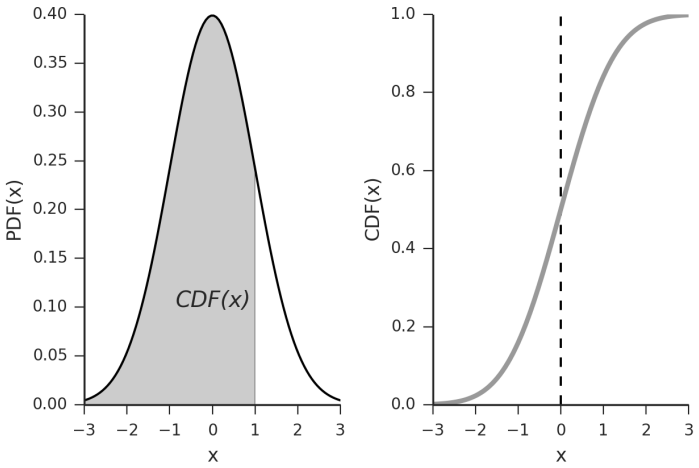


Figure 3: confronto

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

2 Quantili e percentili

Oltre alla mediana, che divide a metà un insieme di dati ordinati, vengono usati anche altri **indici di posizione che dividono le distribuzioni in determinate percentuali** detti quartili, quantili e percentili. Questi sono detti indici di posizione non centrale. Viene anche chiamata funzione del punto percentuale o funzione di distribuzione cumulativa inversa.

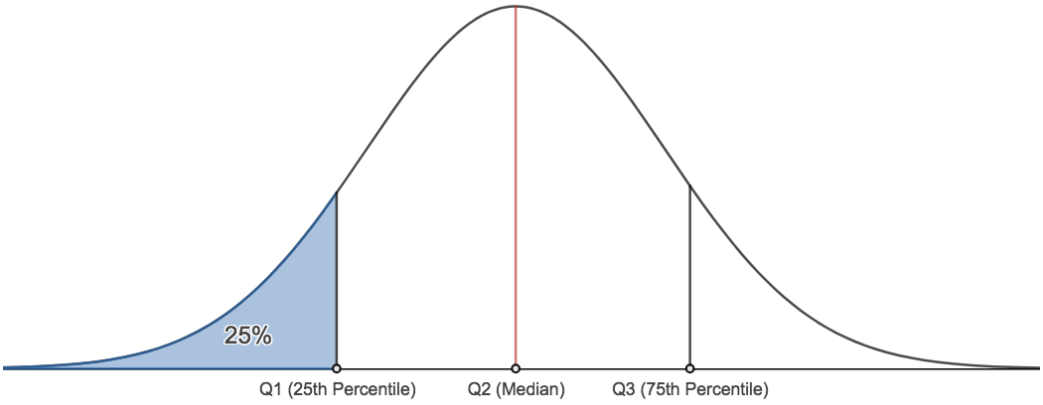


Figure 4: percentili

$\tilde{x}_{\mathrm{p}}$ tale che $F(\tilde{x_{\mathrm{p}}}) = p$

$\tilde{x}_{\mathrm{p}} = F^{-1}(p)$ se $f > 0$

$$\int_a^{med} f(x)dx = 0.50$$

3 Valore atteso e variaziona di una VCC

Il **valore atteso**, che viene chiamato anche media o speranza matematica della distribuzione di una variabile casuale, è un **indice di posizione**. Il valore atteso di una variabile casuale rappresenta il **valore previsto che si potrà ottenere in un gran numero di prove**. Con la locuzione "gran numero di prove" s'intende un numero sufficientemente grande di prove così che sia possibile prevedere, mediante la probabilità, le frequenze relative dei vari eventi.

Proprietà:

- $E(a + bX) = a + bE(X)$
- $E(aX + bY) = aE(X) + bE(Y)$
- $E(X - \mu) = 0$
- $E((aX)^2) = a^2E(X^2)$

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx = \mu$$

La varianza indica **quanto sono "dispersi" i valori della variabile aleatoria relativamente al suo valore medio**. Data una variabile casuale X qualsiasi sia E(X) il suo valore atteso. Consideriamo la variabile casuale X − E(X), i cui valori sono le **"distanze" tra i valori di X e il valore atteso E(X)**. La variabile casuale X − E(X) si chiama scarto, oppure deviazione oppure variabile casuale centrata.

Proprietà:

- $Var(X) \geq 0$
- $Var(a + bX) = b^2 Var(X)$

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

4 Momenti

Valore atteso elevato alla k-esima potenza

- Momenti dall'origine, $k \geq 0$ (se esistono finiti)

$$\mu_{\mathrm{k}} = E(X^{\mathrm{k}}) = \int_{-\infty}^{\infty} x^{\mathrm{k}} f(x) dx$$

- Momenti centrati

$$\bar{\mu}_{\mathrm{k}} = E((X - \mu)^{\mathrm{k}}) = \int_{-\infty}^{\infty} (x - \mu)^{\mathrm{k}} f(x) dx$$

Proprietà

- $\mu_{\mathrm{k}} < \infty$ allora μ_{r} con $r \leq k$ esiste finito

5 VCC Rettangolare R(0,1)

Densità costante:

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & altrimenti \end{cases}$$

CDF: funzione di ripartizione:

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 < x < 1 \\ 1 & x > 1 \end{cases}$$

Considerazioni:

- con $x = 0$ F(0) = P($x \leq 0$) = 0
- con $0 < x < 1$

  $$F(x) = \int_0^x f(x)dx = 0$$

  la funzione in x è sempre pari a 1 per cui

  $$F(x) = \int_0^x 1dx = x$$

  da cui ricaviamo che l'area può essere calcolata come ***base x altezza***
- con $x > 1$ F(x) = 1

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 x \cdot 1dx = \frac{1}{2}$$

$$Var(X) = E((x-\mu)^2) = E(x^2 - \mu^2) = E(x^2) - \mu^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Vcc generica R(a, b)

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

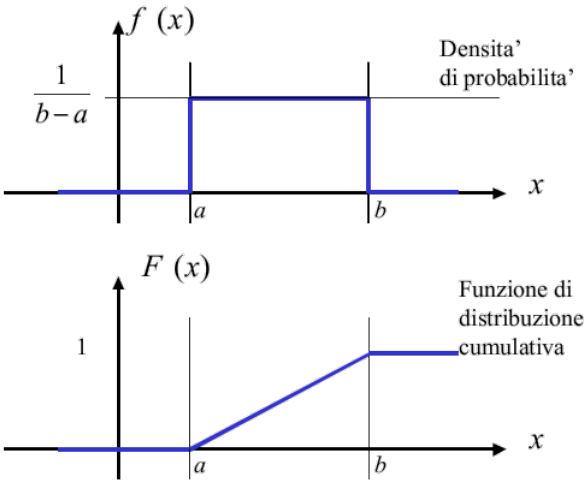


Figure 5: andamento delle funzioni

6 VCC Normale $N(\mu, \sigma^2)$

La distribuzione gaussiana è spesso usata come prima approssimazione per descrivere variabili casuali continue a valori reali che tendono a concentrarsi attorno a un singolo valor medio.
Il grafico della funzione di densità di probabilità associata è **simmetrico rispetto a** μ e ha una forma a campana, nota come campana di Gauss (o anche come curva degli errori, curva a campana, ogiva).

Una distribuzione di probabilità è **simmetrica** quando la sua **funzione di densità** di probabilità (nel caso continuo) siano simmetriche rispetto ad un particolare valore x_0 :

$$f(x_0 + x) = f(x_0 - x)$$

- Densità $X \equiv N(\mu, \sigma^2)$

$$f_{\mu,\,\sigma^2}(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot (\frac{x-\mu}{\sigma})^2} = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})$$

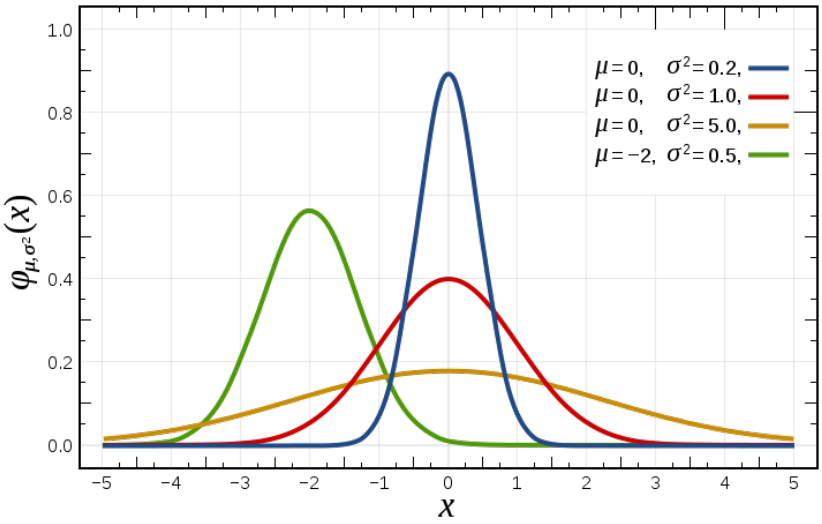


Figure 6: distribuzione della normale

- Ripartizione di X indicata anche come Φ

$$F_{\mu,\,\sigma^2}(x) = \Phi(\frac{x-\mu}{\sigma})$$

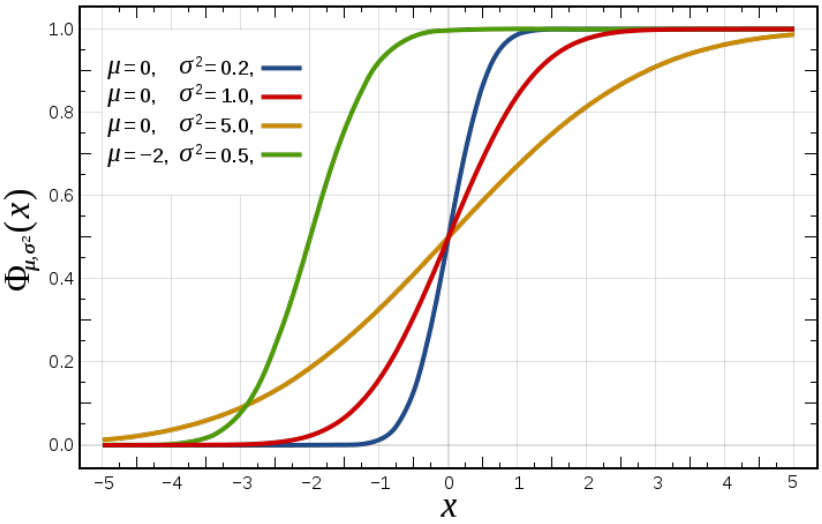


Figure 7: CDF della normale

Momenti

$$E(X) = \mu$$
$$Var(X) = \sigma^2$$
$$\overline{\mu}_3 = E((X-\mu)^3) = 0$$
$$\overline{\mu}_4 = E((X-\mu)^4) = 3\sigma^4$$

Unità di misura della gaussiana

L'unità di misura della gaussiana $N(\mu, \sigma^2)$ è σ.

σ, scarto quadratico medio o deviazione standard, esprimere la dispersione dei dati intorno ad un indice di posizione in questo caso μ.
Ha pertanto la stessa unità di misura dei valori osservati (al contrario della varianza che ha come unità di misura il quadrato dell'unità di misura dei valori di riferimento).

$$Var(X) = \sigma^2$$
$$\sigma = \sqrt{\sigma^2}$$

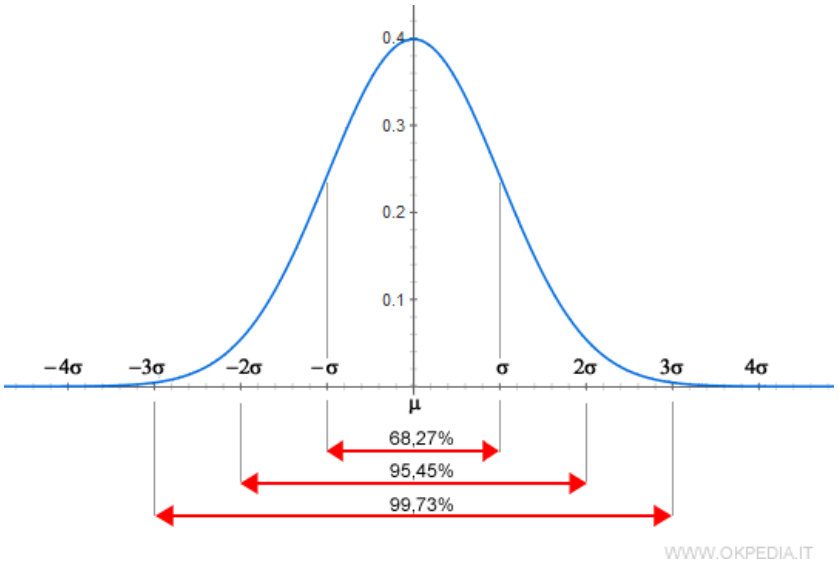


Figure 8: sigma

$$P(\mu - 1\sigma < X < \mu + 1\sigma) = P(|X - \mu| < \sigma) = 0.683$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = P(|X - \mu| < 2\sigma) = 0.95$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = P(|X - \mu| < 3\sigma) = 0.997$$

Indici di forma

Un **indice di asimmetria** di una distribuzione è un valore che cerca di fornire una misura della sua mancanza di simmetria.

Le distribuzioni simmetrice presentano le seguenti proprietà:

- il valore atteso, la mediana e la moda coincidono;
- i momenti centrati di ordine dispari sono nulli, perchè tengono conto del segno negativo ed essendo simmetriche tra loro si annullano

$$P(X - \mu < a) = P(Z - \mu > -a)$$

$$\Phi(\frac{x-\mu}{\sigma}) = 1 - \Phi(-\frac{x-\mu}{\sigma})$$

$$Sk = E((\frac{X-\mu}{\sigma})^3) = \frac{E((X-\mu)^3)}{\sigma^3} = \frac{\int_{-\infty}^{\infty}(x-\mu)^3 \cdot f(x)dx}{\sigma^3} = 0$$

sk = indice di asimmetria

L'indice di curtosi fa riferimento alla maggiore o minore gibbosità di una curva in prossimità delle code.

$$k = E((\frac{X-\mu}{\sigma})^4) = \frac{E((X-\mu)^4)}{\sigma^4} = 3$$

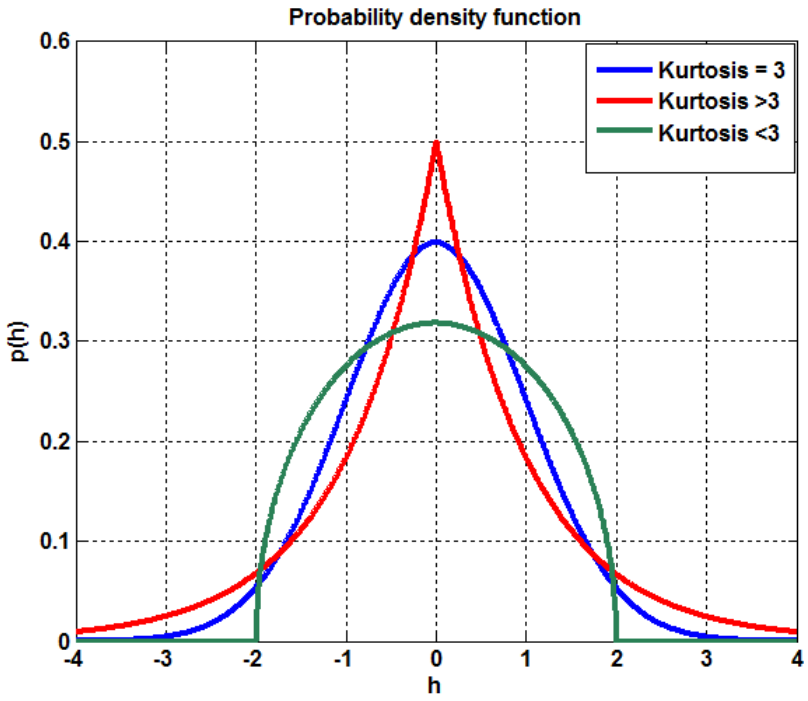


Figure 9: kurtosis

- con k = 3 è la normale
- con k < 3 distribuzione leptocurtica
- con k > 3 distribuzione platicurtica

7 VCC normale standard Z ≡ N(0,1)

Densità di Z (la densità della normale la indichiamo con Z)

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

Ripartizione di Z

$$\Phi(x) = \int_{-\infty}^{x} \phi(x)dx$$

Non esiste una primitiva, ossia una formula, per cui viene utilizzata un'approssimazione numerica. Normcdf (matlab).

Perché non siamo in grado di calcolare l'integrale

Momenti di Z (ricordandosi che $\mu = 0$)

$$E(Z) = 0$$
$$Var(Z) = E(Z^2) = 1$$
$$sk(Z) = E(Z^3) = 0$$
$$k = E(Z^4) = 3$$

Problema diretto per Z: **aree = probabilità**

$$P(a < x < b) = \int_{a}^{b} \phi(x)dx = \Phi(b) - \Phi(a)$$

Problema inverso per Z: **quantili** (percentili):

Quale punto del dominio presenta quella percentuale

$$z_{\mathrm{a}} = \Phi^{-1}(1-a) = \tilde{z}_{1\text{ - a}}$$
$$P(?) = \tilde{z}_{1\text{ - a}}$$

Nb: senza tilde viene chiama **valore critico e corrisponde alla coda destra** ossia all'area dal punto verso ∞, mentre con la tilde da -∞ al punto

Standardizzazione di X, con X ≡ N(μ,σ^2)

Riconduce una variabile aleatoria distribuita secondo una media μ e varianza σ^2, ad una variabile aleatoria con distribuzione "standard", ossia di media zero e varianza pari a 1.

$$Z = \frac{X - \mu}{\sigma} = N(0,1)$$

Riscalatura di Z, con Z(0,1)

$$X = \mu + \sigma \cdot Z \equiv N(\mu, \sigma^2)$$

Problema diretto per X: aree = probabilità

$$P(a < x < b) = P(\frac{a-\mu}{\sigma} < x < \frac{b-\mu}{\sigma}) = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \phi(x)dx =$$

$$= \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

Problema inverso per X: quantili (percentili)

$$x_{\mathrm{a}} = \mu + \sigma \cdot z_{\mathrm{a}} = \mu + \sigma \cdot \Phi^{-1}(1-a)$$

$$= \mu + \sigma \cdot \tilde{z_{\mathrm{a}}} = \tilde{x}_{1\ -\ \mathrm{a}}$$

8 VCC esponenziale negativa $e^{-\lambda}$

La variabile casuale esponenziale negativa è utilizzata per descrivere i tempi di attesa affinchè si verifichi un evento: un utente richieda un servizio...

Distribuzione e ripartizione(con x > 0, $\lambda > 0$)

$$f_\lambda = \lambda \cdot e^{-\lambda \cdot x}$$

$$F_\lambda = 1 - e^{-\lambda \cdot x}$$

Momenti

$$E_\lambda(X) = \frac{1}{\lambda}$$

$$Var_\lambda(X) = \frac{1}{\lambda^2}$$

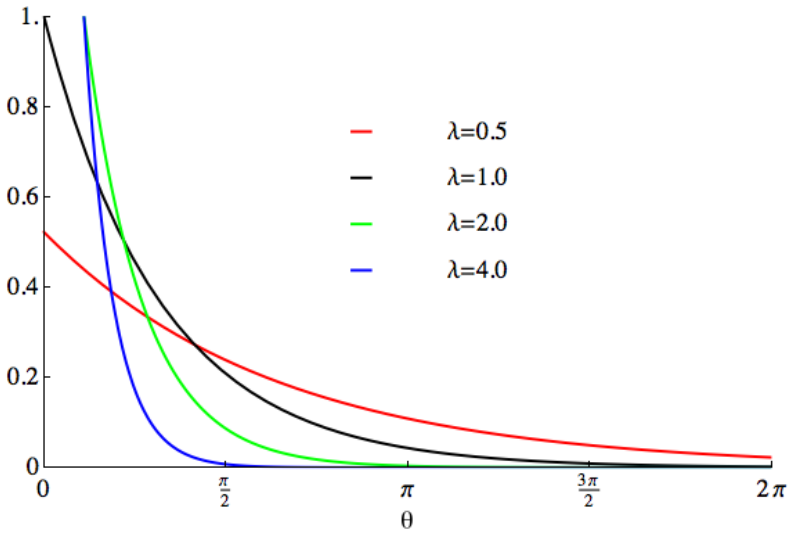


Figure 10: distribuzione esponenziale negativa

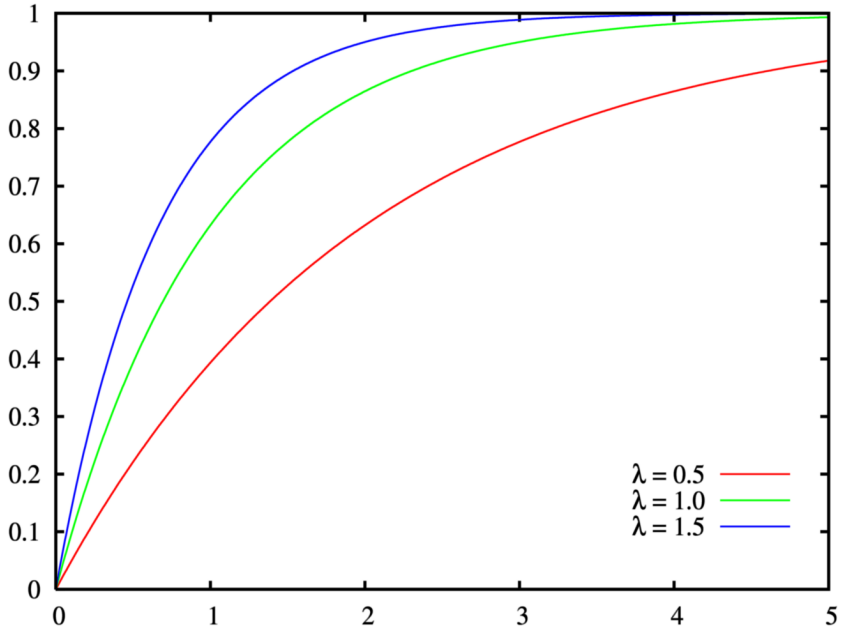


Figure 11: CDF

9 VCC gamma $\Gamma(r, \lambda)$

Comprende, come casi particolari, anche le distribuzioni esponenziale e chi quadrato per determinati valori di r, λ.

Distribuzione

- con $x > 0$
- con $\lambda > 0$
- con $r = 1, 2, ...$
- $P(X > 0) = 1$
- generalizza l'esponenziale:

$$f(x; r, \lambda) = \frac{\lambda}{\Gamma(r)} \cdot (\lambda \cdot x)^{\mathrm{r}\,-\,1} \cdot e^{-\lambda x}$$

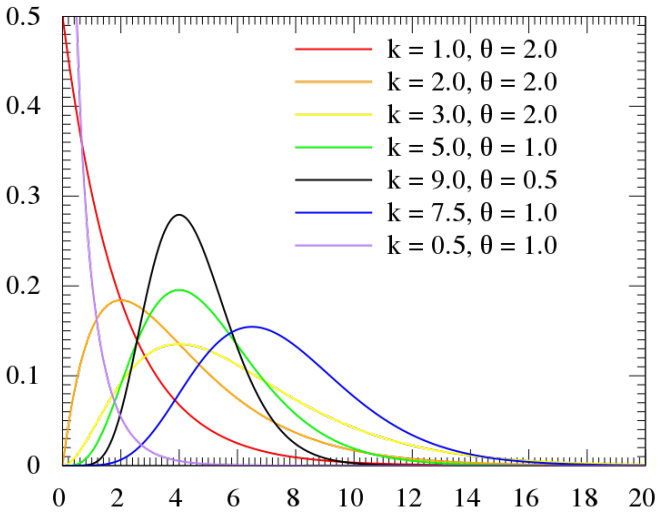


Figure 12: distribuzione gamma

Momenti

$$E(X) = \frac{r}{\lambda}$$
$$Var(X) = \frac{r}{\lambda^2}$$

Si richiama la funzione gamma completa

$$\Gamma(a) = \int_0^\infty x^{\text{a - 1}} \cdot e^{\text{-x}} dx$$

Mentre la funzione di gamma incompleta

$$G(t, a) = \int_0^t x^{\text{a - 1}} \cdot e^{\text{-x}} dx$$
$$G(\infty, a) = \Gamma(a)$$

Proprietà

- $\Gamma(1) = 1$
- $\Gamma(a) = (a-1)\Gamma(a-1), \quad a > 1$
- $\Gamma(n) = (n-1)!, \quad n\ intero$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
- $\Gamma(1, \lambda) = e^{-\lambda} \qquad E(X) = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2}$
- $\Gamma(\frac{n}{2}, \frac{1}{2}) = \chi^2(n)$

10 Teorema limite centrale - Somma

Il teorema limite centrale afferma che la distribuzione della somma di un numero elevato di variabili casuali **indipendenti e identicamente distribuite** (**iid**) tende a distribuirsi **normalmente**, indipendentemente dalla distribuzione delle singole variabili.

Se:

$$X_1, ..., X_{\mathrm{n}} \quad con \quad (\mu, \sigma^2) \quad finiti$$

allora, per $\mathrm{n} \to \infty$

$$T_{\mathrm{n}} = X_1 + ... + X_{\mathrm{n}} \cong N(n\mu, n\sigma^2)$$

indipendentemente da F

Nella teoria della probabilità, una sequenza di variabili casuali è detta indipendente e identicamente distribuita (i.i.d.) se:

- le variabili hanno tutte la stessa distribuzione di probabilità;
- le variabili sono tutte statisticamente indipendenti.
- varianze sono tutte uguali
- valori attesi sono tutti uguali

11 VCC campionarie

Si definisce campione aleatorio un insieme di numeri aleatori X1,X2,...,Xn indipendenti e con la stessa distribuzione f.
In altre parole, ogni numero aleatorio, non è altro che un campione di ampiezza n estratto da una popolazione, a cui è associato una densità di probabilità f. Per ogni campione Xi, quindi, possiamo calcolare una data statistica, come la media o la varianza, che differisce dalle statistiche degli altri campioni.

Campione casuale da F

$$X_1, ..., X_\mathrm{n} \quad iid \quad F$$
$$E(Xi) = \mu$$
$$Var(Xi) = \sigma^2$$

Media campionaria

$$\bar{X} = \frac{1}{n}\sum_j X_\mathrm{j} = \frac{X_1 + ... + X_\mathrm{n}}{n}$$

Teorema delle 3M: La media della media è la media

$$E(\bar{X}) = E(\frac{X_1 + ... + X_\mathrm{n}}{n}) = \frac{1}{n} \cdot (\mu + ... + \mu_\mathrm{n}) = \mu$$
$$E(\bar{X}) = \mu$$

Varianza di $\bar{X}$

$$Var(\bar{X}) = Var(\frac{X_1 + ... + X_\mathrm{n}}{n}) = \frac{1}{n^2} \cdot n \cdot \sigma^2$$
$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Distribuzione di $\bar{X}$

Se presi n campioni dalla normale

$$X_1, ..., X_{\mathrm{n}} \quad iid \quad N(\mu, \sigma^2)$$

allora, per ogni $n \geq 1$

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

$$\bar{X} - \mu \equiv N(0, \frac{\sigma^2}{n})$$

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \equiv N(0, 1)$$

NB:

- $\frac{\sigma^2}{n} = \frac{\sigma}{\sqrt{n}}$ l'errore standard della media
- è stata effettuata una Standardizzazione per avere un N
- $\bar{X}$ è il valore della media dei campioni ed ha una sua distribuzione
- da compioni normali la $\bar{X}$ è una normale gerica non standardizzata

12 Teorema limite centrale - media

Se presi n campioni iid, indipendenetemente dalla distribuzione

$$X_1, ...X_{\mathrm{n}} \quad sono \quad iid \quad (\mu, \sigma^2)$$

allora, per n $\to \infty$

$$\bar{X}_{\mathrm{n}} \cong N(\mu, \frac{\sigma^2}{n})$$

Esempio:

- $Bin(n, p) \cong N(np, np \cdot (1 - p))$

* viene approssimata in questo modo per p intermedia tra 1 e 0, conferma il teorema del limite centrale (somma)

$$y = B(n, p)$$

$$E(X) = p$$

$$Var(X) = p(1 - p)$$

$$y = \sum_{i=1}^{n} xi \cong N(np, np(1 - p)), \quad xi \ \ sono \ \ iid \quad B(1, P)$$

13 Legge dei grandi numeri

Grazie alla legge dei grandi numeri, possiamo fidarci che la media sperimentale, che calcoliamo a partire da un numero sufficiente di campioni, sia sufficientemente vicina alla media vera, ovvero quella calcolabile teoricamente.

Se $X_1, ..., X_{\mathrm{n}}$ iid μ

$$P(\lim_{n\to\infty} \bar{X}_{\mathrm{n}} = \mu) = 1$$

inoltre se $\sigma^2 < \infty$

$$E((\bar{X}_{\mathrm{n}} - \mu)^2) = \frac{\sigma^2}{n} \to 0$$

14 Varianza campionaria

con $X_1, ..., X_n \quad iid \quad (\mu, \sigma^2)$

$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})^2$$

$$\hat{\theta}^2 = S^2 \cdot \frac{n-1}{n} = \frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X})^2$$

Media di S^2

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

Considerazioni

- varianza popolazione $\hat{\theta}^2$
- μ viene sostituito da $\bar{X}$, questo perchè μ è sconosciuto e deve essere quindi stimato, lo stimatore naturale di μ è $\bar{X}$
- nel calcolo della media viene usato il divisore $n-1$ invece di n, questo perchè stimando μ con $\bar{X}$, viene introdotta una piccola distorsione in $(X_i - \bar{X})^2$.
  Infatti:
  $$E[(X_i - \bar{X})^2] = \frac{n-1}{n} \cdot \sigma^2$$
  da cui:
  $$E[\sum_{i=1}^{n} (Y_i - \bar{Y})^2] = n \cdot E \cdot [(X_i - \bar{X})^2] = (n-1) \cdot \sigma^2$$

  Dividendo per n-1 nella formula della varianza campionaria abbiamo corretto i gradi di libertà: la stima della media causa la perdità di 1 grado di libertà nei dati, cosicchè, ne rimangono solo n-1..

  I gradi di libertà di una variabile aleatoria o di una statistica in genere esprimono il numero minimo di dati sufficienti a valutare la quantità d'informazione contenuta nella statistica. Infatti, quando un dato non è indipendente, l'informazione che esso fornisce è già contenuta implicitamente negli altri($\bar{X}$ ***è un dato calcolato dai campioni per cui dipendente*** $\Rightarrow$ ***n - 1***).

Teorema delle 3M:
La varianza dei n campioni è la varianza campionaria e il valore atteso della varianza campionaria è la varianza

$$E(S^2) = \sigma^2$$
$$E(\hat{\theta}^2) = \frac{n}{n-1}\sigma^2$$

15 Distribuzione chi quadrato: χ^2_n

Sia data una popolazione normale avente varianza σ^2 e da essa si estraggono campioni casuali di ampiezza n; allora, la variabile

$$Z_1, ..., Z_n \ \ iid \ \ N(0,1) \Rightarrow \sum_{i=1}^{n} Z^2_i \equiv \chi^2_n$$

$$S^2_n \equiv \frac{\sigma^2}{n-1} \cdot \chi^2_{n-1}$$

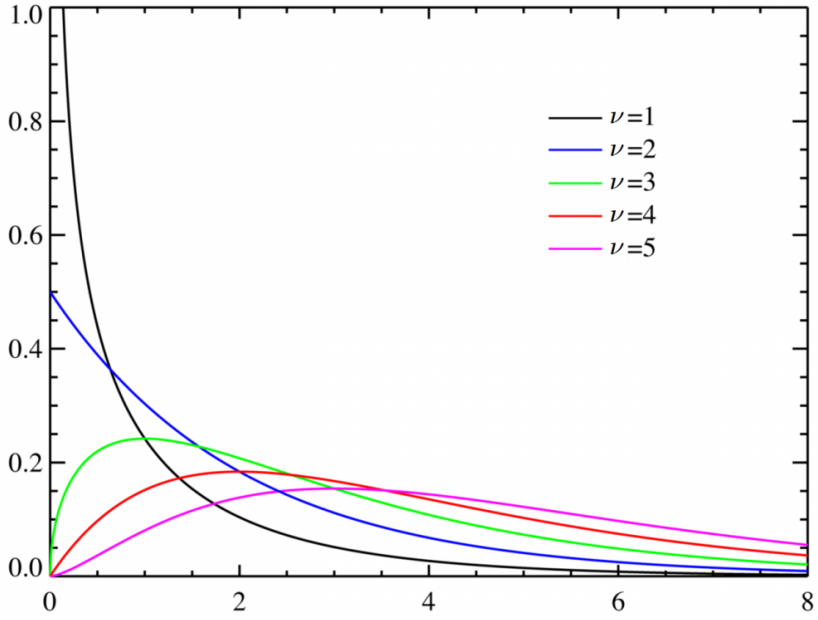


Figure 13: chi distribuzione

Proprietà

- Asimmetrica: $P(\chi^2_n > 0) = 1$ f(x) è asimmetrica
- il parametro n è detto gradi di libertà: $\chi^2_n = n - 1$
- Gamma: $\chi^2_n \equiv \Gamma(\frac{n}{2}, \frac{1}{2})$
- Momenti: $E(\chi^2_n) = n \qquad Var(\chi^2_n) = 2n$
- Normalità asintotica: per $n \to \infty \qquad \chi^2_n \cong N(n, 2n)$

16 Distribuzione t di student: t_n

Motivazioni

$$t = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{\mathrm{n}}}{n}}} \equiv tn$$

se numeratore e denominatore sono indipendeti

Invece:
Data una popolazione normale avente media μ da cui si estraggono campioni casuali ampiezza n, indicando con $\bar{X}$ la media campionaria e con S^2 la varianza campionaria

$Se \;\; X_1, ...X_{\mathrm{n}} \;\; sono \;\; iid \;\; (\mu, \sigma^2)$

$$t = \frac{\bar{X} - \mu}{\sqrt{(\frac{S^2}{n})}} \equiv t_{\mathrm{n}\text{ - }1}$$

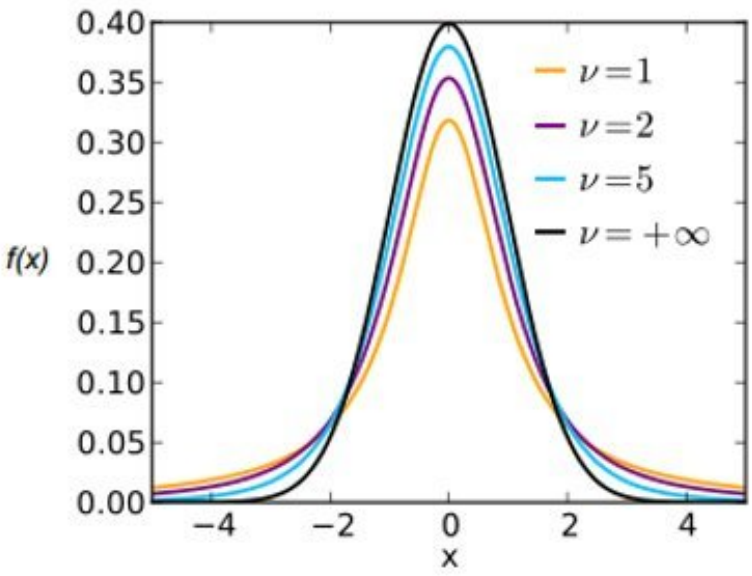


Figure 14: distribuzione t student, v = gradi di libertà

Proprietà

$$f_{\mathrm{n}}(t) = k_{\mathrm{n}}(1 + \frac{t^2}{n})^{-\frac{n+1}{2}}$$

- $f_{\mathrm{n}}(t)$ è simmetrica attorno alla media $\mu = 0$ e campanulare
- Momenti: $\mu_{\mathrm{k,n}} = E(t^{\mathrm{k}}{}_{\mathrm{n}})$ esiste finito solo per $k < n$
- $E(t_{\mathrm{n}}) = 0$ per $n > 1$
- $V(t_{\mathrm{n}}) = \frac{n}{n-2}$ per $n > 2$
- $V(t_{\mathrm{n}}) \to 1$ per $n \to \infty$
- $K_{\mathrm{n}} = E(\frac{t^4{}_{\mathrm{n}}}{\sigma \cdot (t_{\mathrm{n}})^4}) = 3 \cdot \frac{n-2}{n-4}$ per $n > 4$
- la t di student ha $K_{\mathrm{n}} > 3$ quanto più si abbassa il grado di libertà
- $K_{\mathrm{n}} \to 3$ per $n \to \infty$
- $t_{\mathrm{n}} \to N(0,1)$ per $n \to \infty$
- $\frac{S^2}{n} = \frac{S}{\sqrt{n}}$ errore standard
- come la distribuzione chi quadro, anche la distribuzione t di Student sono distribuzioni dipendenti con grado di libertà n - 1

Studentizzazione

$$t_{\mathrm{n}\text{ - }1} = \frac{\bar{X} - \mu}{\sqrt{(\frac{S^2}{n})}}$$

$$t_{\mathrm{n}\text{ - }1} = \frac{\bar{X} - \mu}{(\frac{S}{\sqrt{n}})}$$

Standardizzazione

$$N(0,1) \equiv \frac{\bar{X} - \mu}{(\frac{\sigma}{\sqrt{n}})}$$

17 Inferenza statistica

L'inferenza statistica è un insieme di metodi con cui si cerca di trarre una conclusione sulla popolazione in base ad informazioni ricavate da un campione.

- Solo eccezionalmente conosciamo direttamente le caratteristiche della popolazione, di solito dobbiamo stimarle a partire dalle caratteristiche dei campioni che sono stati estratti dalla popolazione.
- L'inferenza statistica mira alla verifica di un'ipotesi relativa alle caratteristiche della popolazione.
- Il procedimento prevede dapprima la formulazione dell'ipotesi e quindi la valutazione della probabilità di ottenere quei risultati nella popolazione se l'ipotesi di partenza fosse vera.
- **Popolazione è identificata da una serie di statistiche $\rightarrow$ parametri che sono θ**
- **θ parametri rimarranno ignoti**

Fasi

- Estrazione di una parte della popolazione
- Calcolo delle statistiche campionarie
- Stima dei parametri nella popolazione in base ai risultati forniti dal campione (inferenza)

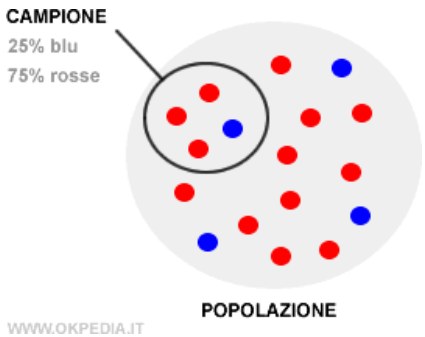


Figure 15: campionamento

18 Stima

Popolazione: X grandezza di interesse con distribuzione $f_\theta(x)$, θ parametro ignoto.

$$\theta = \theta \cdot (f)$$

Campione casuale semplice (ossia con reiseriemento ed equiprobabilità) da X(oppure da f, oppure da F)

$$X_1, ..., X_\mathrm{n} \quad iid \quad f_\theta(x)$$

Stima di θ:

$$\theta = \theta(X_1, ..., X_\mathrm{n})$$

$\Rightarrow \hat{\theta}$ è una particaolare V.C. detta statistica
$\Rightarrow$ Incerteza sull'errore di stima

$$\hat{\theta} - \theta$$

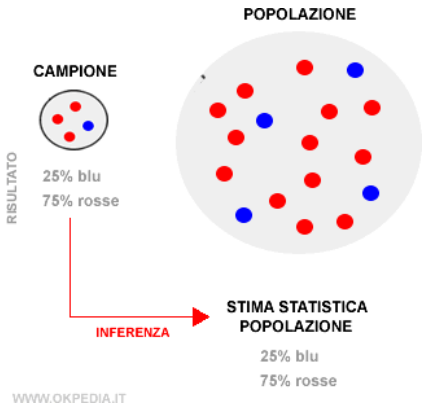


Figure 16: stima

19 Principio del campionamento ripetuto

Si valutano le proprietà di $\hat{\theta}$ nell'ipotesi di ripetere il processo di campionamento un gran numero di volte. Sono rilenvati in quest'ottica l'interpretazione frequentista della probabilità, ***la legge dei grandi numeri ed il metodo Monte Carlo.***

Esempio: sono interessato alla distribuzione di $F(\hat{\theta})$ ma non la conosciamo.

Effettuo m campionamenti

$$X_{1,1}, ..., X_{1,\mathrm{n}} \to \hat{\theta}_1$$

$$X_{2,1}, ..., X_{2,\mathrm{n}} \to \hat{\theta}_2$$

$$X_{\mathrm{m},1}, ..., X_{\mathrm{m,n}} \to \hat{\theta}_{\mathrm{m}}$$

$$E(\hat{\theta}) = \theta = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta} i \quad con \quad m \to \infty$$

20 Stima della media

- $X_1, ..., X_n$ con iid F
- $E(X) = \mu$
- $Var(X) = \sigma^2$

La media campionaria è una stima di μ:

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} Xi$$

Proprietà:

- Teorema delle 3M: $E(\bar{X}) = \mu$
- Varianza della media campionaria $Var(\bar{X}) = \frac{\sigma^2}{n}$
- Distribuzione di $\bar{X} = N(\mu, \frac{\sigma^2}{n})$
- Se $F(t) = \Phi(\frac{\bar{X}-\mu}{\sigma})$ questa distribuzione vale per ogni n (in generale valore per $n \to \infty$ teorema limite centrale)

21 Stima della varianza

- $X_1, ..., X_{\mathrm{n}}$ con iid F
- $E(X) = \mu$
- $Var(X) = \sigma^2$

La varianza campionaria è una stima di σ^2:

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (Xi - \bar{X})^2$$

Proprietà:

- Distribuzione χ^2 con n - 1 gradi di libertà: se $X_1, ..., X_{\mathrm{n}}$ iid $N(\mu, \sigma^2)$ allora:
$$S^2 \cdot \frac{n-1}{\sigma^2} = \chi^2{}_{\mathrm{n\ -\ 1}}$$
- Usando le proprietà del χ^2 si ottiene facilmente che :
  - $E(S^2) = \sigma^2$
  - $Var(S^2) \cong \frac{2 \cdot \sigma^4}{n-1}$

$$S^2 = \chi^2{}_{\mathrm{n\ -\ 1}} \cdot \frac{\sigma^2}{n-1}$$

$$V(S^2) = V(\chi^2{}_{\mathrm{n\ -\ 1}} \cdot \frac{\sigma^2}{n-1})$$

$$V(S^2) = (\frac{\sigma^2}{n-1}))^2 \cdot V(\chi^2{}_{\mathrm{n\ -\ 1}})$$

$$V(S^2) = \frac{\sigma^4}{(n-1)^2} \cdot 2(n-1)$$

$$Var(S^2) \cong \frac{2 \cdot \sigma^4}{n-1}$$

22 Stima di una percetuale

1. Binomiale
Schema di un campionamento: *n estrazioni con reinserimento*, da un'urna binaria con composizione

$$\pi = \frac{\#(A)}{N}$$

dove π esprime una % di popolazione che presenta un certo carattere (π è incognita).

All'iesima estrazione si pone

$$Xi = \begin{cases} 1 & \textit{se evento } A \\ 0 & \textit{se evento } \bar{A} \end{cases}$$

da cui

$$X_1, ..., X_{\mathrm{n}} \ \textit{iid Bin(1,}\pi\textit{)}$$

allora, il numero di eventi favorevoli *#(A)* nel campione

$$S = X_1 + .. + X_{\mathrm{n}} \quad \textit{è} \quad \textit{Bin(n,}\pi\textit{)}$$

se $\boldsymbol{n > 120}$ *prove* lo stimatore migliore per $\hat{\pi}$ è $\bar{X}$

$$\hat{\pi} = \bar{X} = \frac{S}{n}$$

con

$$E(\hat{\pi}) = \pi$$

$$Var(\hat{\pi}) = \frac{\pi \cdot (1 - \pi)}{n}$$

Asintoticamente

$$\bar{X} \to N(\pi, \frac{\pi \cdot (1 - \pi)}{n})$$

2. Ipergeometrica
Schema di campionamento *n estrazioni senza reinserimento*

Allora (*n = estrazioni, N = casi possibili, Nπ = casi favorevoli*)

$$S = X_1 + .. + X_\mathrm{n} \quad è \quad IG(n,\ N,\ N\pi)$$

e

$$(\hat{\pi}) = \frac{S}{n}$$

è stima di π :

$$E(\hat{\pi}) = \pi$$

$$Var(\hat{\pi}) = (\frac{\pi \cdot (1-\pi)}{n}) \cdot (1 - \frac{n-1}{N-1})$$

23 Teoria generale della stima

Considerando un campione

$$X_1, ..., X_n \ \ con \ \ iid \ \ f_\theta(x)$$

ed uno stimatore

$$\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$$

dove per **stimatore** si intede:
In statistica uno stimatore (puntuale) è una funzione che associa ad ogni possibile campione un valore del parametro da stimare. Il valore assunto dallo stimatore in corrispondenza a un particolare campione è detto stima.

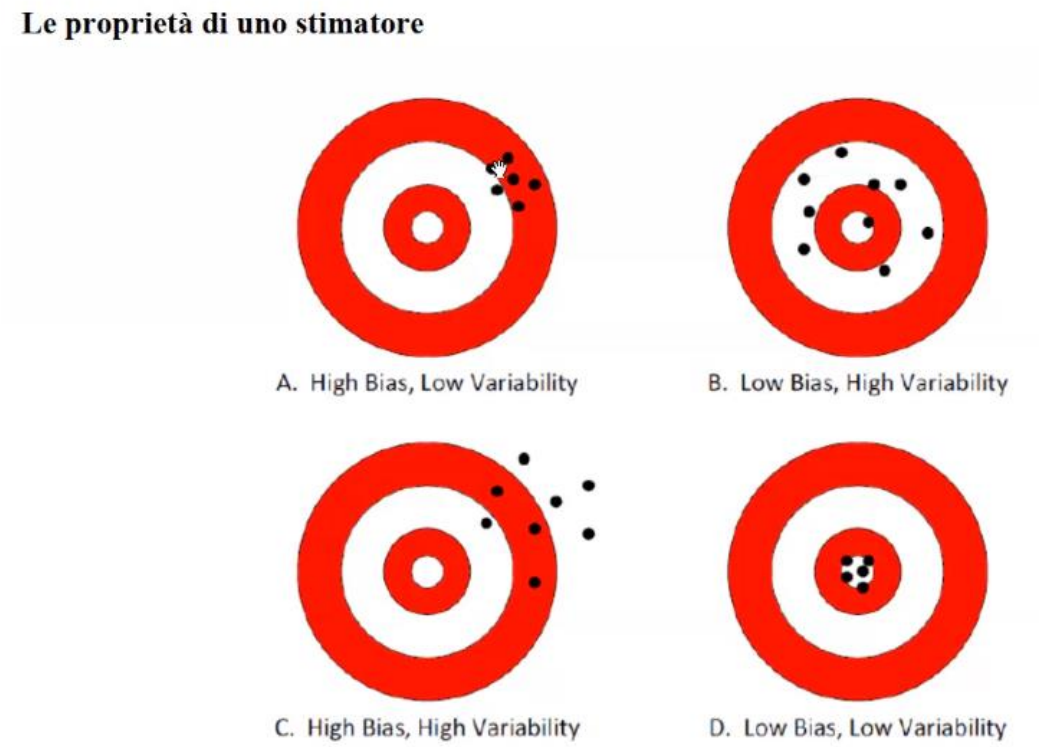


Figure 17: stima

Correttezza o non distorsione

$$E_\theta(\hat{\theta_{\mathrm{n}}}) = \theta$$

Correttezza asintotica

Non si richiede che lo stimatore $\hat{\theta}_{\mathrm{n}}$ basato sul campione $X_1, ..., X_{\mathrm{n}}$ abbia valore atteso uguale a θ, ma ci si accontenta che il suo valore atteso tenda a θ all'aumentare della dimensione del campione

$$\lim_{n\to\infty} E_\theta(\hat{\theta_{\mathrm{n}}}) = \theta$$

Bias o distorsione

Uno stimatore distorto è uno stimatore che per qualche ragione ha valore atteso diverso dalla quantità che stima; uno stimatore non distorto è detto stimatore corretto.

Si supponga di voler simare il parametro θ tramite uno stimatore $\hat{\theta}$

$$b(\hat{\theta}) = E(\hat{\theta_{\mathrm{n}}}) - \theta = 0$$

Errore quadratico medio

indica la discrepanza quadratica media fra i valori dei dati osservati ed i valori dei dati stimati.

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + b(\hat{\theta})^2$$

Efficienza

dati due stimatori $\hat{\theta}_{\mathrm{a}}$ e $\hat{\theta}_{\mathrm{b}}$ l'Efficienza è così definita (n stimatori MSE minore):

$$e(A, B) = \frac{MSE(\hat{\theta}_{\mathrm{a}})}{MSE(\hat{\theta}_{\mathrm{a}})}$$

$$e(A, B) \begin{cases} > 1 & \hat{\theta}_{\mathrm{a}} \textit{ è più efficiente} \\ = 1 & \hat{\theta}_{\mathrm{a}} \textit{ e } \hat{\theta}_{\mathrm{b}} \textit{ sono equivalenti} \\ < 1 & \hat{\theta}_{\mathrm{a}} \textit{ è meno efficiente} \end{cases}$$

Consistenza

Si dice che $\hat{\theta}$ è una stima consistente se "l'incertezza di θ scompare per $n \to \infty$" cioè se:

$$\hat{\theta} \to \theta \quad per\ n \to \infty$$

ossia che la probabilità che $\forall \varepsilon > 0$ valga 0, ossia che lo scostamento da θ sia paria a 0 e $\hat{\theta}$ quindi con $\forall \varepsilon = 0$:

$$P(|\hat{\theta} - \theta| \geq \varepsilon) \to 0$$

Condizione sufficiente per la consistenza

$$\lim_{n\to\infty} E_\theta(\hat{\theta_{\mathrm{n}}}) = \theta$$

$$\lim_{n\to\infty} Var_\theta(\hat{\theta_{\mathrm{n}}}) = 0$$

Corollario

$$\lim_{n\to\infty} MSE(\hat{\theta_{\mathrm{n}}}) = 0$$

$$\lim_{n\to\infty} E(\hat{\theta} - \theta)^2 = \lim_{n\to\infty} (Var(\hat{\theta}) + b(\hat{\theta})^2) = 0$$

24 Considerazioni per modulo II

Proprietà VCC

- una funzione è sempre positiva nel dominio definito:

$$f(x) \geq 0$$

- la densità di una funzione definita in un determinato dominio

$$\int_{-\inf}^{+\infty} f(x) = 1$$

- valore atteso di una VCC:

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx = \mu$$

- varianza di una VCC:

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

- probabilità delle singole applicazioni

$$P(X = 26) = \int_{26}^{26} f(x) = 0$$

- combinazione lineare x = una normale

$$y = \pm a \pm b \cdot x$$

$$E(y) = E(\pm a \pm b \cdot x) = \pm a \pm b \cdot E(x)$$
$$Var(y) = Var(\pm a \pm b \cdot x) = b^2 \cdot Var(x)$$
$$y \equiv N(E(y), Var(y))$$

- coefficiente di variazione

$$CV(X) = \frac{\sqrt{var(x)}}{E(x)}$$

- semplificare

$$\Phi(x) = 0.9$$
$$x = \Phi^{-1}(0.9)$$

25 Matlab

Comandi di Matlab

- densità di distribuzione: $normcdf(x, \mu, \sigma)$
- densità di distribuzione normalizzata: $normcdf(\frac{x-\mu}{\sigma})$
- percentile: $norminv(x, \mu, \sigma)$
- t studen percentile: *tinv(percentile, gradi libertà)*

26 Intervalli di Confidenza

In statistica, quando si stima un parametro, è spesso insufficiente individuare un singolo valore. È opportuno allora accompagnare la stima con un intervallo di valori plausibili per quel parametro, definito intervallo di confidenza.

1 - α: **livello di confidenza**, esprime la % in cui l'intervallo comprende il vero parametro μ

α: **livello di significatività,** esprime la % in cui il parametro non è compreso nell'intervallo

- Stimatore migliore della media $\mu = \bar{X}$

$$\bar{X} = \frac{1}{n}\sum_{j} X_{\mathrm{j}} = \frac{X_1 + ... + X_{\mathrm{n}}}{n}$$

- Stimatore migliore della varianza $\sigma^2 = S^2$

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{\mathrm{j}} - \bar{X})^2$$

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n} X_{\mathrm{j}} - (n \cdot \bar{X}^2)$$

- il quantile $1 - \frac{\alpha}{2}$ può essere usato come estremo destro e sinistro nelle distribuzioni simmetrice quindi: $N(\mu, \sigma^2), t - student$
- nelle distribuzioni asimmetriche $\mathrm{E_{sx}} = \frac{\alpha}{2}$ mentre $\mathrm{E_{dx}} = 1 - \frac{\alpha}{2}$ questo per la χ^2
- Gaussiana: coda destra = coda sinsitra calcolata nel punto cambiata di segno

IC per la media noto σ^2

$$\bar{X} - \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\bar{X} - \Delta \leq \mu \leq \bar{X} + \Delta$$

come se io scrivessi

$$\mu^* - \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \mu^* + \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Ampiezza intervallo

$$a = 2 \cdot \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 2 \cdot \Delta$$

$$a = E_{\text{dx}} - E_{\text{sx}} = (\bar{X} + \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) - (\bar{X} - \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}})$$

Margine di errore ε

$$\varepsilon = \frac{a}{2} = \Delta = \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Numerosità del campione, n

$$n \geq (\frac{\tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}}{\varepsilon})$$

$$n \in N$$

n è un numero naturale si deve sempre aprossimare per eccesso al prossimo numero intero 27,1 = 28

NB

$$P\{|\bar{X} - \mu| \leq K\} = 0.95$$

$k = \varepsilon =$ margine di errore
$0.95 =$ intervallo di confidenza

IC per la media noto σ^2

$$\bar{X} - \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \quad \mu \quad \leq \bar{X} + \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

IC per la media, ignota σ^2

$$\bar{X} - \tilde{t}_{n-1,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \leq \quad \mu \quad \leq \bar{X} + \tilde{t}_{n-1,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$$

IC per una percentuale π

$$\bar{X} - \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\bar{X} \cdot (1-\bar{X})}}{\sqrt{n}} \leq \quad \pi \quad \leq \bar{X} + \tilde{z}_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\bar{X} \cdot (1-\bar{X})}}{\sqrt{n}}$$

IC per σ^2

$$S^2 \cdot \frac{(n-1)}{\tilde{\chi}^2_{\frac{a}{2},n-1}} \leq \quad \sigma^2 \quad \leq S^2 \cdot \frac{(n-1)}{\tilde{\chi}^2_{1-\frac{a}{2},n-1}}$$

Intervalli di confidenza asitotici
Se abbiamo uno stimatore di θ Asintoticamente normale

$$\hat{\theta} \cong N(\theta, \frac{\tau^2}{n})$$

$$per\ n \to \infty$$

$$\hat{\theta} \mp z_{\frac{a}{2}} \frac{\tau}{\sqrt{n}}$$

Statistica modulo III

Silviu Filote

December 17, 2020

Contents

1 Riepilogo

Possibili Risultati Verifica di Ipotesi		
	Stato di Natura	
Decisione	H_0 Vera	H_0 Falsa
Non Rifiutare H_0	No errore $(1 - \alpha)$	Errore Secondo Tipo (β)
Rifiutare H_0	Errore Primo Tipo (α)	No Errore $(1 - \beta)$

$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$	test BILATERALE
$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$	test UNILATERALE DESTRO
$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$	test UNILATERALE SINISTRO

Figure 1: Dò sempre per vera H_0 per calcolare H_1

2 P-Value

È la probabilità di ottenere risultati uguali o meno probabili di quello osservato durante il test, supposta vera l'ipotesi nulla.
Il dato di riferimento è $\bar{x}/\sigma_{\text{dati}}$.

- P-Value = significatività osservata = $\hat{\alpha}$
- Regione rifiuto = errore I tipo = significatività nominale = α
- Accetto H_0 se $\quad \hat{\alpha} \geq \alpha$
- Rifiuto H_0 se $\quad \hat{\alpha} \leq \alpha$

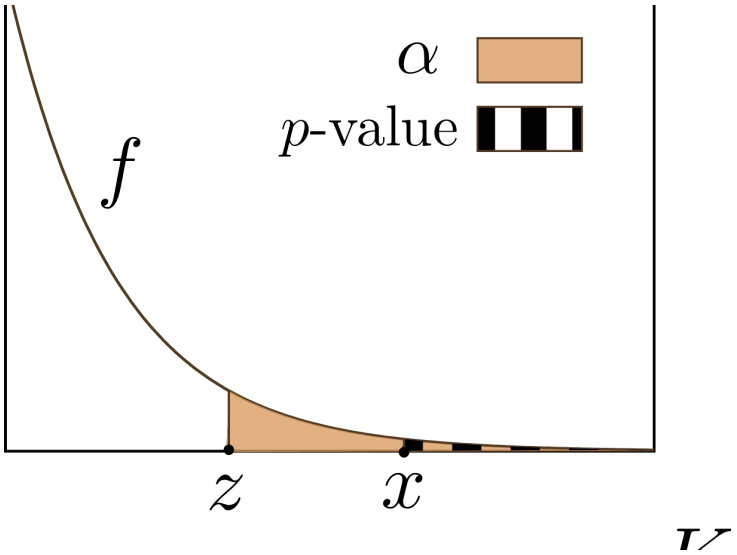


Figure 2: x = $\bar{x}_{\text{dati}}$

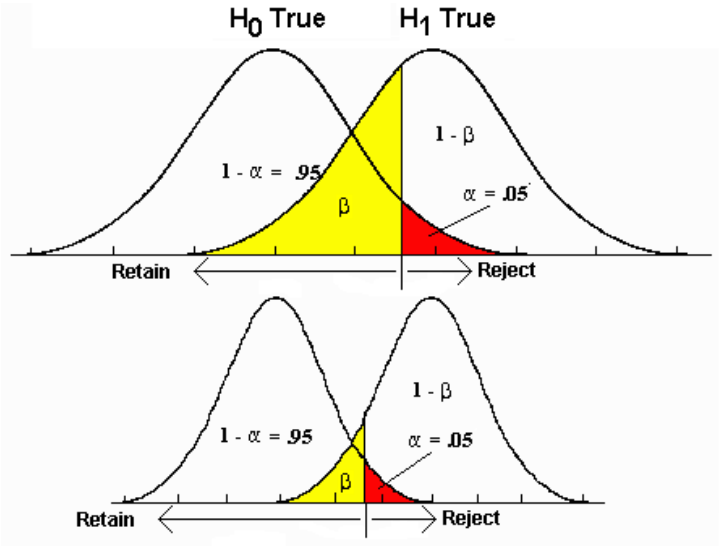
Inoltre:

- P-value > 0.10 Accetto H_0
- 0.05 < P-value < 0.10 Rifiuto H_0 al 10% = α (Rigetto debole)
- 0.01 < P-value < 0.05 Rifiuto H_0 al 5% = α (Rigetto medio)
- P-value < 0.01 Rifiuto H_0 al 1% = α (Rigetto forte)

3 Potenza

È la capacità che ha un test di riconoscere la falsità di H_0

$$\pi(\mu) = 1 - \beta(\mu)$$



$$\alpha = 1 - \Phi(z_{\mathrm{a}}) = 1 - \Phi(\frac{x_{\mathrm{crit}} - \mu_0}{\frac{\sigma}{\sqrt{n}}})$$

$$\beta = \Phi(\frac{x_{\mathrm{crit}} - \mu_1}{\frac{\sigma}{\sqrt{n}}})$$

$$\pi(\mu_1) = 1 - \beta = 1 - \Phi(\frac{x_{\mathrm{crit}} - \mu_1}{\frac{\sigma}{\sqrt{n}}})$$

Proprietà

- Tesi si dice **corretto o non distorno se**

$$\pi(\mu) > \alpha$$

- Test si dice **consistente se**, fissato α, $\forall \delta > 0$ si ha

$$\pi_{\mathrm{n}} \to 1 \quad per \quad n \to \infty$$

4 Test sulla media nota σ

Nota la varianza della popolazione viene utilizzata la **normale**.

test unilaterale destro

Regione di Rifiuto - $(R\ |H_0) = \{\bar{x} > x_{\text{crit}}\}$

$$\{\bar{x}_{\text{dati}} > \mu_0 + z_{1-a} \cdot \frac{\sigma}{\sqrt{n}}\}$$

Regione di accettazione - $(A\ |H_0) = \{\bar{x} < x_{\text{crit}}\}$

$$\{\mu_0 - z_{1-a} \cdot \frac{\sigma}{\sqrt{n}} < \bar{x}_{\text{dati}}\}$$

P-Value:

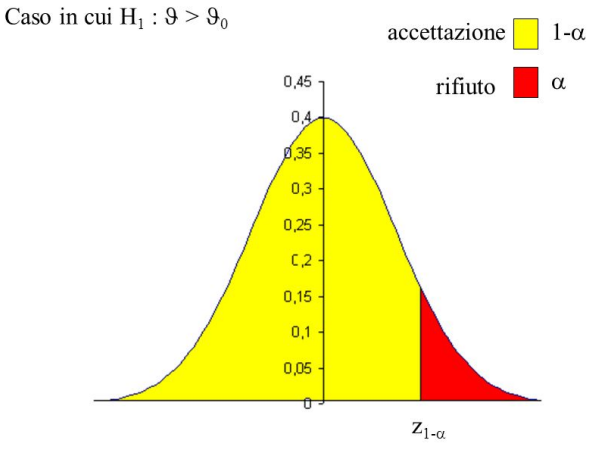
$$P(\bar{x} > \bar{x}_{\text{dati}}|H_0\ vera)$$

Valori critici:

$$x_{\text{crit}} = \mu_0 + z_{1-a} \cdot \frac{\sigma}{\sqrt{n}}$$

Potenza:

$$P(\bar{x} > x_{\text{crit}}|H_1) = 1 - \Phi(\frac{x_{\text{crit}} - \mu_1}{\frac{\sigma}{\sqrt{n}}})$$



Test unilaterale destro

test unilaterale sinistro

Regione di Rifiuto - (R $|H_0$) = $\{\bar{x} < x_{\text{crit}}\}$

$$\{\bar{x}_{\text{dati}} < \mu_0 - z_{1-a} \cdot \frac{\sigma}{\sqrt{n}}\}$$

Regione di accettazione - (A $|H_0$) $= \{\bar{x} > x_{\text{crit}}\}$

$$\{\bar{x}_{\text{dati}} > \mu_0 - z_{1-a} \cdot \frac{\sigma}{\sqrt{n}}\}$$

P-Value:

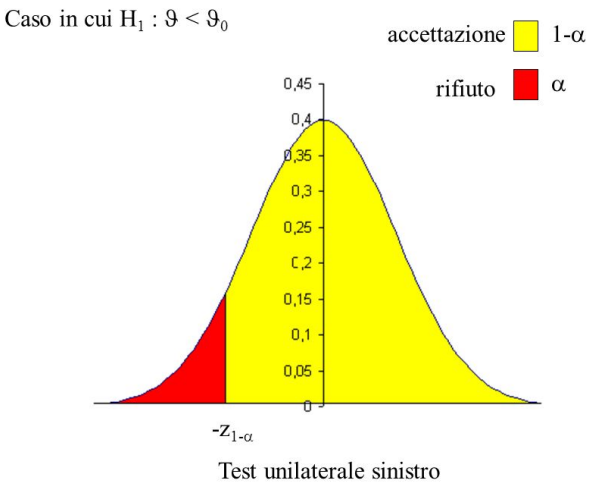
$$P(\bar{x} < \bar{x}_{\text{dati}} | H_0 \ vera)$$

Valori critici:

$$x_{\text{crit1}} = \mu_0 - z_{1-a} \cdot \frac{\sigma}{\sqrt{n}}$$

Potenza:

$$P(\bar{x} > x_{\text{crit}} | H_1) = \Phi(\frac{x_{\text{crit}} - \mu_1}{\frac{\sigma}{\sqrt{n}}})$$



Test unilaterale sinistro

test bilaterale

Regione di Rifiuto - $(R\ |H_0) = \{\bar{x} < x_{\text{crit1}} \ \vee \ \bar{x} > x_{\text{crit2}}\}$

$$\{\bar{x}_{\text{dati}} < \mu_0 - z_{1-\frac{a}{2}} \cdot \frac{\sigma}{\sqrt{n}}\} \quad \vee \quad \{\bar{x}_{\text{dati}} > \mu_0 + z_{1-\frac{a}{2}} \cdot \frac{\sigma}{\sqrt{n}}\}$$

Regione di Rifiuto, $\bar{x}$ stand. - $(R\ |H_0) =$

$$\{\bar{z}_{\text{dati}} < -z_{1-\frac{a}{2}}\} \quad \vee \quad \{\bar{z}_{\text{dati}} > z_{1-\frac{a}{2}}\}$$

Regione di accettazione - $(A\ |H_0) =$

$$\{\mu_0 - z_{1-\frac{a}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \bar{x}_{\text{dati}} < \mu_0 + z_{1-\frac{a}{2}} \cdot \frac{\sigma}{\sqrt{n}}\}$$

P-Value:

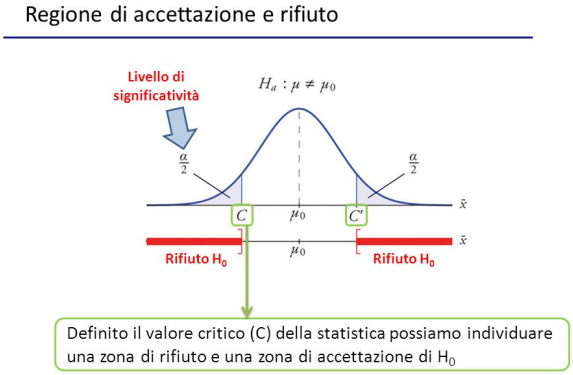
$$2 \cdot P(\bar{x} > \bar{x}_{\text{dati}} | H_0 \ vera)$$
$$2 \cdot P(\bar{z} > \bar{z}_{\text{dati}} | H_0 \ vera)$$

Valori critici:

$$x_{\text{crit1/2}} = \mu_0 \pm z_{1-\frac{a}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Potenza

$$P(\bar{x} < x_{\text{crit1}} \vee \bar{x} > x_{\text{crit2}} | H_1)$$
$$\Phi(\frac{x_{\text{crit1}} - \mu_1}{\frac{\sigma}{\sqrt{n}}}) + 1 - \Phi(\frac{x_{\text{crit2}} - \mu_1}{\frac{\sigma}{\sqrt{n}}})$$



13

5 Test sulla media ignota la σ

Ignota la varianza della popolazione viene utilizzata la t di student (distribuzione simmetrica).

- studentizzazione
- $\sigma = \mathrm{S}$
- potenza, pvalue .... = tcdf - tinv (matlab)

test unilaterale destro

Regione di Rifiuto - (R $|H_0$) = $\{\bar{x} > t_{\text{crit}}\}$

$$\{\bar{x}_{\text{dati}} > \mu_0 + t_{1-a,n-1} \cdot \frac{S}{\sqrt{n}}\}$$

Regione di accettazione - (A $|H_0$) $= \{\bar{x} < t_{\text{crit}}\}$

$$\{\mu_0 + t_{1-a,n-1} \cdot \frac{S}{\sqrt{n}} < \bar{x}_{\text{dati}}\}$$

P-Value:

$$P(\bar{x} > \bar{x}_{\text{dati}} | H_0 \ vera)$$

$$P(t_{\text{n - 1}} > \bar{t}_{\text{dati}} | H_0 \ vera) = P(t_{\text{n - 1}} > \frac{\bar{x}_{\text{dati}} - \mu_0}{\frac{S}{\sqrt{n}}} | H_0 \ )$$

Valori critici:

$$x_{\text{crit}} = \mu_0 + t_{1-a,n-1} \cdot \frac{S}{\sqrt{n}}$$

Potenza:

$$P(\bar{x} > x_{\text{crit}} | H_1) = 1 - T(\frac{x_{\text{crit}} - \mu_1}{\frac{S}{\sqrt{n}}})$$

6 Test sulla varianza

La distribuzione utilizzata in questo caso è un χ^2, distribuzione asimmetrica.

test unilaterale destro

Regione di Rifiuto - (R $|H_0$) =

$$\{S^2{}_{\text{dati}} > \frac{\sigma^2{}_0}{n-1} \cdot \chi^2{}_{1\text{ - a, n - 1}}\}$$

Regione di accettazione - (A $|H_0$) =

$$\{S^2{}_{\text{dati}} < \frac{\sigma^2{}_0}{n-1} \cdot \chi^2{}_{1\text{ - a, n - 1}}\}$$

P-Value:

$$P(S^2 > S^2{}_{\text{dati}} | H_0 \ vera)$$

$$P(\chi^2{}_{\text{n - 1}} > S^2{}_{\text{dati}} \cdot \frac{(n-1)}{\sigma^2{}_0} | H_0 \ vera)$$

test bilaterale

Regione di Rifiuto - (R $|H_0$) =

$$\{S^2{}_{\text{dati}} < \frac{\sigma^2{}_0}{n-1} \cdot \chi^2{}_{\frac{a}{2}, n-1}\} \vee \{S^2{}_{\text{dati}} > \frac{\sigma^2{}_0}{n-1} \cdot \chi^2{}_{1-\frac{a}{2}, n-1}\}$$

Regione di accettazione - (A $|H_0$) =

$$\{\frac{\sigma^2{}_0}{n-1} \cdot \chi^2{}_{\frac{a}{2}, n-1} < S^2{}_{\text{dati}} < \frac{\sigma^2{}_0}{n-1} \cdot \chi^2{}_{1-\frac{a}{2}, n-1}\}$$

valore critico

$$S^2{}_{\text{crit}} = \frac{\sigma^2{}_0}{n-1} \cdot \chi^2{}_{1\text{ - a, n - 1}}$$

$$\chi_{\text{crit}} = \frac{S^2 * (n-1)}{\sigma^2{}_0}$$

7 Test su una percentuale

Viene utilizzata come distribuzione la normale.

test unilaterale destro

$$P(A|H_0) = \hat{\pi}_{\text{dati}} < \pi_0 + z_{1\text{ - a}} \cdot \sqrt{\frac{\pi_0 \cdot \ (1-\pi_0)}{n}}$$

test unilaterale sinistro

$$P(A|H_0) = \hat{\pi}_{\text{dati}} > \pi_0 - z_{1\text{ - a}} \cdot \sqrt{\frac{\pi_0 \cdot \ (1-\pi_0)}{n}}$$

test unilaterale bilaterale

$$P(A|H_0) = \pi_0 - z_{1-\frac{a}{2}} \cdot \sqrt{\frac{\pi_0 \cdot \ (1-\pi_0)}{n}} < \hat{\pi}_{\text{dati}} < \pi_0 + z_{1-\frac{a}{2}} \cdot \sqrt{\frac{\pi_0 \cdot \ (1-\pi_0)}{n}}$$

valore critico

$$x_{\text{crit}} = \pi_0 - z_{1-\frac{a}{2}} \cdot \sqrt{\frac{\pi_0 \cdot \ (1-\pi_0)}{n}}$$

$$p_{\text{crit}} = \frac{\hat{\pi}_{\text{dati}} - \pi_0}{\sqrt{\frac{\pi_0 \cdot \ (1-\pi_0)}{n}}}$$

$$p_{\text{crit}} = z_{1-\frac{a}{2}}$$

Potenza $\pi_1|H_1$

$$\pi(\mu_1) = 1 - \Phi(\frac{x_{\text{crit}} - \pi_1}{\sqrt{\frac{\pi_1 \cdot \ (1-\pi_1)}{n}}})$$

8 Informazioni

- $x_{crit1...critn}$ vengolo calcolati tramite l'ipotesi H_0
- $x_{crit1...critn}$ possono essere standardizzati, studentizzati..... alla distribuzione che si sta utilizzando
- il PV viene calcolato a partire dal dato calcolato, osservato (μ, σ^2, π)
- per il calcolo della potenza unilaterale destro viene utilizzata la formula

$$1 - \beta = \pi = 1 - \Phi(z_{\text{crit}} - \sqrt{n} \cdot \delta) = \Phi(\sqrt{n} \cdot \delta - z_{\text{crit}})$$

$$\delta = \frac{\mu_1 - \mu_0}{\sigma}$$

- Per il calcolo della Potenza del bilatateraleponiamo il μ_1 in un range $(\mu_0 - 2) : 0.01 : (\mu_0 + 2)$
- ricordarsi di scrivere sempre n
- fare accettazione a S2 e S
- per accettare o meno la veridicità del testo H_0 si guardano i dati osservati, calcolati
- diamo sempre per vera l'ipotesi H_0 per il calcolo di H_1
- potenza si calcola partendo da x_{crit}

Statistica modulo IV

Silviu Filote

January 7, 2021

Contents

1 Esercizio con i dati

Nel caso in cui i valori di x e y sono definiti direttamente da testo.
Procedura:

$$x = [.....]$$
$$y = [.....]$$

a) Calcolare il coefficiente di correlazione lineare di X e Y;

$$matlab : cov(y, x)$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

- $a = var(y) = S^2{}_{\mathrm{y}} = \hat{\sigma}^2{}_{\mathrm{y}}$
- $d = var(x) = S^2{}_{\mathrm{x}} = \hat{\sigma}^2{}_{\mathrm{x}}$
- $c = \hat{\sigma}_{\mathrm{xy}}$
- $d = \hat{\sigma}_{\mathrm{xy}}$
- NB: se prima inserisco y in "cov($\mathbf{y}$, x)" allora avrò a = var(y)

$$r = \frac{\sum_{t=1}^{n}(x_{\mathrm{t}} - \bar{x}) \cdot (y_{\mathrm{t}} - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_{\mathrm{t}} - \bar{x})^2 \cdot \sum_{t=1}^{n}(y_{\mathrm{t}} - \bar{y})^2}} = \frac{\hat{\sigma}_{\mathrm{xy}}}{\hat{\sigma}_{\mathrm{x}}{}^2 \cdot \hat{\sigma}_{\mathrm{y}}{}^2}$$

$$oppure$$

$$matlab : corrcoef(y, x)$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

- $a = 1$ correlazione lineare y - y
- $d = 1$ correlazione lineare x - x
- $c = b = r$

c) Calcolare con il metodo dei minimi quadrati ordinari i parametri del modello di regressione e le rispettive standard deviations;

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

$$\bar{y} = \bar{a} + b \cdot \bar{x}$$

$$b = \hat{\beta}_1 = \frac{\hat{\sigma}_{\mathrm{xy}}}{\hat{\sigma}_{\mathrm{x}}{}^2}$$

$$a = lm.Coefficients.Estimate(1) = \bar{y} - b \cdot \bar{x}$$

$$var(\hat{\beta}_1) = \frac{\sigma_{\varepsilon}{}^2}{\sum (x - \bar{x})^2}$$

$$var(\hat{\beta}_0) = \frac{\sigma_{\varepsilon}{}^2}{\sum (x - \bar{x})^2} \cdot m_2(X)$$

Matlab:

- lm = fitlm(x,y)
- $\hat{\beta}_0$ = lm.Coefficients.Estimate(1)
- $\hat{\beta}_1$ = lm.Coefficients.Estimate(2)
- $std(\hat{\beta}_0)$ = lm.Coefficients.SE(1) = sqrt(lm.CoefficientCovariance(1,1))
- $std(\hat{\beta}_1)$ = lm.Coefficients.SE(2) = sqrt(lm.CoefficientCovariance(2,2))

d) Stima della varianza residua

$$s_\varepsilon{}^2 = MSE = \frac{1}{n-2} \cdot \sum e^2{}_\mathrm{t} = \frac{1}{n-2} \cdot \sum y_\mathrm{t} - \hat{y}_\mathrm{t}$$

$$RMSE = \sqrt{s_\varepsilon} = \sqrt{MSE}$$

$$s_\varepsilon = \frac{D_\mathrm{res}}{n-2} = \frac{\sum e^2{}_\mathrm{i}}{n-2} = (1 - r^2) \cdot \frac{D_\mathrm{tot}}{n-2}$$

- $s_\varepsilon{}^2 = \sigma_\varepsilon{}^2$
- SST (Sum of squared total) = $\mathrm{D_{tot}}$
- SSR (Sum of squared regression) = $\mathrm{D_{sp}}$
- SSE (Sum of squared error) = $\mathrm{D_{res}}$
- MSE (Mean squared error) = varianza residua
- RMSE (Root mean squared error) = $\sqrt{\textit{varianza residua}}$

$$D_\mathrm{tot} = D_\mathrm{res} + D_\mathrm{sp}$$

$$\frac{D_\mathrm{tot}}{D_\mathrm{tot}} = \frac{D_\mathrm{res}}{D_\mathrm{tot}} + \frac{D_\mathrm{sp}}{D_\mathrm{tot}}$$

$$R^2 = \frac{D_\mathrm{sp}}{D_\mathrm{tot}} = 1 - \frac{D_\mathrm{res}}{D_\mathrm{tot}}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Matlab:

- $\mathrm{s^2}$ = varianza residua = lm.MSE
- $\mathrm{s^2}$ = lm.SSE / (n-2)
- $\mathrm{s^2}$ = lm.RMSE^2

e) Stimare la varianza spiegata del modello e della varianza totale;

Matlab:

- s^2_{sp} = lm.SSR/(n-2) (corretta per i gradi di libertà)
- s^2_{tot} = lm.SST/(n-2) (corretta per i gradi di libertà)

h) Calcolare il coefficiente di determinazione lineare e commentare;

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = r^2$$

Matlab:

- r2 = lm.Rsquared.Ordinary

f) Determinare l'intervallo di confidenza al 95% per β_1;

$$\hat{\beta}_1 - t_{1-\frac{a}{2},n-2} \cdot std(\hat{\beta}_1) \;<\; \beta_1 \;<\; \hat{\beta}_1 + t_{1-\frac{a}{2},n-2} \cdot std(\hat{\beta}_1)$$

Matlab:

- IC = coefCI(lm,0.05) = (modello stimato, %)
- $IC_{\beta 1}$ = IC(2,:)

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$(a\ b) \rightarrow IC\ di\ \beta_0$$

$$(c\ d) \rightarrow IC\ di\ \beta_1$$

g) Test d'ipotesi

$$\alpha = 0.05$$

$$H_0 : \beta_1 = 0$$

$$H_0 : \beta_1 \neq 0$$

$$t_{\text{crit1}} = -t_{1-\frac{a}{2}, n-2}$$

$$t_{\text{crit2}} = +t_{1-\frac{a}{2}, n-2}$$

$$PV = 2 \cdot tcdf(t_{\text{oss}})$$

Matlab:

- t_{oss} = lm.Coefficients.tStat(2)
- t_{crit1} = - tinv(1-alpha/2,n-2)
- t_{crit2} = + tinv(1-alpha/2,n-2)
- regione di accettazione: $t_{\text{crit1}} < A < t_{\text{crit2}}$
- PV = lm.Coefficients.pValue(2)

2 Esercizio senza dati

Nel caso in cui i valori di x e y non sono definiti direttamente da testo.
Procedura:

$$\sum_{i=1}^{12} x_{\mathrm{i}}$$

$$\sum_{i=1}^{12} y_{\mathrm{i}}$$

$$\sum_{i=1}^{12} {x_{\mathrm{i}}}^2$$

$$\sum_{i=1}^{12} {y_{\mathrm{i}}}^2$$

$$\sum_{i=1}^{12} x_{\mathrm{i}} \cdot y_{\mathrm{i}}$$

Resoluzione:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{12} x_{\mathrm{i}}$$

$${s_{\mathrm{x}}}^2 = \frac{1}{n} \cdot (\sum_{i=1}^{12} {x_{\mathrm{i}}}^2) - (\bar{x})^2$$

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{12} y_{\mathrm{i}}$$

$${s_{\mathrm{y}}}^2 = \frac{1}{n} \cdot (\sum_{i=1}^{12} {y_{\mathrm{i}}}^2) - (\bar{y})^2$$

$$\hat{\sigma}_{\mathrm{xy}} = \frac{1}{n} \cdot (\sum_{i=1}^{12} x_{\mathrm{i}} \cdot y_{\mathrm{i}}) - (\bar{x} \cdot \bar{y})$$

$$r = \frac{\hat{\sigma}_{\mathrm{xy}}}{\sqrt{{s_{\mathrm{x}}}^2 \cdot {s_{\mathrm{y}}}^2}}$$

$$codevianza_{\mathrm{xy}} = S_{\mathrm{xy}} = \hat{\sigma}_{\mathrm{xy}} \cdot n$$

$$devianza_{\mathrm{xx}} = {S_{\mathrm{xx}}}^2 = {s_{\mathrm{x}}}^2 \cdot n$$

$$devianza_{\mathrm{yy}} = {S_{\mathrm{yy}}}^2 = {s_{\mathrm{y}}}^2 \cdot n$$

$$r = \frac{S_{\mathrm{xy}}}{\sqrt{{S_{\mathrm{xx}}}^2 \cdot {S_{\mathrm{yy}}}^2}}$$

Stima parametri osservati:

$$b = \hat{\beta}_1 = \frac{\hat{\sigma}_{\mathrm{xy}}}{{s_{\mathrm{x}}}^2}$$

$$b = \hat{\beta}_1 = \frac{S_{\mathrm{xy}}}{{S_{\mathrm{xx}}}^2}$$

$$a = \hat{\beta}_0 = \bar{y} - b \cdot \bar{x}$$

$$Var - residua = {s_\varepsilon}^2 = (1 - r^2) \cdot \frac{D_{\mathrm{tot}}}{n-2} = (1 - r^2) \cdot \frac{{S_{\mathrm{yy}}}^2}{n-2}$$

$$var(\beta_1) = \frac{{s_\varepsilon}^2}{{S_{\mathrm{xx}}}^2} = \frac{{\sigma_\varepsilon}^2}{{S_{\mathrm{xx}}}^2}$$

$$var(\beta_0) = \frac{{s_\varepsilon}^2}{{S_{\mathrm{xx}}}^2} \cdot (\frac{1}{n} \cdot \sum_{i=1}^{12} {x_{\mathrm{i}}}^2) \rightarrow M_2$$

$$std(\beta_1) = SE_{\beta_1} = \sqrt{var(\beta_1)}$$

$$std(\beta_0) = SE_{\beta_0} = \sqrt{var(\beta_0)}$$

$$R^2 = r^2$$