Esercizio 1 Probabilità che Oneto dia 30L (Lode)
$n = 120$
$\sum_i^{120} x_i = 18$
$\hat{P} = \frac{18}{120} = 0.15 \rightarrow 15\%$

Esercizio 2 N studenti da 30 e lode
$n_1 = 18 \leftarrow$ Oneto
$n_2 = 20 \leftarrow$ Anguita
$n_{1,2} = 10 \leftarrow$ 30L sia con Oneto che con Anguita
$N =?$ Studenti da **30 e Lode**

$$\hat{P}_1 \approx \frac{n_1 2}{n_2} \qquad \hat{P}_1 \approx \frac{n_1}{N} \qquad \frac{n_{1,2}}{n_2} = \frac{n_1}{N}$$

$\Longrightarrow N = \frac{n_1 \cdot n_2}{n_{1,2}} \rightarrow \frac{18 \cdot 20}{10} = 36$

Esercizio 3 Stima del numero di incidenti medio in auto n = 10
$x_1 = \{4, 0, 6, 5, 2, 1, 2, 0, 4, 3\}$
$\hat{\lambda} = \frac{\sum_i x_i}{n} = \frac{27}{10} = 2.7$

$$P\{x \leq 2\} = e^{-2.7}(\frac{2.7^0}{0!} + \frac{2.7^1}{1!} + \frac{2.7^2}{2!}) \approx .4936 \rightarrow 49.36\%$$

Probabilità che non ci siano più di **2 incidenti**

Esercizio gaussiana primo
$x_1 = 1.7$
$x_2 = 1.82$
$x_3 = 1.73$
$x_4 = 1.7$

$x_5 = 1.8$

$$\hat{\mu} = \frac{\sum_i x_i}{n} = \frac{1.7 + 1.82 + 1.73 + 1.7 + 1.8}{5} = 1.75$$

$$\hat{\sigma} = \sqrt{\frac{0.05^2 + 0.07^2 + 0.02^2 + 0.05^2 + 0.05^2}{5}} \approx 0.051$$

Esempio: Sistema di comunicazione $\sigma^2 = 4 \quad n = 9$

$$x_1 = \{5.85, 12, 15, 7, 9, 7.5, 6, 5, 10.5\}$$

$$\hat{\mu} = \frac{1}{n} \sum_i^n x_i = \frac{1}{9} \sum_i^n x_i = \frac{81}{9} = 9$$

$$P\left(9 - 1.96\frac{\sigma}{\sqrt{m}} < \mu < 9 + 1.96\frac{\sigma}{\sqrt{m}}\right) = 0.95$$

$$p\left(9 - 1.96\frac{2}{3} < \mu < 9 + 1.96\frac{2}{3}\right) = 0.95$$

$\longrightarrow [7.693, 10.31] \rightarrow \mu$ si trova tra 7.693 e 10.31

In generale Prob $= 1 - \alpha$

$(\overline{x} - z_a \frac{\sigma}{\sqrt{n}}, \overline{x} + z_a \frac{\sigma}{\sqrt{n}}) \rightarrow$ Si rileva dalle tavole

0.1 Intervalli di confidenza (Unilaterali)

0.2 Esempio:

Pesca stagionale dei salmoni (*Fisso intervallo -> trovo* n)
Ad ogni stagione il peso medio dei salmoni è diverso ma $\sigma = 0.3$ Kg
Intervallo di confidenza al 95%, quindi $\alpha = 0.05$

$$(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

$$1.96\frac{\sigma}{\sqrt{n}} \geq 0.1 \quad \sqrt{n} \geq \frac{1.96}{0.1}\sigma$$

$$n \geq (\frac{1.96}{0.1}0.3)^2 = 5.88^2 \approx 34.6 \leftarrow \text{salmoni}$$

0.3 Intervallo di confidenza

con *media* e *varianza* **incognite**

Esempio: Trasimttente (μ) e ricevitore (μ + rumore)

$$95\%(7.69, 10.31) \quad \hat{\mu} = 9, \sigma^2 = 4$$

$X_i\{5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5\}$
$\hat{\mu} = \overline{X} = \frac{1}{9}\sum_i^n X_i = \frac{81}{9} = 9$
$s^2 = \frac{1}{8}\sum_i(X_i^2 - 9.81) \approx 9.5 \quad s = 3.082$

$$\mu \in (9 - 2.306\frac{3.082}{3}, 9 + 2.306\frac{3.082}{3}) = (6.63, 11.37)$$

Si può dimostrare che $T_{\frac{\alpha}{2}\cdot n-1}\mathbb{E}[S] \geq z_\alpha\sigma$

0.4 Intervallo di confidenza nella varianza

Esempio: Laminatoio n = 4 $X_i = \{0.123, 0.124, 0.126, 0.12\}$spessore in **mm**

Svolgimento

$$\frac{1}{4}\sum_{i}^{4} X_i = \frac{0.493}{4} = 0.12325$$

$$\frac{1}{4-1}\sum_{i}^{4} (X_i - 0.12325)^2 = 1.875 \cdot 10^{-5}$$

$$\sigma^2 \in \left(\frac{s^2(n-1)}{9.348}, \frac{s^2(n-1)}{0.216}\right)$$

Dove **9.348** e **0.216** sono ricavati dalle tabelle

Facciamo la radice:
$\sigma \in (0.0014, 0.0093) \rightarrow 95\%$

0.5 Intervallo di confidenza

della differenza di due medie:

N campioni	**M** campioni
$X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$	$Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$
$\overline{X} = \frac{1}{n}\sum_{i}^{n} X_i$	$\overline{Y} = \frac{1}{m}\sum_{i}^{m} Y_i$

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

$$\mathcal{N}(0,1) \sim \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

$\mu_1 - \mu_2 \in \left(\overline{X} - \overline{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \overline{X} - \overline{Y} + z_{\frac{\alpha}{2}}, \overline{X} - \overline{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right)$

Se σ_1^2, σ_2^2 non sono note:

$S_1^2 = \frac{1}{n-1}\sum_i^n (X_i - \overline{X})$ $\qquad$ $s_2^2 = \frac{1}{m-1}\sum_i^m (X_i - \overline{Y})$

$(n-1)\frac{S_1^2}{\sigma_1^2} \sim \mathcal{X}_{n-1}^2$ $\qquad$ $(n-1)\frac{s_2^2}{\sigma_2^2} \sim \mathcal{X}_{n-1}^2$

Possiamo andare avanti solo se $\boldsymbol{\sigma_1^2 = \sigma_2^2 = \sigma^2}$

$$(n-1)\frac{s_1^2}{\sigma^2} + (n-1)\frac{s_2^2}{\sigma^2} \sim \mathcal{X}_{n+m-2}^2$$

$$\underset{\sim \mathcal{N}(0,1)}{\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{m})}}} \longrightarrow \underset{\sim T_{n+m-2}}{\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p(\frac{1}{n} + \frac{1}{m})}}}$$

$S_p = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$

Se σ sono ignote ma uguali

$\mu_1 - \mu_2 \in (\overline{X} - \overline{Y} - T_{\frac{\alpha}{2}, n+m-2}\sqrt{s^2(\frac{1}{n} + \frac{1}{m})})$
$\overline{X} - \overline{Y} + T_{\frac{\alpha}{2}, n+m-2}\sqrt{s^2(\frac{1}{n} + \frac{1}{m})}$

0.6 Intervallo di previsione

Esempio smartwatch contapassi $\boldsymbol{n = 7}$

$$\begin{array}{lll} LUN & 6922 & X_1 \\ MAR & 5333 & X_2 \\ MER & 7420 & X_3 \end{array} \qquad \begin{array}{lll} GIO & 7432 & X_4 \\ VEN & 6252 & X_5 \\ SAB & 7005 & X_6 \end{array}$$

$$DOM \quad 6752 \quad X_7$$

$$\overline{X}_n = \frac{1}{n}\sum_{i}^{m} X_i = \frac{47016}{7} \approx 6717$$

$$1-\alpha = 95\% \quad \alpha = 5\%$$

$$t_{0.0025,6} = 2.997$$

$$S_n = \sqrt{S_n^2} = 7.333.8$$

$$x_{n+1} \in (6717 - 2.447 \cdot 733397\sqrt{1+\frac{1}{7}}, 6717 + 2.447 \cdot 73397\sqrt{1+\frac{1}{7}})$$

$X_{n+1} \in (9796, 8637)\mu \in (6037, 7396)$

0.7 Qualità di uno stimatore

$X = X_1 \dots X_n \quad \theta \leftarrow$ parametro $\qquad d(x) \leftarrow$ stimatore di θ

$$\boldsymbol{(d(x)-\theta)^2} \qquad\qquad \mathbb{E}\left[\boldsymbol{(d(x)-\theta)^2}\right]$$

Errore Quadratico (*misura della qualità*) Errore Quadratico Medio (*M.S.E*)
Rischio $r(d,\theta) = \mathbb{E}\left[(d-\theta)^2\right]$ Lo stimatore "ottimo" sarà quello con il rischio minimo -> d con r minimo θ

Esempio $\quad d^*(x) = 4 \quad$ se$\theta = 4 \Rightarrow d^* =$ stimatore ottimo(per tutti gli altri valori non va

0.8 Proprietà di uno stimatore

Def: $b_\theta(d) = \mathbb{E}[d] - \theta \rightarrow$ bias o polarizzazione Uno stimatore non è **polarizzato** se $b_\theta(d) = 0$

Esempio : $X_1 \dots X_n \quad \theta$media
$d_1(X_1 \dots X_n) = X_1$
$d_2(X_1 \dots X_n) = \frac{X_1+X_2}{2}$
$d_3(X_1 \dots X_n) = \frac{X_1+X_2+\dots X_n}{n}$ Tutti questi sono **unbiased**

0.9 Stimatore unbaieseo

$r(d,\theta) = \mathbb{E}\left[(d(x)-\theta)^2\right] = \mathbb{E}\left[(d(x)-\mathbb{E}[d(x)])^2\right] = Var(d)$
tra gli stimatori non polarizzati di ottimo è quello con la varianza minima

0.10 Valutazione di uno stimatore

$$X = X_1 \dots X_n \quad \theta =?$$

Dove θ è un *parametro* e $d(x)$ è uno *stimatore* di θ

$r(d,\theta)$(mse) rischio $\quad b_\theta(d) = \mathbb{E}[d] - \theta$
se $b_\theta(d) = 0 \Rightarrow r(d,\theta) = Var(d)$

se $b_\theta(d) \neq 0$? $r(d,\theta)$ = ?

$$\begin{aligned} r(d,\theta) = \mathbb{E}\left[(d(x)-\theta)^2\right] &= \mathbb{E}\left[(d(x)-\mathbb{E}[d]+\mathbb{E}[d]-\theta)^2\right] \\ &= \mathbb{E}\left[(d-\mathbb{E}[d])^2+(\mathbb{E}[d]-\theta)^2-2(d-\mathbb{E}[d])(\mathbb{E}[d]-\theta)\right] \\ &= \mathbb{E}\left[(d-\mathbb{E}[d])^2\right]+\mathbb{E}\left[(\mathbb{E}[d]-\theta)^2\right]-2(\mathbb{E}[d]-\theta)\cdot\mathbb{E}\left[(d-\mathbb{E}\right. \end{aligned}$$

$$\begin{aligned} r(d,\theta) &= \mathbb{E}\left[(d-\mathbb{E}[d])^2\right]+\mathbb{E}\left[(\mathbb{E}[d]-\theta)^2\right] \\ &= Var(d) + b_\theta(d)^2 \leftarrow \text{bias}^2 \end{aligned}$$

0.11 Esempio:

Stimatore della media di una *distribuzione uniforme*

$\mathbb{E}\left[X_i\right] = \theta/2$
$X_1, X_2 \dots X_n$

$d_1 = 2\frac{1}{n}\sum_i^n X_i$
$d_2 = \max X_i$

$d_1 : \mathbb{E}\left[d_1\right] = \frac{2}{n}\sum_i \mathbb{E}\left[X_i\right] = \frac{2}{n}n\frac{\theta}{2} = \theta$
$r(d_1, \theta) = Var(d_1) = \frac{4}{n^2}nVar(X_i) = \frac{4}{n}\frac{\theta^2}{12} = \boldsymbol{\frac{\theta^2}{3n}} \Leftarrow$ *Unbiased*

$$F_2(x) = P_r\{d_2(x) \leq x\} = P_r\{\max X_1 \leq x\}$$
$$= P_r\{X_1 \leq, \forall i \in 1\} = \prod_{i=1}^n P_r\{X_i \leq x\} = (\frac{x}{\theta})^n$$
$$f_2(x) = \frac{d}{dx}F_2(x) = n\frac{x^{n-1}}{\theta^n} \quad x \leq \theta$$

$$\mathbb{E}\left[d_2\right] = \int_0^\theta x f_x(x)\, dx = \int_0^\theta \frac{n}{\theta^n}x^n\, dx = \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\bigg|_0^\theta\right] = \frac{n}{n+1}\theta$$

$$\mathbb{E}\left[d_2^2\right] = \frac{n}{\theta^n}\int_0^\theta x^2 f(x)\, dx = \frac{n}{\theta^n}\int_0^\theta x^{n+1}\, dx = \frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2}\bigg|_0^\theta\right] = \frac{n}{n+2}\theta^2$$

$Var(d_2) = \mathbb{E}\left[d^2\right] - \mathbb{E}\left[d_2\right]^2 = \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \boldsymbol{\frac{n}{(n+2)(n+1)^2}\theta^2}$

$r(d_2,\theta) = Var(d_2) + (\mathbb{E}[d_2] - \theta)^2 = \frac{\mathbf{2 \cdot \theta^2}}{\mathbf{(n+1)(n+2)}}$

$n \geq 4 \quad r(d_2,\theta) < r(d_1,\theta) \qquad d_3 = \frac{n+1}{n} d_2$

In sintesi

$$r(d_1,\theta) = \frac{\theta^2}{3n} \Leftarrow \text{Unbiased}$$
$$r(d_2,\theta) = \frac{2\theta^2}{(n+1)(n+2)} \Leftarrow \text{Biased}$$
$$r(d_3,\theta) = \frac{\theta^2}{n^2+2n} \Leftarrow \text{Unbiased}$$
$$r(d_4,\theta) = \frac{\theta^2}{(n+1)^2} \Leftarrow \text{Biased}$$

1 Test di ipotesi

Ipotesi: Affermazione rispetto a uno o più parametri di una distribuzione Ipotesi da confutare: H_0 (ipotesi nulla)

Esempio :

$$\begin{aligned} X_1 \dots X_n \sim \mathcal{N}(\mu, \sigma^2) \\ H_0 : \mu = 0 \\ H_a : \mu \neq 0 \end{aligned} \tag{1}$$

Diamo per scontato che l'ipotesi sia **vera**
Dobbiamo cercare di *confutarla*

Definizione Regione critica tale che:

$(X_1 \dots X_n) \in C \rightarrow H_0$è rifiutata
$(X_1 \dots X_n) \notin C \rightarrow H_0$è accettata
$\alpha =$ Livello di **significatività** del test ($\alpha = 10\%, 5\% \dots$)

Procedimento

- Fisso alpha
- Suppongo che α sia vera
- calcolo stima di μ
- verifico che non sia "*troppo distante*"

$X_1 \dots X_n \sim \mathcal{N}(\mu, \sigma^2)$
$H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0$
$\overline{X} = \frac{1}{n} \sum_i X_i$

Regione critica $\{(X_1 \dots X_n) : |\overline{X} - \mu_0| > c\}$
$P_{r_{\mu_0}} \left\{|\overline{X} - \mu_0| > c\right\} = \alpha$
$P_{r_{\mu_0}} \left\{\frac{|\overline{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} > \frac{c}{\frac{\sigma}{\sqrt{n}}}\right\} = \alpha$
$P_{r_{\mu_0}} \{|z| > z_\alpha\} = \alpha$

Esempio (5 transimissioni)

$\boldsymbol{n = 5}$ $\quad$ $H_0 : \mu = 8$

$\overline{X} = 9.5$ $\quad$ $\alpha = 5\%$

Ipotizzando che H_0 sia vera:

$$\frac{|\overline{X} - \mu|}{\frac{\sigma}{\sqrt{n}}} = \frac{|9.5 - 8|}{\frac{2}{\sqrt{5}}} \approx 1.68$$

Se:

$\alpha = P_r$(rifiuto H_0 / H_0 vera)

$\alpha \uparrow$ più *"facile"* rifiutare l'ipotesi

$\alpha \downarrow$ più *"difficile"* rifiutare l'ipotesi

1.1 Metolodogia alternativa

$$Ts = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \rightarrow \text{Statistica di test}$$

$P-$value = **probabilità** di ottenere un valore più "anomalo" di quello osservato

Esempio: $X_i \sim \mathcal{N}(\mu, 4)$

$n = 5$ $\quad$ $H_0 : \mu = 8$

$\overline{X} = 8.5$ $\quad$ $H_a : \mu \neq 8$

$$\frac{|\overline{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} = \frac{|8.5 - 8|}{\frac{2}{\sqrt{5}}} = \frac{\sqrt{5}}{2} 0.5 \approx 0.559$$

$$P\{|z| > 0.559\} = 2P\{z > 0.559\} \approx 2 \cdot 0.288 = 0.579 \rightarrow \text{P-value}$$

Se $\overline{\boldsymbol{X}} = \mathbf{11.5}$:

$$\frac{|\overline{X} - \mu_0|}{\frac{\sigma}{\sqrt{n}}} = \frac{|11.5 - 8|}{\frac{2}{\sqrt{5}}} \approx 3.913$$

$$P\{|z| > 3.913\} = 2P\{z > 3.913\} \leq 0.00005 \rightarrow \underline{\text{Rifiuto ipotesi } H_0}$$

1.2 Test di Hp unilaterale

$$H_0 : \mu = \mu_0 (\mu \leq \mu_0) \qquad\qquad H_a : \mu > \mu_0$$

$$C = \{(X_1 \ldots n) \cdot \overline{X} - \mu_0 > c\}$$

$$P_{r_{\mu_0}}\{\overline{X} - \mu_0 > c\} = P_r\{\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{c}{\frac{\sigma}{\sqrt{n}}}\} = P_{r_{\mu_0}}\{z > z_a\} = \alpha$$

Statistica test $\boldsymbol{\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq z_\alpha}$ accetto

1.3 Test di ipotesi

H_0	H_a	TS	Livello α	P - Value
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{1}}$	Rifiuto H_0 se $TS > \frac{z\alpha}{2}$	$2P(z \geq \lvert TS \rvert)$

Altre ipotesi :

H_0	H_a	TS	Livello α	P - Value
$\mu < \mu_0$	$\mu > \mu_0$	$\frac{\overline{X}-\mu_0}{\sigma/\sqrt{2}}$	$H_0 \quad z_\alpha > TS$	$P(z \geq TS)$
$\mu \geq \mu_0$	$\mu < \mu_0$	//	$H_0 \quad z_\alpha < -TS$	$P(z \leq TS)$

1.4 Uguaglianza media di due popolazioni

$X_1 \ldots X_n \sim \mathcal{N}(\mu_1, \sigma_2^2)$
$\overline{X} = \frac{1}{n}\sum_i^n X_i$
$S_x^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$

$Y_i \ldots Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$
$\overline{Y} = \frac{1}{m}\sum_i^m Y_i$
$S_y^2 = \frac{1}{m-1}\sum_i (Y_i - \overline{Y})^2$

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m}$$

H_0	H_a	TS	Livello α	P - Value
$\mu_1 = \mu_2$	$\mu \neq \mu_2$	$\frac{\overline{X}-\overline{Y}}{\sqrt{\sigma_{1/n}^2 + \sigma_{2/m}^2}}$		
$\mu_1 = \mu_2$	$\mu \neq \mu_2$	$\frac{\overline{X}-\overline{Y}}{\sqrt{S_p^2(\frac{1}{n}+\frac{1}{m})}}$	rif. $\lvert TS\rvert > z_{\frac{\alpha}{2}}$	$2P(z \geq \lvert TS\rvert)$
$\mu_1 = \mu_2$	$\mu \neq \mu_2$	$S_i \in$ T-student		

4) T-test per coppie di dati Se X_1 e X_2 **NON** sono indipendenti

$W_i = X_i - Y_i$

$X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$
$Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$

ES Manutenzione (n guasti) tagliand

$H_0 : \mu_a - \mu_b \geq 0 \quad \overline{W} = \frac{1}{5}(-7.5 + 2.5 - 2.5 - 3.5 - 1.5) = -2.5$

$S_W^2 = \frac{1}{4}(W_i - \overline{W})^2 = 13$

$Ts = \frac{\frac{\overline{W}}{S_W}}{\sqrt{n}} = \frac{\frac{-2.5}{\sqrt{13}}}{\sqrt{5}} = 1.55$

$$P_r\{T_{n-1} \leq Ts\} = \{T_4 \leq Ts\}$$

5) Test sulla varianza

$$H_0 : \sigma^2 = \sigma_0^2 \quad H_a : \sigma^2 \neq \sigma_0^2$$

$$\frac{(n-1)}{\sigma_0^2} \sim \mathcal{X}_{n-1}^2$$

$Pr\{\mathcal{X}^2_{1-\frac{\alpha}{2},n-1} \leq \frac{(n-1)s^2}{\sigma_0^2} \leq \mathcal{X}^2_{\frac{\alpha}{2},n-1}\} = 1 - \alpha$

Uguaglianza di varianza :
$X_1 \dots X_n \qquad Y_1 \dots Y_n \quad H_0 : \sigma_x^2 = \sigma_y^2 \qquad H_a : \sigma_x^2 \neq \sigma_y^2$

$S_x^2 - S_y^2 \quad Ts = \frac{\frac{S_x^2}{\sigma_x^2}}{\frac{S_y^2}{\sigma_y^2}} = \frac{S_x^2}{S_y^2}$

$\frac{S_x^2}{S_y^2} \sim F_{n-1,m-1} \qquad Pr\{F_{1-\frac{\alpha}{2},n-1,m-1} \leq \frac{S_x^2}{S_y^2} \leq -F_{1-\frac{\alpha}{2},n-1,m-1}\}$
Non rifiuto se soddisfa la disuguaglianza

Test parametro Bernoulli (Var discrete.)
$H_0 : p \leq p_0 \quad H_a ip > p_0$

$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$ *n campioni (Bernoulli)*

Binomiale $\sim$ **Gaussiana** (quando n è grande)
X n eventi favorevoli
$\mathbb{E}[X] = np \qquad Var(X) = np(1-p) \quad \mathcal{N}(np, np(1-p))$

Esempio Difetti di fabbricazione:
n = 300 $H_o ip \leq p_0 \quad p_0 = 2\%$
X = 10 n difetti

$$\frac{X - np}{\sqrt{nP_0(1-p_0)}} = \frac{10 - 300 \cdot 0.02}{\sqrt{300 \cdot 0.02 \cdot 0.98}} = 1.65$$

$$Pr\{z > 1.65\} = 0.0495$$

1.5 Modelli previsionali

1.5.1 Modelli di regressione previsionale

$Y_i = \alpha + \beta x_i + e_i \qquad e_i \sim \mathcal{N}(0, \sigma^2)$

Problema $\{x_i, y_i\}_{i=1}^n \quad \alpha, \beta =?$

Sum of square -> SS
$SS = \sum_i^n (y_i - \alpha + \beta x_i)^2$*Dove B e A -> var aleatoria*

$$\begin{cases} \dfrac{dSS}{dA} = -2\sum_i^n (y_i - A - Bx_i) = 0 \\ \dfrac{dSS}{dB} = -2\sum_i^n (y_i - A - Bx_i)^2 x_i = 0 \end{cases}$$

$$\begin{cases} \sum_i^n y_i = nA + B\sum_i x_i \\ \sum_i^n x_i y_i = n\sum_i^n + B\sum_i^n x_i \end{cases}$$

$$A = \frac{1}{n}\sum_i y_i - B\frac{1}{n}\sum x_i = \overline{y} - \beta\overline{x}$$

1.5.2 Regressione lineare

$y = \alpha + \beta x \quad e \sim (0,1) \quad y_i = A + \beta x$

$\mathbb{E}[B] = \beta \qquad \mathbb{E}[A] = \alpha$
$Var[B] = \frac{\sigma^2}{\sum_i x_i^2 - n\overline{x}} \qquad Var[A] = \frac{\sigma^2 \sum x_i^2}{n(\sum x_i^2 - n\overline{x}^2)}$

$SS_R = \sum_i (y_i - (A + Bx_i))^2$ (Somma dei quadrati dei residui)

$\frac{SS_R}{\sigma^2} \sim \mathcal{X}^2_{n-2}$

$\mathbb{E}\left[\frac{SS_R}{\sigma^2}\right] = n-2 \qquad \mathbb{E}\left[\frac{SS_R}{n-2}\right] = \sigma^2$

MLE :

$$f_{y_1\ldots y_n}(y_1 \ldots y_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\sum(y-(\alpha+\beta x_i 0))^2/2\sigma^2}$$

MSE = MLE

Notazione :

$$\begin{aligned} S_{xy} &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \ldots = \sum_i x_i y_i - n\bar{x}\bar{y} \\ S_{xx} &= \sum_i (x_i - \bar{x})^2 = \ldots = \sum_i x_i^2 - n\bar{x} \\ S_{yy} &= \sum_i (y_i - \bar{y})^2 = \cdots = \sum_i y_i^2 - n\bar{y} \end{aligned} \tag{2}$$

S_{xy} (Dispersione di x e y) S_{xy} (Dispersione di x)

S_{xy} (Dispersione di y)

$A = \overline{y} - B\overline{x}$ $B = \frac{S_{xy}}{S_{xx}}$ $SS_R = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}}$

Inferenza su β $\frac{B-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim \mathcal{N}(0,1)$ $\frac{SS_R}{\sigma^2} \sim \mathcal{X}^2_{n-2}$

$$\frac{\frac{B-\beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{-\sqrt{\frac{SS_R}{j^2(n-2)}}} \sim t_{2-2}$$

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) \sim t_{n-2}$$

$\beta \in B \pm \sqrt{\frac{SS_R}{(n-2)S_{xx}}} \quad t_{\frac{\alpha}{2},n-2} \rightarrow$ Livello di confidenza

Inferenza su α $\quad \frac{A-\alpha}{\sqrt{\frac{\sigma^2 \sum x_i^2}{nS_{xx}}}} \sim \mathcal{N}(0,1) \qquad \frac{SS_R}{\sigma^2} \sim \mathcal{X}^2_{n-2}$
coeffieciente della retta:

$\alpha \in A \pm \frac{SS_R \sum x_i^2}{\sqrt{n(n-2)S_{xx}}} \sim t_{\frac{\alpha}{2},n-2} \to$ Livello di confidenza

Interferenza su $\alpha + \beta x_0$:

$\mathbb{E}[A + Bx_0] = \mathbb{E}[A] + x_0\mathbb{E}[B] = \alpha + \beta x_0$
$Var(A + Bx_0) = \ldots = \sigma^2[\frac{1}{n} + \frac{(\overline{x}-x_0)^2}{S_{xx}}]$

Distribuzione $A + Bx_0$ **?**
$A + Bx_0 \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2[\frac{1}{n} + \frac{(\overline{x}-x_0)^2}{S_{xx}}])$
Stima di $\alpha + \beta x_0$

$$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{(\frac{1}{n} + \frac{(x_0-x)^2}{S_{xx}}(\frac{SS_R}{n-2}))}} \sim t_{n-2}$$

$$\alpha + \beta x_0 \in A + Bx_0 \pm t_{\frac{\alpha}{n},n-2}\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}(\frac{SS_R}{n-2})}$$

Piccolo se i punti sono vicini alla media

1.5.3 Regressione Lineare (e non)

$\{x_i, y_i\}_{i=1}^2 \leftarrow$ punti stocastici

Inferenza $\alpha + \beta x_0 = \mathbb{E}[y] \rightarrow$ non so niente del valore della y in quel punto
Inferenza $y_0 = y(x_0)\theta$

$$\alpha + \beta x_0 \in A + Bx_0 \pm t_{\frac{\alpha}{2},n-2}\sqrt{(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}})\frac{SS_R}{n-2}}$$

$\alpha + \beta \mathbf{x_0} \rightarrow$ Il punto x_0 che sta sulla retta $\alpha + \beta x_0$

Inferenza $y_0 = y(x_0) \rightarrow$ predittivo

$$y \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2)$$

$$A + Bx_0 \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}))$$

$$y_0 - (A + Bx_0) \sim \mathcal{N}(\sigma, \sigma^2(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}))$$

$$y_0 = y(x_0) = A + Bx_0 \pm t_{\frac{\alpha}{2},n-2}\sqrt{(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}})\frac{SS_R}{n-2}}$$

Coefficiente di determinazione

Definizione: La verifica dei miei valori

Formula generica: $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \rightarrow$ dispersione di y
La dispersione è data da due fattori:

- Retta (regressione)

- Rumore

$SS_R = \sum_{i=1}^{n}(y_i - (A + Bx_i))^2 \rightarrow$ Dipende dalla porzione non spiegata della retta
Utilizzo coefficienti di determinazione:

$$R^2 = \frac{S_{yy} - SS_R}{S_{yy}} = 1 - \frac{SS_R}{S_{yy}} \quad 0 \leq R^2 \leq 1$$

Se $\mathbf{R^2 = 1}$ la dispersione è data solo dalla retta *(regressione)*
Se $\mathbf{R^2 = 0}$ la dispersione è data solo dal *rumore*

La retta è migliore più R^2 è vicino a **1**

Coefficiente di correlazione

Formula generica:

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \ldots = 1 - \frac{SS_R}{S_{yy}} \rightarrow$ Dimostrazione matematica di R^2

Analisi dei residui $\quad y - (A + Bx) \rightarrow$ verifico tutti gli errori residui
Per la non linearità

Trasformazione lineare

$$W(t) = ce^{-dt}$$

Dove ce e $-dt$ sono *parametri*

$\log(W(t)) = \log(c) - dt \rightarrow$ Prob. soluzione al non lineare $y = \alpha + \beta x$

Rimedio al caso eteroschedastico :
$y_i = \alpha + \beta x_i + e_i \qquad e_i \sim \mathcal{N}(0, \sigma_i^2) \rightarrow$ errore in crescita x
$Var(e_i) = \dfrac{\sigma^2}{W_i} \sum W_i(y - (A + Bx_0))^2$

- Regressione lineare multipla
  - $\mathbf{y} = \beta_\mathbf{0} + \beta_\mathbf{1}\mathbf{x_1} + \beta_\mathbf{2}\mathbf{x_2} \ldots \beta_\mathbf{k}\mathbf{x_k} + \mathbf{e}$
  - $min \sum_i (y_i - (B_0 + B_1 x_{i1} + B_2 x_{i2} + \ldots + B_k x_{ik}))^2$
- Regressione (lineare) polinomiale
  - $y = \beta_0 = \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + e$
  - $\{\underline{x_i}, y_i\}_{i=1}^{n}$

$$\frac{\mathrm{d}}{\mathrm{d}B_0} = 0 = \sum_i (y_i - 1 - B_1 x_{i1} + B_2 x_{i2} + \ldots + B_k x_{ik})$$

$$\frac{\mathrm{d}}{\mathrm{d}B_1} = 0 = \sum_i x_{i1}(y_i - B_0 - B_1 x_{i1} + B_2 x_{i2} + \ldots + B_k x_{ik})$$

$$x^t x \underline{\beta} = x^t \underline{y} \Longrightarrow \underline{\beta} = (x^x x)^{-1} x^t \underline{y}$$

AN.O.VA (analysis of variance)
Analisi delle varianze / estensione del test di ipotesi sulle medie

Esempio voti medi degli anni scolastici

Anno.	**Voti medi.**
2020-2021 lockdown	μ_a
2021-2022 lockdown parziale	μ_b
2022-2023 presenza	μ_c

$H_0 : \mu_a = \mu_b = \mu_c$

1) stimatore di σ^2:

$$\sum_{i=1} \sum_{j} \frac{(x_{ij} - \mathbb{E}[x_{ij}])^2}{\sigma^2} = \sum_{i=1} \sum_{j=1} \frac{(x_{ij} - \mu_i)^2}{\sigma^2} \sim \mathcal{X}^2_{m \cdot n}$$

$$SS_W = \sum_{i=1} \sum_{j=1} \frac{(x_{ij} - x_{i*})^2}{\sigma^2} \sim \mathcal{X}^2_{m \cdot n - m}$$

$\mathbb{E}\left[\frac{SS_w}{\sigma^2}\right] = n \cdot m - m$

$\mathbb{E}\left[\frac{SS_w}{nm-m}\right] = \sigma^2$ stimatore 1

2) stimatore di σ^2 supponendo $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_m = \mu$

$$n \sum_{i=1}^{m} \frac{(x_{i*} - \mu)^2}{\sigma^2} \sim \mathcal{X}_m^2 \qquad x_{**} = \frac{\sum_{i=1}^{m} \sum_{i=1}^{n} x_{ij}}{m \cdot n}$$

$SS_b = n \sum_{i=1}^{m} (x_{i*} - x_{**})^2 \sim \mathcal{X}_{m-1}^2$
$\mathbb{E}\left[\frac{SS_b}{m-1}\right] = \sigma^2 \rightarrow$ Stimatore 2

Verifico stimatori :

$$Ts = \frac{SS_b/m-1}{SS_W/nm-m} \rightarrow \text{intorno a 1 va bene}$$

F Distribution: $F_{m-1}, mn - m, \alpha$

ANOVA :
Se i gruppi sono uguali : n camp $= n \cdot m$
Se sono diversi : n camp $= \sum_i n_i$

Life testing (Misura di affidabilità)
$x \geq 0|$ *tempo di vita* $\qquad \lambda(t) = \frac{f(t)}{1-F(t)}$
$f(t) =$ Densità di popolazione
$\lambda(t) =$ Intensità di rottura (failure rate)

$$\begin{aligned} P(x \in (t, t + \triangle t) | x > t) &= \frac{P(x \in (t, t + \triangle t), x > t)}{P(x > t)} \\ &= \frac{P(x \in (t, t + \triangle t))}{P(x > t)} \\ &\approx \frac{F(t) \triangle t}{1 - F(t)} \end{aligned}$$

Intensità di rottura

Definizione: Densità condizionale di probabilità che un oggetto funzionante almeno fino a t si guasti "subito dopo"

Formula generica:

$$\lambda(t) = \frac{F(t)}{1 - F(t)}$$

Se la distribuzione è esponenziale:

$$\lambda(t) = \frac{\lambda e^{-\lambda t}}{1 - 1 + e^{-\lambda t}} = \lambda \rightarrow \text{dove } \lambda \text{ è una costante}$$

Proprietà $\lambda(t) \Rightarrow F(t)$

$$\lambda(s) = \frac{f(s)}{1 - F(s)} = \frac{F'(s)}{1 - F(s)} = \frac{d}{dS}[-\log(1 - F(s))]$$

$$\int_0^t \lambda(s) = -\log(1 - F(s)) + \log(1 - F(s)) = 1 - F(t) = e^{-\int_0^t \lambda(s) ds}$$

Esempio Tasso di mortalità di un fumatore ($\lambda_{\mathbf{s}}$) e di un <u>non</u> fumatore ($\lambda_{\mathbf{n}}$)

$$\lambda_s(t) = 2\lambda_n(t)$$

$$
\begin{aligned}
&= P(\text{Non fumatore di età } \mathbf{A} \text{ vive fino a } \mathbf{B}) \\
&= P(\text{Non fumatore vive fino a } \mathbf{B} \mid \text{è vissuto fino } \mathbf{A}) \\
&= \frac{P(\text{Non fumatore viva fino a } \mathbf{B})}{P(\text{Non fumatore viva fino a } \mathbf{A})} \\
&= \frac{1 - F_N(B)}{1 - F_N(A)} \\
&= \frac{e^{-\int_0^B \lambda(t)\,dt}}{e^{-\int_0^A \lambda(t)\,dt}}
\end{aligned}
$$

Quindi:
$P(\text{Non fumatore di eta } \mathbf{A} \text{ vive fino a } \mathbf{B}) = e^{-\int_A^B \lambda(t)\,dt}$

Per i non fumatori invece:
$P(\text{Fumatore di eta } \mathbf{A} \text{ vive fino a } \mathbf{B}) = e^{-\int_A^B \lambda(t)\,dt} = Ps$

Dove $Ps = (Pn)^2 \rightarrow$ quindi la probabilità di soppravivenza del fumatore è uguale alla probabilità di soppravivenza del non fumatore *al quadrato*

Probabilità che un non fumatore arrivi ai 60 anni sapendo che è arrivato ai 50:

$$\lambda_N(t) = \frac{1}{20} \qquad 50 \leq t \leq 60$$

$P_N = e^{-\int_{50}^{60} \frac{1}{20}\,dt} = e^{-\frac{1}{20}(60-50)} = e^{-\frac{1}{2}} \approx 0.607 \approx 61\%$
$P_{\leq}(e^{-\frac{1}{2}})^2 = e^{-1} \approx 0.368 \approx 37\%$

Stima di affidabilità N oggetti che si possono guastare *indipendenti* tra di loro
Tempi di vita: $\lambda e^{-\lambda t}$ $\qquad \lambda = \frac{1}{\theta} \Rightarrow \frac{1}{\theta} e^{-\frac{t}{\theta}}$

Dati a disposizione: $x_1 \leq x_2 \leq x_3 \leq x_4 = r \quad i_1 = 2, i_2 = 3, i_3 = n, i_4 = 1$

Studio la variabile aleatoria X_i, i_j indica quale *oggetto si è guastato* per j-esimo all'istante x_j

(n-r) non si sono guastati ⇒ per questi $X_i > x_r$

$$fx_1, x_2 \dots x_r(x_1, x_2, \dots, x_r) = \prod_j^r \frac{1}{\theta} e^{-\frac{x_j}{\theta}} = \frac{1}{\theta^r} e^{-\frac{\sum_j x_j}{\theta}}$$

Ora per i non guasti:

$$P\left(X_j > x_j \text{ con } j \notin \{i_1, \dots, i_r\}\right) = \prod_{r+1}^n (1 - F_{X_j}(x_r)) = \left[1 - (1 - e^{\frac{-x_r}{\theta}})\right]$$

$$\log L = -r \log \theta - \frac{1}{\theta} \left[\sum_i^r x_i + (n-r)x_r\right]$$

$$\frac{d \log}{d\theta} = -\frac{r}{\theta} + \frac{1}{\theta^2} \left[\sum_i^r x_i + (n-r)x_r\right] = 0$$

$$-\theta r + \left[\sum_i^r x_i + (n-r)x_r\right] = 0$$

$$\hat{\theta} = \frac{\sum_i^r x_i + (n-r)x_r}{r} = \frac{t}{r} = \frac{TTT}{r} \rightarrow \text{Total Time Test}$$