www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

FORMULARIO STATISTICA

- A∪B = "Accada A, accada B o entrambi"
- A∩B = "Accada A in contemporanea a B"
- A^C = "Accade la negazione di A"

Frequenza= numero di volte con cui un dato appare

Ogni classe di intervallo ha un intervallo con ampiezza W	$w = \frac{V_{max} - V_{min}}{N_{classi}}$
Frequenza relativa	$F_{rel} = \frac{F}{F_{tot}}$
Percentuale	$N\% = F_{rel} \cdot 100$
Frequenza cumulata	$F_{c_n} = F_{n-1} + F_n$
Percentuale cumulata P_c	$P_{c_n} = P_{n-1} + P_n$
Media aritmetica	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Campo di variazione	$C_v = V_{max} - V_{min}$
Differenza interquartile (elimina outliers)	$D_{IQR} = Q_3 - Q_1$
Varianza popolazione	$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$
Varianza campione	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Scarto quadratico medio popolazione	$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}}$
Scarto quadratico medio campione	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
Coefficiente di variazione popolazione	$C_{V\,pop} = \left(\frac{\sigma}{\lvert\mu\rvert}\right) \cdot 100$
Coefficiente di variazione campione	$C_{V\,camp} = \left(\frac{s}{\lvert\bar{x}\rvert}\right) \cdot 100$
Media pesata	$\bar{x}_p = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$

Moda= valore più frequente

Mediana= valore centrale $\rightarrow m_1 = \frac{n+1}{2}$ se m_1 è dispari è il valore associato a quella posizione, se pari allora la mediana è la media dei due valori in posizione n/2 e ((n/2)+1).

PROBABILITÀ'

Approccio frequentista	$P(A) = \frac{N\ volte\ che\ si\ verifica\ A}{N\ ripetizioni\ dell'esperimento}$
Approccio combinatorio	$P(A) = \frac{N\ volte\ che\ si\ verifica\ A}{N\ eventi\ elementari}$
Assiomi Kolmogorov	$P(A) \in R\ t.c.\ 0 \leq P(A) \leq 1$ $P(\Omega) = 1$ A,B t.c. A∩B ≠ Ø allora $P(A \cup B) = P(A) + P(B)$
Regola della somma	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Regola della somma con eventi indipendenti	$P(A \cup B) = P(A) + P(B)$
Probabilità del complementare	$P(A^c) = 1 - P(A)$
Probabilità condizionata	$P(A\|B) = \frac{P(A \cap B)}{P(B)}$

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Per def. di eventi indipendenti la loro probabilità condizionata è:	$P(A\|B) = P(A)$ $P(B\|A) = P(B)$
Regola prodotto con eventi indipendenti	$P(A \cap B) = P(A) \cdot P(B)$
TEOREMA DI BAYES (eventi non specificato se indipendenti)	$P(A\|B) = \frac{P(B\|A) \cdot P(A)}{P(B\|A) \cdot P(A) + P(B\|A^c) \cdot P(A^c)}$

VARIABILI

Una variabile casuale è una funzione che associa ad ogni variabile descrittiva un numero.
Variabili Casuali Discrete: è una variabile discreta quando l'immagine di questa funzione X: Ω → ℝ ha al più un'infinità numerabile di valori [Im(X)⊆ ℝ].

p_j è la collezione distribuzione di probabilità	$p_j = P(X = x_j) = P(\{\omega \in \Omega \mid X(\omega) = x_j\})$
Media della variabile (n ∈ R oppure n=+∞)	$\mu_x = \sum_{j=1}^{n} x_j P(x_j) = \sum_{j=1}^{n} x_j p_j$
Varianza della variabile	$\sigma^2 = \sum_{j=1}^{n} (x_j - \mu_x)^2 . P(X = x_j) = \sum_{j=1}^{n} (x_j - \mu_x)^2 \cdot p_j$
Deviazione standard della variabile	$\sigma = \sqrt{\sum_{j=1}^{n} (x_j - \mu_x)^2 \cdot p_j}$

1. La variabile di Bernoulli, o a distribuzione Bernoulliana, mi da due opzioni: 0 se insuccesso, 1 se successo con relativa probabilità, per zero avrò (1-p) e per 1 avrò p. (N.B. p∈(0,1)
   La medie della variabile → $\mu_x = p$
   La varianza → $\sigma^2 = p(1-p)$ La deviazione standard → $\sigma = \sqrt{p(1-p)}$

Nota: media della somma è somma delle medie. Se gli eventi sono indipendenti alche la somma della varianza è la varianza della somma (se dipendenti non vale).

2. La variabile binomiale prendiamo n varibili tali che ognuna sia di Bernoulli ma per definizione è: X binomiale B(n,p) se
   $X = \sum_{i=1}^{n} X_i$ Dove $X_i \sim Bern(p)$
   Inoltre se dovessimo chiederci la probabilità di quando X=k introduciamo il coefficiente binomiale e avremo $P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$
   Dove $\binom{n}{k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} = \frac{n!}{k! \cdot (n-k)!}$ è chiamato coefficiente binomiale.
   La media della variabile → $\mu_x = np$ La varianza → $\sigma^2 = np(1-p)$

3. La variabile di Poisson è una variabile con lo scopo di dare un modello per contare l'accadere di un evento in una cera area (geografica, temporale, spaziale, …). Si indica con X ~ Po(μ).
   $$P(X = i) = \frac{\mu^i \cdot e^{-\mu}}{i!}$$
   La media della variabile → $\mu_x = \mu$ La varianza → $\sigma^2 = \mu$
   μ è l'intensità accadere evento (numero eventi che usiamo/totale dei fattori (n bombe/n regioni))
4. La variabile casuale geometrica descrive il "tempo d'attesa" ovvero il numeri di tentativi prima che si verifichi un successo. Anch'essa è bernoulliana ovvero $X_i \sim Bern(p)$ e ogni X_i indipendente.
   $P(X = i) = (1-p)^{i-1} \cdot p$ Dove i indica il numero dei tentativo della vittoria. Questa prende il nome di serie geometrica e svolgendo i calcoli troviamo che:
   La media della variabile → $\mu_x = \frac{1}{p}$ La varianza → $\sigma^2 = \frac{1-p}{p^2}$

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Variabili Casuali Continue

Variabile casuale continua caratterizzata da densità di probabilità f: (a,b) →ℝ continua e tale che ∀x ho f(x)≥0 e anche $\int_a^b f(x)dx = 1$. Avremo quindi che

Media: $\mu_x = \int_a^b x \cdot f(x)dx$ Varianza: $\sigma_x^2 = \int_a^b (x-\mu_x)^2 \cdot f(x)dx$

1. ***Variabile casuale uniforme***
   Si scrive X~ U[a,b] se la densità f(x) sarà $f(x) = \frac{1}{(b-a)}$ ∀x ∈ [a,b] e f(x)=0 ∀x ∉ [a,b].
   La medie è il punto medio dell'intervallo ovvero: $\mu_x = \frac{b+a}{2}$
   La varianza coincide invece con $\sigma_x^2 = \frac{1}{12} \cdot (b-a)^2$

2. ***Variabile casuale con distribuzione normale***
   X~N(0,1) se $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
   - Normale Standard
     Si ottiene quando μ=0 e σ²=1 ed è centrata nell'origine
   - Come passare da una normale a quella standard?
     Se dovessi avere una X ~ N(μ, σ²) definisco $Z = \frac{X-\mu}{\sigma}$ tale che Z ~ N(0,1)
     P(a ≤ X ≤ b) = P(a-μ ≤ X-μ ≤b-μ)= P($\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}$) = P($\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}$) ma ora la parte centrale è proprio Z che è nota (ho le tavole).

3. ***Variabile Esponenziale***
   X~E(λ) quando densità è f è definita su [0, +∞] tale che $f(x) = \lambda e^{-\lambda x}$ con $\lambda > 0$. Anche questa verificherà il tempo d'attesa per un arrivo.

   Media: $\mu_x = \frac{1}{\lambda}$ Varianza: $\sigma_x^2 = \frac{1}{\lambda^2}$

Indici di posizione

Percentile P_k: valore tra un gruppo di dati ordinati che lascia a sx il k% di dati e a dx il (n-k)%.
Definiamo L come $L = n \cdot \frac{k}{100}$ con k>0 intero. Se L è intero il percentile P_k è la media tra il dato in posizione L e L+1. Se L non è intero il percentile P_k è dato dal valore nella posizione della parte intera di L sommando 1 $P_k = x_{[L]+1}$. Per comodità si indicano $P_{25} = Q_1$, $P_{50} = Q_2$ e $P_{75} = Q_3$.

Gli Outliers

- Outlier debole x se $x \leq Q_1 - 1{,}5 \cdot (Q_3 - Q_1)$ oppure $x \geq Q_1 + 1{,}5 \cdot (Q_3 - Q_1)$;
- Outlier forte x se $x \leq Q_1 - 3 \cdot (Q_3 - Q_1)$ oppure $x \geq Q_1 + 3 \cdot (Q_3 - Q_1)$.

Intervalli di confidenza

Prendo campione variabili di eguale distribuzione (tutte media μ e varianza σ²) indipendenti. Definiamo $\bar{X}$ lo stimatore e $\bar{x}$ la stima della media. Mentre $\bar{S}$ è lo stimatore e $\bar{s}$ la stima della varianza.

Stimatore $\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i \rightarrow$ Stima $= \frac{1}{n} \cdot \sum_{i=1}^n x_i$

Stimatore $\bar{S} = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2 \rightarrow$ Stima $\bar{s} = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})^2$

(Per le proporzioni) Se cercassimo quanto la stima differenzi dalla media mera possiamo fare un esempio con una Benoulliana, se riprendiamo la Z ~ N(0,1) (con $\hat{p}$ la stima e p è la media vera) troviamo che $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

In questo caso l'intervallo di confidenza $IC_\alpha = \left[\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}};\ \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}\right]$ e l'errore è $\varepsilon = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$.

... per media assumendo varianza e deviazione standard note

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

Fisso livello di confidenza (1-α) definito come $P = (\bar{x} - \varepsilon \leq \mu \leq \bar{x} + \varepsilon)$. Sfruttando solita Z trovo che $\varepsilon = z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$. svolto l'esperimento ho $\bar{x}$ perciò posso calcolare $IC_{1-\alpha} = [\bar{x} - \varepsilon;\ \bar{x} + \varepsilon]$

… per media assumendo con deviazione standard incognita

Sfrutto una variabile T~T-student con (n-1) gradi di libertà. Anche qui fissato un livello di confidenza (1-α) trovo che $1 - \alpha = P\left(\bar{X} - t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right) \rightarrow 1 - \alpha = P\left(\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right)$ capisco che

$\varepsilon = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$ e quindi $IC_{1-\alpha} = [\bar{x} - \varepsilon; \bar{x} + \varepsilon]$.

… per la varianza

Qui abbiamo che $\frac{(n-1)\cdot S^2}{\sigma^2} \sim \chi^2_{n-1}$ è una distribuzione asimmetrica. Definisco allora due punti:

- χ^2_R tale che $P(\chi^2_{n-1} \leq \chi^2_R) = 1 - \frac{\alpha}{2}$
- χ^2_L tale che $P(\chi^2_L \leq \chi^2_{n-1}) = \frac{\alpha}{2}$

In questo caso abbiamo che $IC^{\sigma^2}_{1-\alpha} = \left[\frac{s^2\cdot(n-1)}{\chi^2_R}; \frac{s^2\cdot(n-1)}{\chi^2_L}\right]$

TEST D'IPOTESI

La mia ipotesi di partenza è l'ipotesi nulla o H_0, valuto l'ipotesi alternativa H_1. Devo fare in modo che $P(rigetto\ H_0 | H_0\ vera) = \alpha$ con α valore molto vicino a 0.
Affermazione → Campione Casuale semplice → Studio → Rigetto o mantengo H_0
Fisso un livello di significatività che andrà a definirmi elle aree di rigetto

	Mantengo H_0	Rigetto H_0
H_0 vera	OK	Prima specie
H_0 falsa	Seconda specie	OK

Verifica di ipotesi su μ

- σ nota sfrutto $Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
- σ non nota sfrutto T= $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \sim T - student$

… con σ nota

Una volta fatto il campionamento trovo $z_{oss} = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$, considero Z una N(0,1) e introduciamo il P-value (P_v).

So che $\mu=\mu_0$		
H_1: $\mu>\mu_0$	H_1: $\mu<\mu_0$	H_1: $\mu \neq \mu_0$
Area di rigetto $\alpha = P(Z \geq z_\alpha) = P\left(\bar{x} \geq -\mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right)$	Area di rigetto $\alpha = P(Z \leq z_\alpha) = P\left(\bar{x} \leq \mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right)$	Aree di rigetto $\alpha = P\left(\bar{x} \leq \mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\ \text{v}\ \bar{x} \geq -\mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right)$
$P_v = P(Z > z_{oss})$	$P_v = P(Z < z_{oss})$	$P_v = 2P(Z > \lvert z_{oss}\rvert)$
Rigetto H_0 se $P_v < \alpha$	Rigetto H_0 se $P_v < \alpha$	Rigetto H_0 se $P_v < \alpha$

… con σ non nota

Sfruttiamo lo stimatore $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2$ e quindi anche al posto di Z utilizzeremo la T~T-student $T = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}$ mentre $t_{oss} = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$

So che $\mu=\mu_0$		
H_1: $\mu>\mu_0$	H_1: $\mu<\mu_0$	H_1: $\mu \neq \mu_0$
Rigetto H_0 se $t_{oss} > t_{\alpha_{n-1}}$	Rigetto H_0 se $t_{oss} < t_{\alpha_{n-1}}$	Rigetto H_0 se $t_{oss} < -t_{\alpha_{n-1}}$ oppure $t_{oss} > t_{\alpha_{n-1}}$
$P_v = P(T_{n-1} > t_{oss})$	$P_v = P(T_{n-1} < t_{oss})$	$P_v = 2P(T_{n-1} > \lvert t_{oss}\rvert)$
Rigetto H_0 se $P_v < \alpha$	Rigetto H_0 se $P_v < \alpha$	Rigetto H_0 se $P_v < \alpha$

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari

TEST D'IPOTESI SU DUE CAMPIONI

1) ***Medie delle popolazioni***:
   a. Campioni dipendenti (stesso gruppo prima e dopo)
   b. Campioni indipendenti (campioni separati poi confrontati)
      - σ_x^2 e σ_y^2 note;
      - σ_x^2 e σ_y^2 non note.

2) ***Proporzioni delle popolazioni***

Per la ***proporzione dei campioni*** uso una variabile (statistica test) chiamata variabile aleatoria (Z).

Prendendo gli stimatori (con x e y): $\widehat{P_x} = \frac{1}{n} \cdot \sum_{i=1}^{n_x} X_i$ e le stime $\widehat{p_x} = \frac{1}{n} \cdot \sum_{i=1}^{n_x} x_i$ allora avremo che

$$Z = \frac{(\widehat{P_x} - \widehat{P_y}) - (P_x - P_y)}{\sqrt{\frac{P_x(1-P_x)}{n_x} + \frac{P_y(1-P_y)}{n_y}}}$$

In quanto assumiamo H_0 (vera $p_x - p_y = 0$) allora $Z = \frac{(\widehat{p_x} - \widehat{p_y})}{\sqrt{\frac{p_0(1-p_0)}{n_x} + \frac{p_0(1-p_0)}{n_y}}}$ con $\widehat{p_0} = \frac{n_x\widehat{p_x} + n_y\widehat{p_y}}{n_x + n_y}$

Regole di decisione		
Test unilaterale sinistro	Test unilaterale destro	Test bilaterale
$H_0: p_x - p_y \geq 0$ $H_1: p_x - p_y < 0$	$H_0: p_x - p_y \leq 0$ $H_1: p_x - p_y > 0$	$H_0: p_x - p_y = 0$ $H_1: p_x - p_y \neq 0$
Rigetto H_0 se $z_{oss} < -z_\alpha$	Rigetto H_0 se $z_{oss} > z_\alpha$	Rigetto H_0 se $z_{oss} < -z_{\frac{\alpha}{2}}$ o $z_{oss} > z_{\frac{\alpha}{2}}$

Medie delle popolazioni (campioni indipendenti)

1. Note σ_x^2 e σ_y^2, abbiamo gli stimatori delle medie $\bar{X}$ e $\bar{Y}$ e la variabile aleatoria $Z = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$

Regole di decisione		
Test unilaterale sinistro	Test unilaterale destro	Test bilaterale
$H_0: \mu_x - \mu_y \geq 0$ $H_1: \mu_x - \mu_y < 0$	$H_0: \mu_x - \mu_y \leq 0$ $H_1: \mu_x - \mu_y > 0$	$H_0: \mu_x - \mu_y = 0$ $H_1: \mu_x - \mu_y \neq 0$
Rigetto H_0 se $z_{oss} < -z_\alpha$	Rigetto H_0 se $z_{oss} > z_\alpha$	Rigetto H_0 se $z_{oss} < -z_{\frac{\alpha}{2}}$ o $z_{oss} > z_{\frac{\alpha}{2}}$

2. Non sono note σ_x^2 e σ_y^2 ma che sappiamo essere diverse.
   Sfruttiamo la T-student $T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$. Per calcolarne i gradi di libertà a mano prendiamo il minimo tra $(n_x\text{-}1)$ e $(n_y\text{-}1)$. Le regole di decisione rimangono uguali.

Medie delle popolazioni (campioni dipendenti)

Ho campioni con misure ripetute ($n_x = n_y$). Definiamo $d_i = (x_i - y_i)$. Abbiamo ancora una statistica test data da una T-student $T = \frac{\bar{d} - d_0}{\frac{s_d}{\sqrt{n}}}$ e il t_{oss} è uguale a T ma senza il fattore d_0.

www.unidocs.it - Appunti e dispense per superare i tuoi esami universitari