

Probabilità e Statistica per l'Informatica

UniShare

Davide Cozzi
@dlcgold

Gabriele De Rosa
@derogab

Federica Di Lauro
@f_dila

Indice

Capitolo 1

Introduzione

Questi appunti sono presi a lezione. Per quanto sia stata fatta una revisione è altamente probabile (praticamente certo) che possano contenere errori, sia di stampa che di vero e proprio contenuto. Per eventuali proposte di correzione effettuare una pull request. Link: <https://github.com/dlclgold/Appunti>.

Grazie mille e buono studio!

La statistica è una disciplina, basata sulla matematica, con finalità lo studio quantitativo e qualitativo di un particolare fenomeno collettivo, in condizioni di incertezza o non determinismo ed è usata in molti ambiti, come ad esempio l'intelligenza artificiale, data science, robotica, domotica e tutte le analisi per poter ottenere ricavare delle informazioni sui dati.

Si ha l'*A-B testing*, per decidere tra due scelte la migliore e per la decisione si analizzano i dati presi da campioni di popolazione, utilizzando il *tasso di conversione*, ossia la percentuale di visitatori unici che hanno effettuato la azione su cui si sta effettuando il test.

In questo corso verranno affrontati e studiati i seguenti argomenti:

1. statistica descrittiva
2. calcolo delle probabilità
3. distribuzioni notevoli
4. teoremi di convergenza
5. stima dei parametri
6. test di ipotesi parametrici
7. test di ipotesi non parametrici

8. regressione lineare

Capitolo 2

Statistica Descrittiva

La statistica descrittiva è una raccolta di metodi e strumenti matematici usati per organizzare una o più serie di dati al fine di trovarne delle simmetrie, periodicità o delle eventuali leggi.

Solitamente i dati disponibili non rappresentano tutta la popolazione ma un numero limitato di osservazioni effettuato su un *campione*, sottoinsieme selezionato della popolazione su cui si effettua l'analisi statistica, la cui efficacia dipende da quale sottoinsieme è stato scelto, infatti non esiste un solo campione ma vi sono diversi modi per sceglierli, più o meno efficaci, per l'analisi statistica.

Quando si effettua un'analisi statistica si vuole affermare qualcosa riguardo i *caratteri* della popolazione, ossia gli elementi su cui si effettua l'analisi statistica, che possono essere:

- *caratteri qualitativi*, indicanti qualità (colori, stili, materiali etc...) e anche dati non numerici in cui solitamente non è definita una *relazione d'ordine*
- *caratteri quantitativi*, maggiormente studiati dal corso, dati numerici in cui vengono definite *relazioni d'ordine*, che possono essere a loro volta divisi in *discreti*, indicanti valori in \mathbb{Z} , e *continui*, con valori nel campo \mathbb{R} .

Supponiamo di considerare n elementi della popolazione e di rilevare, per ognuno di essi, il dato relativo al carattere quantitativo da esaminare, ossia definiamo l'insieme di dati $E = \{x_1, x_2, \dots, x_n\}$ con la numerosità, il numero di elementi considerati, pari a n .

In caso il carattere è discreto è comodo raggruppare i dati considerando l'insieme di tutti i valori assumibili, detta *modalità del carattere* ed associare ad ognuno di esso il numero di volte che esso compare in E .

Si ha quindi il numero di totalità N del carattere e si definisce l'insieme di modalità $S = \{s_1, \dots, s_N\}$ su cui si definiscono i seguenti valori statistici:

frequenza assoluta numero di volte f_j che si ha un elemento di un campione

frequenza cumulata assoluta somma delle frequenze assolute di tutte le modalità, indicato con F_j , calcolato con la seguente formula:

$$F_j = \sum_{k:s_k \leq s_j} f_k$$

frequenza relativa rapporto tra la frequenza assoluta e il numero di elementi, indicata con p_j , calcolata come

$$p_j = \frac{f_j}{n}$$

frequenza cumulativa relativa P_j somma delle frequenze relative di tutte le modalità, indicato con P_j , calcolata come

$$P_j = \sum_{k:s_k \leq s_j} p_k$$

Si definisce *distribuzione di frequenza* una funzione $F : S \rightarrow \mathbb{N}$ che associa ad ogni modalità la corrispondente frequenza, per cui esiste la distribuzione di frequenza assoluta, relativa, frequenza cumulativa assoluta e relativa.

Quando il carattere da studiare è continuo o discreto con un gran numero di valori, è conveniente ricondursi a raggruppamenti come quelli appena trattati, per cui si suddivide l'insieme delle modalità S , in alcune classi, sottoinsiemi di S , che formano una partizione.

La scelta delle classi con cui si suddivide l'insieme S è del tutto arbitraria anche se è necessario che esse formino una partizione di S .

Le partizioni devono essere significative e sufficientemente numerose ed inoltre ad ogni classe si associano le grandezze: confini superiori ed inferiori, l'ampiezza e il valore centrale della classe.

Nel caso in cui il carattere esaminato sia continuo occorre specificare come le classi sono chiuse, a destra o a sinistra, ossia specificare se gli elementi dell'indagine il cui dato coincide con il confine della classe sono da raggruppare all'interno della classe stessa oppure no.

2.1 Indici di tendenza Generale

Fino ad ora abbiamo visto come rappresentare i dati, ora iniziamo ad analizzare gli indici che ci forniscono un valore che rappresenta un certo aspetto della serie di dati, incominciando dagli *indici di tendenza generale*:

media è la media aritmetica tra tutti i valori dei dati osservati

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + \cdots + x_n}{n}$$

Considerando le distribuzioni di frequenza definite, possiamo fornire definizioni equivalenti di media:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n s_j f_j = \sum s_j p_j$$

La dimostrazione dell'uguaglianza di queste definizioni alternative è banale e si riconduce alla definizione di frequenza relativa ed assoluta.

mediana è l'elemento in mezzo ai valori dei dati, ordinati in maniera crescente in cui se il numero degli elementi n è dispari è l'elemento $\frac{n+1}{2}$ altrimenti è la media degli elementi di posto $\frac{n}{2}$ e $\frac{n}{2} + 1$.

moda valore o classe, indicato con \tilde{x} , corrispondente alla massima frequenza assoluta e viene usata solitamente in caso sia impossibile definire la media e la mediana.

La moda non è unica infatti parliamo di *distribuzione uni-modale*, nel caso di un'unica moda, altrimenti di *distribuzione multi-modale*.

Gli indici di tendenza centrale non sono utili per fornire informazioni circa l'omogeneità dei dati, in quanto forniscono informazioni sui valori centrali e medi del campione statistico, per cui per risolvere sto problema introduciamo i seguenti indici:

varianza è la media dello scarto quadratico di ogni elemento dalla sua media

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

La varianza ovviamente è tanto più grande quanto i singoli elementi si discostano dalla media, ossia significa che i dati in tal caso sono molto disomogenei.

Come abbiamo già visto per la media sono presenti le seguenti definizioni alternative di varianza:

$$s^2 = \frac{1}{n} \sum_{j=1}^N f_j (s_j - \bar{x})^2$$

$$s^2 = \sum_{j=1}^N p_j (s_j - \bar{x})^2$$

$$s^2 = \sum_{j=1}^n x_j - \bar{x}^2$$

Le prime due definizioni alternative derivano dalla definizione di frequenza assoluta e frequenza mentre l'ultima proviene da passaggi algebrici, dimostrati di seguito formalmente:

Dimostrazione.

$$\begin{aligned} s^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n} \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n} (\sum x_i^2 - 2\bar{x} \sum x_i + \sum \bar{x}^2) \\ &= \frac{1}{n} (\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2) \\ &= \frac{1}{n} \sum x_i^2 - \bar{x}^2 \end{aligned}$$

Si dimostra $\sum x_i = n\bar{x}$ in quanto $\bar{x} = \frac{1}{n} \sum x_i$ e il resto sono soltanto passaggi algebrici elementari \square

scarto quadratico medio misura quanto sono distanti gli elementi di un campione ed è calcolata come:

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

Nel calcolo della varianza si utilizza il quadrato per la differenza tra l'elemento e la sua media in quanto si ha $\sum (x_i - \bar{x}) = 0$, dato che la media è il valore

la cui distanza è minima tra tutti gli elementi del campione.

Per evitare questo problema si eleva la differenza tra un elemento e la sua media al quadrato.

La varianza è definito come il momento secondo rispetto alla media, come analizzeremo nel capitolo 3, espresso tramite la formula:

$$M_{k,y} = \frac{1}{n} \sum (x_i - y)^2$$

Fino ad ora noi abbiamo considerato il caso unidimensionale ma molte analisi richiedono di analizzare due o più caratteri del campione contemporaneamente, per riconoscere leggi ed analogie tra i diversi caratteri.

Considereremo solo due caratteri contemporanei, sia perché un'analisi con più caratteri si ha gli stessi aspetti e soprattutto per evitare di aggravare troppo la rappresentazione dei dati e assumiamo inoltre che entrambi i caratteri sono di tipo quantitativo e discreto, in quanto se fossero quantitativi continui subirebbero prima un raggruppamento a classi.

L'insieme dei dati viene rappresentato come l'insieme delle coppie $E = \{(x_i, y_i) \mid \forall i \leq n\}$ mentre l'insieme delle coppie di valori assumibili sono rappresentati con l'insieme

$$S = \{(s_j, u_k), j = 1 \dots N \quad k = 1 \dots M\}$$

Come abbiamo fatto anche per il caso unidimensionale definiamo le seguenti quantità:

frequenza assoluta è la quantità f_{jk} corrispondente al numero di elementi con valore (s_j, u_k)

frequenza relativa rapporto tra la frequenza e il numero di elementi, calcolato come

$$p_{jk} = \frac{f_{jk}}{n}$$

frequenza cumulata assoluta somma delle frequenze assoluta, calcolata come segue

$$F_{jk} = \sum_{r:s_r \leq s_j; l:u_l \leq u_k} f_{rl}$$

frequenza cumulata relativa somma delle frequenze relative, calcolata come segue

$$P_{jk} = \sum_{r:s_r \leq s_j; l:u_l \leq u_k} p_{rl}$$

frequenza

Si definisce *distribuzione di frequenza doppia* una qualsiasi funzione f, F, p, P che associa ad ogni coppia (s_j, u_k) la corrispondente frequenza ma non esistono solo queste funzioni, infatti noi vediamo anche le *distribuzioni marginali*, in cui si analizza la distribuzione dei singoli caratteri, presi indipendentemente dagli altri.

Le distribuzioni marginali hanno la definizione delle seguenti funzioni:

frequenza assoluta marginale quantità di elementi f_{xj} data dagli elementi di E , il cui primo carattere ha valore s_j

frequenza relativa marginale rapporto tra la frequenza assoluta marginale e il numero di osservazioni n .

frequenza cumulata assoluta marginale F_{xj} somma delle frequenze assolute marginali di tutti gli s_k con $s_k \leq s_j$

frequenza cumulata relativa marginale P_{xj} somma delle frequenze relative marginali di tutti gli s_k con $s_k \leq s_j$

Oltre a quelli definiti fino ad ora, esiste un indice che fornisce un grado di interdipendenza tra i due caratteri, importante in quanto molti problemi concreti necessitano di analizzare gradi di correlazione tra due o più serie di dati, iniziando prima di tutto da un esempio.

Considerando due serie $\{x_i\}$ e $\{y_i\}$, con $i = 1 \dots n$, e le coppie di scarti $x_i - \bar{x}$ e $y_i - \bar{y}$, di tutti i valori della serie rispetto alla media, si ha una relazione di dipendenza tra i due caratteri se i due scarti corrispondono sistematicamente o quasi valori positivi o negativi.

Si definisce quindi la **covarianza** c_{xy} , dei dati o campionaria, come

$$c_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

La covarianza assume un valore positivo (negativo), che diviene grande in valore assoluto, nel caso in cui i termini prodotto abbiano segni concordi e in questo caso si parla di serie statistiche fortemente correlate o per meglio dire di dati delle serie fortemente correlati.

Nel caso opposto vale a dire nel caso in cui i dati delle serie siano incorrelati avremo che i prodotti avranno segni diversi (saranno discordi in segno) e la covarianza, per come definita, risulterà piccola in valore assoluto, prossima al valore 0.

Si ha anche la seguente formula per la covarianza:

$$c_{xy} = \frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}$$

Dimostrazione.

$$\begin{aligned}
 c_{xy} &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\
 &= \frac{1}{n} \sum_{i=1}^n (x_i y_i - x_i \bar{y} - y_i \bar{x} + \bar{x} \bar{y}) \\
 &= \frac{1}{n} \left(\sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + \sum \bar{x} \bar{y} \right) \\
 &= \frac{1}{n} \left(\sum x_i y_i - n \bar{y} \bar{x} - n \bar{y} \bar{x} + n \bar{y} \bar{x} \right) \\
 &= \frac{1}{n} \left(\sum x_i y_i - n \bar{x} \bar{y} \right) \\
 &= \frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}
 \end{aligned}$$

□

Nel caso in cui i dati si riferiscano a caratteri quantitativi discreti, di cui è nota la distribuzione di frequenza doppia, è possibile utilizzare le seguenti formule per il calcolo della covarianza:

$$\begin{aligned}
 cxy &= \sum_1^N \sum_1^M (s_j - \bar{x})(u_k - \bar{y}) p_{jk} \\
 cxy &= \sum_1^N \sum_1^M s_j u_k p_{jk} - \bar{x} \bar{y}
 \end{aligned}$$

Date due serie di dati si ha le seguenti proprietà:

- le due serie di dati *statisticamente incorrelate* se la loro covarianza è nulla
- le due serie di dati sono *statisticamente indipendenti* se vale:

$$\forall j = 1, \dots, N \quad k = 1, \dots, M \quad p_{jk} = p_j p_k$$

con p_{jk} frequenza relativa doppia mentre le altre sono le frequenze relative marginali.

Si ha che la proprietà di indipendenza è più forte dell'incorrelazione, infatti se le due serie di dati sono indipendenti, risultano anche incorrelate mentre il contrario non è detto, infatti risulta:

$$\sum \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) \sum (y_i - \bar{y}) = 0$$

cosa che non è possibile leggerla al contrario.

Nel caso bidimensionale, con variabili x e y , la covarianza si può rappresentare attraverso una matrice 2×2 :

$$C = \begin{vmatrix} c_{xx} & c_{xy} \\ c_{xy} & c_{yy} \end{vmatrix} = \begin{vmatrix} \text{var}(x) & \text{cov}(x, y) \\ \text{cov}(x, y) & \text{var}(y) \end{vmatrix}$$

Per una misura indipendente dalla variabilità delle grandezze si usa la matrice di correlazione:

$$\text{Corr} = \begin{vmatrix} \frac{c_{xx}}{\sigma_x^2} & \frac{c_{xy}}{\sigma_x \sigma_y} \\ \frac{c_{xy}}{\sigma_x \sigma_y} & \frac{c_{yy}}{\sigma_y^2} \end{vmatrix} = \begin{vmatrix} 1 & \text{corr}(x, y) \\ \text{corr}(x, y) & 1 \end{vmatrix}$$

che ovviamente può crescere in m dimensioni.

2.2 Regressione Lineare

In molti casi ci si pone la questione se tra dei caratteri x ed y esista un legame/relazione di tipo funzionale che ne descriva in modo soddisfacente corretto il legame realmente esistente.

Si parla di un'*analisi di regressione*, in caso in cui si considera ad uno dei due caratteri come variabile indipendente e si cerca una funzione che stabilisce la relazione tra i due caratteri.

Se fisso x , come *variabile indipendente*, cerco $y = f(x)$ in modo che essa descriva al meglio il legame tra la variabile indipendente x e il carattere y , che a questo punto viene interpretato come *variabile dipendente*.

Si determina quindi la funzione f che minimizza le distanze tra i valori osservati del carattere y e quelli ottenibili se la relazione tra x e y fosse proprio quella descritta da f , quindi si cerca la funzione f che minimizza la quantità:

$$g(f) = \sum [f(x_i) - y_i]^2$$

dove il quadrato si utilizza affinché le distanze vengano tutte considerate con segno positivo.

Se f è vincolata ad essere una funzione lineare allora si parla di **regressione lineare**, con la retta rappresentata da $y = mx + q$, tale per cui risulti minima la quantità:

$$g(m, q) = \sum [mx_i + q - y_i]^2$$

con $mx_i + q = f(x_i)$ che sono l'approssimazione alle y_i mediante f .

Si ha che:

$$m = \frac{c_{xy}}{s_x^2}$$

$$q = \bar{y} - \frac{c_{xy}}{s_x^2} \bar{x}$$

Questo metodo consente di determinare la retta che meglio descrive la relazione tra i due caratteri senza peraltro fornire alcuna indicazione circa il grado di approssimazione che è in grado di offrire.

Per tale motivo è stata introdotta una nuova grandezza detta **coefficiente di correlazione lineare**:

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$

L'importanza di tale coefficiente deriva dal fatto che esso assume valori sempre appartenenti all'intervallo $[-1, 1]$ ed inoltre il coefficiente è nullo se le serie sono statisticamente incorrelate.

il valore assoluto risulta tendente a 1 se le coppie sono tutte sulla retta $y = mx + q$, quindi rappresenta il grado di allineamento delle coppie di dati.

Abbiamo accennato in precedenza al fatto che non si è sempre vincolati alla scelta di una retta tra le funzioni che possono descrivere la relazione tra le due serie di dati ma quanto esposto in precedenza può essere applicato anche nel caso in cui si considerino relazioni funzionali di diversa natura, la cui scelta può essere suggerita da una qualche impressione derivante da ispezioni visive dei dati o da altre forme di conoscenza circa il fenomeno analizzato, avendo quindi il modello non lineare di regressione.

Molte relazioni funzionali non lineari possono essere ricondotte a tali (lineari) con opportune trasformazioni delle variabili, infatti prendendo per esempio la relazione:

$$y = a \cdot e^{bx}$$

che si può riscrivere come:

$$\tilde{y} = \beta \cdot \tilde{x} + \alpha$$

con:

$$\tilde{y} = \log(y)$$

$$\tilde{x} = x$$

$$\alpha = \log(a)$$

$$\beta = b$$

si ottiene quindi una sorta di curva e non più una retta.

La determinazione dei coefficienti a e b che meglio permettono di approssimare una serie di punti $\{x_i, y_i\}$ può essere effettuata riconducendosi ad una regressione lineare ovvero determinando i coefficienti α, β che meglio approssimano, linearmente, la serie dei punti $\{\tilde{x}_i, \tilde{y}_i\}$, con:

$$\tilde{y}_i = \log(y_i)$$

$$\tilde{x}_i = x_i$$

Una volta determinati tali coefficienti il calcolo di a e b risulta immediato.

Ecco alcune funzioni riconducibili a lineari:

$$y = a \log(x) + b$$

$$y = ax^b$$

$$y = \frac{1}{a + b \cdot e^{-x}}$$

Capitolo 3

Calcolo delle probabilità

La probabilità è la disciplina di carattere matematico che permette di affrontare l'analisi delle situazioni che hanno un esito imprevedibile a priori e pertanto con conseguenze incerte ed è lo strumento di base della statistica, che invece trae conclusioni su una popolazione, utilizzando i dati osservati su una collezione di individui appartenenti alla popolazione, basandosi inoltre su considerazioni probabilistiche.

Si hanno 4 impostazioni per la probabilità:

1. *classica*, in cui la probabilità di un evento A è data dal rapporto tra i casi favorevoli e il numero di casi possibili, supponendo tutti i casi ugualmente possibili.
2. *frequentista*, in cui la probabilità di un evento è il limite del rapporto tra gli esperimenti favorevoli all'evento e il totali di quelli effettuati, supponendo di averli ripetuti nella stessa condizione.
3. *soggettivista*, in cui la probabilità di un evento è la misura del grado di fiducia che un individuo attribuisce al verificare dell'evento A e questa impostazione è basata su quanto uno scommettitore pagherebbe per il verificare dell'evento.
4. *assiomatica*

Iniziamo a considerare l'impostazione assiomatica, supponendo di voler studiare una situazione con un insieme Ω , detto *spazio campione*, di possibili esiti ben distinti tra loro, in cui tutti i suoi sottoinsiemi A sono eventi, elementari in caso contengono solo un elemento altrimenti sono composti, e ad ogni evento si ha una quantità numerica $P(A)$ detta *probabilità*.

Un passo fondamentale nel superare le polemiche sull'interpretazione del concetto di probabilità fu compiuto da Kolmogorov (1933) che abbandonò il tentativo di fondare la teoria della probabilità su una interpretazione sperimentale del concetto e costruire una teoria secondo una *impostazione assiomatica*, trascendendo il significato effettivo di probabilità, rinviandone l'interpretazione al momento delle applicazioni.

Due eventi sono *incompatibili* se non hanno elementi in comune, formando un'intersezione vuota e l'insieme delle parti di Ω , $\wp(\Omega)$, definisce ovviamente l'insieme di tutti i sottoinsiemi.

Si dice *misura di probabilità* ogni applicazione $P : \wp(\Omega) \rightarrow \mathbb{R}_0^+$ che associa un valore reale ad ogni sottoinsieme di Ω e per cui valgono le seguenti proprietà:

- esiste ed è un unico numero $P(A) \geq 0 \quad \forall A \subseteq \Omega$
- $P(\Omega) = 1$
- data la famiglia $\{A_i \in I \subseteq N\}$ di eventi incompatibili vale:

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i)$$

Ogni misura di probabilità è una funzione che assegna valori numerici a sottoinsiemi di Ω e non ai suoi elementi, come contrariamente si è portati a credere intuitivamente, infatti avviene che la funzione di misura di probabilità associa un valore agli eventi elementari come una conseguenza dell'assegnazione di valori ad eventi composti.

La definizione di probabilità non fornisce indicazioni su quali valori numerici devono essere assegnati dato che dipende, come sempre dal particolare problema da analizzare.

Dalla definizione assiomatica derivano facilmente le seguenti proprietà aggiuntive:

Definizione 1. *Sia P una misura di probabilità definita sull'insieme delle parti $\wp(\Omega)$ di uno spazio campione Ω allora:*

- $\forall A, B \subseteq \Omega$ eventi anche incompatibili risulta che $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Dimostrazione. Si dimostra che:

$$P(A \cup B) = P(A \cap \overline{B}) + P(A \cap B) + P(\overline{A} \cap B)$$

Si ricava e si sa che sono soddisfatte le seguenti equazioni

$$P(A) = P(A \cap \overline{B}) + P(A \cap B)$$

$$P(B) = P(A \cap B) + P(\overline{A} \cap B)$$

quindi si ricava dall'equazione iniziale, sostituendo $P(A)$ e $P(B)$ la seguente equazione:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

□

- $\forall A \subseteq \Omega \rightarrow P(\overline{A}) = 1 - P(A)$

Dimostrazione.

$$\forall A \subseteq \Omega \rightarrow P(\overline{A}) = 1 - P(A)$$

↓

$$1 = P(A \cup \overline{A}) = P(A) + P(\overline{A}) - P(A \cap \overline{A}) = P(A) + P(\overline{A}) - 0$$

□

- $\forall A \subseteq \Omega$ risulta che $P(A) \leq 1$
- $\forall A, B \subseteq \Omega, A \subseteq B$ risulta che $P(A) \leq P(B)$

Si ha che il terzo assioma della probabilità generalizza la definizione di $P(A \cup B)$, nel caso di tutti gli eventi incompatibili, mentre la terza e la quarta proprietà si dimostrano facilmente, considerando gli elementi basilari di teoria degli insiemi e dei 3 assiomi della probabilità.

Si ha uno spazi campione con elementi equiprobabili se dato $\Omega = \{1, 2, \dots, N\}$ tale che:

$$P(\{1\}) = \dots = P(\{N\}) = p$$

Da questa formula si ricava facilmente le seguenti due equazioni, fondamentali per il calcolo della probabilità:

$$P(\{1\}) + \dots + P(\{N\}) = Np = p(\Omega) = 1$$

$$P(\{i\}) = p = \frac{1}{n}$$

Da quest'ultima equazione, considerando anche il terzo assioma della probabilità si ricava che la probabilità di evento E è data dal rapporto tra i numeri degli eventi elementari dell'evento e la numerosità degli elementi dello spazio campione.

Si ha inoltre che la probabilità di un evento E è data dalla seguente formula

$$P(E) = \frac{\text{num. eventi elementari } E}{N}$$

In caso si facciano due esperimenti: se l'esperimento 1 può avere n possibili esiti equiprobabili, e l'esperimento 2 può avere m possibili esiti equiprobabili, allora i due esperimenti hanno $n \times m$ possibili esiti.

Se si espande a r esperimenti si avranno $n_1 \times n_2 \times \dots \times n_r$ possibili esiti, come tutti sappiamo dalla regola del prodotto del calcolo combinatorio.

3.1 Accenni di Calcolo Combinatorio

Definizione 2. Si definisce permutazione di n oggetti a_1, a_2, \dots, a_n un ordinamento degli n oggetti e il numero di permutazioni possibili in un insieme di n oggetti è uguale a $n!$.

Dato un insieme di n elementi, per calcolare il numero di modi di scegliere k elementi dall'insieme di n elementi, senza possibilità di ripetizione, vi sono due modi:

- *disposizione semplice*, quando è importante l'ordine di scelta degli elementi e viene calcolato come segue

$$D(n, k) = n * (n - 1) * (n - 2) * \dots * (n - k + 1)$$

Questa definizione è molto intuitiva, infatti il primo elemento può essere scelto tra n elementi, il secondo tra $n - 1$ elementi e così via fino ad arrivare a prelevare da $n - (k - 1)$ elementi.

- *combinazioni semplici*, quando l'ordine tra gli elementi scelti è irrilevante e si utilizza come simbolo $\binom{n}{k}$, che indica il numero di sottoinsiemi di k elementi scelti dall'insieme con cardinalità n .

Il coefficiente binomiale, come si dovrebbe conoscere da altri corsi, si calcola come

$$\binom{n}{k} = \frac{D(n, k)}{k!} = \frac{n!}{k!(n - k)!}$$

Dalla definizione e dal significato di $\binom{n}{k}$ appare chiaro che

$$\binom{n}{0} = 1 = \binom{n}{n} \quad \binom{n}{1} = n$$

Valgono inoltre anche le seguenti proprietà:

$$\binom{n}{k} = \binom{n}{n-k}$$

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$$

Quest'ultima formula ci permette di definire il triangolo di tartaglia, che tutti gli studenti scientifici ed informatici dovrebbero già conoscere, e di inoltre $\binom{n}{k}$ si chiama *coefficiente binomiale* perchè compaiono come coefficiente nella formula di Newton, per calcolare lo sviluppo di un binomio $(x+y)^n$, come si può notare

$$(x+y)^n = \sum_{i=0}^n \binom{n}{i} x^{n-i} y^i$$

3.2 Probabilità condizionata

In molti casi, quando si desidera studiare un fenomeno con comportamenti aleatori, oppure si vuole effettuare un esperimento con esiti imprevedibili a priori, si utilizzano alcune informazioni complementari per restringere il campo dei possibili risultati e per poter trattare matematicamente queste situazioni viene introdotto il concetto di *probabilità condizionata*.

Definizione 3. *Siano dati uno spazio campione Ω ed una misura di probabilità P definita su $\wp(\Omega)$ e secondo l'impostazione assiomatica di probabilità, considerati due eventi, A e B con $P(B) > 0$, viene definita probabilità dell'evento A condizionata dall'evento B la quantità:*

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

Questa definizione non è limitata solo alla probabilità assiomatica, infatti secondo l'approccio frequentista si definisce la probabilità dell'evento A , limitandosi a considerare solo i casi appartenenti a B come totale dei casi possibili.

Con la probabilità condizionata vi possono essere 3 diversi casi:

- la probabilità dell'evento B sotto condizionamento diviene maggiore rispetto a quella che avrebbe assunto senza condizionamento
- il valore di probabilità diminuisca a fronte del condizionamento
- può accadere che il condizionamento rispetto ad un evento non inficia in alcun modo la probabilità di un altro evento ed in sto caso i due eventi sono *stocasticamente indipendenti*.

Due eventi $A, B \in \wp(\Omega)$ sono stocasticamente indipendenti in caso in cui $P(A) = P(A|B) = P(B)$ da cui risultano i seguenti risultati:

- $P(A \cap B) = P(A)P(B)$
- $P(A \cap B) = P(A|B)P(B)$

L'ultima formula presentata può essere generalizzata ad n eventi, venendo chiamata *formula del prodotto*, nel seguente modo:

Definizione 4. Sia $n \in \mathbb{N}_+$ e data la famiglia $\{A_i | i = 1, \dots, n\}$ di sottoinsiemi di Ω , allora risulta

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1})$$

Considerata una partizione dello spazio campione Ω , si hanno le seguenti formule:

- $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$ (formula delle probabilità totali)
- $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}$ (formula di Bayes)

3.3 Variabili Aleatorie

In base a quanto visto fino ad ora per affrontare problemi di calcolo delle probabilità occorre di volta in volta definire in maniera appropriata lo spazio campione e la misura di probabilità.

Questo fatto comporta delle difficoltà se ci si pone come obiettivo la formulazione di una teoria generale della probabilità, in quanto spazio campione e misura di probabilità sono diversi ogni volta ed inoltre in quasi tutti i problemi concreti si ha a che fare con situazioni il cui esito, imprevedibile a priori, è di tipo numerico.

Queste considerazioni hanno portato i matematici ad introdurre delle opportune trasformazioni, chiamate *variabili aleatorie*, che consentono di ricondursi sempre ad \mathbb{R} come spazio campione ed a considerare quali suoi

sottoinsiemi tutti gli intervalli del tipo (a, b) o $[a, b]$ con $-\infty \leq a < b \leq +\infty$ in cui sono comprese tutte le possibili unioni ed intersezioni, finite o infinite, e i loro complementi.

Dato uno spazio campione Ω si definisce *variabile aleatoria o casuale* un'applicazione $X : \Omega \rightarrow \mathbb{R}$ che associa un numero reale ad ogni elemento di Ω .

In base a questa definizione è possibile assegnare delle probabilità ad eventi del tipo $X \in B \subseteq \mathbb{R}$ in quanto

$$P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\})$$

ma ciò appesantisce troppo la notazione per cui utilizziamo per comodità $P(X \in B)$ con però $X \in B$ un evento in Ω , invece di rappresentare la funzione della variabile aleatoria.

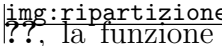
Dal punto di vista intuitivo la definizione di variabile aleatoria è poco chiara ma nel nostro caso assumiamo, senza andare ad affrontare in maniera dettagliata, che essa rappresenta le quantità d'interesse determinate dall'esecuzione dell'esperimento su cui si vuole conoscere la probabilità di avvenimento dell'evento.

Come notazione adotteremo quella solitamente utilizzata in campo statistico, indicando con lettere maiuscole le variabili aleatorie e con lettere minuscole le rispettive possibili realizzazioni.

Essendo imprevedibile a priori il valore assunto da una variabile aleatoria, tutto ciò che si può fare relativamente ad essa è esprimere delle valutazioni di tipo probabilistico sui valori che essa assumerà e per tale ragione ad ogni variabile aleatoria X è associata una funzione, la *funzione di ripartizione* $F_X : \mathbb{R} \rightarrow [0, 1] \subset \mathbb{R}$ definita come:

$$F_X(t) = P(X \leq t) \quad \forall t \in \mathbb{R}$$

che ci indica la probabilità che la variabile casuale X assuma un valore minore o uguale a t .

Come si può notare nella Figura , la funzione di ripartizione è una funzione a gradini, in cui ad ogni valore intero si ha la presenza di un punto di discontinuità, con un salto in avanti.

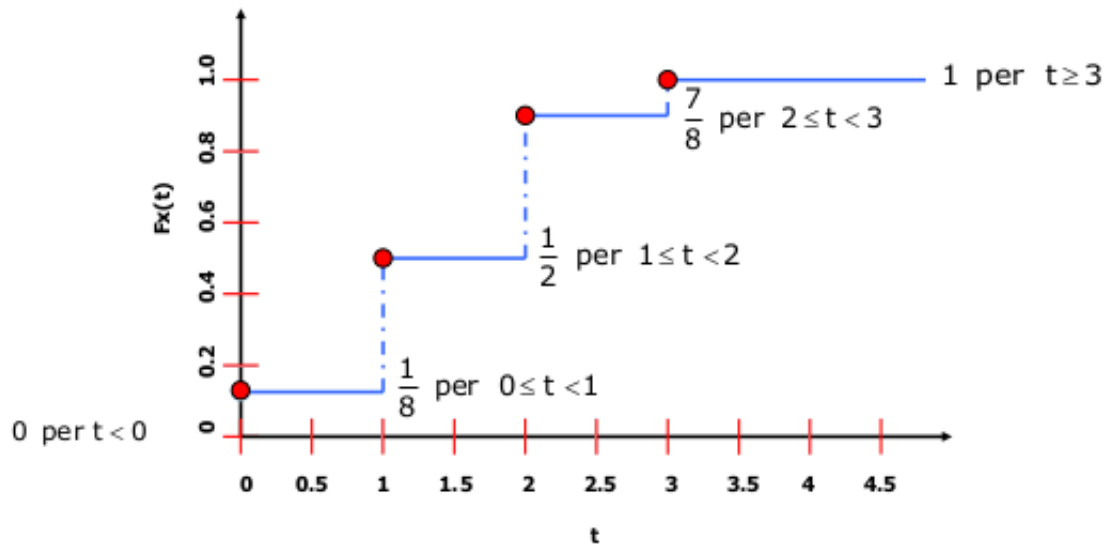
Avendo definito la funzione di ripartizione, adesso tutte le questioni riguardanti la variabile casuale X possono trovare una soluzione attraverso di essa, infatti supponiamo di calcolare $P(\{a < X \leq b\})$ con l'evento $X \leq b$ esprimibile come l'unione dei due eventi indipendenti $X \leq a$ e $a < X \leq b$ da cui si ricava, usando il terzo assioma della probabilità la seguente formula:

$$\begin{aligned} P(\{X \leq b\}) &= P(\{X \leq a\}) + P(\{a < X \leq b\}) \text{ da cui si ricava} \\ P(\{a < X \leq b\}) &= F(b) - F(a) \end{aligned}$$

Figura 3.1: Funzione di ripartizione

img:ripartizione

La **FUNZIONE DI RIPARTIZIONE** risulta essere descritta dal grafico nella figura sottostante



Questa formula ci permette di determinare la probabilità che una variabile casuale possa assumere valori in intervalli reali e ciò ha un notevole utilizzo in statistica e nella probabilità.

In genere la funzione di ripartizione non è nota, altrimenti tutti gli eventi della nostra vita sarebbero facilmente analizzabili senza nessuna incertezza, ossia ad esempio si sarebbe in grado di prevedere come vincere al superenalotto e tutti i giochi d'azzardo, per cui l'obiettivo della statistica è di determinarla o determinare le grandezze ad essa associate mentre la probabilità e le sue applicazioni assumono che essa sia sempre nota.

Si dimostra che sono delle funzioni di ripartizione tutte e sole le funzioni del tipo $F : \mathbb{R} \rightarrow [0, 1]$ che godono simultaneamente delle seguenti proprietà:

- F è monotona crescente
- $\lim_{t \rightarrow +\infty} F(t) = 1$
- $\lim_{t \rightarrow -\infty} F(t) = 0$
- $\lim_{t \rightarrow t_0^+} F(t) = F(t_0), \forall t_0 \in \mathbb{R}$

Le variabili aleatorie si distinguono in due categorie, in base a che valori può assumere l'insieme di valori S di supporto:

- *variabili aleatorie discrete*: l'insieme S è finito oppure costituito da un insieme infinito di valori discreti.
- *variabili aleatorie continue*: l'insieme S assume valori infiniti continui

Iniziamo ad analizzare prima le variabili discrete, più semplici da analizzare per poi considerare il caso continuo, in cui si estendono i valori assumibili dalle variabili.

3.3.1 Variabile Aleatoria Discreta

Come già visto, una variabile aleatoria è detta *variabile aleatoria discreta* nel caso in cui l'insieme dei valori S che essa può assumere è finita oppure da un infinito di valori discreti, in cui si associa anche, oltre alla funzione di ripartizione, una funzione di probabilità assunta da valori specifici.

Sia S il supporto della variabile aleatoria X e si definisce *distribuzione discreta di probabilità* la funzione: $p_X : \mathbb{R} \rightarrow [0, 1]$ così definita:

$$p_X = \begin{cases} P(X = t) & \forall t \in S \\ 0 & \text{altrimenti} \end{cases}$$

Una funzione rappresenta una distribuzione di probabilità, in caso siano soddisfatte entrambe le seguenti proprietà:

•

$$p_X(t) \geq 0 \quad \forall t \in R$$

•

$$\sum_{s \in S} p_X(s) = 1$$

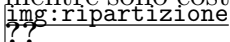
Tra le funzioni di ripartizione e le distribuzioni discrete esiste una corrispondenza biunivoca, in quanto si hanno le seguenti equivalenze:

•

$$F_X(t) = \sum_{s \in S: s \leq t} p_X(s) \quad \forall t \in \mathbb{R}$$

•

$$p_X(s) = F_X(s) - \lim_{t \rightarrow s^-} F_X(t) \quad \forall s \in S$$

Dalla prima di tali relazioni se ne deduce che le funzioni di ripartizione delle variabili aleatorie discrete presentano dei *salti* in corrispondenza dei valori s mentre sono costanti per gli altri valori, come si può anche notare nella figura 

3.3.2 Variabile Aleatoria Continua

Come già affermato in precedenza, una variabile aleatoria è detta continua nel caso in cui la corrispondente funzione di ripartizione F_X sia continua, ed in particolare viene detta *assolutamente continua* se esiste una funzione $f_X : \mathbb{R} \rightarrow \mathbb{R}_+$ tale che

$$F_X(t) = \int_{-\infty}^t f_X(u) du \quad \forall t \in \mathbb{R}$$

Una tale funzione, in caso in cui è definito l'integrale, viene detta *densità di probabilità* di X .

È detto poi *supporto* della variabile X l'insieme $S = \{t \in \mathbb{R} : f_X(t) \neq 0\}$ e si osservi che se la densità di probabilità esiste allora la sua funzione di ripartizione è una sua primitiva.

Per semplicità supporremo nel seguito che le variabili aleatorie assolutamente continue abbiano funzione di ripartizione derivabile e che la funzione di densità di probabilità sia la derivata della funzione di ripartizione.

Come anche per le distribuzioni discrete di probabilità, le funzioni di densità di probabilità per essere tali devono soddisfare le seguenti due proprietà:

1.

$$f_X(t) \geq 0 \quad \forall t \in \mathbb{R}$$

2.

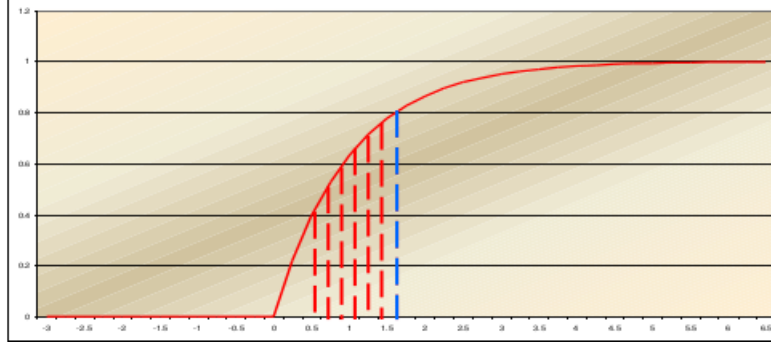
$$\int_{-\infty}^{+\infty} f_X(t) dt = 1$$

La probabilità che una variabile aleatoria continua (o assolutamente continua) assuma un ben determinato valore è sempre nulla, in quanto se X è una variabile aleatoria continua allora per ogni $t_0 \in \mathbb{R}$ risulta

$$\begin{aligned} P(X = t_0) &= P(X \leq t_0) - \lim_{t \rightarrow t_0^-} P(X \leq t) \\ &= F_X(t_0) - \lim_{t \rightarrow t_0^-} F_X(t) \\ &= F_X(t_0) - F_X(t_0) = 0 \end{aligned}$$

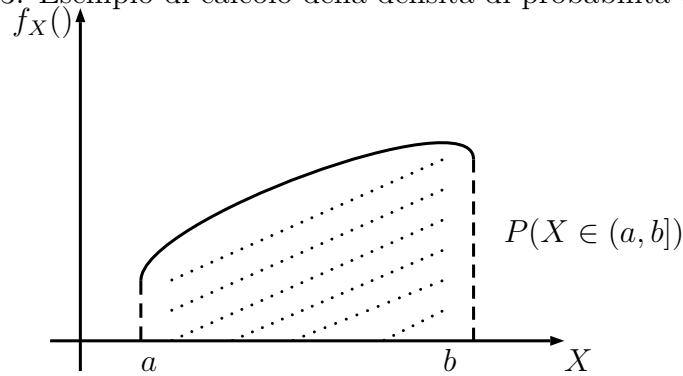
Pertanto, quando si pensa a variabili aleatorie continue, non ha mai senso domandarsi quale sia la probabilità che esse assumano valori esatti, ma conviene invece domandarsi quale è la probabilità che essi assumano specifici valori appartenenti in specifici intervalli dell'asse reale.

Figura 3.2: Funzione di ripartizione continua



img:ripartizione

Figura 3.3: Esempio di calcolo della densità di probabilità continua



graficoRipartizione

Per calcolare la probabilità che una variabile casuale continua X assuma un valore in un intervallo $(a, b] \subseteq \mathbb{R}$ è possibile far ricorso alla formula:

$$P(X \in (a, b]) = \int_a^b f_X(u) du \text{ da cui si ricava}$$

$$\begin{aligned} P(X \in (a, b]) &= F_X(b) - F_X(a) \\ &= \int_{-\infty}^b f_X(u) du - \int_{-\infty}^a f_X(u) du \end{aligned}$$

In pratica la probabilità che sia soddisfatto l'evento $X \in (a, b]$ corrisponde all'area sottesa dalla densità f_X nell'intervallo $(a, b]$, come si può notare nella figura ??, per cui

$$P(X \in (a, b]) = \int_a^b f_X(u) du \quad \forall a, b \in \mathbb{R}, \quad a < b$$

Essendo inoltre $P(X = a) = 0$ si ha:

$$P(X \in (a, b]) = P(X \in [a, b]) \quad \forall a, b \in \mathbb{R}, \quad a < b$$

Si ha che la funzione di ripartizione è la funzione integrale della funzione di densità di probabilità, quindi si ottiene la funzione di densità di probabilità tramite derivazione:

$$\frac{d}{dt}F_X(t) = f_X(t)$$

3.3.3 Variabili Aleatorie Multidimensionali

In molti casi è lecito considerare situazioni (esperimenti) il cui esito è rappresentato, anziché da un valore numerico, da una coppia o da una n-pla di valori, in tal caso si parla di *variabili aleatorie multidimensionali*.

Anche qui le variabili sono definite da uno spazio campione Ω a \mathbb{R}^n , con n dimensione della variabile, ed è comodo pensare a queste variabili aleatorie come a risultati esprimibili da n-ple di valori numerici.

Consideriamo quindi le *variabili aleatorie bidimensionali assolutamente continue*, anche se ovviamente si può estendere a n variabili anche non continue.

Sia quindi una variabile aleatoria $(X, Y) : \Omega \rightarrow \mathbb{R}^2$ con Ω , uno spazio campione al quale è associata una probabilità P definita sui sottoinsiemi di Ω .

Si definisce la *funzione di ripartizione congiunta* la funzione bidimensionale $F_{X,Y}(t, s) : \mathbb{R}^2 \rightarrow [0, 1] \subseteq \mathbb{R}$ definita come:

$$F_{X,Y} = P(\{X \leq t\} \cap \{Y \leq s\} \quad \forall (t, s) \in \mathbb{R}^2$$

inoltre se la variabile (X, Y) è assolutamente continua esiste la *funzione di densità congiunta* $f_{X,Y} : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ tale che

$$F_{X,Y}(t, s) = \int_{-\infty}^t \int_{-\infty}^s f_{X,Y}(u, v) du dv \quad \forall (t, s) \in \mathbb{R}^2$$

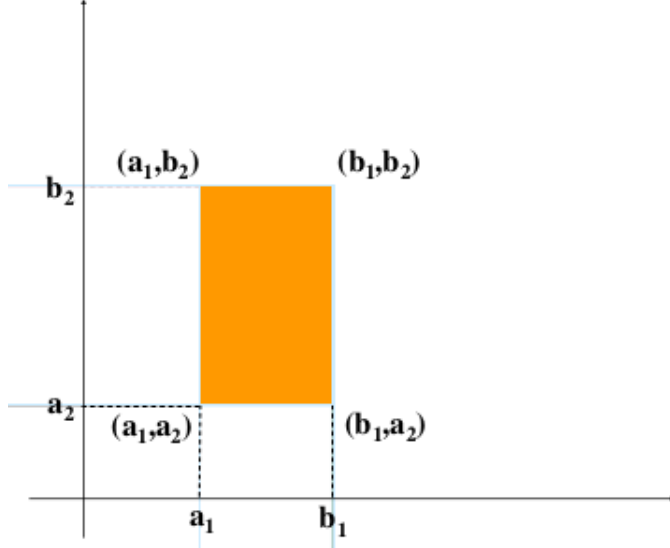
Conoscendo una delle due funzioni sopra è possibile determinare la probabilità che la coppia (X, Y) assuma valori in un qualsiasi sottoinsieme rettangolare $(a_1, b_1] \times (a_2, b_2] \in \mathbb{R}^2$, in cui risulta:

$$\begin{aligned} P((X, Y) \in (a_1, b_1] \times (a_2, b_2]) &= F_{X,Y}(b_1, b_2) - F_{X,Y}(a_1, b_2) \\ &\quad - F_{X,Y}(b_1, a_2) + F_{X,Y}(a_1, a_2) \\ &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} f_{X,Y}(u, v) du dv \quad \forall (a_1, b_1] \times (a_2, b_2] \in \mathbb{R}^2 \end{aligned}$$

In molti casi, benché ci si trovi di fronte a situazioni i cui esiti sono di tipo multidimensionale, capita di essere interessati ai valori che possono essere assunti solamente da una delle variabili per cui sono state introdotte le *funzioni marginali*.

Figura 3.4: Densità di probabilità multidimensionale

img:densitaMulti



Data una variabile aleatoria bidimensionale (X, Y) assolutamente continua, avente funzione di ripartizione congiunta $F_{X,Y}$ e funzione di densità disgiunta $f_{X,Y}$ è detta *funzione di ripartizione marginale di X* in caso si ha

$$\begin{aligned} F_X(t) &= P(\{X \leq t\} \cap \{Y \leq +\infty\}) \\ &= F_{X,Y}(t, +\infty) \end{aligned}$$

mentre si definisce la funzione p detta *funzione di densità marginale di X* come

$$f_X(t) = \int_{-\infty}^{+\infty} f_{X,Y}(t, s) ds$$

Data una variabile bidimensionale (X, Y) diciamo che le due variabili considerate singolarmente sono *stocasticamente indipendenti* se e solo se per ogni $(t, s) \in \mathbb{R}^2$ vale

$$F_{X,Y}(t, s) = F_X(t) \cdot F_Y(s)$$

che discende dalla definizione di stocasticamente indipendente fornita nella teoria assiomatica della probabilità.

3.4 Indici delle variabili aleatorie

Iniziamo ad affrontare gli indici associati alle variabili aleatorie, partendo prima dagli indici centrali, grandezze numeriche associate alle variabili aleatorie, in grado di sintetizzare, con un solo valore, le principali caratteristiche

delle loro distribuzioni.

Gli indici risultano strettamente legati agli indici introdotti nella prima parte in relazione alla statistica descrittiva, il più importante degli indici di tendenza centrale è detto *valore atteso* corrispondente alla media matematica dei dati statistici.

Data una variabile aleatoria unidimensionale X avente supporto $S \subseteq \mathbb{R}$ è detto valore atteso di X la quantità:

$$E[X] = \begin{cases} \sum_{s \in S} s \cdot p_X(s) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} u \cdot f_X(u) du & \text{se } X \text{ è assolutamente continua} \end{cases}$$

In effetti il valore atteso, così come la media di una serie di dati, va pensato come una *media pesata* dei valori assumibili dalla variabile, e fornisce un'indicazione di massima del posizionamento della variabile lungo l'asse dei numeri reali.

È possibile volere calcolare il valore atteso da una funzione $g(X)$, anch'essa una variabile aleatoria, per cui esiste una distribuzione e che è comparabile alla conoscenza della distribuzione di X

Questo calcolo effettuato su una funzione $g(X)$ è una generalizzazione della definizione della definizione del valore atteso effettuato su una variabile X e ciò ci permette di definire e dimostrare una serie di proprietà utili.

La definizione di valore atteso di una funzione $g(X)$ di una variabile aleatoria è la seguente:

$$E[g(X)] = \begin{cases} \sum_{x \in S} g(x) \cdot p_X(x) & \text{se } g(X) \text{ è discreta} \\ \int_{-\infty}^{+\infty} g(x) \cdot f_X(x) dx & \text{se } g(X) \text{ è assolutamente continua} \end{cases}$$

Il valore atteso gode delle seguenti tre proprietà:

1. $\forall a \in \mathbb{R}$, se $X = a$ con probabilità uguale ad 1, allora $E[X] = a$
2. $E[a \cdot X + b] = a \cdot E[X] + b$ per ogni variabile X e per ogni $a, b \in \mathbb{R}$

Dimostrazione. Iniziamo a dimostrare $E[aX + b]$ nel caso discreto per poi farlo nel caso continuo

$$\begin{aligned} E[aX + b] &= \sum (ax + b)p(x) \\ &= \sum axp(x) + \sum bp(x) \\ &= a \sum xp(x) + b \sum p(x) \\ &= aE[x] + b \end{aligned}$$

Nel caso continuo avviene lo stesso procedimento, solo che si utilizzano gli integrali invece della sommatoria, come si può vedere nella riga successiva.

$$\begin{aligned}
 E[aX + b] &= \int_{-\infty}^{+\infty} (ax + b)f(x)dx \\
 &= \int_{-\infty}^{+\infty} axf(x)dx + \int_{-\infty}^{+\infty} bf(x)dx \\
 &= a \int_{-\infty}^{+\infty} xf(x)dx + b \int_{-\infty}^{+\infty} f(x)dx \\
 &= aE[X] + b
 \end{aligned}$$

□

Il valore atteso non è detto che esista, infatti in caso l'integrale o la sommatoria non convergono il valore atteso risulta non definito e come avevamo già visto con la media, il valore atteso è in realtà un caso particolare di momento centrale di ordine $r = 1$, la cui formula è definita come:

$$E[X^n] = \begin{cases} \sum x^n \cdot p(n) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} x^n \cdot f_X(x) dx & \text{se } X \text{ è assolutamente continua} \end{cases}$$

Un altro indice di tendenza centrale importante è la *moda* di una variabile aleatoria X , indicata con $\tilde{X} \in \mathbb{R}$, corrispondente al valore per cui è massima la distribuzione discreta di probabilità, se X è discreta, oppure rappresenta la funzione di densità; tale valore non è detto che sia unico, ma può avere la presenza di più valori di moda, e ciò porta a parlare di distribuzione multimodale.

Un terzo indice di tendenza centrale è la *mediana* di una variabile aleatoria X , indicata con $\hat{X} \in \mathbb{R}$, che soddisfa la disequaglianza:

$$\lim_{t \rightarrow \hat{X}^-} F_X(t) \leq \frac{1}{2} \leq F_X(\hat{X})$$

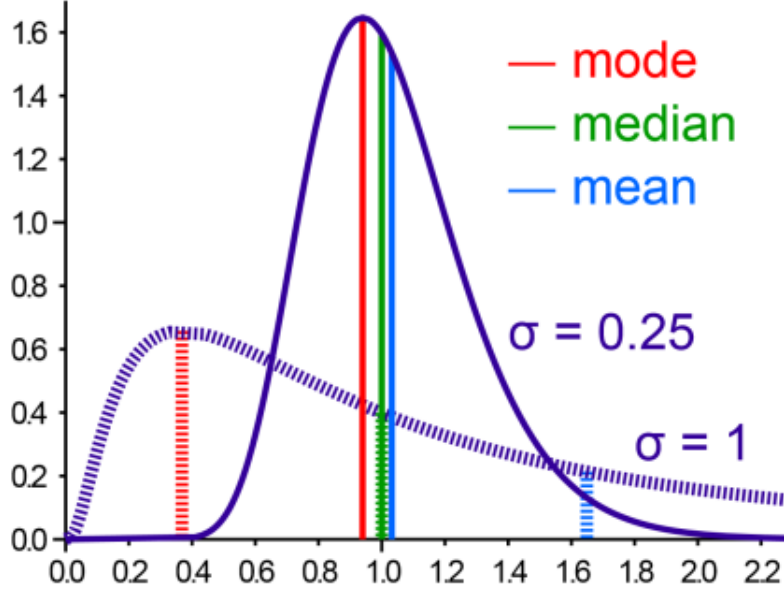
Nel caso in cui la funzione di ripartizione della variabile sia continua ed invertibile allora: $\hat{X} = F_X^{-1}(\frac{1}{2})$ mentre nel caso di variabili discrete la mediana è il valore dell'ascissa in cui la funzione di ripartizione passa da un valore minore di $\frac{1}{2}$ ad uno superiore.

La mediana può non essere unica, e ciò avviene in caso esistano più valori t per cui risulta $F_X(t) = \frac{1}{2}$.

Unitamente alla mediana è possibile considerare altri indici definiti in maniera simile e che dividono la retta dei reali in due intervalli di probabilità

Figura 3.5: grafico di moda, mediana e media

fig:centralValue



assegnata e che sono detti *quantili*, come si nota nella Figura [fig:quantile](#) ???. Dato un valore $P \in [0, 1] \subseteq \mathbb{R}$ è detto *quantile p-esimo della variabile aleatoria X* il valore $x_p \in \mathbb{R}$ tale che

$$\lim_{t \rightarrow x_p^-} F_X(t) \leq p \leq F_X(x_p)$$

Nel caso in cui la funzione di ripartizione sia continua ed invertibile, allora $x_p = F_X^{-1}(p)$ e questa definizione ci porta a pensare a x_p come ad un valore in cui risulta $P(X \leq x_p) = p$ e $P(X > x_p) = 1 - p$.

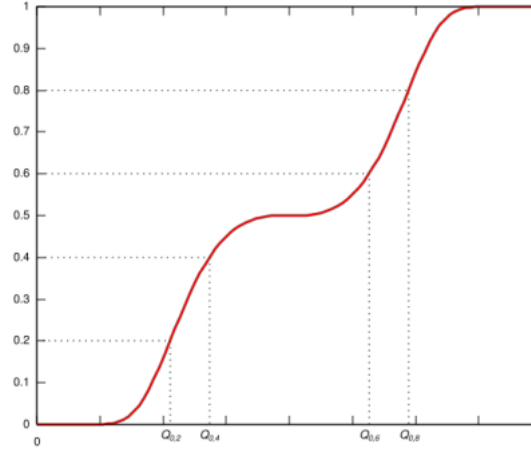
Come visto in precedenza, nella statistica descrittiva, oltre agli indici di tendenza centrale è conveniente anche considerare indici che forniscano un'idea del grado di dispersione dei valori assumibili da una variabile aleatoria. Questi vengono detti *indici di variabilità*: il più famoso tra tutti è la *varianza* di una variabile aleatoria X , indicata con σ_X^2 , definita come:

$$V[X] = \begin{cases} \sum_x (x - E[X])^2 \cdot p_X(x) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} (x - E[X])^2 \cdot f_X(x) dx & \text{se } X \text{ è assolutamente continua} \end{cases}$$

Un'altra definizione della varianza, molto utile da calcolare e da usare nelle

Figura 3.6: Grafico del quantile

fig:quantile



dimostrazione è la seguente:

$$\begin{aligned}
 V[X] &= E[(X - \mu)^2] \\
 &= E[X^2 - 2\mu X + \mu^2] \\
 &= E[X^2] - 2\mu E[X] + \mu^2 \\
 &= E[X^2] - 2\mu^2 + \mu^2 \\
 &= E[X^2] - (E[X])^2
 \end{aligned}$$

Così come il valore atteso anche la varianza talvolta può non esistere, quando la sommatoria o l'integrale divergono.

La radice della varianza è anch'esso un altro indice importante, detto *deviazione standard*, il cui vantaggio rispetto alla varianza è quello di avere la stessa unità di misura del valore atteso.

Prima di andare a definire le proprietà della varianza, andiamo a definire il valore atteso di 2 variabili come

$$E[g(X, Y)] = \begin{cases} \sum_y \sum_x g(x, y) p(x, y) & \text{se } X \text{ è discreta} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y) f(x, y) dx dy & \text{se } X \text{ è continua} \end{cases}$$

Riguardo al valore atteso di due variabili aleatorie si ha la seguente proprietà:

Teorema 1.

$$E[X + Y] = E[X] + E[Y]$$

$$E[X_1 + X_2 + \cdots + X_n] = \sum_{i=1}^n E[X_i]$$

Dimostrazione.

$$\begin{aligned}
 E[X + Y] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x + y)f(x, y)dx dy \\
 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xf(x, y)dx dy + \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} yf(x, y)dy \\
 &= E[X] + E[Y]
 \end{aligned}$$

In maniera analoga si dimostra nel caso discreto, in cui si usa ovviamente il simbolo di sommatoria mentre nel caso di somma di 3 o più variabili si applica più volte la somma di 2 variabili, fino ad arrivare a sommare n variabili aleatorie. \square

La varianza ha le seguenti proprietà:

1. $\forall a \in \mathbb{R}$, se $X = a$ con probabilità uguale ad 1 allora $V[X] = 0$
2. $V[a \cdot X + b] = a^2 \cdot V[X]$ per ogni variabile X e per ogni $a, b \in \mathbb{R}$
3. $V[X] = E[X^2] - (E[X])^2 \forall X$ variabile aleatoria

Dimostrazione.

$$\begin{aligned}
 V[aX + b] &= E[(aX + b - E[aX + b])^2] \\
 &= E[(aX + b - aE[X] - b)^2] \\
 &= E[(aX - a\mu)^2] \\
 &= a^2 E[(X - \mu)^2] \\
 &= a^2 V[X]
 \end{aligned}$$

\square

Nel caso della varianza non è generalmente vero che la varianza della somma di variabili aleatorie coincide con la somma delle varianze delle singole variabili aleatorie.

In caso le variabili aleatorie sono indipendenti la somma di variabili aleatorie coincide alla somma delle varianze ma per poterlo dimostrare e definire formalmente dobbiamo prima introdurre il concetto di covarianza.

La covarianza, come già visto nel capitolo sulla statistica descrittiva, ci fornisce il grado di indipendenza di due o più variabili aleatorie ed è definita con le seguenti due definizioni:

•

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

•

$$\begin{aligned}
Cov(X, Y) &= E[(X - \mu_x)(Y - \mu_y)] \\
&= E[XY - \mu_y X - \mu_x Y + \mu_x \mu_y] \\
&= E[XY] - \mu_y E[X] - \mu_x E[Y] + \mu_x \mu_y \\
&= E[XY] - \mu_x \mu_y - \mu_x \mu_y + \mu_x \mu_y \\
&= E[XY] - E[X]E[Y]
\end{aligned}$$

La covarianza prevede le seguenti proprietà, utile per dimostrazioni e per calcolare le varianze composte:

•

$$Cov(aX, Y) = aCov(X, Y)$$

Dimostrazione.

$$\begin{aligned}
Cov(aX, Y) &= E[(aX - \mu_x)(Y - \mu_y)] \\
&= E[aXY - aX\mu_y - \mu_x Y + \mu_x \mu_y] \\
&= aE[XY] - a\mu_y E[X] - \mu_x E[Y] + \mu_x \mu_y \\
&= aE[XY] - a\mu_x \mu_y \\
&= a(E[XY] - \mu_x \mu_y) \\
&= aCov(X, Y)
\end{aligned}$$

□

•

$$Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$$

Dimostrazione.

$$\begin{aligned}
Cov(X + Z, Y) &= E[(X + Z)Y - E[X + Z]E[Y]] \\
&= E[XY] + E[ZY] - E[X]E[Y] - E[Z]E[Y] \\
&= Cov(X, Y) + Cov(Z, Y)
\end{aligned}$$

□

•

$$Cov\left(\sum_{i=1}^n X_i, Y\right) = \sum_{i=1}^n Cov(X_i, Y)$$

Dimostrazione. Per $n = 2$ il risultato era stato appena dimostrato, per cui proviamo a ripetere lo stesso procedimento per $n = 3$

$$\begin{aligned} Cov(X_1 + X_2 + X_3, Y) &= E[(X_1 + X_2 + X_3)Y - E[X_1 + X_2 + X_3]E[Y]] \\ &= E[X_1Y] + E[X_2Y] + E[X_3Y] - E[X_1]E[Y] - E[X_2]E[Y] - E[X_3]E[Y] \\ &= Cov(X_1, Y) + Cov(X_2, Y) + Cov(X_3, Y) \end{aligned}$$

Applicando lo stesso procedimento per $n > 3$ si ottiene il procedimento appena dimostrato \square

•

$$Cov\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m Cov(X_i, Y_j)$$

Dimostrazione.

$$\begin{aligned} Cov\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) &= \sum_{i=1}^n Cov(X_i, \sum_{j=1}^m Y_j) \text{ per la proprietà precedente dimostrata} \\ &= \sum_{i=1}^n Cov\left(\sum_{j=1}^m Y_j, X_i\right) \text{ per la simmetria della funzione} \\ &= \sum_{i=1}^n \sum_{j=1}^m Cov(Y_j, X_i) \\ &= \sum_{i=1}^n \sum_{j=1}^m Cov(X_i, Y_j) \text{ per la simmetria della covarianza} \end{aligned}$$

 \square

•

$$V\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n V[X_i] + \sum_{i=1}^n \sum_{j=1, j \neq i}^n Cov(X_i, X_j)$$

Dimostrazione. La dimostrazione deriva dalla proprietà precedentemente dimostrata, impostando $m = n$ e $Y_j = X_j$ per $j = 1 \dots n$ \square

•

$$V[X + Y] = V[X] + V[Y] + 2Cov(X, Y)$$

Dimostrazione. Questo è un corollario del teorema appena dimostrato, ponendo $n = 2$ e sapendo che $Cov(X, Y) = Cov(Y, X)$ \square

- Date n variabili aleatorie indipendenti X_1, X_2, \dots, X_n si ha che

$$V\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n V[X_i]$$

Dimostrazione. Partendo dal caso semplice, ossia $n = 2$, sappiamo dal corollario precedente che

$$V[X_1 + X_2] = V[X_1] + V[X_2] + 2Cov(X_1, X_2)$$

Essendo X_1 e X_2 indipendenti risulta che $Cov(X_1, X_2) = 0$ e quindi $V[X_1 + X_2] = V[X_1] + V[X_2]$.

Applicando questo aspetto per le altre $n - 2$ variabili aleatorie indipendenti si verifica la proposizione, ossia $V[\sum_{i=1}^n X_i] = \sum_{i=1}^n V[X_i]$. \square

In caso la covarianza è nulla le due variabili aleatorie vengono dette *incorrelate*, relazione meno forte della indipendenza stocastica, come già avevamo notato nel capitolo sulla statistica descrittiva.

Dalla sua definizione si nota che $Cov(X, Y) = Cov(Y, X)$ e che $Cov(X, X) = E[(X - \mu)(X - \mu)] = V[X]$. Anche per le variabili multidimensionali ed in particolare per quelle bidimensionali esistono indici di tendenza centrale e variabilità.

Sia (X, Y) una variabile aleatoria bidimensionale discreta o continua, sono detti *valori attesi marginali* e *varianza marginali* le quantità $E[X]$ $E[Y]$ $V[X]$ $V[Y]$ ottenute considerando le distribuzioni marginali di X ed Y ed integrando (o sommando) in accordo ai sistemi visti sopra.

Un ultimo indice è quello detto *coefficiente di correlazione di Pearson*, strettamente legato alla covarianza ed utilizzato per esprimere più chiaramente il grado di dipendenza tra due variabili:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{V[X] \cdot V[Y]}} = \frac{Cov(X, Y)}{\sigma_x \cdot \sigma_y}$$

questo indice possiede le seguenti proprietà:

1. $Corr(XY) = 0$ se X e Y sono incorrelate
2. $|Corr(XY)| = 1$ se vale la relazione $Y = aX + b \quad \forall a, b \in \mathbb{R}$

La funzione generatrice $\phi(t)$ del momento di una variabile X è definita come

$$\phi(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx} p(x) & \text{con } X \text{ una variabile discreta} \\ \int_{-\infty}^{+\infty} e^{tx} f(x) dx & \text{con } X \text{ variabile assolutamente continua} \end{cases}$$

La funzione generatrice non è detto che esista, in quanto il valore atteso $E[e^{tX}]$ potrebbe non esistere, anche se nel momento $\phi(0)$ la funzione esiste sempre ed è uguale a 1.

La funzione generatrice $\phi(t)$ ci fornisce il momento n -esimo di una variabile aleatoria X , attraverso la differenziazione della variabile, come si può notare ora nel proseguo

$$\phi'(t) = \frac{d}{dt} E[e^{tx}] = E\left[\frac{d}{dt} e^{tx}\right] = E[xe^{tx}]$$

Calcolando il momento primo rispetto a 0 otteniamo $\phi'(0) = E[X]$, cosa che sapevamo già dalla statistica descrittiva mentre similmente se deriviamo ancora il $\phi'(t)$ otteniamo il momento secondo come segue

$$\phi''(t) = \frac{d}{dt} \phi'(t) = \frac{d}{dt} E[xe^{tx}] = E\left[\frac{d}{dt} (xe^{tx})\right] = E[x^2 e^{tx}]$$

Calcolando il momento secondo rispetto a 0 otteniamo che $\phi''(0) = E[X^2]$ ed in generale risulta

$$\phi^n(0) = E[X^n] \quad n \geq 1$$

Teorema 2. *Date due variabili indipendenti X e Y risulta che $\phi_{X+Y}(t) = \phi_X(t) \cdot \phi_Y(t)$*

Dimostrazione.

$$\begin{aligned} \phi_{X+Y}(t) &= E[e^{t(x+y)}] \\ &= E[e^{tx} e^{ty}] \\ &= E[e^{tx}] E[e^{ty}] \\ &= \phi_X(t) \cdot \phi_Y(t) \end{aligned}$$

□

Un'altro importante aspetto della funzione generatrice è che determina la distribuzione di una variabile, quindi può essere usata come modo alternativo per definire la funzione di ripartizione di una variabile.

Capitolo 4

Distribuzioni Notevoli

Essendoci una correlazione, come notato nel precedente capitolo, tra la funzione di ripartizione e la sua distribuzione/densità di probabilità, si parla di **distribuzione** di una variabile intendendo indifferentemente la sua ripartizione o la sua densità/distribuzione.

Con $X \sim F$ si indica che la variabile X è distribuita secondo la distribuzione F .

Adesso affrontiamo una serie di distribuzioni famose, che hanno un notevole successo ed utilizzo in campo statistico e probabilistico, incominciando dalle distribuzioni discrete e poi si analizzano quelle assolutamente continue.

4.1 Distribuzioni discrete

Le distribuzioni discrete maggiormente utilizzate sono le seguenti:

- bernoulliana
- binomiale
- poisson
- geometrica

Una variabile aleatoria X è detta distribuita secondo una Bernoulliana di parametro $p \in [0, 1]$, indicata con $X \sim B(p)$ se essa può assumere solo i valori 1 e 0 rispettivamente con probabilità p e $(1 - p)$.

Questa distribuzione presenta le seguenti funzioni di ripartizione e la sua corrispondente distribuzione di probabilità:

$$\bullet \quad p_X(t) = \begin{cases} 1-p & \text{se } t=0 \\ p & \text{se } t=1 \\ 0 & \text{altrimenti} \end{cases}$$

$$\bullet \quad F_X(t) = \begin{cases} 0 & \text{se } t < 0 \\ 1-p & \text{se } 0 \leq t < 1 \\ 1 & \text{se } t \geq 1 \end{cases}$$

L'importanza di questa semplice distribuzione è ovvia, in quanto sono variabili di Bernoulli tutte quelle che individuano il verificarsi di uno specifico evento e che valgono 1 se questo si verifica e 0 altrimenti.

Attraverso l'applicazione della definizione di valore atteso e di varianza si ottiene:

$$E[X] = 0 \cdot (1-p) + 1 \cdot p = p$$

$$V[X] = [0^2 \cdot (1-p) + 1^2 \cdot p] - (1-p) \cdot p$$

Siano X_1, \dots, X_n n variabili Bernoulliane di identico parametro p e stocasticamente indipendenti tra loro, e sia anche $X = \sum X_i$, variabile definita distribuita secondo una binomiale di parametri n e p , espressa con $X \sim \text{Bin}(n, p)$, se tale variabile può assumere qualsiasi valore intero k compreso tra 0 e n , in accordo con la seguente probabilità:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

La parte $p^k \cdot (1-p)^{n-k}$ fornisce la probabilità che k delle n variabili X_i assumano il valore 1 e che le restanti $n-k$ variabili assumano valore 0 mentre la prima parte $\binom{n}{k}$, come ovvio dal corso di matematica discreta, fornisce il numero di combinazioni possibili delle variabili k .

In questa distribuzione vengono definite le seguenti funzioni di ripartizione e di distribuzione:

•

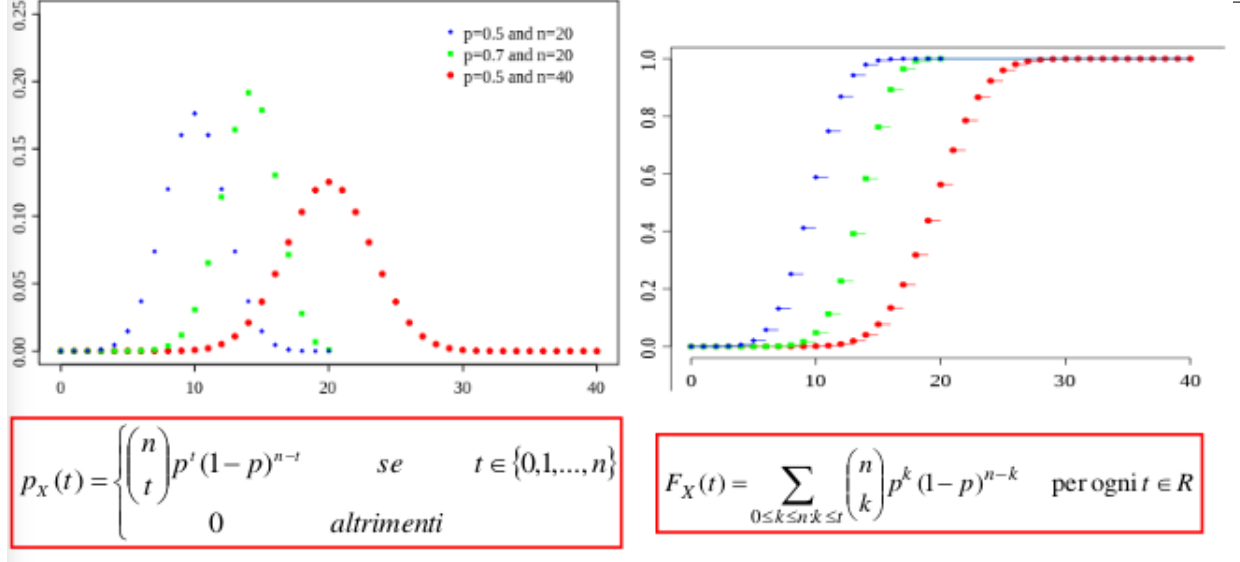
$$p_X(t) = \begin{cases} \binom{n}{t} p^t (1-p)^{n-t} & \text{se } t \in \{0, \dots, n\} \\ 0 & \text{altrimenti} \end{cases}$$

•

$$F_X(t) = \sum_{0 \leq k \leq n; k \leq t} \binom{n}{k} p^k (1-p)^{n-k}, \quad \forall t \in \mathbb{R}$$

Figura 4.1: Distribuzione binomiale

fig:binomiale



Le variabili X_i sono indipendenti quindi posso calcolare il valore atteso e la varianza di X :

$$E[X] = E[X_1 + \dots + X_n] = E[X_1] + \dots + E[X_n] = n \cdot p$$

$$V[X] = V[X_1 + \dots + X_n] = V[X_1] + \dots + V[X_n] = n \cdot (1-p) \cdot p$$

La principale applicazione della distribuzione binomiale consiste nella definizione di variabili che "contano" le realizzazioni di eventi quando questi siano da considerarsi indipendenti e con identica probabilità di verificarsi.

La distribuzione di Poisson può essere usata per approssimare una Binomiale quando il numero di variabili X_i che compaiono in $X = \sum X_i$ tende ad infinito mentre il valore del parametro p tende a zero, in modo tale che il prodotto $\lambda = n \cdot p$ resti costante.

In caso ciò viene rispettato definiamo X distribuita secondo una Poisson con parametro $\lambda \in \mathbb{R}_+$, indicata con $X \sim Poi(\lambda)$, applicabile sui valori in \mathbb{R} . La probabilità associata a questa distribuzione è la seguente:

$$P(X = k) = \lim_{n \rightarrow +\infty} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda} \quad \forall k \in \mathbb{N}$$

la *distribuzione di probabilità* e la *funzione di ripartizione* risultano essere quindi:

$$p_X(t) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & \text{se } t \in \{0, \dots, n\} \\ 0 & \text{altrimenti} \end{cases}$$

$$F_X(t) = \sum_{k \in \mathbb{N}: k \leq t} \frac{\lambda^k}{k!} e^{-\lambda}, \quad \forall t \in \mathbb{R}$$

La distribuzione Poisson può essere definita tramite la funzione generatrice $\phi(t)$ come

$$\begin{aligned} \phi(t) &= E[e^{tx}] = \sum_{i=0}^{+\infty} \\ &= e^{-\lambda} \sum_{i=0}^{+\infty} \frac{(\lambda e^t)^i}{i!} \\ &= e^{-\lambda e^{\lambda e^t}} \\ &= e^{\lambda(e^t - 1)} \end{aligned}$$

Derivando una e due volte la funzione generatrice otteniamo la funzione generatrice prima e seconda

$$\phi'(t) = \lambda e^t e^{\lambda(e^t - 1)}$$

$$\phi''(t) = (\lambda e^t)^2 e^{\lambda(e^t - 1)} + \lambda e^t e^{\lambda(e^t - 1)}$$

Valutando a $t = 0$ otteniamo il valore atteso e la varianza come

$$E[X] = \phi'(0) = \lambda$$

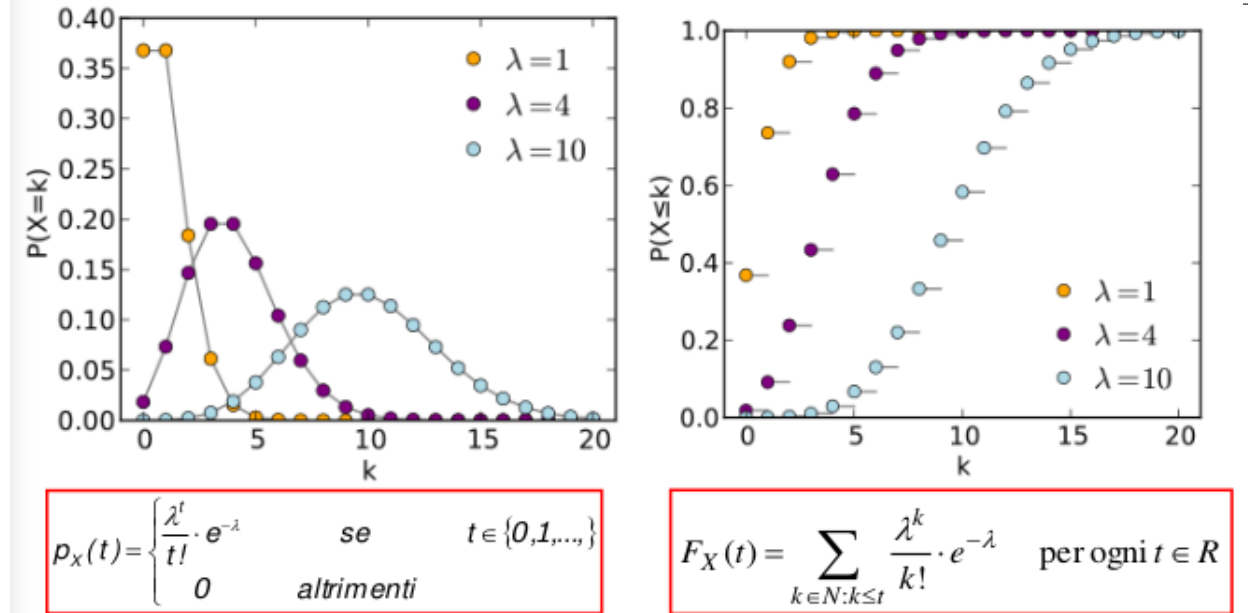
$$V[X] = \phi''(0) - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

La funzione generatrice della somma di variabili poisson indipendenti X_1 e X_2 è definita come

$$\begin{aligned} \phi_{X_1+X_2}(t) &= E[e^{t(x_1+x_2)}] \\ &= E[e^{tx_1} e^{tx_2}] \\ &= E[e^{tx_1}] \cdot E[e^{tx_2}] \\ &= e^{\lambda_1(e^t - 1)} \cdot e^{\lambda_2(e^t - 1)} \\ &= e^{(\lambda_1 + \lambda_2)(e^t - 1)} \end{aligned}$$

Figura 4.2: Distribuzione di Poisson

fig:poisson



Essendo la funzione generatrice appena trovata una poisson con $\lambda = \lambda_1 + \lambda_2$ possiamo concludere che la somma di due variabili poisson indipendenti è anch'essa una poisson, con parametro λ uguale alla somma dei loro parametri λ_1 e λ_2 .

La distribuzione di Poisson viene utilizzata quando si considerino grandi popolazioni di individui in cui ogni individuo ha una probabilità p molto piccola di essere soggetto ad uno specifico evento in esame, come approssimazione di una binomiale, e per tale ragione la distribuzione di Poisson viene anche detta degli eventi rari.

Una variabile aleatoria X è detta distribuita secondo una Geometrica di parametro $p \in [0, 1]$, indicata con $X \sim Geo(p)$, se può assumere qualsiasi valore intero non negativo k con probabilità $P(X = k) = p \cdot (1 - p)^k$.

La distribuzione di probabilità e la funzione di ripartizione risultano essere quindi:

$$p_X(t) = \begin{cases} p \cdot (1 - p)^t & \text{se } t \in \mathbb{N} \\ 0 & \text{altrimenti} \end{cases}$$

$$F_X(t) = \sum_{k \in \mathbb{N}: k \leq t} p \cdot (1 - p)^k, \quad \forall t \in \mathbb{R}$$

con valore atteso e varianza:

$$E[X] = \frac{1 - p}{p}$$

$$V[X] = \frac{1-p}{p^2}$$

Questa distribuzione ha la proprietà di **assenza di memoria**, ossia risulta $P(X = k + m | X \geq m) = P(X = k)$.

Per comprenderne il significato, supponiamo che X sia il tempo di vita di una macchina soggetta a guasti, che possono avvenire solo in corrispondenza di intervalli di tempo unitari, e supponiamo di aver rilevato che per m unità di tempo essa non si sia guastata.

La proprietà di assenza di memoria asserisce che la probabilità che la macchina si guasti all'istante $(k + m)$ -esimo, condizionata dall'evento $X \geq m$, è uguale alla probabilità iniziale che essa si guasti all'istante k -esimo.

Quindi questa proprietà asserisce che il tempo trascorso da quando abbiamo iniziato ad esaminare il funzionamento della macchina non influisce sulla distribuzione del tempo restante al verificarsi del guasto.

4.2 Distribuzioni continue

Le distribuzioni continue, affrontate in questo corso sono le seguenti:

- uniforme
- triangolare
- esponenziale
- normale(o gaussiana)

Parliamo ora della *distribuzione uniforme*, che rappresenta la più semplice distribuzione assolutamente continua e viene adottata nel caso in cui la variabile considerata possa assumere qualsiasi valore compreso in un dato intervallo con probabilità costante.

Si dice che la variabile X è distribuita secondo una uniforme di supporto $[a, b]$, indicata con $X \sim U[a, b]$ se essa è assolutamente continua con densità e funzione di ripartizione:

$$f_X(t) = \begin{cases} \frac{1}{b-a} & \text{se } t \in [a, b] \\ 0 & \text{altrimenti} \end{cases}$$

$$F_X(t) = \begin{cases} 0 & \text{se } t < a \\ \frac{t-a}{b-a} & \text{se } t \in [a, b] \\ 1 & \text{se } t > b \end{cases}$$

e con semplici integrazioni è possibile ricavare il valore atteso e la varianza:

Teorema 3.

$$E[X] = \frac{a+b}{2}$$

$$V[X] = \frac{(b-a)^2}{12}$$

Dimostrazione.

$$\begin{aligned} E[X] &= \int_{\alpha}^{\beta} \frac{x}{\beta - \alpha} dx \\ &= \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} \\ &= \frac{(\beta + \alpha)(\beta - \alpha)}{2(\beta - \alpha)} \\ &= \frac{(\beta + \alpha)}{2} \end{aligned}$$

□

Dimostrazione.

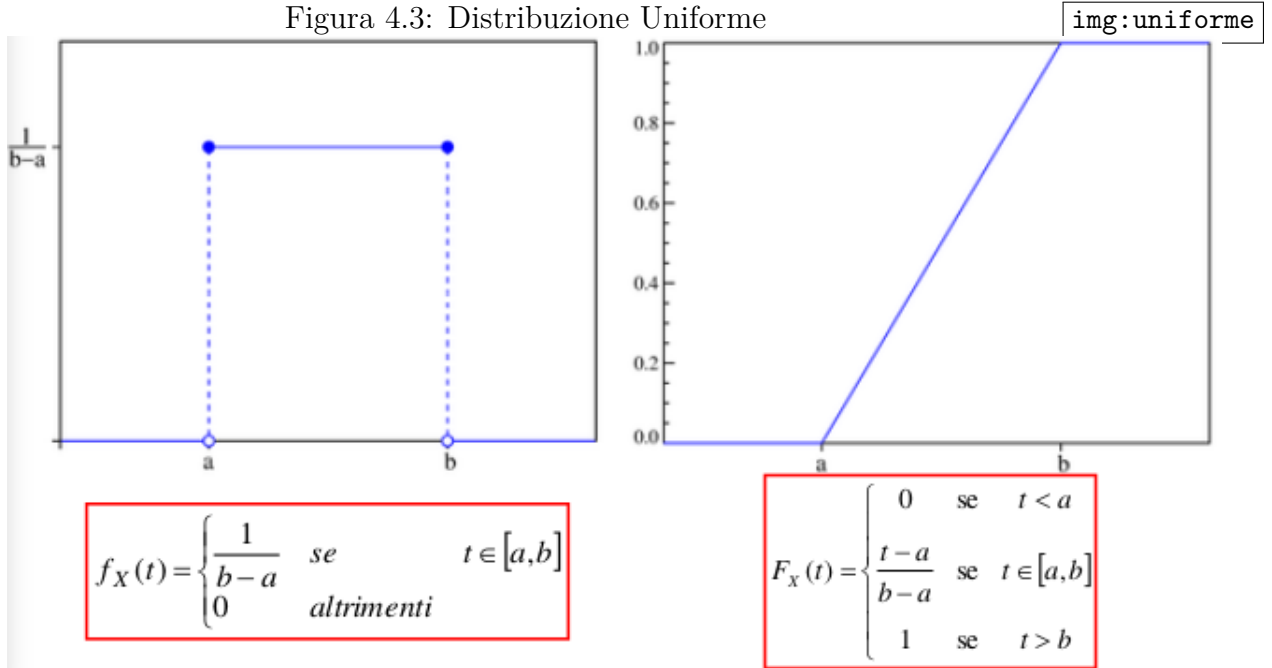
$$\begin{aligned} E[X^2] &= \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} x^2 dx \\ &= \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} \\ &= \frac{\beta^2 + \alpha\beta + \alpha^2}{3} \end{aligned}$$

Sfruttando il fatto appena dimostrato si dimostra che

$$\begin{aligned} V[X] &= E[X^2] - (E[X])^2 \\ &= \frac{\beta^2 + \alpha\beta + \alpha^2}{3} - \left(\frac{\beta + \alpha}{2}\right)^2 \\ &= \frac{\beta^2 + \alpha^2 - 2\alpha\beta}{12} \\ &= \frac{(\beta - \alpha)^2}{12} \end{aligned}$$

□

Figura 4.3: Distribuzione Uniforme



Come accennato in precedenza l'interesse in questa distribuzione è giustificato dal fatto che essa descrive bene situazioni nelle quali le variabili possono assumere valori in intervalli finiti di \mathbb{R} con probabilità uniforme, ovvero tale da essere identica per intervalli di medesima ampiezza, purché contenuti nel supporto della variabile stessa.

Ovviamente non è certo che i valori assumibili abbiano tutti la stessa probabilità di presentarsi, per questo sono stati introdotti alcune generalizzazioni della distribuzione uniforme; una di queste è la **distribuzione triangolare**, che assegna alla densità di probabilità valori maggiori al centro del supporto e minori in prossimità degli estremi.

Formalmente diciamo che la variabile X è distribuita secondo una Triangolare di supporto $[a, b]$, indicata con $X \sim T[a, b]$ se essa è assolutamente continua con le seguenti funzioni di densità e ripartizione:

$$f_X(t) = \begin{cases} \frac{4(t-a)}{(b-a)^2} & \text{se } t \in \left[a, \frac{a+b}{2}\right) \\ \frac{4(b-t)}{(b-a)^2} & \text{se } t \in \left[\frac{a+b}{2}, b\right] \\ 0 & \text{altrimenti} \end{cases}$$

$$F_X(t) = \begin{cases} 0 & \text{se } t < a \\ \frac{2(t-a)^2}{(b-a)^2} & \text{se } t \in \left[a, \frac{a+b}{2}\right) \\ 1 - 2\frac{(b-t)^2}{(b-a)^2} & \text{se } t \in \left[\frac{a+b}{2}, b\right] \\ 1 & \text{se } t > b \end{cases}$$

Attraverso l'integrazione delle funzioni presentate si ottiene il valore atteso e la varianza:

$$E[X] = \frac{a+b}{2}$$

$$V[X] = \frac{(b-a)^2}{24}$$

Passiamo alla **distribuzione esponenziale**, importante nello studio di variabili che descrivono i tempi necessari per il verificarsi di un evento. Formalmente, una variabile aleatoria X è distribuita secondo una Esponenziale di parametro $\lambda \in \mathbb{R}$, indicata con $X \sim \text{Exp}(\lambda)$ se essa è assolutamente continua con le seguenti funzioni:

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & \text{se } t \geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

$$F_X(t) = \begin{cases} 1 - e^{-\lambda t} & \text{se } t \geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

La distribuzione esponenziale può essere definita mediante la funzione generatrice come segue

$$\begin{aligned} \phi(t) &= E[e^{tx}] \\ &= \int_0^{+\infty} e^{tx} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^{+\infty} e^{-(\lambda-t)x} dx \\ &= \frac{\lambda}{\lambda-t} \text{ con } t < \lambda \end{aligned}$$

Derivando una e due volte otteniamo la derivata prima e la derivata seconda

$$\phi'(t) = \frac{\lambda}{(\lambda-t)^2}$$

$$\phi''(t) = \frac{2\lambda}{(\lambda - t)^3}$$

Calcolando il momento nel punto $t = 0$ siamo in grado di trovare la varianza e il valore atteso

$$E[X] = \phi'(0) = \frac{1}{\lambda}$$

$$V[X] = \phi''(0) - (\phi'(0))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

L'importanza della distribuzione esponenziale in numerosi campi applicativi è dovuta al fatto che essa è l'unica distribuzione assolutamente continua che gode della **proprietà di assenza di memoria**, vista anche nella distribuzione geometrica, di tipo discreto.

Teorema 4. *Siano X_1, X_2, \dots, X_n delle variabili aleatorie indipendenti esponenziali, aventi parametri $\lambda_1, \dots, \lambda_n$, allora $\min(X_1, X_2, \dots, X_n)$ è un esponenziale con parametri $\sum_{i=1}^n \lambda_i$*

Dimostrazione. Dato che il minor valore di una serie di numeri è più grande di X se e solo se tutti i suoi valori sono più grandi di X abbiamo che

$$\begin{aligned} P(\min(X_1, X_2, \dots, X_n) > X) &= P(X_1 > X, X_2 > X, \dots, X_n > X) \\ &= \prod_{i=1}^n P(X_i > X) \text{ per l'indipendenza} \\ &= \prod_{i=1}^n e^{-\lambda_i x} \\ &= e^{-\sum_{i=1}^n \lambda_i x} \end{aligned}$$

□

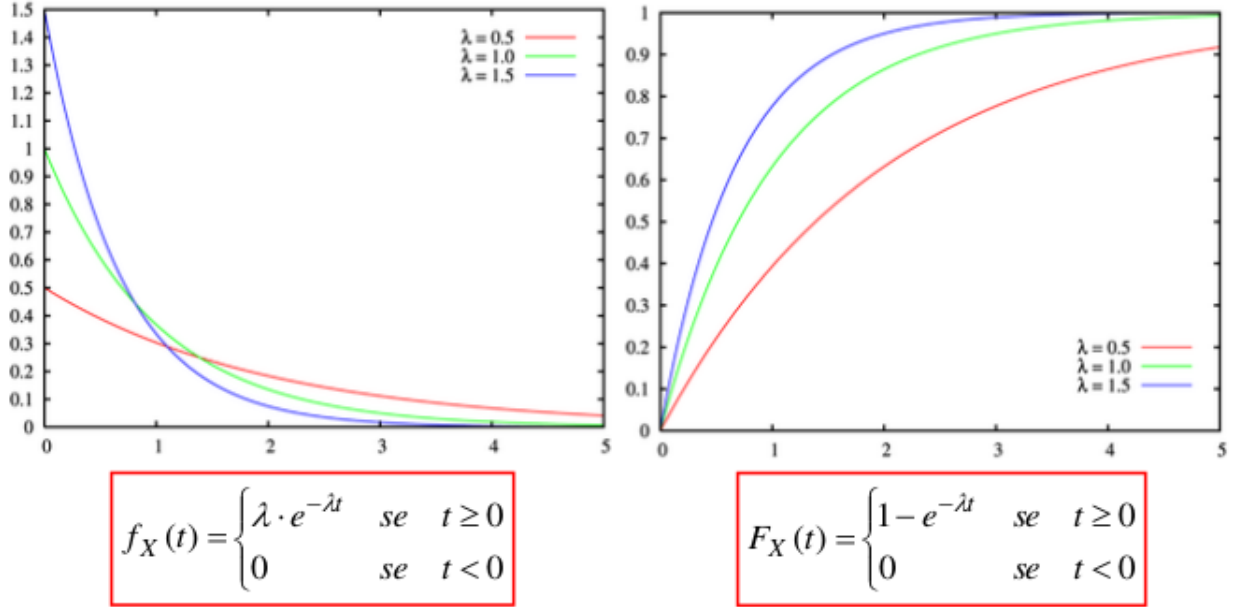
4.2.1 Distribuzione Normale

Una variabile aleatoria X è detta distribuita secondo una *Normale* con parametri $\mu \in \mathbb{R}$ e $\sigma \in \mathbb{R}_+$, indicata attraverso $X \sim N(\mu, \sigma)$, se essa è assolutamente continua con densità:

$$f_X(x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\frac{(t-\mu)^2}{2 \cdot \sigma^2}} \quad \forall t \in \mathbb{R}$$

La funzione di densità di $f_X(x)$ è una curva simmetrica rispetto a μ e avente come massimo $\frac{1}{\sigma \dots \sqrt{2\pi}} = \frac{399}{n\sigma}$ con $x = \mu$.

Figura 4.4: Distribuzione esponenziale



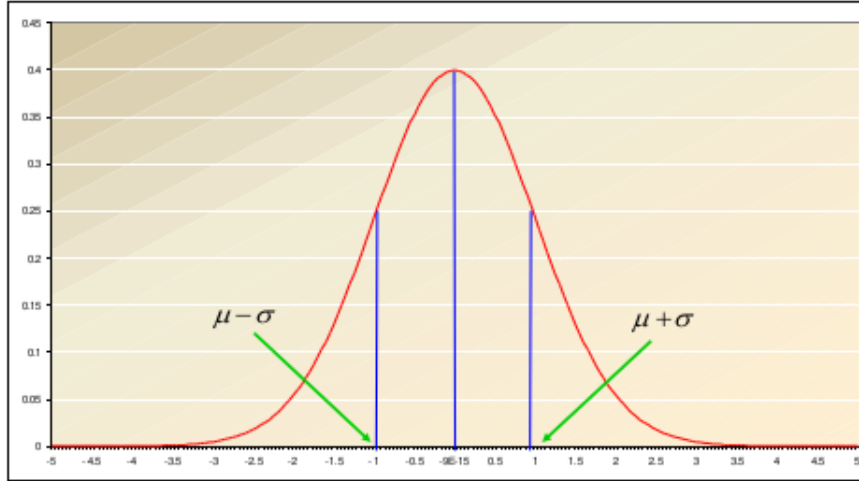
Questa distribuzione è fondamentale per il *teorema centrale del limite*, studiata nel prossimo capitolo e poi perchè molti fenomeni studiati dalla statistica e dalla probabilità si comportano come una normale, chiamata anche *gaussiana* infatti una normale viene usata per approssimare una binomiale, con n grande.

La funzione generatrice momento di una normale è definita come, ponendo $y = \frac{x-\mu}{\sigma}$,

$$\begin{aligned}
 \phi(t) &= E[e^{tx}] \\
 &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
 &= \frac{1}{\sqrt{2\pi}} e^{\mu t} \int_{-\infty}^{+\infty} e^{t\sigma y} e^{-\frac{y^2}{2}} dy \\
 &= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\frac{-y^2 - 2t\sigma y}{2}} dy \\
 &= e^{\frac{\mu t + \sigma^2 t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(y-t\sigma)^2}{2}} dy \\
 &= e^{\frac{\mu t + \sigma^2 t^2}{2}}
 \end{aligned}$$

L'ultimo passaggio deriva dalla definizione di normale, con parametri $t\mu$ e 1

Graficamente la densità di una Normale risulta come quella presentata sotto



Si osservi che la moda coincide con la media e che in corrispondenza dei valori $\mu - \sigma$ $\mu + \sigma$ vi sono dei punti di flesso.

e derivando una e due volte si ottiene

$$\phi'(t) = (\mu + t\sigma^2)e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

$$\phi''(t) = \sigma^2 e^{\mu t + \frac{\sigma^2 t^2}{2}} + e^{\mu t + \frac{\sigma^2 t^2}{2}} (\mu + t\sigma^2)$$

Dalle funzioni momento appena calcolate otteniamo il valore atteso e la varianza come

$$E[X] = \phi'(0) = \mu e^0 + 0 = \mu$$

$$V[X] = \phi''(0) - (\phi'(0))^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

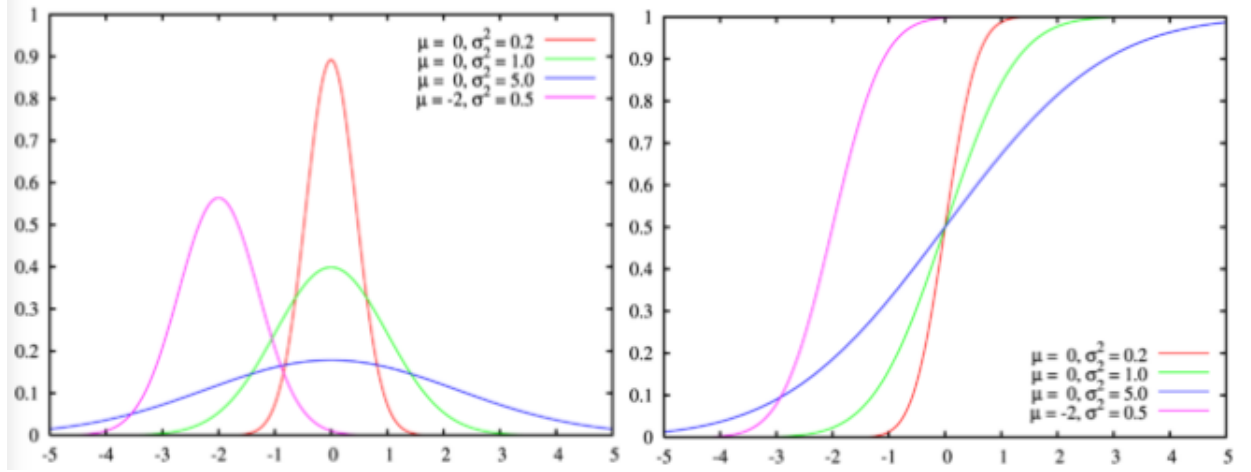
Calcolare la probabilità P , assunta da una variabile X in un intervallo $[a, b]$ è notevolmente difficile in quanto si deve risolvere la seguente equazione:

$$P(X \in [a, b]) = \int_a^b \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot e^{-\frac{(t-\mu)^2}{2 \cdot \sigma^2}} dt$$

Per questa ragione si ricorre ad opportune tavole che si riferiscono alla distribuzione normale standard ovvero con parametri $\mu = 0$ e $\sigma = 1$ e che forniscono i valori di $\int_0^z f_X(t) dt$ per un elevato numero di valori $z \in \mathbb{R}$.

Quando si sia interessati a determinare delle probabilità associate ad una generica normale ci si riconduce al caso sopra osservando che la variabile $Z = \frac{X-\mu}{\sigma}$ è distribuita secondo una *Normale Standard*

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



quindi ogni variabile $X \sim N(\mu, \sigma)$ può essere ricondotta ad una può essere ricondotta ad una Normale standardizzata ovvero ancora per ogni a, b ovvero ancora $\forall [a, b] \subseteq \mathbb{R}$ si avrà:

$$P(X \in [a, b]) = P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) = P\left(Z \in \left[\frac{a - \mu}{\sigma}, \frac{b - \mu}{\sigma}\right]\right)$$

La distribuzione gaussiana possiede la seguente proprietà:

Definizione 5. se $X_1 \sim N(\mu_1, \sigma_1)$ e $X_2 \sim N(\mu_2, \sigma_2)$ e se X_1 e X_2 sono indipendenti allora la variabile $Y = X_1 + X_2$ tale che:

$$Y \sim N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$$

In altri termini la variabile somma di due variabili aleatorie stocasticamente indipendenti con distribuzioni normali è ancora una variabile aleatoria distribuita secondo una normale i cui parametri sono ricavabili facilmente da quelli delle distribuzioni degli addendi.

La tabella della z

una volta ottenuto $P\left(Z \in \left[\frac{a-\mu}{\sigma}, \frac{b-\mu}{\sigma}\right]\right)$ poniamo per semplicità $\gamma = \frac{a-\mu}{\sigma}$ e $\delta = \frac{b-\mu}{\sigma}$. Per simmetria notiamo che è indifferente valutare γ e δ sia che siano positivi che negativi e per comodità la tabella presenta unicamente valori positivi. Valutemo quindi $|\gamma|$ e $|\delta|$. Sappiamo che $P(Z \in [\gamma, \delta]) = P(Z \in [0, |\gamma|]) + P(Z \in [0, |\delta|])$. Per calcolare, per esempio, $P(Z \in [0, |\gamma|])$

prendo la cifra intera e la prima cifra decimale e trovo la riga corrispondente nella prima colonna (se $\gamma = 1.35$ cercherò 1.3) e poi cerco scelgo la colonna corrispondente al valore della seconda cifra decimale presente nella prima riga (nel caso di prima cerco nella prima riga il valore 0.05 e ne scelgo la colonna). L'incrocio fra la riga scelta prima e la colonna scelta dopo mi daranno il valore ricercato. Ecco un esempio con 0.08 e 1.35:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Regole di calcolo per normali standardizzati da tabelle per integrali:

- integrali della forma $\int_{-\infty}^b f(u) du$:

– $b > 0$, finito:

$$\int_{-\infty}^b f(u) du = \frac{1}{2} + \int_0^b f(u) du$$

– $b < 0$, finito:

$$\int_{-\infty}^b f(u) du = \frac{1}{2} - \int_0^{-b} f(u) du$$

- integrali della forma $\int_a^{+\infty} f(u) du$:

$$\int_a^{+\infty} f(u) du = 1 - \int_{-\infty}^a f(u) du$$

- integrali della forma $\int_a^b f(u) du$:

$$\int_a^b f(u) du = \int_{-\infty}^b f(u) du - \int_{-\infty}^a f(u) du$$

Data una normale di parametri μ e σ^2 abbiamo che $Y = \alpha X + \beta$ è una normale, con media $\alpha\mu + \beta$ e varianza $\alpha^2\sigma^2$ e ciò si dimostra attraverso la funzione generatrice

$$\begin{aligned} E[e^{t(\alpha X + \beta)}] &= e^{t\beta} E[e^{\alpha t X}] \\ &= e^{t\beta} e^{\mu\alpha t + \frac{\sigma^2(\alpha t)^2}{2}} \\ &= e^{\beta + \mu\alpha)t + \frac{\alpha^2\sigma^2 t^2}{2}} \end{aligned}$$

Da questo segue che se $X \sim N(\mu, \sigma^2)$ allora $Z = \frac{X - \mu}{\sigma}$ è una normale standard, con media 0 e varianza 1 che ci permette di essere calcolato tramite delle tavole standard di Z , dato che calcolare la funzione di ripartizione di una normale non è semplice nè immediato.

La somma di variabili normali indipendenti è anch'essa una normale, come mostriamo ora con il seguente teorema

Teorema 5. *Supponiamo che X_i , con $i = 1 \dots n$, siano variabili normali indipendenti, con media μ_i e varianza σ_i^2 , abbiamo che $\sum_{i=1}^n X_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$*

Dimostrazione. la funzione generatrice di $\sum_{i=1}^n X_i$ è la seguente

$$\begin{aligned} E[e^{t \sum_{i=1}^n X_i}] &= E[e^{tx_1} e^{tx_2} \dots e^{tx_n}] \\ &= \prod_{i=1}^n E[e^{tx_i}] \text{ per l'indipendenza delle variabili} \\ &= \prod_{i=1}^n e^{\mu_i t + \frac{\sigma_i^2 t^2}{2}} \\ &= e^{\mu t + \frac{\sigma^2 t^2}{2}} \end{aligned}$$

Avendo $\sum_{i=1}^n X_i$ la stessa funzione generatrice di una normale possiamo concludere che $\sum X_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ \square

Dopo aver considerato le distribuzioni principali continue, consideriamo 3 distribuzioni utili per la statistica inferenziale:

- **chi-quadro:** siano X_1, \dots, X_n variabili con distribuzione normale standard ed indipendenti tra loro e sia X una variabile aleatoria definita come

$$X = \sum X_i^2$$

si dice distribuita secondo una chi-quadro con n gradi di libertà, indicata con $X \sim \chi_n^2$, che ovviamente essendo definita come somma di quadrati può assumere solo valori non negativi.

Teorema 6. Se X_1 e X_2 sono 2 variabili chi-quadro, con n_1 e n_2 gradi di libertà, allora $X_1 + X_2$ è una chi-quadro con $n_1 + n_2$ gradi di libertà.

Dimostrazione. Questa proprietà si può facilmente dimostrare dalla definizione di chi-quadro, ponendo $X = X_1 + X_2$, e a loro volta sostituire X_1 e X_2 con la loro definizione, risultando la somma di $n_1 + n_2$ variabili normali elevate al quadrato, da cui segue che è chi-quadro con $n_1 + n_2$ gradi di libertà. \square

Se X è una variabile chi-quadro con n gradi di libertà, allora $\forall \alpha \in (0, 1)$ la quantità $\chi_{\alpha, n}^2$ è definita come $P(X \geq \chi_{\alpha, n}^2) = \alpha$. Per calcolare la χ si usano delle tavole standard che mostrano $\chi_{\alpha, n}^2$, per una certa varietà di α e n .

- **t di student:** siano $Z \sim N(0, 1)$ e $Y \sim \chi_n^2$ due variabili indipendenti e sia X una variabile aleatoria definita come

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

si dice distribuita secondo una t di student con n gradi di libertà, indicata con $X \sim t_n$.

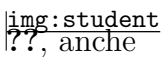
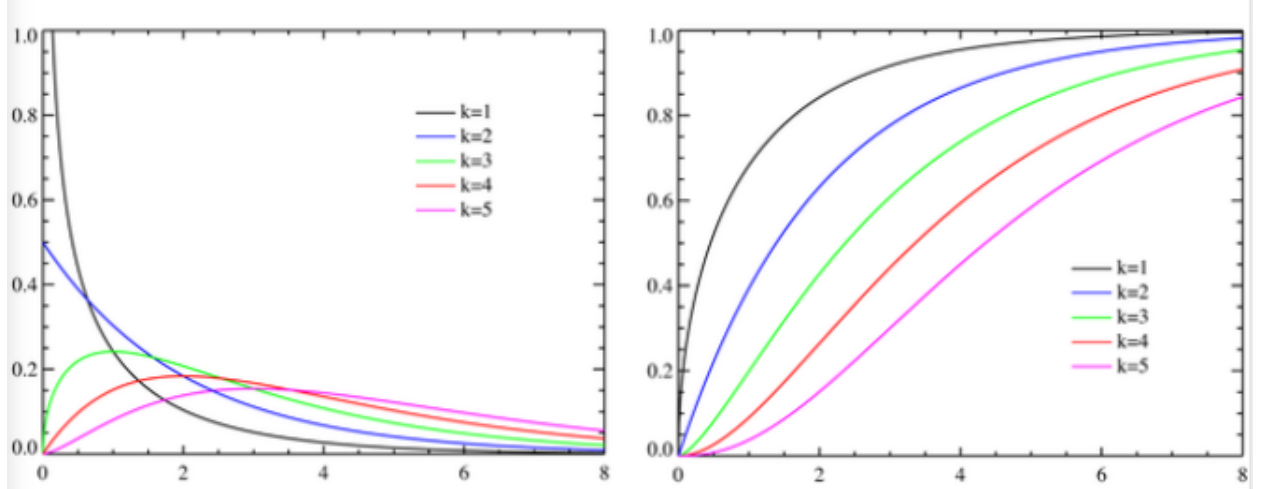
Come la normale standard, come si può notare nella Figura , anche la densità di t è simmetrica a 0 e in aggiunta con n grande la sua densità diviene molto simile alla densità di una normale standard, per la legge dei grandi numeri, dato che $E[\chi^2]$ converge a n , essendo le variabili

Figura 4.5: Distribuzione chi-quadro



formanti la chi-quadro convergono a 1 da cui segue che la t può essere approssimata dalla variabile Z , variabile aleatoria normale standard.

La distribuzione t possiede i seguenti valori atteso e varianza

$$E[T_n] = 0 \quad n > 1$$

$$V[T_n] = \frac{n}{n-2} \quad n > 2$$

Anche la distribuzione t viene calcolata mediante le tavole standard, che ci forniscono molti valori di $t_{\alpha,n}$ per dei selezionati α e n .

- **distribuzione f**: siano $U \sim \chi_m^2$ e $V \sim \chi_n^2$ due variabili indipendenti, si ha che X una variabile aleatoria definita come

$$X = \frac{\frac{U}{m}}{\frac{V}{n}}$$

si dice distribuita secondo una F con m e n gradi di libertà, indicata con $X \sim F(n, m)$.

Anche la distribuzione f viene calcolata mediante le tavole standard con $\alpha \leq \frac{1}{2}$, sfruttando il fatto che

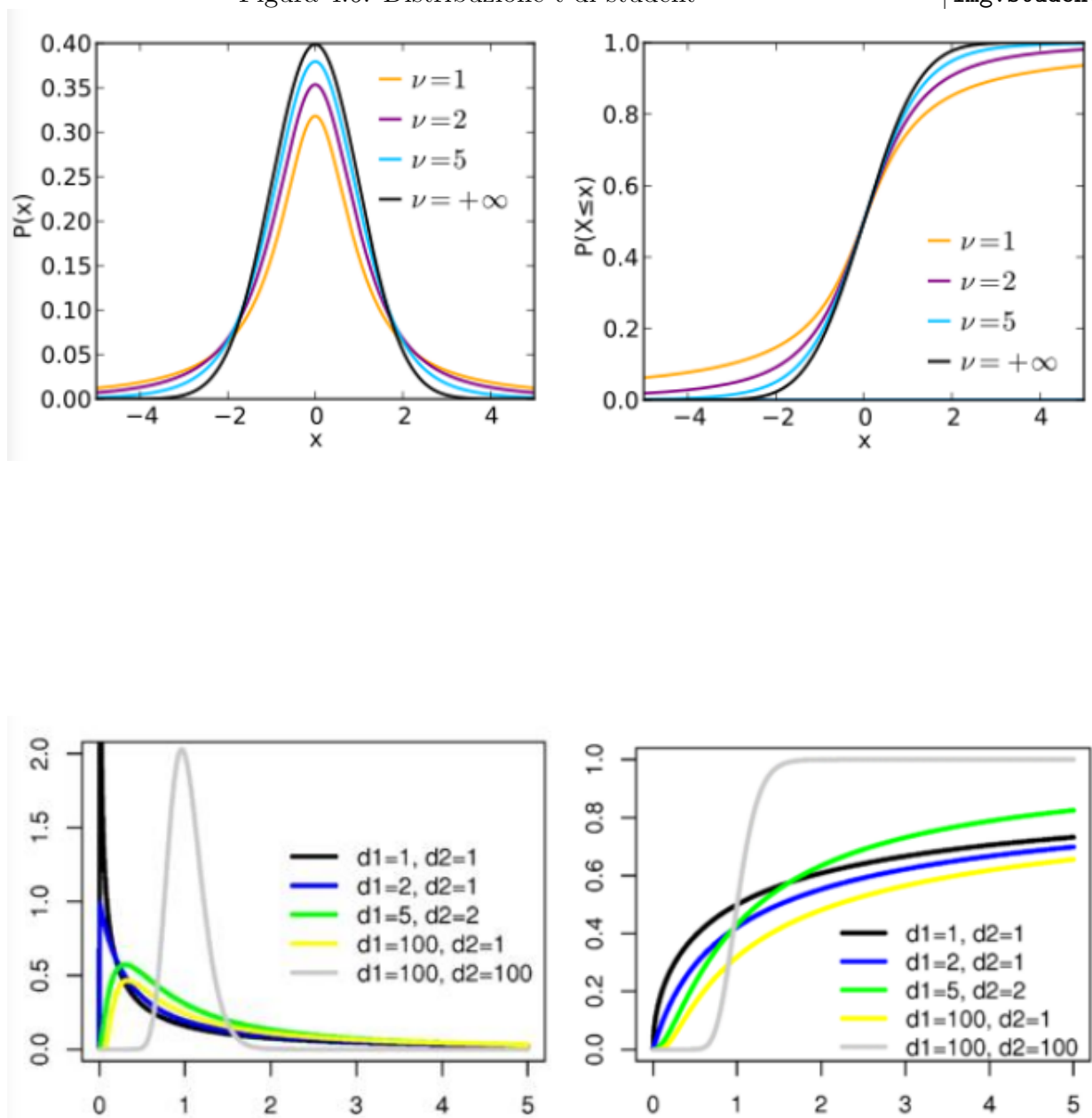
$$P(F_{n,m} > F_{\alpha,n,m}) = \alpha$$

In caso $\alpha > \frac{1}{2}$ si può risolvere andando a calcolare

$$P(F_{n,m} \geq \frac{1}{F_{\alpha,n,m}}) = 1 - \alpha$$

Figura 4.6: Distribuzione t di student

img:student



Capitolo 5

Teoremi di Convergenza

Consideriamo una successione $\{X_n, n \in \mathbb{N}\}$ di variabili aleatorie e sia F_n la funzione di ripartizione della generica variabile X_n della successione.

Diremo che la successione converge in distribuzione alla variabile X avente funzione di ripartizione F se vale:

$$\lim_{n \rightarrow \infty} F_n(t) = F(t)$$

per ogni $t \in \mathbb{R}$, punto di continuità per la funzione di ripartizione F .

Si useranno le seguenti notazioni:

$$X_n \xrightarrow{d} X$$

$$F_n \xrightarrow{d} F$$

Nella definizione di convergenza in distribuzione vengono esclusi, nel passaggio al limite, i punti in cui la funzione di ripartizione limite F è discontinua, per avere un concetto di convergenza simile all'intuizione.

Consideriamo ad esempio una successione di numeri reali $\{a_n, n \in \mathbb{N}\}$ tale che:

$$a_n \xrightarrow{d} a \in \mathbb{R} \text{ per } n \rightarrow \infty$$

e pensiamo alle variabili aleatorie X_n aventi funzione di ripartizione:

$$F_n(t) = \begin{cases} 0 & \text{se } t < a_n \\ 1 & \text{se } t \geq a_n \end{cases}$$

ovvero la generica variabile X_n assume valore a_n con probabilità 1.

Da un punto di vista intuitivo siamo portati a pensare che valga $X_n \xrightarrow{d} X$, con X avente la funzione di ripartizione:

$$F(t) = \begin{cases} 0 & \text{se } t < a \\ 1 & \text{se } t \geq a \end{cases}$$

ovvero $X = a$ con probabilità uguale a uno ma la successione $\{a_n, n \in \mathbb{N}\}$ tale che:

$$\begin{cases} a < a_n & \text{con } n \text{ pari} \\ a_n < a & \text{con } n \text{ dispari} \end{cases}$$

risulta quindi:

$$\begin{cases} F_n(a) = 0 & \text{per } n \text{ pari} \\ F_n(a) = 1 & \text{per } n \text{ dispari} \end{cases}$$

il limite $\lim_{n \rightarrow \infty} F_n(a)$ non esiste quindi è scorretto dire che la successione converge in distribuzione in quanto non è definibile in a la funzione ripartizione limite F ; per evitare questi problemi si esclude, nella definizione di convergenza, i valori in cui la F limite non è continua.

Segnaliamo che la convergenza in distribuzione non è l'unico tipo di convergenza tra variabili aleatorie definito in letteratura, come ad esempio sono importanti convergenza quasi certa e in probabilità ma in questo corso di statistica e probabilità abbiamo deciso di non considerarle, in quanto esula dagli scopi e finalità del nostro corso.

5.1 Legge dei Grandi Numeri

Considero la successione $\{X_i \in \mathbb{N}_+\}$ di variabili aleatorie indipendenti ed identicamente distribuite e considero poi la variabile aleatoria, detta *media aritmetica n -esima della successione*

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

Avendo $E[X_i] = \mu$ e $V[X_i] = \sigma^2$, si ha allora $\bar{X}_n \xrightarrow{d} M$, con M variabile aleatoria che assume il valore μ con probabilità 1.

La proprietà appena introdotta costituisce una forma debole del risultato, noto con il nome di *Legge dei Grandi Numeri*.

Nel dettaglio questa legge asserisce che:

Teorema 7. *Se considero una successione di variabili aleatorie indipendenti ed identicamente distribuite $\{X_i, i \in \mathbb{N}_+\}$ per cui esistono finito valore atteso e varianza allora si può affermare che la successione:*

$$\{\bar{X}_n, n \in \mathbb{N}_+\}$$

delle corrispondenti medie aritmetiche tende, al crescere di n , ad una variabile che assume certamente il valore:

$$E[X_i] = \mu$$

Quindi se considero una successione $\{x_i, i \in \mathbb{N}_+\}$ di realizzazioni delle variabili $\{X_i, i \in \mathbb{N}_+\}$ e considero la successione $\{\bar{x}, n \in \mathbb{N}_+\}$ delle corrispondenti realizzazioni delle medie aritmetiche, abbiamo che questa seconda successione tende, per n tendente ad infinito, al valore $E[X_i] = \mu$.

Nella realtà le ipotesi della legge dei grandi numeri possono essere indebolite rispetto a quelle definite da noi, infatti esistono versioni alternative in cui non è richiesta l'ipotesi che le variabili X_i siano identicamente distribuite.

La legge dei grandi numeri assicura la convergenza $\bar{X}_n \xrightarrow{d} M$ ma non ci fornisce alcuna informazione riguardo la rapidità con cui ciò avviene, infatti non sappiamo per quale n è lecito supporre che la realizzazione \bar{x}_n assuma valore μ o prossimo ad esso.

È intuitivo pensare che la convergenza avvenga con maggiore rapidità in caso di una varianza molto piccola, e questo è il *teorema del limite centrale*, in cui viene specificata quale sia la distribuzione della variabile aleatoria \bar{X}_n , con n sufficientemente grande, e quali siano il valore atteso e la varianza della stessa.

Sia $\{X_i, i \in \mathbb{N}_+\}$ una successione di variabili aleatorie che soddisfa le ipotesi della Legge dei Grandi Numeri, ovvero siano le X_i indipendenti ed identicamente Distribuite ed aventi valore atteso μ e varianza σ^2 , entrambi esistenti e finiti.

Si consideri ora la variabile aleatoria S_n , $S_n = \sum X_i$, in cui vale:

$$S_n \xrightarrow{d} X \sim N(n \cdot \mu, \sqrt{n} \cdot \sigma) = N(n \cdot \mu, n \cdot \sigma^2)$$

. Ovvero S_n converge in distribuzione ad una variabile distribuita come una normale di media $n \cdot \mu$ e deviazione standard $\sqrt{n} \cdot \sigma$.

Osservato che vale $\bar{X}_n = \frac{S_n}{n}$ dalla formula:

$$S_n \xrightarrow{d} X \sim N(n \cdot \mu, \sqrt{n} \cdot \sigma) = N(n \cdot \mu, n \cdot \sigma^2)$$

si ha che $E[a \cdot X + b] = a \cdot E[X] + b$, $\forall X, \forall a, b \in \mathbb{R}$ e quindi:

$$E[\bar{X}_n] = E\left[\frac{S_n}{n}\right] = \frac{1}{n}E[S_n] = \frac{1}{n}E[X_1 + \dots + X_n] = \frac{1}{n}(n\mu) = \mu$$

inoltre si ha che $V[a \cdot X + b] = a^2 \cdot V[X]$, $\forall X \forall a, b \in \mathbb{R}$ e quindi

$$V[\bar{X}_n] = V\left[\frac{S_n}{n}\right] = \frac{1}{n^2}V[S_n] = \frac{1}{n^2}V[X_1 + \dots + X_n] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

quindi si conclude che:

$$S_n \xrightarrow{d} X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(\mu, \frac{\sigma^2}{n}\right)$$

quindi per n sufficientemente grande possiamo approssimare le variabili S_n e \bar{X}_n con delle variabili aventi distribuzione normale i cui parametri dipendono da quelli delle variabili X_i .

Si ha inoltre la seguente approssimazione:

n è ritenuto sufficientemente grande sse $n \geq 30$

Capitolo 6

Stime di Parametri

La *statistica inferenziale* ci consente di dedurre particolari caratteristiche di una popolazione, considerando ed analizzando un campione finito e preferibilmente piccolo di suoi individui.

In caso si ha caratteristiche esprimibili numericamente si parla di *parametri*, mentre per stima di parametri si intende il problema della deduzione di caratteristiche di tipo numerico di una popolazione, attraverso l'analisi di un suo campione opportunamente scelto, in maniera casuale.

Per *campionamento casuale* si intende un campione, in cui si assegna la stessa probabilità di essere estratto ad ogni individuo della popolazione.

Per generare un campione casuale si assegna, in maniera progressiva, un numero ad ogni individuo della popolazione e poi si estrae, tramite un qualsiasi generatore casuale, tanti numeri quanti devono essere gli elementi del campione; questa procedura è dispensiosa dal punto di vista temporale ma è necessaria, dato che ci assicura di avere un campione casuale, i cui elementi sono stati scelti senza alcun discriminazione e/o considerazione.

Un'altra considerazione che occorre sempre fare durante un'operazione di campionamento riguarda la possibilità di estrarre più volte uno stesso individuo, detto *campionamento con ripetizione o senza ripetizione*.

La scelta tra queste due alternative diviene rilevante quando la popolazione considerata è di numerosità limitata e diviene trascurabile nel caso di popolazioni di vaste dimensioni o infinite e questo è il caso che verrà trattato d'ora in avanti.

Le tecniche per effettuare la stima di parametri sono basate sulla conoscenza delle *distribuzioni campionarie*, vale a dire distribuzioni di particolari indici statistici associati alle caratteristiche del campione.

Denotiamo con X il carattere della popolazione considerato, il cui valore assunto varia a seconda dell'individuo considerato, per cui conviene pensare

ad X come ad una variabile aleatoria, la cui distribuzione, sconosciuta, corrisponde a quella che si otterrebbe facendo ricorso alle tecniche di statistica descrittiva.

Si considera per cui i valori assunti dai singoli individui come a delle realizzazioni di X e formalmente un campione casuale di numerosità n è un n -pla (X_1, dots, X_n) di variabili aleatorie stocasticamente indipendenti, aventi ognuna la stessa distribuzione del carattere X della popolazione mentre i valori (x_1, \dots, x_n) sono una realizzazione della n -pla (X_1, \dots, X_n) .

Un **parametro** è un valore numerico che descrive una caratteristica della popolazione, e come tale è una grandezza associata alla sua distribuzione mentre una *stima* è una misura che descrive una caratteristica del campione, ossia sono del tipo $H_n = h(X_1, \dots, X_n)$ dove h è una funzione in n variabili. Le variabili così definite sono dette *statistiche campionarie* e le loro distribuzioni sono dette campionarie.

Per generare un campione casuale si può usare il campionamento *stratificato proposizionale*, in cui si presuppone una suddivisione preventiva della popolazione in gruppi omogenei e gli individui che costituiscono il campione vengono poi estratti da ogni gruppo in proporzione alla numerosità del gruppo stesso.

Il vantaggio di questo metodo è che ci permette di ridurre la numerosità finale del campione, ma richiede un tempo di campionamento maggiore, per cui è possibile usare il *campionamento a grappoli*, che prevede anch'esso una suddivisione in gruppi, ma sta volta in gruppi eterogenei, in maniera tale che ogni singolo gruppo sia rappresentativo dell'intera popolazione, per cui è sufficiente limitarsi ad estrarre un singolo gruppo quale campione.

Questo metodo presenta dei vantaggi nella raccolta dei dati ma risulta meno efficiente in termini di inferenze e solo a titolo informativo esistono i campionamento *longitudinale* e il campionamento *doppio* e infine solitamente si usano più metodi di campionamento contemporaneamente quando si effettua un'indagine statistica.

Possiamo pensare al carattere della popolazione su cui vogliamo fare delle inferenze come ad una variabile aleatoria X , avente una funzione di ripartizione F sconosciuta, ma corrispondente alla distribuzione di frequenza cumulata di tale carattere, che si potrebbe ottenere se fosse possibile analizzare per intero la popolazione.

Una stima di un parametro F è il valore assunto da una funzione di un campione casuale (X_1, \dots, X_n) di variabili stocasticamente indipendenti ed aventi tutte distribuzione F , in corrispondenza di una specifica realizzazione (x_1, \dots, x_n) di tale campione casuale; inoltre ogni statistica campionaria, essendo una funzione di variabili aleatorie, è una variabile aleatoria e come tale avrà una sua distribuzione.

Considero un campione (X_1, \dots, X_n) estratto da una popolazione con distribuzione F , valore medio μ e deviazione standard σ , e definiamo come *media campionaria* il valore:

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

invece si chiama *distribuzione campionaria delle media (campionaria) n -sima* la distribuzione della variabile \bar{X}_n .

Cerchiamo ora valore atteso e varianza di \bar{X}_n e si assume l'indipendenza delle variabili X_i , con valore atteso μ e varianza σ^2 .

Teorema 8. *Dati un insieme (X_1, \dots, X_n) di variabili aleatorie indipendenti e con tutte valore atteso μ e varianza σ^2 si ottiene le seguenti proprietà:*

$$E[\bar{X}_n] = \mu$$

$$V[\bar{X}_n] = \frac{\sigma^2}{n}$$

In caso in cui la numerosità del campione tende ad infinito, le realizzazioni \bar{x}_n saranno vicine a μ e grazie al teorema del limite centrale, per n abbastanza grande, la variabile \bar{X}_n può essere approssimata con una variabile avente distribuzione normale $N(\mu, \frac{\sigma^2}{n})$, indipendentemente dall'espressione della distribuzione F .

Dimostrazione. Il fatto che il valore atteso di \bar{X}_n sia uguale a μ proviene dal seguente fatto

$$\begin{aligned} E[\bar{X}_n] &= \frac{1}{n} \sum_{i=1}^n X_i \\ &= \frac{1}{n} (E[X_1] + E[X_2] + \dots + E[X_n]) \\ &= \frac{1}{n} n\mu = \mu \end{aligned}$$

Il fatto che la varianza di \bar{X}_n sia pari a $\frac{\sigma^2}{n}$ proviene dal fatto che

$$\begin{aligned} V[\bar{X}_n] &= V\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] \\ &= \frac{1}{n^2} (V[X_1] + V[X_2] + \dots + V[X_n]) \\ &= \frac{\sigma^2}{n^2} n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

□

In caso in cui la distribuzione sia già normale in partenza, attraverso la proprietà di chiusura rispetto alla somma di variabili aleatorie con distribuzione normale, la media campionaria è anch'essa una variabile aleatoria

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

Introduciamo ora una seconda statistica usata per stimare la varianza della popolazione, ovvero la *varianza campionaria n-sima*, definita dalla variabile:

$$S_n^2 = \frac{1}{n-1} \sum (X_i - \bar{X}_n)^2$$

mentre viene detta *distribuzione campionaria della varianza* la distribuzione della variabile S_n^2 .

Si possono determinare il valore atteso e la varianza, come si può notare:

$$E[S_n^2] = \frac{n-1}{n} \cdot \sigma^2$$

Dimostrazione.

$$\begin{aligned} (n-1)E[S^2] &= \sum X_i^2 - n\bar{X}^2 \\ &= E[\sum X_i^2] - nE[\bar{X}^2] \\ &= nE[X_1^2] - nE[\bar{X}^2] \\ &= nV[X_1] + n(E[X_1])^2 - nV[\bar{X}] - n(E[\bar{X}])^2 \\ &= n\sigma^2 + n\mu^2 - n(\frac{\sigma^2}{n} - n\mu^2) \\ &= (n-1)\sigma^2 \end{aligned}$$

Da cui si ricava dividendo per $(n-1)$ che $E[S^2] = \sigma^2$

□

$$V[S_n^2] = \frac{1}{n} \left(E[(X - \mu)^4] - \frac{n-3}{n-1} \cdot \sigma^4 \right)$$

Anche per questa statistica è possibile dimostrare, facendo ricorso al Teorema Limite Centrale, che per n sufficientemente grande la sua distribuzione può essere approssimata con una normale, con parametri dati dalle formule precedenti.

Ora scopriamo un'importante considerazione riguardante \bar{X} e S^2 , importantissimi per tutta la parte di stima intervallare e/o verifica di ipotesi, definita dal seguente teorema

Teorema 9. \bar{X} e S^2 sono due variabili aleatorie indipendenti, con $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ mentre S^2 è una chi-quadro con $n - 1$ gradi di libertà.

Dimostrazione. Il fatto che \bar{X} è approssimabile mediante una normale di parametri μ e $\frac{\sigma^2}{n}$ è già stato mostrato precedentemente, mostrando che \bar{X} è una normale tramite il teorema del limite centrale con valore atteso μ e varianza $\frac{\sigma^2}{n}$.

Per la varianza, incominciamo ponendo per i numeri $x_1, x_2, \dots, x_n, y_i = x_i - \mu$ con $i = 1 \dots n$ da cui per l'identità

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

che risulta il seguente risultato

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

Essendo X_1, X_2, \dots, X_n un campione di variabili normali, aventi media μ e varianza σ^2 , allora si ricava che

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

che può essere scritto in maniera equivalente come

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{\sum (X_i - \bar{X})^2}{\sigma^2} + \left[\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right]^2$$

Essendo $\frac{X_i - \mu}{\sigma}$, per $i = 1 \dots n$ delle variabili normali indipendenti, segue che la parte sinistra dell'equazione sia una chi-quadro con n gradi di libertà.

Anche $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$ è una variabile normale standard, per cui la sua elevazione al quadrato è una chi-quadro con 1 gradi di libertà e dato che la somma di variabili chi-quadro è una chi-quadro con $n_1 + n_2$ gradi di libertà segue che $\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2}$ è una chi-quadro di $n - 1$ gradi di libertà. \square

Da questo teorema discende un importante corollario, usato nel proseguo del capitolo per determinare la stima intervallare, con una dimostrazione abbastanza banale per uno studente universitario

Teorema 10. Sia X_1, X_2, \dots, X_n un campione di variabili normali, con media μ , se \bar{X} denota la media e S la deviazione standard allora

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

dove t è una variabile di *t* student, con $n - 1$ gradi di libertà.

Dimostrazione. Una variabile di distribuzione t per definizione è il rapporto tra $\frac{Z}{\sqrt{\frac{\chi_n^2}{n}}}$, dove Z è una normale indipendente da χ_n^2 , variabile chi-quadro di grado n .

Segue ovviamente che

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\sqrt{n}(\bar{X}-\mu)}{S}$$

che è ovviamente una t variabile, di grado $n-1$. □

Inoltre valgono anche le seguenti relazioni:

$$\hat{S}_n^2 = \frac{n}{n-1} \cdot S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

6.1 Stimatori e stime puntuali

Sia ϑ un parametro incognito della popolazione X , una statistica campionaria $H_n = h(X_1, \dots, X_n)$ è detta *stimatore puntuale* in caso viene usata per stimare il parametro incognito ϑ .

Si definisce invece *stima puntuale* di ϑ il valore $\hat{\vartheta} = h(x_1, \dots, x_n)$, assunto dallo stimatore $H_n = h(X_1, \dots, X_n)$ nella realizzazione (x_1, \dots, x_n) del campione casuale.

Una statistica campionaria può essere considerata uno stimatore puntuale deve rispettare le seguenti proprietà:

1. *proprietà di correttezza* in caso lo stimatore $H_n = h(X_1, \dots, X_n)$ di ϑ , qualsiasi valore di ϑ abbia, risulta

$$E[H_n] = \vartheta$$

2. *proprietà di consistenza*, in caso lo stimatore $H_n = h(X_1, \dots, X_n)$ di ϑ , qualsiasi valore esso sia, risulta

$$\lim_{n \rightarrow \infty} P(|H_n - \vartheta| \leq \epsilon) = 1 \quad \forall \epsilon > 0$$

dove $H_n = h(X_1, \dots, X_n)$ è lo stimatore basato su un campione di numerosità n .

Non sempre è facile dimostrare che uno stimatore sia consistente per cui si ha questo teorema in soccorso:

Teorema 11. *Dato uno stimatore corretto $H_n = h(X_1, \dots, X_n)$ risulta anche consistente se vale*

$$\lim_{n \rightarrow \infty} V[H_n] = 0$$

Si ha che \hat{S}_n^2 è uno stimatore corretto di σ^2 ma la sua radice non lo è per la deviazione standard, proprio per tale ragione nelle analisi statistiche troviamo sempre riportate le stime delle varianze anziché quelle delle deviazioni standard.

Per stimare un generico parametro ϑ possono essere definiti diversi stimatori e spesso si stabilisce un ordine di precedenza, con stimatori preferibili ad altri infatti tra due stimatori $H_{1,n}$ e $H_{2,n}$ entrambi corretti, si ha che $H_{1,n}$ è più efficiente di $H_{2,n}$ se:

$$V[H_{1,n}] \leq V[H_{2,n}]$$

per ogni numerosità del campione n e per ogni effettivo valore del parametro ϑ da stimare.

La scelta del miglior stimatore per ogni distribuzione è effettuato tramite il metodo della *verosimiglianza* (likelihood) che ci porta affermare il seguente teorema:

Teorema 12. *Supponiamo di avere un campione X_1, X_2, \dots, X_n di variabili indipendenti esponenziali, la cui funzione di ripartizione è definita eccetto un parametro sconosciuto θ , aventi ognuna media sconosciuta θ .*

La massima verosimiglianza di θ su un campione di variabili esponenziali, è

Dimostrazione. La funzione di ripartizione di un campione di variabili indipendenti esponenziali X_1, X_2, \dots, X_n , aventi tutti media θ , è data dalla seguente formula

$$\begin{aligned} f(X_1, X_2, \dots, X_n) &= f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_n}(x_n) \text{ con } 0 < x_i < \infty \text{ si può asserire che} \\ &= \frac{1}{\theta} e^{-\frac{x_1}{\theta}} \dots \frac{1}{\theta} e^{-\frac{x_n}{\theta}} \end{aligned}$$

L'obiettivo è stimare θ mediante i dati di X_1, X_2, \dots, X_n e il metodo della massima verosimiglianza consiste nel trovare il massimo valore $\hat{\theta}$ della funzione di verosimiglianza $f(x_1, x_2, \dots, x_n | \theta)$, dove x_i sono i valori osservati dalle variabili aleatorie X_i .

Dai concetti appresi nel corso di analisi sappiamo che $f(x_1, x_2, \dots, x_n)$ e $\log f(x_1, x_2, \dots, x_n)$ presentano il massimo per lo stesso valore di θ per cui si trova il massimo stimatore mediante i logaritmi.

Supponiamo di avere un esperimento di n variabili indipendenti

$$f(X_1$$

□

Da quello che abbiamo appena dimostrato per approssimare la media e la varianza di un campione di variabili, normalmente distribuiti, usiamo le variabili \bar{X} e S^2 , che ovviamente non ci forniscono un valore effettivo e/o sempre verificato da tutta la popolazione ma che la nostra media e varianza sull'intera popolazione si avvicina al valore medio e di varianza calcolato sul campione casuale della popolazione e con n abbastanza grande ciò si avvicina con un elevato grado di verosomiglianza.

6.2 Stime Intervallari

Alle stime puntuali, che non forniscono informazioni sul grado di approssimazione delle stesse, vengono preferite quando possibile determinarle le *stime intervallari* che sono stime espresse sotto forma di intervalli *fiduciari*, all'interno dei quali con buona probabilità si trova il valore vero del parametro da stimare.

Supponiamo che X_1, X_2, \dots, X_n siano un campione di variabili indipendenti normali, con μ e σ^2 sconosciuta, sappiamo che $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ è il massimo stimatore per μ ed essendo \bar{X} una variabile normale, di media μ e varianza $\frac{\sigma^2}{n}$ segue che

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \sqrt{n} \frac{\bar{X} - \mu}{\sigma}$$

è una normale standard e supponiamo di volere che uno stimatore possieda una proprietà con $100(1 - \alpha)$ di confidenza, attraverso cui si può discostarsi dal valore stimato di $\pi m \frac{\alpha}{2}$.

Supponiamo di voler avere un intervallo, con confidenza 95% per cui $\alpha = 0.05$, per la variabile \bar{X} da cui segue che

$$P(-z_{1-\frac{\alpha}{2}} < \sqrt{n} \frac{\bar{X} - \mu}{\sigma} < z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Calcolando $z_{0.975}$ tramite le tavole standard e sostituendo il valore di α otteniamo

$$P(-1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \frac{\sigma}{\sqrt{n}})$$

Moltiplicando per -1 e sostando \bar{X} dall'altra parte dell'espressione otteniamo

$$P(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}})$$

Quindi essendo $\bar{X} = \bar{x}$ possiamo che l'intervallo con confidenza 0.95 di μ è il seguente

$$(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}})$$

In caso si volesse considerare solo una direzione dell'intervallo, ossia considerare solo l'intervallo superiore e/o inferiore, è possibile attraverso considerazioni similari che si ha con confidenza 0.95 i seguenti intervalli

$$(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, +\infty) \text{ intervallo di confidenza superiore}$$

$$(-\infty, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}) \text{ intervallo di confidenza inferiore}$$

Ovviamente questi intervalli sono definiti con $\alpha = 0.05$ ma ovviamente con valori diversi la struttura la stessa cambierà solo il valore $z_{1-\frac{\alpha}{2}}$, che nel nostro caso valeva $z_{0.975} = 1.96$

Supponiamo X_1, X_2, \dots, X_n un campione normale, di parametri sconosciuti μ e σ^2 e vogliamo costruire un $100(1 - \alpha)$ intervallo di confidenza percentuale per μ .

Dato che σ è sconosciuto non possiamo stabilire se

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

sia una variabile normale standard e ponendo

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

per un corollario dimostrato nel capitolo precedente sappiamo che

$$\frac{\sqrt{n}(\bar{X} - \mu)}{s}$$

è una variabile t con $n - 1$ gradi di libertà.

Per la simmetria della funzione di densità di t segue che per ogni $\alpha \in (0, \frac{1}{2})$

$$P(-t_{\frac{\alpha}{2}, n-1} < \frac{\sqrt{n}(\bar{X} - \mu)}{s} < t_{\frac{\alpha}{2}, n-1}) = 1 - \alpha$$

Effettuando dei normali e semplici passaggi algebrici e osservando che $\bar{X} = \bar{x}$ e $S = s$ allora possiamo stabilire l'intervallo di confidenza $100(1 - \alpha)$ come

$$\mu \in (\bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}})$$

Il nostro calcolo dell'intervallo di confidenza $100(1 - \alpha)$ per una popolazione di media μ ha assunto che la distribuzione sia normale ma in caso ciò non lo sia per $n > 30$, per il teorema del limite centrale, è possibile approssimarlo ad una normale standard.

Data una distribuzione normale di parametri sconosciuti μ e σ^2 possiamo costruire un intervallo di confidenza σ^2 , usando il fatto che

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi_{n-1}^2$$

da cui segue che

$$P(-\chi_{1-\frac{\alpha}{2}, n-1}^2 \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi_{1-\frac{\alpha}{2}, n-1}^2) = 1 - \alpha$$

Effettuando degli elementari passaggi algebrici e osservando che $S^2 = s^2$ otteniamo che un intervallo di confidenza per σ^2 è dato come

$$\left(\frac{(n-1)s^2}{\chi_{\frac{\alpha}{2}, n-1}^2}, \frac{(n-1)s^2}{\chi_{1-\frac{\alpha}{2}, n-1}^2} \right)$$

In maniera simile è possibile definire gli intervalli di confidenza superiore ed inferiore.

Teorema 13. Sia X_1, X_2, \dots, X_n un campione normale di lunghezza n , con media μ_1 e varianza σ_1^2 e sia Y_1, Y_2, \dots, Y_m un campione normale di lunghezza m , con media μ_2 e varianza σ_2^2 .

Supponiamo che i due campioni siano indipendenti dagli altri e noi siamo interessati a stimare $\mu_1 - \mu_2$.

Dato che $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$ e $\bar{Y} = \sum_{i=1}^m \frac{Y_i}{m}$ sono la massima verosimiglianza per μ_1 e μ_2 segue che $\bar{X} - \bar{Y}$ è la massima verosimiglianza per $\mu_1 - \mu_2$.

Dimostrazione. Per poter ottenere un intervallo di confidenza per $\mu_1 - \mu_2$ dobbiamo determinare la distribuzione di $\bar{X} - \bar{Y}$.

Sapendo che $\bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{n})$ e $\bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{m})$ per il teorema sulla somma di variabili normali segue che

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

Assumendo di conoscere σ_1^2 e σ_2^2 segue che

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

Per quello che si conosce sulla distribuzione normale e degli stimatori segue che

$$P(-z_{\frac{\alpha}{2}} < \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} < z_{\frac{\alpha}{2}}) = 1 - \alpha$$

Attraverso dei passaggi algebrici e supponendo che $\bar{X} = \bar{x}$ e $\bar{Y} = \bar{y}$ allora un intervallo di confidenza $100(1 - \alpha)$ per $\mu_1 - \mu_2$ è il seguente

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

Gli intervalli per un solo verso di $\mu_1 - \mu_2$ sono ottenuti in maniera similare.

Supponiamo ora invece che i valori σ_1^2 e σ_2^2 sono sconosciuti e sapendo che S^2 è la funzione di massima verosimiglianza per σ^2 possiamo ottenere che la variabile che definisce il nostro intervallo di confidenza viene fornita dalla variabile

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

Per utilizzare il risultato provato precedentemente abbiamo bisogno di sapere la distribuzione, che non deve dipendere dai parametri sconosciuti σ_1^2 e σ_2^2 . Sfortunatamente questa distribuzione è difficile da ottenere ed inoltre solo con $\sigma_1^2 = \sigma_2^2$ che siamo in grado di determinare un intervallo.

Supponiamo che la varianza, ancora sconosciuta, sia unica ed uguale a σ^2 e per il teorema sulla distribuzione di una varianza segue che $(n-1)\frac{S_1^2}{\sigma^2} \sim \chi_{n-1}^2$ e $(m-1)\frac{S_2^2}{\sigma^2} \sim \chi_{m-1}^2$.

Dato che il campione è indipendente, segue che le 2 variabili aleatorie sono indipendenti e sapendo che la somma di variabili chi_quadro è anch'essa una chi_quadro segue che

$$(n-1)\frac{S_1^2}{\sigma^2} + (m-1)\frac{S_2^2}{\sigma^2} \sim \chi_{n+m-2}^2$$

Sapendo che $\bar{X} - \bar{Y}$ è una normale otteniamo che

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

Da questo fatto otteniamo che

$$S_p^2 = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}}$$

ha distribuzione t con $n + m - 2$ gradi di libertà.

Per le proprietà della distribuzione t segue che

$$P(-t_{\frac{\alpha}{2}, n+m-2} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \leq t_{\frac{\alpha}{2}, n+m-2}) = 1 - \alpha$$

Ponendo $\bar{X} = \bar{x}$, $\bar{Y} = \bar{y}$ e $S_p = s_p$ otteniamo l'intervallo di confidenza

$$(\bar{x} - \bar{y} - t_{\frac{\alpha}{2}, n+m-2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{x} - \bar{y} + t_{\frac{\alpha}{2}, n+m-2} s_p \sqrt{\frac{1}{n} + \frac{1}{m}})$$

Gli intervalli di confidenza con solo un verso sono ottenuti in maniera simile.

□

Capitolo 7

Ipotesi Parametriche

Supponiamo di avere un campione di una popolazione, specificato eccetto per un vettore di parametri sconosciuti che deve essere osservato ed invece di stimare i parametri sconosciuti, analizzato e considerato nel capitolo precedente, in questo capitolo verifichiamo/testiamo alcune ipotesi, affermazioni riguardo un insieme di parametri della popolazione, sul campione analizzato. L'obiettivo di questo capitolo è quello di determinare se un campione casuale è consistente con l'ipotesi fatta per cui consideriamo una popolazione, con distribuzione F_0 , dove abbiamo il parametro sconosciuto θ , e supponiamo di testare una specifica ipotesi riguardo θ , denotata con H_0 e chiamata *ipotesi nulla*.

Se F_0 è distribuita secondo una normale, con media θ e varianza pari a 1 allora due possibili ipotesi sono:

- $H_0 : \theta = 1$: specifica la distribuzione della popolazione e quindi si definisce *ipotesi semplice*
- $H_0 : \theta \leq 1$: non specifica la distribuzione della popolazioni, per cui si definisce *ipotesi composta*

Supponiamo di testare l'ipotesi nulla H_0 , osservando un campione X_1, X_2, \dots, X_n ed attraverso i valori delle realizzazioni del campione dobbiamo decidere se accettare o meno l'ipotesi H_0 .

Un test su un ipotesi H_0 consiste nel definire la regione C in uno spazio n -esimo in cui l'ipotesi viene accettata se il campione X_1, X_2, \dots, X_n è fuori dalla regione altrimenti viene rifiutata.

Praticamente H_0 viene accettato se $(X_1, X_2, \dots, X_n) \notin C$ mentre H_0 viene rifiutato se $(X_1, X_2, \dots, X_n) \in C$.

Un test comune su $\theta = 1$, media di una normale con varianza 1, ha la regione critica data da

$$C = \{(X_1, X_2, \dots, X_n) : \left| \frac{\sum_{i=1}^n X_i - 1}{\sqrt{n}} \right| > \frac{1.96}{\sqrt{n}}\}$$

Questo test viene rifiutato quando la media campionaria differisca da 1 per più di $\frac{1.96}{\sqrt{n}}$.

Quando si effettua un test di ipotesi possono generarsi due tipologie di errori:

- *errore di prima specie* se il test ci porta a rifiutare H_0 mentre in realtà sarebbe corretto
- *errore di seconda specie* se il test ci porta ad accettare H_0 mentre in realtà sarebbe da rifiutare

L'obiettivo della statistica inferenziale è quello di stabilire se H_0 è consistente rispetto ai dati e sembra ragionevole che H_0 sia rifiutato se i dati sono distanti quando H_0 è vero.

Il modo classico di verificarlo consiste nello specificare un valore α e richiediamo che il test ha la possibilità che se H_0 sia vero, allora la probabilità di effettuare un errore, e quindi rifiutarlo, è minore o uguale a α .

Il valore α viene chiamato *valore di confidenza* del test e viene settato in anticipo, con valori soliti 0.01, 0.05 e 0.001.

Per stabilire se l'ipotesi nulla $H_0 : \alpha \in w$ è corretta si effettuano i seguenti due passi:

- determinare lo stimatore di α , chiamata $d(X)$, in cui si rifiuta H_0 se $d(X)$ è distante dalla regione w
- determinare la distribuzione di $d(X)$ quando H_0 risulta vero, dato che ciò ci permette di stabilire l'appropriata regione critica per fare il test, con significato α .

Supponiamo che X_1, X_2, \dots, X_n siano un campione normale di lunghezza n , avente media sconosciuta μ e varianza anch'essa sconosciuta σ^2 e supponiamo inoltre di essere interessati all'ipotesi nulla $H_0 : \mu = \mu_0$ contro l'ipotesi alternativa $H_1 : \mu \neq \mu_0$, dove μ_0 è una costante significativa.

Dato che $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$ è lo stimatore naturale di μ è ragionevole accettare H_0 se non è troppo distante da μ_0 , per cui la regione critica deve essere, per qualche c stabilito,

$$C = \{X_1, X_2, \dots, X_n : |\bar{X} - \mu_0| > c\}$$

Se vogliamo affermare che il test ha significato α , allora dobbiamo determinare il valore c , tale per cui si rende l'errore di prima specie uguale a α per cui risulta ovviamente

$$P_{\mu_0}(|\bar{X} - \mu_0| > c) = \alpha$$

Quando risulta $\mu = \mu_0$ risulta \bar{X} una normale con media μ_0 e varianza σ^2 per cui

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

Sostituendo questo risultato nella formula precedente otteniamo che

$$P(|Z| > \frac{c\sqrt{n}}{\sigma}) = \alpha$$

Sapendo che $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ e con passaggi algebrici otteniamo che

$$c = \frac{z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}$$

Il test di significato α viene rifiutato se risulta $|\bar{X} - \mu_0| > \frac{z_{\frac{\alpha}{2}} \sigma}{\sqrt{n}}$ altrimenti viene accettato.

Un corretto livello di significato da usare dipende dalle circostanze, infatti se ad esempio rifiutare un'ipotesi costa un'enormità, che sarebbe uno spreco in caso di un errore, allora si usano dei valori α molto piccoli mentre in caso siamo sicuri della correttezza di H_0 allora per avere un'evidenza del contrario usiamo dei valori di α abbastanza grandi.

Dall'equazione precedente segue che possiamo determinare se accettare o meno l'ipotesi nulla attraverso la computazione prima del valore statistico v e poi della probabilità che una normale unitaria sia in valore assoluto $> v$. Questa probabilità, chiamata p -value di un test, fornisce il livello critico di accettazione, nel senso che H_0 è accettato se il livello di accettazione α è minore o uguale al p -value altrimenti viene rifiutato.

Il livello di accettazione di solito non viene settato in anticipo ma lo si cerca di determinare, quando si effettua l'analisi e il test dell'ipotesi.

La probabilità di avere un errore di seconda specie, in cui accettiamo l'ipotesi nulla H_0 quando $\mu \neq \mu_0$, dipende dal valore di μ per cui definiamo $\beta(\mu)$ come

$$\begin{aligned} \beta(\mu) &= P_{\mu}(H_0 \text{ è accettato}) \\ &= P_{\mu}\left(\left|\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| \leq z_{\frac{\alpha}{2}}\right) \\ &= P_{\mu}\left(-z_{\frac{\alpha}{2}} \leq \left|\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| \leq z_{\frac{\alpha}{2}}\right) \end{aligned}$$

La funzione $\beta(\mu)$ è chiamata la curva *caratteristica operativa* (OC) e rappresenta la probabilità che H_0 viene accettata quando la media vera è μ . Attraverso dei passaggi algebrici e della definizione di approssimazione di una normale la valutazione della funzione β consiste nella differenza della valutazione, mediante le tavole normali standard, di $\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_{\frac{\alpha}{2}}$ e $\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} - z_{\frac{\alpha}{2}}$. Per un livello fissato α , la curva OC risulta simmetrica rispetto a μ_0 e dipenda da μ solo attraverso $\frac{\sqrt{n}}{\sigma}|\mu - \mu_0|$.

La funzione $1 - \beta(\mu)$ è chiamata la funzione *potenza* del test, ossia per un dato μ , la potenza del test è uguale alla proprietà di rifiuto quando μ è il valore corretto.

La funzione OC è utile per determinare la lunghezza del campione casuale per cui possiamo fare alcune specifiche sugli errori di seconda specie.

Nell'effettuare il test sull'ipotesi nulla $H_0 : \mu = \mu_0$ abbiamo scelto un test che rifiuta quando \bar{X} è distante da μ_0 con valore minore o maggiore di μ_0 . Se in caso volessimo avere come ipotesi alternativa $H_1 : \mu > \mu_0$, in cui non vogliamo rifiutare H_0 quando \bar{X} sia minore di μ_0 , dato che in quel caso è più verosimile che H_0 sia corretto rispetto a H_1 .

Per questo motivo quando si effettua un test $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$ la regione critica deve essere

$$C = \{(X_1, X_2, \dots, X_n) : \bar{X} - \mu_0 > c\}$$

Dato che la probabilità di rifiutare deve essere uguale a α quando H_0 è vero dobbiamo richiedere che

$$P(\bar{X} - \mu_0 > c) = \alpha$$

ma dato che $Z = \frac{\sqrt{n}\bar{X} - \mu_0}{\sigma} \sim N(0, 1)$ quando H_0 è vero, segue che

$$P_{\mu_0}(\bar{X} - \mu_0) = P(Z > \frac{c\sqrt{n}}{\sigma}) = \alpha \text{ quando } Z \sim N(0, 1)$$

Essendo $P(Z > z_\alpha) = \alpha$ notiamo che $c = \frac{z_\alpha \sigma}{\sqrt{n}}$. Il test di ipotesi H_0 analizzato ora viene rifiutato se $\bar{X} - \mu_0 > \frac{z_\alpha \sigma}{\sqrt{n}}$ altrimenti viene accettato ossia equivalentemente risulta che H_0 è accettato se $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) \leq z_\alpha$ altrimenti viene accettato H_1 . Questa viene chiamata regione critica a solo un verso e il corrispondente test viene chiamato *test a un solo verso*.

Per calcolare il p -value nell'equazione precedente prima usiamo i dati effettivi per calcolare il valore della statistica $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$ e il p -value è uguale alla probabilità che l'ipotesi H_0 sia verificata, per cui nel caso di un campione normalmente distribuito deve essere almeno largo

La funzione OC per i test a un solo verso è definita come

$$\begin{aligned}
 \beta(\mu) &= P_\mu(H_0 \text{ verificato}) \\
 &= P_\mu(\bar{X} \leq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}) \\
 &= P_\mu\left(\frac{\bar{X} - \mu}{\sigma} \sqrt{n} \leq \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} + z_\alpha\right) \\
 &= P\left(Z \leq \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right)
 \end{aligned}$$

Risulta che la funzione $\beta(\mu)$ viene calcolata usando le tavole standard della normale, come avevamo visto nella definizione della funzione β per i test a doppio verso.

All'aumento del valore del suo argomento, la funzione $\beta(\mu)$ porta a decrescere in μ e questo proviene dall'intuizione che è ragionevolmente vero che lontani dalla media μ è molto meno probabile che si conclude $\mu \leq \mu_0$ ed inoltre risulta $\beta(\mu_0) = 1 - \alpha$. Il test dato dall'equazione 8.3.10., disegnato per il test $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$ può essere usato come test, con significatività α , per il test di ipotesi ad una grandezza $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$.

Per verificare che rimane un test di confidenza α dobbiamo mostrare che la probabilità di rifiuto di H_0 non sarà più grande di α quando H_0 sarebbe vero. Per fare questo bisogna mostrare che $1 - \beta(\mu) \leq \alpha$ per ogni $\mu \leq \mu_0$ ma abbiamo già mostrato nell'equazione 8.3.10 che $\beta(\mu)$ diminuisce in μ e $\beta(\mu_0) = 1 - \alpha$ da cui siamo in grado di ottenere $\beta(\mu) \geq \beta(\mu_0) = 1 - \alpha$ per ogni valore $\mu \leq \mu_0$. Oltre alla definizione del test a un verso del tipo $H_0 : \mu \leq \mu_0$, è possibile definire un test ad un solo verso del tipo $H_0 : \mu = \mu_0 (\mu \geq \mu_0)$ con test alternativo $H_1 : \mu < \mu_0$, avente significato α , tale per cui H_0 risulta accettato se $\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu_0) \geq -z_\alpha$ altrimenti H_0 è rifiutato.

Questo test può essere effettuato in maniera alternativa, calcolando prima il valore statistico $\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$ e poi calcolare il p -value, indicante la probabilità che l'ipotesi H_0 sia verificata, per cui si rifiuta l'ipotesi H_0 in caso il nostro livello di significato α sia maggiore o uguale al p -value.

Fino ad ora abbiamo supposto che σ^2 non fosse sconosciuto ma il caso comune avviene con sia μ e σ^2 sconosciuti.

Supponiamo ora di considerare un test di ipotesi $H_0 : \mu = \mu_0$ con l'alternativa $H_1 : \mu \neq \mu_0$ e si dovrebbe notare che l'ipotesi nulla H_0 non è più semplice, in quanto non conosciamo la distribuzione di σ^2 .

Come prima, sembra ragionevole rigettare H_0 quando \bar{X} è distante da μ ma la distanza necessaria per rifiutarlo dipende dal valore di σ^2 .

Sapendo che σ^2 ha come massimo stimatore la variabile

$$S^2 = \frac{\sum_{i=0}^n (\bar{X} - \mu)^2}{n-1}$$

da cui segue che si rifiuta H_0 quando è largo $|\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}|$.

Per stabilire per quale valore dobbiamo rifiutare, al fine di avere un test con significato α , dobbiamo determinare la funzione di ripartizione di $|\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}|$ quando H_0 è verificato.

Come già notato e dimostrato nel capitolo precedente la statistica T , definita come

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S}$$

quando $\mu = \mu_0$ risulta una t -student con $n-1$ gradi di libertà.

Da questo aspetto si ricava che

$$P(-t_{\frac{\alpha}{2}, n-1} \leq \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \leq t_{\frac{\alpha}{2}, n-1}) = 1 - \alpha$$

dove $t_{\frac{\alpha}{2}, n-1}$ è il $100(\frac{\alpha}{2})$ quantile della distribuzione t con $n-1$ gradi di libertà.

Dall'ultima equazione si può notare che un test con significato α di $H_0 : \mu = \mu_0$ con ipotesi alternativa $H_1 : \mu \neq \mu_0$, con σ^2 sconosciuto che ci pone ad affermare che H_0 risulta accettato se $|T| \leq t_{\frac{\alpha}{2}, n-1}$ altrimenti H_0 risulta rifiutato.

È possibile effettuare un test t ad un solo verso per testare l'ipotesi $H_0 : \mu = \mu_0 (\mu \leq \mu_0)$ con ipotesi alternativa $H_1 : \mu > \mu_0$ avente significato α , in cui H_0 risulta accettato se $|T| \leq t_{\frac{\alpha}{2}, n-1}$ altrimenti si rifiuta H_0 .

Il test avente significato α , $H_0 : \mu = \mu_0 (\mu \geq \mu_0)$ con ipotesi alternativa $H_1 : \mu < \mu_0$ porta ad accettare H_0 se $\frac{\sqrt{n}(\bar{X}-\mu)}{S} \geq -t_{\frac{\alpha}{2}, n-1}$ altrimenti H_0 viene rifiutato.

Una situazione normale è stabilire se due campioni, aventi distribuzione normale, di una popolazione hanno la stessa media, per questo consideriamo il test di ipotesi $H_0 : \mu_x = \mu_y$, con ipotesi alternativa $H_1 : \mu_x \neq \mu_y$, effettuato sui campioni (X_1, X_2, \dots, X_n) e (Y_1, Y_2, \dots, Y_m) indipendenti provenienti da una popolazione normalmente distribuita, aventi media sconosciuta μ_x e μ_y ma varianze conosciute σ_x^2 e σ_y^2 .

Essendo \bar{X} il naturale stimatore di μ_x e \bar{Y} il naturale stimatore di μ_y , segue che $\bar{X} - \bar{Y}$ può essere usato per stimare $\mu_x - \mu_y$.

In quanto il test di ipotesi può essere scritto come $H_0 : \mu_x - \mu_y = 0$ sembra ragionevole rifiutare $\bar{X} - \bar{Y}$ quando è distante da zero.

La forma del test dovrebbe portarci ad accettare H_0 se $|\bar{X} - \bar{Y}| \leq c$ altrimenti a rifiutare H_0 e per determinare il valore di c , necessario per verificare il test di significato α , dobbiamo conoscere la distribuzione di $\bar{X} - \bar{Y}$ quando H_0 è vero.

Come abbiamo già dimostrato nel capitolo precedente, abbiamo che $\bar{X} - \bar{Y} \sim N(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m})$ che implica che

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0, 1)$$

e quindi si ottiene la probabilità di effettuare un errore di prima specie attraverso

$$P_{H_0}(-z_{\frac{\alpha}{2}} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

Da questa equazione otteniamo che H_0 risulta accettato se

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \leq z_{\frac{\alpha}{2}}$$

altrimenti H_0 viene rifiutato.

Per il test di ipotesi $H_0 : \mu_x = \mu_y$ ($\mu_x \leq \mu_y$), con ipotesi alternativa $H_1 : \mu_x > \mu_y$ si decide di accettare H_0 se $\bar{X} - \bar{Y} \leq z_{\alpha} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$ ed ovviamente rifiutarlo in caso contrario.

Supponiamo ora che i due campioni (X_1, X_2, \dots, X_n) e (Y_1, Y_2, \dots, Y_m) abbiano tutti i parametri sconosciuti, anche nelle due varianze σ_x^2 e σ_y^2 , e di voler analizzare il test d'ipotesi $H_0 : \mu_x = \mu_y$ con ipotesi alternativa $H_1 : \mu_x \neq \mu_y$.

Per poter determinare il test di ipotesi nulla H_0 effettuiamo l'assunzione che le varianze sconosciute σ_x^2 e σ_y^2 siano uguali, scrivendo σ^2 come valore comune. Come prima si rifiuta H_0 quando $\bar{X} - \bar{Y}$ sono distanti da zero e per poterlo determinare definiamo le variabili S_x^2 e S_y^2 come gli stimatori delle varianze dei due campioni, la cui definizione è uguale a quella di S^2 , e come avevamo già notato nel capitolo precedente abbiamo che

$$\frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \sim t_{n+m-2}$$

dove S_p^2 è la pooled variance è data da

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

Quando H_0 è verificato e quindi $\mu_x = \mu_y$ la variabile statistica

$$T \equiv \frac{(\bar{X} - \bar{Y})}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}}$$

ha una distribuzione t , avente $n+m-2$ gradi di libertà.

Viene accettata l'ipotesi H_0 in caso $|Z| \leq t_{\frac{\alpha}{2}, n+m-2}$ altrimenti lo si rifiuta.

In maniera alternativa è possibile determinare l'accettazione del test mediante il p -value, in cui si pone prima $v = T$ e il p -value è definito come

$$\begin{aligned} p - value &= P(|T_{n+m-2}| \geq |v|) \\ &= 2P(T_{n+m-2} \geq |v|) \end{aligned}$$

Ovviamente, essendo il p -value la probabilità che l'ipotesi H_0 sia verificata, il p -value deve essere superiore ad α per avere l'accettazione del test, con confidenza $100(1 - \alpha)$.

Supponiamo che la varianza σ_x^2 e σ_y^2 non solo sono sconosciute ma anche diseguali ed essendo S_x^2 e S_y^2 i normali stimatori di σ_x^2 e σ_y^2 sembra ragionevole basare il nostro test $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x \neq \mu_y$ sulla statistica

Capitolo 8

Ipotesi non Parametriche

Nel precedente capitolo abbiamo visto ed analizzato i test di ipotesi effettuati su dei parametri di una popolazione, su cui si effettuano delle assunzioni sulla tipologia di distribuzioni ma ora ci occupiamo dei test di ipotesi non parametriche, in cui non si effettua nessuna assunzione sulla distribuzione del campione/popolazione.

In questo capitolo verranno introdotti i seguenti test:

- *test per la bontà dell'adattamento*, che servono per verificare se una popolazione segue una distribuzione prestabilita
- *test per confrontare le distribuzioni di due popolazioni*, quando queste non sono necessariamente delle normali
- test per verificare se sussiste *indipendenza o incorrelazione* tra due diversi caratteri di una popolazione

Anche per questi test valgono le considerazioni viste in merito agli errori di prima e seconda specie ed alle regioni critiche e di accettazione nel capitolo precedente sulle ipotesi parametriche.

8.1 Test per la bontà dell'adattamento

I test relativi alla bontà dell'adattamento pongono a capire se una popolazione X è distribuita secondo una funzione di ripartizione F , ossia sono dei test la cui ipotesi nulla è $H_0 : F_X(t) = F(t) \forall t \in \mathbb{R}$ con ipotesi alternativa $H_1 : F_X(t) \neq F(t)$ per almeno un t in \mathbb{R} .

I due test riguardanti la bontà dell'adattamento analizzati in questo corso sono i seguenti:

- *test di Kolmogorov-Smirnov* si basa sulla popolazione empirica, introdotta nel seguito del paragrafo, e sulla statistica

$$D_n = \sup_{t \in \mathbb{R}} |F(t) - \bar{F}_{X,n}(t)|$$

che specifica l'estremo superiore delle distanze in valore assoluto tra la funzione di ripartizione che vogliamo controllare come possibile per X e quella empirica, ottenuta tramite il campione disponibile.

Si può dimostrare che quando vale H_0 e F è una funzione continua allora tale statistica è indipendente dalla forma di F , ovvero ha la stessa distribuzione qualunque sia F e pertanto è possibile fissare delle regioni di accettazione e critiche al variare di n e del livello di significatività.

La distribuzione della statistica campionaria D_n al variare di n è stata studiata dagli inventori, che hanno fornito apposite tabelle per determinare i quantili.

Si può notare che è logico aspettarsi che D_n assume valori piccoli se l'ipotesi nulla è vera, ed assume valori grandi quando H_0 è falsa.

Infatti, per ogni ampiezza α del test la regione critica risulta essere

$$C = (d_{1-\alpha}, 1]$$

dove il quantile $d_{1-\alpha}$ è quel valore per cui risulta

$$P(D_n \leq d_{1-\alpha}) = 1 - \alpha$$

che può essere determinato sulle apposite tavole (si noti che D_n non può sicuramente assumere valori maggiori di 1)

Si osservi che la massima distanza tra la funzione di ripartizione F e la funzione di ripartizione empirica viene raggiunta sempre in corrispondenza di uno dei salti della distribuzione empirica.

Questo fatto può essere utile nella determinazione di D_n , quando non è possibile fare uso di rappresentazioni grafiche.

Osserviamo che il test di Kolmogorov-Smirnov non richiede particolari assunti sui dati o sulla distribuzione F per essere impiegato, se non quello di continuità della F stessa.

Dato un campione (X_1, X_2, \dots, X_n) estratto da una popolazione, si definisce *funzione di ripartizione empirica* della popolazione X , basata sul campione (X_1, X_2, \dots, X_n) , la funzione aleatoria

$$\bar{F}_{X,n}(t) = \frac{1}{n} \sum_{i=1}^n U_{(-\infty, t]}(X_i) \forall t \in \mathbb{R}$$

dove si definisce

$$U_{-\infty, t]}(X_i) = \begin{cases} 1 & \text{se } X_i \in (-\infty, t] \\ 0 & \text{altrimenti} \end{cases}$$

Si osservi che le funzioni di ripartizione empiriche sono sempre delle funzioni crescenti a gradino ed è facile rendersi conto poi che al crescere della numerosità n del campione la funzione di ripartizione empirica assomiglia sempre di più alla reale funzione di ripartizione della popolazione, fino a coincidere quando n corrisponde alla numerosità dell'intera popolazione.

- *test del chi-quadro* può essere usato senza porre condizioni sulla funzione di ripartizione F . Anche per poter descrivere questo test occorre introdurre alcune nozioni e notazioni e come al solito supporremo di poter estrarre un campione (X_1, X_2, \dots, X_n) di numerosità n . Un approccio classico per ottenere la bontà di adattamento di un test consiste nel partizionare i valori possibili di una variabile aleatoria in un numero finito di regioni, i cui il numero di valori presenti viene poi determinato e valutato con il valore atteso sotto una specifica distribuzione probabilistica e quando sono significativamente distanti l'ipotesi nulla H_0 viene rifiutata.

Supponiamo che n variabili aleatorie (Y_1, Y_2, \dots, Y_n) ognuna avente uno dei valori $1 \dots k$, vengono osservate e noi siamo interessati all'ipotesi nulla $H_0 : P(Y = i) = p_i, i = 1 \dots k$ vs $H_1 : P(Y = i) \neq p_i$.

Per effettuare il seguente test poniamo $X_i, i = 1 \dots k$ a denotare il numero di Y_j che sono uguali a i e ogni Y_j indipendenti tra di loro è uguale a i con probabilità $P(Y = i)$.

Segue che sotto H_0 , X_i è una binomiale con parametri n e p_i ed in caso di H_0 verificato si ha $E[X_i] = np_i$ e quindi $(X_i - np_i)^2$ indica quanto è simile che p_i è uguale a $P(Y = i)$ e quando questo è largo in relazione con np_i allora H_0 non è corretta per cui consideriamo il seguente test statistico

$$T = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i}$$

Per determinare la regione critica dobbiamo prima specificare un livello di significato α e poi determinare un valore critico c tale che $P_{H_0}(T \geq c) = \alpha$.

Il test è portato a rifiutare l'ipotesi nulla, con un livello di confidenza α quando si ha $T \geq c$ altrimenti accetta con $T < c$.

L'approccio classico per determinare c consiste nell'usare il risultato che

con n largo e H_0 verificato, T viene approssimato con una chi-quadro con $k - 1$ gradi di libertà

8.2 Test per l'uguaglianza di distribuzione

Di seguito vengono presentati 3 test per verificare o rifiutare ipotesi che le distribuzioni di due distinte popolazioni X ed Y siano identiche, non necessariamente gaussiane.

In questo paragrafo siamo interessati ad effettuare il test $H_0 : F_X(t) = F_Y(t) \forall t \in \mathbb{R}$ con ipotesi alternativa $H_1 : F_X(t) \neq F_Y(t) \exists t \in \mathbb{R}$ e lo effettuiamo mediante tre diversi possibili test:

- *sign test*, basato sui segni negativi o positivi delle differenze tra le coppie di elementi presi dai campioni estratti dalle due popolazioni considerate.

Siano (X_1, X_2, \dots, X_n) e (Y_1, Y_2, \dots, Y_n) due campioni appaiati, estratti da X e da Y e ciò significa che la coppia (X_i, Y_i) è relativa allo stesso individuo, ossia ad esempio una valutazione prima e dopo un avvenimento/esperimento.

Stabiliamo ora le seguenti tre quantità:

- $S^+ = \#i : X_i > Y_i$
- $S^- = \#i : X_i < Y_i$
- $S^= = \#i : X_i = Y_i$

Se l'ipotesi nulla H_0 è verificata è logico aspettarsi che le quantità S^+ e S^- non si discostano molto tra di loro, infatti se H_0 è vera e se $n - S^= \geq 10$ la quantità S_n risulta

$$S_n = S^+ - S^- \sim N(0, \frac{n - S^=}{2})$$

Da questa considerazione è possibile costruire una regione critica per S_n ragionando nel solito modo rifiutando H_0 quando S_n si discosta troppo dallo zero per essere una normale e ragionando come nel capitolo sulla stima dei parametri, la regione di accettazione, con confidenza α , per S_n risulta essere

$$(-z_{1-\frac{\alpha}{2}} \sqrt{\frac{n - S^=}{2}}, +z_{1-\frac{\alpha}{2}} \sqrt{\frac{n - S^=}{2}})$$

dove $z_{1-\frac{\alpha}{2}}$ risulta essere il quantile di ordine $1 - \frac{\alpha}{2}$ della normale standardizzata.

È possibile effettuare il test anche quando la quantità $n - S^=$ è piccola ed in questo caso si può considerare la statistica S^+ e tener conto che quando sussiste l'ipotesi H_0 essa è distribuita come una binomiale di parametri $n - S^=$ e $\frac{1}{2}$.

Segnaliamo l'esistenza di un test simile a quello dei segni ma più elaborato, nel senso che tiene conto non solo dei segni delle differenze ma anche delle ampiezze di queste differenze, e che porta il nome di *test dei segni di Wilcoxon*.

- *Wilcoxon ranked sign test* che è molto utile da usare anche per campioni non appaiati ed è molto potente.

Fornire un'espressione funzionale della statistica su cui questo test è basato è abbastanza complesso pertanto ci limiteremo a descrivere il procedimento da seguire per effettuare il test senza cercare di fornire un'interpretazione anche solo intuitiva di esso.

Siano quindi (X_1, X_2, \dots, X_n) e (Y_1, Y_2, \dots, Y_m) due campioni di lunghezza n e m , estratti rispettivamente da X e da Y .

Le fasi di procedura del test sono i seguenti:

- Si ordina in senso crescente insieme di tutti i dati e si associa a ciascun dato il proprio rango, ovvero la posizione in cui si trova nella sistemazione in ordine crescente dei dati.
- Si sommano separatamente i ranghi relativi ai due campioni e siano R_X e R_Y le loro somme.
- Si calcolano le statistiche $U_X = n * m + \frac{n(n-1)}{2} - R_X$ e $U_Y = n * m + \frac{n(n-1)}{2} - R_Y$ e se i conti sono corretti la somma di U_X ed U_Y deve essere uguale al prodotto tra le numerosità dei due campioni, vale a dire che deve essere $U_X + U_Y = n * m$.
- Si considera poi la statistica $U = \min(U_X, U_Y)$ e si può dimostrare che per n ed m sufficientemente grandi, in genere maggiori di 8, quando vale H_0 la U è approssimabile tramite una normale con parametri $\mu_U = \frac{n*m}{2}$ e $\sigma_U^2 = \frac{n*m*(n+m+1)}{12}$.

Tramite quello visto ed analizzato nei capitoli precedenti si riesce a mostrare che l'intervallo di accettazione del test, con significato α , risulta essere $(-z_{1-\frac{\alpha}{2}}, +z_{1-\frac{\alpha}{2}})$.

Nel caso in cui non sia verificata la condizione $n, m \geq 8$ allora è possibile ricorrere ad apposite tavole per determinare la regione critica per U .

- il terzo test è l'*adattamento al test di Kolmogorov* al caso di due campioni, che non richiede che i due campioni siano appaiati, ma in compenso può essere utilizzato solo quando ci siano validi motivi per pensare che la distribuzione delle popolazioni da confrontare sia continua.

Siano quindi (X_1, X_2, \dots, X_n) e (Y_1, Y_2, \dots, Y_m) due campioni di numerosità n e m , estratti rispettivamente da X e Y e siano poi $F_{X,n}$ e $F_{Y,m}$ le funzioni di ripartizione empiriche di X e Y , ricavate tramite i due campioni e sia poi

$$D_{n,m} = \sup_{t \in \mathbb{R}} |F_{X,n}(t) - F_{Y,m}(t)|$$

la statistica che specifica l'estremo superiore delle distanze, in valore assoluto, tra le due funzioni di ripartizione empiriche.

Anche in questo caso si può dimostrare che quando H_0 è vera e quando F_X è una funzione continua, allora $D_{n,m}$ è indipendente dalla forma di F_X , ovvero ha la stessa distribuzione qualunque sia F_X .

Anche per la distribuzione della statistica campionaria $D_{n,m}$ al variare di n ed m esistono apposite tabelle per determinare i quantili ed è quindi possibile fissare al solito delle regioni di accettazione e critiche al variare di n ed m e del livello di significatività σ , in maniera uguale a quelli stabiliti per il test Kolmogorov-Smirnov nel caso di un solo campione.

8.3 Test per l'indipendenza

Un problema che si pone frequentemente nelle applicazioni è quello di stabilire se due caratteri di una popolazione bidimensionale sono tra loro stocasticamente indipendenti oppure no.

Per rispondere a questa domanda sono stati inventati diversi test ma uno in particolare viene sempre utilizzato ed esso prende il nome di *Test del chi-quadro per l'indipendenza* e come già nome ci porta a pensare ricorda molto la formulazione del test chi-quadro per la bontà di adattamento.

Si consideri una popolazione bidimensionale (X, Y) e si supponga di voler estrarre un campione casuale $((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n))$ e si vuole controllare con livello di significatività α l'ipotesi H_0 : i caratteri X e Y sono indipendenti, con ipotesi alternativa H_1 : i caratteri X e Y non sono indipendenti.

Per effettuare il test del Chi-Quadro occorre inizialmente fare una partizione dei due supporti dei caratteri X ed Y in un numero finito di intervalli ciascuno e definiamo I_k^X $k \in 1, 2, \dots, M_X$ e I_j^Y $j \in 1, 2, \dots, M_Y$ gli intervalli che definiscono una partizione del supporto di X e Y . Per ogni coppia (k, j)

consideriamo le quantità $n_k^X, n_j^Y, n_{k,j}$ indicanti il numero di elementi calcolati sui intervalli I_k^X, I_j^Y e $I_k^X \times I_j^Y$ e su cui è possibile calcolare le relative frequenze relative.

La quantità $f_{k,j}$ è detta *frequenza relativa osservata* delle regioni $I_k^X \times I_j^Y$ e notiamo che per ogni coppia (k, j) il prodotto $f_k * f_j$ fornisce invece una *frequenza relativa attesa* di elementi del campione che devono ricadere nell'intervallo $I_k^X \times I_j^Y$ se fosse vera l'ipotesi nulla H_0 , poichè in questo caso la frequenza di ogni regione deve essere uguale al prodotto delle frequenze marginali di quella regione.

Se H_0 è vero le differenze $f_{k,j} - f_k * f_j$ dovrebbero essere piccole in valore assoluto e consideriamo la statistica

$$W = n \sum_{k=1}^{M_X} \sum_{j=1}^{M_Y} \frac{(f_{k,j} - f_k * f_j)^2}{f_k * f_j}$$

Si può mostrare che quando l'ipotesi nulla è vera e quando le $n_{k,j}$ sono sufficientemente grandi (almeno maggiori o uguali a 5) allora W è approssimativamente distribuita secondo una chi-quadro con $(M_k - 1) * (M_j - 1)$ gradi di libertà.

La regola di decisione del test segue in base alle considerazioni fatte sopra, saremo portati a rifiutare l'ipotesi nulla quando la W assume valori troppo lontani dallo zero per essere una Chi-Quadro con opportuni gradi di libertà. Come nel caso del test Chi-Quadro per la bontà dell'adattamento anche il test Chi-Quadro per l'indipendenza ha il vantaggio di essere utilizzabile quando si considerano caratteri con distribuzioni non continue.

Addirittura esso può essere utilizzato considerando caratteri con modalità non numeriche.

8.4 Test per l'incorrelazione

Il test di incorrelazione che presentiamo ora non è basato sul coefficiente di correlazione lineare di Pearson ma su una statistica campionaria che porta il nome di *coefficiente di correlazione dei ranghi R_s di Spearman*.

Esso viene incluso nei test di tipo non-parametrico in quanto la determinazione di R_s non coinvolge direttamente i valori numerici assunti dai dati campionari ma solo i loro ranghi di cui viene fornita la definizione.

Consideriamo il campione $((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n))$ di numerosità n estratto in modo casuale dalla popolazione bidimensionale (X, Y) .

Ordiniamo in senso crescente prima l'insieme dei dati di tipo X_i e successivamente quelli di tipo Y_i ed è detto *rango* di ciascun dato la posizione che esso assume nella sequenza così ottenuta di dati dello stesso tipo.

Ad ogni coppia (X_i, Y_i) associamo ora la corrispondente coppia di ranghi (r_i^X, r_i^Y) e denotiamo la loro differenza con

$$d_i = r_i^X - r_i^Y$$

Si definisce *coefficiente di correlazione dei ranghi R_s di Spearman* la statistica

$$R_s = 1 - \frac{6 * \sum_{i=1}^n d_i^2}{n^3 - n}$$

Il coefficiente R_s soddisfa proprietà simili a quelle di cui gode il coefficiente di correlazione lineare di Pearson e consente di definire un nuovo tipo di incorrelazione (che chiameremo incorrelazione nel senso di Spearman) che si ha quando R_s assume valore zero.

La possibilità di utilizzare R_s in un test di ipotesi deriva dal fatto che indipendentemente dalla forma della distribuzione congiunta (X, Y) quando i caratteri X ed Y sono incorrelati (nel senso di Spearman) e la numerosità del campione è maggiore di 10 allora la statistica

$$\bar{T}_n = R_s * \sqrt{\frac{n-2}{1-R_s^2}}$$

risulta essere approssimativamente distribuita come una t di Student con $(n-2)$ gradi di libertà.

Dovendo valutare l'ipotesi H_0 : i caratteri X e Y sono incorrelati secondo Spearman con ipotesi alternativa H_1 : i caratteri X e Y sono correlati secondo Spearman e possiamo pensare di accettare H_0 quando \bar{T}_n assume valori non troppo distanti da zero e di rifiutarla in favore di H_1 in caso contrario per cui la nostra ipotesi H_0 viene accettata, con confidenza α , se la statistica \bar{T}_n risiede nell'intervallo $(-t_{1-\frac{\alpha}{2}}, +z_{1-\frac{\alpha}{2}})$ altrimenti si accetta l'ipotesi H_1 .

Capitolo 9

Regressione Lineare

Considero una popolazione $(m + 1)$ dimensionale: $(X_1, X_2, \dots, X_m, Y)$.

In molti casi applicativi risulta particolarmente utile o interessante stabilire se tra i caratteri della popolazione sussistano legami di dipendenza che descrivano uno di essi come espressione funzionale degli altri ovvero se esista una relazione del tipo:

$$Y = f(X_1, X_2, \dots, X_m)$$

dove f può essere indifferentemente deterministica o con parametri aleatori. Trovarsi a conoscenza di una relazione di questo tipo consente infatti di risolvere molti problemi pratici soprattutto quando si incontrano difficoltà a rilevare i valori assunti dal carattere Y .

Un equazione del tipo:

$$Y = f(X_1, X_2, \dots, X_m)$$

viene detta **equazione di regressione del carattere Y rispetto ai caratteri X_1, \dots, X_m** . La Y viene detta **variabile di risposta** mentre le X_k sono dette **variabili esplicative (o indipendenti o regressori)**.

Con **Analisi di regressione** si intende la verifica dell'ammissibilità dell'equazione e in caso affermativo la determinazione di una stima di f che lega Y alle X_k . Si noti che un'analisi di regressione deve essere effettuata facendo ricorso all'estrazione di un campione di numerosità finita e quindi verificando l'ammissibilità dell'equazione ricorrendo ai metodi della statistica inferenziale, poichè come al solito non possiamo pensare di poter analizzare tutti gli individui appartenenti alla popolazione. Lo stesso vale per la determinazione di una stima di f .

Studiamo solamente l'equazione di regressione, ovvero il caso in cui f sia una funzione lineare nelle variabili esplicative X_k e in un addendo aleatorio

E ovvero il caso in cui la relazione cercata è del tipo:

$$Y = \alpha_0 + \alpha_1 \cdot X_1 + \dots + \alpha_m \cdot X_m + E$$

con le α_k , $k = 1, \dots, m$ costanti realci e $E \sim N(0, \sigma^2)$.

Ci si riferisce alla **regressione lineare semplice** se nella relazione sopra compare una sola variabile esplicativa e alla **regressione lineare multipla** in caso contrario.

Il termine aleatorio E è detto **residuo o errore**. Considerando il valore che ci si deve attendere per la variabile di risposta conoscendo le realizzazioni degli altri caratteri:

$$\hat{Y} = \alpha_0 + \alpha_1 \cdot X_1 + \dots + \alpha_m \cdot X_m$$

il residuo descrive scostamenti da tale valore atteso imputabili a cause aleatorie e non controllabili quali ad esempio, errori di misura, dipendenze da variabili non considerate o valutazioni soggettive.

Notiamo poi una cosa: la scrittura

$$Y = \alpha_0 + \alpha_1 \cdot X_1 + \dots + \alpha_m \cdot X_m + E$$

potrebbe risultare ambigua nel senso che potrebbe portare a pensare che il residuo abbia lo stesso valore per ogni singolo individuo della popolazione. **In realtà il residuo ha solo la stessa distribuzione per ogni singolo individuo.**

Quindi l'equazione:

$$Y = \alpha_0 + \alpha_1 \cdot X_1 + \dots + \alpha_m \cdot X_m + E$$

descrive infatti la distribuzione del carattere Y come distribuzione della somma di $m+1$ variabili, tra cui una che esprime gli scostamenti aleatori. Per ovviare al rischio di incomprensioni preferiamo quindi riscrivere la precedente relazione esplicitando il fatto che tanto il residuo quanto i valori assunti dalle variabili esplicative variano per ogni individuo della popolazione, considerando dora in poi lequazione:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{1,i} + \dots + \alpha_m \cdot X_{m,i} + E_i$$

con $i = 1, \dots, N$ e N numero totale degli individui della popolazione.

Si procede nella seguente maniera: istintivamente verrebbe naturale provare subito a controllare la validità del modello lineare e solo successivamente in caso di risultati positivi determinare le costanti che in esso compaiono, per le seguenti motivazioni:

1. lo statistico non effettua mai un test per vedere se una certa ipotesi è vera, bensì per vedere se tale ipotesi non può essere rifiutata. Nella regressione lineare in genere si assume a priori la validità della relazione:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{1,i} + \dots + \alpha_m \cdot X_{m,i} + E_i$$

e solo successivamente si controlla se i dati portano ad un rifiuto di tale ipotesi

2. i test di ammissibilità di

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{1,i} + \dots + \alpha_m \cdot X_{m,i} + E_i$$

i basano principalmente su particolari statistiche la cui realizzazione può essere determinata solo dopo aver effettuato le stime delle costanti (varianza dei residui e costanti α_i)

9.1 Stima delle Costanti del Modello

Consideriamo una popolazione bidimensionale (X, Y) ed assumiamo che tra i due caratteri esista un legame di dipendenza esprimibile tramite un'equazione di regressione lineare semplice. per ogni individuo $i = 1, \dots, N$ sia quindi:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_i + E_i$$

Cerco quindi i valori delle costanti α_0 e α_1 . Ovviamente, a meno di non disporre dei dati relativi a tutti gli N individui della popolazione, di tali costanti non potremo fare altro che determinarne delle stime numeriche che denoteremo con a_0 e a_1 . Abbiamo un campione di numerosità n :

$$\{(X_1, Y_1); (X_2, Y_2); \dots (X_n, Y_n)\}$$

Si hanno 3 ipotesi per i residui:

1. $E[E_i] = 0 \quad \forall i = 1, 2, \dots, N$
2. $V[E_i] = \sigma^2 \quad \forall i = 1, 2, \dots, N$, che viene detta **ipotesi di omoschedasticità**
3. le variabili E_i sono incorrelate tra loro

Però non conosciamo il valore della varianza che viene per ora comunque indicato come costante nota. Gli stimatori puntuali utilizzati per stimare le costanti sono le statistiche:

$$A_0 = \bar{Y} - A_1 \cdot \bar{X}$$

$$A_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

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

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$$

È possibile osservare che tali stimatori sono definiti analogamente alle costanti q ed m della retta dei minimi quadrati:

$$Y = m \cdot X + q$$

per descrivere al meglio un legame lineare tra due caratteri di una serie di dati.

Gli stimatori A_0 e A_1 vengono determinati allo stesso modo come valori per cui risulta minima la quantità:

$$\sum_{i=1}^n [A_1 \cdot X_i + A_0 - Y_i]^2$$

Supposto quindi che $\{(x_1, y_1); (x_2, y_2); \dots (x_n, y_n)\}$ sia una realizzazione del campione $\{(X_1, Y_1); (X_2, Y_2); \dots (X_n, Y_n)\}$ due stime puntuali di α_0 e α_1 sono:

$$a_0 = \bar{y} - a_1 \cdot \bar{x}$$

$$a_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

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

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

Inoltre gli stimatori, sotto le 3 ipotesi sopra, presentano le seguenti proprietà:

$$E[A_0] = \alpha_0$$

$$V[A_0] = \sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$E[A_1] = \alpha_1$$

$$V[A_1] = \sigma^2 \cdot \left[\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

Quindi A_0 e A_1 sono stimatori corretti e adatti a fare inferenze su α_0 e α_1 . Inoltre, sempre se valgono le 3 ipotesi, si può dire qualcosa di ancora più forte infatti in questo caso A_0 ed A_1 sono, tra gli stimatori corretti per α_0 e α_1 i più efficienti ovvero quelli cui corrisponde varianza minima. Questo è il **Teorema di Gauss-Markov**.

Notiamo che i due stimatori dipendono dalla numerosità n del campione ma anche dalla disposizione sull'asse delle x delle osservazioni x . Se la dispersione di tali osservazioni sull'asse è piccola allora la quantità:

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

risulta piccola e le varianze diventano grandi.

In questo caso allora cresce la probabilità che le stime a_0 e a_1 siano lontane dai valori reali α_0 e α_1 . Per questa ragione sarebbe preferibile avere campioni in cui siano presenti grandi scostamenti dei valori assunti dal carattere X .

Supponiamo ora di voler effettuare delle stime intervallari per le costanti α_0 , α_1 . Per farlo aggiungiamo una nuova ipotesi sui residui: **le variabili E_i sono normalmente distribuite**.

Notiamo che essa aggiunta alle altre equivale a dire che ogni residuo è normalmente distribuito con media nulla ed identica varianza σ^2 . Questa non è una ipotesi molto restrittiva o irrealistica, in effetti è logico aspettarsi che gli errori siano normalmente distribuiti in quanto somma di numerosi fattori casuali indipendenti. Si ricordi infatti che per il Teorema del Limite Centrale la somma di numerosi fattori aleatori tende a distribuirsi come una normale. Per quattro ipotesi:

$$E[E_i] = 0, \quad \forall i = 1, \dots, N$$

$$V[E_i] = \sigma^2, \quad \forall i = 1, \dots, N$$

le variabili E_i sono incorrelate tra loro
le variabili E_i sono normalmente distribuite

si ha che A_0 e A_1 sono normalmente distribuiti con:

$$A_0 \cong N \left(\alpha_0, \sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \right)$$

$$A_1 \cong N \left(\alpha_1, \sigma^2 \cdot \left[\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \right)$$

La conoscenza della distribuzione dei due stimatori potrebbe essere utilizzata per determinare degli intervalli di confidenza dei parametri ma σ^2 non è noto. Usiamo quindi una stima data dalla realizzazione statistica per sostituirlo:

$$S_{RES}^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

dove \hat{Y}_i sono i valori teorici che dovrebbero essere assunti dal carattere Y se si trovassero sulla retta di regressione stimata ovvero:

$$\hat{Y}_i = a_0 + a_1 \cdot X_i, \quad \forall i = 1, 2, \dots, n$$

Questa statistica non è altro che una varianza campionaria che descrive lo scostamento tra le osservazioni y_i del carattere Y e le loro stime $\hat{y}_i = a_0 + a_1 \cdot x_i$ ottenute dalla regressione. Questa statistica è quindi la **varianza campionaria dei residui**:

$$E_i = Y_i - (\alpha_0 + \alpha_1 \cdot X_i)$$

$\frac{1}{n-2}$ permette di rendere corretto lo stimatore S_{RES}^2 di σ^2 , ovvero:

$$E[S_{RES}^2] = \sigma^2$$

e sostituendo nelle distribuzioni degli stimatori si ha:

$$T_0 = \frac{A_0 - \alpha_0}{\sqrt{S_{RES}^2 \cdot \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}}$$

$$T_1 = \frac{A_1 - \alpha_1}{\sqrt{S_{RES}^2 \cdot \left[\frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}}$$

e sempre per le consuete ipotesi si ha che: *le variabili T_0 e T_1 risultano essere distribuite come delle t di Student con $n - 2$ gradi di libertà.*

Possiamo avere anche gli intervalli di confidenza per α_0 e α_1 :

$$a_0 - t_{1-\frac{\alpha}{2}} \cdot \sqrt{s_{RES}^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{s_x^2} \right]}, a_0 + t_{1-\frac{\alpha}{2}} \cdot \sqrt{s_{RES}^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{s_x^2} \right]}$$

$$\left[a_1 - t_{1-\frac{\alpha}{2}} \cdot \sqrt{s_{RES}^2 \cdot \frac{1}{s_x^2}}, a_1 + t_{1-\frac{\alpha}{2}} \cdot \sqrt{s_{RES}^2 \cdot \frac{1}{s_x^2}} \right]$$

con:

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

mentre $t_{1-\frac{\alpha}{2}}$ è il quantile di ordine $1 - \frac{\alpha}{2}$ della t di student con $n - 2$ gradi di libertà.

È nota la distribuzione di S_{RES}^2 infatti il rapporto $\frac{(n-2) \cdot S_{RES}^2}{\sigma^2}$ è distribuito come una Chi-Quadro con $n - 2$ gradi di libertà, sempre in base alle solite quattro ipotesi.

Un intervallo di confidenza per σ^2 con un livello di fiducia α è:

$$\left[\frac{(n-2) \cdot s_{RES}^2}{q_{1-\frac{\alpha}{2}}}, \frac{(n-2) \cdot s_{RES}^2}{q_{\frac{\alpha}{2}}} \right]$$

con $q_{1-\frac{\alpha}{2}}$ e $q_{\frac{\alpha}{2}}$ sono i due quantili della Chi-Quadro con $n - 2$ gradi di libertà.

9.2 Attendibilità del Modello Lineare

Abbiamo già detto che dopo aver stimato i parametri di un modello lineare occorre verificare l'attendibilità del modello stesso, ovvero verificare ipotesi che le relazioni tra i due caratteri siano esprimibili tramite equazione:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_i + E_i$$

Un primo criterio adottabile per procedere in queste verifiche consiste nell'effettuare uno dei test di ipotesi descritti nel capitolo precedente. Se con l'uso di questi test si dovesse rifiutare l'ipotesi di incorrelazione allora ha senso provare ad effettuare test più approfonditi sulla validità del modello lineare, altrimenti si dovrebbe passare a considerare qualche altro modello.

Se si hanno dati accurati uso il test basato sulla statistica:

$$\hat{T}_n = R_n \cdot \sqrt{\frac{n-2}{1-R_n^2}}$$

se non si hanno dati accurati si usa:

$$\tilde{T}_n = R_S \cdot \sqrt{\frac{n-2}{1-R_S^2}}$$

Un test più specifico per la regressione lineare è basato su un approccio di analisi della varianza del carattere dipendente Y :

dato un campione $(X_1, Y_1); (X_2, Y_2); \dots (X_n, Y_n)$ estratto da (X, Y) si considerino i soliti stimatori e quindi:

$$\hat{Y}_i = A_0 + A_1 \cdot X_i$$

con le seguenti statistiche:

$$D_{TOT} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad \text{Devianza Totale}$$

$$D_{SP} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad \text{Devianza Spiegata}$$

$$D_{RES} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad \text{Devianza dei Residui}$$

Queste tre quantità sono gli indici degli scostamenti tra le osservazioni campionarie del carattere Y , la media di queste osservazioni e le loro stime tramite la regressione. In particolare:

- la **devianza totale** si riferisce agli scostamenti tra le osservazioni Y_i e la loro media campionaria \bar{Y}
- la **devianza spiegata** si riferisce agli scostamenti tra le stime \hat{Y}_i e la \bar{Y}
- la **devianza residua** si riferisce agli scostamenti tra le osservazioni Y_i e le loro stime \hat{Y}_i

inoltre si ha che:

$$D_{TOT} = D_{SP} + D_{RES}$$

Possiamo allora affermare che le variazioni tra le osservazioni Y_i e la loro media \bar{Y} sono da attribuire in parte alle variazioni spiegate dalla retta di regressione, ed in parte al fatto che le osservazioni non si trovano esattamente su tale retta (cioè ai residui). **In altri termini possiamo dire che se non fosse per i residui allora la regressione spiegherebbe lo scostamento totale tra le osservazioni e la loro media.**

È logico pensare che il modello è adatto a descrivere variazioni del carattere Y se queste sono causate principalmente dalla regressione, ovvero se il rapporto:

$$R^2 = \frac{D_{SP}}{D_{TOT}} = 1 - \frac{D_{RES}}{D_{TOT}} \sim 1$$

In effetti si può verificare che R^2 è il quadrato del coefficiente di correlazione lineare r_{XY} introdotto nei capitoli precedenti cioè vale $\sqrt{R^2} = |r_{XY}|$.

Mentre però il coefficiente di correlazione lineare ha senso solo se si considera una regressione lineare semplice, il coefficiente R^2 presenta il vantaggio di essere utilizzabile come misura di adattamento qualunque sia la forma della curva di regressione e qualunque sia il numero di variabili esplicative considerate. Per tale ragione è chiamato **coefficiente di correlazione generalizzato**.

Inoltre si ha che il modello lineare è adatto a descrivere le variazioni del carattere Y se $D_{RES} \ll D_{SP}$ quindi $D_{SP}D_{RES}$ è molto grande. Un test sull'attendibilità del modello lineare molto utilizzato si basa proprio su questo rapporto e specificamente sul fatto che la statistica:

$$\tilde{F} = (n - 2) \cdot \frac{D_{SP}}{D_{RES}} = \frac{D_{SP}}{S_{RES}^2}$$

è distribuita come una F con $(1, n - 2)$ gradi di libertà quando siano soddisfatte le solite quattro ipotesi e quando valga $\alpha_1 = 0$ (la dipendenza tra i valori assunti dal carattere X e quelli assunti dal carattere Y).

È possibile pertanto effettuare un test per ipotesi nulla, per una regressione non significativa, $H_0 : \alpha_1 = 0$ rifiutandola quando la realizzazione \tilde{f} della statistica $\tilde{F} = (n - 2) \cdot \frac{D_{SP}}{D_{RES}} = \frac{D_{SP}}{S_{RES}^2}$ assume valori grandi, ovvero nel caso di un test di ampiezza α quando $\tilde{f} > F_{1-\alpha}$ dove $F_{1-\alpha}$ è il quantile di ordine $1 - \alpha$ della $F(1, n - 2)$.

Un secondo test che si può adottare per verificare l'attendibilità del modello lineare si basa sul fatto che assumere l'esistenza di una relazione lineare tra i due caratteri X ed Y corrisponde ad assumere $\alpha_1 \neq 0$, in quanto se fosse $\alpha_1 = 0$ non si potrebbe avere $Y_i = \alpha_0 + \alpha_1 \cdot X_i + E_i$, allora le variazioni delle Y non dipenderebbero da quelle delle X . Si può quindi effettuare un test avente come ipotesi nulla la:

$$H_0 : \alpha_1 = 0$$

e come ipotesi alternativa $H_1 : \alpha_1 \neq 0$. In tal caso in base alle solite quattro ipotesi sui residui la variabile T :

$$T_1 = \frac{A_1 - \alpha_1}{\sqrt{S_{RES}^2 \cdot \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right]}}$$

è distribuita come una t di Student con $(n - 2)$ gradi di libertà.

Rifiuteremo ipotesi nulla quando la sua realizzazione \tilde{t} assume valori troppo distanti dallo zero per poter pensare che essa sia distribuita secondo una t di Student con $(n - 2)$ gradi di libertà.

9.3 Analisi dei residui

Nei test si fa uso del livello di significatività, il quale implicitamente definisce la regione critica: se il valore della statistica calcolato dai dati campionari cade nella regione critica, si rifiuta l'ipotesi. Non si sa però, se si modificasse la significatività, se l'ipotesi continuerebbe a esser rifiutata o no, salvo ricalcolare la regione: non si sa cioè se il valore ottenuto sia significativamente o solo marginalmente interno alla regione critica predefinita. Un modo alternativo e più informativo è quello di usare il **p-value**.

Sia t' il valore assunto dalla statistica sui dati campionari e si considerino le regioni critiche a seconda dell'ipotesi alternativa:

- $H_1 : \vartheta \neq \vartheta_0$ per il test bidirezionale
- $H_1 : \vartheta < \vartheta_0$ per il test unidirezionale con coda a sinistria
- $H_1 : \vartheta > \vartheta_0$ per il test unidirezionale con coda a destra

si ha, se H_1 è l'ipotesi alternativa:

$$C' = \left(-\infty, -t_{1-\frac{\alpha}{2}}^*\right) \cup \left(t_{1-\frac{\alpha}{2}}^*, +\infty\right)$$

$$C'' = \left(-\infty, -t_{1-\alpha}^*\right)$$

$$C''' = \left(t_{1-\alpha}^*, +\infty\right)$$

Il p-value è definito come la probabilità che la statistica valida sotto l'ipotesi nulla assuma valore nella regione critica così definita. Il p-value misura quindi la **verosimiglianza** (o linverosimiglianza) del valore campionario ottenuto, dalla distribuzione della statistica sotto l'ipotesi nulla. Allora è chiaro che, quanto più piccolo è il p-value, tanto più inverosimile è che il risultato campionario della statistica provenga dalla distribuzione della statistica valida sotto l'ipotesi nulla: quindi con tanta maggior sicurezza si rifiuta l'ipotesi nulla in favore dell'ipotesi alternativa. Invece, quanto più grande è il p-value, tanto più verosimile è che si verifichi quel tal risultato campionario della statistica, sotto l'ipotesi nulla: quindi con tanta maggior sicurezza si esclude che l'evidenza ottenuta contrasti con l'ipotesi nulla.

Per l'**ipotesi di incorrelazione** tra i residui il metodo più utilizzato è il **test di Durbin-Watson**. In esso viene considerata la statistica, con e_i i-esimo residuo:

$$D = \frac{\sum_{i=2}^n (E_i - E_{i-1})^2}{\sum_{i=1}^n E_i^2}$$

otto l'ipotesi di incorrelazione tale statistica ha una particolare distribuzione, le cui tavole si trovano in numerosi testi introduttivi all'analisi di regressione.

È possibile quindi determinare delle regioni critiche per D e rifiutare l'ipotesi di incorrelazione se la realizzazione di D dovesse cadere in tali regioni. Si ha quindi che se $p - \text{value}(Prob < D)$ è grande non si rifiuta H_0 in quanto i residui sono incorrelati, altrimenti lo si rifiuta.

Si hanno varie tecniche per controllare l'ipotesi di normalità, la migliore è ricorrere ai test per la bontà dell'adattamento, ovvero:

- test di Kolmogorov-Smirnov
- test del Chi-Quadro
- test di Shapiro Wilk

In tutti e tre si controlla l'ipotesi che i residui campionari e_1, \dots, e_n ottenuti dal campione $\{(X_1, Y_1); (X_2, Y_2); \dots (X_n, Y_n)\}$.

siano distribuiti normalmente con media nulla e varianza s_{RES}^2 .

- **test di Kolmogorov-Smirnov-Lillefors**, che è basato sulla differenza tra ripartizione empirica e ripartizione della normale:

$$K_n = \sup_x |F_e(x) - F_n(x)| \sqrt{n}$$

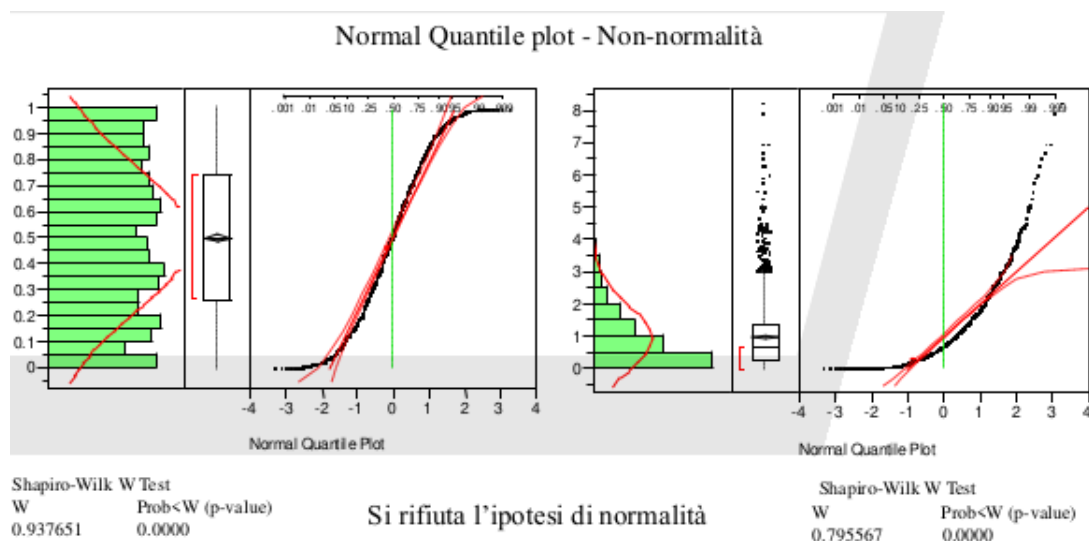
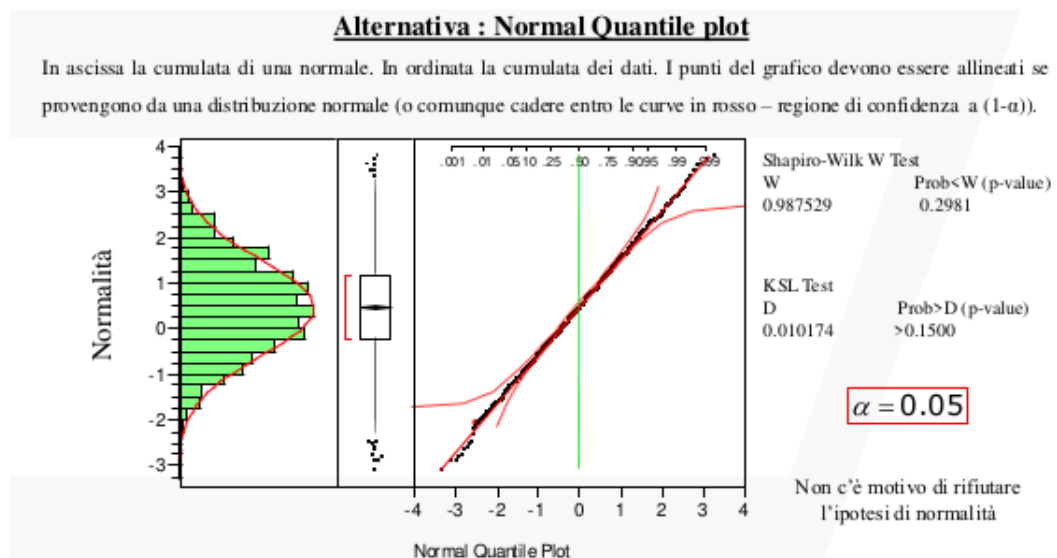
In questo caso si rifiuta l'ipotesi di normalità se K_n è grande. Questo test richiede campioni di elevata numerosità

- **test di Shapiro Wilk** che serve per campioni di bassa numerosità. La statistica W misura l'aretilineità del plot:

$$W = \frac{[\sum_1^n w_i e_i]^2}{\sum_1^n (e_i - \bar{e})^2}$$

con i w_i sono le quantità tabulate e i e_i sono valori campionari ordinati. Si rifiuta se W è piccolo.

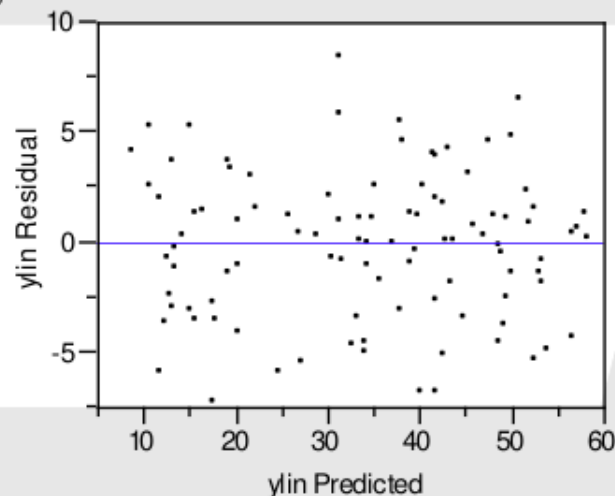
Ovvero:



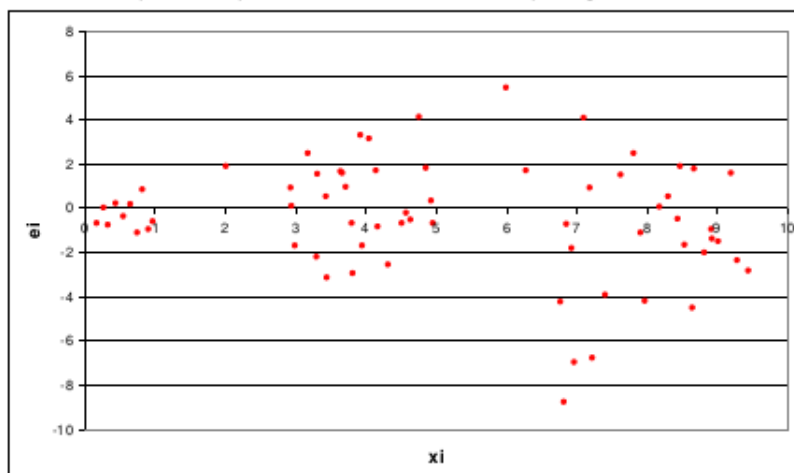
Le maggiori difficoltà si incontrano nella verifica di omoschedasticità, ovvero di identica varianza per ogni residuo. Test veramente significativi per tale ipotesi possono essere realizzati se il numero di valori assumibili dal carattere X è limitato e se per ognuno di tali valori esistono numerose osservazioni, ma è una situazione rara e si deve quindi ricorrere a metodi grafici.

Tra questi, comunque non sempre disprezzabili, segnaliamo quello basato sulla rappresentazione dei residui in funzione dei valori assunti dalla variabile X o Y . Per ogni elemento del campione si individui il corrispondente punto su un piano cartesiano avente i valori del carattere X (o Y) su un asse ed i valori del residuo sull'altro

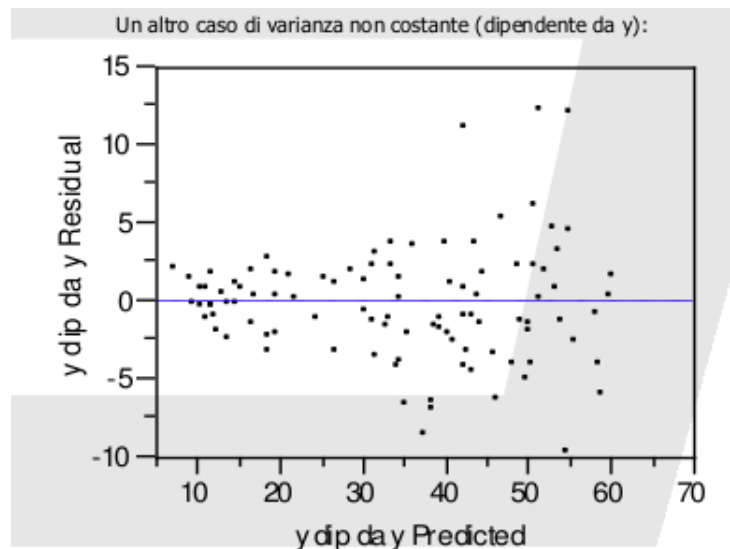
ACCETTIAMO L'IPOTESI DI OMOSCHEDASTICITÀ se i punti così individuati si presentano a forma di una nuvola, come riportato sotto



Rifiutiamo invece l'ipotesi se i punti assumono un andamento più regolare come mostrato sotto.



In questo caso è lecito pensare che la varianza del residuo aumenti all'aumentare del valore assunto dal carattere X .

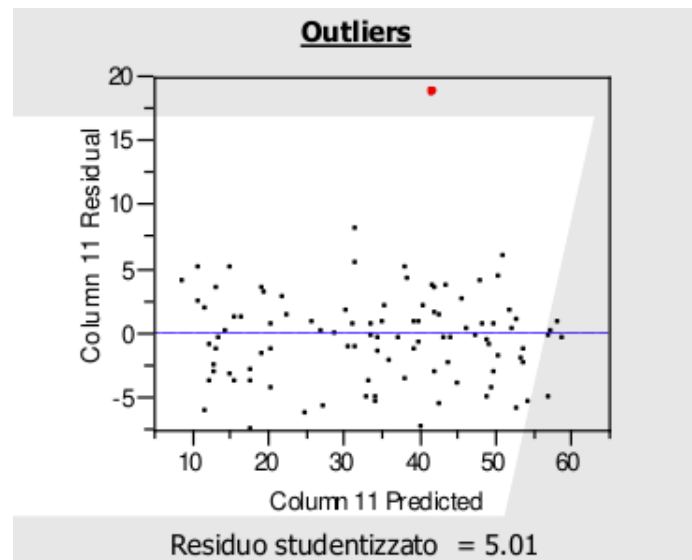


9.3.1 Outliers

Uno dei problemi che si presentano nella pratica è la presenza di osservazioni anomale (dovute ad esempio a errori di misura). Questi valori anomali, detti outliers, possono provocare un peggioramento nella qualità della regressione, in quanto i valori dei parametri stimati vengono distorti per tenere conto anche di questi valori errati. Per riconoscere i possibili outliers ed eliminarli dal calcolo un modo è quello di calcolare i **Residui Studentizzati**:

$$r_i = \frac{e_i}{\sqrt{MSE \left(1 - \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}}$$

con e_i Si dimostra che essi sono distribuiti secondo una t di student e se il residuo stu



Non tutti gli outliers hanno però un ugual effetto di modifica dei parametri della regressione: per riconoscere quelli che la influenzano effettivamente, si usa una misura detta **D-Cook**, che indica il cambio di pendenza se si omette un'osservazione; un sospetto outlier è influente se $D > 1$.

Capitolo 10

Regressione Lineare Multipla

Torniamo quindi a considerare una popolazione $(m + 1)$ dimensionale.

$$(X_1, X_2, \dots, X_m, Y)$$

ed è un'equazione di **equazione di regressione del tipo**:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{1,i} + \dots + \alpha_m \cdot X_{m,i} + E_i, \quad i = 1, \dots, N = \text{numero totale degli individui della popolazione}$$

Presupponiamo le solite tre ipotesi e supponiamo pertanto di poter affermare che tali ipotesi sono soddisfatte e supponiamo di poter disporre di un campione:

$$(x_{1,1}, X_{2,1}, \dots, X_{m,1}, Y_1) \dots, (X_{1,n}, X_{2,n}, \dots, X_{m,n}, Y_n)$$

di numerosità n maggiore di m estratto dalla popolazione $(X_1, X_2, \dots, X_m, Y)$ con $X_{k,i}$ denota il carattere k -mo dell'individui i -mo con $k = 1, \dots, m$ e $i = 1, \dots, n$.

Si può pervenire alla determinazione delle stime delle costanti del modello ragionando esattamente come nel caso semplice, ovvero rendendo minima la somma dei quadrati o dei residui:

$$\sum_{i=1}^n E_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (\alpha_0 + \alpha_1 \cdot X_{1,i} + \dots + \alpha_m \cdot X_{m,i}))^2$$

Si ottengono così le stime delle costanti per la cui descrizione conviene introdurre una notazione matriciale. A tal fine sia:

$$\mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} & \dots & X_{m,1} \\ 1 & X_{1,2} & X_{2,2} & \dots & X_{m,2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{1,n} & X_{2,n} & \dots & X_{m,n} \end{pmatrix}$$

la matrice contenente le m variabili esplicative degli n individui del campione (aggiunta della prima colonna contenente solo unità).

Siano poi $Y = (Y_1, Y_2, \dots, Y_n)^T$ il vettore con i caratteri Y degli individui del campione e $A = (A_0, A_1, A_2, \dots, A_m)^T$ il vettore colonna contenente gli stimatori di $\alpha_0, \alpha_1, \dots, \alpha_m$. Denotiamo infine con M^T la trasposta di una matrice o un vettore M e con M^{-1} la matrice inversa di una matrice M quadrata e non singolare. Il modello regressivo può allora esser riscritto in forma vettoriale:

$$Y = X \cdot A + E$$

Si parte dalle solite tre ipotesi. Gli stimatori corretti di varianza minima di $\alpha_0, \alpha_1, \dots, \alpha_m$ sono dati da $A = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$.

Si dimostra poi che uno stimatore corretto della varianza dei residui è:

$$S_{RES}^2 = \frac{1}{n - m - 1} \cdot \sum_{i=1}^n E_i^2$$

Quindi se $(x_{1,1}, x_{2,1}, \dots, x_{m,1}, y_1), \dots, (x_{1,n}, x_{2,n}, \dots, x_{m,n}, y_n)$ è una realizzazione del campione $(X_{1,1}, X_{2,1}, \dots, X_{m,1}, Y_1), \dots, (X_{1,n}, X_{2,n}, \dots, X_{m,n}, Y_n)$ allora una stima del vettore dei coefficienti $(\alpha_0, \alpha_1, \dots, \alpha_m)^T$ è data dal vettore:

$$a = (a_0, a_1, \dots, a_m)^T$$

ottenuto come.

$$a = (x^T \cdot x)^{-1} \cdot x^T \cdot y$$

Una stima della varianza è invece data da:

$$s_{RES}^2 = \frac{1}{n - m - 1} \cdot \sum_{i=1}^n e_i^2$$

dove gli e_i sono le realizzazioni dei residui E_i , ovvero:

$$e_i = y_i - \hat{y}_i = y_i - (a_0 + a_1 \cdot x_{1,i} + \dots + a_m \cdot x_{m,i})$$

Nel caso in cui risulti valida anche l'ipotesi di normalità sarà possibile computare anche gli intervalli di confidenza per le stime dei parametri in modo del tutto analogo a quanto fatto nel caso della regressione lineare semplice.

Si denoti con $C_{k,k}$ la k -ma componente della diagonale principale della matrice $(X^T \cdot X)^{-1}$. Gli intervalli di confidenza per le costanti si ottengono tenendo conto del fatto che quando sono soddisfatte le ipotesi sui residui allora le variabili:

$$T_k = \frac{A_k - \alpha_k}{\sqrt{S_{RES}^2 \cdot C_{k,k}}}$$

sono distribuite secondo delle t di Student con $(n - m - 1)$ gradi di libertà.

Abbiamo già accennato al fatto che passando a considerare la regressione lineare multipla alcune quantità utilizzabili per valutare la bontà dell'adattamento di un modello di regressione lineare semplice perdono di significato. Tra di esse in particolare ricordiamo il coefficiente di correlazione generalizzato definito come:

$$R^2 = 1 - \frac{D_{RES}}{D_{TOT}}$$

il modello di regressione:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{1,i} + \dots + \alpha_m \cdot X_{m,i} + E_i$$

e cui costanti vengono stimate come mostrato in precedenza dovrà essere considerato tanto più attendibile quanto più $R^2 \rightarrow 1$ ed essere considerato inattendibile quando esso è prossimo allo 0. Un vero e proprio test si può effettuare quando siano verificate tutte quattro le ipotesi sui residui. In questo caso si può far ricorso alla statistica:

$$\tilde{F} = \frac{D_{SP}}{S_{RES}^2} = (n - m - 1) \cdot \frac{D_{SP}}{D_{RES}}$$

che risulta essere distribuita secondo una F con $(m, n - m - 1)$ gradi di libertà quando non esiste dipendenza tra il carattere risposta Y ed i caratteri esplicativi X_k ovvero quando $\alpha_1 = \dots = \alpha_m = 0$ e rifiuteremo l'ipotesi nulla $H_0 : \alpha_1 = \dots = \alpha_m = 0$ quando la realizzazione \tilde{f} della statistica \tilde{F} In particolare rifiuteremo l'ipotesi nulla (per un test di ampiezza α) quando $\tilde{f} > F_{1-\alpha}$ dove $F_{1-\alpha}$ rappresenta il quantile di ordine $1 - \alpha$ della $F(m, n - m - 1)$.

Per quanto riguarda le ipotesi sui residui continuano a valere le considerazioni espresse relativamente alla regressione lineare semplice.

Non tutte le variabili X_k che compaiono nel modello di regressione lineare multipla:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_{1,i} + \dots + \alpha_m \cdot X_{m,i} + E_i$$

possono essere realmente esplicative, nel senso che non necessariamente tutte contribuiscono a spiegare i valori assunti dal carattere dipendente Y . Potrebbe infatti essere $\alpha_k = 0$ per qualche $k = 1, \dots, m$. Può essere utile venire a conoscenza di tale fatto onde evitare in futuro di rilevare i valori assunti da X_k per determinare delle stime di Y .

Occorre allora poter effettuare dei test specifici per accertare eventuale indipendenza tra una singola variabile esplicativa X_k e la variabile dipendente Y . Per questo si ricorre al fatto già citato che le variabili:

$$k = \frac{A_k - \alpha_k}{\sqrt{S_{RES}^2 \cdot C_{k,k}}}$$

hanno distribuzione se si hanno le solite quattro ipotesi.

In particolare se vale anche l'ipotesi nulla $H_0 : \alpha_k = 0$ avremo la statistica:

$$\tilde{T}_k = \frac{A_k}{\sqrt{S_{RES}^2 \cdot C_{k,k}}}$$

sarà distribuita come una t di Student con $(n - m - 1)$ gradi di libertà.

Si rifiuta l'ipotesi nulla qualora la realizzazione \tilde{t}_k di \tilde{T}_k assuma valori troppo distanti dallo zero per poter pensare che essa abbia distribuzione t di Student con $(n - m - 1)$ gradi di libertà.

Per esempio, per un test di ampiezza α , rifiutiamo H_0 quando $|\tilde{t}_k| > t_{1-\frac{\alpha}{2}}$ dove $t_{1-\frac{\alpha}{2}}$ rappresenta il quantile di ordine $1 - \frac{\alpha}{2}$ per una t di Student con $(n - m - 1)$ gradi di libertà.

Si ha il problema della **multicollinearità**. Notiamo che per determinare il vettore delle stime $a = (a_0, a_1, \dots, a_m)^T$ è necessario calcolare la matrice inversa di $(x^T \cdot x)$ e per questo occorre che tale prodotto matriciale sia non singolare, ovvero con determinante non nullo. Ciò avviene se nessuna delle variabili X_k è combinazione lineare delle altre. **Nella scelta delle variabili esplicative occorre quindi essere certi che non si verifichi questo fenomeno.**

In realtà si presentano dei problemi non solo se una delle X_k è combinazione lineare delle altre ma anche se ci si trova di fronte a forti dipendenze lineari tra le variabili esplicative. In questi casi occorre quindi eliminare dal modello una o più di esse. Non è studio del corso valutare se esiste un problema di multicollinearità. Una cosa che comunque si può fare senza troppe difficoltà è considerare la matrice di correlazione delle variabili esplicative (ovvero la matrice le cui componenti sono i coefficienti di correlazione lineare di tutte le coppie di variabili esplicative). Una forte correlazione tra due variabili esplicative è condizione sufficiente per affermare che esiste un fenomeno di multicollinearità. Conviene in questo caso eliminare una delle due variabili dal modello