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/dlcgold/Appunti`.
Grazie mille e buono studio!

Capitolo 2

Breve Introduzione

La statistica è una discliplina, basata sulla matematica, con il fine 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.

Ormai i dati sono pervasivi e un loro studio è diventato necessario ed inoltre si parla spesso di target marketing, con una selezione dei possibili clienti infatti è usata in maniera massiccia nel mondo dello shopping online. 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 codesto corso vengono effettuati 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 3

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, ossia si effettua una descrizione delle informazioni implicite ai dati.

Ovviamente 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 intende effettuare l'analisi statistica e 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.

Si vuole affermare qualcosa riguardo i **caratteri** della popolazione, ossia gli elementi su cui effettuo l'analisi statistica, che possono essere:

- **caratteri qualitativi**, indicanti qualità (colori, stili, materiali etc...) e non dati numerici in cui solitamente non è definita una *relazione d'ordine*
- **caratteri quantitativi**, maggiormente studiati dal corso, in cui vengono definite *relazioni d'ordine*:
  - **discreti**, come i lanci di un dado, rappresentanti valori in $\mathbb{Z}$
  - **continui**, che assumono valori reali, come la temperatura, in $\mathbb{R}$

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

$$E = \{x_1, x_2, \ldots, x_n\}$$

con la numerosità, il numero di elementi considerati, pari a n. In caso il carattere è quantitativo discreto è comodo raggruppare i dati considerando l'insieme di tutti i valori assumibili, **modalità del carattere** ed associare ad ognuno di esso il numero di volte che esso compare in E.
Si ha quindi N il numero di totalità del carattere e si definisce l'insieme di modalita:

$$S = \{s_1, ..., s_N\}$$

su cui si definiscono i seguenti valori statistici:

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

frequenza cumulata assoluta F_j somma delle frequenze assolute di tutte le modalità:

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

frequenza relativa p_j rapporto tra la frequenza assoluta e il numero di elementi

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

frequenza cumulativa relativa P_j somma delle frequenze relativa di tutte le modalità:

$$P_j = \sum k : s_k \leq s_j p_k$$

Si definisce **distribuzione di frequenza** una funzione $F : S \to \mathbb{N}$ 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 S, l'insieme delle modalità, in alcune classi (sottoinsiemi di S) che formano una partizione e 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:

- confine superiore e inferiore (valori estremi della classe)
- ampiezza (differenza tra confine superiore ed inferiore)
- valore centrale (media tra i due confini)

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.

3.1 Indici di tendenza Generale

Fino ad ora abbiamo visto come rappresentare i dati, sia discreti che continui, 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

$$\overline{x} = \frac{1}{n}\sum x_i = \frac{x_1 + \cdots + x_n}{n}$$

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

$$\overline{x} = \frac{1}{n}\sum s_j f_j = \sum s_j p_j$$

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 somma degli elementi di posto $\frac{n}{2}$ e $\frac{n}{2} + 1$.

moda $\widetilde{x}$ valore o classe 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 in cui vi sia un unica moda
- **distribuzione multi-modale** nel caso in cui vi siano più mode

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

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

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:

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

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

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

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

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

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

Nel calcolo della varianza si utilizza il quadrato per la differenza tra l'elemento e la sua media in quanto per come è definita la media si ha $\sum(x_i - \overline{x}) = 0$ e per evitare ciò si eleva la differenza tra un elemento e la sua media al quadrato.

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

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

3.2 Il caso bidimensionale

Fino ad ora noi abbiamo considerato il caso unidimensionale in cui consideriamo solo un carattere del campione ma molte analisi richiedono di analizzare due o più caratteri del campione contemporaneamente per riconoscere leggi ed analogie tra i diversi caratteri.
Consideremo solo due caratteri contemporanei, sia perchè un analisi con più caratteri si comporta uguale e sia per non aggravare troppo la rappresentazione dei dati negli esempi e assumiamo 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

$$E = \{(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)\}$$

mentre l'insieme delle coppie di valori assumibili si rappresentati con l'insieme

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

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

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

frequenza relativa

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

rapporto tra la frequenza e il numero di elementi

frequenza cumulata assoluta

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

frequenza cumulata relativa

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

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 esistono anche altri tipi di distribuzioni, infatti noi vediamo le **distribuzioni marginali**: distribuzioni dei singoli caratteri presi indipendemente degli 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 agli indici definiti fino ad ora, esiste un indice che fornisce un grado di interdipendenza tra i due caratteri, indice 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.

Considero due serie $\{x_i\}, \{y_i\},\ i = 1, ..., n$ ponendo a confronto le variazione delle coppie di dati rispetto ai corrispondenti valori medi, considerando le coppie di scarti:

$$x_i - \overline{x}$$

$$y_i - \overline{y}$$

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, delle due serie di dati :

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

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:

$$c_{xy} = \frac{1}{n}\sum x_i y_i - \overline{xy}$$

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:

$$cxy = \sum_1^N \sum_1^M (s_j - \overline{x})(u_k - \overline{y})p_{jk}$$

$$cxy = \sum_1^N \sum_1^M s_j u_k p_{jk} - \overline{xy}$$

Date due serie di dati si ha che sono:

- **statisticamente incorrelate** se la loro covarianza è nulla
- **statisticamente indipendenti** se vale:

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

con:

$$p_{jk} = \frac{f_{jk}}{n}$$
$$p_j = \frac{f_j}{n}$$
$$p_k = \frac{f_k}{n}$$

inoltre due serie di dati statisticamente indipendenti sono incorrelate mentre non è necessariamente vero il contrario, infatti:

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

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

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

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

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

che ovviamente può crescere in m dimensioni.

3.3 Regressione Lineare

La regressione lineare

Prendo un campione di coppie di dati:

$$E = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

In molti casi ci si pone la questione se tra tali caratteri x ed y esista un legame di tipo funzionale o una relazione di tipo funzionale che ne descriva in modo soddisfacente corretto il legame realmente esistente.
Si parla di un'**analisi di regressione**, in cui si pensa 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 che si otterrebbero per il carattere y se la relazione che lega il carattere y ad x fosse proprio quella descritta da f, quindi cerco 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:

$$y = mx + q$$

con q intercetta e m coefficiente angolare, 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:

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

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

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 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

3.4 Regressione non Lineare

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:

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

con:

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

$$\widetilde{x} = x$$

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

$$\beta = b$$

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 $\{\widetilde{x}_i, \widetilde{y}_i\}$, con:

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

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

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

Ecco alcune funzioni riconducibili a lineari:

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

$$y = ax^b$$

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