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

Ormai i dati sono pervasivi e un loro studio è diventato necessario. Inoltre si parla spesso di target marketing, con una selezione dei possibili clienti. Inoltre la statistica è usata in maniera massiccia nel mondo dello shopping online. Si ha l'*A-B testing*, per decidere tra due scelte la migliore. Per scegliere si analizzano i dati presi da campioni di popolazione. Si ha il *tasso di conversione*. Un altro ambito importante è la robotica e la domotica. Ovviamente la statistica è importante anche nel mondo dei videogames. Anche genetica e biologia fanno uso abbondante dei dati statistici.
Si hanno i seguenti 8 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 trovare:

- simmetrie
- periodicità
- leggi varie

Si descrivono quindi le informazioni implicite ai dati. Solitamente la serie di dati di cui si dispone è costituita da un numero limitato di **osservazioni** che devono essere rappresentative di un'ampia **popolazione**, che è quindi l'insieme a cui si riferisce l'indagine statistica. Un **campione** è un sottoinsieme selezionato della popolazione, che viene analizzato per dire qualcosa sulla popolazione da cui è stato prelevato. Non esiste un solo campione di una popolazione ma si hanno molti modi diversi di scegliere un campione. Si vuole affermare qualcosa riguardo i **caratteri/caratteristiche** della popolazione. Si hanno:

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

Supponiamo di considerare n elementi della popolazione e di rilevare, per ognuno di essi, il dato relativo al carattere quantitativo da esaminare. Si ha un insieme di dati:

$$E = \{x_1, ..., x_n\}$$

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

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

Si chiama **frequenza assoluta** s_j. f_j è il numero di volte che si presenta un elemento di un campione ovvero è il numero di elementi di E che hanno lo stesso valore s_j.
Si ha la **distribuzione di frequenza assoluta dei dati** unzione che associa ad ogni modalità la corrispondente frequenza assoluta:

$$F : S \to N$$

Si definisce **frequenza cumulata assoluta** per la modalità s_j la somma delle frequenze assolute di tutte le modalità:

$$s : k \in S : s_k \leq s_j$$

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

frequenza relativa:

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

frequenza cumulativa relativa:

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

Si dicono distribuzione di frequenza cumulata assoluta, relativa e cumulata relativa dei dati osservati, le funzioni F, p, P che associano ad ogni modalità frequenze s_j le relative frequenze F_j, p_j, P_j, con n numero di osservazioni Quando il carattere da studiare è continuo (o discreto con un gran numero di valori) è conveniente ricondursi a raggruppamenti come quelli appena trattati. Si suddivide S, l'insieme delle modalità, in alcune classi (sottoinsiemi di S) che formano una partizione (ogni famiglia di classi tra loro disgiunte la cui

unione dà tutto S). 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. Ad ogni classe si asscoiano delle grandezze:

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

Nel caso in cui il carattere esaminato sia continuo occorre specificare quando le classi sono chiuse, a destra o a sinistra, ovvero 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.

3.0.1 Indici di tendenza Generale

Si cerca un modo di rappresentare una qualche serie di dati con un solo valore. Si usano gli **indici di tendenza generale** che sono quantità in grado di sintetizzare con un solo valore numerico i valori assunti dai dati. Uno di questi è la **media aritmetica** $\overline{x}$. Ho un campione $x_1, ..., x_n$. Si ha la media:

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

Nel caso in cui i dati siano di tipo quantitativo discreto allora avremo:

$$\overline{x} = \frac{1}{n} \sum s_j f_j = \frac{s_1 f_j + \cdots + s_n f_n}{n}$$

$$\overline{x} = \sum s_j p_j = s_1 p_1 + \cdots + s_N p_N$$

si ha il momento k-esimo rispetto ad y:

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

la media è anche il **momento primo rispetto all'origine**, con $y = 0$ e ovviamente $k = 1$.

Un secondo indice di tendenza è rappresentato dalla **mediana** definita come *quel numero reale che precede tanti elementi della serie di dati quanti ne segue.*

Se ordino $x_1, ..., x_n$ in ordine crescente ottengo la serie:

$$x_{(1)}, ..., x_{(n)}$$

e la **mediana** $\hat{x}$ è:

- $\frac{n+1}{2}$ se n è dispari
- dalla media aritmetica tra l'elemento di posto $\frac{n}{2}$ e quello di posto $\frac{n}{2}+1$ se n è pari

la **moda** $\widetilde{x}$ è quel valore o classe cui corrisponde la massima frequenza assoluta. La moda viene spesso utilizzata nel caso di dati qualitativi ovvero quando risulti impossibile definire media e mediana. Si ha che non è garantita l'unicità della moda infatti parleremo di:

- **distribuzione uni-modale** nel caso in cui vi sia un unica moda
- **distribuzione multi-modale** nel caso in cui vi siano più mode

La moda è anche il valore con massima frequenza assoluta Gli indici di tendenza centrale non sono utili per fornire informazioni circa l'omogeneità dei dati. Questo aspetto viene fornito da altri indici. Uno di questi è la **varianza**:

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

ed è quindi il momento secondo rispetto alla media, quindi con $k = 2$:

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

viene introdotto il quadrato perché in caso contrario, sostituendo col valore della media:

$$\frac{1}{n}\sum(x_i - \overline{x}) = 0$$

La varianza è tanto più grande quanto più i singoli dati si scostano dalla media, vale a dire tanto più i dati risultano tra loro disomogenei. E pertanto la misura varianza consente di rappresentare il grado di disomogeneità della serie di dati. Ovviamente si ha anche:

$$s^2 = \frac{1}{n}\sum x_i^2 - \overline{x}^2$$

Con dati quantitativi discreti, di cui si sa la distribuzione frequenza:

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

$$s^2 = \sum (s_j - \overline{x})^2 p_j$$

ovviamente si ha anche:

$$s^2 = \sum s_j^2 p_j - \overline{x}^2$$

Poiché la dimensione della varianza è il quadrato di quella dei dati, in molti casi si preferisce una diversa misura detta scarto quadratico medio:

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

La varianza mi dice quindi quanto disperse sono le osservazioni.

3.0.2 Il caso bidimensionale

In molti casi si è interessati a studiare fenomeni che coinvolgono due o più caratteri della popolazione tali da non potersi considerare separatamente. Ci limitiamo al caso di due caratteri contemporanei, quindi con l'insieme dei dati sarà un insieme di coppie:

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

Ipotizziamo inoltre che entrambe i caratteri siano di tipo quantitativo e discreto. Se fossero stati quantitativi continui subirebbero prima un raggruppamento a classi. Sia S l'insieme delle coppie di valori assumibili dalla coppia di caratteri analizzati:

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

Viene detta **frequenza assoluta** di (s_j, u_k) la quantità f_{jk} che è il numero di elementi di E aventi valori (s_j, u_k).
La **distribuzione di frequenza doppia** la funzione f che associa ad ogni coppia (s_j, u_k) la corrispondente frequenza f_{jk}.

Si hanno i vari tipi di frequenza:

- **frequenza cumulata assoluta:**

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

- **frequenza relativa:**

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

- **frequenza cumulata relativa:**

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

Con **distribuzione di frequenza doppia** si intende infine una qualsiasi delle funzioni f, F, p, P che associ ad una coppia (s_j, u_k) la corrispondente frequenza.
Si hanno anche altri tipi di distribuzione. Vediamo le **distribuzioni marginali**, che sono distribuzioni dei singoli caratteri presi indipendentemente dagli altri. Nel caso ci si riferisca al primo carattere, per ogni valore assumibile da esso, sia s_j è detta **frequenza assoluta marginale** la quantità f_x data dal numero di elementi di E il cui il primo carattere ha valore s_j. Quindi f_{xj} è il numero di elementi di E aventi valore $(s_j, \cdot)$. Analogamente si ha:

- **frequenza cumulata assoluta marginale:** F_{xj} che è la somma delle frequenze assolute marginali di tutti gli s_k $s_k \leq s_j$
- **frequenza relativa marginale:** p_{xj} che è il rapporto tra la frequenza assoluta marginale e la numerosità delle osservazioni
- **frequenza cumulata relativa marginale:** $FOxj$ che è la somma delle frequenze relative marginali di tutti gli s_k $s_k \leq s_j$

considero due serie $\{x_i\}, \{y_i\},\ i = 1, ..., n$. Pongo 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 (positivi o negativi che siano). 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 (non vi è dipendenza degli uni dagli altri) 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 (variabili x e y) la covarianza si può rappresentare attraverso una matrice 2×2:

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

E' dipendente dalla grandezza delle varianze. 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

3.0.3 Regressione Lineare

Prendo un campione di coppie di dati:

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

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 soddisfacentemente corretto il legame realmente esistente Si parla di un'**analisi di regressione**. Si pensa ad uno dei due caratteri come variabile indipendente e cerco 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 (una retta) 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 $[-1, 1]$ inoltre è nullo se le serie sono statisticamente incorrelate ed è in valore assoluto apri a 1 se le coppie sono tutte sulla retta $y = mx + q$. Quindi rappresenta il grado di allineamento delle coppie di dati

3.0.4 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. 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. Si ha quindi il **modello non lineare di regressione**.
Molte relazioni funzionali non lineari possono essere ricondotte a tali (lineari) con opportune trasformazioni delle variabili. Si ha 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:

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

$$y = ax^b$$

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