

# Glossario minimo

Giorgio Marrubini e Camillo Melzi

07 giugno 2020

Le voci seguenti sono in ordine alfabetico e riportano alcune definizioni di termini e concetti di base che ricorrono nelle dispense. Le definizioni sono tratte dai riferimenti citati in bibliografia.

## **Aleatorio, campione a., evento a., intervallo a., variabile a.**

Il termine aleatorio è sinonimo di casuale. Deriva da “alea” che in latino significa dado. Quindi aleatorio è aggettivo di un campione, un evento o altro la cui natura è legata al caso (v. anche la voce *caso*, *casuale*).

## **Analisi di correlazione**

Verifica del fatto che due variabili siano correlate tra loro. V. *Correlazione*.

## **Analisi di regressione**

Definizione del modello funzionale per cui una proprietà  $Y$  dipende da un fattore  $X$  e quindi che valga la relazione  $Y = f(X)$ . Nel caso di più fattori,  $X_1, X_2, \dots, X_n$ , si parla di regressione multipla e si verifica quindi la validità della relazione  $Y = f(X_1, X_2, \dots, X_n)$ .

**Autocorrelazione** v. anche *Correlazione*. L'autocorrelazione è il grado di dipendenza tra i valori che assumono le variabili in ascissa.

## **Campionario, media c., varianza c.**

Campionario è detto di proprietà relativa al campione.

## **Caso, casuale**

Il caso è un termine con cui si indica un evento che si verifica indipendentemente (almeno in apparenza) da una causa oggettiva, oppure un evento di cui non si conoscono le cause.

## **Confidenza, intervallo di c., livello di c.**

In statistica è sinonimo di “fiducia”. Indica la probabilità o grado di fiducia che la stima di un parametro sulla base di un campione (per es. la media) sia una buona approssimazione del parametro della popolazione. Più in dettaglio, fissato un valore di probabilità  $1-\alpha$  (con  $0 < \alpha < 1$ ), detto *livello di confidenza* (es.  $1-\alpha = 95\%$ ), l'*intervallo di confidenza* è l'intervallo  $[\theta_1, \theta_2]$  all'interno del quale si trova il valore del parametro  $\theta$  da stimare, con probabilità  $1-\alpha$ . Il numero  $\alpha$  è detto *livello di significatività* (v. anche *Significatività*) ed esprime la probabilità di compiere un errore cosiddetto di I tipo, affermando che il valore del parametro da stimare non appartiene all'intervallo di confidenza quando in realtà ciò è vero (rifiuto l'ipotesi nulla  $H_0$  quando questa è vera).

## **Correlazione c. lineare, c. seriale, c. multipla**

Legame di interdipendenza tra due o più variabili statistiche quantitative. Tra due variabili, esiste correlazione, se al variare dell'una anche l'altra varia in modo non casuale.

Se il legame tra due variabili è approssimabile con una funzione lineare, cioè rappresentabile mediante una retta, si parla di c. lineare o di collinearità.

## **Covarianza**

Legame di interdipendenza tra due variabili aleatorie.

## **Deterministico, modello d.**

Un modello deterministico è una equazione che a partire da certe condizioni iniziali note (es. una legge fisica) consente di conoscere con buona approssimazione il risultato del fenomeno in osservazione (es. legge di caduta dei gravi: lascio cadere un grave e osservo che esso raggiunge il suolo con una certa accelerazione).

## **Disegno**

*d. sperimentale, d. fattoriale, d. di miscele, d. fattoriale frazionario, d. D-ottimale, d. di screening*

Sinonimo di progetto, piano di studio, piano degli esperimenti.

**Economia degli effetti, principio di *Sparsity of effects principle*** Principio basato sulla osservazione empirica, fino ad oggi mai confutata, secondo cui la maggior parte dei sistemi (chimici e fisici) è regolata da pochi effetti principali e interazioni di ordine 2; la maggior parte degli effetti riconducibili a interazioni di ordine superiore a 2 è trascurabile. Li et al., Regularities in data from factorial experiments. Complexity (2006) 11(5):33-45. Bergquist et al., A Bayesian Analysis of Unreplicated Two-Level Factorials Using Effects Sparsity, Hierarchy, and Heredity. Quality Engineering (2011) 23(2):152-166.

### **Effetto**

Variazione della risposta sperimentale che è prodotta da uno o più fattori, o dalle loro interazioni.

### **Fattore**

Ognuna delle  $m$  variabili aleatorie, non correlate tra loro, che si ricavano da un insieme più numeroso di  $k$  ( $k > m$ ) variabili che si suppongono interdipendenti.

### **Intervallo di confidenza, v. confidenza**

### **Ipotesi, i. statistica**

Enunciato formulato per indagarne le conseguenze a prescindere dalla sua verità fattuale. Nello studio di un problema, l'ipotesi è la proprietà che si suppone vera e dalla quale mediante una verifica o una dimostrazione obiettiva, si deducono altre proprietà. In statistica, nella procedura di verifica delle ipotesi (test di ipotesi), l'ipotesi nulla  $H_0$  (*null hypothesis*), è l'ipotesi di partenza che costituisce la proposizione espressa sotto forma di equazione verificabile quantitativamente che si formula prima di predisporre un esperimento e analizzare i risultati di un test statistico. Accanto alla ipotesi nulla è formulata la sua negazione denominata ipotesi alternativa e indicata con  $H_1$ .

### **Livello di confidenza, v. Confidenza**

### **Minimi quadrati, metodo dei m.q.**

Metodo di stima usato nei modelli di regressione in cui una variabile dipendente  $y$  è espressa attraverso una funzione (lineare o non lineare) di una o più variabili indipendenti. Il metodo dei minimi quadrati consiste nello scegliere come stime dei parametri che figurano nell'equazione i valori che rendono minima la somma dei quadrati delle differenze tra i valori della variabile  $y$  stimata come dipendente (valori osservati sperimentalmente  $y_i$ , in corrispondenza dei valori di  $x - i$ ) e quelli stimati mediante la funzione. Per fissare le idee, se  $(x_i, y_i)$  sono  $n$  coppie di osservazioni sulle variabili  $X$  e  $Y$  e la relazione ipotizzata tra  $X$  e  $Y$  è lineare, la funzione che lega le due variabili è  $Y = a + bX$ . In corrispondenza di ogni valore  $x_i$  si ha un valore reale osservato  $y_i$  e un valore teorico, detto valore atteso  $\hat{y}_i = a + bx_i$ . Tra ogni valore atteso e ogni valore osservato c'è uno scarto  $d$  calcolato dalla formula:  $d_i = |y_i - \hat{y}_i| = |y_i - (a + bx_i)|$ . La somma dei quadrati di tutti gli scarti  $d_i$  è una misura della distanza tra il modello teorico scelto e i dati osservati. Il metodo di stima dei minimi quadrati porta quindi a scegliere  $a$  e  $b$  in modo tale che sia minima la quantità

$$\sum_{i=1}^n [y_i - (a + bx_i)]^2$$

La retta individuata dai parametri  $a$  e  $b$  così ottenuti prende il nome di *retta dei minimi quadrati* o *retta di regressione*. Il principio dei minimi quadrati assicura di determinare la funzione che con maggiore probabilità si adatta ai dati rilevati. Il metodo di calcolo dei coefficienti  $a$  e  $b$  consiste in un procedimento di approssimazioni successive che, partendo dal valore della media dei valori osservati, attraverso il calcolo degli scarti e successive correzioni della media, permette di stabilire il valore più probabile della stima di  $a$  e  $b$  e fornisce un indice della sua precisione (lo scarto più piccolo appunto).

### **Parametro**

Sinonimo di dato numerico, numero.

### **Percentile, v. quantile**

### **Probabilità**

Valutazione numerica attribuita al possibile verificarsi di un evento aleatorio, cioè casuale.

## Quantile

Indice di posizione che fornisce informazioni sulla struttura della distribuzione di una serie di dati. In una successione di numeri posti in ordine non crescente o non decrescente i quantili dividono la successione in  $n$  gruppi contenenti un uguale numero di osservazioni. In particolare, si parla di *quartile*, *decile* o *percentile* a seconda che si ottengano quattro, dieci o cento gruppi di dati. Se i dati sono ordinati (ad es. dal più piccolo al più grande), i *percentili* sono i 99 valori che dividono l'insieme dei dati in 100 intervalli (da 1 a 100) di uguale numerosità. Il cinquantesimo percentile coincide con la mediana della distribuzione. I percentili, come i decili e i quartili fanno parte del concetto generale di suddivisione di una distribuzione ordinata in  $q$  parti uguali delle *quantili*. Quindi ad esempio se la serie di dati in esame è  $x_1, x_2, \dots, x_9$ , ordinata dal numero più piccolo al numero più grande, allora  $x_1$  è il primo decile, e il 10% dei dati è compreso tra 0 e  $x_1$ , mentre  $x_9$  è il 9° decile e il 90% dei dati è minore di  $X_9$ .

## Risposta

*r. sperimentale, r. di un sistema, r. di un apparecchio, r. di un dispositivo, r. di un esperimento*, il modo con cui il sistema (apparecchio, dispositivo, esperimento) esplica il processo in osservazione, al variare delle condizioni di operazione.

## Scarto, s. interquartile, s. quadratico medio

In statistica si indica con *scarto* il valore di una differenza, per esempio tra un valore osservato e il valore calcolato da una funzione di regressione, oppure tra un valore assunto da una variabile e un valore fisso (es. media o mediana). Il termine *scarto* può essere usato anche per indicare una misura dell'insieme di più differenze: in questo caso il termine *scarto* è seguito da da aggettivi che specificano come sia stata realizzata la sintesi delle differenze e assume il significato di un indice di variabilità come ad esempio lo *scarto* o *differenza interquartile*, che è la differenza tra il terzo e il primo quartile di una distribuzione ( $Q_3 - Q_1$ ). Lo *scarto quadratico medio* è la varianza.

## Significatività

*livello di s.*

Probabilità di commettere un errore di prima specie in un test di verifica di ipotesi, vale a dire che la *s.* è la probabilità di rifiutare l'ipotesi nulla quando questa è vera. E' un numero indicato con la lettera dell'alfabeto greco  $\alpha$ , generalmente fissato pari a 0.05 (e si dice allora che il test è significativo al livello del 5%), oppure a 0.01 (test significativo al livello dell'1%). Nella stima di un parametro sulla base di un campione, il livello di significatività  $\alpha$  è strettamente legato al livello di probabilità dell'intervallo di confidenza prefissato in quanto ne è il complemento a 1 (equivalente a dire al 100%).

## Statistica

1) In generale, disciplina della matematica che si occupa della analisi quantitativa delle osservazioni di un qualsiasi fenomeno soggetto a variazione. In particolare, la statistica è la scienza che si occupa della raccolta, della analisi, della interpretazione, della presentazione e della organizzazione di dati numerici.

2) Risultati numerici di una analisi di dati (es. la deviazione standard relativa percentuale di una serie di misure, RSD%)

3) Valore numerico di un parametro statistico (es. il valore di  $t$  calcolato per una serie di dati).

## Stima, s. puntuale, s. per intervallo

La stima è la assegnazione sulla base dei dati campionari di uno o più valori numerici ad un parametro ignoto che caratterizza una popolazione (es. statura media della popolazione maschile italiana nel 1999). In statistica inferenziale, si distingue tra i valori della popolazione (detti parametri) e i corrispondenti valori numerici che si ricavano dal campione (che rappresentano le stime dei parametri di cui sopra). Quindi sulla base di un campione aleatorio si vuole trovare un valore, o un insieme di valori, che sia la migliore approssimazione possibile del valore incognito del parametro della popolazione. Quando è assegnato un unico valore si parla di *stima puntuale*; se invece è assegnato un insieme di valori, si parla di *stima per intervallo*.

## Test di ipotesi, v. Ipotesi

## Trattamento

Dato un fattore  $X$ , stabiliti i due livelli -1 e +1 entro cui esso può variare, si definiscono trattamenti i due

esperimenti in cui  $X$  assume i valori assegnati,  $X = -1$  o  $X = +1$ . Per più fattori,  $X_1, X_2, \dots, X_n$ , ogni trattamento corrisponde a una combinazione dei livelli ( $\pm 1$ ) degli  $n$  fattori.

**Variabile**, *v. dipendente, v. indipendente*

Ente che può identificarsi con ciascuno degli elementi di un insieme assegnato. Una variabile può assumere tutti i valori all'interno di tale insieme. In una equazione in una incognita, la variabile è l'incognita stessa. La notazione funzionale indica il valore di una variabile dipendente al variare di una o più variabili indipendenti, come ad esempio in  $y = f(x)$ , in cui  $x$  è la variabile indipendente e  $y$  la variabile dipendente da  $x$  secondo la relazione funzionale “ $f$ ”. Se due variabili sono *statisticamente* indipendenti si dice anche che esse sono *incorrelate* (v. *Correlazione*).

**Verifica delle ipotesi**, **v. Ipotesi**

## Bibliografia

Le Garzantine, “Matematica”, edizione in 2 volumi, Garzanti libri s.r.l., Milano 2014.

Everitt B.S., Skrondal A., “The Cambridge Dictionary of Statistics”, fourth edition, Cambridge University press, New York, 2010.

Wonnacott T.H., Wonnacott R.J., “Introduzione alla statistica”, 18a edizione, FrancoAngeli s.r.l., Milano 2002.