Università Ca' Foscari di Venezia

Federica Giummolè

Informazioni generali

Probabilità e Statistica

A.A. 2013/2014

Informazioni generali

• Docente del corso: Federica Giummolè
email: `giummole@unive.it`
web: `http://www.dst.unive.it/~giummole`

• Ricevimento studenti: Martedì 12:00–13:00, via Torino stanza ospiti
`http://www.dst.unive.it/~giummole/avvisi.html`

• Lezioni: Martedì e Giovedì 8:45–10:15, via Torino aula 2

• Tutorato (Marco Fiorucci): Martedì 12:15–13:45, via Torino aula 2 o lab 5

• Moodle: `http://moodle.unive.it/` per materiale didattico e test di autovalutazione

• Testi di riferimento:
Ross, S.M. (2007). *Calcolo delle probabilità*, seconda edizione, Apogeo.
Ross, S.M. (2009). *Probabilità e statistica per l'ingegneria e le scienze*, seconda edizione, Apogeo.

• L'esame consiste in una prova scritta con 10 domande a risposta multipla con sbarramento e alcuni esercizi teorici e pratici, anche sull'uso di R.

Laboratorio con R

R è un ambiente di sviluppo specifico per l'analisi statistica dei dati che utilizza un linguaggio di programmazione derivato e in larga parte compatibile con S.

- R è open-source e può essere scaricato gratuitamente dal sito `http://cran.r-project.org/`
- R funziona sotto UNIX, Windows e Mac
- R ha un *help* approfondito e dettagliato
- R ha eccellenti capacità grafiche
- R è un linguaggio di programmazione con molte funzioni predefinite e la possibilità di costruirne di nuove
- R è mantenuto e aggiornato da una squadra internazionale di esperti. Tutti possono contribuire con *packages* sempre nuovi

Università Ca' Foscari di Venezia

Federica Giummolè

Introduzione alla statistica

Probabilità e Statistica

A.A. 2013/2014

La statistica nella società dell'informazione

- Tutti dicono che viviamo nella società dell'informazione. Ma molti si lamentano che le informazioni sono troppe. E' facile raccoglierle, memorizzarle, distribuirle. E' difficile verificarle ed interpretarle.

- La statistica è la *tecnologia* necessaria per risolvere queste difficoltà.

- Uno statistico sa combinare informazioni di tipo differente, valutarne l'affidabilità, sintetizzare e presentare molti dati in maniera tale da evidenziare la storia che raccontano, costruire modelli (=visioni stilizzate di una parte di mondo) che facilitano l'interpretazione, e, ad esempio, permettono di calcolare previsioni o di formulare ipotesi di decisione.

Un po' di terminologia

- Un insieme (di individui o animali o oggetti o aziende o...) costituisce la parte del mondo che interessa, quella su cui dobbiamo produrre nuove conoscenze, quella che è coinvolta nelle decisioni da prendere. Questo insieme viene chiamato convenzionalmente la **popolazione di riferimento**. Gli elementi della popolazione sono chiamati genericamente **unità statistiche**.

- Alcune caratteristiche di tutte o di una parte delle unità statistiche vengono rilevate/misurate. Il risultato di questo rilevare/misurare costituisce quello che chiamiamo i **dati**. Le unità statistiche sono disomogenee rispetto ai fenomeni rilevati.

- L'obbiettivo è quello di trasformare i dati in nuove conoscenze o ipotesi di decisione. Ovvero, di trasformare i dati in affermazioni sul mondo (sulla popolazione di riferimento).

Un po' di terminologia

- Le caratteristiche rilevate sulle unità statistiche vengono chiamate **variabili**.

- I valori distinti assunti da una variabile sono chiamate le **modalità** della variabile stessa.

- Se le variabili di interesse non sono rilevate su tutte le unità statistiche, il sottoinsieme della popolazione oggetto della rilevazione è chiamato il **campione**.

Tipi di variabili

In statistica si parla di variabili:

- **qualitative** o **categoriali** quando le modalità utilizzate per descrivere il fenomeno analizzato prendono la forma di aggettivi o di altre espressioni verbali. Le variabili qualitative possono essere

  - **sconnesse** se non esiste nessun ordinamento naturale tra le modalità; esempi di variabili sconnesse sono: (i) il sesso, (ii) il tipo di servizio offerto da un albergo;

  - **ordinali** nel caso in cui un ordinamento naturale esiste; esempi di variabili qualitative ordinali sono: (i) il titolo di studio, (ii) il parere di un intervistato (ad es. classificato come "mediocre", "discreto", "buono").

  Quando le modalità sono solamente due (esempi (i) maschio vs. femmina, (ii) vivo vs. morto; (iii) buono vs. difettoso) si parla di variabili **dicotomiche** o **binarie**.

- **numeriche** quando le modalità sono espresse da numeri. Dal punto di vista dei modelli e delle tecniche utilizzate le variabili numeriche si suddividono a loro volta in

  - **discrete** o **intere** quando le modalità sono esprimibili da numeri interi; esempi sono: (i) il numero di clienti, (ii) il numero di pezzi prodotti;

  - **continue** o **reali** quando le modalità sono esprimibili da numeri reali; esempi sono: (i) il tempo d'attesa ad uno sportello, (ii) il peso di un manufatto.

Piccolo esempio (per fissare la terminologia)

Vogliamo avere un'idea sul numero di clienti e sul volume di vendite dei negozi di una città per tre categorie merceologiche ritenute simili. La popolazione di riferimento è l'insieme di tutti i negozi secondo le tre categorie merceologiche. Le unità statistiche sono i negozi. I dati si presentano in questa forma:

negozio	clienti	vendite	categoria
1	907	11.2	A
⋮	⋮	⋮	⋮
10	420	6.12	B
11	679	7.63	B
⋮	⋮	⋮	⋮
19	1010	11.77	C
20	621	7.41	A

Le variabili considerate nello studio sono tre:

clienti le cui *modalità* sono numeriche e discrete;

vendite (in migliaia di euro) le cui *modalità* sono numeriche e (con approssimazione) continue.

categoria le cui *modalità* sono sconnesse (A, B e C.)

Il modo in cui sono raccolti i dati può condizionare il loro tipo

Si consideri una macchina che deve forare delle lastre di metallo. Il diametro nominale dei fori è di $1mm$ con una tolleranza di $0{,}06mm$. Ovvero un foro è *ben fatto* se il suo diametro è compreso tra $0{,}94mm$ e $1{,}06mm$.

Allora, dati sulla *qualità* della produzione della macchina, potrebbero essere disponibili nella forma

1. "buono" vs. "difettoso" (dati dicotomici);

2. "troppo piccolo", "buono", "troppo grande" (dati qualitativi ordinali);

3. lunghezza del diametro (dati numerici continui).

Si osservi che le differenze non sono semplicemente dovute a come i dati vengono registrati ma possono anche essere dovute a come *i diametri vengono effettivamente misurati*. Ad esempio, raccogliere dati sui diametri nella forma (2) è più rapido e richiede strumenti meno costosi (bastano due bastoncini metallici di diametro rispettivamente uguale ai due estremi dell'intervallo di tolleranza) di quanto richiesto dalla forma (3).

Dati sperimentali vs dati osservazionali

Nell'analizzare dei dati è bene poi tenere presente il tipo di studio in cui sono stati rilevati. In particolare, è importante la distinzione tra:

- **studi sperimentali** ovvero situazioni in cui i dati sono stati raccolti in situazioni replicabili e controllate (esempio classico sono gli esperimenti di laboratorio);

- **studi osservazionali** ovvero situazioni in cui il ricercatore semplicemente rileva dei dati già esistenti (esempio: il numero di presenze alberghiere in una stagione, il prezzo di un'azione,...).

Il problema principale degli studi osservazionali è che non controllando i fattori che possono influenzare il fenomeno sotto indagine risulta difficile essere certi di averli individuati appropriatamente.

Metodi di raccolta dei dati

1. Esperimenti in laboratorio

2. Interviste telefoniche

3. Questionari inviati per posta

4. Social network

5. Carte fedeltà

6. ...

Il **campionamento** e il **disegno degli esperimenti** si occupano delle problematiche connesse con la raccolta dei dati.

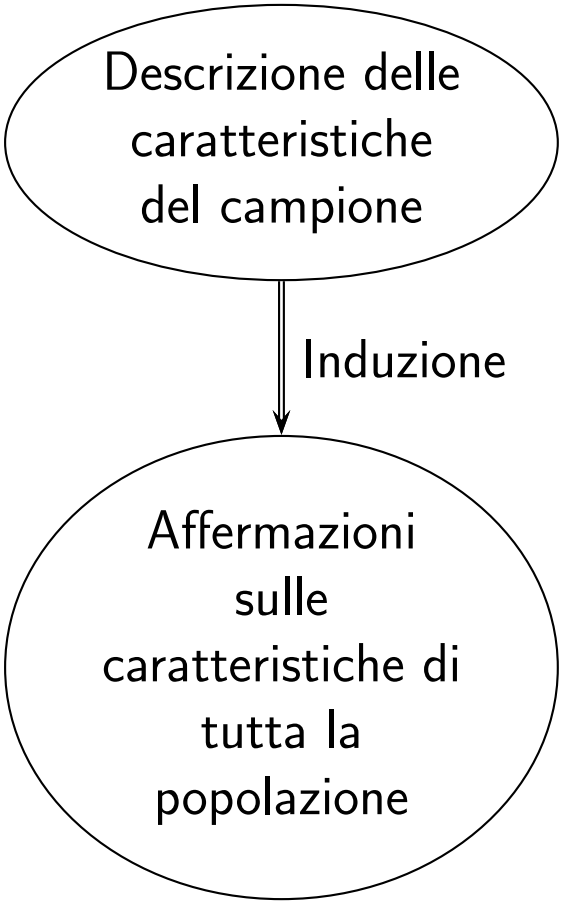
Statistica Descrittiva e Inferenza Statistica

Descrittiva: metodi per rappresentare, sintetizzare ed evidenziare le caratteristiche più significative di un insieme di dati. Usualmente si dispone di dati su tutta la popolazione di riferimento.

Inferenza: i dati disponibili sono stati rilevati solamente su una parte delle unità statistiche (il campione, da cui *indagini campionarie*). Si vogliono utilizzare le informazioni del campione per fare delle affermazioni sulle caratteristiche generali di tutta la popolazione.

Statistica Descrittiva e Inferenza Statistica

Statistica Descrittiva ed **Inferenza Statistica** nelle applicazioni non sono facilmente separabili. Infatti i problemi di *inferenza* vengono normalmente affrontati in accordo allo schema



La statistica descrittiva viene dunque utilizzata per un'analisi preliminare delle caratteristiche del campione.

Calcolo delle probabilità

Perché l'inferenza porti a risultati sensati, bisogna che sia noto il legame fra popolazione e campione. In particolare, il campione deve essere scelto in modo che rappresenti, in piccolo, la popolazione.

La relazione fra campione e popolazione si descrive attraverso il **calcolo delle probabilità**.

E' il calcolo delle probabilità che fornisce gli strumenti per l'inferenza e che permette di quantificare gli errori che commettiamo nel passaggio dal particolare (campione) al generale (popolazione).