Introduzione all'inferenza statistica

Carla Rampichini

Dipartimento di Statistica "Giuseppe Parenti" - Firenze - Italia

carla@ds.unifi.it - www.ds.unifi.it/rampi/

Indice degli argomenti trattati

- Statistica e metodologia della ricerca
- Introduzione ai diversi approcci all'inferenza statistica
- Inferenza statistica classica:
  - Principi di riduzione dei dati
  - Teoria degli stimatori
  - Test delle ipotesi
  - Modelli lineari: stima e test delle ipotesi

Bibliografia essenziale

- Casella G. e Berger R. L. (2002), *Statistical Inference*, 2nd Edition, Duxbury Press.
- Piccolo D. (2000), *Statistica*, Il Mulino, Bologna.

Per approfondimenti:

- Barnett V. (1999), *Comparative Statistical Inference*, 3rd Edition, John Wiley and Sons.
- De Groot M. H. (1970), *Optimal Statistical Decisions*, New-York: MacGraw-Hill.
- Hoel P. G. ,Port S. C. e Stone C.J. (1971) Introduction to statistical theory, Boston: Houghton Mifflin.
- Lehmann E.L. (1986), *Testing Statistical Hypotheses*, 2nd Edition, New York: Wiley.
- Lehmann E.L. and Casella G. (1998), *Theory of point estimation*, 2nd Edition, New York: Springer-Verlag.
- Lindley D. V. (1965), *Introduction to probability and statistics from a bayesian viewpoint*, Cambridge: Cambridge University Press.
- Lindley D. V. (1985), *Making decisions*, Chichester: Wiley.
- Rubin D.B. e Little R.J.A. (2002), *Statistical analysis with missing data*, New-York: Wiley.

Statistica e metodologia della ricerca

Definiamo STATISTICA un metodo per il trattamento dell'informazione che consente di riflettere su e dare un'indicazione per l'azione in situazioni di incertezza.

Situazione di incertezza

1. c'è più di un possibile risultato
2. il risultato non è noto in anticipo, è indeterminato

Siamo interessati a:

- conoscere quale sarà il risultato
- decidere come agire in base al risultato che si presenterà

$\rightsquigarrow$ Costruzione di un modello formale: teoria del comportamento in situazioni di incertezza

$\rightsquigarrow$ formulazione del concetto di probabilità: per distinguere tra i risultati in base al grado di incertezza.

Modello probabilistico

Il modello è una semplificazione della realtà.
Obiettivo dello statistico è la costruzione di un modello adeguato per la **descrizione** e/o per la guida nelle **decisioni**.

Per la costruzione del modello è necessario specificare:

- l'insieme dei possibili risultati
- il meccanismo probabilistico che genera i dati

$\rightsquigarrow$ La situazione reale viene sostituita dal modello. Se il modello è **adeguato** è possibile:

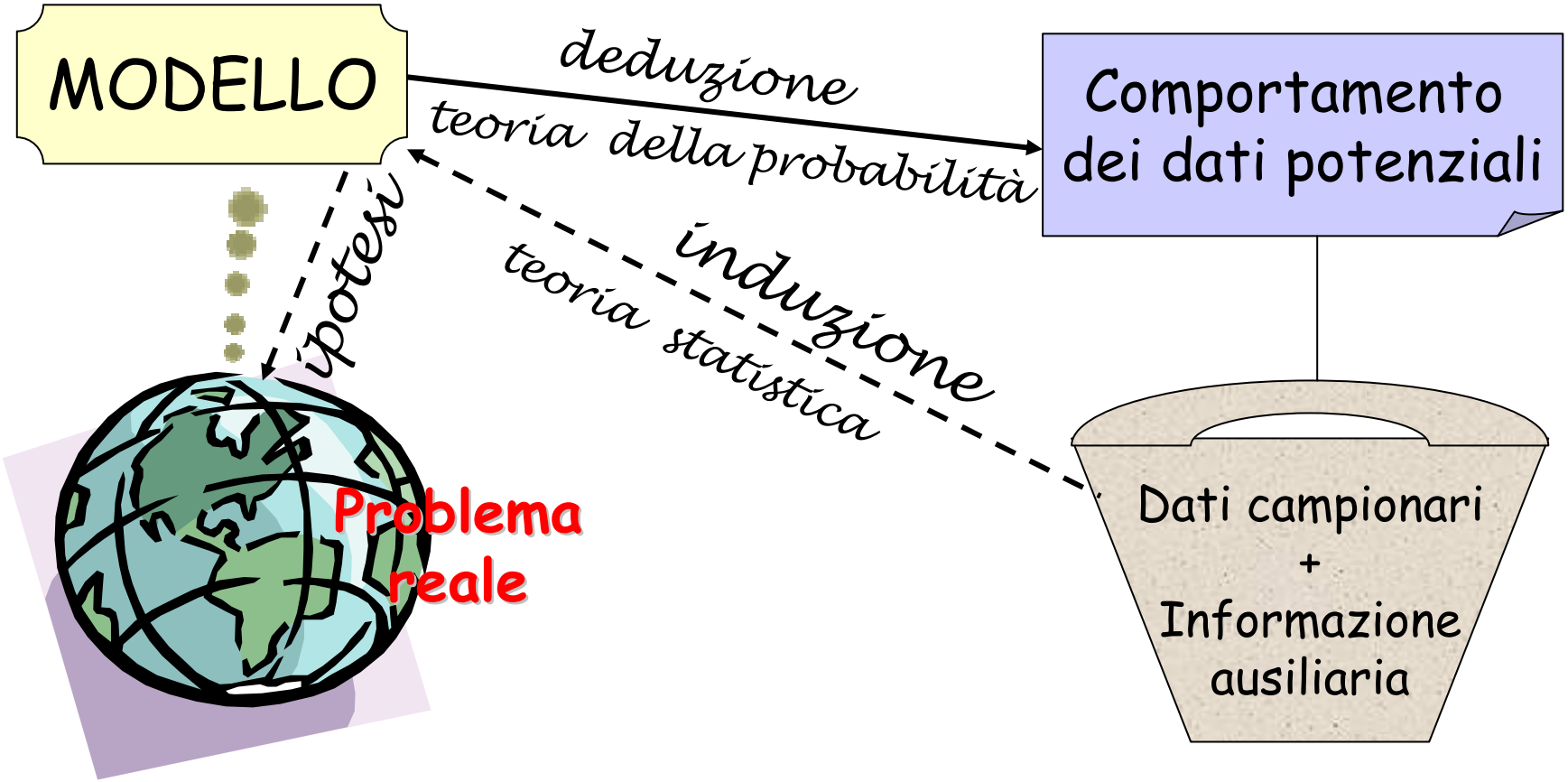
- **dedurre** il comportameto caratteristico dei dati derivanti dal modello e quindi, per *assunzione*, dalla situazione reale.
- utilizzare una procedura statistica, p.e. analisi della varianza, per il problema **inverso**: usare i dati osservati per STIMARE o CONVALIDARE il modello.

Le componenti della teoria statistica

Gli *ingredienti* della teoria statistica sono:

- fenomeno (situazione pratica)
- modello
- informazione

Queste componenti sono legate tra loro in maniera deduttiva o induttiva.



Legame tra le componenti della teoria statistica

Informazione rilevante

A seconda delle circostanze i seguenti 3 tipi di informazione possono essere rilevanti per l'analisi statistica

- passata (o esterna): informazione a priori (esperienza)
- situazione attuale: dati campionari
- futuro: possibili conseguenze (utilità: modello razionale del comportamento umano, come si compie la scelta tra possibili alternative in situazione di incertezza)

Informazione a priori e conseguenze possono essere difficili da quantificare e sono spesso *soggettive*.

Comunque sia quantificata l'informazione, per utilizzarla sono necessarie metodologie statistiche in grado di incorporarla.

Diversi approcci all'inferenza statistica

Funzioni dell'analisi statistica:

- descrivere un fenomeno e/o
- fornire regole per l'azione nel contesto di tale fenomeno.

Inferenza statistica: utilizza l'informazione per ottenere una descrizione del fenomeno attraverso un modello probabilistico

Decisione statistica: procedura inferenziale che suggerisce l'azione da intraprendere

Approcci principali all'inferenza statistica

- I dati campionari come unica fonte di informazione: Inferenza statistica classica
- Informazione a priori: Statistica bayesiana
- Costi e conseguenze: Teoria Statistica delle decisioni

Statistica classica

Origina dai lavori di Fisher, Neyman, Pearson e altri.

Include le procedure di:

- stima puntuale e per intervalli
- test di significatività e delle ipotesi

Si basa su:

- dati campionari rappresentati attraverso la funzione di *verosimiglianza*
- impostazione *frequentista* della probabilità
- *distribuzioni campionarie*

La bontà delle procedure è valutata in base alle caratteristiche delle distribuzioni campionarie (p.e. stimatori puntuali corretti o consistenti).

Statistica bayesiana

Si basa su:

- dati campionari+informazione a priori
- impostazione *frequentista* + *soggettiva* della probabilità

L'informazione a priori è modificata dai dati campionari attraverso l'utilizzo del **teorema di Bayes** (Lindley, 1965).

Inferenza espressa attraverso **distribuzioni di probabilità a posteriori**, incorpora una misura della propria accuratezza.

Fondamentali i concetti di *coerenza* (razionalità degli individui in situazioni di incertezza) e *scambiabilità* (degli eventi).

Teoria statistica delle decisioni

Introdotta da Wald (1950).

Fornisce **regole di decisione** in situazioni di *incertezza*.

Considera le *conseguenze* di azioni *alternative*, espresse attraverso la *teoria dell'utilità* sotto forma di **funzioni di perdita**.

Obiettivo: scegliere la decisione cui è associato il **rischio minimo**.

Si basa su:

- dati campionari+informazione a priori+conseguenze
- non è richiesta un'impostazione particolare della probabilità
- approccio inferenziale può essere classico o bayesiano

Principali caratteristiche dei 3 approcci

Approccio	*Funzione*	*Probabilità*	*Informazione*
Classico	Inferenziale (prevalente)	frequentista	dati campionari
Bayesiano	Inferenziale	grado di fiducia soggettivista frequentista	dati campionari a priori
Teoria delle decisioni	Decisioni	frequentista (soggettiva se incorpora a priori)	dati campionari conseguenze (perdite o utilità) (a priori)

Inferenza statistica e causalità

Se due variabili sono *associate statisticamente*, c'è un legame di **causa-effetto**?

p.e. gli incidenti stradali sono cresciuti negli anni (fino a un certo periodo) e sono aumentati anche i camion per uso commerciale circolanti. Questa è una *relazione statistica*. Possiamo affermare che l'incremento del trasporto merci su strada ha **causato** l'aumento degli incidenti?

♠ L'inferenza statistica consente di esaminare l'associazione tra fattori. Tale associazione **non implica** necessariamente causalità.

♠ L'inferenza statistica può essere utilizzata per l'analisi causale a particolari condizioni (Rubin *et al.*, 2002; Cox, 1992)

Riduzione dei dati

- Si vuole fare inferenza su un parametro incognito θ in base all'informazione fornita dal campione casuale $\mathbf{X} = (X_1, X_2, \ldots, X_n)$
- il campione osservato è $\mathbf{x} = (x_1, x_2, \ldots, x_n)$
- l'informazione campionaria è sintetizzata attraverso una statistica, funzione dei dati campionari (media, varianza, valore più piccolo e più grande, ecc.)

Ogni statistica $T(\mathbf{X})$ definisce una sintesi o *riduzione* dei dati.

♣ Se si fa inferenza utilizzando la statistica al posto dell'intero campione, due campioni $\mathbf{x}$ e $\mathbf{y}$ tali che $T(\mathbf{x}) = T(\mathbf{y})$ forniscono la stessa sintesi dei dati e sono equivalenti

Statistica e spazio campionario

- La statistica $T(\mathbf{x})$ definisce una partizione dello spazio campionario $\mathcal{X}$
- L'immagine di $\mathcal{X}$ attraverso $T(\mathbf{X})$ è:
  $\mathcal{T} = \{t : t = T(\mathbf{X}),$ per qualche $\mathbf{x}\epsilon\mathcal{X}\}$
- $T(\mathbf{X})$ ripartisce $\mathcal{X}$ nei sottinsiemi $A_t, t\epsilon\mathcal{T}$, definiti da $A_t = \{\mathbf{x} : T(\mathbf{X}) = t\}$

$T(\mathbf{X})$ sintetizza i dati nel senso che riporta solo $T(\mathbf{x}) = t$ o $\mathbf{x}\epsilon A_t$ invece dell'intero campione.

$\rightsquigarrow$ Tale sintesi comporta vantaggi e conseguenze.

Principi di riduzione dei dati

$\rightsquigarrow$ La statistica $T(\mathbf{X})$ sintetizza i dati.

∇ Interessano metodi che non scartino informazione rilevante per θ e che scartino informazione che non serve.

Vedremo 3 principi di riduzione dei dati:

- Principio di sufficienza: sintetizza i dati senza scartare informazione su θ
- Principio di verosimiglianza: funzione dei parametri determinata dal campione osservato che conserva l'informazione su θ contenuta nel campione
- Principio di equivarianza: opera una sintesi dei dati conservando alcune caratteristiche rilevanti del modello.

Principio di Sufficienza

Una *statistica sufficiente* per θ è una statistica che cattura tutta l'informazione su θ contenuta nel campione.

PRINCIPIO DI SUFFICIENZA: Se $T(\mathbf{X})$ è una statistica sufficiente per θ, allora ogni inferenza su θ dipende dal campione $\mathbf{X}$ solo attraverso il valore $T(\mathbf{X})$.
Cioè, se $\mathbf{x}$ e $\mathbf{y}$ sono due punti campionari tali che $T(\mathbf{x}) = T(\mathbf{y})$, allora l'inferenza su θ deve essere la stessa che si osservi $\mathbf{X} = \mathbf{x}$ o $\mathbf{X} = \mathbf{y}$.

DEFINIZIONE: Una statistica $T(\mathbf{X})$ è una *statistica sufficiente* per θ se la distribuzione condizionata del campione $\mathbf{X}$ dato il valore $T(\mathbf{X})$ non dipende da θ.

Statistica sufficiente

- sia t un valore possibile di $T(\mathbf{X})$: $P_\theta(T(\mathbf{X}) = t) > 0$
- ci interessa $P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t)$
- se $T(\mathbf{x}) \neq t \Rightarrow P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) = 0$
- quindi ci interessa $P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x}))$
- se $T(\mathbf{X})$ è sufficiente per θ
  $\Rightarrow P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})) = P(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x}))$
  non dipende da θ, $\theta \epsilon \Theta$.

Una statistica sufficiente cattura tutta l'informazione su θ in questo senso.

$T(\mathbf{X})$ è sufficiente?

✓Se $T(\mathbf{X})$ è sufficiente per θ, $P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t)$ deve essere costante al variare di $\theta \epsilon \Theta$, per ogni valore dato di $\mathbf{x} \epsilon \mathcal{X}$ e $t \epsilon \mathcal{T}$.

✓ poichè $P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) = 0$ per tutti i valori $t \neq T(\mathbf{x})$
$\Rightarrow$ basta verificare che $P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x}))$ non dipende da θ.

✓ L'evento $\{\mathbf{X} = \mathbf{x}\}$ è un sottoinsieme dell'evento $\{T(\mathbf{X}) = T(\mathbf{x})\}$,

$$
\begin{aligned}
P_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = T(\mathbf{x})) &= \frac{P_\theta(\mathbf{X}=\mathbf{x} \quad \& \quad T(\mathbf{X})=T(\mathbf{x}))}{P_\theta(T(\mathbf{X})=T(\mathbf{x}))} \\
&= \frac{P_\theta(\mathbf{X}=\mathbf{x})}{P_\theta(T(\mathbf{X})=T(\mathbf{x}))} = \frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{X}|\theta))}
\end{aligned}
$$

- $p(\mathbf{x} \mid \theta)$ è la funzione di massa di probabilità congiunta di $\mathbf{X}$
- $q(T(\mathbf{X}) \mid \theta)$ è la funzione di massa di probabilità di $T(\mathbf{X})$

$\Rightarrow$ $T(\mathbf{X})$ è sufficiente per θ **se e solo se** $\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{X}|\theta))}$ è costante al variare di θ per ogni **x**.

$T(\mathbf{X})$ è sufficiente? (2)

TEOREMA Se $p(\mathbf{x} \mid \theta)$ è la funzione di densità (*pdf*) o di massa di probabilità (*pmf*) congiunta di $\mathbf{X}$ e $q(T(\mathbf{X} \mid \theta))$ è la *pdf* o la *pmf* di $T(\mathbf{X})$, allora $T(\mathbf{X})$ è una statistica sufficiente per θ se, per ogni $\mathbf{x}\epsilon\mathcal{X}$, il rapporto $p(\mathbf{x} \mid \theta)/q(T(\mathbf{X} \mid \theta))$ è costante al variare di θ.

Alcuni esempi di statistiche sufficienti:

- distribuzione binomiale: n. di *successi* $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$
- distribuzione normale, σ^2 nota: *media campionaria* $T(\mathbf{X}) = (1/n)\sum_{i=1}^{n} X_i = \overline{x}$
- statistiche d'ordine: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$