INFERENZA STATISTICA

Nell'analisi di un fenomeno reale di natura casuale (altezza o altro carattere di una popolazione di individui, lunghezza delle sbarre prodotte da una catena di fabbriche, o piú in generale caratteristiche di determinati oggetti) la legge di probabilità cui esso obbedisce é in generale sconosciuta (a meno che non si abbiano informazioni precise a priori). In casi particolari potremmo sapere che il fenomeno ha una distribuzione normale (ma non conosciamo μ e σ), o binomiale (prove ripetute indipendenti, ma non conosciamo la probabilità p di successo in ogni prova).
Come determinare la legge del fenomeno (o quantomeno il valore di alcune grandezze significative relative al fenomeno), sulla base di informazioni ricavate da un campione estratto dalla popolazione?
Di questo problema si occupa la Statistica inferenziale la quale, attraverso l'analisi dei dati forniti dal campione, ha lo scopo di definire univocamente (se possibile), la legge probabilistica incognita che descrive la natura aleatoria del fenomeno che studiamo.
In alcuni casi la tipologia di problema porta a supporre che il fenomeno segua una determinata legge di cui bisogna determinare i parametri (ad esempio media e varianza). In altre situazioni non si può ipotizzare nulla sulla distribuzione del fenomeno: é utile stimare quantomeno media e varianza dello stesso, perché ció permette di calcolare la probabilità di eventi ad esso legati senza conoscerne la distribuzione
Esempi di indagini: distribuzione delle età delle persone che frequentano le sale cinematografiche, distribuzione delle età delle persone affette da una determinata patologia, distribuzione dello spessore delle lastre prodotte da una ditta, distribuzione della quantità di liquido contenuto nelle bottiglie di coca cola.
Tali distribuzioni possono avere andamenti disparati e non possiamo a priori stabilire che seguano una distribuzione nota. I problemi di questo tipo si dicono di inferenza non parametrica. Quelli in cui possiamo a priori stabilire che tipo di distribuzione segua il fenomeno si dicono di inferenza parametrica.
Entrambi hanno in comune il fatto che bisogna in qualche modo stimare la media e la varianza del fenomeno.
Data una popolazione X, su cui vogliamo effettuare una indagine statistica, la media μ e la varianza σ^2 della grandezza oggetto di studio, sono ignote. Per conoscerle dovremmo analizzare l'intera popolazione e ciò non é possibile. In alcuni casi la popolazione é infinita (intervalli temporali ad esempio). Possiamo prelevare un campione e fare indagini su di esso. Diciamo ampiezza del campione il numero di individui che lo compongono. Se potessimo estrarre tutti i possibili campioni di ampiezza prefissata e farne la media campionaria otterremmo una nuova distribuzione

$$\overline{X}_n = \frac{X_1 + \ldots\ldots X_n}{n} \tag{1}$$

che si chiama distribuzione delle medie campionarie. **Attenzione: é una nuova variabile aleatoria!** Le X_i sono "copie" di X. Una variabile aleatoria come quella introdotta in (1) si chiama statistica (termine infelice...): tale denominazione si utilizza per tutte le **variabili aleatorie che dipendono dai dati di un campione**.
Introdurremo in seguito la Varianza campionaria, un'altra statistica.
La distribuzione delle medie campionarie entra in gioco anche in problemi nei quali é nota la legge di X (quindi μ e σ), ma bisogna calcolare delle probabilitá per cui

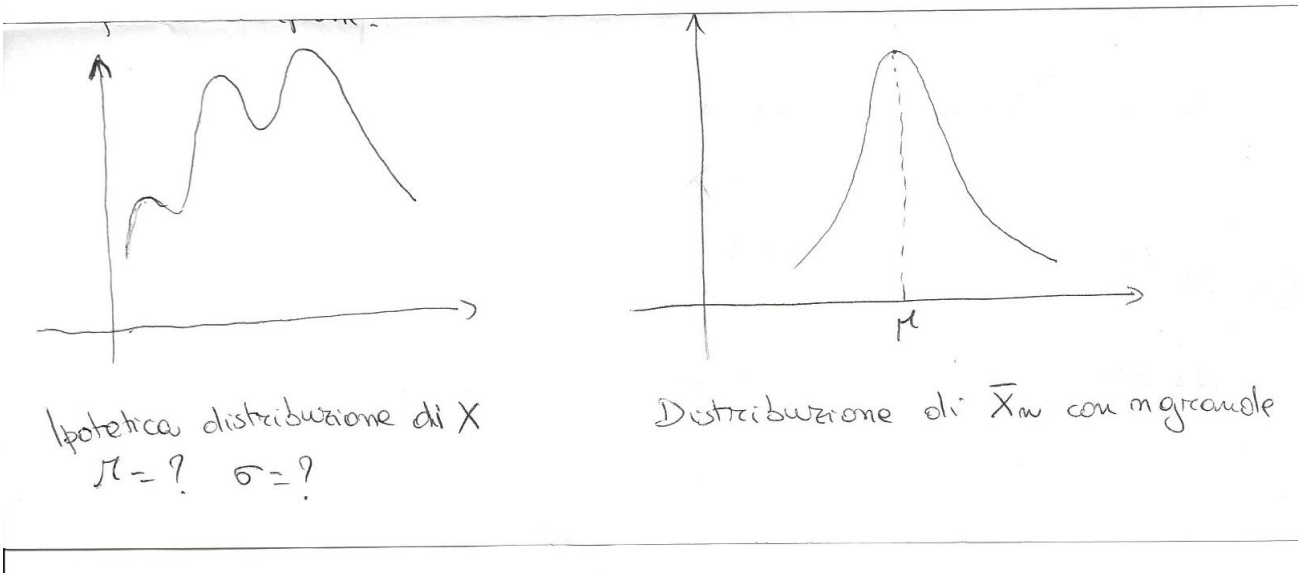
serve $\overline{X}_n$. Tali questioni sono piú semplici da affrontare.
Ovviamente non è possibile costruire materialmente $\overline{X}_n$ (non possiamo prendere TUTTI i campioni di una prefissata ampiezza), ma per i problemi in cui μ e σ sono incognite ci verranno in aiuto dei risultati astratti che consentono di **valutare μ a partire dalla media $\overline{x}_n$ e dalla varianza campionaria s_n di un singolo campione**. Ricordiamo che

$$\overline{x}_n = \frac{x_1 + \ldots\ldots x_n}{n} \tag{2}$$

e

$$s_n = \sqrt{\frac{\sum_{i=1}^n (x_i - \overline{x}_n)^2}{n-1}}\,. \tag{3}$$

A volte scriveremo $\overline{x}$ e s anziché $\overline{x}_n$ e s_n.



Vedremo in seguito quanto deve essere ampio il campione.
Calcoliamo media e varianza di $\overline{X}_n$. Ricordiamo che poiché le X_i sono copie di X ciascuna di esse ha media μ e varianza σ^2 (generalmente incognite).

$$E(\overline{X}_n) = E\left(\frac{X_1 + \ldots\ldots X_n}{n}\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n}\sum_{i=1}^n \mu = \mu\,. \tag{4}$$

$$Var(\overline{X}_n) = Var\left(\frac{X_1 + \ldots\ldots X_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}\,. \tag{5}$$

La (4) ci dice che $\overline{X}_n$ é uno stimatore corretto per μ. Ogni singolo valore $\overline{x}_n$ si dice stima puntuale.

Definizione 0.1. *Uno stimatore di una grandezza incognita é una v.a. avente come scopo quello di stimare il parametro incognito.*

Definizione 0.2. *Uno stimatore si dice corretto (o non distorto) se la sua media coincide con il parametro incognito da stimare. Altrimenti si dice distorto.*

Definizione 0.3. *Una successione di stimatori Y_n di una grandezza incognita ϑ si dice consistente in probabilità se, per qualsiasi ϑ compatibile con il modello statistico, essa converge al valore teorico ϑ con probabilità 1 quando la numerosità n del campione tende a infinito:*

$$\lim_{n\to+\infty} P(|Y_n - \vartheta| < \varepsilon) = 1 \ \forall \varepsilon > 0\,. \tag{6}$$

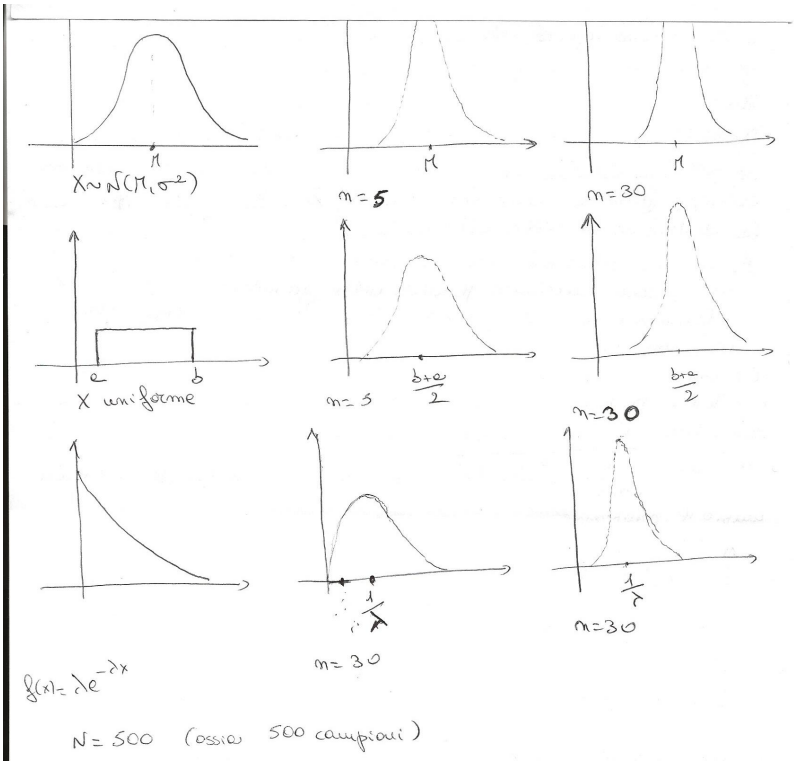
Definizione 0.4. *Una successione di stimatori Y_n di una grandezza incognita ϑ si dice consistente in media quadratica se, per la successione degli errori quadratici medi si ha*

$$\lim_{n\to+\infty} E[(Y_n - \vartheta)^2] = 0\ . \tag{7}$$

Uno stimatore (o meglio una successione di stimatori) consistente in media quadratica é consistente in probabilità. Se Y_n é uno stimatore corretto per ϑ allora, $E[(Y_n - \vartheta)^2] = \sigma^2(Y_n)$, per cui se $\lim_{n\to+\infty} \sigma^2(Y_n) = 0$ allora Y_n é consistente in media quadratica.
Nel caso di $\overline{X}_n$, la (5) ci dice che $\overline{X}_n$ é consistente in media quadratica.

Definizione 0.5. *Date due successioni di stimatori corretti $Y_{1,n}$, $Y_{2,n}$ di una grandezza incognita ϑ, diciamo che $Y_{1,n}$ é più efficiente di $Y_{2,n}$ se*

$$\sigma^2(Y_{1,n}) \le \sigma^2(Y_{2,n}) \ \forall n \in N\,. \tag{8}$$



Teorema 0.6 (Teorema del limite centrale, TLC)**.** *Sia* $\{X_n\}_{n\in N}$ *una successione di variabili aleatorie indipendenti ed equidistribuite (ossia aventi la medesima legge), di media* μ *e varianza* σ^2 *finite. Allora la variabile standardizzata*

$$S_n^* = \frac{X_1 + \ldots\ldots X_n - n\mu}{\sigma\sqrt{n}} = \frac{n\left(\frac{X_1+\ldots\ldots X_n}{n} - \mu\right)}{\sigma\sqrt{n}} = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

tende in legge a una $\mathcal{N}(0,1)$*, ossia*

$$\lim_{n\to+\infty} P\left(\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \le z\right) = \Phi(z)\,. \tag{9}$$

Il TLC ci dice che per n "grande" $\frac{\overline{X}_n-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$:

$$P\left(\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \le z\right) \approx \Phi(z) \tag{10}$$

e $X_1 + \ldots\ldots X_n \sim \mathcal{N}(n\mu, n\sigma^2)$

(11)

$$P(X_1 + \ldots\ldots X_n \le x) = P\left(\frac{X_1 + \ldots\ldots X_n - n\mu}{\sigma\sqrt{n}} \le \frac{x - n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

(perché $X_1 + \ldots\ldots X_n$ ha media $n\mu$ e varianza $n\sigma^2$).

Indipendentemente dalla distribuzione di X, la distribuzione di $\overline{X}_n$ é approssimativamente normale ($\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$). Basta già $n \ge 30$ per visualizzare l'andamento. Se la distribuzione di X é simmetrica allora $\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ già per $n \ge 5$ (vedi figura 2).

Soffermiamoci sull'ipotesi **variabili aleatorie indipendenti ed equidistribuite.** La condizione di equiprobabilità delle estrazioni degli elementi del campione (ossia che le variabili X_i siano indipendenti) implica che nella costruzione del campione, l'estrazione di ciascuno degli n elementi dalla popolazione debba essere effettuata con rimpiazzo. Le estrazioni di un campione aleatorio vengono invece fatte senza rimpiazzo (e questo é chiaro se pensiamo allo scopo dell'indagine). Abbiamo visto che l'estrazione senza rimpiazzo dà luogo a v.a. dipendenti (e con distribuzioni diverse). Nel caso di popolazione N poco numerosa (cosa che accade RARAMENTE nei fenomeni che noi studiamo), o quando la dimensione n del campione non é trascurabile rispetto a quella dell'intera popolazione (N), i risultati dovranno essere corretti. La correzione diventa trascurabile quando la numerosità N degli individui della popolazione é elevata, ed é teoricamente nulla per $N = \infty$. In ogni caso i numeri con cui lavoriamo ci consentiranno **SEMPRE di trattare le** X_i **come v.a. indipendenti ed equidistribuite** (riguardate l'esempio delle estrazioni con e senza rimpiazzo).

Il TLC si utilizza in piú modi:

1) per valutare proabilità relative a fenomeni che seguono una distribuzione di media e deviazione standard note;

2) per fornire un valore approssimato di una media incognita (intervalli di confidenza). La stima di μ diventa via via più precisa al crescere di n, ma dipende da $\frac{\sigma}{\sqrt{n}}$ e se σ é incognita dobbiamo riuscire a stimare anche σ se vogliamo stimare

μ. Si dimostra (vedremo dei cenni di ciò nelle prossime lezioni) che per $n >> 1$ la deviazione standard campionaria (3) (ottenuta su un singolo campione!!!!) approssima bene σ per cui possiamo sostiuirla a σ. Il campione va però scelto nel modo giusto (una scelta random é abbastanza corretta): se selezioniamo ad esempio n individui che presentano lo stesso dato x_i allora $s_n = 0$ e non va bene. Chiariremo meglio in seguito,
3) per verificare se il valor medio fornito per una grandezza è plausibile (test d'ipotesi).

ESERCIZI (caso 1)
1) Una compagnia ha 20000 polizze attive e sa che il risarcimento dovuto annualmente per ogni assicurato é una v.a. con media 320 euro e deviazione standard 540 euro. Quanto vale la probabilità che in un anno le richieste di indennizzo superino i $7,5$ milioni?
Siano $X_i,\ i = 1 \dots 20000$ le v.a. che danno il risarcimento dovuto annualmente per ogni assicurato (media e varianza note). Il risarimento annuo sarà $\sum_{i=1}^{20000} X_i$. Usando la (11) (con il $>$)

$$P(X_1 + \ldots\ldots + X_{20000} > 7,5 \cdot 10^6)$$
$$= P\left(\frac{X_1 + \ldots\ldots + X_{20000} - 20000 \cdot 320}{540\sqrt{20000}} > \frac{7,5 \cdot 10^6 - 6,4 \cdot 10^6}{76367,5}\right)$$
$$\approx 1 - \Phi\left(\frac{1,1 \cdot 10^6}{7,6 \cdot 10^4}\right) \approx \ldots\ldots \quad \square$$

Prima di svolgere ulteriori esercizi introduciamo quella che si chiama correzione di continuità. Si fa quando le X_i assumono solo valori interi (ad esempio numero di giorni di durata di un macchinario). A differenza del caso precedente, in cui il risarcimento si può ipotizzare come un continuo (le cifre sono notevoli e la differenza tra due valori successivi é dell'ordine del centesimo di euro) in un quesito sulle settimane (o giorni o mesi) di durata di un dispositivo si ragiona per unità temporali. Supponiamo che ci venga chiesta $P(X_1 + \ldots + X_n > 10$ giorni), ossia $P(X_1 + \ldots + X_n \geq 11$ giorni). Se ragioniamo per unità

$$P(X_1 + \ldots + X_n \geq 11 \text{ giorni}) + P(X_1 + \ldots + X_n \leq 10 \text{ giorni}) = 1\,.$$

Nel passaggio al continuo se lasciamo il 10 e l'11 perdiamo l'area tra 10 e 11. Sostituiamo alla linea rappresentativa $P(X = k)$ il rettangolo di base $[k - 0,5,\ k + 0,5]$. Cosi $P(X_1 + \ldots + X_n \geq 11$ giorni) si legge $P(X_1 + \ldots + X_n \geq 10,5$ giorni) e ad essa corrisponderà l'area da $10,5$ a ∞. Quindi

$$P(X_1 + \ldots + X_n > 10 \text{ giorni}) = P((X_1 + \ldots + X_n > 10,5 \text{ giorni})$$

e

$$P(X_1 + \ldots + X_n \leq 10 \text{ giorni}) = P(X_1 + \ldots + X_n \leq 10,5 \text{ giorni})\,.$$

2) Il numero di settimane di funzionamento di un certo tipo di batterie é una v.a. con media 7 e deviazione standard $1,6$. Calcolare approssimativamente la probabilità che in un anno (52 settimane) si debbano impiegare 8 o più batterie. Stessa domanda ma con 12 o più
Siano $X_i,\ i = 1 \dots 7$, le v.a che danno i tempi di durata di sette diverse batterie. Se in un anno ne devo usare 8 o più allora la somma delle durate delle 7 batterie è

inferiore a 52 settimane.

$$P(X_1 + \ldots + X_7 < 52) = P(X_1 + \ldots + X_7 \leq 51,5)$$
$$= P\left(\frac{X_1 + \ldots + X_7 - 7 \cdot 7}{1,6\sqrt{7}} \leq \frac{51,5 - 49}{1,6\sqrt{7}}\right) \approx \Phi\left(\frac{2,5}{4,23}\right) \approx ..... \quad \square$$

3) Una casa produttrice di sigarette dichiara per un tipo specifico un contenuto medio di nicotina di $2,2\ mg$ e deviazione standard $0,3\ mg$. Qual é la probabilità che su uncampione di 100 sigarette si trovi un contenuto medio di nicotina pari o superiore a $3,1\ mg$?
Ora il problema é sulla media e non sulla semplice somma. Siano $X_i,\ i = 1 \ldots 100$, le v.a che danno il contenuto di nicotina di ogni sigaretta. Sia $\overline{X}_{100}$ la loro media. Allora

$$P(\overline{X}_{100} \geq 3,1) = P\left(\frac{\overline{X}_{100} - 2,2}{0,3}\sqrt{100} \geq \frac{3,1 - 2,2}{0,3} \cdot 10\right) \approx 1 - \Phi\,(30) \approx 0\,. \quad \square$$

4) Il tempo ~~medio~~ di vita di un componente elettrico é una v.a. con media 100 *ore* e deviazione standard 20 *ore*. Qual é la probabilità che il tempo ~~medio~~ di vita di 16 componenti sia compreso tre 96 e 102 ore?
Siano $X_i,\ i = 1 \ldots 16$, le v.a che danno i tempi di vita di 16 diversi componenti.

$$P(96 \leq \overline{X}_{16} \leq 102) = P(95,5 \leq \overline{X}_{16} \leq 102,5) =$$
$$P\left(\frac{95,5 - 100}{20} \cdot 4 \leq Z \leq \frac{102,5 - 100}{20} \cdot 4\right) \approx \Phi\,() - \Phi\,() \approx ..... \quad \square$$