PROBABILITA' ELEMENTARE E CENNI DI STATISTICA

CAPITOLO 1

STATISTICA DESCRITTIVA

La statistica descrittiva si occupa del rapporto tra gli avvenimenti con risultati non esattamente predicibili del mondo reale e la loro rappresentazione in termini scientificamente utilizzabili. Ad esempio, quando si si fanno osservazioni o si registrano i risultati di esperimenti la rappresentazione delle informazioni ottenute in termini trasmissibili ed analizzabili richiede il prendere molte decisioni spesso altamente opinabili. Per esempio, quando si analizzano i dati relativi a migliaia di pazienti per studiare certe patologie o certe terapie, spesso questi sono scarabocchiati su fogli da medici diversi e vanno letti (e qui già ci vorrebbe l'aiuto di un medico), interpretati, trasferiti in forma unificata indovinando possibili differenze tra i diversi redattori e spesso rendendo numeriche delle valutazioni qualitative. Poi i dati vanno accuratamente trascritti in un data base, a quel punto analizzati per cercare delle indicazioni significative; una volta che ci siano delle osservazioni apparentemente sensate queste vanno giustificate e successivamente presentate.

E' chiaro che tutto questo è un processo lungo e delicato e che vi sono innumerevoli articoli e trattati che lo studiano. Noi qui ci occuperemo brevemente solo di un aspetto particolare: supponendo di avere dati già in forma numerica ci poniamo il problema di generare un grafico o alcuni valori rappresentativi.

Supponiamo quindi di avere un *campione* $x_1, \ldots, x_n \in \mathbb{R}$ e poniamo

$$x_{min} = \min\{x_i, i = 1, \ldots, n\}$$

e

$$x_{max} = \max\{x_i, i = 1, \ldots, n\}$$

così che $[x_{min}, x_{max}]$ è il range del campione.

1.1. L'istogramma

L'*istrogramma* dei dati, detto usualmente istogramma, è una rappresentazione grafica dei dati che rappresenta frequenze attraverso le aree. Descriveremo ora una modalità di realizzazione degli istogrammi, che come si vedrà prevede comunque numerose scelte arbitrarie.

Dividiamo $[\lfloor x_{min} \rfloor, \lfloor x_{max} \rfloor + 1]$ in intervallini scegliendo

$$\lfloor x_{min} \rfloor = d_0 < d_1 < \cdots < d_m = \lfloor x_{max} \rfloor + 1$$

e formando intervallini semiaperti a destra (scegliendoli semiaperti a destra si semplificano leggermente le definizioni):

$$[d_k, d_{k+1}[$$

per $k = 0, \ldots, m-1$.

Determinati gli intervallini, o separati comunque i dati in classi, per ogni intervallo o classe k si dice:

- **frequenza assoluta** $|\{i : x_i \in [d_k, d_{k+1}[\}|$,
- **frequenza relativa**$|\{i : x_i \in [d_k, d_{k+1}[\}|/n$

e **istogramma** la funzione che per $x \in [d_k, d_{k+1}[$ vale

$$I(x) = \frac{|\{i : x_i \in [d_k, d_{k+1}[\}|}{n(d_{k+1} - d_k)}.$$

Inoltre si chiama **frequenza cumulativa** o **funzione di ripartizione empirica** in x la funzione $F(x) = |\{i : x_i \leq x\}|/n$

Esercizio 1. *Con dati* $0, 1, 2, 2, 1, 1, 0, 2, 1$ *si disegnino gli istogrammi ottenuti con la scelta* $d_0 = 0, d_1 = 1, d_2 = 3$ *e con la scelta* $d_0 = 0, d_1 = 1, d_2 = 2, d_3 = 3$.

Osservazione 1. *Attraverso l'istogramma possiamo capire e mostrare dove i dati si concentrano di più di meno, anche se l'arbitrarietà della scelta degli intervalli e del comportamento agli estremi degli intervalli lascia spazio a molti fraintendimenti e possibili rappresentazioni falsate.*

Osservazione 2. *Se cominciamo ad immaginare, come faremo nella statistica inferenziale, che i dati siano realizzazioni di un qualche meccanismo casuale, ad esempio variabili aleatorie continue indipendenti, allora una delle cose che possono essere suggerite dall'istogramma è la forma della densità di tali variabili aleatorie.*

1.2. Elementi di Statistica Descrittiva

Per descrivere sinteticamente i dati si utilizzano le definizioni seguenti. Per dati univariati o comunque per ogni singolo tipo di dato:

(1) **valor medio empirico** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$;
(2) **mediana** ogni m tale $|\{i : x_i \leq m\}|/n \geq 50\%$ e $|\{i : x_i \geq m\}|/n \geq 50\%$
(3) **quantile-p** con $p \in [0, 1]$ ogni y_p tale che $|\{i : x_i \leq y_p\}|/n \geq p$ e $|\{i : x_i \geq y_p\}|/n \geq 1 - p$
(4) **1° (o 3°) quartile** il quantile-0.25 (o 0.75)
(5) **varianza empirica** $\sigma_n^2 = Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$;
(6) **SD empirica** $SD = \sigma_n = \sigma_n(x) = \sqrt{\sigma_n^2}$;
(7) **scarto interquartile** $y_{0.75} - y_{0.25}$;
(8) **campo di variazione** $x_{max} - x_{min}$;
(9) **MAD** la mediana di $|x_i - \bar{x}|$;
(10) **momento empirico k-simo** $m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$;

(11) **dati standardizzati** $z_i = \frac{x_i - \bar{x}}{\sigma_n}$;

per motivi che vedremo si utilizzano spesso σ_{n-1}^2 e σ_{n-1}, definiti come i precedenti con $n-1$ al posto di n. Si noti che la $m = y_{0.5}$ e che

LEMMA 1. *Media e mediana minimizzano rispettivamente lo scarto quadratico medio e lo scarto assoluto medio.*

DIMOSTRAZIONE. Si ha che

$$\frac{\partial}{\partial t}\sum_{i=1}^{n}(x_i - t)^2 = 2(\sum_{i=1}^{n} x_i - nt) = 0$$

implica $t = \bar{x}$, che è un punto di minimo essendo la funzione nonnegativa e divergente asintoticamente. Invece se m è una mediana

$$\sum_{i=1}^{n} |x_i - t| - \sum_{i=1}^{n} |x_i - m| = \sum_{i=1,\dots,n:x_i \le m} (t-m)$$
$$+ \sum_{i=1,\dots,n:x_i \ge t} (m-t) + \sum_{i=1,\dots,n:m<x_i<t} (t - x_i - |x_i - m|) \ge 0$$

essendo che per $x_i \ge m$ si ha $t - m \ge |x_i - m|$. □

Per coppie di dati (x_i, y_i):

(1) **diagramma di dispersione**: il grafico dei punti (x_i, y_i) in $\mathbb{R}^2$;

(2) **correlazione** si considerano: $r = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sigma_n(x)} \frac{y_i - \bar{y}}{\sigma_n(y)}$

e la regressione che ora descriviamo.

Ipotizziamo che $y_i = a + bx_i + r_i$ con r_i un errore relativamente piccolo e tale abbia media nulla per ogni valore fissato di x siano circa nulli; questo significa che detto $\bar{y}_i|x_i = \frac{\sum_{j=1,\dots,n:x_j=x_i} y_j}{\{j:x_j=x_i\}}$ sia ha $\bar{y}_i|x_i \approx a + bx_i$. Determinando a e b in modo che minimizzino lo scarto quadrato medio $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ si ottiene: minimizzando prima in a a b fissato: $a = \frac{1}{n}\sum_{i=1}^{n}(y_i - bx_i) = \bar{y} - b\bar{x}$ per quanto visto sopra sulla media empirica. Sostituendo e derivando:

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - \bar{y} + b(\bar{x} - x_i))^2 = \sum_{i=1}^{n}(\bar{x} - x_i)(y_i - \bar{y} + b(\bar{x} - x_i)) = 0$$

implica

$$b = \frac{\sum_{i=1}^{n}(\bar{x} - x_i)(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)},$$

che è un punto di minimo essendo la funzione nonnegativa e divergente asintoticamente, e quindi $a = \bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x}$. Per cui il modello lineare prevede

$$y_i = \bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x} + \frac{Cov(x,y)}{Var(x)}x_i$$

o anche

$$\frac{y_i - \bar{y}}{SD(y)} = r\frac{x_i - \bar{x}}{SD(x)}$$

che si memorizza dicendo che in variabili standardizzate il coefficiente della retta interpolatrice è r. Quindi una rappresentazione dei dati si ottiene dalla **retta di regressione** di equazione:

$$y = \bar{y} - \frac{Cov(x,y)}{Var(x)}\bar{x} + \frac{Cov(x,y)}{Var(x)}x$$

oppure in variabili standardizzate

$$y = rx.$$

Si dicono poi **residui** i valori $r_i = y_i - a - bx_i$ e, ponendo $C_{x,y} = Cov(x,y)$ e $\sigma_x^2 = Var(x) = \sigma_n(x)$, **varianza residua del modello lineare**:

$$\begin{aligned}
V_M &= \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2 \\
&= \sum_{i=1}^{n}\big((y_i - \bar{y})^2 + (\bar{y} - a - bx_i)^2 + 2(y_i - \bar{y})(\bar{y} - a - bx_i)\big) \\
&= \sum_{i=1}^{n}\left((y_i - \bar{y})^2 + (\frac{C_{x,y}}{\sigma_x^2}(\bar{x} - x_i))^2 + 2(y_i - \bar{y})(\frac{C_{x,y}}{\sigma_x^2}(\bar{x} - x_i))\right) \\
&= \sigma_y^2 + \frac{C_{x,y}^2}{\sigma_x^2} - 2\frac{C_{x,y}^2}{\sigma_x^2} = \sigma_y^2 - \frac{C_{x,y}^2}{\sigma_x^2} \leq \sigma_y^2.
\end{aligned}$$

Per cui la varianza residua del modello lineare è sempre minore della varianza originale delle y: $\sigma_y^2 - V_M$ è detta **varianza spiegata** e può essere una misura dell'efficiacia del modello lineare, per cui si definisce **coefficiente di determinazione** il valore:

$$R^2 = \frac{(\sigma_y^2 - V_M)}{\sigma_y^2} = \frac{C_{x,y}^2}{\sigma_x^2\sigma_y^2} = r^2.$$

Per cui si trova un nuovo significato di r come misura della capacità del modello lineare di predire la variabile dipendente.
Si chiama poi **regressione non lineare** l'uso della regressione lineare non tra le variabili osservate x ed y, ma tra loro funzioni. Se si sostituisce x con $g(x)$ il modello si adatta immediatamente facendo la regressione $y = a + bg(x)$. Mentre è più interessante il caso in cui si sostituisce y con $f(y)$, per una funzione f invertibile. Si può allora considerare la regressione $f(y) = a + bx$ e stimare a e b, ottenendo la predizione $y_i = f^{-1}(a + bx_i)$ e la varianza del modello $V_M = \sum_{i=1}^{n}(y_i - f^{-1}(a + bx_i))^2$. Si noti che ora non è detto che $V_M \leq \sigma_y^2$, ossia che $R^2 \geq 0$, anche se naturalmente un valore negativo vuol dire che il modello è peggio di predire y_i con $\bar{y}$.

CAPITOLO 2

STATISTICA INFERENZIALE

Immaginiamo ora che i dati siano realizzazioni di un qualche meccanismo casuale descrivibile nell'ambito del calcolo delle probabilità. Consideriamo un caso semplice, in cui i dati del campione che stiamo considerando siano realizzazioni di variabili aleatorie indipendenti ed identicamente distribuite, discrete o continue a seconda che i dati osservati siano numeri interi o multipli di una costante o meno.

Definizione 1. *Un modello statistico è una quaterna*

$$(\Omega, \mathcal{A}, \Theta, P_\theta)$$

con Ω un insieme, $\mathcal{A}$ una σ-algebra su Ω, Θ un insieme di parametri e per ogni $\theta \in \Theta$, P_θ una probabilità su $(\Omega, \mathcal{A})$. I dati x_i si considerano realizzazioni di variabili aleatorie X_i i.i.d. con distribuzione P_θ.

Esempio 1. *Nel caso binomiale si può prendere come modello statistico $(\{0,1\}, \mathcal{P}(\{0,1\}), [0,1], P_p)$ con $p \in [0,1]$ e $P_p(1) = 1 - P_p(0) = p$.*

Esempio 2. *Nel caso normale o gaussiano si può prendere come modello statistico $(\mathbb{R}, \mathcal{B}, \mathbb{R} \times \mathbb{R}^+, \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \lambda)$ con $\mathcal{B}$ σ-algebra di Borel in $\mathbb{R}$ e λ misura di Lebesgue.*

Esempio 3. *Possiamo immaginare che i dati dell'esempio 1 di cui sono stati fatti gli istogrammi nel paragrafo precedente siano realizzazioni di variabili i.i.d. $B(2,p)$, ossia ognuno dei dati corrisponda al numero di successi in 2 prove indipendenti ognuna con probabilità di successo p; oppure che siano $B(r,p)$. In tal caso il modello statistico sarà $(\{0,1,\ldots,r\}, \mathcal{P}(\{0,1,\ldots,r\}), [0,1], P_p)$ con $p \in [0,1]$ e $P_p \sim B(r,p)$.*

2.1. Stima di parametri

Una volta decisa la distribuzione delle variabili aleatorie ossia il modello statistico rimangono dei parametri da determinare, come ad esempio la probabilità di successo p nell'ultimo esempio, uno o più parametri in Θ.

Definizione 2. *Dato un modello statistico $(\Omega, \mathcal{A}, \Theta, P_\theta)$ uno stimatore dei parametri è una funzione*

$$\hat{\theta} : \Omega^n \to \Theta$$

che a campioni $x_1, \ldots, x_n$ *associa* $\hat{\theta}(x_1, \ldots, x_n) \in \Theta$.

Sono stati sviluppati vari metodi per identificare delle funzioni che possano servire come stimatori, e qui ne vedremo due, quello dei momenti e quello di massima verosimiglianza. Sono poi stati sviluppati anche metodi di valutazione della efficacia o bontà dei vari stimatori, e di questi vedremo solo brevemente il concetto di distorsione.

2.2. Il metodo dei momenti

Si considerano allora i momenti k-simi $E(X^k) = E_{\theta_1}(X^k)$ della distribuzione degli X_i; dalla legge debole dei grandi numeri applicata ad X^k, se $E(X^{2k}) < \infty$ allora per ogni $\epsilon > 0$

$$\lim_n P(|\frac{1}{n}\sum_{i=1}^n X_i^k - E(X_1^k)| > \epsilon) = 0.$$

Per cui una stima dei parametri si ottiene ugualiando i momenti k-simi ai momenti k-simi empirici corrispondenti, formando un numero di equazioni sufficiente a permettere di determinare la stima dei parametri. Quando questo funziona si pone $\hat{\theta}$ uguale alle soluzioni del sistema composto di tali equazioni $m_k = E(X^k)$.

Esempio 4. *Con i dati dell'esercizio 1 si può stimare la probabilità* p *della binomiale con il metodo dei momenti. Ora se* $X \sim B(2, p)$ *allora* $E(X) = 2p$ *ed* $m_1 = \bar{x} = 10/9$, *per cui* $\hat{p} = 10/18$.

Esempio 5. *Se invece interpretiamo i dati* $-2, 3, 1$ *come realizzazioni di variabili aleatorie i.i.d. continue* $X_i \sim U([-a, a])$ *allora* $f_{X_i}(x) = \frac{1}{2a}\mathbb{I}_{[-a,a]}(x)$ *e* $E(X_i) = 0$ *per simmetria e*

$$E(X_i^2) = \frac{1}{2a}\int_{-a}^{a} x^2 dx = \frac{a^2}{3}$$

per cui $\hat{a} = \sqrt{14} \approx 3.74$. *Se* $\hat{a}$ *fosse risultato minore di* x_{max} *ne avremmo concluso che o il metodo di stima dei parametri non era adeguato o le assunzioni fatte non costituivano un buon modello per i dati.*

Vediamo ora le principali distribuzioni, i loro parametri e la relazione con alcuni momenti idonei alla stima.

2.3. Stima di parametri: il metodo della massima verosimiglianza

Un altro metodo di stima puntuale dei parametri si basa sul determinare i valori dei parametri che massimizzano la probabilità o la densità di probabilità delle osservazioni: tale probabilità è detta verosimiglianza.

Distribuzione	Densità o funzione di distribuzione	parametri	relazione con i momenti
Binomiale $B(n,p)$	$P_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, \dots, n$	p	$m_1 = np$
Geometrica $Geom(p)$	$P_p(k) = p(1-p)^{k-1}$, $k = 0, \dots$	p	$m_1 = 1/p$
Poisson $Poiss(\lambda)$	$P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$, $k = 0, \dots$	λ	$m_1 = \lambda$
Esponenziale $Exp(\lambda)$	$f_\lambda(t) = \lambda e^{-\lambda t}$, $t \geq 0$	λ	$m_1 = 1/\lambda$
Gamma $\Gamma(k,\lambda)$	$f_{k,\lambda}(t) = \frac{\lambda(\lambda t)^{k-1}}{\Gamma(k)} e^{-\lambda t}$, $t \geq 0$	k, λ	$m_1 = k/\lambda$ $m_2 = \frac{k(k+1)}{\lambda^2}$
Gaussiana $N(\mu,\sigma)$	$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $t \in \mathbb{R}$	μ, σ	$m_1 = \mu$ $m_2 = \sigma^2 + \mu^2$
Lognormale $LogN(\lambda,\zeta)$	$f_{\lambda,\zeta}(t) = \frac{1}{\sqrt{2\pi}\zeta t} e^{-\frac{(\ln t - \lambda)^2}{2\zeta^2}}$, $t \geq 0$	λ, ζ	$m_1 = e^{\lambda + \zeta^2/2}$ $m_2 = e^{2(\lambda+\zeta^2)}$
Uniforme $U(a,b)$	$f_{U(a,b)}(t) = \frac{1}{b-a}$, $t \in [a,b]$	a, b	$m_1 = \frac{a+b}{2}$ $m_2 = \frac{(a+b)^2}{3}$
Beta $B(q,r)$	$f_{B(q,r)}(x)$ $= \frac{\Gamma(q+r)}{\Gamma(q)\Gamma(r)} x^{q-1}(1-x)^{r-1}$, $t \geq 0$	q, r	$m_1 = \frac{q}{q+r}$ $m_2 = \frac{(q+1)q}{(q+r+1)(q+r)}$

TABELLA 1. Principali distribuzioni, parametri e momenti.

DEFINIZIONE 3. *Date osservazioni $x_1, \dots, x_n$ ed un modello statistico basato su una distribuzione discreta P_θ o continua con densità $f_{(X_1,\dots,X_n),\theta}$ con parametri $\theta \in \Theta$ si dice* **verosimiglianza** *il valore:*

$$L_\theta(x_1, \dots, x_n) = P_\theta(X_1 = x_1, \dots, X_n = x_n)$$

nel caso discreto o

$$L_\theta(x_1, \dots, x_n) = f_{(X_1,\dots,X_n),\theta}(x_1, \dots, x_n)$$

nel caso continuo.

ESEMPIO 6. *Se le X_i sono ipotizzate i.i.d. continue con densità $f_{\theta_1,\dots,\theta_k}$ allora $L_{\theta_1,\dots,\theta_k}(x_1, \dots, x_n) = \prod_{i=1}^n f_{\theta_1,\dots,\theta_k}(x_i)$.*

DEFINIZIONE 4. *Date osservazioni $x_1, \dots, x_n$ si dice stima di massima versomiglianza dei parametri $\theta \in \Theta$ l'argomento di $max_\theta L_\theta(x_1, \dots, x_n)$.*

Poiché la verosimiglianza è spesso un prodotto si utilizza spesso il logaritmo, per cui in generale la stima di massima verosimiglianza si

ottiene risolvendo equazioni del tipo

$$\frac{\partial}{\partial \theta_j} \log L_\theta(x_1, \ldots, x_n) = 0$$

dove $\theta = (\theta_1, \ldots, \theta_l) \in \Theta$.

Esempio 7. *Determinare nel caso di distribuzioni esponentiali* (λ) *la stima di massima verosimiglianza di* λ *e confrontare con la stima del metodo dei momenti.*

Esempio 8. *Determinare nel caso di distribuzioni lognormali* (λ, ζ) *la stima di massima verosimiglianza di* λ *e* ζ *e confrontare con la stima del metodo dei momenti.*

2.4. Distorsione di stimatori

Un metodo di valutazione di uno stimatore $\hat{\theta}(x_1, \ldots, x_n)$ in un modello statistico $(\Omega, \mathcal{A}, \Theta, P_\theta)$ è quello di considerare $\hat{\theta}$ come funzione delle variabili aleatorie X_i e prendere il valore atteso:

Definizione 5. *Uno stimatore* $\hat{\theta}$ *si dice non distorto se per ogni* $\theta \in \Theta$

$$E_\theta(\hat{\theta}(X_1, \ldots, X_n)) = \theta.$$

Esempio 9. *Se* $E_\theta(X_1)$ *è un parametro da stimare, allora per la linearità del valore atteso lo stimatore* $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ *è non distorto, infatti per ogni* $\theta \in \Theta$

$$E_\theta(\frac{1}{n}\sum_{i=1}^n X_i) = E_\theta(X_1).$$

D'altra parte, se $Var(X_1)$ è un parametro, allora per cercare uno stimatore non distorto utilizziamo il lemma seguente.

Lemma 2. *Se* $E(X_1^2) < \infty$

$$E(\sum_{i=1}^n (X_i - (\frac{1}{n}\sum_{j=1}^n X_j))^2) = (n-1)Var(X_1)$$

Dimostrazione. Poniamo $Y_i = X_i - E(X_i)$, allora $E(Y_i) = 0$ e $E(Y_iY_j) = 0$ se $i \neq j$. Inoltre

$$\begin{aligned}
E(\sum_{i=1}^n (X_i - (\frac{1}{n}\sum_{j=1}^n X_j))^2) &= E(\sum_{i=1}^n (Y_i - (\frac{1}{n}\sum_{j=1}^n Y_j))^2) \\
&= \sum_{i=1}^n (E(Y_i^2)(1-1/n)^2 + \frac{1}{n^2}\sum_{j\neq i} E(Y_j^2)) \\
&= \sum_{i=1}^n (E(Y_i^2)((\frac{n-1}{n})^2 + \frac{n-1}{n}^2 \\
&= (n-1)Var(X_1)
\end{aligned}$$

□

Per cui è $\hat{\sigma}^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ lo stimatore non distorto di $Var(X_1)$.

2.5. Test di ipotesi

Ci poniamo ora in una prospettiva diversa. Immaginiamo di avere un modello teorico per un esperimento con risultati casuali e dei dati raccolti dopo la realizzazione dell'esperimento. Ci potremmo chiedere se i dati confermino o meno il modello teorico, ma da un punto di vista scientifico (si veda per esempio K. Popper, Logica della scoperta scientifica, Einaudi, Torino, 1970) è più sensato chiedersi all'opposto se i dati contraddicano o meno l'ipotesi teorica. Questo è il senso del test di ipotesi.
In generale, un modello teorico contiene molte ipotesi, anche non esplicitate: per esempio, quando discutiamo del risultato del lancio di dadi stiamo assumendo che in ogni esperimento riusciamo a leggere il risultato e che i risultati stessi siano riportati in modo fedele e non per esempio alterati dopo la realizzazione dell'esperimento, ecc. ecc.
Qui presenteremo una versione semplificata di questa problematica, in cui si testa semplicemente il valore dei parametri di una distribuzione o, tutt'al più , la scelta della distribuzione stessa. Per cui si assume un modello statistico $(\Omega, \mathcal{A}, \Theta, P_\theta)$ e l'ipotesi riguarderà l'assunzione che il valore reale del parametro è $\theta_0 \in \Theta$.
Lo schema che presentiamo è detto test di ipotesi e si realizza come segue:

- (0) Si fissa un **livello del test**, detto α, tipo $\alpha = 0.05$ oppure $\alpha = 0.01$.
- (1) Si descrive l'ipotesi che si vuole testare, detta **Ipotesi Nulla** H_0, ed un'**ipotesi alternativa**, con cui l'ipotesi nulla viene confrontata, indicata con H_1. Per semplificare assumiamo che le due ipotesi identifichino una famiglia di distribuzioni ed un valore $\theta_0 \in \Theta$. Si noti che, in questo caso, H_1 sarà semplicemente il complemento di H_0.
- (2) Si determina una certa funzione $Q = Q(x_1, \ldots, x_n; \theta) \in \mathbb{R}$ dei dati e dei parametri con la proprietà che per ogni distribuzione identificata dall'ipotesi nulla e per ogni $\theta \in \Theta_0 \quad Q(X_1, \ldots, X_n; \theta)$, con X_i variabili aleatorie i.i.d. distribuite secondo la distribuzione scelta, ha una distribuzione nota, almeno approssimativamente. Si determina poi per ogni $\theta \in \Theta_0$ la **statistica del test** $a(\theta) = Q(x_1, \ldots, x_n; \theta)$. La statistica del test viene indicata con lettere tipo z, t, χ^2 ecc.
- (3) Su questa base siamo ora in grado di calcolare la probabilità o la densità di probabilità dell'evento $Q = a$.

Tuttavia, abbiamo visto che questa (densità di) probabilità non è utile ai fini di valutare H_0. Si cercano quindi degli insiemi $I(a) \subset \mathbb{R}$ (generalmente intervalli o unioni di intervalli) con la proprietà che
- $P(Q \in I(a))$ è monotono in a;
- $P(Q \in I(a)) >> 0$ per certi valori di a;
- per θ_0, $P(Q \in I(a))$ tende a 0 quando a tende ad un estremo del condominio $S_{Q(X_1,\ldots,X_n;\theta)}$ di Q.

e si definisce **valore p del test** o p-value il valore

$$p = P_{\theta_0}(Q(X_1, \ldots, X_n; \theta) \in I(a(x_i, \theta_0))$$

dove $a(x_i, \theta_0)$ è la statistica del test.

- (4) Si confrontano p e α. Se $p < \alpha$ si rigetta H_0 e si assume H_1 ed altrimenti non si rigetta.
- (3') Si può ottenere una procedura alternativa al calcolo del valore p invertendo la parte finale del ragionamento precedente. Una volta fissato il livello del test α ogni possibile realizzazione dei dati porta ad un valore di a tale che H_0 si rigetta o meno, e quindi si può definire la **regione di rigetto**

$$D = D(\alpha) = \{(x_1, \ldots, x_n) : \ H_0 \text{ si rigetta al livello } \alpha\}$$

- (4') Per determinare se rigettare o meno H_0 si verifica semplicemente se per i dati osservati vale $(x_1, \ldots, x_n) \in D(\alpha)$

Si osservi che in un test si può
- Erroneamente rigettare H_0, detto errore di prima specie, questo evento ha probabilità α.
- Erroneamente accettare H_0, detto errore di seconda specie.

Per ridurre la probabilità di errore di prima specie occorre ridurre α e per controllare l'errore di seconda specie si deve aumentare α.
Vediamo ora vari esempi di test.

Osservazione 3. *Qualche testo costruisce il test di ipotesi partendo senza definire preventivamente il livello del test e quindi direttamente dalla regione di rigetto D. A quel punto il livello del test dipende da D ed è definita come*

$$\alpha(D) = sup_{\theta \in \Theta_0} P_\theta((X_1, \ldots, X_n) \in D).$$

Test per la media di osservazioni gaussiane a varianza nota. Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. $N(\mu, \sigma^2)$ con σ^2 nota e vogliamo testare il valore di μ che assumiamo essere un certo valore μ_0 mentre i dati sembrano suggerire un valore più piccolo; vogliamo limitare l'errore di prima specie. Allora ha senso porre

- $\alpha = 0.1$

- H_0: $\mu = \mu_0$ e H_1: $\mu < \mu_0$;
- $z = \frac{\sum x_i - \mu_0}{\sqrt{n}\sigma}$ con $Z = \frac{\sum X_i - \mu_0}{\sqrt{n}\sigma} \sim N(0,1)$;
- $p = P(Z \leq z)$ e confrontare con 0.01;
- oppure $D = \{x_i : \frac{\sum x_i - \mu_0}{\sqrt{n}\sigma} \leq -\phi_{0.99}\}$.

Test per la probabilità di successo a due code. Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. $N(\mu, \sigma^2)B(1,p)$ e vogliamo testare il valore di p che assumiamo essere un certo valore p_0 mentre i dati potrebbero suggerire un valore diverso; vogliamo limitare l'errore di prima specie. Allora ha senso porre

- $\alpha = 0.1$
- H_0: $p = p_0$ e H_1: $p \neq p_0$;
- $z = \frac{\sum x_i - p_0}{\sqrt{np_0(1-p_0)}}$ con $Z = \frac{\sum X_i - p_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1)$ approssimativamente;
- $p = P(|Z| \geq |z|)$ e confrontare con 0.01;
- oppure $D = \{x_i : |\frac{\sum x_i - p_0}{\sqrt{np_0(1-p_0)}}| \geq \phi_{0.995}\}$.

Test a una coda per il confronto delle medie in due popolazioni. Assumiamo che vi siano due popolazioni con dati x_i, $i = 1, \ldots, n_1$ e y_j, $j = 1, \ldots, n_2$ generati da variabili aleatorie X_i ed Y_j, rispettivamente, entrambe i.i.d. ed independenti tra loro. Vogliamo testare se $E(X) = E(Y)$ mentre i dati suggeriscono che sia minore. Il test si realizza nel modo seguente:

- $\alpha = 0.1$
- H_0: $E(X) = E(Y)$ e H_1: $E(X) \leq E(Y)$;
- $t = \frac{(\sum x_i - \sum y_j)\sqrt{n_1+n_2-2}}{\sqrt{((n_1-1)S_X^2 + (n_2-1)S_Y^2)(1/n_1+1/n_2)}}$ con $T = \frac{(\sum X_i - \sum X_j)\sqrt{n_1+n_2-2}}{\sqrt{((n_1-1)S_X^2 + (n_2-1)S_Y^2)(1/n_1+1/n_2)}} \sim Stud(n_1+n_2-2)$ approssimativamente;
- $p = P(T \leq t)$ e confrontare con 0.01;
- oppure $D = \{x_i : t \leq -\phi_{0.99}\}$.

Tutti questi erano z test, nel senso che la distribuzione della statistica era normale. In altri casi questo non succede: nel prossimo esempio la distribuzione è la t di Student.

Test per la media di osservazioni gaussiane a varianza non nota. Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. $N(\mu, \sigma^2)$ con σ^2 non nota e vogliamo testare il valore di μ che assumiamo essere un certo valore μ_0 mentre i dati sembrano suggerire un valore più piccolo; vogliamo limitare l'errore di seconda specie. Allora ha senso porre

- $\alpha = 0.5$
- H_0: $\mu = \mu_0$ e H_1: $\mu < \mu_0$;

- $t = \frac{\sum x_i - \mu_0}{\sqrt{n}\sigma}$ con $T = \frac{\sum X_i - \mu_0}{\sqrt{n}\sigma} \sim Stud(n-1)$;
- $p = P(T \leq t)$ e confrontare con 0.05;
- oppure $D = \{x_i : \frac{\sum x_i - \mu_0}{\sqrt{n}\sigma} \leq -t_{n-1,0.95}\}$ con $t_{n-1,x}$ quantile di ordine x della distribuzione di Student con $(n-1)$ gradi di libertà .

Test per la varianza di osservazioni gaussiane . Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. $N(\mu, \sigma^2)$ e vogliamo testare il valore di σ^2 che assumiamo essere un certo valore σ_0 mentre i dati sembrano suggerire un valore più grande; vogliamo limitare l'errore di seconda specie. Allora ha senso porre

- $\alpha = 0.5$
- H_0: $\sigma = \sigma_0$ e H_1: $\sigma > \sigma_0$;
- $\chi^2 = \frac{(n-1)\sum(x_i - \bar{x})^2}{\sigma^2}$ con $K = \frac{(n-1)\sum(X_i - \bar{x})^2}{\sigma^2} \sim \chi^2(n-1)$;
- $p = P(K \geq \chi^2)$ e confrontare con 0.05;
- oppure $D = \{x_i : \frac{(n-1)\sum(x_i - \bar{x})^2}{\sigma^2} \geq \chi^2_{n-1,0.95}\}$ con $\chi^2_{n-1,x}$ quantile di ordine x della distribuzione del Chi-quadro con $n-1$ gradi di libertà o $\Gamma((n-1)/2, 1/2)$.

Test del chi-quadro per l'adattamento di una variabile aleatoria ad una distribuzione finita. Assumiamo che le variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. con una data distribuzione finita
$P = P(\theta_1, \ldots, \theta_r) = (p_1(\theta_1, \ldots, \theta_r), \ldots, p_k(\theta_1, \ldots, \theta_r))$ nota a meno di r parametri incogniti sullo spazio $S = \{s_1, \ldots, s_k\}$. Vogliamo testare se la distribuzione sia adeguata a rappresentare i dati limitando l'errore di seconda specie. Poniamo n_j il numero di osservazioni di j nel campione. Allora ha senso porre

- $\alpha = 0.5$
- H_0: la distribuzione è $P(\theta_1, \ldots, \theta_r)$ per qualche valore dei parametri e H_1: la negazione di H_0;
- $\chi^2 = \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j}$ con $K = \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \sim \chi^2(k-r-1)$ approssimativamente, avendo posto Y_j variabile aleatoria che dà il numero di osservazioni di j in un campione di dimensione n;
- $p = P(K \geq \chi^2)$ e confrontare con 0.05;
- oppure $D = \{x_i : \sum_{j=1}^{k} \frac{(n_j - np_j)^2}{np_j} \geq \chi^2_{k-r-1,0.95}\}$ con $\chi^2_{m,x}$ quantile di ordine x della distribuzione del Chi-quadro con m gradi di libertà o $\Gamma(m/2, 1/2)$.

Test del chi-quadro per l'adattamento di una variabile aleatoria ad una distribuzione continua. Se si assume che le variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. con una data distribuzione continua con densità f_X allora ci si può

rifare al caso discreto dividendo in k classi I_j, $j = 1, \ldots, k$ e prendendo $p_j(\theta_1, \ldots, \theta_r) = P(X \in I_j) = \int_{I_j} f_X(t)dt$ sempre a meno di r parametri incogniti. Ci si può rifare quindi al caso precedente, sempre utilizzando la distribuzione χ^2 con $k - r + 1$ gradi di libertà . Si noti che i gradi di libertà dipendono dal numero di intervalli scelti, ma naturalmente la soglia determinata dal χ^2 cresce al crescere del numero di intervallini (quindi si ottiene un fitting più difficile confrontato con una soglia maggiore).

Test di Kolmogorov-Smirnov per l'adattamento di variabili aleatorie qualsiasi. Assumiamo che la variabili aleatorie $X_1, \ldots, X_n$ che hanno generato i dati siano i.i.d. con una data funzione di distribuzione F_X. Vogliamo testare se la distribuzione sia adeguata a rappresentare i dati limitando l'errore di seconda specie. Sia F_n la funzione di distribuzione empirica (che vale k/n in $[x_{(k)}, x_{(k+1)}]$) Allora ha senso porre

- $\alpha = 0.5$
- H_0: la distribuzione è $P(\theta_1, \ldots, \theta_r)$ per qualche valore dei parametri e H_1: la negazione di H_0;
- $d_n = \max_t |F_n(t) - F_X(t)|$ con D_n una distribuzione tabulata:
- $p = P(D_n \geq d_n)$ e confrontare con 0.05;
- oppure $D = \{x_i : \max_t |F_n(t) - F_X(t)| \geq D_n(0.95)\}$ con $D_n(x)$ quantile di ordine x.

Come esempi numerici consideriamo lo z-test ad una coda da cui eravamo partiti.

Esempio 10. *Ritornando all'esempio 1, se esce testa 459 volte su 1000 lanci di una moneta, allora si esegue un test di ipotesi relativamente alla correttezza della moneta ponendo:*

- $\alpha = 0.2$
- H_0*:* $p = 1/2$ *e* H_1*:* $p \leq 1/2$*;*
- $z = \frac{\sum_{i=1}^{1000} x_i - 1/2}{\sqrt{1000}/2} \approx -2.59$*;*
- *dal TCL si ha* $p = P(Z \leq -2.59) \approx \int_{-\infty}^{-2.59} \frac{\sqrt{2\pi}}{e^{-\frac{x^2}{2}}} \approx 0.0048 < 0.01$ *quindi si rigetta* H_0*, ossia si conclude che l'insieme delle ipotesi fatte non è consistente con i dati osservati.*

Osservazione 4. *Avevamo già visto che questo numero di teste non era accettabile; la terminlogia del test di ipotesi ha permesso di chiarire bene la procedura e le conclusioni.*

Ora possiamo concludere l'analisi dell'esempio 2 attraverso uno z-test approssimato.

Esempio 11. *Si ha*

- $\alpha = 0.2$

- H_0*: I tempi di vita sono realizzazioni di variabili aleatorie i.i.d. esponenziali di parametro* λ *tale che il tempo medio di funzionamento previsto corrisponde al valor medio della distribuzione, ossia* $\frac{1}{\lambda} = 1000$*, quindi* $\lambda = \frac{1}{1000}$ *e* H_1*:* $\lambda \leq \frac{1}{1000}$*;*
- $z = \frac{\sum_{i=1}^{1000} x_i - 100000}{\sqrt{100} \cdot 1000} = \frac{90200 - 100 \cdot 1000}{\sqrt{100} \cdot 1000} \approx -0.98$*;*
- *dal TCL si ha* $p \approx \int_{-\infty}^{-0,98} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \approx 16,35\% > 0.01$ *quindi non si rigetta* H_0*, ossia si conclude che i dati non contraddicono l'insieme delle ipotesi fatte.*

Quindi un tempo totale di funzionamento di 90000 ore era perfettamente in linea con la durata media prevista.

Esercizio 2. *Verificare che con una soglia dell'*1% *la regione di rigetto D corrisponde ad un funzionamento totale minore di 76700 ore e quindi solo sotto tale soglia si dovrebbe considerare falsa la dichiarazione sulla durata media.*

Esercizio 3. *Perchè questa volta si tende a considerare inappropriato un risultato ancora perfettamente adeguato?*