1 Test per la media di una gaussiana

Due persone esaminano una grandezza aleatoria X, che supponiamo gaussiana, di varianza nota σ (altrimenti la teoria si generalizza usando la t di Student). La prima persona afferma che la media di X è μ_0. La seconda possiede un campione sperimentale $X_1, ..., X_n$ e vuole decidere se, sulla base di tale campione, accetta o rifiuta l'affermazione dell'altra persona. L'ipotesi (detta *ipotesi nulla*) è che la media sia μ_0 e si sottopone ad un *test* questa ipotesi. In parole povere, il test consiste semplicemente nel vedere se il valore $\overline{X}$ del campione cade nell'intervallo di confidenza

$$\left[\mu_0 - \frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}, \mu_0 + \frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}\right]$$

per un valore scelto a priori (e piccolo) di α. Infatti, se la media vera è proprio μ_0, $\overline{X}$ deve cadere in tale intervallo con grande probabilità (pari a $1-\alpha$). Se scopriamo che $\overline{X}$ non cade in tale intervallo, rifiutiamo l'ipotesi nulla. Se invece $\overline{X}$ cade in tale intervallo, affermiamo di non aver trovato alcuna contraddizione tra l'ipotesi nulla ed il campione.

Per uniformità tra varie procedure di test, si preferisce vedere se la grandezza

$$Z = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$$

cade nell'intervallo $\left[-q_{1-\alpha/2}, q_{1-\alpha/2}\right]$, cosa equivalente alla precedente. Si confronta quindi $|Z|$ con $q_{1-\alpha/2}$. Tutti i test vengono alla fine formulati come confronto con un quantile. Se risulta $|Z| > q_{1-\alpha/2}$ il valore non cade nell'intervallo giusto e quindi si rifiuta l'ipotesi. Si usa dire che *il test è significativo*.

L'esito del test (significativo o meno) dipende dal valore prefissato di α, che viene detta *significatività*. Più α viene scelto piccolo, più è difficile che risulti $|Z| > q_{1-\alpha/2}$, e quindi se questo accade rifiutiamo l'ipotsi nulla con maggior sicurezza. Per queste ragioni, è interessante sapere qual'è il più piccolo valore di α per cui avremmo rifiutato l'ipotesi nulla. Questo valore viene detto *p-value*, valore p. Se il p-value è molto piccolo, significa che avremmo rifiutato l'ipotesi nulla con un valore molto piccolo di α, cosa che era difficile da ottenere per puro caso, quindi siamo molto sicuri della nostra conclusione (il rifiuto). Siccome è uso prendere $\alpha = 0.05$, quando il SW fornisce il p-value e questo è inferiore a 0.05, tendiamo a rifiutare l'ipotesi nulla. Altrimenti affermiamo che non c'è contraddizione tra ipotesi e dati.

Il valore p si calcola nel seguente modo. Indichiamo con $\widehat{Z}$ il valore sperimentale di Z ottenuto dal campione. Abbiamo detto che ogni scelta α tale che $\left|\widehat{Z}\right| > q_{1-\alpha/2}$ conduce al rifiuto dell'ipotesi, mentre ogni scelta tale che $\left|\widehat{Z}\right| < q_{1-\alpha/2}$ non permette di rifiutare l'ipotesi. Quindi il valore p è il punto di separazione tra questi due regimi e si trova uguagliando

$$\left|\widehat{Z}\right| = q_{1-p/2}$$

intesa come equazione nell'incognita p. Si riconosce che questa equazione equivale a $P\left(Z < \left|\widehat{Z}\right|\right) = 1 - p/2$ ed infine

$$p = P\left(|Z| > \left|\widehat{Z}\right|\right).$$

Nel caso specifico questa equazione si risolve esplicitamente, ma in ogni caso queste due equazioni equivalenti sono il modo di trovare il valore p: si cerca il valore di α per cui c'è l'uguaglianza nel confronto tra grandezza sperimentale e quantile, oppure equivalentemente si calcola la probabilità che la variabile con cui si esegue il test (qui la $|Z|$) assuma valori più estremi di quello sperimentale. Perseguiremo nel prossimo paragrafo questa seconda visione.

1.1 Riformulazione

Un'altro modo di capire la strategia precedente, utile per uniformità con altre strategie di test, è la seguente.

Se il campione proviene da una $N(\mu_0, \sigma^2)$, la grandezza $Z = \frac{\overline{X}-\mu_0}{\sigma}\sqrt{n}$ ha una certa distribuzione, in questo caso una $N(0,1)$. Se il campione sperimentale ha dato il valore $\widehat{Z}$, che probabilità c'è che dalla distribuzione di Z esca un simile valore? Se è molto improbabile, rifiutiamo l'ipotesi (che il campione provenga da una $N(\mu_0, \sigma^2)$).

Purtroppo questa idea così chiara va raffinata: la probabilità che Z valga $\widehat{Z}$ è zero. Quindi ogni valore sperimentale risulterebbe così improbabile, secondo questa logica, da far rifiutare l'ipotesi. Tutto questo sembra assurdo, ma è identico all'usuale paradosso: prefissato un valore, questo esce da una distribuzione continua con probabilità zero, ma effettuato un esperimento, questo ha dato un valore (che sarebbe uscito con probabilità zero). Uno dei modi di uscire dal paradosso è: sapendo che è uscito quel valore, che

probabilità c'è che esca? Uno. Quindi basta poco per ribaltare le cose. Avvertiamo però che questa è una spiegazione semplicistica ed incompleta: solo con concetto di potenza si capisce tutto.

Torniamo alla logica descritta poco sopra, inopportuna, e cerchiamone una versione utile. Invece di calcolare la probabilità che Z valga $\widehat{Z}$, che è zero, calcoliamo la probabilità che Z assuma valori ancora più estremi del valore sperimentale $\widehat{Z}$. Nella visione bilatera perseguita fino ad ora (intervalli di confidenza bilateri), calcoliamo

$$P\left(|Z| > \left|\widehat{Z}\right|\right)$$

e prendiamo questo come indice della verosimiglianza dell'ipotesi. Ecco quindi un modo ragionevole, anche se suscettibile di critiche, di arrivare al valore p, alternativo a quello del paragrafo precedente. Questo modo è in un certo senso più diretto e si adatta meglio alle generalizzazioni. Abbiamo già evidenziato alcune critiche, tra le quali esplicitiamo quella della scelta bilatero-unilatero: potevano calcolare

$$P\left(Z > \widehat{Z}\right)$$

e questa ci sarebbe sembrata altrettanto sensata, forswe ancor di più. Su che base logica dobbiamo effetturare queste scelte? Solo con il concetto di *potenza* si può rispondere in modo completo e non vago.

1.2 Altri test

Con R abbiamo eseguito il testi di Kolmogorov-Smirnov, tramite il comando *ks.test(x,"pnorm",mu,s)*. Esso fornisce un p-value. Ecco il suo significato: l'ipotesi nulla è che il campione x derivi da una gaussiana di media *mu* e deviazione s. Se il p-value è inferiore a 0.05, rifiutiamo questa ipotesi e quindi riteniamo che il campione non provenga da una gaussiana. Il p-value nell'esempio della terza lezione è 0.0778, quindi non possiamo rifiutare l'ipotesi di gaussianità. Certo è piuttosto vicino a 0.05, quindi non siamo lontani dal rifiuto. Sembra strano dal Q-Q plot, ma si deve considerare che il campione è numerosissimo e questo aumenta la possibilità che il test riconosca una deviazione dall'ipotesi. Torniamo su KS sotto, con più precisione.

Altri test si trovano leggendo *summary(reg)* nell'ottava lezione. Nell'analisi dei risultati di una regressione multipla vengono automaticamente eseguiti

vari test. Uno di questi consiste nel giudicare l'importanza di ogni singolo predittore, inclusa l'intercetta. L'ipotesi nulla è che il modello senza quel predittore dia risultati - come bontà del fit - paragonabili al modello col predittore. Un basso valore p indica allora che questa ipotesi è falsa, cioè quel predittore è utile. Nella risposta del SW non si legge la solita frase "p-value" ma

$$\Pr\left(> |t|\right).$$

Questo è un simbolo sintetico analogo a $P\left(|Z| > \left|\widehat{Z}\right|\right)$ visto sopra. E' però riferito ad una certa distribuzione, la t di Student, che interviene in quello specifico test. Insomma, come abbiamo capito sopra, $\Pr\left(> |t|\right)$ è il valore p.

Più in basso tra le risposte del SW, si parla della F-statistics (un'altra particolare grandezza come Z) e si esegue il seguente test: si giudica il modello nel suo complesso, cioè si assume come ipotesi nulla che il modello nullo, cioè quello con con $\alpha = \beta = 0$, dia risultati simili al modello trovato. In altre parole, i valori predetti $\widehat{y}_i$ dal modello trovato o quelli che avrebbe prodotto un modello con $\alpha = \beta = 0$ erano all'incirca ugualmente vicini ai valori y_i dei dati, in media quadratica. Si vede un p-value enormemente piccolo: 4.058e-08. Trovando un p-value pari a 4.058e-08 rifiutiamo senza esitazione questa ipotesi, avvalorando enormemente la validità del modello. Più precisamente, non viene avvalorato il modello nel senso di dire che quel particolare modello è veritiero, ma solo nel senso che è più veritiero del niente (cioè del modello a coefficienti nulli). In altre parole, è fuor di dubbio che questo modello spieghi qualcosa della struttura dei dati. Che poi sia il modello giusto, è un'altro problema.

2 Il test di Kolmogorov-Smirnov

Viene usato per capire se un campione proviene da una certa distribuzione. Ad esempio, dato un campione, ci chiediamo se sia ragionevolmente gaussiano. Allora si prende la gaussiana avente media e deviazione pari a quelle sperimentali del campione, e si esegue il test di KS tra il campione e quella gaussiana.

In astratto, data una v.a. X con c.d.f. $F(x)$ ed un campione $x_1, ..., x_n$, ci chiediamo se $x_1, ..., x_n$ provienga ragionevolmente da X oppure no.

Ordiniamo il campione: $x'_1, ..., x'_n$. Calcoliamo la $\widehat{F}(x)$:

$$\widehat{F}(x) = \sum_{k=1}^{n-1} \frac{k}{n} 1_{x \in [x_k, x_{k+1})} + 1_{x \geq x_n}$$

Poniamo

$$D_n := \sup_{-\infty < x < \infty} \left| F(x) - \widehat{F}(x) \right|.$$

Vale

$$D_n = \sup_{k=1,...,n} \left| F(x_k) - \frac{k-1}{n} \right| \vee |1 - F(x_n)|.$$

Questa è la grandezza che il SW calcola per eseguire il test. Il SW fornisce il valore di D_n ed il valore p. Valori grandi di D_n corrispondono a forti discrepanze tra $F(x)$ e $\widehat{F}(x)$, quindi inducono a rifiutare l'*ipotesi nulla*, cioè l'ipotesi che il campione provenga da quella distribuzione. A valori grandi di D_n corrispondono dei piccoli valori p. Come sempre, quando p è molto piccolo tendiamo a rifiutare l'ipotesi nulla (in questo caso quindi, se il valore p è piccolo, tendiamo a pensare che il campione *non* abbia distribuzione $F(x)$).

2.1 La logica dietro il test

La distribuzione della v.a.

$$D_n = \sup_{k=1,...,n} \left| F(X_k) - \frac{k-1}{n} \right| \vee |1 - F(X_n)|$$

sotto l'ipotesi nulla, ovvero quando le X_k hanno c.d.f. $F(x)$, *non dipende da* $F(x)$, ma solo da n. Infatti, le v.a.

$$U_k := F(X_k)$$

hanno distribuzione uniforme (U è a valori in $[0,1]$ e vale $P(F(X) \leq t) = P(X \leq F^{-1}(t)) = F(F^{-1}(t)) = t$) e sono indipendenti. Il calcolo analitico di questa distribuzione è difficile, ma tramite simulazione non è difficile costruirsi una tavola dei quantili ed eseguire un test.

Si noti che, al giorno d'oggi, se anche la distrubuzione di D_n fosse stata dipendente da F, avremmo potuto calcolarne i quantili volta per volta per simulazione. Il test di KS ha il vantaggio di non richiedere necessariamente questo volta per volta.

2.2 Il test per simulazione

Col metodo di simulazione Monte Carlo si può facilmente eseguire questo test con le seguenti operazioni: dato n,

- chiamiamo $\widehat{D}_n$ il valore relativo al campione sperimentale;
- generiamo una tabella $n \times 10000$ di numeri uniformi e costruiamo con essa un campione $d_1, ..., d_{10000}$ della D_n;
- ordiniamo il campione: $d'_1, ..., d'_{10000}$;
- prendiamo in questo campione ordinato il valore più vicino a $\widehat{D}_n$ e chiamiamo k il suo numero d'ordine;ù
- calcoliamo la proporzione $\frac{10000-k}{10000}$: questo è il valore p (essndo la proporzione di casi in cui è venuto un valore di D_n più estremo di quello sperimentale).

3 Considerazioni generali sui test. Potenza di un test

Schematizziamo concettualmente la teoria dei test statistici.

In prima approssimazione, l'idea è questa: se i dati $x_1, ..., x_n$ provengono da una certa distribuzione $F_X(x)$, un certo riassunto $T = T(x_1, ..., x_n)$ dei dati avrà una certa distribuzione $F_T(x)$ e quindi certi sui valori sono improbabili. Ora, se il campione sperimentale esaminato produce un valore di $\widehat{T}$ improbabile, siamo indotti a concludere che sia falsa l'ipotesi che i dati $x_1, ..., x_n$ provengano da $F_X(x)$.

Esempio (variante t di Student del test per la media descritto sopra): se i dati $x_1, ..., x_n$ provengono da una gaussiana di media μ_0, il riassunto $T = \frac{\overline{X}-\mu_0}{S}\sqrt{n}$, con $\overline{X}$ ed S pari a media e deviazione empiriche, ha distribuzione t di Student a $(n-1)$ gradi di libertà. Se il campione in esame produce il valore $\widehat{T}$, si calcola la probabilità che una t di Student a $(n-1)$ gradi di libertà superi $\widehat{T}$: questo è il valore p e rappresenta la probabilità che, nell'ipotesi che $x_1, ..., x_n$ provengano da una gaussiana di media μ_0, il riassunto T assuma un valore $\geq \widehat{T}$. Se il valore p è piccolo, è improbabile che sotto tale ipotesi esca il valore $\widehat{T}$, quindi siamo indotti a rifiutare l'ipotesi.

Due elementi di questo ragionamento lasciano perplessi: i) siamo passati dal valore $\widehat{T}$ all'insieme di valori $\geq \widehat{T}$; ii) potevamo arbitrariamente scegliere altri ragionamenti, es. calcolare la probabilità che una t di Student a $(n-1)$ gradi di libertà assuma valori in $\left[\widehat{T}-1, \widehat{T}\right]$, oppure già avevamo evidenziato l'arbitrarietà nella scelta tra bilatero e unilatero. Oppure, per esemplificare di nuovo (ii), potevamo calcolare l'indicatore $T = X_1$, che è una gaussiana di media μ_0 e calcolare la probabilità che una gaussiana di media μ_0 e deviazione S superi x_1.

Cosa aiuta in questa selva di obiezioni?

3.1 Potenza

Riassumiamo esemplificando con il test per la media di una gaussiana. Dato il campione sperimentale $x_1, ..., x_n$, ad esso possiamo associare molte variabili aleatorie potenzialmente utili per eseguire un test:

$$Z, \quad |Z|, \quad t, \quad F, \quad X_1$$

e sicuramente altre (non abbiamo descritto t ed F in dettaglio). Inoltre, possiamo usarle potenzialmente in vari modi per eseguire un test:

$$P\left(|Z| > \left|\widehat{Z}\right|\right), \quad P\left(Z > \widehat{Z}\right), \quad P\left(Z < \widehat{Z}\right),$$
$$P\left(Z = \widehat{Z}\right), \quad P\left(Z \in \left[\widehat{Z}-1, \widehat{Z}\right]\right) \quad P\left(|X_1| > \left|\widehat{X_1}\right|\right)$$

e così via. Ci sono quindi, a priori, infiniti test eseguibili. Appena abbiamo a disposizione il valore sperimentale $\widehat{Z}$, scelto a priori una significatività α, alcuni di questi test porteranno al rifiuto dell'ipotesi, altri no.

Intanto non c'è una contraddizione in questo fatto, contrariamente alla regola generale della logica per cui un'affermazione può solo essere vera o falsa. La conclusione di un test, cioè l'affermazione che l'ipotesi è falsa oppure no, non è una conclusione logica, ma un'estrapolazione ragionevole, in quanto c'è in gioco il livello α prescelto. α è uguale alla probabilità che pur valendo l'ipotesi, essa venga rifiutata dal test (cosidetto errore di prima specie). Quindi la conclusione del test (se è un rifiuto) è soggetta ad un errore di probabilità α. Per questo non siamo all'interno della logica usuale.

Tornando alla sostanza, cioè a come dobbiamo giudica questi infiniti test che hanno dato risposte discordanti, la grandezza decisiva è la *potenza*.