Appunti di Probabilità e Statistica

a.a. 2014/2015 C.d.L. Informatica – Bioinformatica
I. Oliva

Lezione 7

1 Statistica Inferenziale

1.1 Test di ipotesi

Ipotesi statistica: assunto relativo ad uno o più parametri (ignoti) che caratterizzano la distribuzione della popolazione di riferimento.

ATTENZIONE: ipotesi $\neq$ affermazione, nel senso che non possiamo essere sicuri *a priori* che essa sia vera.

Dunque, risulta fondamentale verificare tale supposizione.

Test d'ipotesi: procedura statistica, atta a verificare se le informazioni ricavate dal c.c.s., dal quale si deduce l'ipotesi formulata, risultino in grado di spiegare il fenomeno per l'intera popolazione.

Si tratta di una procedura molto importante dal punto di vista applicativo, in quanto le ipotesi formulate sulla distribuzione generatrice dei dati possono avere profonde implicazioni sulla comprensione di un fenomeno o sulle decisioni operative che lo riguardano.

Una ipotesi statistica può essere *accettata* o *rifiutata* con una certa probabilità. Accettare l'ipotesi statistica non vuol dire che essa sia sicuramente vera, ma, semplicemente che i dati a disposizione (c.c.s.) la avvalorano.

Data una popolazione di cui conosciamo la distribuzione, ma non i valori dei parametri ad essa associati e considerato un c.c.s., l'insieme dei possibili valori assunti da tali parametri si chiama *spazio parametrico.*

Nel momento in cui viene formulata una ipotesi statistica, tale spazio viene diviso in due parti distinte: una contiene i valori dei parametri che soddisfano tale ipotesi, l'altra quelli che non la soddisfano.

Struttura di un test di ipotesi.

Indicato con θ il parametro che si sta stimando, sia H_0 l'ipotesi formulata (*ipotesi nulla*). Ad essa, si aggiunge l'ipotesi contraria H_1 (*ipotesi alternativa*). Mettendo insieme, si ottiene un *sistema di ipotesi:*

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases} \quad ; \quad \begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \geq \theta_0 \end{cases} \quad ; \quad \begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \leq \theta_0 \end{cases}$$

Nel primo caso, si parla di ipotesi nulla *semplice* ed ipotesi alternativa *bilaterale,* negli altri due di ipotesi nulla *semplice* ed ipotesi alternativa *unilaterale.* L'ipotesi nulla si dice *composita* se, quando è verificata, identifica più di un parametro e, dunque, più di una v.a.

Ricordiamo che si sta parlando di accettare (o rifiutare) una data ipotesi statistica, sulla base dei dati ricavati dal campione, dunque esiste la possibilitacc di commettere un errore "di valutazione":

Errore di I tipo: si rifiuta H_0 quando H_0 è vera

$$\alpha := P(Rifiuto\, H_0 | H_0\, vera)$$

Errore di II tipo: si accetta H_0 quando H_1 è vera

$$\beta := P(Accetto\, H_0 | H_1\, vera)$$

La quantità $1 - \beta$ si chiama *potenza del test,* la quantità α è il *livello di significatività del test.*

Da un punto di vista teorico, una procedura di verifica di ipotesi statistiche è ottima se riesce a mantenere molto bassi i valori delle probabilità associate ai due tipi di errori.

Da un punto di vista pratico, però, ciò non avviene mai, in quanto, al diminuire dell'uno, aumenta l'altro. La scelta ricade sulla tecnica seguente.

Essendo H_0 l'ipotesi congetturata da noi, mentre H_1 si ricava di conseguenza, allora è più grave commettere un errore del I tipo, rispetto a commetterne uno del secondo. Per questo motivo, fissiamo a priori α su un livello piuttosto piccolo.

Esempio 1.1. *In un laboratorio farmaceutico, si sta sperimentando l'utilizzo di un nuovo farmaco. Si sottopone a test statistico la maggiore o minore bontà di tale farmaco, rispetto a quello attualmente in commercio.*

$$\begin{cases} H_0: \textit{ Il nuovo farmaco non è migliore del vecchio} \\ H_1: \textit{ Il nuovo farmaco è migliore} \end{cases}$$

Gli errori che si possono commettere sono i seguenti:

- *rifiutare l'ipotesi nulla, sapendo che è vera, ossia, produrre il nuovo farmaco, nonostante non sia migliore (I tipo)*
- *accettare l'ipotesi nulla, sapendo che l'ipotesi alternativa è vera, ossia, mantenere in produzione il vecchio farmaco nonostante il nuovo sia migliore (II tipo)*

Costruzione di un test di ipotesi.

- Si sceglie il livello di significatività α
- Si identifica uno stimatore T_n di θ
- Si determina la *statistica test,* con distribuzione di probabilità nota
- Si determina un *valore soglia* C tale che $|T_n - \theta_0| > C$
- Si calcola la *regione critica,* i.e., l'insieme di valori assunti dalla statistica test rispetto al valore soglia:

$$P(|T_n - \theta_0| > C \,|\, H_0) = \alpha$$

Vediamo ora alcuni esempi di costruzione di test di ipotesi.

Verifica d'ipotesi sulla media di una popolazione gaussiana con varianza nota

Sia $\{X_1, \ldots, X_n\}$ un c.c.s. con $X_i \sim \mathscr{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$. Scegliamo, quale ipotesi nulla, la relazione $\mu = \mu_0$.

Lo stimatore da usare è la media campionaria

$$\bar{X} \sim \mathscr{N}\left(\mu, \frac{\sigma^2}{n}\right) \sim \mathscr{N}\left(\mu_0, \frac{\sigma^2}{n}\right) ,$$

dunque la statistica test è

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathscr{N}(0, 1) .$$

e dobbiamo determinare il valore soglia C tale che, fissato α, si abbia

$$\begin{aligned}\alpha &= P\left(\frac{\bar{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} > C \,|\, H_0\right)\\ &= 1-\Phi\left(\frac{C}{\sqrt{\frac{\sigma^2}{n}}}\right)\\ \Rightarrow C &= z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}\end{aligned}$$

Resta da determinare la regione critica (RC):

- se $H_1 : \mu > \mu_0$, allora
$$RC = \{z \in \mathbb{R} \,|\, z > z_{1-\alpha}\}$$
- se $H_1 : \mu < \mu_0$, allora
$$RC = \{z \in \mathbb{R} \,|\, z < -z_{1-\alpha}\}$$
- se $H_1 : \mu \neq \mu_0$, allora
$$RC = \{z \in \mathbb{R} \,|\, z < -z_{1-\alpha/2} \text{ e } z > z_{1-\alpha/2}\}$$

Verifica d'ipotesi sulla media di una popolazione gaussiana con varianza ignota

Sia $\{X_1, \ldots, X_n\}$ un c.c.s. con $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$. Scegliamo, quale ipotesi nulla, la relazione $\mu = \mu_0$.

Poichè la varianza è ignota, sostituiamo a σ^2 il valore della corrispondente statistica campionaria $\bar{S}^2$, mentre lo stimatore da usare è ancora la media campionaria $\bar{X}$, dunque la statistica test è

$$T = \frac{\bar{X}-\mu_0}{\sqrt{\frac{\bar{S}^2}{n}}} \sim t_{n-1} \ .$$

e dobbiamo determinare il valore soglia C tale che, fissato α, si abbia

$$\begin{aligned}\alpha &= P\left(\frac{\bar{X}-\mu_0}{\sqrt{\frac{\bar{S}^2}{n}}} > C \,|\, H_0\right)\\ \Rightarrow C &= t_{n-1;1-\alpha}\sqrt{\frac{\sigma^2}{n}}\end{aligned}$$

Resta da determinare la regione critica (RC):

- se $H_1 : \mu > \mu_0$, allora

$$RC = \{t \in \mathbb{R} \,|\, t > t_{n-1;1-\alpha}\}$$

- se $H_1 : \mu < \mu_0$, allora

$$RC = \{t \in \mathbb{R} \,|\, t < -t_{n-1;1-\alpha}\}$$

- se $H_1 : \mu \neq \mu_0$, allora

$$RC = \{t \in \mathbb{R} \,|\, t < -t_{n-1;1-\alpha/2} \text{ e } t > t_{n-1;1-\alpha/2}\}$$

Verifica d'ipotesi sulla frazione di probabilità di una popolazione bernoulliana

Sia $\{X_1, \ldots, X_n\}$ un c.c.s. con $X_i \sim Bin(1, p)$, $i = 1, \ldots, n$. Scegliamo, quale ipotesi nulla, la relazione $p = p_0$.

Lo stimatore da usare è

$$\pi \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \sim \mathcal{N}\left(p_0, \frac{p_0(1-p_0)}{n}\right) ,$$

dunque la statistica test è

$$Z = \frac{\pi - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \mathcal{N}(0,1) .$$

e dobbiamo determinare il valore soglia C tale che, fissato α, si abbia

$$\begin{aligned} \alpha &= P\left(\frac{\pi - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > C \,|\, H_0\right) \\ &= 1 - \Phi\left(\frac{C}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right) \\ \Rightarrow C &= z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}} \end{aligned}$$

Resta da determinare la regione critica (RC):

- se $H_1 : p > p_0$, allora

$$RC = \{z \in \mathbb{R} \,|\, z > z_{1-\alpha}\}$$

- se $H_1 : p < p_0$, allora

$$RC = \{z \in \mathbb{R} \,|\, z < -z_{1-\alpha}\}$$

- se $H_1 : p \neq p_0$, allora

$$RC = \{z \in \mathbb{R} \,|\, z < -z_{1-\alpha/2} \text{ e } z > z_{1-\alpha/2}\}$$

Verifica d'ipotesi sulla varianza di una popolazione gaussiana Sia $\{X_1, \ldots, X_n\}$ un c.c.s. con $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$. Scegliamo, quale ipotesi nulla, la relazione $\sigma^2 = \sigma_0^2$.

Lo stimatore da usare è $\bar{S}^2$, dunque la statistica test è

$$\chi^2 = \frac{(n-1)\bar{S}^2}{\sigma_0^2} \sim \chi^2_{n-1} \,.$$

Resta da determinare la regione critica (RC):

- se $H_1 : \sigma^2 > \sigma_0^2$, allora

$$RC = \{z \in \mathbb{R} \,|\, z > \frac{\sigma_0^2}{n-1}\chi^2_{n-1;\alpha}\}$$

- se $H_1 : \sigma^2 < \sigma_0^2$, allora

$$RC = \{z \in \mathbb{R} \,|\, z < \frac{\sigma_0^2}{n-1}\chi^2_{n-1;1-\alpha}\}$$

- se $H_1 : \sigma^2 < \sigma_0^2$, allora

$$RC = \{z \in \mathbb{R} \,|\, z < \frac{\sigma_0^2}{n-1}\chi^2_{n-1;1-\alpha/2} \text{ e } z > \frac{\sigma_0^2}{n-1}\chi^2_{n-1;\alpha/2}\}$$

Verifica d'ipotesi sulla varianza di una popolazione gaussiana Sia $\{X_1, \ldots, X_n\}$ un c.c.s. con $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$. Scegliamo, quale ipotesi nulla, la relazione $\sigma^2 = \sigma_0^2$.

Lo stimatore da usare è $\bar{S}^2$, dunque la statistica test è

$$\chi^2 = \frac{(n-1)\bar{S}^2}{\sigma_0^2} \sim \chi^2_{n-1} \,.$$

Resta da determinare la regione critica (RC):

- se $H_1 : \sigma^2 > \sigma_0^2$, allora

$$RC = \{z \in \mathbb{R} \,|\, z > \frac{\sigma_0^2}{n-1}\chi^2_{n-1;\alpha}\}$$

- se $H_1 : \sigma^2 < \sigma_0^2$, allora

$$RC = \{z \in \mathbb{R} \,|\, z < \frac{\sigma_0^2}{n-1}\chi^2_{n-1;1-\alpha}\}$$

- se $H_1 : \sigma^2 < \sigma_0^2$, allora

$$RC = \{z \in \mathbb{R} \,|\, z < \frac{\sigma_0^2}{n-1}\chi^2_{n-1;1-\alpha/2} \text{ e } z > \frac{\sigma_0^2}{n-1}\chi^2_{n-1;\alpha/2}\}$$