Statistica inferenziale

Test di ipotesi

Introduzione

Abbiamo visto che l'esito di una procedura di stima di un parametro consiste in un numero o in un intervallo di numeri.

In questa lezione affronteremo il problema del test di ipotesi e vedremo che l'esito di un test è una decisione tra ipotesi alternative.

Per introdurre le nozioni fondamentali di un test di ipotesi facciamo un esempio.

Esempio

Supponiamo di voler stabilire se una moneta è truccata oppure equa (non truccata).

A tale scopo effettuiamo n=100 lanci della moneta (esperimento) ed osserviamo 38 volte testa (T). Da questo risultato deduciamo che la probabilità empirica di ottenere T in un lancio è $\overline{p}$=38/100.

Possiamo affermare che la moneta è truccata oppure dobbiamo attribuire al caso il fatto di aver osservato 38 T in 100 lanci di una moneta non truccata?

Esempio (cont.)

Analizziamo gli effetti del caso. Supponiamo che la moneta sia equa, ovvero p = ½ . In questa ipotesi il numero di T (N_T) in n=100 lanci è una v.a. Binomiale B(100, ½). Ricorda che $E(N_T)=np=50$. Osserva che la deviazione rispetto alla media dei nostri dati sperimentali è 50 – 38 = 12.

Qual è la probabilità, nel caso di moneta equa, di ottenere risultati sperimentali uguali o più estremi di quelli osservati?

Calcoliamo:

$$P\{|N_T-50|\geq 12\}=P\{\{N_T-50\leq -12\}\cup\{N_T-50\geq 12\}\}=$$
$$P\{N_T-50\leq -12\}+P\{N_T-50\geq 12\}=P\{N_T\leq 38\}+P\{N_T\geq 62\}=0.02$$

Nel caso di moneta equa solo due volte su 100 osserverò risultati uguali o peggiori di quelli osservati e quindi sono portato a concludere che la moneta è truccata.

Esempio (cont.)

Inoltre, osserva che 0.02 è la probabilità di commettere un errore, ossia è la probabilità di affermare che la moneta è truccata quando invece essa è equa.

Domanda:

quante volte devo osservare T su 100 lanci della moneta per poter affermare che la moneta è truccata con una probabilità di 0.05 di commettere un errore?

$$
\begin{aligned}
&P\{|N_T - 50| \geq \delta\} = 0.05 \Rightarrow \\
&P\{\{N_T - 50 \leq -\delta\} \cup \{N_T - 50 \geq \delta\}\} = \\
&P\{N_T - 50 \leq -\delta\} + P\{N_T - 50 \geq \delta\} = \\
&P\{N_T \leq 50 - \delta\} + P\{N_T \geq 50 + \delta\} = 0.05
\end{aligned}
$$

Devo osservare un numero di $T \leq 40$ oppure $T \geq 60$.

Esempio (cont.)

Iniziamo con l'osservare che ci sono due possibili ipotesi alternative:

H_0: la moneta è equa;

H_1: la moneta è truccata.

A ciascuna ipotesi, in generale, associamo il valore di qualche parametro rilevante per il problema in oggetto.

In questo caso il parametro potrebbe essere N_T (numero di T in 100 lanci) oppure la probabilità p di ottenere T in un lancio.

Nel primo caso assoceremo $N_T = 50$ ad H_0 e $N_T \neq 50$ ad H_1, nel secondo assoceremo $p = ½$ ad H_0 , e $p \neq ½$ ad H_1 .

Esempio (cont.)

Poiché ogni lancio della moneta può assumere uno di due possibili valori, allora introduciamo una v.a. X di Bernoulli $B(1,p)$ che prende valori 1 o 0 con probabilità p e $1-p$ seconda che in un lancio esce T o C.

Il risultato del nostro esperimento è costituito dai valori di n=100 v.a. di Bernoulli indipendenti X_k, con k=1,…,n. Per la LGN possiamo stimare p attraverso $\overline{X}$ il cui valore è $\overline{p}$ = 38/100 dato dal rapporto del numero di T sui 100 lanci.

Vogliamo utilizzare questi dati sperimentali per decidere se il valore del parametro p è uguale a ½ o no, cioè se accettare H_0 o se accettare H_1.

Esempio (cont.)

La maniera più semplice è quella di fissare una soglia $\delta>0$ e di confrontare $|\ \bar{p} - ½\ |$ con tale quantità. Se $|\ \bar{p} - ½\ | > \delta$ allora la differenza tra $\bar{p}$ e ½ è troppo grande per poter affermare che l'ipotesi H_0 è vera e quindi accetteremo l'ipotesi H_1.

In generale, le ipotesi alternative saranno formulate in termini dei valori di un parametro θ di qualche distribuzione. Si divide l'insieme di tutti i possibili valori che θ può assumere in due parti Θ_0 e Θ_1. Un'ipotesi è $\theta\in\Theta_0$, mentre l'ipotesi alternativa è $\theta\in\Theta_1$.

Nel nostro esempio, p è il parametro di $B(1,p)$ con $p\in(0,1)$ e quindi le nostre due ipotesi sono definite da $\Theta_0=\{½\}$ mentre $\Theta_1= (0, ½)\cup(½,1)$.

Osservazione

Nota che le decisioni prese sulla base di test statistici sono solo probabilisticamente corrette. Avremo sempre la possibilità di commettere degli errori ed è importante calcolare la probabilità di commettere tali errori.

Vedremo il legame tra il valore della soglia δ e la probabilità di errore.

Nel nostro esempio, se il numero di T in 100 lanci è 80 allora sono portato a dire che H_1 è vera. Tale decisione è solo probabilisticamente corretta in quanto anche nell'ipotesi in cui H_0 è vera posso osservare 80 T in 100 lanci.

Esempio (cont.)

Nel nostro esempio fissiamo $\delta=0.1$ e calcoliamo dai dati empirici $\bar{p}$. Se $|\bar{p} - ½| < \delta$ allora accetteremo H_0; se invece $|\bar{p} - ½| > \delta$ allora accetteremo H_1.

È evidente che il test ci può condurre a prendere delle decisioni errate in uno dei due seguenti modi:

il valore vero è $p=½$ ma i dati sperimentali, essendo $|\bar{p}-½|>\delta$, ci portano ad accettare H_1 e quindi a concludere che $p\neq½$;

oppure

realmente $p\neq½$ ma essendo $|\bar{p} - ½| < \delta$ allora siamo portati ad accettare H_0 e quindi a concludere che $p=½$.

Osservazione

Quindi in generale possiamo commettere due tipi di errori:

affermare che H_0 è falsa quando H_0 è vera;

affermare che H_0 è vera quando H_0 è falsa.

È importante controllare le probabilità di commettere tali errori.

Nel nostro esempio, tali probabilità dipendono dalla soglia δ.

Se δ è grande allora $P\{| \bar{p} - ½ | > \delta \}$ è piccola se H_0 è vera (p=½) e quindi la probabilità di commettere un errore del primo tipo è piccola. Ma aver scelto un δ grande ha come conseguenza quella di affermare spesso che H_0 è vera anche quando essa è falsa, e quindi di aumentare la probabilità di commettere un errore del secondo tipo. Quindi il valore di δ deve essere scelto opportunamente.

Ipotesi nulla e ipotesi alternativa

Chiameremo H_0 **ipotesi nulla** e H_1 **ipotesi alternativa**.

In generale lo sperimentatore sceglie come ipotesi nulla quella più conservativa e utilizza i dati sperimentali per verificare se è possibile rigettare l'ipotesi nulla in favore di quella alternativa.

Ad esempio, nello studio dell'espressione di un gene X in una patologia Y, supponiamo che il livello di espressione di X è uguale in soggetti sani (H) ed in soggetti affetti dalla patologia (D) (ipotesi nulla). Dopo, misuriamo il livello di espressione di X in un gruppo di soggetti sani X_H ed in un gruppo di soggetti malati X_D (dati sperimentali). Se X_H-X_D è grande allora decido di rigettare H_0 in favore dell'ipotesi che X si esprime in modo diverso tra sani e malati (ipotesi alternativa).

Tipi di errore

Chiameremo **errore di prima specie** quello che si commette rigettando H_0 quando H_0 è vera.

Chiameremo **errore di seconda specie** quello che si commette accettando H_0 quando H_0 è falsa.

Questi due tipi di errori non hanno lo stesso peso: è più grave commettere un errore di prima specie che uno di seconda specie (è molto più grave dire che un farmaco è efficace quando non lo è piuttosto che dire il contrario). Inoltre poiché l'errore di prima specie si commette sotto l'ipotesi nulla vera, sarà sempre possibile, in generale, calcolare la probabilità di tale errore. Viceversa, l'errore di seconda specie si commette sotto l'ipotesi alternativa vera e la sua probabilità non è sempre possibile calcolare.

Definizione

Sia $\theta \in \Theta$ un parametro. H_0 è definita dai valori di θ in Θ_0, dove $\Theta_0 \subset \Theta$, e H_1 dai valori di θ nel complementare Θ_1 di Θ_0.

Chiameremo **regione critica** l'evento *D* che conduce al rifiuto dell'ipotesi nulla H_0.

Chiameremo **livello del test** la quantità:

$$\alpha = \max_{\theta \in \Theta_0} P_\theta \{D\}$$

Chiameremo **potenza del test** la funzione:

$$\pi(\theta) = P_\theta \{D\} \qquad \theta \in \Theta_1$$

Chiameremo **significatività del test** il più piccolo valore α_s del livello che, per un campione dato, conduce al rifiuto di H_0.

Osservazioni

La regione critica D è un evento del tipo $\{T \geq t\}$ definito da una statistica T la cui distribuzione è nota sotto l'ipotesi nulla vera.

Il livello α del test è la massima probabilità di commettere un errore di prima specie ed è fissato dallo sperimentatore (valori possibili sono α=0.05, α=0.01, α=0.10).

La potenza del test è la probabilità di rigettare H_0 quando H_0 è falsa, ovvero è la probabilità di accettare H_1 quando H_1 è vera.

La probabilità β di commettere un errore di seconda specie (accettare H_0 quando H_0 è falsa) è $\beta = 1 - \pi(\theta)$ con $\theta \in \Theta_1$.

La significatività del test è una quantità che dipende dai dati ed è il più piccolo livello del test che porta ad un rifiuto di H_0. La significatività del test, detta anche **p-value**, è la probabilità di ottenere valori di T più estremi di quelli realmente ottenuti.

Un test è buono quando α e α_s sono piccoli e $\pi(\theta)$ è grande.

Esempio (cont.)

Introduciamo n=100 v.a. i.i.d. X_k, k=1,2,…,n, B(1,p) con $p \in (0,1)$ con $X_k=1$ se nel k° lancio esce T, e $X_k=0$ altrimenti.

Introduciamo la v.a. $N_T=X_1+\ldots+X_n$ (numero di T in 100 lanci) che sappiamo essere B(100,p). Il nostro esperimento ci ha restituito una osservazione di N_T uguale a 38. In base a questo risultato sperimentale vogliamo decidere quale ipotesi accettare:

$$H_0 : p = \frac{1}{2}, \qquad H_1 : p \neq \frac{1}{2}.$$

Esempio (cont.)

Abbiamo:

$N_T = \sum_{k=1}^{n} X_k$ e $\overline{X} = \frac{N_T}{n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ i cui valori sperimentali sono

$n_T = 38$ e $\overline{p} = \frac{38}{100} = 0.38$. Poiché N_T è $B(n, p)$ allora:

$$E(N_T) = np, \qquad Var(N_T) = np(1-p),$$

$$E(\overline{X}) = p, \qquad Var(\overline{X}) = \frac{p(1-p)}{n}.$$

Per il TLC si ha che $\overline{X} : N\left(p, \frac{p(1-p)}{n}\right)$ e quindi:

$$Z = \sqrt{n}\,\frac{\overline{X} - p}{\sqrt{p(1-p)}} : N(0,1).$$

Esempio (cont.)

$$\text{Consideriamo } \delta > 0 \text{ e calcoliamo:}$$

$$P\left\{\left|\overline{X} - \frac{1}{2}\right| \geq \delta\right\} = P\left\{\left\{\overline{X} - \frac{1}{2} \leq -\delta\right\} \cup \left\{\overline{X} - \frac{1}{2} \geq \delta\right\}\right\}$$

$$= P\left\{\overline{X} - \frac{1}{2} \leq -\delta\right\} + P\left\{\overline{X} - \frac{1}{2} \geq \delta\right\}.$$

$$\text{Sostituendo: } \overline{X} = \sqrt{\frac{p(1-p)}{n}} Z + p \text{ si ha:}$$

$$= P\left\{\sqrt{\frac{p(1-p)}{n}} Z + p \leq \frac{1}{2} - \delta\right\} + P\left\{\sqrt{\frac{p(1-p)}{n}} Z + p \geq \frac{1}{2} + \delta\right\}$$

$$= P\left\{Z \leq \left(\frac{1}{2} - p - \delta\right)\sqrt{\frac{n}{p(1-p)}}\right\} + P\left\{Z \geq \left(\frac{1}{2} - p + \delta\right)\sqrt{\frac{n}{p(1-p)}}\right\}$$

Esempio (cont.)

$$= P\left\{ Z \le \left(\frac{1}{2} - p - \delta\right)\sqrt{\frac{n}{p(1-p)}} \right\} + 1 - P\left\{ Z \le \left(\frac{1}{2} - p + \delta\right)\sqrt{\frac{n}{p(1-p)}} \right\}$$

Poiché Z è approssimativamente $N(0,1)$ allora:

$$\approx \Phi\left(\left(\frac{1}{2} - p - \delta\right)\sqrt{\frac{n}{p(1-p)}}\right) + 1 - \Phi\left(\left(\frac{1}{2} - p + \delta\right)\sqrt{\frac{n}{p(1-p)}}\right).$$

Sotto l'ipotesi H_0 vera $\left(p = \frac{1}{2}\right)$ si ha:

$$P\left\{ \left| \overline{X} - \frac{1}{2} \right| \ge \delta \right\} \approx \Phi\left(-2\delta\sqrt{n}\right) + 1 - \Phi\left(2\delta\sqrt{n}\right).$$

Ricordando la proprietà $\Phi(-x) = 1 - \Phi(x)$ si ha:

$$P\left\{ \left| \overline{X} - \frac{1}{2} \right| \ge \delta \right\} \approx 1 - \Phi\left(2\delta\sqrt{n}\right) + 1 - \Phi\left(2\delta\sqrt{n}\right) = 2\left[1 - \Phi\left(2\delta\sqrt{n}\right)\right].$$

Esempio (cont.)

Avendo fissato $\delta = 0.1$ allora la regione critica del test è:

$$D = \left\{ \left| \overline{X} - \frac{1}{2} \right| \geq 0.1 \right\}.$$

Inoltre, avendo osservato un valore di $\overline{X}$ uguale a $\overline{p} = 0.38$, si ha:

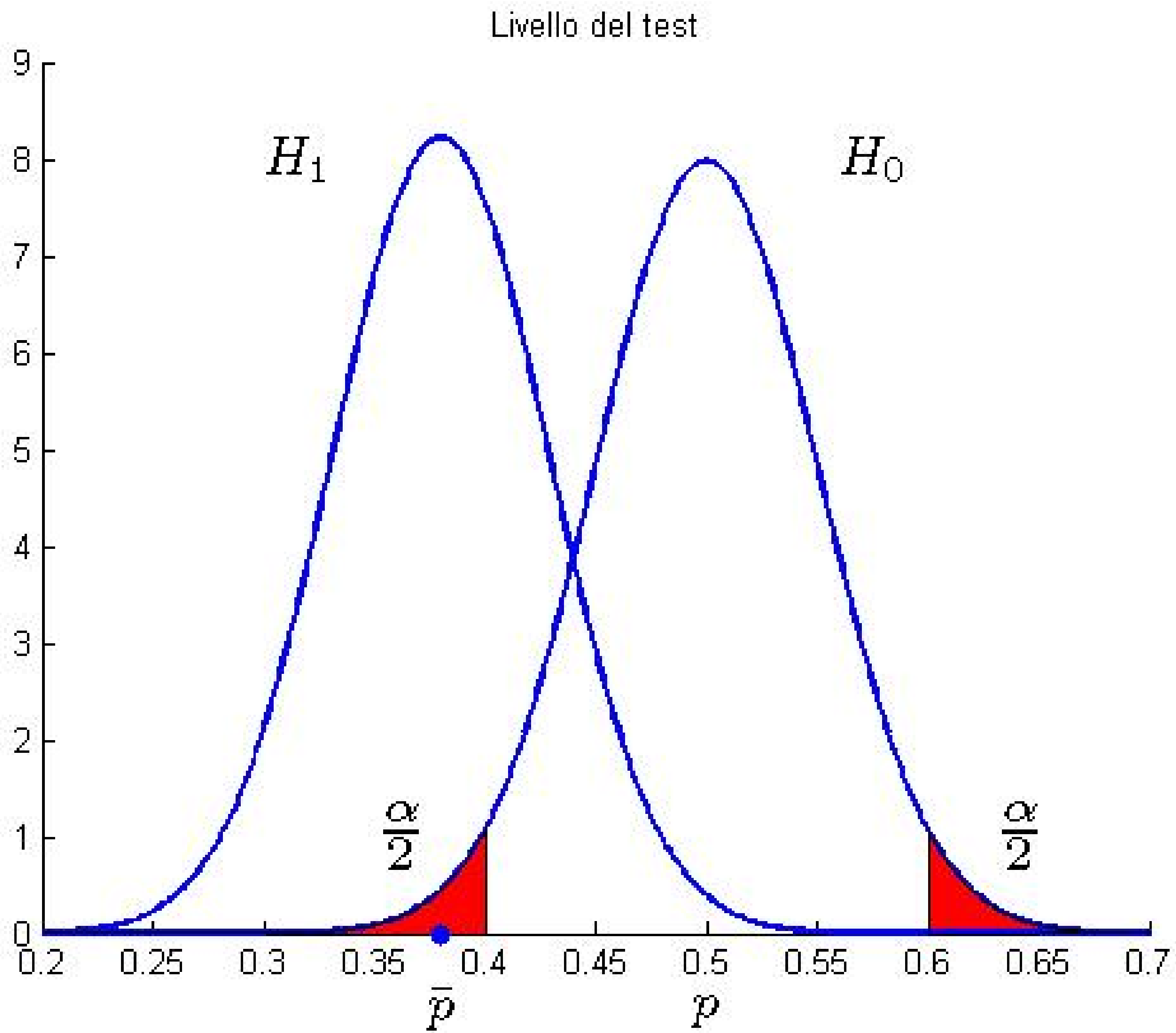
$$\left| \overline{p} - \frac{1}{2} \right| = |0.38 - 0.5| = 0.12 > 0.1$$

Quindi l'esito del test è che noi rigetteremo H_0 in favore di H_1.

Calcoliamo il livello α del test in corrispondenza del valore di δ fissato. Infatti, sotto l'ipotesi H_0 vera, per $n = 100$ si ha:

$$\alpha = P\{D\} = P\left\{ \left| \overline{X} - \frac{1}{2} \right| \geq 0.1 \right\} = 2[1 - \Phi(2)] = 2[1 - 0.97725] = 0.046.$$

Quindi aver fissato $\delta = 0.1$ produce un test di livello $\alpha = 0.046$ che ci porta a rigettare H_0.

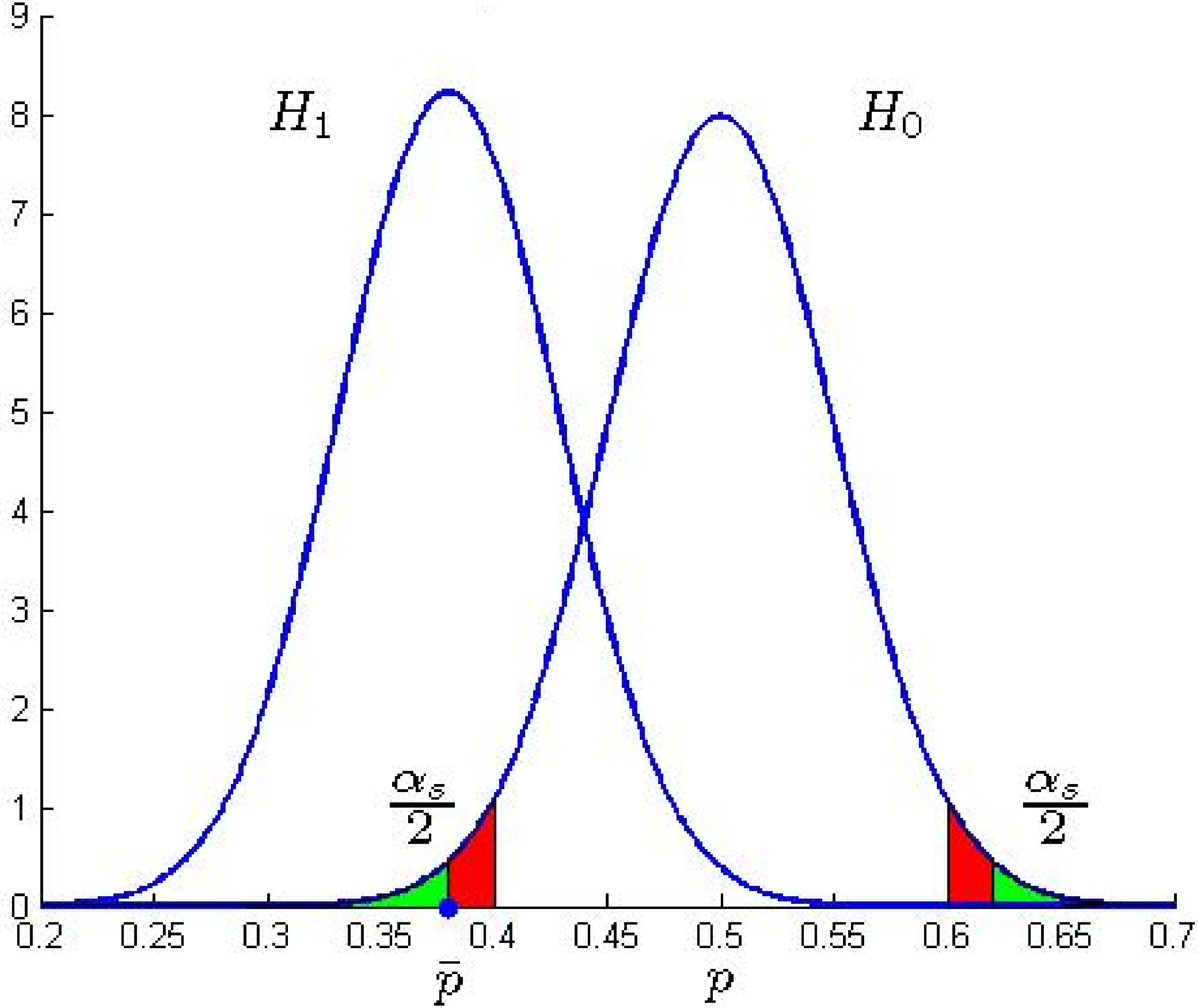

Livello del test
9
8
7
6
5
4
3
2
1
0
H_1
H_0
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
$\bar{p}$
p

Esempio (cont.)

Calcoliamo il p - value:

$$p - value = P\left\{\left|\overline{X} - \frac{1}{2}\right| \geq \left|\overline{p} - \frac{1}{2}\right|\right\} = 2\left[1 - \Phi\left(2.4\right)\right] = 0.016.$$

Significatività del test



Esempio (cont.)

Nella pratica prima si fissa il livello α del test ed in base a questo valore si stabilisce la regione critica. Infatti :

$$\alpha = P\{D\} = P\left\{\left|\overline{X} - \frac{1}{2}\right| \geq \delta\right\} = 2\left[1 - \Phi\left(2\delta\sqrt{n}\right)\right]$$

da cui segue $\Phi\left(2\delta\sqrt{n}\right) = 1 - \frac{\alpha}{2}$ e quindi il quantile di ordine $1 - \frac{\alpha}{2}$ è :

$\varphi_{1-\frac{\alpha}{2}} = 2\delta\sqrt{n}.$

Quindi il valore di δ che definisce D a livello α è :

$\delta = \frac{1}{2\sqrt{n}}\varphi_{1-\frac{\alpha}{2}}.$

Per $\alpha = 0.05$ e $n = 100$ si ha $\delta = 0.098$. Poichè $\left|\overline{p} - \frac{1}{2}\right| = 0.12 > \delta$

allora rigetto H_0.

Esempio (cont.)

Alternativamente, nella pratica prima si fissa il livello α del test e poi si calcola il p - value. Se $p-value \leq \alpha$ allora rigetto H_0 altrimenti accetto H_0.

Nel nostro esempio, essendo $p-value = 0.016$ ed $\alpha = 0.05$, allora rigetto H_0 e concludo che la moneta è truccata.

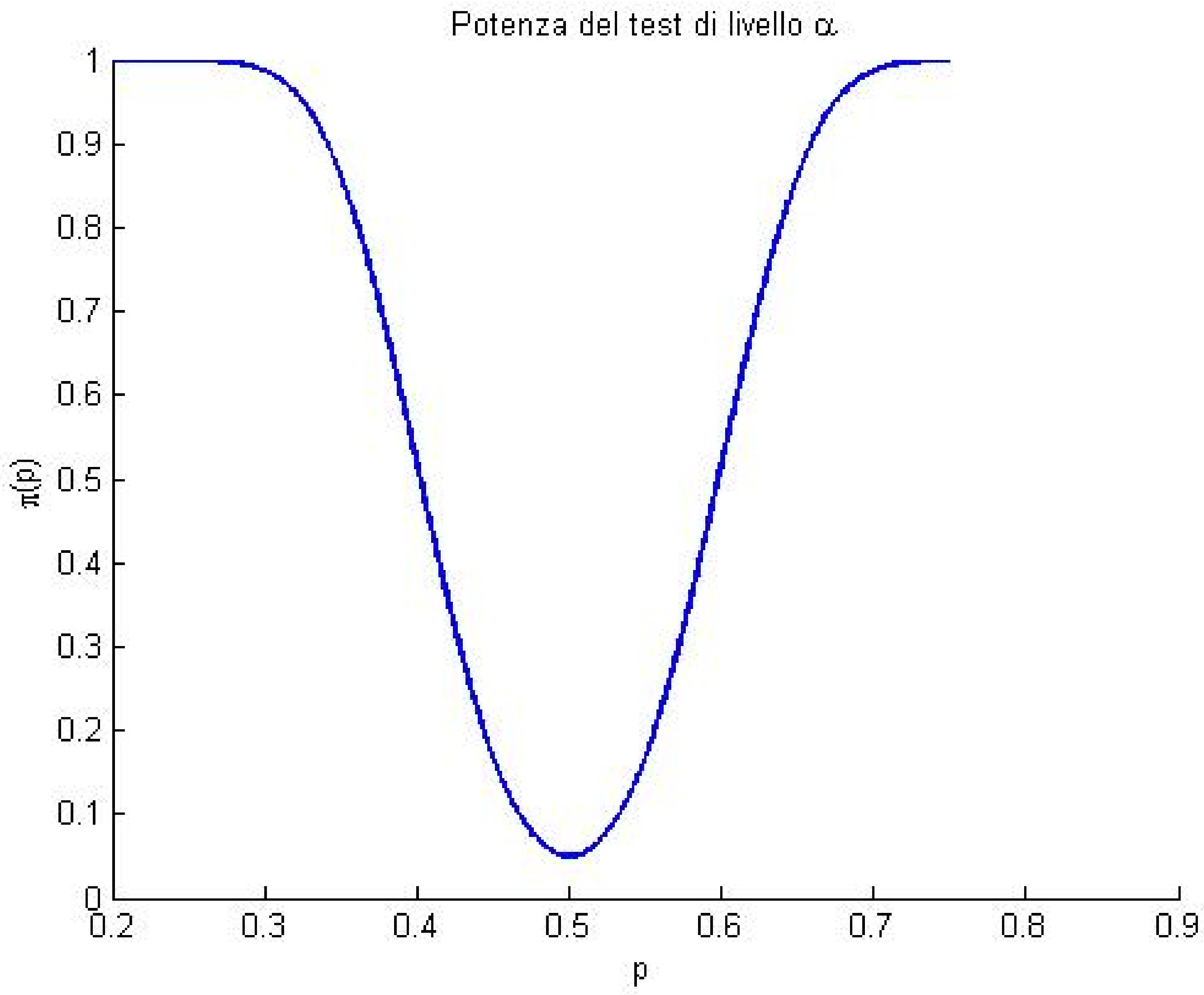
Esempio (cont.)

Calcoliamo la potenza del test con livello $\alpha = 0.05$, ossia $\delta = 0.098$.
Infatti, nell'ipotesi che H_1 sia vera si ha :

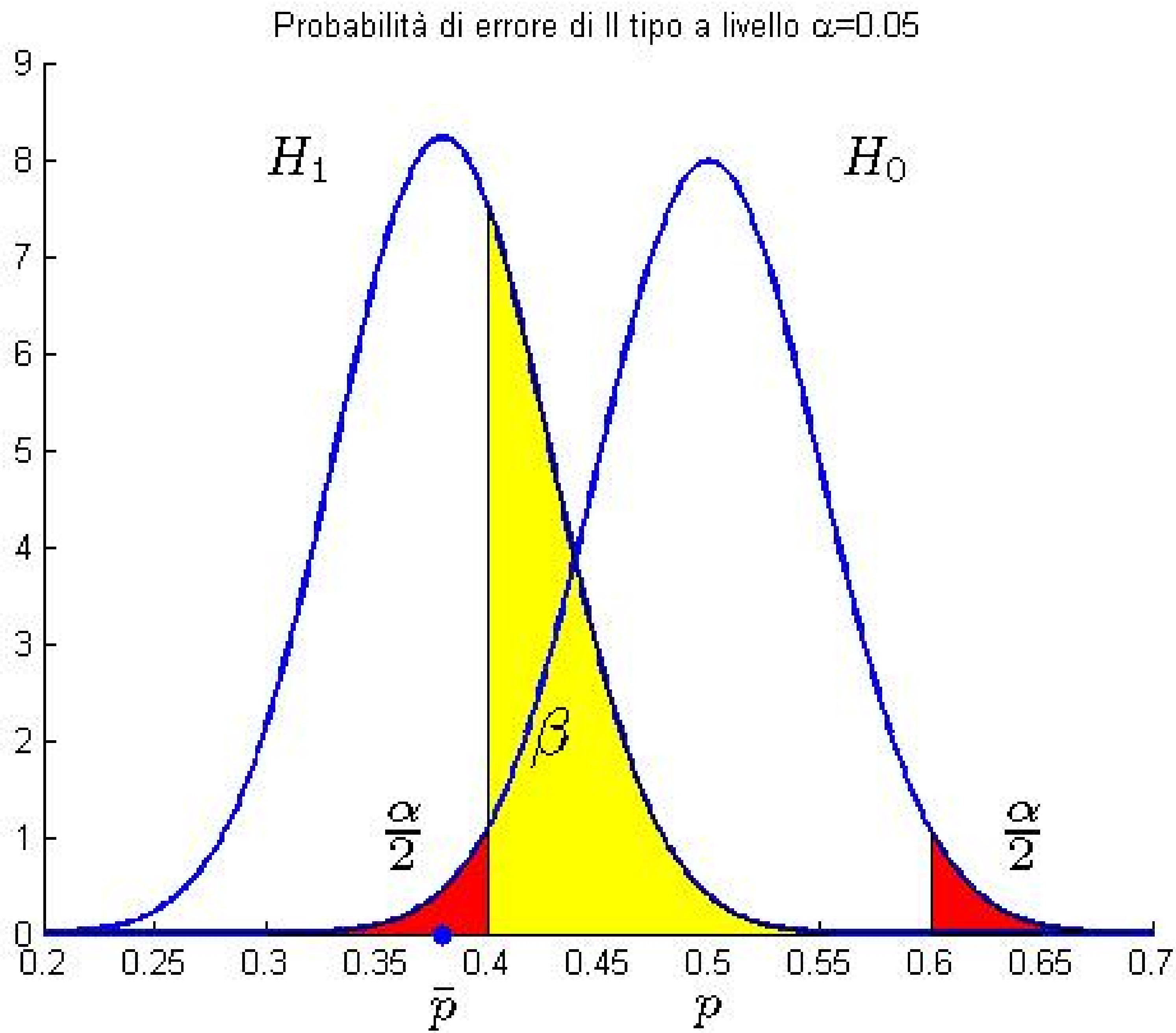
$$\pi(p) = P\{D\} = P\left\{\left|\overline{X} - \frac{1}{2}\right| \geq \delta\right\}$$

$$= \Phi\left(\left(\frac{1}{2} - p - \delta\right)\sqrt{\frac{n}{p(1-p)}}\right) + 1 - \Phi\left(\left(\frac{1}{2} - p + \delta\right)\sqrt{\frac{n}{p(1-p)}}\right).$$

Per n $= 100$ e $\delta = 0.098$ il grafico di $\pi(p)$ è il seguente :



Potenza del test di livello α
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
π(p)
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
p


Probabilità di errore di II tipo a livello α=0.05
H_1
H_0
β
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
0 1 2 3 4 5 6 7 8 9
0.2 0.25 0.3 0.35 0.4 0.45 0.5 0.55 0.6 0.65 0.7
$\bar{p}$
p