

Corso di Statistica Facoltà di Economia

a.a. 2010-2011

francesco mola

Lezione n° 19

Sommario

- Test o verifica delle ipotesi
 - *fondamenti dei test*
 - *costruzione dei test*
 - *Errori e probabilità di errori*
- Test parametrici
 - *Test sul valore medio (e sulla proporzione campionaria)*
 - *Test sulla varianza*
 - *Test sulla differenza tra medie (e sulla differenza tra proporzioni)*
- Test non parametrici
 - *Test sulla forma*
 - *Test sull'indipendenza*

a.a. 2010-2011

statistica-francesco mola

2

Obiettivi della verifica delle ipotesi

- “...prendere decisioni in condizioni di incertezza, fornendo un giudizio probabilistico circa le decisioni prese...”
- Differenza tra test parametrici e non parametrici

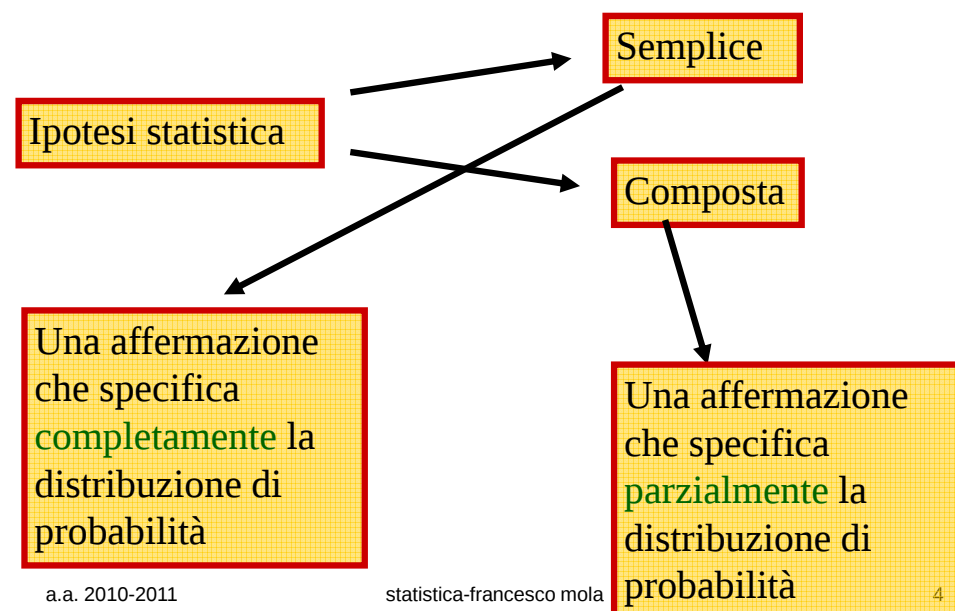
● Un test è una regola istituita sullo spazio campionario che ci permette di rifiutare o non rifiutare una ipotesi. Tale ipotesi prende il nome di ipotesi nulla o H_0

a.a. 2010-2011

statistica-francesco mola

3

Tipi di ipotesi statistiche



a.a. 2010-2011

statistica-francesco mola

4

Sia $X = f(X, \theta)$

e
 $\Omega(\theta)$ spazio parametrico di X

Sia $\omega_0 \subset \Omega(\theta)$ l'insieme dei
valori specificati da H_0

Si deve scegliere tra :

$$H_0 : \theta \in \omega_0$$



Ipotesi nulla

$$H_1 : \theta \notin \omega_0$$



Ipotesi alternativa

Se H_0 è vera $\Rightarrow \theta \in \omega_0$
Se H_0 è falsa $\Rightarrow \theta \notin \omega_0$

Es.: Ipotesi semplice

$$H_0 : \theta = \theta_0$$

Ipotesi composte

$$H_1 : \theta \neq \theta_0$$



bidirezionale

$$H_1 : \theta < \theta_0$$



unidirezionale

$$H_1 : \theta > \theta_0$$

...riflessioni...

- Si noti che l'ipotesi nulla è anche l'**ipotesi preesistente**, cioè lo **status quo**
- L'ipotesi nulla e l'ipotesi alternativa sono **esaustive** e **disgiunte**, quindi o vale l'una o vale l'altra

• Siccome il nostro obiettivo è cercare una regola **fondata sui risultati campionari**, avremo campioni che ci porteranno ad accettare l'ipotesi nulla e campioni che ci indurranno a rifiutarla.

Dividiamo lo spazio campionario C in due regioni:

$$C_0 \subset C \text{ e } C_1 \subset C \text{ con } C_0 \cup C_1 = C$$

Avremo campioni sia in C_0 che in C_1

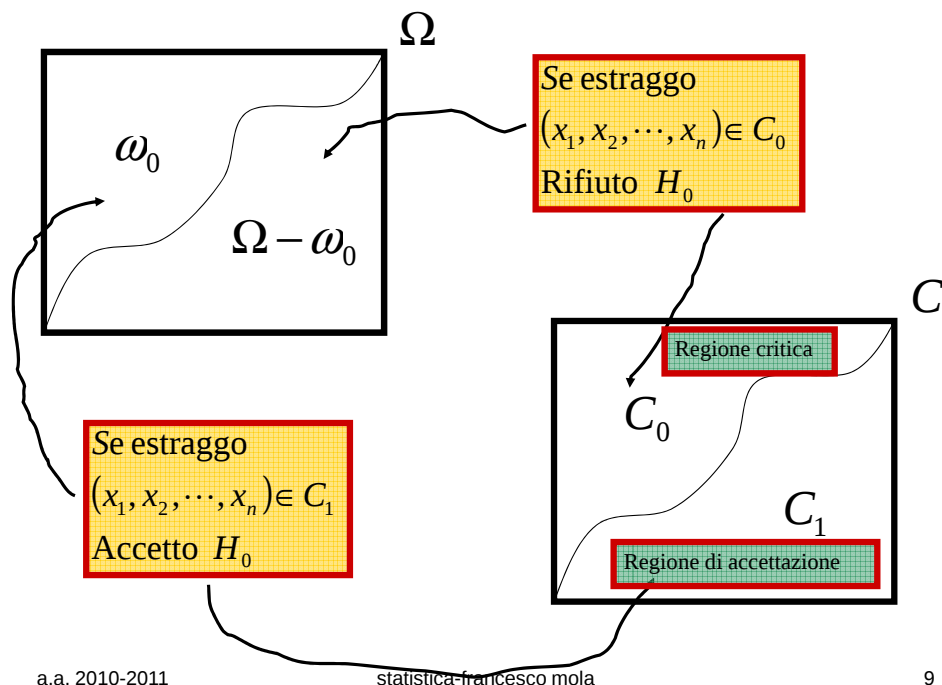
Se $(x_1, x_2, \dots, x_n) \in C_0 \Rightarrow$ dobbiamo rifiutare H_0

Se $(x_1, x_2, \dots, x_n) \in C_1 \Rightarrow$ non dobbiamo rifiutare H_0

C_0 = regione di rifiuto o regione critica (RC)

C_1 = regione di accettazione

La regola per decidere se accettare o rifiutare l'ipotesi nulla equivale ad una bipartizione dello spazio campionario C



..si noti che...

Le assunzioni che abbiamo fatto sono comunque probabilistiche, nel senso che non è certo che:

$$(x_1, x_2, \dots, x_n) \in C_0$$

implica necessariamente $\theta \notin \omega_0$

come non è certo che

$$(x_1, x_2, \dots, x_n) \in C_1$$

implica necessariamente $\theta \in \omega_0$

Questa corrispondenza è corretta con una certa probabilità!!!

a.a. 2010-2011

statistica-francesco mola

10

Abbiamo quattro possibili soluzioni quando consideriamo una qualsiasi struttura sperimentale!

$$G_1 \equiv (x_1, x_2, \dots, x_n) \in C_1 \mid \theta \in \omega_0 \quad \text{Accettare } H_0 \text{ quando essa è vera}$$

$$G_2 \equiv (x_1, x_2, \dots, x_n) \in C_0 \mid \theta \notin \omega_0 \quad \text{Rifiutare } H_0 \text{ quando essa è falsa}$$

$$E_1 \equiv (x_1, x_2, \dots, x_n) \in C_0 \mid \theta \in \omega_0 \quad \text{Rifiutare } H_0 \text{ quando essa è vera}$$

$$E_2 \equiv (x_1, x_2, \dots, x_n) \in C_1 \mid \theta \notin \omega_0 \quad \text{Accettare } H_0 \text{ quando essa è falsa}$$

$$G_1 \text{ e } G_2 \Rightarrow \text{Decisioni giuste}$$

$$E_1 \Rightarrow \text{Errore di prima specie (o primo tipo)}$$

$$E_2 \Rightarrow \text{Errore di seconda specie (o secondo tipo)}$$

a.a. 2010-2011

statistica-francesco mola

11

Consideriamo adesso le probabilità associate alle decisioni possibili che possiamo prendere.

$$P(E_1) = P(\text{rifiutare } H_0 \mid H_0 \text{ è vera}) = P((x_1, x_2, \dots, x_n) \in C_0 \mid \theta \in \omega_0) = \alpha$$

$$P(E_2) = P(\text{accettare } H_0 \mid H_0 \text{ è falsa}) = P((x_1, x_2, \dots, x_n) \in C_1 \mid \theta \notin \omega_0) = \beta$$

a.a. 2010-2011

statistica-francesco mola

12

$$P(G_1) = P(\text{accettare } H_0 \mid H_0 \text{ è vera}) = P((x_1, x_2, \dots, x_n) \in C_1 \mid \theta \in \omega_0) = 1 - \alpha$$

$$P(G_2) = P(\text{rifiutare } H_0 \mid H_0 \text{ è falsa}) = P((x_1, x_2, \dots, x_n) \in C_0 \mid \theta \notin \omega_0) = 1 - \beta$$

α = Livello di significatività

$1 - \beta$ = Potenza del test

Quale errore è meno grave??? E_1 o E_2

E_2 Non modifica la realtà!!! Non cambia lo *status quo*

Se proviamo ad esprimere le probabilità di commettere errori in funzione solo di C_0 abbiamo:

$$\alpha = P(E_1) = P((x_1, x_2, \dots, x_n) \in C_0 \mid \theta \in \omega_0)$$

$$\beta = P(E_2) = 1 - P((x_1, x_2, \dots, x_n) \in C_0 \mid \theta \notin \omega_0)$$

Sembra evidente quindi che un buon test si traduce nella ricerca di una regione critica

C_0 : α e β siano minimi!!

Si tratta quindi di individuare regioni critiche ottimali (RCO).

Purtroppo non è possibile minimizzare α e β contemporaneamente perché queste probabilità sono legate da una relazione tale che se :

$$\alpha \downarrow \Rightarrow \beta \uparrow$$

$$\alpha \uparrow \Rightarrow \beta \downarrow$$

Significa anche che :

se C_0 aumenta, C_1 diminuisce

se C_0 diminuisce, C_1 aumenta

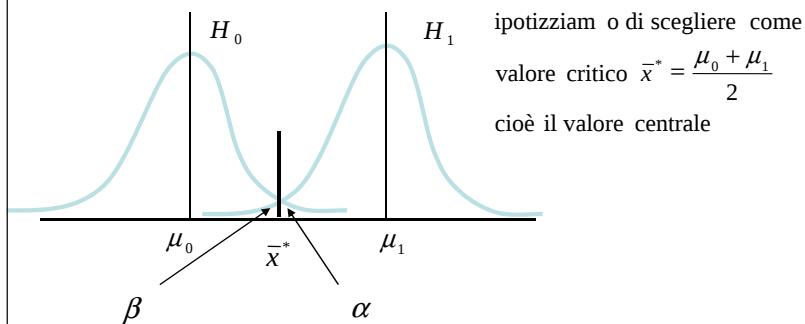
Es.:

$X \sim N(\mu, \sigma^2)$ con μ incognita

$H_0 : \mu = \mu_0$

$H_1 : \mu = \mu_1$ con $\mu_1 > \mu_0$

$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ oppure $\bar{X} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$



$$C_1 = (\bar{x} : \bar{x} < \bar{x}^*) \quad C_0 = (\bar{x} : \bar{x} \geq \bar{x}^*)$$

se sposto $\bar{x}^*$ riduco un'area e quindi diminuisco la probabilità di commettere un errore, ma allo stesso tempo aumento l'altra area e quindi aumento la probabilità di commettere l'altro errore!!

Come si fa a scegliere la regione critica ottimale?

Uno dei criteri proposti è fornito da Neyman e Pearson, secondo cui :

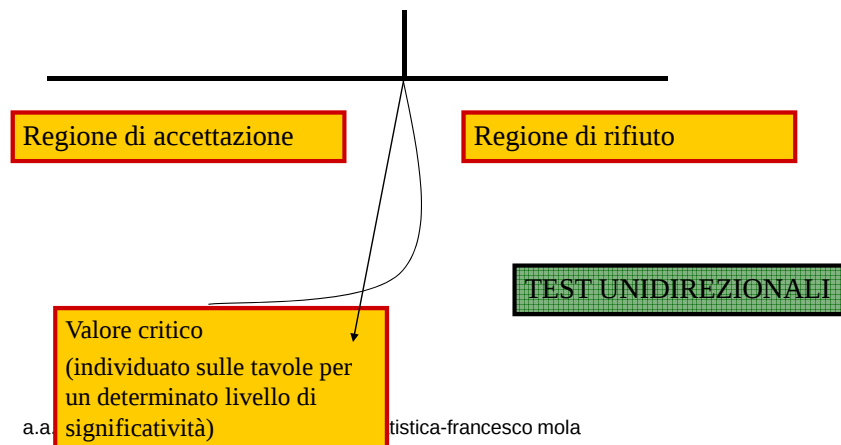
Una RCO di ampiezza α è una Regione Critica C_0 con probabilità di errore di prima specie pari ad α e tale che fra tutte le possibili regioni critiche con la stessa probabilità di errore di prima specie, è quella che possiede il minimo β .

Come si opera praticamente?

- Si fissa il livello di significatività α di solito pari a 0.01 o 0.05 o 0.001
- Si formulano le ipotesi nulla e quella alternativa
- Si considera la statistica test

Come si opera praticamente? (cont.)

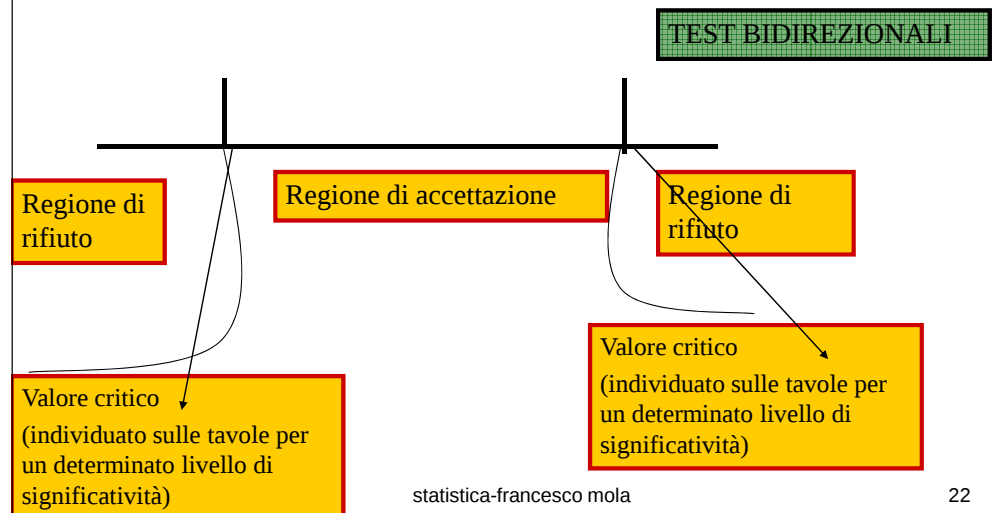
d) Se il test è unidirezionale si identifica la regione di accettazione e di rifiuto



21

Come si opera praticamente? (cont.)

d') Se il test è bidirezionale si identifica la regione di accettazione e le regioni di rifiuto



22

Come si opera praticamente? (cont.)

e) Si confronta la statistica campionaria con il valore critico e si propende per confermare l'ipotesi nulla oppure per rifiutare l'ipotesi nulla.

Principali test parametrici

Consideriamo una variabile casuale distribuita come una Normale:

$$X \sim N(\mu, \sigma^2) \quad \text{con } \sigma^2 = \text{nota}$$

Si estrae un campione di ampiezza n

Si fissa il livello di significatività α

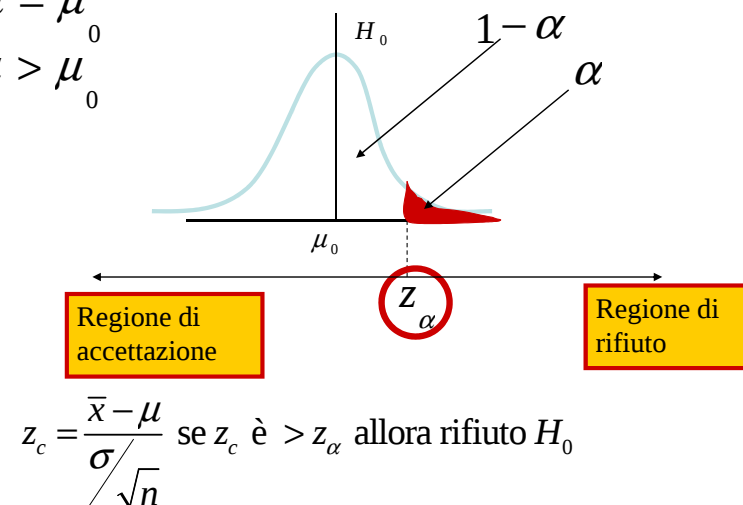
Test sul valore medio con varianza nota:

Abbiamo tre possibili situazioni:

- 1) $H_0 : \mu = \mu_0$
 $H_1 : \mu > \mu_0$ RCO (α): $\bar{X} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$
- 2) $H_0 : \mu = \mu_0$
 $H_1 : \mu < \mu_0$ RCO (α): $\bar{X} \leq \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$
- 3) $H_0 : \mu = \mu_0$
 $H_1 : \mu \neq \mu_0$ RC (α): $\begin{cases} \bar{X} \geq \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ \bar{X} \leq \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \end{cases}$

Test sul valore medio con varianza nota:

- 1) $H_0 : \mu = \mu_0$
 $H_1 : \mu > \mu_0$



Esempio 1: Su un campione casuale di 81 negozi di articoli sportivi la media settimanale di scarpette vendute da ciascun negozio risultata pari a 100. Si sa che nella popolazione la deviazione standard delle vendite settimanali è pari a 8. Si assuma l'ipotesi di normalità del numero di scarpe vendute in ogni negozio.

Da precedenti studi sulla popolazione dei negozi di articoli sportivi si sa che la media settimanale di scarpette vendute è uguale a 96. Si può affermare, ad un livello di significatività del 5%, che il numero medio di scarpette vendute nella popolazione dei negozi sia aumentato?

Soluzione:

$$n = 81$$

$$\mu_0 = 96$$

$$\bar{x} = 100$$

$$\sigma = 8$$

$$\sigma^2 = 64$$

$$\alpha = 0.05$$

$$z_\alpha = -1.64$$

$$H_0 : \mu = 96$$

$$H_0 : \mu > 96$$

valore della media nella popolazione sotto l'ipotesi H_0

valore della media nel campione

valore della popolazione

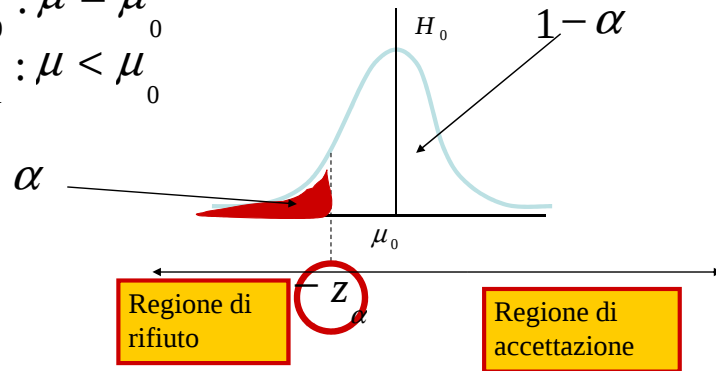
trovato sulle tavole della Z

$$z_c = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{100 - 96}{8 / 9} = \frac{4}{0.89} = 4.49$$

$$z_c > 1.645 \text{ allora rifiuto } H_0$$

Test sul valore medio con varianza nota:

$$2) \begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu < \mu_0 \end{aligned}$$



$$z_c = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \text{ se } z_c \text{ è } < -z_\alpha \text{ allora rifiuto } H_0$$

Esempio 2: Su un campione casuale di 81 negozi di articoli sportivi la media settimanale di scarpe vendute da ciascun negozio risultata pari a 100. Si sa che nella popolazione la deviazione standard delle vendite settimanali è pari a 8. Si assuma l'ipotesi di normalità del numero di scarpe vendute in ogni negozio.

Da precedenti studi sulla popolazione dei negozi di articoli sportivi si sa che la media settimanale di scarpe vendute è uguale a 105. Si può affermare, ad un livello di significatività del 5%, che il numero medio di scarpe vendute nella popolazione dei negozi sia diminuito?

Soluzione:

$$n = 81$$

$$\mu_0 = 105$$

$$\bar{x} = 100$$

$$\sigma = 8$$

$$\sigma^2 = 64$$

$$\alpha = 0.05$$

$$z_\alpha = -1.64$$

trovato sulle tavole della Z

$$H_0: \mu = 105$$

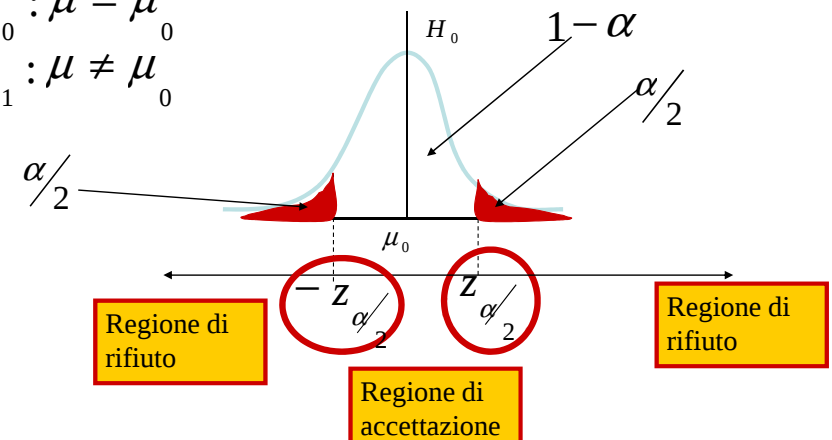
$$H_1: \mu < 105$$

$$z_c = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{100 - 105}{8 / 9} = -\frac{5}{0.89} = -5.61$$

$$z_c < -1.645 \text{ allora rifiuto } H_0$$

Test sul valore medio con varianza nota:

$$3) \begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$



$$z_c = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \text{ se } z_c \text{ è } < -z_{\alpha/2} \text{ oppure } > z_{\alpha/2} \text{ allora rifiuto } H_0$$

Esempio 3: Su un campione casuale di 81 negozi di articoli sportivi la media settimanale di scarpette vendute da ciascun negozio risultata pari a 100. Si sa che nella popolazione la deviazione standard delle vendite settimanali è pari a 8. Si assuma l'ipotesi di normalità del numero di scarpe vendute in ogni negozio.

Da precedenti studi sulla popolazione dei negozi di articoli sportivi si sa che la media settimanale di scarpette vendute è uguale a 99. Si può affermare, ad un livello di significatività del 5%, che il numero medio di scarpette vendute nella popolazione dei negozi sia cambiato?

Soluzione:

$$n = 81$$

$$\mu_0 = 99$$

$$\bar{x} = 100$$

$$\sigma = 8$$

$$\sigma^2 = 64$$

$$\alpha = 0.05$$

$$z_{\alpha} = -1.64$$

$$H_0 : \mu = 99$$

$$H_0 : \mu \neq 99$$

valore della media nella popolazione sotto l'ipotesi H_0

valore dell'amedia nel campione

valore della popolazione

trovato sulle tavole della Z

$$z_c = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{100 - 99}{\frac{8}{\sqrt{9}}} = \frac{1}{0.89} = 1.12$$

$$-1.645 < z_c < 1.645 \quad \text{allora accetto } H_0$$

Per questo test si utilizza la distribuzione Normale Standardizzata

Le tavole da utilizzare sono quelle della Normale Standardizzata

Test sulla proporzione (n grande):

Si opera allo stesso modo del test sulla media sostituendo alla distribuzione della media campionaria quella della proporzione campionaria (come per gli intervalli di Confidenza)

Per questo test si utilizza la distribuzione Normale Standardizzata

Le tavole da utilizzare sono quelle della Normale Standardizzata

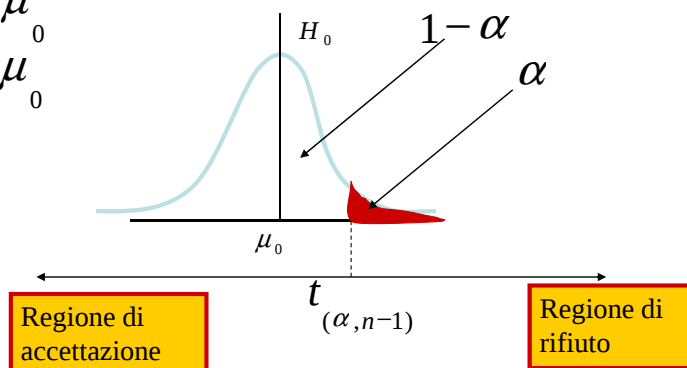
Test sul valore medio con varianza non nota:

Abbiamo tre possibili situazioni:

- 1) $H_0 : \mu = \mu_0$
 $H_1 : \mu > \mu_0$ RCO (α): $\bar{X} \geq \mu_0 + t_{(\alpha, n-1)} \frac{\tilde{s}}{\sqrt{n}}$
- 2) $H_0 : \mu = \mu_0$
 $H_1 : \mu < \mu_0$ RCO (α): $\bar{X} \leq \mu_0 - t_{(\alpha, n-1)} \frac{\tilde{s}}{\sqrt{n}}$
- 3) $H_0 : \mu = \mu_0$
 $H_1 : \mu \neq \mu_0$ RC (α): $\begin{cases} \bar{X} \geq \mu_0 + t_{(\alpha/2, n-1)} \frac{\tilde{s}}{\sqrt{n}} \\ \bar{X} \leq \mu_0 - t_{(\alpha/2, n-1)} \frac{\tilde{s}}{\sqrt{n}} \end{cases}$

Test sul valore medio con varianza non nota:

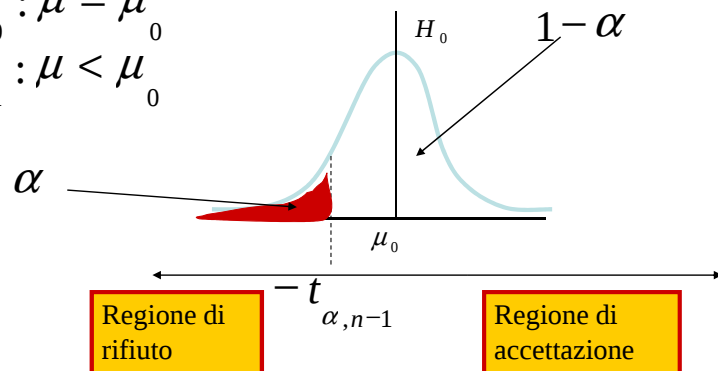
- 1) $H_0 : \mu = \mu_0$
 $H_1 : \mu > \mu_0$



$$t_c = \frac{\bar{x} - \mu}{\tilde{s}/\sqrt{n}} \text{ se } t_c \text{ è } > t_{(n-1, \alpha)} \text{ allora rifiuto } H_0$$

Test sul valore medio con varianza non nota:

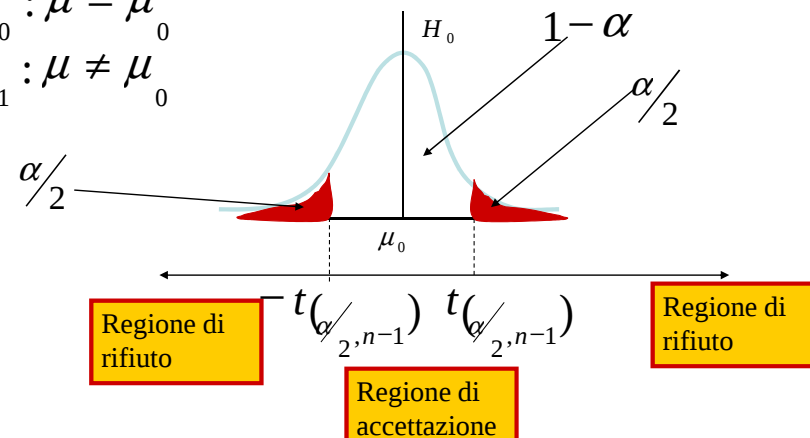
- 2) $H_0 : \mu = \mu_0$
 $H_1 : \mu < \mu_0$



$$t_c = \frac{\bar{x} - \mu}{\tilde{s}/\sqrt{n}} \text{ se } t_c \text{ è } < -t_{(n-1, \alpha)} \text{ allora rifiuto } H_0$$

Test sul valore medio con varianza non nota:

- 3) $H_0 : \mu = \mu_0$
 $H_1 : \mu \neq \mu_0$



$$t_c = \frac{\bar{x} - \mu}{\tilde{s}/\sqrt{n}} \text{ se } t_c \text{ è } < -t_{(n-1, \alpha/2)} \text{ oppure } > t_{(n-1, \alpha/2)} \text{ allora rifiuto } H_0$$

Test sulla differenza tra medie con varianze note

$$1) \begin{array}{ll} H_0: \mu_1 - \mu_2 = 0 & H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 - \mu_2 > 0 & H_1: \mu_1 > \mu_2 \end{array}$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Dato che la differenza tra le medie della popolazione sotto l'ipotesi nulla è pari a 0, abbiamo:

Test sulla differenza tra medie con varianze note (cont.)

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \Rightarrow \text{RCO}(\alpha): Z \geq z_\alpha$$

$$2) \begin{array}{ll} H_0: \mu_1 - \mu_2 = 0 & \text{RCO}(\alpha): Z \leq -z_\alpha \\ H_1: \mu_1 - \mu_2 < 0 & \end{array}$$

$$3) \begin{array}{ll} H_0: \mu_1 - \mu_2 = 0 & \text{RC}(\alpha): |Z| \geq z_{\alpha/2} \\ H_1: \mu_1 - \mu_2 \neq 0 & \end{array}$$

Test sulla differenza tra medie con varianze NON note ma supposte uguali

$$1) \begin{array}{ll} H_0: \mu_1 - \mu_2 = 0 & H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 - \mu_2 > 0 & H_1: \mu_1 > \mu_2 \end{array}$$

$$T = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\Rightarrow \text{RCO}(\alpha): T \geq t_{(\alpha, n^1 + n^2 - 2)}$$

Test sulla differenza tra medie con varianze NON note Ma supposte uguali (cont.)

$$2) \begin{array}{ll} H_0: \mu_1 - \mu_2 = 0 & \text{RCO}(\alpha): T \leq -t_{(\alpha, n^1 + n^2 - 2)} \\ H_1: \mu_1 - \mu_2 < 0 & \end{array}$$

$$3) \begin{array}{ll} H_0: \mu_1 - \mu_2 = 0 & \text{RC}(\alpha): |T| \leq t_{(\alpha/2, n^1 + n^2 - 2)} \\ H_1: \mu_1 - \mu_2 \neq 0 & \end{array}$$

Se $n \rightarrow \infty \Rightarrow T \sim N(0,1)$

(già con $n > 30$ si ha una buona approssimazione)

Test sulla differenza tra proporzioni (n grande):

Si opera allo stesso modo del test sulla differenza tra medie sostituendo alla distribuzione della differenza delle medie campionarie quella della differenza tra proporzioni campionarie (come per gli intervalli di confidenza)

Per questo test si utilizza la distribuzione Normale Standardizzata

Le tavole da utilizzare sono quelle della Normale Standardizzata

Test sulla varianza con media nota:

$$\frac{1}{n} \sum (X_i - \mu)^2 = \bar{S}^2 = \text{stimatore MLE}$$

$$1) \begin{matrix} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{matrix} \quad \text{RCO}(\alpha): \sum (X_i - \mu)^2 \geq \sigma_0^2 \chi_{\alpha, n}^2$$

$$2) \begin{matrix} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{matrix} \quad \text{RCO}(\alpha): \sum (X_i - \mu)^2 \leq \sigma_0^2 \chi_{1-\alpha, n}^2$$

$$3) \begin{matrix} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{matrix} \quad \text{RC}(\alpha): \begin{cases} \sum (X_i - \mu)^2 \geq \sigma_0^2 \chi_{\alpha/2, n}^2 \\ \sum (X_i - \mu)^2 \leq \sigma_0^2 \chi_{1-\alpha/2, n}^2 \end{cases}$$

Per questo test si utilizza la distribuzione del χ^2

I gradi di libertà sono n

Le tavole da utilizzare sono quelle del χ^2

Test sulla varianza con media non nota:

$$\frac{1}{n-1} \sum (X_i - \bar{X})^2 = \tilde{S}^2 = \text{stimatore corretto}$$

$$1) \begin{matrix} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{matrix} \quad \text{RCO}(\alpha): \tilde{S}^2 \geq \frac{\sigma_0^2}{n-1} \chi_{\alpha, n-1}^2$$

$$2) \begin{matrix} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{matrix} \quad \text{RCO}(\alpha): \tilde{S}^2 \leq \frac{\sigma_0^2}{n-1} \chi_{1-\alpha, n-1}^2$$

$$3) \begin{matrix} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{matrix} \quad \text{RC}(\alpha): \begin{cases} \tilde{S}^2 \geq \frac{\sigma_0^2}{n-1} \chi_{\alpha/2, n-1}^2 \\ \tilde{S}^2 \leq \frac{\sigma_0^2}{n-1} \chi_{1-\alpha/2, n-1}^2 \end{cases}$$

Per questo test si utilizza la distribuzione del χ^2

I gradi di libertà sono $n-1$

Le tavole da utilizzare sono quelle del χ^2

Principali test non parametrici

Test sulla bontà di adattamento
Test sull'ipotesi funzionale

Obiettivo:

Verificare se i dati si adattano ad un modello prefissato a priori

$$X \sim f(x)$$

$$H_0 : f(x) = f_0(x)$$

$$H_1 : f(x) \neq f_0(x)$$

$f_0(x)$ è la supposizione

Caso 1. $f_0(x)$ Non contiene parametri incogniti (è completamente specificata).

- ✓ Si divide la X in intervallini (k classi) e siano $p_1, p_2, \dots, p_k$ le probabilità associate a tali classi sotto l'ipotesi nulla.
- ✓ Si estrae un campione, si divide nelle k classi e si osservano i valori $\frac{n_1}{n}, \frac{n_2}{n}, \dots, \frac{n_k}{n}$ ovvero le frequenze relative osservate

classi	p_i	$f_i = n_i/n$	$p_i - f_i$
1	p_1	n_1/n	$p_1 - f_1$
2	p_2	n_2/n	$p_2 - f_2$
.			
.			
.			
k	p_k	n_k/n	$p_k - f_k$

Se $f_0(x)$ è effettivamente la distribuzione di X allora le differenze $p_i - f_i$ sono piccole e comunque dovute a fluttuazioni campionarie

Poiché si è formulata una ipotesi, allora le differenze possono essere attribuite al fatto che $f_0(x)$ è diverso da $f(x)$

La statistica test proposta è:

$$\chi^2 = \sum_{i=1}^k \left[\frac{(f_i - p_i)^2}{p_i} \right] = \frac{\sum_i (n_i - np_i)^2}{np_i}$$

che per n grande $\sim \chi^2_{k-1}$

regola pratica : se $np_i \geq 5 \forall i = 1, 2, \dots, k$ allora la statistica proposta è valida

Consideriamo la distribuzione di 1000 clienti di un supermercato tra le 10 casse attive

cassa	n_i	$f_i = n_i/n$	p_i	$(n_i/n) - p_i$	$((n_i/n) - p_i)^2 / p_i$
1	80	0.080	0.1	0.080-0.1	$(0.080-0.1)^2/0.1$
2	92	0.092	0.1	0.092-0.1	$(0.092-0.1)^2/0.1$
3	106	0.106	0.1	0.106-0.1	$(0.106-0.1)^2/0.1$
4	94	0.094	0.1	0.084-0.1	$(0.084-0.1)^2/0.1$
5	110	0.110	0.1	0.110-0.1	$(0.110-0.1)^2/0.1$
6	120	0.120	0.1	0.120-0.1	$(0.120-0.1)^2/0.1$
7	98	0.098	0.1	0.098-0.1	$(0.098-0.1)^2/0.1$
8	110	0.110	0.1	0.110-0.1	$(0.110-0.1)^2/0.1$
9	90	0.090	0.1	0.090-0.1	$(0.090-0.1)^2/0.1$
10	100	0.100	0.1	0.100-0.1	$(0.100-0.1)^2/0.1$
tot	1000	1	1		12.4

$$H_0 : f(x) = f_0(x)$$

$$H_1 : f(x) \neq f_0(x)$$

$$\text{con } f_0(x) = U(K)$$

$$\alpha = 0.05$$

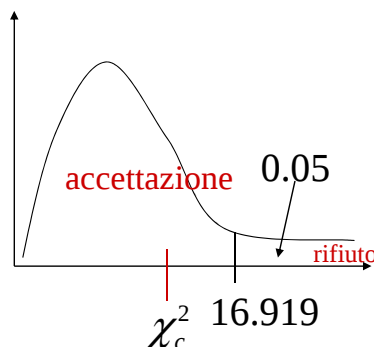
$$g.l. = k - 1 = 10 - 1 = 9$$

$$p_i = 0.1 \quad \forall i = 1, 2, \dots, 10$$

$$\chi^2_c = 12.4$$

$$\chi^2_{(0.05,9)} = 16.919$$

accetto H_0



Caso 2. Se si hanno parametri da stimare allora i gradi di libertà saranno $k-m-1$, dove m = numero dei parametri da stimare

mentre si calcolano $k-1$ g.l. perché esiste un vincolo:

$$\sum p_i = \sum f_i$$

▪ Se $f_0(x) = N(\mu, \sigma^2)$

✓ Si standardizza

✓ Si trova dalle tavole l'area a sinistra del punto z (limite superiore della classe)

✓ Si moltiplica l'area per n ottenendo le frequenze

Esempio

Fissato $\alpha = 0.05$

si verifichi l'ipotesi che i redditi dichiarati dai 500 contribuenti, riportati nella tabella seguente, provengano da una distribuzione normalmente distribuita con media $\mu = 1050$ e varianza $\sigma^2 = 18598$

Classi di reddito	Contribuenti
Meno di 700	2
700 -800	18
800 -900	40
900 -1000	110
1000 -1100	150
1100 -1200	125
1200 -1300	35
1300 -1400	17
1400 e più	3
Totale	500

Lez20_a.a. 2010-2011

statistica-francesco mola

57

Alcune quantità da calcolare

$$z_i = \frac{x_i - \mu}{\sigma} \quad \text{con } x_i = \text{estremo superiore della classe}$$

$F(z_i)$ = percentile della normale standardizzata in z_i
(dedotto dalle tavole della Z)

$$p_i = F(z_i) - F(z_{i-1}) = \text{area tra i due valori di } z$$

$\hat{n}_i = p_i * n$ = numero osservazioni che teoricamente dovrebbero aversi per un modello teorico perfetto.

Attenzione, quando questa frequenza teorica è inferiore a 5 va accorpata con una precedente o successiva coerentemente all'andamento della distribuzione.

$$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = \text{quantità per la determinazione della statistica test}$$

I gradi di libertà vanno considerati in funzione degli accorpamenti fatti.

Lez20_a.a. 2010-2011

statistica-francesco mola

58

Classi di reddito	Contribuenti	Z_i	$F(Z_i)$	$p_i = F(Z_i) - F(Z_{i-1})$	$\hat{n}_i$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
Meno di 700	2	-2.57	0.0050	0.0059	2.504	0.804
700 -800	18	-1.84	0.0327	0.0277	13.866	
800 -900	40	-1.10	0.1337	0.1010	50.52	2.190
900 -1000	110	-0.37	0.3536	0.2198	109.94	0.000
1000 -1100	150	0.35	0.6369	0.2860	143.04	0.337
1100 -1200	125	1.09	0.8623	0.2226	111.30	1.683
1200 -1300	35	1.82	0.9659	0.1035	51.78	5.439
1300 -1400	17	2.55	0.9947	0.0287	14.39	0.520
1400 e più	3		1.0000	0.0052	2.63	
Totale	500			1		10.976

$$\chi_c^2 = 10.976$$

$$\chi_{6;0.05}^2 = 12.59$$

Poiché il valore osservato è minore del valore empirico, NON si può rifiutare l'ipotesi di normalità della popolazione da cui proviene il campione.



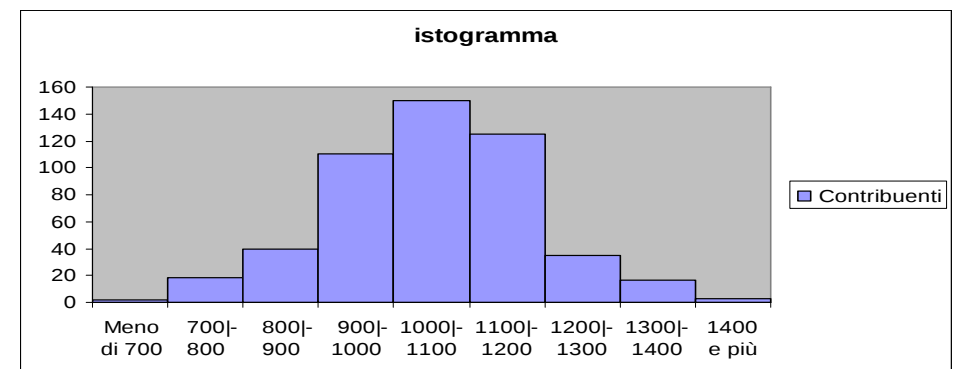
Lez20_a.a. 2010-2011

statistica-francesco mola

59

considerazioni

Il risultato del test è coerente con l'evidenza empirica risultante dall'istogramma costruito a partire dai dati osservati



Lez20_a.a. 2010-2011

statistica-francesco mola

60

Si opera analogamente per le altre distribuzioni:

$$f_0(x) = P(\lambda)$$

$$f_0(x) = B(n, p)$$

Test sull'indipendenza stocastica

$$A = a_1, a_2, \dots, a_k$$

$$B = b_1, b_2, \dots, b_h$$

A \ B	b_1	b_2	b_h	
a_1	$p_{i,j}$			$p_{i,\bullet}$
a_2				
a_k				
	$p_{\bullet,j}$			1

$$p_{i,j} = \text{probabilità congiunta di } a_i \text{ e } b_j$$

$$p_{\bullet,j} = \sum_i p_{i,j}$$

$$p_{i,\bullet} = \sum_j p_{i,j}$$

$$p_{i,j} = p_{i,\bullet} p_{\bullet,j} \Rightarrow A \text{ e } B \quad \forall i, j$$

sono stocasticamente indipendenti

$$H_0 : p_{i,j} = p_{i,\bullet} p_{\bullet,j} \quad \forall i, j$$

$$H_1 : p_{i,j} \neq p_{i,\bullet} p_{\bullet,j} \quad \text{per almeno una coppia di } (i, j)$$

$$\frac{n_{i,j}}{n} = \text{frequenze relative osservate}$$

$$\frac{n_{i,\bullet}}{n} \text{ e } \frac{n_{\bullet,j}}{n} = \text{frequenze relative marginali osservate}$$

A \ B	b_1	b_2	b_h	
a_1	$n_{i,j}$			$n_{i,\bullet}$
a_2				
a_k				
	$n_{\bullet,j}$			1

$$c_{i,j} = \frac{n_{i,\bullet} n_{\bullet,j}}{n}$$

$\hat{n}_{i,j}$ (indicated by a red arrow pointing to $c_{i,j}$)

$$X^2 = \sum_{i=1}^k \sum_{j=1}^h \left[\frac{(n_{i,j} - c_{i,j})^2}{c_{i,j}} \right]$$

$$X^2 = \sum_{i=1}^k \sum_{j=1}^h \left[\frac{(n_{i,j} - \hat{n}_{i,j})^2}{\hat{n}_{i,j}} \right]$$

$$X^2 \sim \chi^2_{[(k-1) \times (h-1)]}$$

