Verifica delle ipotesi

Introduzione

- I metodi statistici che abbiamo discusso fino ad ora sono stati sviluppati per dedurre la conoscenza alcuni aspetti del modello di distribuzione che rappresenta i dati di interesse.

- Tali deduzioni spesso assumono la forma di stime numeriche, come singolo numero o intervalli di fiducia.

- Capita a volte che la conclusione da trarre non sia espressa numericamente, ma sia piuttosto la scelta tra due teorie, o ipotesi.
  - valutare se la durata di un certo tipo di cuscinetto a sfera si discosti o meno dalla durata garantita dal produttore;
  - sapere se è una perforazione a secco è più veloce o meno di una a liquido.

- Il processo di formulazione di conclusioni sulla base di dati sperimentali scegliendo tra due alternative è noto come verifica di ipotesi

Sommario della lezione 10

- Introduzione alla verifica delle ipotesi.
  - Ingredienti necessari alla formulazione di un test
  - Ingredienti necessari alla attuazione di un test
  - Errori possibili
  - Test sulla media di una VA
  - Test sulla varianza
  - Test sulla differenze tra medie.

- Analisi dell'adeguatezza della regressione
  - Verifica di ipotesi sui parametri della regressione
  - Analisi dei residui

Obiettivo

Problema:

- Sulla base dei dati di un campione possiamo prendere decisioni che riguardano la popolazione?

Esemplificazione:

- Le misure sono VA: fino a quando posso invocare il caso per dire che due campioni sono differenti pur provenendo dalla stessa popolazione?

Test statistico

- Verifica di ipotesi fatte **sulla popolazione** (sconosciuta) sulla base di un campione aleatorio

TEST DELLE IPOTESI

- Una ipotesi statistica è una assunzione circa **la distribuzione** di una VA. Per esempio: una VA ha media=20.3.

- Il test statistico di una ipotesi è una procedura in cui un campione è usato per determinare se si possa non rigettare l'ipotesi o se la si debba rigettare. Per esempio: due reattori sono uguali?

- Sorgenti di ipotesi:
  - Qualità richiesta
  - Esperienza pregressa
  - Teoria

Test delle Ipotesi - Introduzione

- In ogni test si usa un campione e cerchiamo di trarre delle conclusioni sulla popolazione corrispondente.
- Una conclusione non può mai essere completamente certa.
- Ogni test può comportare un certo rischio di errore

Test delle Ipotesi - Introduzione

Esempio

- Una fabbrica produce un tipo di catalizzatore per l'industria la cui durata di vita media è $\mu = 1200\,\text{h}$ e la deviazione standard è $\sigma = 300$ h

- Un'altra fabbrica produce (con un'altra procedura) lo stesso tipo di catalizzatore. Un campione di 100 catalizzatori ha rilevato una media di vita $\bar{y} = 1265$ ore

- È possibile che la nuova procedura produca un catalizzatore di maggior durata oppure la differenza è legata semplicemente alla natura aleatoria dei dati?

Test delle Ipotesi

- Si vuole verificare l'ipotesi che i risultati di questa sperimentazione siano esiti di VA che abbiano media $\mu = 1200$

- Questa affermazione implica l'introduzione di una ipotesi:

- IPOTESI $H_0 : \quad \mu = \mu_0 = 1200$

- L'ipotesi che tentiamo di verificare viene detta nel gergo statistico IPOTESI NULLA

Ipotesi nulla

Test delle Ipotesi

- Una altra possibilità (plausibile) è che il nuovo catalizzatore sia effettivamente più reattivo di quello della vecchia produzione

- Questa altra possibilità fornisce una IPOTESI ALTERNATIVA:

$$H_1: \quad \mu \neq \mu_0$$

Ipotesi alternativa

- Tutti i test delle ipotesi statistici richiedono la formulazione di un'ipotesi nulla e di un'ipotesi alternativa.

- Abbiamo quindi esaminato il primo ingrediente per l'effettuazione del test dell'ipotesi: LA FORMULAZIONE DI IPOTESI

- Il secondo passo è la DEFINIZIONE STATISTICA del test

Test Ipotesi

- La DEFINIZIONE STATISTICA del test identifica la distribuzione che consente di ottenere una risposta ai test
  - Normale
  - t Student
  - Chi-quadro
  - F di Fisher

Definizione statistica

- Questo passaggio ha a che fare con la natura della VA che ha generato gli esiti

Test Ipotesi

- Supponiamo di avere effettuato 10 esperimenti di misura di viscosità Newtoniana ad una fissata temperatura.

- Ipotesi nulla il liquido ha una viscosità pari a 50P.

- Ipotesi alternativa il liquido ha una viscosità diversa da 50P

- Intuitivamente: se la media del campione è prossima a 50P possiamo non rigettare l'ipotesi nulla.
  – p.e. fissiamo un intervallo [48.5, 51.5]: se la media del campione cade dentro questo intervallo non rigettiamo l'ipotesi nulla.

- Possiamo valutare la probabilità di commettere errori quando procediamo in questo modo.

Test Ipotesi – Errore del 1° e 2° tipo

- Come abbiamo già detto data la natura statistica del processo di verifica, ad esso è connesso un rischio di commettere errori.

- Si definisce ERRORE DEL 1° TIPO l'errore che si commette quando si rigetta una ipotesi nulla quando è vera.
  - Rischio del produttore

- Si definisce ERRORE DEL 2° TIPO l'errore che si commette quando si non si rigetta una ipotesi nulla quando essa è falsa.
  - Rischio del consumatore

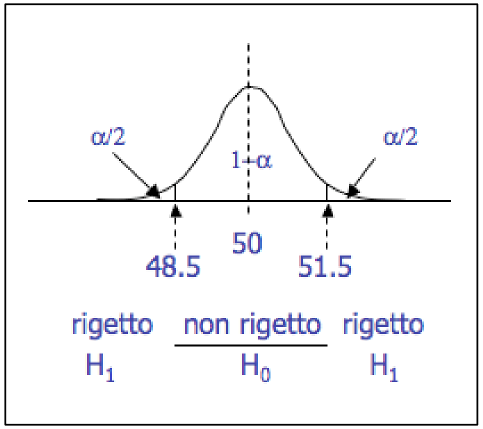
- La nostra decisione è basata su una VA, quindi possiamo valutare la probabilità di commettere i due tipi di errore.

Test Ipotesi – Errore del I° e 2° tipo

		True state of nature	
		H_0 is true	H_1 is true
Our decision on the basis of the data	Reject H_0	*Type I error*	Correct decision
	Not reject H_0	Correct decision	*Type II error*

Test Ipotesi – Errore del I° tipo

- Nell'esempio che stiamo considerando l'errore di tipo I lo commettiamo se la media del **campione osservato** cade fuori dell'intervallo [48.5, 51.5] pur essendo la popolazione genitrice a media 50.



Misure Gaussiane

Media Gaussiana

- Le due code con probabilità complessiva α rappresentano la probabilità di commettere l'errore di Tipo I

Test Ipotesi – Errore del I° tipo

- Gli esperimenti sono Gaussiani con deviazione standard sperimentale pari a 2.5P. La deviazione standard della media è 0.79

- Allora:

$$\alpha = P\left(\overline{Y} \le 48.5\right)\Big|_{\mu=50} + P\left(\overline{Y} \ge 51.5\right)\Big|_{\mu=50} = P\left(Z \le -1.90\right) + P\left(Z \ge 1.90\right) = 0.0574$$

- Questo vuole dire che la scelta che avevamo fatto dell'intervallo di accettazione genera una probabilità del 5.74% di rigettare l'ipotesi quando il valore vero della media è effettivamente 50P.

- E' semplice notare che
  - se allarghiamo l'intervallo si riduce la probabilità di commettere un errore di tipo I.
  - Se aumentiamo la dimensione del campione riduciamo la probabilità di commettere un errore di tipo I

Test Ipotesi – Errore del I° tipo

- Un modo usato da molti programmi di statistica per valutare un test relativamente all'errore del I tipo è il P-value.

- Il P-value è il più piccolo livello di significatività che porterebbe al rigetto dell'ipotesi nulla per un fissato set di dati.

- Il P-value si calcola sulla base della statistica su cui si fa il test.

- Nel caso della media con varianza nota dell'esempio sulla viscosità, la media del campione sia 51.3P. Data la simmetria della Gaussiana scegliamo come intervallo critico

$$\left[50-(51.3-50),\, 51.3\right]$$

Test Ipotesi – Errore del I° tipo

- Il P-value è calcolato come:

$$P - value = 1 - P\left(48.7 \leq \bar{Y} \leq 51.3\right)\Big|_{\mu=50} = 0.038$$

- Per cui l'ipotesi nulla sarà rigettata per ogni $\alpha>0.038$ ma non rigettata per $\alpha<0.038$.

- Un aspetto utile di questo modo di procedere è che abbiamo una idea quantitativa di quanto l'α scelto sia lontano dalla condizione critica.

Test Ipotesi – Errore del 2° tipo

- ERRORE DEL 2° TIPO l'errore che si commette quando si non si rigetta una ipotesi nulla quando essa è falsa.

- La probabilità di commettere tale errore viene spesso indicata con la lettera β.

- Per valutare la probabilità di commettere un errore di tipo II dobbiamo specificare anche l'alternativa.

- Per esempio supponiamo che sia importante rigettare l'ipotesi nulla m=50P ogni qual volta m è maggiore di 52P o minore di 48P.

- Possiamo calcolare la probabilità β usando la VA con μ=48 o con μ=52 (Data la simmetria è equivalente)

Test Ipotesi – Errore del 2° tipo

- In figura abbiamo due gaussiane una relativa alla H_0 μ=50, e l'altra alla H_1 μ=52.

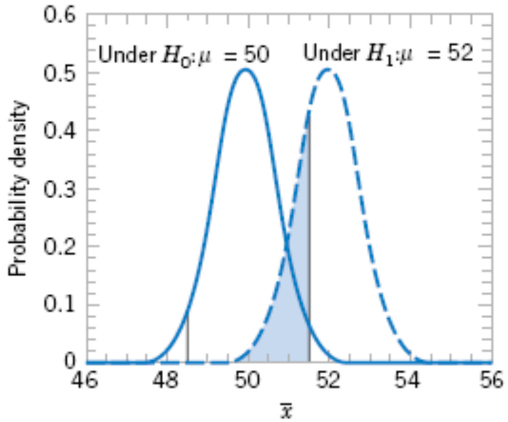


Figure 9-3 The probability of type II error when $\mu = 52$ and $n = 10$.

- L'errore di tipo II sarà commesso se la media del campione cade tra 48.5 e 51.5 (i limiti critici scelti) quando la media vera è μ=52 (area ombreggiata in figura.

Test Ipotesi – Errore del 2° tipo

- La probabilità β è quindi:

$$\beta = P\left(48.5 \le \bar{Y} \le 51.5\right)\Big|_{\mu=52} = P\left(-4.43 \le Z \le -0.63\right)$$
$$= P\left(Z \le -0.63\right) - P\left(Z \le -4.43\right) = 0.2643 - 0.000 = 0.2643$$

- Per simmetria se il valore vero è 48P b vale sempre 0.2643.

- Cosa succede a β se facciamo diminuire α? Perché?

- Se la dimensione del campione cresce diminuisce la probabilità di commettere un errore di tipo II.
  - Si può usare N per fissare β

Test Ipotesi – Potenza del test

- La **potenza (power)** di un test statistico è la probabilità di rigettare l'ipotesi nulla quando questa è falsa.

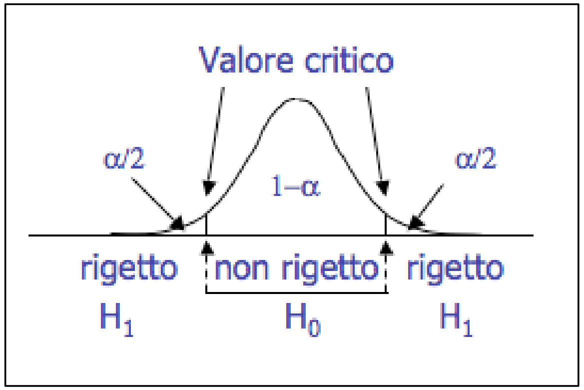
- La potenza è calcolata come 1-β.

- La potenza è una misura descrittiva e concisa della sensitività di un test statistico.

Riassumendo

- Il terzo ingrediente è rappresentato dalla REGOLA DI ACCETTAZIONE ovvero dal livello di SIGNIFICATIVITA' del test

- Dobbiamo identificare due tipi di regioni relative alla distribuzione genitrice:
  - La regione di rigetto
  - La regione di non rigetto



Il livello di significatività del test è la probabilità di rigettare l'ipotesi nulla nonostante essa sia vera

- Il **livello di significatività** α è la probabilità delle regioni di rigetto.
  - La regione di non rigetto ha probabilità $1-\alpha$.

Test Ipotesi

La fase progettuale del test delle ipotesi è completa:

1. Formulazione ipotesi
2. Definizione della statistica del test
3. Regola di accettazione

- Dobbiamo passare alla fase esecutiva

Ingredienti della fase esecutiva:

1. Raccolta dati
   - Viene condotta la campagna sperimentale che fornirà il campione
2. Decisione
   - Rigetto o non rigetto dell'ipotesi nulla

Test Ipotesi – Media

- Come già visto nella stima per intervallo le cose si differenziano se conosciamo o meno il valore esatto della varianza.

$$Z = \frac{\overline{Y} - \mu}{\sigma / \sqrt{N}} \qquad\qquad T_{n-1} = \frac{\overline{Y} - \mu}{S / \sqrt{N}}$$

- Nel secondo caso usiamo una T di Student ad N-1 gdl.

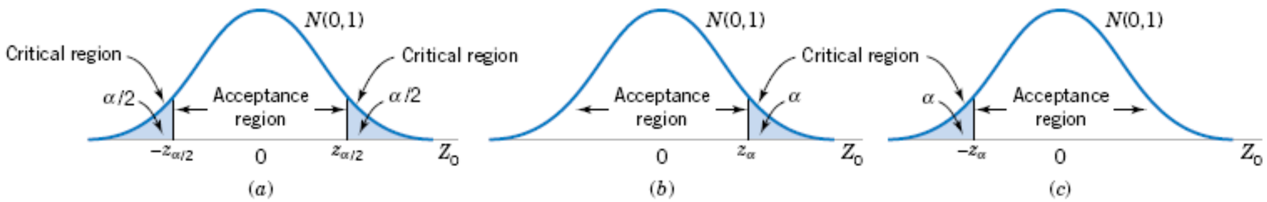


Figure 9-6 The distribution of Z_0 when H_0: $\mu = \mu_0$ is true, with critical region for (a) the two-sided alternative $H_1 : \mu \neq \mu_0$, (b) the one-sided alternative $H_1 : \mu > \mu_0$, and (c) the one-sided alternative $H_1 : \mu < \mu_0$.

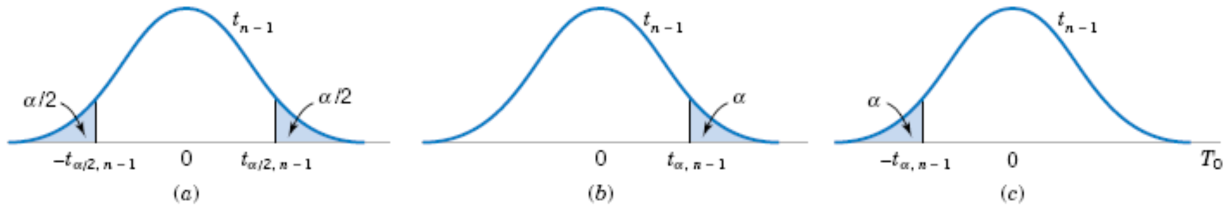


Figure 9-8 The reference distribution for H_0: $\mu = \mu_0$ with critical region for (a) H_1: $\mu \neq \mu_0$, (b) H_1: $\mu > \mu_0$, and (c) H_1: $\mu < \mu_0$.

TEST DELLE IPOTESI

- Nel seguito considereremo soltanto campioni con dimensione fissata (N)

- Ipotesi semplice: la distribuzione della VA è nota
- Ipotesi composita: la distribuzione della VA non è nota.

- Test sulla media di una Gaussiana $\mu=\mu_0$.
  - Se la varianza è nota **hp. Semplice**
  - Se la varianza non è nota **hp. Composita**

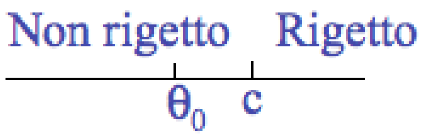
- Dobbiamo fissare un livello di significatività e ricavare un valore critico.
  - Il valore critico dipende dalla hp e dalla alternativa.

Ipotesi semplice

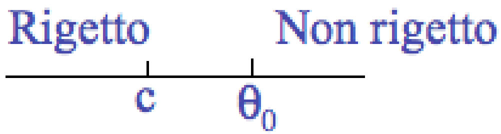
Ipotesi composita

TEST DELLE IPOTESI

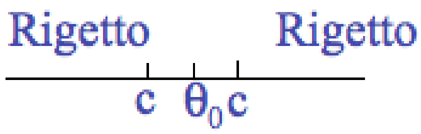
- Esempi di ipotesi e alternative:
  - H_0: $\theta=\theta_0$
  - alt. 1: $\theta>\theta_0$



  - alt. 2: $\theta<\theta_0$



  - alt. 3: $\theta\neq\theta_0$



TEST SULLA MEDIA

- Media con varianza nota:
$$\hat{\mu} = \frac{\sum y_i}{N}, \qquad \sigma^2_{\hat{\mu}} = \frac{\sigma^2}{N}$$
  - $\sigma^2=9$
  - N=10
  - H_0: $\mu=\mu_0=24$
  - H_1: $\mu>\mu_0$
- Stima della media e significatività ci fanno calcolare c:

$$P(\hat{\mu} > c)\Big|_{\mu=24} = \alpha = 0.05$$

$$P(\hat{\mu} \le c)\Big|_{\mu=24} = 1-\alpha = .95$$

$$\Phi\left(\frac{c-24}{\sqrt{0.9}}\right) = .95$$

$$c = 25.56$$

$$se\,\hat{\mu} \le c\; l'ipotesi\; nulla\; è\; non\; rigettabile$$

TEST SULLA MEDIA

- Media con varianza non nota ma stimata
  - s=115
  - N=16
  - Stima media 4482
  - significatività 0.05
  - H_0: $\mu=\mu_0=4500$
  - H_1: $\mu=4400$

$$T_{N-1} = \sqrt{N}\,\frac{\hat{\mu}-\mu}{s}$$

$$P\left(T<c\right)_{\mu=4500} = 0.05$$

$$c = -1.75$$

$$t = \sqrt{N}\,\frac{\hat{\mu}-\mu}{s} = -0.626$$

$$Non\ rigettiamo\ l'ipotesi$$

Test Ipotesi – Esempio

- Ad una raffineria è richiesta come specifica la produzione di benzina con numero di ottano 98.

- La raffineria verifica la qualità con una significatività α pari a 0.1 attraverso un campione di 10 misure e prende una decisione in merito al mettere o meno in commercio la benzina.

- Campione sperimentale:

$$y=(98.2, 97.3, 97.0, 98.6, 96.8, 97.3, 97.9, 98.8, 97.0, 96.1)$$

- Il test che la raffineria esegue è un test sulla media della popolazione

Test Ipotesi – Esempio

1. Formulazione ipotesi:
   - H_0=98 (ipotesi nulla), $H_1 \neq$98 (ipotesi alternativa)

2. Definizione della statistica del test:
   - La media del campione ha una distribuzione normale con media μ e varianza σ^2/N
   - La varianza non è nota ma stimata quindi la statistica per la media è

$$T = \frac{\overline{Y} - \mu}{S / \sqrt{N}}$$

3. Regola di accettazione: α=0.1
   - Determinazione del valore critico
     - La T è simmetrica con media zero

$$P\left(-t_{9,0.95} < T\right) = 0.05 \qquad \Rightarrow \qquad t_{9,0.95} = 1.833$$
$$P\left(T < t_{9,0.95}\right) = 0.95$$

Test Ipotesi – Esempio

4. Raccolta dati:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = 97.5, \qquad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 = 0.73, \qquad s = 0.86$$

$$t = \frac{\bar{y} - \mu}{s/\sqrt{N}} = -1.85$$

- La stima della media differisce dal valore ipotizzato per la aleatorietà del campione o perché l'ipotesi è errata?

5. Decisione

- Dobbiamo confrontare il valore di t ottenuto dai dati con il valore limite basato sulla scelta del livello di significatività.
- $t = -1.85 < -t_{9,0.95} \left(= -1.833\right)$
- L'ipotesi nulla viene rigettata
- La probabilità che l'aver misurato un campione avente numero di ottano medio pari a 97.5 estratto da una popolazione con numero di ottano medio 98 è inferiore al 5%.

TEST SULLA VARIANZA

- Test per verificare l'ipotesi sulla varianza.
  - H_0: $\sigma^2 = \sigma_0^2$
  - H_1: $\sigma^2 > \sigma_0^2$
- Nell'ipotesi nulla la distribuzione della s^2 è nota $\chi_0^2 = \dfrac{(N-1)s^2}{\sigma_0^2}$
- I limiti di accettazione sono calcolati come avevamo fatto per l'intervallo fiduciario

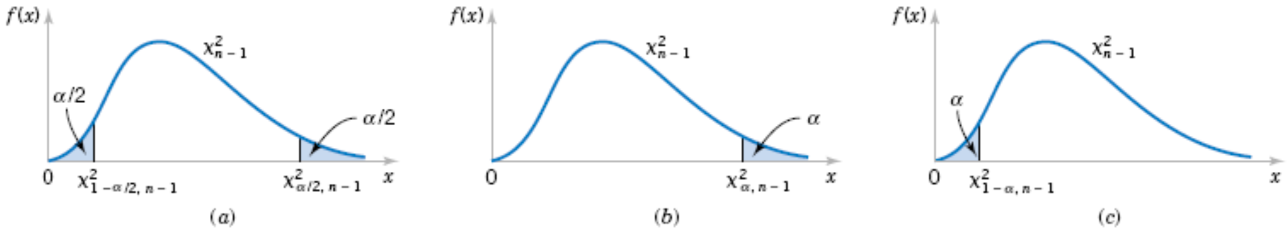


Figure 9-10 Reference distribution for the test of H_0: $\sigma^2 = \sigma_0^2$ with critical region values for (a) H_1: $\sigma^2 \neq \sigma_0^2$, (b) H_1: $\sigma^2 > \sigma_0^2$, and (c) H_1: $\sigma^2 < \sigma_0^2$.

TEST SULLA VARIANZA

PROCEDIMENTO

- Scegliamo una significatività
- Determiniamo c dalla tabella per la chi-quadro ad N-1 gdl.
- Confrontiamo c con il valore di $\dfrac{s^2(N-1)}{\sigma_0^2}$
- Se $\dfrac{s^2(N-1)}{\sigma_0^2} > c$ rigettiamo l'ipotesi
- Se $\dfrac{s^2(N-1)}{\sigma_0^2} \leq c$ non rigettiamo l'ipotesi

TEST SULLA VARIANZA

- Una macchina per il riempimento automatico di bottiglie è usata per produrre confezioni di detergente. Viene raccolto un campione di 20 bottiglie cui compete una varianza $s^2=0.0153$ lt^2. Se la varianza eccede 0.01 lt^2, la porzione di bottiglie prodotte con sovrariempimento o sottoriempimento è inaccettabile. L'evidenza sperimentale suggerisce che ci sia tale problema? Usate $\alpha=0.05$ e assumete che il volume di riempimento abbia una distribuzione Gaussiana.
  - H_0: $\sigma^2=0.01$, H_1: $\sigma^2 > 0.01$, $\alpha=0.05$
- La statistica del test: $\chi_0^2 = \dfrac{s^2(N-1)}{\sigma_0^2} = \dfrac{19\times 0.0153}{0.01} = 29.7$
- H_0 rigettata se: $\chi_0^2 > \chi^2_{0.05,19}, \quad da\, tavole\, \chi^2_{0.05,19} = 30.14$
- Non rigettiamo l'ipotesi nulla

Test sulla differenza di due medie

- Si considerino due variabili aleatorie X e Y
- Si effettuino n_1 e n_2 osservazioni delle variabili aleatorie X e Y
- In base alle suddette osservazioni si vuole testare l'eventualità che le due variabili aleatorie siano caratterizzate dalla stessa media, vale a dire

$$\mu_X = \mu_Y$$

- Le varianze di X ed Y non sono calcolate ma si assume che esse siano uguali.

Test delle ipotesi sulla differenza di due medie

- Caso A:
  - I campioni hanno le stesse dimensioni.
  - Inoltre ogni valore di un campione corrisponde precisamente ad un valore dell'altro.

- Esempio:
  - Due misure effettuate sullo stesso oggetto, ma con strumenti differenti

- In questo caso il procedimento è molto semplice.

- È sufficiente fare le differenze dei valori corrispondenti e testare l'ipotesi che la popolazione corrispondente abbia media 0.

Test delle ipotesi sulla differenza di due medie

- Caso B:
  - I due campioni sono indipendenti e non necessariamente delle stesse dimensioni.

- In questo caso le ipotesi che dobbiamo verificare sono

$$
\begin{aligned}
H_0: &\quad \mu_X = \mu_Y \\
H_1: &\quad \mu_X > \mu_Y
\end{aligned}
$$

- Ovviamente è possibile considerare anche le altre ipotesi alternative.

Test delle ipotesi sulla differenza di due medie - Teoria

- Come sappiamo le medie di n misure sperimentali sono delle variabili aleatorie:

$$\bar{X} = \frac{1}{n_1}(X_1 + \ldots + X_2) \qquad \bar{X} = N\left(\mu_X, \frac{\sigma^2}{n_1}\right)$$

$$\bar{Y} = \frac{1}{n_2}(Y_1 + \ldots + Y_2) \qquad \bar{Y} = N\left(\mu_Y, \frac{\sigma^2}{n_2}\right)$$

Test delle ipotesi sulla differenza di due medie - Teoria

- Inoltre si può dimostrare che:

$$V_1 = \frac{1}{\sigma^2}(n_1 - 1)s_X^2 \qquad \left(s_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_i - \bar{X}\right)^2 \right)$$

$$V_2 = \frac{1}{\sigma^2}(n_2 - 1)s_Y^2 \qquad \left(s_Y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(Y_i - \bar{Y}\right)^2 \right)$$

  – sono delle variabili aleatorie di tipo χ^2 a (n_1-1) e (n_2-1) gradi di libertà.

- La loro somma V_1+V_2 sarà quindi anche essa una variabile aleatoria di tipo χ^2 a (n_1+n_2-2) gradi di libertà.

Test delle ipotesi sulla differenza di due medie - Teoria

- Con qualche passaggio si può arrivare a dimostrare che la variabile aleatoria

$$T = \frac{Z}{\sqrt{\frac{V}{n_1 + n_1 - 2}}}$$

  – è una variabile aleatoria di tipo T di Student a (n_1+n_2-2)

Test delle ipotesi sulla differenza di due medie - Ricetta 1/2

1. Scegliere un livello di significatività a del test

2. Calcolare le medie dei campioni di X e Y:

$$\bar{x} = \frac{\sum x_i}{n_1} \qquad \bar{y} = \frac{\sum y_i}{n_2}$$

3. Determinare il numero c tale che: $P(T \leq c) = 1 - \alpha$
   – Dove T è la distribuzione T di Student a (n_1+n_2-2) gradi di libertà

Test delle ipotesi sulla differenza di due medie - Ricetta 2/2

4. Calcolare

$$t_0 = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}}$$

5. se t_0 è minore di c, non rigettare l'ipotesi
   - Nota bene: nel caso di $n_1=n_2$, l'espressione si riduce a

$$t_0 = \sqrt{n} \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2 + s_y^2}}$$

Test su modelli di regressione lineare

- E' importante poter stabilire l'adeguatezza di regressioni lineari. Questa operazione viene condotta costruendo test statistici sui parametri del modello.

- Supponiamo di voler verificare l'ipotesi sulla pendenza di un modello lineare $\beta_0 + \beta_1 x$

- Possiamo procedere come abbiamo appena visto sfruttando ciò che sappiamo dello stimatore per il parametro β_1.

$$H_0 : \beta_1 = \beta_{10}$$
$$H_1 : \beta_1 \neq \beta_{10}$$

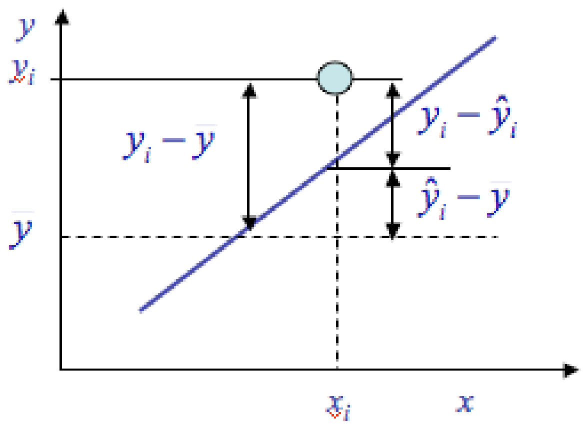
- Ricordiamo che lo stimatore per la pendenza è $\hat{\beta}_1 := N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

Test su modelli di regressione lineare

- La statistica $T = \dfrac{\hat{\beta}_1 - \beta_{10}}{S \big/ \sqrt{S_{xx}}}$ è distribuzione t di Student con n-2 gdl
- Rigetteremo l'ipotesi nulla se $|t_0| > t_{\alpha/2, n-2}$
- Un simile procedimento si può usare anche per l'intercetta

Analisi della Varianza

- Un metodo chiamato analisi della varianza può essere usato per verificare la significatività della regressione. La procedura suddivide la variabilità totale nella variabile dipendente in componenti.



$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

- $y_i - \bar{y}$: Deviazione totale del punto misurato dal punto centrale del campione
- $y_i - \hat{y}_i$: Deviazione non spiegata: residuo tra valore misurato e valore predetto
- $\hat{y}_i - \bar{y}$: Deviazione del valore predetto dal punto centrale del campione

Analisi della Varianza

- Facendo il quadrato della precedente relazione e sommando su tutti i punti sperimentali
  - I termini misti vanno a zero

$$\underbrace{\sum_{i=1}^{N}(y_i-\bar{y})^2}_{\substack{\text{Deviazione totale}\\ SS_T\\ \text{gdl } N-1}} = \underbrace{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}_{\substack{\text{Deviazione non spiegata}\\ SS_E\\ N-2}} + \underbrace{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}_{\substack{\text{Deviazione spiegata}\\ SS_R\\ 1}}$$

- SS_E divisa per i suoi gradi di libertà è una stima della variabilità sperimentale

$$E\left(\frac{SS_E}{n-2}\right)=\sigma^2$$

- Se il modello è corretto questa stima è una buona stima della varianza sperimentale.

Analisi della Varianza

- Immaginiamo di conoscere la varianza sperimentale per altra via

$$\frac{SS_E / (N-2)}{s^2} := F_{N-2,Ns}$$

- Test statistico per verificare la correttezza del modello:

$$H_0 : \; SS_E / (N-2) \leq s^2$$
$$H_1 : \; SS_E / (N-2) > s^2$$

- La variabilità nei dati non è sovrastimata

- E' un F test. Provate a "configurarlo"

Analisi della Varianza

- La verifica su β_1 si può anche condurre con un F test. La seguente VA è infatti una F di Fisher:

$$T^2 = \frac{\left(\hat{\beta}_1 - \beta_{10}\right)^2}{\sigma^2 \Big/ \left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right)} \frac{1}{\frac{1}{\sigma^2} \frac{SS_E}{(N-2)}}$$

- Ricordando che $SSR = \hat{\beta}_1^2 \left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right)$ la F appena considerata può essere riscritta come:

$$F_0 = \frac{SS_R/1}{SS_E/(N-2)}$$

  - Rappresenta il rapporto tra la "varianza" (dispersione) dei dati motivati dalla regressione e la "varianza" non spiegata dalla regressione.

Analisi della Varianza

Ricetta operativa:

- Si fissano le ipotesi: $H_0: \quad \beta_1 = 0$

  $H_1: \quad \beta_1 \neq 0$

- Si fissa un livello di significatività del test α

- Si calcola il valore c della F di Fisher a (1, n-2) gradi di libertà tale che: $P\left(F_{1,n-2} \leq c\right) = 1 - \alpha$

- Si calcola il valore di f_0: $f_0 = \dfrac{SS_R}{\left(\dfrac{SS_E}{n-2}\right)}$

- L'ipotesi è rigettata se f_0 è maggiore di c. Altrimenti l'ipotesi nulla non è rigettata

Sommario

- Introduzione alla verifica delle ipotesi.
  - Ingredienti necessari alla formulazione di un test
  - Ingredienti necessari alla attuazione di un test
  - Errori di 1° e 2° tipo

- Test sulla media di una VA
  - Con varianza nota
  - Con varianza non nota

- Test sulla varianza

- Test sulla differenze tra medie.

- Analisi dell'adeguatezza della regressione
  - Verifica di ipotesi sui parametri della regressione