Inferenza statistica parametrica: verifica d'ipotesi con due popolazioni

7 giugno 2017

Verificare se due popolazioni gaussiane hanno la stessa media oppure no

Esercizio. Vengono messe a confronto le durate di batterie (espresse in ore) di due tipologie diverse 1 e 2. Si può ritenere si tratti di campioni normali indipendenti con medie ignote μ_1 e μ_2 rispettivamente e con deviazioni standard note e pari a $\sigma_1 = 0.12$ e $\sigma_2 = 0.15$ rispettivamente. Si osservano i seguenti dati:

Tipo 1:	1.49	1.50	1.46	1.51	1.54	1.33	1.37	1.63
Tipo 2:	1.49	1.47	1.40	1.17	1.1	1.31		

1. Eseguire un test per l'ipotesi $\mathbb{H}_0$ che la durata media delle batterie di tipo 1 e quelle di tipo 2 sia uguale contro l'alternativa che sia diversa, con un livello di significatività 10%.

$X_1, \ldots, X_n$ i.i.d. $X_k \sim \mathcal{N}(\mu_X, \sigma_X^2)$ con μ_X incognita, $Y_1, \ldots, Y_m$ i.i.d. $Y_k \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ con μ_Y incognita e $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ campioni indipendenti. Consideriamo il problema di verifica d'ipotesi:

$$\mathbb{H}_0 : \mu_X = \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y = 0)$$

contro l'ipotesi alternativa

$$\mathbb{H}_1 : \mu_X \neq \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y \neq 0).$$

1. Quando le varianze σ_X^2 e σ_Y^2 sono note.

Idea per costruire il test. Ricordiamo che $\overline{X}_n$ è uno stimatore non distorto di μ_X e $\overline{Y}_m$ di μ_Y e quindi $\overline{X}_n - \overline{Y}_m$ è uno stimatore non distorto di $\mu_X - \mu_Y$. È ragionevole rifiutare $\mathbb{H}_0$ se $\overline{X}_n - \overline{Y}_m$ è lontano da 0, cioè se $|\overline{X}_n - \overline{Y}_m| \geq c$ dove, come al solito, c viene fissato in base al livello di significatività del test.

N.B.

$$\overline{X}_n - \overline{Y}_m \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

$$\Downarrow$$

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0,1)$$

$$\Downarrow$$

sotto $\mathbb{H}_0 : \mu_X - \mu_Y = 0$

$$\Downarrow$$

$$\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathcal{N}(0,1)$$

e quindi

$$P_{\mathbb{H}_0}\left(\left|\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}\right| \geq z_{\alpha/2}\right) = \alpha.$$

Perciò un test di livello di significatività α per verificare l'ipotesi nulla

$$\mathbb{H}_0 : \mu_X = \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y = 0)$$

contro l'ipotesi alternativa

$$\mathbb{H}_1 : \mu_X \neq \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y \neq 0).$$

è quello che, avendo osservato $\overline{X}_n = \bar{x}_n$ e $\overline{Y}_m = \bar{y}_n$:

$$\text{si rifiuta } \mathbb{H}_0 \text{ se } \left| \frac{\bar{x}_n - \bar{y}_n}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \right| \geq z_{\alpha/2}$$

$$\text{si accetta } \mathbb{H}_0 \text{ se } \left| \frac{\bar{x}_n - \bar{y}_n}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \right| < z_{\alpha/2}.$$

Risolvere l'esercizio iniziale

Soluzione. I dati:

Tipo 1:	1.49	1.50	1.46	1.51	1.54	1.33	1.37	1.63
Tipo 2:	1.49	1.47	1.40	1.17	1.1	1.31		

Si tratta di dati estratti da popolazioni normali indipendenti con varianze note $\sigma_1^2 = (0.12)^2$ e $\sigma_2^2 = (0.15)^2$. Le numerosità dei due campioni sono $n = 8$ ed $m = 6$. Calcoliamo $\overline{X}_n = 1.47875$ e $\overline{Y}_m = 1.32\bar{3}$. Un test di livello $\alpha = 0.1$ per
$\mathbb{H}_0$: $\mu_1 = \mu_2$ contro $\mathbb{H}_1$: $\mu_1 \neq \mu_2$ ha come regione di rifiuto

$$|Z_0| := \left| \frac{\overline{X}_n - \overline{Y}_n}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \right| \geq z_{0.05} \simeq 1.64.$$

Nel caso in questione $|Z_0| \simeq 2.09$ e l'ipotesi nulla viene rifiutata.

Si ha che p-value $\simeq 2(1 - \Phi(2.09)) \simeq 0.04$, quindi al livello 5% l'ipotesi va rifiutata mentre al livello 1% va accettata.

$X_1, \ldots, X_n$ i.i.d. $X_k \sim \mathcal{N}(\mu_X, \sigma_X^2)$ con μ_X incognita,
$Y_1, \ldots, Y_m$ i.i.d. $Y_k \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ con μ_Y incognita e
$X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ campioni indipendenti.

$\mathbb{H}_0 : \mu_X = \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y = 0)$
contro l'ipotesi alternativa
$\mathbb{H}_1 : \mu_X \neq \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y \neq 0)$.

2. Quando le varianze sono incognite ma $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

Poniamo

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{e} \quad S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2$$

e

$$\frac{(n-1)S_x^2}{\sigma^2} \sim \chi^2(n-1) \quad \text{e} \quad \frac{(m-1)S_y^2}{\sigma^2} \sim \chi^2(m-1).$$

Posto

$$S_p^2 := \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

(stimatore pooled di σ^2)

$$\Longrightarrow \qquad \frac{(n+m-2)S_p^2}{\sigma^2} = \frac{(n-1)S_x^2 + (m-1)S_y^2}{\sigma^2} \sim \chi^2(n+m-2)$$

e tenendo presente che

$$\frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim \mathcal{N}(0,1)$$

$$\Longrightarrow \qquad \frac{\overline{X}_n - \overline{Y}_m - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t(n+m-2).$$

Quindi, se $\mathbb{H}_0 : \mu_X = \mu_Y$ è vera

$$\frac{\overline{X}_n - \overline{Y}_m}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t(n+m-2)$$

$$\Longrightarrow \qquad P_{\mathbb{H}_0}\left(\left|\frac{\overline{X}_n - \overline{Y}_m}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}\right| \geq t_{\alpha/2,n+m-2}\right) = \alpha.$$

Perciò un test di livello di significatività α per verificare l'ipotesi nulla

$$\mathbb{H}_0 : \mu_X = \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y = 0)$$

contro l'ipotesi alternativa

$$\mathbb{H}_1 : \mu_X \neq \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y \neq 0).$$

è quello che, avendo osservato il valore $\overline{X}_n = \bar{x}_n$, $\overline{Y}_m = \bar{y}_n$ e $S_p^2 = s_p^2$:

$$\text{si rifiuta } \mathbb{H}_0 \text{ se } \left| \frac{\bar{x}_n - \bar{y}_n}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \right| \geq t_{\alpha/2, n+m-2}$$

$$\text{si accetta } \mathbb{H}_0 \text{ se } \left| \frac{\bar{x}_n - \bar{y}_n}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \right| < t_{\alpha/2, n+m-2}.$$

Se osservo il valore $\dfrac{\overline{X}_n - \overline{Y}_m}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} = \bar{t}$ della statistica-test allora

$$p-value = P_{\mathbb{H}_0}\left(\left|\frac{\overline{X}_n - \overline{Y}_m}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}\right| \geq |\bar{t}|\right) = P(|T_{n+m-2}| \geq |\bar{t}|)$$

dove $T_{n+m-2} \sim t(n+m-2)$ cioè ha la stessa distribuzione della statistica-test sotto l'ipotesi $\mathbb{H}_0$.

$X_1, \ldots, X_n$ i.i.d. $X_k \sim \mathcal{N}(\mu_X, \sigma_X^2)$ con μ_X incognita,
$Y_1, \ldots, Y_m$ i.i.d. $Y_k \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ con μ_Y incognita e
$X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ campioni indipendenti.

$\mathbb{H}_0 : \mu_X = \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y = 0)$
contro l'ipotesi alternativa

$\mathbb{H}_1 : \mu_X \neq \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y \neq 0)$.

3. Quando le varianze σ_X^2 e σ_Y^2 sono incognite e non si possono assumere uguali.

Sembra sensato basare il test sulla statistica

$$\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}},$$

ma questa statistica, sotto $\mathbb{H}_0$, ha una distribuzione complicata e dipendente dai parametri incogniti. Per questi motivi non può essere utilizzata in generale.

Tuttavia, se n e m sono "grandi" si può dimostrare che la sua distribuzione, sotto $\mathbb{H}_0$, può essere approssimata da una gaussiana standard. In simboli

$$\frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \approx \mathcal{N}(0,1) \qquad \text{se } \mathbb{H}_0 \text{ è vera}$$

e quindi

$$P_{\mathbb{H}_0}\left(\left| \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \right| \geq z_{\alpha/2} \right) \simeq \alpha$$

Perciò un test di livello di significatività approssimativamente uguale ad α per verificare l'ipotesi nulla

$$\mathbb{H}_0 : \mu_X = \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y = 0)$$

contro l'ipotesi alternativa

$$\mathbb{H}_1 : \mu_X \neq \mu_Y \ (\Leftrightarrow \mu_X - \mu_Y \neq 0).$$

è quello che, avendo osservato il valore $\overline{X}_n = \bar{x}_n$, $\overline{Y}_m = \bar{y}_n$ e $S_x^2 = s_x^2$ e $S_y^2 = s_y^2$:

$$\text{si rifiuta } \mathbb{H}_0 \text{ se } \left| \frac{\bar{x}_n - \bar{y}_n}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \right| \geq z_{\alpha/2}$$

$$\text{si accetta } \mathbb{H}_0 \text{ se } \left| \frac{\bar{x}_n - \bar{y}_n}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \right| < z_{\alpha/2}.$$

In modo analogo si ottengono i test unilateri.

Verificare se due popolazioni gaussiane hanno la stessa varianza oppure no

Esercizio. Due macchine diverse producono filo di rame che deve avere diametro costante. Per controllare la qualità del processo vengono eseguite misure precise del diametro in punti casuali del filo prodotto dall'una e dell'altra macchina. Le osservazioni cosí ottenute possono ritenersi provenienti da una legge normale di media e varianza incognita.

13 misure effettuate sulla prima macchina hanno fornito una varianza campionaria $s_x^2 = 0.001225$.

11 misure effettuate sulla seconda macchina hanno fornito una varianza campionaria $s_y^2 = 0.003844$.

1. Si sottoponga ad un test l'ipotesi $\mathbb{H}_0$: c'è uguaglianza tra le varianze delle due macchine contro $\mathbb{H}_1$: le varianze sono diverse ad un livello di significatività del 5%.
2. Verificare l'ipotesi $\mathbb{H}_0$: la varianza della prima macchina è maggiore o uguale di quella della seconda.

Il problema può essere formalizzato nel modo seguente: siano $X_1, \ldots, X_n$ i.i.d. $X_k \sim \mathcal{N}(\mu_X, \sigma_X^2)$ con μ_X e σ_X^2 incognite, $Y_1, \ldots, Y_m$ i.i.d. $Y_k \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ con μ_Y σ_X^2 incognite e $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ campioni indipendenti. Si consideri il problema di verifica d'ipotesi:

$\mathbb{H}_0 : \sigma_X^2 = \sigma_Y^2$

contro l'ipotesi alternativa

$\mathbb{H}_1 : \sigma_X^2 \neq \sigma_Y^2$.

Idea per costruire il test. Si considerino come al solito le varianze campionarie

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{e} \quad S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y}_m)^2$$

e abbiamo visto che

$$\frac{(n-1)S_x^2}{\sigma_X^2} \sim \chi^2(n-1) \quad \text{e} \quad \frac{(m-1)S_y^2}{\sigma_Y^2} \sim \chi^2(m-1)$$

e ovviamente sono indipendenti.

$$\Downarrow$$

$$\frac{S_x^2}{\sigma_X^2} \Big/ \frac{S_y^2}{\sigma_Y^2} \sim \mathbb{F}(n-1, m-1).$$

Quindi, sotto $\mathbb{H}_0 : \sigma_X^2 = \sigma_Y^2$,

$$\Downarrow$$

$$\frac{S_x^2}{S_y^2} \sim \mathbb{F}(n-1, m-1)$$

(distribuzione F di Fisher con $n-1, m-1$ gradi di libertà) e

$$P_{\mathbb{H}_0}\left(f_{1-\alpha/2,n-1,m-1} < \frac{S_x^2}{S_y^2} < f_{\alpha/2,n-1,m-1}\right) = 1-\alpha$$

Perciò un test di livello di significatività α è:

$$\text{si rifiuta } \mathbb{H}_0 \text{ se } \frac{s_x^2}{s_y^2} \leq f_{1-\alpha/2,n-1,m-1} \text{ oppure}$$

$$\text{se } \frac{s_x^2}{s_y^2} \geq f_{\alpha/2,n-1,m-1}$$

$$\text{si accetta } \mathbb{H}_0 \text{ se } f_{1-\alpha/2,n-1,m-1} < \frac{S_x^2}{S_y^2} < f_{\alpha/2,n-1,m-1} .$$

dove s_x^2 e s_y^2 sono i valori osservati di S_x^2 e S_y^2, rispettivamente.

Esercizio 1. Mostrare che, se $\tilde{c}$ è il valore assunto dalla statistica del test $\frac{S_x^2}{S_y^2}$ in corrispondenza dei dati, allora

$$p-value = 2 \min \left\{ P(F \leq \tilde{c}), P(F \geq \tilde{c}) = 1 - P(F \leq \tilde{c}) \right\}$$

dove $F \sim \mathbb{F}(n-1, m-1)$, cioè ha la stessa distribuzione della statistica-test sotto $\mathbb{H}_0$.

Si procede analogamente per i test unilateri.

Soluzione dell'esercizio iniziale.

La statistica-test, in corrispondenza dei dati, assume il valore

$$\frac{S_x^2}{S_y^2} = \frac{s_x^2}{s_y^2} = \frac{0.001225}{0.003844} \simeq 0.3187. \ (n = 13,\ m = 11)$$

1. Considerata $\mathbb{H}_0 : \sigma_X^2 = \sigma_Y^2$ contro $\mathbb{H}_0 : \sigma_X^2 \neq \sigma_Y^2$, rifiuto $\mathbb{H}_0$ se

$$0.3187 \simeq \frac{S_x^2}{S_y^2} \geq f_{\alpha/2,n-1,m-1} = f_{0.025,12,10} \simeq 3.62$$

oppure se

$$0.3187 \simeq \frac{S_x^2}{S_y^2} \leq f_{1-\alpha/2,n-1,m-1} = f_{0.975,12,10} = \frac{1}{f_{0.025,10,12}} \simeq 0.2967.$$

Poiché nessuna delle due disuguaglianze è verificata non posso rifiutare l'ipotesi bilatera $\mathbb{H}_0$ al 5%.

2. Rifiuto invece $\mathbb{H}_0 : \sigma_X^2 \geq \sigma_Y^2$ al 5% se

$$0.3187 \simeq \frac{S_x^2}{S_y^2} \leq f_{1-\alpha,n-1,m-1} = f_{0.95,12,10} = \frac{1}{f_{0.05,10,12}} \simeq 0.3636$$

Quindi rifiuto l'ipotesi unilatera $\mathbb{H}_0$ al 5%.

t-test per campioni di coppie di dati

Abbiamo finora considerato il caso in cui si estraggono due campioni casuali da due popolazioni gaussiane indipendenti. Talvolta capita di avere a che fare con due gruppi di n osservazioni, estratti da popolazioni gaussiane che, per il loro significato, risultano naturalmente "accoppiate".

Esempio.
1) temperature minime e massime di n città in un dato giorno;
2) consumo di n macchine prima e dopo l'installazione di un dispositivo contro l'inquinamento;
3) pesi di n persone prima e dopo una cura dimagrante.

Esercizio. Un gruppo di 6 pazienti vengono sottoposti ad una cura dimagrante. Si osservano i seguenti dati:

Prima:	77	87	104	98	91	78
Dopo:	75	88	97	99	83	70

1. C'è evidenza sperimentale, ad un livello del 5%, che la cura è stata efficace (cioè che, mediamente, c'è stato un calo di peso)?
2. Se la cura promette un calo medio di più di 5 kg, c'è evidenza sperimentale che è stata efficace?

Formalmente siamo in presenza di un campione di dimensione n di coppie, cioè

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

dove, per $i = 1, \ldots, n$, (X_i, Y_i) sono vettori aleatori indipendenti ma le componenti X_i e Y_i non possono essere considerate indipendenti tra loro.

Nell'esercizio:
X_i peso dell'individuo i prima della cura dimagrante;
Y_i peso dell'individuo i dopo la cura dimagrante
Un possibile approccio per verificare se la cura dimagrante sia efficace o no è quello di considerare come dati
$W_i = Y_i - X_i,\ i = 1, \ldots, n$.
Se la cura dimagrante non fosse efficace il peso medio dopo la cura sarebbe uguale (o maggiore o uguale) a quello prima della cura dimagrante:

$$\mu_W = \mathbb{E}(W_i) = \mathbb{E}(Y_i - X_i) = 0 \ \ (o \ \ \mu_W \geq 0)$$

Perciò siamo interessati a testare:

$\mathbb{H}_0 : \mu_W = 0 \ \ (o \ \ \mu_W \geq 0)$
contro l'ipotesi alternativa
$\mathbb{H}_1 : \mu_W < 0$.

Se i vettori del campione (X_i, Y_i) (i.i.d.) sono gaussiani, allora $W_i = Y_i - X_i$, $i = 1, \ldots, n$, sono v.a. i.i.d gaussiane con una certa media μ_W e varianza σ^2_W che assumiamo incognite.

Quindi siamo in presenza di un test sulla media di un campione gaussiano con media e varianza incognite. Quindi il t-test di livello α introdotto precedentemente fornisce la regola:

$$\text{si rifiuta } \ \mathbb{H}_0 \ \text{ se } \frac{\overline{W}_n}{\sqrt{\frac{S^2_w}{n}}} \leq -t_{\alpha,n-1}$$

$$\text{si accetta } \ \mathbb{H}_0 \ \text{ se } \frac{\overline{W}_n}{\sqrt{\frac{S^2_w}{n}}} > -t_{\alpha,n-1}.$$

dove $\overline{W}_n$ e S^2_w sono, rispettivamente, la media e la varianza campionarie delle W_i.

Soluzione dell'esercizio iniziale.

1. Vogliamo testare $\mathbb{H}_0 : \mu_W = \mu_0 = 0$ (o $\mu_W \geq 0$ cioè cura inutile) contro $\mathbb{H}_1 : \mu_W < \mu_0 = 0$.
   Con i dati a disposizione il valore della statistica test è $\frac{\overline{W}_n \sqrt{n}}{S_w} \simeq -2.156$. Si rifiuta $\mathbb{H}_0$ se il valore della statistica-test $\frac{\overline{W}_n \sqrt{n}}{S_w} < -t_{0.05,n-1} = -t_{0.05,5} \simeq -2.015$.
   Quindi, con i dati a disposizione, possiamo rifiutare $\mathbb{H}_0$ al 5%.
2. In questo caso

$$\mathbb{H}_0 : \mu_W \geq \mu_0 = -5 \text{ contro } \mathbb{H}_1 : \mu_W < \mu_0 = -5.$$

   Quindi il t-test di livello $\alpha = 0.05$ rifiuta $\mathbb{H}_0$ se il valore osservato della statistica test

$$\frac{\overline{W}_n + 5}{S_w} \sqrt{n} \simeq 0.656 < -t_{0.05,5} \simeq -2.015$$

   Quindi non possiamo rifiutare $\mathbb{H}_0$ al 5% con questi dati.

Test per i parametri di due popolazioni bernoulliane indipendenti

Consideriamo due campioni indipendenti di dimensione n_1 e n_2, rispettivamente, estratti da due popolazioni Bernoulliane con probabilità di successo p_1 e p_2, rispettivamente. Siano S_{n_1} e S_{n_2} le v.a. che contano il numero di successi nel primo e nel secondo campione, rispettivamente.
Allora sappiamo che

$$S_{n_1} \sim Bin(n_1, p_1) \quad e \quad S_{n_2} \sim Bin(n_2, p_2)$$

e sono indipendenti.
Quindi se n_1 e n_2 sono "grandi", per il teorema di De Moivre-Laplace (o il TCL)

$$\frac{S_{n_1}}{n_1} \approx \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad e \quad \frac{S_{n_2}}{n_2} \approx \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

e quindi

$$\frac{S_{n_1}}{n_1} - \frac{S_{n_2}}{n_2} \approx \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

Costruiamo un test di livello approssimato α per

$$\mathbb{H}_0 : p_1 = p_2 \text{ contro } \mathbb{H}_1 : p_1 \neq p_2.$$

Sotto $\mathbb{H}_0$

$$\frac{\frac{S_{n_1}}{n_1} - \frac{S_{n_2}}{n_2}}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \approx \mathcal{N}(0,1) \qquad (p = p_1 = p_2)$$

Ma questa non è una statistica perché contiene il parametro incognito p.

Stimiamo p usando i campione delle due popolazioni:
Sotto $\mathbb{H}_0 : p_1 = p_2$ sappiamo che $S_{n_1} + S_{n_2} \sim Bin(n_1 + n_2, p)$ e per la legge forte dei grandi numeri

$$\frac{S_{n_1} + S_{n_2}}{n_1 + n_2} \approx p \qquad \text{per } n_1 \text{ e } n_2 \text{ grandi.}$$

Quindi si prende come statistica test

$$\tilde{Z}_0 := \frac{\frac{S_{n_1}}{n_1} - \frac{S_{n_2}}{n_2}}{\sqrt{\frac{S_{n_1}+S_{n_2}}{n_1+n_2}\left(1 - \frac{S_{n_1}+S_{n_2}}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

che è una statistica basata sul campione e si può dimostrare che ancora, per n_1 e n_2 grandi, sotto $\mathbb{H}_0$

$$\tilde{Z}_0 \approx \mathcal{N}(0, 1).$$

Concludendo: la statistica del test è

$$\tilde{Z}_0 := \frac{\frac{S_{n_1}}{n_1} - \frac{S_{n_2}}{n_2}}{\sqrt{\frac{S_{n_1}+S_{n_2}}{n_1+n_2}\left(1 - \frac{S_{n_1}+S_{n_2}}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \overset{\mathbb{H}_0}{\approx} \mathcal{N}(0,1)$$

$\mathbb{H}_0 : p_1 = p_2$ vs $\mathbb{H}_1 : p_1 \neq p_2 \Rightarrow$ rifiuto $\mathbb{H}_0$ se $|\tilde{Z}_0| \geq z_{\alpha/2}$;

Analogamente

$\mathbb{H}_0 : p_1 \leq p_2$ (o $p_1 = p_2$) vs $\mathbb{H}_1 : p_1 > p_2 \Rightarrow$ rifiuto $\mathbb{H}_0$ se $\tilde{Z}_0 \geq z_\alpha$;

$\mathbb{H}_0 : p_1 \geq p_2$ (o $p_1 = p_2$) vs $\mathbb{H}_1 : p_1 < p_2 \Rightarrow$ rifiuto $\mathbb{H}_0$ se $\tilde{Z}_0 \leq -z_\alpha$

N.B.

$$\frac{S_{n_1}+S_{n_2}}{n_1+n_2} = \frac{n_1\hat{P}_1 + n_2\hat{P}_2}{n_1+n_2} = \frac{n_1}{n_1+n_2}\hat{P}_1 + \left(1 - \frac{n_1}{n_1+n_2}\right)\hat{P}_2$$

dove

$$\hat{P}_1 = \frac{S_{n_1}}{n_1} \qquad \hat{P}_2 = \frac{S_{n_2}}{n_2}$$

sono la media campionaria delle prime n_1 e delle seconde n_2 prove di Bernoulli, rispettivamente.

Esercizio. Un ufficio studi di una certa assicurazione ha constatato che nella località A, dove conta 250 automobili assicurate, vi sono stati in un certo lasso di tempo 50 furti di auto, nella località B, a fronte di 450 auto assicurate, vi sono stati 80 furti di auto. Testare al 5%

$\mathbb{H}_0$: le due località sono ugualmente pericolose
contro
$\mathbb{H}_1$: non lo sono.

ovvero, indicate con p_1 la probabilità di furto in A e p_2 quella in B

$\mathbb{H}_0$: $p_1 = p_2$
contro
$\mathbb{H}_1$: $p_1 \neq p_2$.

Soluzione.

$$\frac{S_{n_1}}{n_1} = \frac{50}{250} = 0.2 \qquad \frac{S_{n_2}}{n_2} = \frac{80}{450} = \frac{8}{45} \qquad \frac{S_{n_1} + S_{n_2}}{n_1 + n_2} = \frac{130}{700} = \frac{13}{70}.$$

Quindi

$$|\tilde{Z}_0| = \left| \frac{\frac{5}{25} - \frac{8}{45}}{\sqrt{\frac{13}{70} \times \frac{57}{70} \times (\frac{1}{250} + \frac{1}{450})}} \right| \simeq 0.6682 < z_{\alpha/2=0.025} \simeq 1.96$$

quindi non possiamo rifiutare l'ipotesi nulla di uguaglianza delle probabilità di furto al 5% con i dati a disposizione.

Esercizio. Un campione di 300 votanti nella zona A e 200 votanti nella zona B, ha mostrato che, rispettivamente, il 56% e il 48% è favorevole ad un certo candidato. Ad un livello di significatività $\alpha = 0.05$ testare

$\mathbb{H}_0$: non c'e' differenza tra le due zone
contro
$\mathbb{H}_1$: il candidato è preferito nella zona A.

Soluzione. Siano p_1 (p_2 risp.) la probabilità che un votante della zona A (zona B risp.) sia favorevole a quel candidato. Si tratta di testare al 5%

$\mathbb{H}_0$: $p_1 = p_2$
contro
$\mathbb{H}_1$: $p_1 > p_2$.

$$\frac{S_{n_1} + S_{n_2}}{n_1 + n_2} = \frac{n_1 \hat{P}_1 + n_2 \hat{P}_2}{n_1 + n_2} = \frac{0.56 \times 300 + 0.48 \times 200}{500} \simeq 0.528$$

con $\hat{P}_1 = \frac{S_{n_1}}{n_1} = 0.56$ e $\hat{P}_2 = \frac{S_{n_2}}{n_2} = 0.48$ (dati).

In corrispondenza dei dati la statistica del test assume il valore

$$\tilde{Z}_0 := \frac{\frac{S_{n_1}}{n_1} - \frac{S_{n_2}}{n_2}}{\sqrt{\frac{S_{n_1}+S_{n_2}}{n_1+n_2}\left(1-\frac{S_{n_1}+S_{n_2}}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$
$$= \frac{0.56 - 0.48}{\sqrt{0.528 \times 0.472(\frac{1}{300}+\frac{1}{200})}} \simeq 1.75$$

e per $\alpha = 0.05$

$$z_\alpha \simeq 1.645$$

quindi rifiuto $\mathbb{H}_0$ al 5% con i dati a disposizione.

$$p - value = 1 - \Phi(1.75) \simeq 1 - 0.9599 = 0.0401$$

quindi rifiuto $\mathbb{H}_0$ per $\alpha \geq 0.0401$.