Metodi statistici per la ricerca sociale

Capitolo 7. Confronto tra Due Gruppi

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM–88 SOCIOLOGIA E RICERCA SOCIALE

Confronto tra due gruppi

- Soggetti cardiopatici trattati con bypass sopravvivono più a lungo di soggetti cardiopatici non trattati con bypass?
- Passano più tempo su internet gli uomini o le donne?
- I laureati guadagnano più dei diplomati?
- Un intervento di microcredito migliora il benessere economico di famiglie che vivono in condizioni di povertà?
- Un corso di formazione professionale aiuta l'inserimento nel mondo del lavoro di giovani svantaggiati?
- Le performance di studenti con lacune in matematica migliorano dopo un corso di recupero?

Confronto tra due gruppi: Confrontare due (sotto-)popolazioni

- Obiettivo: Stabilire se due popolazioni sono uguali (o diverse)

- Confronto tra due campioni: I due campioni provengono dalla stessa popolazione ossia da popolazioni aventi un parametro caratteristico di uguale valore?

- Esiste evidenza (*significatività statistica*) che le osservazioni campionarie siano generati da due popolazioni diverse?

Variabili esplicative e variabili risposta

- Le sotto-popolazioni sono in genere definite da una variabile che prende il nome di *variabile esplicativa*

- La variabile che interessa confrontare prende il nome di *variabile risposta*

Variabili esplicative e variabili risposta: Esempi

- Un intervento di microcredito migliora il benessere economico di famiglie che vivono in condizioni di povertà?
  - ✓ Variabile Esplicativa = Microcredito (Popolazione delle famiglie che hanno accesso a un microcredito *versus* Popolazione delle famiglie che non hanno accesso a un microcredito)
  - ✓ Variabile risposta = Indicatore di benessere economico della famiglia (La famiglia vive al di sopra o al di sotto della soglia di povertà)
- Un corso di formazione professionale aiuta l'inserimento nel mondo del lavoro di giovani svantaggiati?
  - ✓ Variabile Esplicativa = Frequenza del corso di formazione (Popolazione di giovani svantaggiati che frequentano il corso di formazione *versus* Popolazione di giovani svantaggiati che non frequentano il corso di formazione)
  - ✓ Variabile risposta = Status occupazionale a sei mesi dalla fine del corso di formazione
- Esempio: Le performance di studenti con lacune in matematica migliorano dopo un corso di recupero?
  - ✓ Variabile Esplicativa = "Tempo: Prima versus dopo" (Popolazione di studenti con lacune in matematica prima del corso di recupero *versus* Popolazione di studenti con lacune in matematica prima del dopo il corso di recupero)
  - ✓ Variabile risposta = Punteggio a un test di matematica

Campioni dipendenti e indipendenti

- Confronto tra popolazioni indipendenti: le unità appartenenti a una popolazione sono indipendenti dalle unità appartenenti all'altra popolazione
- I soggetti nelle due popolazioni sono diversi e non esiste alcune forma di "appaiamento" tra i due gruppi
- Campioni indipendenti: Campioni selezionati in modo indipendente da popolazioni indipendenti
- Popolazioni dipendenti: Esiste una relazione ("accoppiamento") naturale tra ciascuna unità di una popolazione e ciascuna unità dell'altra popolazione
- Campioni dipendenti: Campioni selezionati da popolazioni dipendenti
- Studi longitudinali, Misure ripetute, Dati appaiati: Coppie di osservazioni relative a una stessa unità statistica
- Esempi di popolazioni dipendenti
  - ✓ Punteggio a un test di matematica di uno stesso studente con lacune in matematica prima e dopo il corso di recupero
  - ✓ Il volume delle vendite di una stessa azienda prima e dopo una specifica campagna pubblicitaria
  - ✓ La stessa unità sperimentale prima delle cura e dopo la cura

Confronto tra gruppi: Schema

- Confronto tra medie di variabili quantitative
  - ✓ Popolazioni: Y_1 con media $\mathbb{E}[Y_1] = \mu_1$; Y_2 con media $\mathbb{E}[Y_2] = \mu_2$
  - ✓ Parametro di interesse $\mu_D \equiv \mu_2 - \mu_1$
  - ✓ Obiettivo: Fare inferenza sulla differenza tra le medie $\mu_D = \mu_2 - \mu_1$
- Confronto tra proporzioni (medie di variabili binarie)
  - ✓ Popolazioni: $Y_1 \sim Ber(\pi_1)$ versus $Y_2 \sim Ber(\pi_2)$
  - ✓ Parametro di interesse $\pi_D \equiv \pi_2 - \pi_1$
  - ✓ Obiettivo: Fare inferenza sulla differenza tra le proporzioni $\pi_D = \pi_2 - \pi_1$

Confronto tra medie di variabili quantitative

- Variabile di interesse: Y (variabile quantitativa)
- Popolazioni: Y_1 con media $\mathbb{E}[Y_1] = \mu_1$; Y_2 con media $\mathbb{E}[Y_2] = \mu_2$
- Parametro di interesse $\mu_D \equiv \mu_2 - \mu_1$
- Obiettivo: Fare inferenza sulla differenza tra le medie $\mu_D = \mu_2 - \mu_1$
- Situazioni:
  1. Popolazioni Normali indipendenti con varianze note
  2. Popolazioni Normali indipendenti con varianze ignote ma uguali
  3. Popolazioni Normali indipendenti con varianze ignote
  4. Popolazioni qualsiasi indipendenti e campioni di grandi dimensioni
  5. Popolazioni Normali dipendenti

Confronto tra medie: Popolazioni Normali indipendenti

- Popolazioni: $Y_1 \sim N(\mu_1, \sigma_1^2)$ versus $Y_2 \sim N(\mu_2, \sigma_2^2)$ indipendenti
- Campioni casuali indipendenti:

$$(Y_{11}, \ldots, Y_{n_1 1})\, i.i.d. \quad (Y_{12}, \ldots, Y_{n_2 2})\, i.i.d. \quad \text{indipendenti}$$

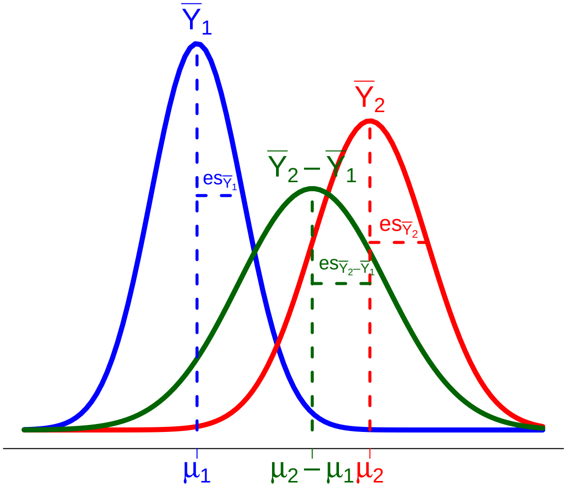
- Medie campionarie: $\overline{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{i1}$ e $\overline{Y}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{i2}$
- Stimatore per la differenza tra medie: $\overline{Y}_D = \overline{Y}_2 - \overline{Y}_1$
- Distribuzioni campionarie

$$\overline{Y}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{e} \quad \overline{Y}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Sotto l'ipotesi di indipendenza

$$\overline{Y}_2 - \overline{Y}_1 \sim N\left(\mu_2 - \mu_1, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \Longleftrightarrow \overline{Y}_D \sim N\left(\mu_D, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Confronto tra medie: Popolazioni Normali indipendenti



Confronto tra medie: Popolazioni Normali indipendenti

- Utilizzando la distribuzione campionaria della differenza tra le medie campionarie si possono costruire intervalli di confidenza e eseguire test delle ipotesi per la differenza tra medie
- Obiettivo: Valutare se $\mu_1 = \mu_2 \iff \mu_2 - \mu_1 = \mu_1 - \mu_2 = 0$
- Intervalli di confidenza:

$$\text{Stima puntuale} \pm \text{Margine di errore}$$

- Ipotesi

$$\begin{array}{lll} H_0 : \mu_1 = \mu_2 & vs & H_a : \mu_2 \neq \mu_1 \\ H_0 : \mu_1 = \mu_2 & vs & H_a : \mu_2 > \mu_1 \\ H_0 : \mu_1 = \mu_2 & vs & H_a : \mu_2 < \mu_1 \end{array} \iff \begin{array}{lll} H_0 : \mu_D = 0 & vs & H_a : \mu_D \neq 0 \\ H_0 : \mu_D = 0 & vs & H_a : \mu_D > 0 \\ H_0 : \mu_D = 0 & vs & H_a : \mu_D < 0 \end{array}$$

Inferenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze note

- Popolazioni: $Y_1 \sim N(\mu_1, \sigma_1^2)$ versus $Y_2 \sim N(\mu_2, \sigma_2^2)$ indipendenti, σ_1^2 e σ_2^2 note
- Campioni casuali indipendenti:

$$(Y_{11}, \ldots, Y_{n_1 1})\, i.i.d. \quad (Y_{12}, \ldots, Y_{n_2 2})\, i.i.d. \quad \text{indipendenti}$$

- Stimatore della differenza tra le medie: $\overline{Y}_2 - \overline{Y}_1$
- Errore standard dello stimatore della differenza tra le medie

$$es(\overline{Y}_2 - \overline{Y}_1) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Distribuzione campionaria dello stimatore della differenza tra le medie

$$\overline{Y}_2 - \overline{Y}_1 \sim N\left(\mu_2 - \mu_1, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \text{ quindi } \frac{(\overline{Y}_2 - \overline{Y}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Intervallo di confidenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze note

Intervallo di confidenza al livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\mu_2 - \mu_1) =$$
$$\left[(\overline{y}_2 - \overline{y}_1) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; (\overline{y}_2 - \overline{y}_1) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Intervallo di confidenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze note: Esempio

- Obiettivo: Confrontare il peso medio (in grammi) di neonati nati da madri non fumatrici (popolazione 1) il peso medio (in grammi) di neonati nati da madri fumatrici (popolazione 2)
- Varianze note: $\sigma_1^2 = \sigma_{\bar{F}}^2 = 50850$ e $\sigma_2^2 = \sigma_F^2 = 60025$
- Sintesi dei dati campionari:

Madre fumatrice:	$n_F = 100$	$\overline{y}_F = 2150$
Madre non fumatrice:	$n_{\bar{F}} = 120$	$\overline{y}_{\bar{F}} = 3375$

- Differenza tra medie campionarie e suo errore standard

$$\overline{y}_2 - \overline{y}_1 = \overline{y}_F - \overline{y}_{\bar{F}} = -1225 \quad \text{e} \quad es(\overline{Y}_F - \overline{Y}_{\bar{F}}) = \sqrt{\frac{50850}{120} + \frac{60025}{100}} = 32$$

- Livello di confidenza: $1 - \alpha = 0.95 \Longrightarrow z_{\alpha/2} = 1.96$
- Intervallo di confidenza

$$IC_{1-\alpha}(\mu_F - \mu_{\bar{F}}) = [-1225 - 1.96 \cdot 32; -1225 + 1.96 \cdot 32] = [-1287.719; -1162.281]$$

- L'intervallo contiene solo valori negativi, quindi si ha evidenza al livello di confidenza del 95% che il peso medio di neonati da madri fumatrici sia inferiore all peso medio di neonati da madri non fumatrici

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze note

- Popolazioni: $Y_1 \sim N(\mu_1, \sigma_1^2)$ versus $Y_2 \sim N(\mu_2, \sigma_2^2)$ indipendenti, σ_1^2 e σ_2^2 note
- Campioni casuali indipendenti:

$$(Y_{11}, \ldots, Y_{n_1 1})\, i.i.d. \quad (Y_{12}, \ldots, Y_{n_2 2})\, i.i.d. \quad \text{indipendenti}$$

- Ipotesi nulla: $H_0 : \mu_2 = \mu_1$ ossia $H_0 : \mu_D = 0$
- Statistica test

$$\overline{Y}_2 - \overline{Y}_1 \quad \longmapsto \quad Z = \frac{(\overline{Y}_2 - \overline{Y}_1) - 0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

- Sotto l'ipotesi nulla

$$\overline{Y}_2 - \overline{Y}_1 \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \Longrightarrow Z \sim N(0, 1)$$

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze note

- Livello di significatività del test = α
- Regione Critica (Regione di Rifiuto)

Ipotesi alternativa	Regione di Rifiuto
$H_a : \mu_2 \neq \mu_1$	$Z \leq -z_{\alpha/2}$ oppure $Z \geq z_{\alpha/2}$
$H_a : \mu_2 > \mu_1$	$Z \geq z_{\alpha}$
$H_a : \mu_2 < \mu_1$	$Z \leq -z_{\alpha}$

- P-valore: Sia z^{oss} il valore osservato della statistica test Z

Ipotesi alternativa	P-valore
$H_a : \mu_2 - \mu_1 \neq 0$	$P(Z < -\lvert z_{oss}\rvert \text{ oppure } Z > \lvert z_{oss}\rvert; H_0) = 2 \cdot (1 - P(Z \leq \lvert z_{oss}\rvert; H_0))$
$H_a : \mu_2 - \mu_1 > 0$	$P(Z \geq z_{oss}; H_0) = (1 - P(Z \leq z_{oss}; H_0))$
$H_a : \mu_2 - \mu_1 < 0$	$P(Z \leq z_{oss}; H_0)$

Nota bene. Il p-valore è calcolato usando la distribuzione Normale standard (la distribuzione campionaria di Z sotto H_0)

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze note: Esempio I

- Obiettivo: Confrontare se il peso medio (in grammi) di neonati nati da madri non fumatrici (popolazione 1) è uguale al peso medio (in grammi) di neonati nati da madri fumatrici (popolazione 2)

$$H_0 : \mu_F = \mu_{\bar{F}} \quad vs \quad H_a : \mu_F \neq \mu_{\bar{F}} \Longleftrightarrow H_0 : \mu_D = 0 \quad vs \quad H_a : \mu_D \neq 0$$

- Varianze note: $\sigma_1^2 = \sigma_{\bar{F}}^2 = 50850$ e $\sigma_2^2 = \sigma_F^2 = 60025$
- Sintesi dei dati campionari:

Madre fumatrice:	$n_F = 100$	$\bar{y}_F = 2150$
Madre non fumatrice:	$n_{\bar{F}} = 120$	$\bar{y}_{\bar{F}} = 3375$

- Differenza tra medie campionarie e suo errore standard

$$\bar{y}_2 - \bar{y}_1 = \bar{y}_F - \bar{y}_{\bar{F}} = -1225 \quad \text{e} \quad es(\overline{Y}_F - \overline{Y}_{\bar{F}}) = \sqrt{\frac{50850}{120} + \frac{60025}{100}} = 32$$

- Livello di significatività del test: $\alpha = 0.05$
- Regione Critica (Regione di Rifiuto): $\alpha = 0.05 \Longrightarrow \alpha/2 = 0.025 \Longrightarrow z_{\alpha/2} = 1.96$
  Quindi

$$RC_{0.05} = Z \leq -1.96 \quad \text{oppure} \quad Z \geq 1.96$$

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze note: Esempio I

- Valore osservato della statistica test

$$
\begin{aligned}
z_{oss} &= \frac{\bar{y}_F - \bar{y}_{\bar{F}}}{\sqrt{\dfrac{\sigma^2_F}{n_F} + \dfrac{\sigma^2_{\bar{F}}}{n_{\bar{F}}}}} = \frac{2150 - 3375}{\sqrt{\dfrac{50850}{120} + \dfrac{60025}{100}}} \\
&= \frac{-1225}{\sqrt{423.75 + 600.25}} = \frac{-1225}{\sqrt{1024}} = \frac{-1225}{32} = -38.28
\end{aligned}
$$

- Decisione: Il valore osservato della statistica test appartiene alla regione di rifiuto: $z_{oss} = -38.28 < -1.96 = -z_{0.025} \Longrightarrow$ I dati mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 5%
- P-valore

$$
\begin{aligned}
P &= 2 \cdot (1 - P(Z \leq |-38.28|; H_0)) = 2 \cdot (1 - P(Z \leq 38.28; H_0)) \\
&= 2 \cdot (1 - 1) = 0.000
\end{aligned}
$$

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze note: Esempio II

- Obiettivo: Valutare se un nuovo farmaco contro l'ipertensione riduce la pressione sanguigna
- Popolazione 1 = Soggetti ipertesi a cui è somministrato il farmaco standard
- Popolazione 2 = Soggetti ipertesi a cui è somministrato il nuovo farmaco

$$H_0 : \mu_2 = \mu_1 \quad vs \quad H_a : \mu_2 < \mu_1 \Longleftrightarrow H_0 : \mu_D = 0 \quad vs \quad H_a : \mu_D < 0$$

- Varianze note: $\sigma_1^2 = 420$ e $\sigma_2^2 = 420$
- Campioni:

u_{i1}	1	2	3	4	5	6	7	8	9	10	11	12
y_{i1}	178	153	186	118	178	174	162	175	158	178	133	157

u_{i2}	1	2	3	4	5	6	7	8	9	10
y_{i2}	117	166	118	168	153	115	140	150	133	156

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze note: Esempio II

- Livello di significatività del test: $\alpha = 0.01$
- Regione Critica (Regione di Rifiuto): $z_{0.01} = 2.33 \Longrightarrow RC_{0.01} = Z \leq -2.33$
- Valore osservato della statistica test

$$\bar{y}_1 = \frac{1950}{12} = 162.5 \qquad \bar{y}_2 = \frac{1416}{10} = 141.6$$

$$\begin{aligned} z_{oss} &= \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{141.6 - 162.5}{\sqrt{\frac{420}{12} + \frac{420}{10}}} \\ &= \frac{20.9}{\sqrt{35 + 42}} = \frac{-20.9}{\sqrt{77}} = \frac{-20.9}{8.775} = -2.38 \end{aligned}$$

- Decisione: Il valore osservato della statistica test appartiene alla regione di rifiuto: $z_{oss} = -2.38 < -2.33 = z_{0.01} \Longrightarrow$ I dati mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 1%
- P-valore: $P = P(Z \leq -2.38; H_0) = 1 - P(Z \leq 2.38; H_0) = 1 - 0.9911385 = 0.008615$

Inferenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali

- Popolazioni: $Y_1 \sim N(\mu_1, \sigma_1^2)$ versus $Y_2 \sim N(\mu_2, \sigma_2^2)$ indipendenti, σ_1^2 e σ_2^2 ignote ma $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$
- Campioni casuali indipendenti:

$$(Y_{11}, \dots, Y_{n_1 1})\, i.i.d. \quad (Y_{12}, \dots, Y_{n_2 2})\, i.i.d. \quad \text{indipendenti}$$

- Stimatore della differenza tra le medie: $\overline{Y}_2 - \overline{Y}_1$
- Errore standard dello stimatore della differenza tra le medie

$$es(\overline{Y}_2 - \overline{Y}_1) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

- Distribuzione campionaria dello stimatore della differenza tra le medie

$$\overline{Y}_2 - \overline{Y}_1 \sim N\left(\mu_2 - \mu_1, \sigma^2 \left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right) \text{ quindi } \frac{(\overline{Y}_2 - \overline{Y}_1) - (\mu_2 - \mu_1)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1)$$

- Problema: la varianza σ^2 non è nota

- Stimatore congiunto (pooled) della varianza: Media ponderata degli stimatori corretti di σ_1^2 e σ_2^2 con pesi proporzionali alla dimensione dei due campioni

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(Y_{i1} - \overline{Y}_1\right)^2 \qquad e \qquad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(Y_{i2} - \overline{Y}_2\right)^2$$

Quindi

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

- Statistica T

$$T = \frac{\left(\overline{Y}_2 - \overline{Y}_1\right) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{\left(\overline{Y}_2 - \overline{Y}_1\right) - (\mu_2 - \mu_1)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

Intervallo di confidenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali

Intervallo di confidenza al livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\mu_2 - \mu_1) = \left[(\bar{y}_2 - \bar{y}_1) - t_{(n_1+n_2-2),\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right.$$

$$\left. (\bar{y}_2 - \bar{y}_1) + t_{(n_1+n_2-2),\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right]$$

Intervallo di confidenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali: Esempio

- Obiettivo: Confrontare il voto medio alla fine del primo anno di studenti di istituti professionali in classi in cui si sono adottati metodi di insegnamento standard (popolazione 1) e in classi in cui si sono adottati metodi di interattivi (popolazione 2)

Metodo di insegnamento standard	Metodo di insegnamento interattivo
6.6	6.5
6.2	7.2
5.7	6.7
6.1	
5.4	

- Assunzione: Campioni estratti da popolazioni normali (indipendenti) di uguale varianza (incognita)
  - ✓ Ipotesi di normalità delle distribuzioni del voto medio
  - ✓ Ipotesi di uguale varianza (omoschedasticità) del voto medio nelle due popolazioni di classi

Intervallo di confidenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali: Esempio

- Stima della differenza tra le medie

$$\overline{y}_2 - \overline{y}_1 = \frac{20.4}{3} - \frac{30}{5} = 6.8 - 6 = 0.8$$

- Stime delle varianze

$$s_1^2 = \frac{0.86}{5-1} = 0.215 \qquad \text{e} \qquad s_2^2 = \frac{0.26}{3-1} = 0.13$$

- Stima della varianza comune

$$s_p^2 = \frac{4 \cdot 0.215 + 2 \cdot 0.13}{5+3-2} = \frac{0.86 + 0.26}{6} = \frac{1.12}{6} = 0.187$$

- Stima dell'errore standard dello stimatore della differenza tra le medie

$$\hat{es}(\overline{Y}_2 - \overline{Y}_1) = \sqrt{0.187\left(\frac{1}{5} + \frac{1}{3}\right)} = \sqrt{0.0996} = 0.3155$$

Intervallo di confidenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali: Esempio

- Livello di confidenza: $1 - \alpha = 0.99$

$$\alpha = 0.01 \Longrightarrow t_{0.005,6} = 3.707$$

- Intervallo di confidenza

$$IC_{0.99}(\mu_2 - \mu_1) =$$
$$\left[0.8 - 3.707 \cdot \sqrt{0.187\left(\frac{1}{5} + \frac{1}{3}\right)}; 0.8 + 3.707 \cdot \sqrt{0.187\left(\frac{1}{5} + \frac{1}{3}\right)}\right] =$$
$$[0.8 - 3.707 \cdot 0.3155; 0.8 + 3.707 \cdot 0.3155] =$$
$$[-1.970; 0.370]$$

- Si noti che l'intervallo di confidenza contiene il valore zero, non escludendo quindi la possibilità che le due medie siano uguali al livello di confidenza del 99%

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali

- Popolazioni: $Y_1 \sim N(\mu_1, \sigma_1^2)$ versus $Y_2 \sim N(\mu_2, \sigma_2^2)$ indipendenti, σ_1^2 e σ_2^2 ignote ma $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$
- Ipotesi nulla: $H_0 : \mu_1 = \mu_2$ ossia $H_0 : \mu_D = 0$
- Campioni: $(Y_{11}, \ldots, Y_{1n_1})\ i.i.d. \quad (Y_{21}, \ldots, Y_{2n_2})\ i.i.d.$ indipendenti
- Statistica test

$$\overline{Y}_2 - \overline{Y}_1 \quad \longmapsto \quad T = \frac{(\overline{Y}_2 - \overline{Y}_1) - 0}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{\overline{Y}_2 - \overline{Y}_1}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- Sotto l'ipotesi nulla $T \sim t_{n_1+n_2-2}$

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali

- Livello di significatività del test = α
- Regione Critica (Regione di Rifiuto)

Ipotesi alternativa	Regione di Rifiuto
$H_a : \mu_2 \neq \mu_1$	$T \leq -t_{(n_1+n_2-2),\alpha/2}$ oppure $T \geq t_{(n_1+n_2-2),\alpha/2}$
$H_a : \mu_2 > \mu_1$	$T \geq t_{(n_1+n_2-2),\alpha}$
$H_a : \mu_2 < \mu_1$	$T \leq -t_{(n_1+n_2-2),\alpha}$

- P-valore: Sia t^{oss} il valore osservato della statistica test T

Ipotesi alternativa	P-valore
$H_a : \mu_2 - \mu_1 \neq 0$	$P(T < -\vert t_{oss}\vert$ oppure $T > \vert t_{oss}\vert; H_0)$
	$= 2 \cdot (1 - P(T \leq \vert t_{oss}\vert; H_0))$
$H_a : \mu_2 - \mu_1 > 0$	$P(T \geq t_{oss}; H_0) = (1 - P(T \leq t_{oss}; H_0))$
$H_a : \mu_2 - \mu_1 < 0$	$P(T \leq t_{oss}; H_0)$

Nota bene. Il p-valore è calcolato usando la distribuzione t-Student con $n_1 + n_2 - 2$ gdl (la distribuzione campionaria di T sotto H_0)

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali: Esempio I

- Obiettivo: Confrontare il voto medio alla fine del primo anno di studenti di istituti professionali in classi in cui si sono adottati metodi di insegnamento standard (popolazione 1) e in classi in cui si sono adottati metodi di interattivi (popolazione 2)

Metodo di insegnamento standard	Metodo di insegnamento interattivo
6.6	6.5
6.2	7.2
5.7	6.7
6.1	
5.4	

- Ipotesi

$$H_0 : \mu_2 = \mu_1 \quad vs \quad H_a : \mu_2 \neq \mu_1 \Longleftrightarrow H_0 : \mu_D = 0 \quad vs \quad H_a : \mu_D \neq 0$$

- Livello di significatività del test: $\alpha = 0.01$
- Regione Critica (Regione di Rifiuto)

$$\alpha = 0.01 \Longrightarrow t_{0.005,6} = 3.707 \Longrightarrow RC_{0.01} = T \leq -3.707 \text{ oppure } T \geq 3.707$$

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali: Esempio I

- Stima della differenza tra le medie
$$\overline{y}_2 - \overline{y}_1 = \frac{20.4}{3} - \frac{30}{5} = 6.8 - 6 = 0.8$$
- Stima della varianza
$$s_1^2 = \frac{0.86}{5-1} = 0.215 \qquad \text{e} \qquad s_1^2 = \frac{0.26}{3-1} = 0.13$$
Quindi
$$s_p^2 = \frac{4 \cdot 0.215 + 2 \cdot 0.13}{5+3-2} = \frac{0.86 + 0.26}{6} = \frac{1.12}{6} = 0.187$$
- Valore osservato della statistica test
$$t_{oss} = \frac{\overline{y}_2 - \overline{y}_1}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.8}{\sqrt{0.187\left(\frac{1}{5} + \frac{1}{3}\right)}} = \frac{0.8}{\sqrt{0.0996}} = \frac{0.8}{0.3155} = 2.535$$
- Decisione: Il valore osservato della statistica test NON appartiene alla regione di rifiuto: $-t_{0.005,6} = -3.707 < t_{oss} = 2.535 < 3.707 = t_{0.005,6}$. I dati NON mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 1%
- P−valore = 0.04435 (p−valore non estremamente piccolo)

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali: Esempio II

- Obiettivo: Valutare se un nuovo farmaco contro l'ipertensione riduce la pressione sanguigna
- Popolazione 1 = Soggetti ipertesi a cui è somministrato il farmaco standard
- Popolazione 2 = Soggetti ipertesi a cui è somministrato il nuovo farmaco

$$H_0: \mu_2 = \mu_1 \quad vs \quad H_a: \mu_2 < \mu_1 \Longleftrightarrow H_0: \mu_D = 0 \quad vs \quad H_a: \mu_D < 0$$

- Dimensioni campionarie: $n_1 = 12$ e $n_2 = 10$
- Livello di significatività del test: $\alpha = 0.05$
- Regione Critica (Regione di Rifiuto)

$$\alpha = 0.01 \Longrightarrow t_{0.05,10+12-2} = t_{0.05,20} = 1.72 \Longrightarrow RC_{0.05} = T \leq -1.72$$

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali: Esempio II

u_{i1}	y_{i1}	$(y_{i1} - \overline{y}_1)^2$
1	178	240.25
2	153	90.25
3	186	552.25
4	118	1980.25
5	178	240.25
6	174	132.25
7	162	0.25
8	175	156.25
9	158	20.25
10	178	240.25
11	133	870.25
12	157	30.25

u_{i2}	y_{i2}	$(y_{i2} - \overline{y}_2)^2$
1	117	605.16
2	166	595.36
3	118	556.96
4	168	696.96
5	153	129.96
6	115	707.56
7	140	2.56
8	150	70.56
9	133	73.96
10	156	207.36

$$\overline{y}_1 = \frac{1950}{12} = 162.5$$

$$s_1^2 = \frac{4553}{12-1} = 413.91$$

$$\overline{y}_2 = \frac{1416}{10} = 141.6$$

$$s_2^2 = \frac{3646.4}{10-1} = 405.16$$

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze ignote ma uguali: Esempio II

- Stima pooled della varianza

$$\begin{aligned} s_p^2 &= \frac{11 \cdot 413.91 + 9 \cdot 405.16}{10 + 12 - 2} \\ &= \frac{4553 + 3646.4}{20} = \frac{8199.4}{20} = 409.97 \end{aligned}$$

- Valore osservato della statistica test

$$\begin{aligned} t_{oss} &= \frac{\overline{y}_2 - \overline{y}_1}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{141.6 - 162.5}{\sqrt{409.97\left(\frac{1}{10} + \frac{1}{12}\right)}} \\ &= \frac{-20.9}{\sqrt{75.16}} = \frac{-20.9}{8.67} = -2.41 \end{aligned}$$

- Decisione: Il valore osservato della statistica test appartiene alla regione di rifiuto: $t_{oss} = -2.41 < -1.72 = -t_{0.05,20} \Longrightarrow$ I dati mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 5%
- P−valore = 0.01283 (p−valore piccolo ma non estremamente piccolo)

Inferenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze non note

- Popolazioni: $Y_1 \sim N(\mu_1, \sigma_1^2)$ versus $Y_2 \sim N(\mu_2, \sigma_2^2)$ indipendenti, σ_1^2 e σ_2^2 ignote
- Campioni casuali indipendenti:

$$(Y_{11}, \ldots, Y_{n_1 1})\, i.i.d. \quad (Y_{12}, \ldots, Y_{n_2 2})\, i.i.d. \quad \text{indipendenti}$$

- Stimatore della differenza tra le medie: $\overline{Y}_2 - \overline{Y}_1$
- Distribuzione campionaria: $\overline{Y}_2 - \overline{Y}_1 \sim N\left(\mu_2 - \mu_1; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
- Stimatori delle varianze

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(Y_{i1} - \overline{Y}_1\right)^2 \qquad e \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(Y_{i2} - \overline{Y}_2\right)^2$$

- Statistica

$$\frac{\left(\overline{Y}_2 - \overline{Y}_1\right) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Inferenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze non note

- Distribuzione campionaria della statistica

$$\frac{(\overline{Y}_2 - \overline{Y}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{gdl}$$

- I *gdl* dipendono dalle stime delle deviazioni standard s_1^2 e s_2^2 e dalle numerosità campionarie n_1 e n_2
  - ✓ Approssimazione Welch-Satterthwaite (complessa)
  - ✓ Se $s_1^2 = s_2^2$ e $n_1 = n_2$ allora $gdl = n_1 + n_2 - 2$
  - ✓ $\min\{(n_1 - 1), (n_2 - 1)\} \leq gdl \leq (n_1 + n_2 - 2)$
- Se n_1 e n_2 sono sufficientemente grandi (almeno 30 ciascuno):

$$\frac{(\overline{Y}_2 - \overline{Y}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{gdl} \approx N(0, 1)$$

Intervallo di confidenza per la differenza tra le medie
Popolazioni Normali indipendenti con varianze non note

Intervallo di confidenza (approssimato) al livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\mu_2 - \mu_1) \approx \left[(\overline{y}_2 - \overline{y}_1) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; (\overline{y}_2 - \overline{y}_1) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

Nota bene. L'approssimazione è adeguata se n_1 e n_2 sono sufficientemente grandi

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze non note

- Popolazioni: $Y_1 \sim N(\mu_1, \sigma_1^2)$ versus $Y_2 \sim N(\mu_2, \sigma_2^2)$ indipendenti, σ_1^2 e σ_2^2 ignote
- Campioni casuali indipendenti:

$$(Y_{11}, \ldots, Y_{n_1 1})\, i.i.d. \quad (Y_{12}, \ldots, Y_{n_2 2})\, i.i.d. \quad \text{indipendenti}$$

- Stimatore della differenza tra le medie: $\overline{Y}_2 - \overline{Y}_1$
- Stimatori delle varianze

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(Y_{i1} - \overline{Y}_1\right)^2 \qquad e \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(Y_{i2} - \overline{Y}_2\right)^2$$

- Statistica test e sua distribuzione campionaria sotto $H_0 : \mu_2 - \mu_1 = 0$

$$Z = \frac{(\overline{Y}_2 - \overline{Y}_1) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{gdl} \text{ sotto l'ipotesi nulla}$$

con *gdl* difficili da calcolare

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze non note

- Se n_1 e n_2 sono sufficientemente grandi, sotto l'ipotesi nulla, $H_0 : \mu_2 - \mu_1 = 0$

$$Z = \frac{(\overline{Y}_2 - \overline{Y}_1) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{gdl} \approx N(0, 1)$$

Nota bene. L'approssimazione è adeguata se n_1 e n_2 sono sufficientemente grandi

Test per la differenza tra le medie
Popolazioni Normali indipendenti con varianze non note

- Livello di significatività del test = α
- Regione di rifiuto (Approssimata)

Ipotesi alternativa	Regione di Rifiuto		
$H_a : \mu_2 \neq \mu_1$	$Z \leq -z_{\alpha/2}$	oppure	$Z \geq z_{\alpha/2}$
$H_a : \mu_2 > \mu_1$		$Z \geq z_\alpha$	
$H_a : \mu_2 < \mu_1$		$Z \leq -z_\alpha$	

- P-valore (approssimato): Sia z^{oss} il valore osservato della statistica test Z

Ipotesi alternativa	P-valore
$H_a : \mu_2 - \mu_1 \neq 0$	$P(Z < -\lvert z_{oss}\rvert \text{ oppure } Z > \lvert z_{oss}\rvert; H_0)$ $= 2 \cdot (1 - P(Z \leq \lvert z_{oss}\rvert; H_0))$
$H_a : \mu_2 - \mu_1 > 0$	$P(Z \geq z_{oss}; H_0) = (1 - P(Z \leq z_{oss}; H_0))$
$H_a : \mu_2 - \mu_1 < 0$	$P(Z \leq z_{oss}; H_0)$

Nota bene. Il p-valore è calcolato usando la distribuzione Normale standard (la distribuzione campionaria approssimata di Z sotto H_0)

Inferenza per la differenza tra le medie
Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni

- Popolazioni:

$$Y_1 : \mathbb{E}(Y_1) = \mu_1; var(Y_1) = \sigma_1^2 \quad versus \quad Y_2 : \mathbb{E}(Y_2) = \mu_2; var(Y_2) = \sigma_2^2$$

  Y_1 e Y_2 indipendenti, σ_1^2 e σ_2^2 ignote

- Campioni casuali indipendenti:

$$(Y_{11}, \ldots, Y_{n_1 1})\ i.i.d. \quad (Y_{12}, \ldots, Y_{n_2 2})\ i.i.d. \quad \text{indipendenti}$$

- Stimatore della differenza tra le medie: $\overline{Y}_2 - \overline{Y}_1$
- Stimatori delle varianze

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{i1} - \overline{Y}_1)^2 \qquad e \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_{i2} - \overline{Y}_2)^2$$

Inferenza per la differenza tra le medie
Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni

- Statistica
$$\frac{\left(\overline{Y}_2 - \overline{Y}_1\right) - (\mu_2 - \mu_1)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$
- Distribuzione campionaria (approssimata) della statistica. Se n_1 e n_2 sono sufficientemente grandi:
$$\frac{\left(\overline{Y}_2 - \overline{Y}_1\right) - (\mu_2 - \mu_1)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \approx N(0, 1)$$

Intervallo di confidenza per la differenza tra le medie
Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni

Intervallo di confidenza (approssimato) al livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\mu_2 - \mu_1) \approx \left[(\overline{y}_2 - \overline{y}_1) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; (\overline{y}_2 - \overline{y}_1) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

Nota bene. L'approssimazione è adeguata se n_1 e n_2 sono sufficientemente grandi

Intervallo di confidenza per la differenza tra le medie

Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni: Esempio

- Obiettivo: Confrontare il livello di dolore (misurato attraverso la scala visuo-analogica del dolore) a 4 ore dalla fine dell'operazione tra due gruppi di pazienti sottoposti a un'operazione chirurgica all'addome: pazienti trattati preliminarmente con morfina e pazienti trattati preliminarmente con un placebo
- Sintesi dei dati campionari:

$$\begin{array}{llll} \text{Gruppo placebo:} & n_1 = 200 & \overline{y}_1 = 45.4 & s_1^2 = 333.2 \\ \text{Gruppo morfina:} & n_2 = 120 & \overline{y}_2 = 28.0 & s_2^2 = 162.7 \end{array}$$

- Intervallo di confidenza al livello di confidenza $1 - \alpha = 0.95$ (approssimato)

$$IC_{0.95}(\mu_2 - \mu_1) \approx \left[-17.4 - 1.96\sqrt{\frac{333.2}{200} + \frac{162.7}{120}}; -17.4 + 1.96\sqrt{\frac{333.2}{200} + \frac{162.7}{120}}\right]$$
$$= [-20.81; -13.99]$$

Test per la differenza tra le medie
Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni

- Popolazioni:

  $$Y_1 : \mathbb{E}(Y_1) = \mu_1; var(Y_1) = \sigma_1^2 \quad versus \quad Y_2 : \mathbb{E}(Y_2) = \mu_2; var(Y_2) = \sigma_2^2$$

  Y_1 e Y_2 indipendenti, σ_1^2 e σ_2^2 ignote

- Se n_1 e n_2 sono sufficientemente grandi, sotto l'ipotesi nulla,

  $$Z = \frac{(\overline{Y}_2 - \overline{Y}_1) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1)$$

Test per la differenza tra le medie
Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni

- Livello di significatività del test = α
- Regione Rifiuto (Approssimata)

Ipotesi alternativa	Regione di Rifiuto
$H_a : \mu_2 \neq \mu_1$	$Z \leq -z_{\alpha/2}$ oppure $Z \geq z_{\alpha/2}$
$H_a : \mu_2 > \mu_1$	$Z \geq z_\alpha$
$H_a : \mu_2 < \mu_1$	$Z \leq -z_\alpha$

- P-valore (approssimato): Sia z^{oss} il valore osservato della statistica test Z

Ipotesi alternativa	P-valore
$H_a : \mu_2 - \mu_1 \neq 0$	$P(Z < -\lvert z_{oss}\rvert$ oppure $Z > \lvert z_{oss}\rvert; H_0)$ $= 2 \cdot (1 - P(Z \leq \lvert z_{oss}\rvert; H_0))$
$H_a : \mu_2 - \mu_1 > 0$	$P(Z \geq z_{oss}; H_0) = (1 - P(Z \leq z_{oss}; H_0))$
$H_a : \mu_2 - \mu_1 < 0$	$P(Z \leq z_{oss}; H_0)$

Nota bene. Il p-valore è calcolato usando la distribuzione Normale standard (la distribuzione campionaria approssimata di Z sotto H_0)

Test per la differenza tra le medie
Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni: Esempio

- Obiettivo: Confrontare il livello di dolore (misurato attraverso la scala visuo-analogica del dolore) a 4 ore dalla fine dell'operazione tra due gruppi di pazienti sottoposti a un'operazione chirurgica all'addome: pazienti trattati preliminarmente con morfina e pazienti trattati preliminarmente con un placebo

$$H_0 : \mu_2 = \mu_1 \quad vs \quad H_a : \mu_2 \neq \mu_1 \Longleftrightarrow H_0 : \mu_D = 0 \quad vs \quad H_a : \mu_D \neq 0$$

- Sintesi dei dati campionari:

$$\begin{array}{llll} \text{Gruppo placebo:} & n_1 = 200 & \bar{y}_1 = 45.4 & s_1^2 = 333.2 \\ \text{Gruppo morfina:} & n_2 = 120 & \bar{y}_2 = 28.0 & s_2^2 = 162.7 \end{array}$$

- Livello di significatività del test: $\alpha = 0.05$
- Regione di rifiuto (Approssimata):
$\alpha = 0.05 \Longrightarrow \alpha/2 = 0.025 \Longrightarrow z_{\alpha/2} = 1.96$. Quindi

$$RC_{0.05} = Z \leq -1.96 \quad \text{oppure} \quad Z \geq 1.96$$

- Valore osservato della statistica test

$$z_{oss} = \frac{\overline{y}_2 - \overline{y}_1}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{28 - 45.4}{\sqrt{\dfrac{333.2}{200} + \dfrac{162.7}{120}}}$$

$$= \frac{-17.4}{\sqrt{1.666 + 1.356}} = \frac{-17.4}{\sqrt{3.022}} = \frac{-17.4}{1.738} = -10.01$$

- Decisione: Il valore osservato della statistica test appartiene alla regione di rifiuto: $z_{oss} = -10.01 < -1.96 = -z_{0.025} \Longrightarrow$ I dati mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 5%

- $P-$valore (approssimato)

$$p = 2 \cdot [1 - P(Z \leq |-10.01|; H_0)] = 0.000$$

($p-$valore piccolo)

Corrispondenza tra intervalli di confidenza e test bilaterali per la media di una variabile quantitativa

- $IC_{1-\alpha}(\mu_2 - \mu_1)$: Intervallo di confidenza al livello di confidenza $1-\alpha$
- Test bilaterale:

$$H_0 : \mu_2 = \mu_1 \quad vs \quad H_a : \mu_2 \neq \mu_1 \Longleftrightarrow H_0 : \mu_D = 0 \quad vs \quad H_a : \mu_D \neq 0$$

  ✓ RC_α: Regione di rifiuto al livello di significatività α

- L'intervallo di confidenza al livello di confidenza $1-\alpha$ *contiene il valore zero* $\Longleftrightarrow$ I dati *non mostrano* evidenza contro H_0 al livello di significatività α
- L'intervallo di confidenza al livello di confidenza $1-\alpha$ *non contiene il valore zero* $\Longleftrightarrow$ I dati *mostrano* evidenza contro H_0 al livello di significatività α

Corrispondenza tra intervalli di confidenza e test bilaterali
Popolazioni Normali indipendenti con varianze note: Esempio

- Obiettivo: Confrontare il peso medio (in grammi) di neonati nati da madri non fumatrici (popolazione 1) il peso medio (in grammi) di neonati nati da madri fumatrici (popolazione 2)
- Varianze note: $\sigma_1^2 = \sigma_{\bar{F}}^2 = 50850$ e $\sigma_2^2 = \sigma_F^2 = 60025$
- Sintesi dei dati campionari:

Madre fumatrice:	$n_F = 120$	$\overline{y}_F = 2150$
Madre non fumatrice:	$n_R = 100$	$\overline{y}_R = 3375$

quindi $\overline{y}_F - \overline{y}_{\bar{F}} = -1225$ e $es(\overline{Y}_F - \overline{Y}_{\bar{F}}) = 32$

- Intervallo di confidenza al livello di confidenza $1 - \alpha = 0.95$

$$IC_{0.95}(\mu_F - \mu_{\bar{F}}) = [-1287.719; -1162.281]$$

- Verifica delle ipotesi $H_0 : \mu_F = \mu_{\bar{F}} \quad vs \quad H_a : \mu_F \neq \mu_{\bar{F}}$ al livello $\alpha = 0.05$:

$$z_{oss} = -38.28 \in RC_{0.05} = Z \leq -1.96 \text{ oppure } Z \geq 1.96$$

- L'intervallo di confidenza NON contiene lo zero: i dati mostrano evidenza contraria alla possibilità che la differenza tra le due medie sia nulla $\Longleftrightarrow$ Il test porta a rifiutare H_0 al livello di significatività del 5%

Corrispondenza tra intervalli di confidenza e test bilaterali
Popolazioni Normali indipendenti con varianze ignote ma uguali: Esempio

- Obiettivo: Confrontare il voto medio alla fine del primo anno di studenti di istituti professionali in classi in cui si sono adottati metodi di insegnamento standard (popolazione 1) e in classi in cui si sono adottati metodi di interattivi (popolazione 2)
- Sintesi dei dati campionari:

$$\begin{array}{lll} n_1 = 5 & n_2 = 3 & \\ \overline{y}_1 = 6.0 & \overline{y}_2 = 6.8 & \overline{y}_2 - \overline{y}_1 = 0.8 \\ s_1^2 = 0.215 & s_2^2 = 0.13 & s_p^2 = 0.187 \end{array}$$

- Intervallo di confidenza al livello di confidenza $1 - \alpha = 0.99$

$$IC_{0.99}(\mu_2 - \mu_1) = [-1.970; 0.370]$$

- Verifica delle ipotesi $H_0 : \mu_2 = \mu_1$ vs $H_a : \mu_2 \neq \mu_1$ al livello $\alpha = 0.01$:

$$T_{oss} = 2.535 \notin RC_{0.01} = T \leq -3.707 \;\; \text{oppure} \;\; T \geq 3.707$$

- L'intervallo di confidenza contiene lo zero: i dati non mostrano evidenza contraria alla possibilità che la differenza tra le due medie sia nulla. Coerentemente non si rifiuta H_0 al livello di significatività del 5%

Corrispondenza tra intervalli di confidenza e test bilaterali

Popolazioni qualsiasi indipendenti – Campioni di grandi dimensioni: Esempio

- Obiettivo: Confrontare il livello di dolore (misurato attraverso la scala visuo-analogica del dolore) a 4 ore dalla fine dell'operazione tra due gruppi di pazienti sottoposti a un'operazione chirurgica all'addome: pazienti trattati preliminarmente con morfina e pazienti trattati preliminarmente con un placebo
- Sintesi dei dati campionari:

$$\begin{array}{llll} \text{Gruppo placebo:} & n_1 = 200 & \overline{y}_1 = 45.4 & s_1^2 = 333.2 \\ \text{Gruppo morfina:} & n_2 = 120 & \overline{y}_2 = 28.0 & s_2^2 = 162.7 \end{array}$$

- Intervallo di confidenza al livello di confidenza $1 - \alpha = 0.95$ (approssimato)

$$IC_{0.95}(\mu_2 - \mu_1) \approx [-20.81; -13.99]$$

- Verifica delle ipotesi $H_0 : \mu_2 = \mu_1$ vs $H_a : \mu_2 \neq \mu_1$ al livello $\alpha = 0.05$:

$$z_{oss} = -10.01 \in RC_{0.05} = Z \leq -1.96 \text{ oppure } Z \geq 1.96$$

- L'intervallo di confidenza NON contiene lo zero: i dati mostrano evidenza contraria alla possibilità che la differenza tra le due medie sia nulla. Coerentemente si rifiuta H_0 al livello di significatività del 5%

Inferenza per la differenza tra le medie
Popolazioni Normali dipendenti (dati appaiati)

- Studi longitudinali → Misure Ripetute → Dati appaiati
- Una popolazione in due tempi successivi:

$$Y_1 \sim N(\mu_1, \sigma_1^2) \quad \textit{versus} \quad Y_2 \sim N(\mu_2, \sigma_2^2)$$

- Campioni casuali: $(Y_{11}, \ldots, Y_{n_1 1})$ $i.i.d.$ e $(Y_{12}, \ldots, Y_{n_2 2})$ $i.i.d.$
- Y_{i1} e Y_{i2}: misure sulla stessa unità in tempi successivi (prima e dopo un trattamento)
- $Y_{iD} = Y_{i2} - Y_{i1} \Longrightarrow Y_{1D} \ldots Y_{iD}, \ldots, Y_{nD}$ $i.i.d.$ da $Y_D \sim N(\mu_D, \sigma_D^2)$
- Si suppone che le varianze σ_1^2 e σ_2^2 (e quindi σ_D^2) siano non note

Inferenza per la differenza tra le medie
Popolazioni Normali dipendenti (dati appaiati)

- Stimatore della differenza tra le medie

$$\overline{Y}_D = \frac{1}{n}\sum_{i=1}^{n} Y_{iD}$$

- Distribuzione campionaria della statistica $\overline{Y}_D$

$$\overline{Y}_D \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right)$$

Inferenza per la differenza tra le medie
Popolazioni Normali dipendenti (dati appaiati)

- Stimatore della varianza delle differenze

$$S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_{iD} - \overline{Y}_D)^2}$$

- Statistica T e sua distribuzione campionaria

$$T = \frac{\overline{Y}_D - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$$

Intervallo di confidenza per la differenza tra le medie Popolazioni Normali dipendenti (dati appaiati)

Intervallo di confidenza al livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\mu_D) = \left[\overline{y}_D - t_{(n-1),\alpha/2}\sqrt{\frac{S_D^2}{n}}; \overline{y}_D + t_{(n-1),\alpha/2}\sqrt{\frac{S_D^2}{n}}\right]$$

Intervallo di confidenza per la differenza tra le medie Popolazioni Normali dipendenti (dati appaiati): Esempio

- Pressione sanguigna per soggetti ipertesi prima e dopo un trattamento farmacologico
- Campione casuale di 5 soggetti prima e dopo il trattamento
- Osservazioni campionarie

u_i	y_{i1}	y_{i2}	y_{iD}
1	177	136	−41
2	177	140	−37
3	166	139	−27
4	169	141	−28
5	169	141	−28

$$\bar{y}_D = \frac{161}{5} = -32.2$$

$$s_D^2 = \frac{5347 - 5 \cdot (-32.2)^2}{5-1} = 40.7$$

- Intervallo di confidenza al livello di confidenza $1 - \alpha = 0.95$

$$\begin{aligned} IC_{0.95}(\mu_D) &= \left[-32.2 - 2.78 \cdot \sqrt{\frac{40.7}{5}}; -32.2 + 2.78 \cdot \sqrt{\frac{40.7}{5}}\right] \\ &= [-40.12; -24.28] \end{aligned}$$

Test per la differenza tra le medie
Popolazioni Normali dipendenti (dati appaiati)

- Una popolazione in due tempi successivi:

$$Y_1 \sim N(\mu_1, \sigma_1^2) \quad \textit{versus} \quad Y_2 \sim N(\mu_2, \sigma_2^2)$$

- Campioni casuali: $(Y_{11}, \ldots, Y_{n_1 1})$ *i.i.d.* e $(Y_{12}, \ldots, Y_{n_2 2})$ *i.i.d.*
- Y_{i1} e Y_{i2}: misure sulla stessa unità in tempi successivi (prima e dopo un trattamento)
- $Y_{iD} = Y_{i2} - Y_{i1} \Longrightarrow Y_{1D} \ldots Y_{iD}, \ldots, Y_{nD}$ *i.i.d.* da $Y_D \sim N(\mu_D, \sigma_D^2)$
- Si suppone che le varianze σ_1^2 e σ_2^2 (e quindi σ_D^2) siano non note
- Test per la media di una popolazione normale con varianza ignota
- Ipotesi

$$\begin{array}{lcl} H_0: \mu_D = 0 & vs & H_a: \mu_D \neq 0 \\ H_0: \mu_D = 0 & vs & H_a: \mu_D > 0 \\ H_0: \mu_D = 0 & vs & H_a: \mu_D < 0 \end{array}$$

Test per la differenza tra le medie
Popolazioni Normali dipendenti (dati appaiati)

- Statistica test
$$T = \frac{\overline{Y}_D - 0}{S_D/\sqrt{n}}$$
- Sotto l'ipotesi nulla $T \sim t_{n-1}$
- Regione rifiuto al livello di significatività α

Ipotesi alternativa	Regione di Rifiuto
$H_a : \mu_D \neq 0$	$T \leq -t_{(n-1),\alpha/2}$ oppure $T \geq t_{(n-1),\alpha/2}$
$H_a : \mu_D > 0$	$T \geq t_{(n-1),\alpha}$
$H_a : \mu_D < 0$	$T \leq -t_{(n-1),\alpha}$

- P-valore: Sia t^{oss} il valore osservato della statistica test T

Ipotesi alternativa	P-valore
$H_a : \mu_D \neq 0$	$P(T < -\lvert t_{oss}\rvert$ oppure $T > \lvert t_{oss}\rvert; H_0)$
$H_a : \mu_D > 0$	$P(T \geq t_{oss}; H_0) = (1 - P(T \leq t_{oss}; H_0))$
$H_a : \mu_D < 0$	$P(T \leq t_{oss}; H_0)$

Nota bene. Il p-valore è calcolato usando la distribuzione t-Student con $n-1$ gdl (la distribuzione campionaria di T sotto H_0)

Test per la differenza tra le medie
Popolazioni Normali dipendenti (dati appaiati): Esempio

- Pressione sanguigna per soggetti ipertesi prima e dopo un trattamento farmacologico
- Campione casuale di 5 soggetti prima e dopo il trattamento

$$\bar{y}_D = \frac{-161}{5} = -32.2 \qquad s_D^2 = \frac{5347 - 5 \cdot (-32.2)^2}{5-1} = 40.7$$

- Sistema di ipotesi

$$H_0 : \mu_D = 0 \quad \textit{versus} \quad H_a : \mu_D \neq 0$$

- Livello di significatività: $\alpha = 0.05$

$$\alpha = 0.05 \Longrightarrow \alpha/2 = 0.025 \Longrightarrow t_{(n-1),0.025} = t_{4,0.025} = 2.78$$

- Regione rifiuto al livello di significatività α

$$T \leq -2.78 \quad \text{oppure} \quad T \geq 2.78$$

Test per la differenza tra le medie
Popolazioni Normali dipendenti (dati appaiati): Esempio

- Valore osservato della statistica test

$$t_{oss} = \frac{-32.2}{\sqrt{\frac{40.7}{5}}} = \frac{-32.2}{\sqrt{8.14}} = \frac{-32.2}{2.85} = -11.29$$

- Decisione: Il valore osservato della statistica test appartiene alla regione di rifiuto: $t_{oss} = -11.29 < -2.78 = -t_{4,0.025} \Longrightarrow$ I dati mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 5%
- P−valore = 0.0003512 (p−valore piccolo)
- Corrispondenza tra intervalli di confidenza e test bilaterali per la differenza tra medie con campioni dipendenti:
  - ✓ L'intervallo di di confidenza al livello di confidenza $1 - \alpha = 0.95$ *non contiene il valore zero* ma solo valori negativi
  - ✓ Coerentemente i dati *mostrano* evidenza contro H_0 al livello di significatività $\alpha = 0.05$

Inferenza per la differenza tra proporzioni

- Variabile di interesse: Y (variabile bianria)
- Popolazioni: $Y_1 \sim Ber(\pi_1)$ versus $Y_2 \sim Ber(\pi_2)$ indipendenti
- Obiettivo: Valutare se $\pi_1 = \pi_2 \iff \pi_D \equiv \pi_2 - \pi_1 = 0$
- Campioni casuali indipendenti:

$$(Y_{11}, \ldots, Y_{n_1 1})\, i.i.d. \quad (Y_{12}, \ldots, Y_{n_2 2})\, i.i.d. \quad \text{indipendenti}$$

- Proporzioni campionarie: $\hat{\pi}_1 = \overline{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{i1}$ e $\hat{\pi}_2 = \overline{Y}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{i2}$
- Stimatore per la differenza tra proporzioni: $\hat{\pi}_2 - \hat{\pi}_1 = \overline{Y}_2 - \overline{Y}_1$
- Distribuzioni campionarie: Se n_1 e n_2 sono sufficientemente grandi

$$\hat{\pi}_1 = \overline{Y}_1 \approx N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right) \quad \text{e} \quad \hat{\pi}_2 = \overline{Y}_2 \approx N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

Sotto l'ipotesi di indipendenza (per n_1 e n_2 sufficientemente grandi)

$$\hat{\pi}_2 - \hat{\pi}_1 \approx N\left(\pi_2 - \pi_1, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

Intervallo di confidenza per la differenza tra proporzioni
Campioni di grandi dimensioni

Intervallo di confidenza al livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\pi_2 - \pi_1) \approx$$
$$\left[(\hat{\pi}_2 - \hat{\pi}_1) - z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}; \right.$$
$$\left. (\hat{\pi}_2 - \hat{\pi}_1) + z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \right]$$

Intervallo di confidenza per la differenza tra proporzioni
Campioni di grandi dimensioni: Esempio

- Obiettivo: Confrontare il tasso di occupazione di giovani svantaggiati che hanno seguito un corso di formazione professionale (π_2) con il tasso di occupazione di giovani svantaggiati che non hanno seguito un corso di formazione professionale (π_1)
- Osservazioni campionarie

Corso di formazione	Numerosità campionaria	Numero occupati	Proporzione campionaria
No	$n_1 = 80$	60	$\hat{\pi}_1 = \frac{60}{80} = 0.75$
Si	$n_2 = 120$	93	$\hat{\pi}_2 = \frac{93}{120} = 0.775$

- Stima della differenza tra proporzioni $\hat{\pi}_2 - \hat{\pi}_1 = 0.775 - 0.75 = 0.025$
- Stima dell'errore standard dello stimatore differenza tra proporzioni

$$\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} =$$

$$\sqrt{\frac{0.75\cdot(1-0.75)}{80} + \frac{0.775\cdot(1-0.775)}{120}} =$$

$$\sqrt{\frac{0.1875}{80} + \frac{0.1744}{120}} = \sqrt{0.0023 + 0.0015} = \sqrt{0.0038} = 0.0616$$

- Intervallo di confidenza al livello di confidenza $1-\alpha = 0.97$

$$\begin{aligned} IC_{0.97}(\pi_2 - \pi_1) &\approx [0.025 - 2.17\cdot 0.0616; 0.025 + 2.17\cdot 0.0616] \\ &= [-0.1087; 0.1587] \end{aligned}$$

Test per la differenza tra proporzioni
Campioni di grandi dimensioni

- Parametro di interesse $\pi_D \equiv \pi_2 - \pi_1$
- Ipotesi nulla e ipotesi alternativa

$$
\begin{array}{lcl}
H_0: \pi_2 = \pi_1 \quad vs \quad H_a: \pi_2 \neq \pi_1 & & H_0: \pi_D = 0 \quad vs \quad H_a: \pi_D \neq 0 \\
H_0: \pi_2 = \pi_1 \quad vs \quad H_a: \pi_2 > \pi_1 & \Longleftrightarrow & H_0: \pi_D = 0 \quad vs \quad H_a: \pi_D > 0 \\
H_0: \pi_2 = \pi_1 \quad vs \quad H_a: \pi_2 < \pi_1 & & H_0: \pi_D = 0 \quad vs \quad H_a: \pi_D < 0
\end{array}
$$

- Sotto l'ipotesi nulla $\pi_1 = \pi_2 \equiv \pi \Longrightarrow$ Sotto H_0 le due popolazioni hanno stessa varianza $\pi(1-\pi)$
- Stimatore congiunto (pooled) della proporzione

$$
\hat{\pi} = \overline{Y}_p = \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1} Y_{i1} + \sum_{i=1}^{n_2} Y_{i2}\right) = \frac{n_1\overline{Y}_1 + n_2\overline{Y}_2}{n_1 + n_2} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2}
$$

Test per la differenza tra proporzioni
Campioni di grandi dimensioni

- Statistica test

$$Z = \frac{(\overline{Y}_2 - \overline{Y}_1) - 0}{\sqrt{\overline{Y}_p(1 - \overline{Y}_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{\sqrt{\hat{\pi}(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- Sotto l'ipotesi nulla, per n_1 e n_2 sufficientemente grandi,

$$Z \approx N(0, 1)$$

Test per la differenza tra proporzioni
Campioni di grandi dimensioni

- Livello di significatività del test = α
- Regione di rifiuto (Approssimata)

Ipotesi alternativa	Regione di Rifiuto
$H_a: \pi_2 \neq \pi_1$	$Z \leq -z_{\alpha/2}$ oppure $Z \geq z_{\alpha/2}$
$H_a: \pi_2 > \pi_1$	$Z \geq z_{\alpha}$
$H_a: \pi_2 < \pi_1$	$Z \leq -z_{\alpha}$

- P-valore (approssimato): Sia z^{oss} il valore osservato della statistica test Z

Ipotesi alternativa	P-valore
$H_a: \pi_2 - \pi_1 \neq 0$	$P(Z < -\lvert z_{oss}\rvert \text{ oppure } Z > \lvert z_{oss}\rvert; H_0)$ $= 2 \cdot (1 - P(Z \leq \lvert z_{oss}\rvert; H_0))$
$H_a: \pi_2 - \pi_1 > 0$	$P(Z \geq z_{oss}; H_0) = (1 - P(Z \leq z_{oss}; H_0))$
$H_a: \pi_2 - \pi_1 < 0$	$P(Z \leq z_{oss}; H_0)$

Nota bene. Il p-valore è calcolato usando la distribuzione Normale standard (la distribuzione campionaria approssimata di Z sotto H_0)

Test per la differenza tra proporzioni
Campioni di grandi dimensioni: Esempio I

- Obiettivo: Confrontare il tasso di occupazione di giovani svantaggiati che hanno seguito un corso di formazione professionale (π_2) con il tasso di occupazione di giovani svantaggiati che non hanno seguito un corso di formazione professionale (π_1)

$$H_0 : \pi_2 = \pi_1 \quad vs \quad H_a : \pi_2 \neq \pi_1 \Longleftrightarrow H_0 : \pi_D = 0 \quad vs \quad H_a : \pi_D \neq 0$$

- Osservazioni campionarie

Corso di formazione	Numerosità campionaria	Numero occupati	Proporzione campionaria
No	$n_1 = 80$	60	$\hat{\pi}_1 = \frac{60}{80} = 0.75$
Si	$n_2 = 120$	93	$\hat{\pi}_2 = \frac{93}{120} = 0.775$

Test per la differenza tra proporzioni
Campioni di grandi dimensioni: Esempio I

- Stima pooled della proporzione

$$\hat{\pi} = \frac{80 \cdot 0.75 + 120 \cdot 0.775}{80 + 120} = \frac{60 + 93}{80 + 120} = \frac{153}{200} = 0.765$$

- Livello di significatività: $\alpha = 0.03 \Rightarrow \alpha/2 = 0.015 \Rightarrow z_{0.015} = 2.17$
- Regione rifiuto: $RC_{0.03} = Z \leq -2.17 \quad$ oppure $\quad Z \geq 2.17$
- Valore osservato della statistica test

$$\begin{aligned} z_{oss} &= \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.775 - 0.75}{\sqrt{0.765 \cdot (1 - 0.765)\left(\frac{1}{80} + \frac{1}{120}\right)}} \\ &= \frac{0.025}{\sqrt{0.1798(0.0125 + 0.0083)}} = \frac{0.025}{\sqrt{0.0037}} = \frac{0.025}{0.0612} = 0.4085 \end{aligned}$$

- Decisione: Il valore osservato della statistica test NON appartiene alla regione di rifiuto: $-z_{0.015} = -2.17 < z_{oss} = 0.4085 < 2.17 = z_{0.015}$. I dati NON mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 3%
- $P-$valore (approssimato) $= 2 \cdot [1 - P(Z \leq |0.4085|; H_0)] = 0.6829$ ($p-$valore grande)

Test per la differenza tra proporzioni
Campioni di grandi dimensioni: Esempio II

- Obiettivo: Confrontare la proporzione di abbandoni tra matricole con e senza borsa di studio in un certo ateneo

$$H_0 : \pi_2 = \pi_1 \quad vs \quad H_a : \pi_2 > \pi_1 \Longleftrightarrow H_0 : \pi_D = 0 \quad vs \quad H_a : \pi_D < 0$$

- Osservazioni campionarie

Borsa di Studio	Numerosità campionaria	Numero di abbandoni	Proporzione campionaria
No	$n_1 = 800$	200	$\hat{\pi}_1 = \frac{200}{800} = 0.250$
Si	$n_2 = 400$	70	$\hat{\pi}_2 = \frac{70}{400} = 0.175$

Test per la differenza tra proporzioni
Campioni di grandi dimensioni: Esempio II

- Stima pooled della proporzione

$$\hat{\pi} = \frac{800 \cdot 0.25 + 400 \cdot 0.175}{800 + 400} = \frac{200 + 70}{1200} = \frac{270}{1200} = 0.225$$

- Livello di significatività: $\alpha = 0.05 \Rightarrow z_{0.05} = 1.645$
- Regione rifiuto: $RC_{0.05} = Z \leq -1.645$
- Valore osservato della statistica test

$$\begin{aligned} z_{oss} &= \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.175 - 0.25}{\sqrt{0.225 \cdot (1 - 0.225)\left(\frac{1}{800} + \frac{1}{400}\right)}} \\ &= \frac{-0.075}{\sqrt{0.1744(0.00125 + 0.0025)}} = \frac{-0.075}{\sqrt{0.0006}} = \frac{0.075}{0.0256} = -2.933 \end{aligned}$$

- Decisione: Il valore osservato della statistica test appartiene alla regione di rifiuto: $z_{oss} = -2.933 \leq -1.645 = -z_{0.05} \Longrightarrow$ I dati mostrano evidenza contraria all'ipotesi nulla al livello di significatività del 5%
- $P-$valore (approssimato) $= P(Z \leq -2.933; H_0) = 0.00168$ ($p-$valore piccolo)