STATISTICA

Esercizi-5

Statistico, detective o romanziere?

« Per risolvere i casi complicati bisogna essere capaci di **costruire una storia**, partendo dagli **indizi disponibili**, che contenga una **spiegazione plausibile** di tutti gli elementi che abbiamo. Ci vuole una certa dose di fantasia ed è un lavoro simile a quello di uno scrittore. Una volta costruita questa storia che è, in sostanza, un'**ipotesi** su come potrebbero essersi svolti i fatti, bisogna andare alla ricerca delle **conferme**.»

G. Carofiglio
Una mutevole verità

Statistico, detec… …anziere?

LINGUAGGIO della PROBABILITA'

« Per risolvere i casi comp… …ati bisogna essere capaci di **costruire una storia**, partendo dagli **indizi disponibili**, che contenga una **spiegazione plausibile** di t… …tti gli elementi che abb… vuole una ce… …i fantasia ed è u… simile a quello di uno scrittore. Una volta costruita questa storia che è, in sostanza, u… come potrebbero essersi svolti i fatt… …bisogna andare alla ricerca delle **conferme**.»

DATI

MODELLI

INFERENZA

Esercizio 6

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

1. Stimare la percentuale di studenti universitari favorevoli al partito Arcobaleno e determinare l'IC(90%).

Esercizio 6

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

1. Stimare la percentuale di studenti universitari favorevoli al partito Arcobaleno e determinare l'IC(90%).

$X_1, \dots, X_{100} \sim b(p)$ $\qquad$ $X_1 + \cdots + X_{100} \sim Binom(100, p)$

$$\hat{p}_{100} = \frac{40}{100} = 0.4$$

Stima della vera % di favorevoli ad Arc. nell'intera pop. studentesca

Esercizio 6

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

1. Stimare la percentuale di studenti universitari favorevoli al partito Arcobaleno e determinare l'IC(90%).

$X_1, \dots, X_{100} \sim b(p)$ $\quad\quad$ $X_1 + \cdots + X_{100} \sim Binom(100, p)$

$\hat{p}_{100} = \dfrac{40}{100} = 0.4$ $\quad\quad$ Stima della vera % di favorevoli ad Arc. nell'intera pop. studentesca

$$IC(90\%):\ \hat{p}_n \mp z_{\frac{0.10}{2}} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} : 0.4 \mp 1.6449 \sqrt{\frac{0.4 \times 0.6}{100}} \Rightarrow (0.319\,, 0.481)$$

Esercizio 6

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

2. Nel 2013, su un campione di 100 studenti si sono detti favorevoli al partito Arcobaleno in 43. C'è abbastanza evidenza per asserire che nel 2013 il consenso al partito è aumentato rispetto al 2012?

Esercizio 6

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

2. Nel 2013, su un campione di 100 studenti si sono detti favorevoli al partito Arcobaleno in 43. C'è abbastanza evidenza per asserire che nel 2013 il consenso al partito è aumentato rispetto al 2012?

$$\left.\begin{array}{ll} X_1, \dots, X_{100} \sim b(p_1) & (2012) \\ Y_1, \dots, Y_{100} \sim b(p_2) & (2013) \end{array}\right\} \text{2 campioni indipendenti, e grandi.} \left\} \begin{array}{l} \hat{p}_1 = 0.4 \\ \hat{p}_2 = 0.43 \end{array}\right.$$

Esercizio 6

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

2. Nel 2013, su un campione di 100 studenti si sono detti favorevoli al partito Arcobaleno in 43. C'è abbastanza evidenza per asserire che nel 2013 il consenso al partito è aumentato rispetto al 2012?

$$\left.\begin{array}{ll} X_1, \dots, X_{100} \sim b(p_1) & (2012) \\ Y_1, \dots, Y_{100} \sim b(p_2) & (2013) \end{array}\right\}$$

2 campioni indipendenti, e grandi.

$$\hat{p}_1 = 0.4 \qquad \hat{p}_2 = 0.43$$

$$\boldsymbol{H_0 : p_1 = p_2} \qquad \boldsymbol{H_1 : p_1 < p_2}$$

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \times (1 - \hat{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$\hat{p}$ = frazione **totale** di "successi"

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

rifiuto H_0 al livello $\alpha\%$ se la statistica test < $-z_\alpha$

Esercizio 6

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

2. Nel 2013, su un campione di 100 studenti si sono detti favorevoli al partito Arcobaleno in 43. C'è abbastanza evidenza per asserire che nel 2013 il consenso al partito è aumentato rispetto al 2012?

$$\left.\begin{array}{ll} X_1, \dots, X_{100} \sim b(p_1) & (2012) \\ Y_1, \dots, Y_{100} \sim b(p_2) & (2013) \end{array}\right\}$$

2 campioni indipendenti, e grandi.

$$\hat{p}_1 = 0.4 \qquad \hat{p}_2 = 0.43$$

$$\boxed{\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \times (1 - \hat{p}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}$$

$\hat{p}$ = frazione **totale** di "successi"

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

$$\hat{p} = \frac{100 \times 0.4 + 100 \times 0.43}{200} = 0.415$$

$$\frac{0.4 - 0.43}{\sqrt{0.415 \times (1 - 0.415) \times \left(\frac{1}{100} + \frac{1}{100}\right)}} = -0.43$$

Esercizio 6

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

2. Nel 2013, su un campione di 100 studenti si sono detti favorevoli al partito Arcobaleno in 43. C'è abbastanza evidenza per asserire che nel 2013 il consenso al partito è aumentato rispetto al 2012?

$X_1, \dots, X_{100} \sim b(p_1)$ (2012)

2 campioni indipendenti, e grandi.

$Y_1, \dots, Y_{100} \sim b(p_2)$ (2013)

$\hat{p}_1 = 0.4$

$\hat{p}_2 = 0.43$

$$\boldsymbol{H_0 : p_1 = p_2} \qquad \boldsymbol{H_1 : p_1 < p_2}$$

rifiuto H_0 al livello $\alpha\%$ se la statistica test < $z_\alpha = -z_{1-\alpha}$

$\boldsymbol{Stat.test = -0.43}$ $\Rightarrow$ $p - value = P(Z < -0.43) = 0.3336$

troppo grande!

non possiamo rifiutare H_0 a favore di un aumento del consenso, a nessun livello

Esercizio 6 (ripasso)

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

3. Se la vera percentuale di favorevoli ad Arc. nella popolazione degli studenti è 0.45, qual è la probabilità che in un campione casuale di 100 di loro i favorevoli siano 40?

Esercizio 6 (ripasso)

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

3. Se la vera percentuale di favorevoli ad Arc. nella popolazione degli studenti è 0.45, qual è la probabilità che in un campione casuale di 100 di loro i favorevoli siano 40?

$$X_1, \dots, X_{100} \sim b(0.45)$$

S = numero di favorevoli nel campione di 100 studenti

$$S \sim Binom(100, 0.45)$$

$$P(S = 40) = \binom{100}{40} 0.45^{40} 0.55^{60} = 0.0488029$$

Esercizio 6 (ripasso)

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

4. Se la vera percentuale di favorevoli ad Arc. nella popolazione degli studenti è 0.45, qual è la probabilità che in un campione casuale di 100 di loro almeno 40 siano favorevoli?

Esercizio 6 (ripasso)

In un sondaggio elettorale nel 2012 condotto su un campione di 100 studenti per stimare la percentuale di coloro che sono favorevoli al partito Arcobaleno, 40 studenti hanno risposto positivamente.

4. Se la vera percentuale di favorevoli ad Arc. nella popolazione degli studenti è 0.45, qual è la probabilità che in un campione casuale di 100 di loro almeno 40 siano favorevoli?

$X_1, \ldots, X_{100} \sim b(0.45)$

$S =$ numero di favorevoli nel campione di 100 studenti

$S \sim Binom(100, 0.45) \approx N(100 \times 0.45, 100 \times 0.45 \times 0.55) = N(45, 24.75)$

$$\boldsymbol{P(S \geq 40) = \sum_{k=40}^{100} \binom{100}{k} 0.45^k 0.55^{100-k} \approx P\left(\frac{S-45}{\sqrt{24.75}} \geq \frac{40-45}{\sqrt{24.75}}\right) = P(Z \geq -1.01)}$$

$$\boldsymbol{= 1 - P(Z \leq -1.01) = 1 - 0.1562 = 0.8438}$$

Esercizio 1

Per il lancio di una nuova pubblicità di una grande azienda d'abbigliamento viene estratto un campione casuale di 1000 unità e viene chiesto di scegliere tra 3 nuovi manifesti stradali (A, B e C) quello preferito. Nella tabella sottostante le risposte sono distinte per fasce d'età:

Età	**A**	**B**	**C**
$18 \dashv 25$	60	10	90
$25 \dashv 35$	115	60	95
$35 \dashv 45$	120	125	75
$45 \dashv 60$	120	90	40

Si vuole valutare se c'è associazione tra l'età dell'individuo e la risposta al quesito.

a) Si specifichino le ipotesi da sottoporre a verifica.
b) Si decida in merito alle ipotesi poste sulla base di un opportuno test.
c) Si commenti il risultato ottenuto.

Esercizio 1

Per il lancio di una nuova pubblicità di una grande azienda d'abbigliamento viene estratto un campione casuale di 1000 unità e viene chiesto di scegliere tra 3 nuovi manifesti stradali (A, B e C) quello preferito. Nella tabella sottostante le risposte sono distinte per fasce d'età:

Età	A	B	C
$18 \dashv 25$	60	10	90
$25 \dashv 35$	115	60	95
$35 \dashv 45$	120	125	75
$45 \dashv 60$	120	90	40

H_0 : indipendenza H_1 : associazione

Rifiutiamo l'ip. di indipendenza se la statistica test è "troppo grande"

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{\left(n_{ij} - n^*{}_{ij}\right)^2}{n^*{}_{ij}}$$

Esercizio 1

Per il lancio di una nuova pubblicità di una grande azienda d'abbigliamento viene estratto un campione casuale di 1000 unità e viene chiesto di scegliere tra 3 nuovi manifesti stradali (A, B e C) quello preferito. Nella tabella sottostante le risposte sono distinte per fasce d'età:

$$\boldsymbol{n_{ij}}^* = n \times \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n} = \frac{\boldsymbol{n_{i\cdot} \times n_{\cdot j}}}{\boldsymbol{n}}$$

Età	A	B	C	$n_{i\cdot}$
$18 \dashv 25$	60	10	90	**160**
$25 \dashv 35$	115	60	95	**270**
$35 \dashv 45$	120	125	75	**320**
$45 \dashv 60$	120	90	40	**250**
$\boldsymbol{n_{\cdot j}}$	**415**	**285**	**300**	$\boldsymbol{n = 1000}$

A	B	C
66.40	45.60	48
112.05	76.95	81
132.80	91.20	96
103.75	71.25	75

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{\left(n_{ij} - n^*{}_{ij}\right)^2}{n^*{}_{ij}} = 113.5579$$

$$\chi^2(3 \times 2)_{0.0001} = 27.8527$$

Esercizio 1

Per il lancio di una nuova pubblicità di una grande azienda d'abbigliamento viene estratto un campione casuale di 1000 unità e viene chiesto di scegliere tra 3 nuovi manifesti stradali (A, B e C) quello preferito. Nella tabella sottostante le risposte sono distinte per fasce d'età:

$$\boldsymbol{n_{ij}}^* = n \times \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n} = \frac{\boldsymbol{n_{i\cdot} \times n_{\cdot j}}}{\boldsymbol{n}}$$

Età	**A**	**B**	**C**	$n_{i\cdot}$
18 ⊣ 25	60	10	90	**160**
25 ⊣ 35	115	60	95	**270**
35 ⊣ 45	120	125	75	**320**
45 ⊣ 60	120	90	40	**250**
$\boldsymbol{n_{\cdot j}}$	**415**	**285**	**300**	$\boldsymbol{n = 1000}$

A	**B**	**C**
66.40	45.60	48
112.05	76.95	81
132.80	91.20	96
103.75	71.25	75

$$\chi^2 = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{\left(n_{ij} - {n^*}_{ij}\right)^2}{{n^*}_{ij}} = 113.5579$$

rifiutiamo l'ip. di indip.

$$\chi^2(3 \times 2)_{0.0001} = 27.8527$$

***p*-valore < 0.0001**

Esercizio 1

Per il lancio di una nuova pubblicità di una grande azienda d'abbigliamento viene estratto un campione casuale di 1000 unità e viene chiesto di scegliere tra 3 nuovi manifesti stradali (A, B e C) quello preferito. Nella tabella sottostante le risposte sono distinte per fasce d'età:

$$\boldsymbol{n_{ij}}^* = n \times \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n} = \frac{\boldsymbol{n_{i\cdot} \times n_{\cdot j}}}{\boldsymbol{n}}$$

Età	A	B	C	$n_{i\cdot}$
$18 \dashv 25$	60	**10**	**90**	**160**
$25 \dashv 35$	115	60	95	**270**
$35 \dashv 45$	120	125	75	**320**
$45 \dashv 60$	120	90	**40**	**250**
$\boldsymbol{n_{\cdot j}}$	**415**	**285**	**300**	$\boldsymbol{n = 1000}$

A	B	C
66.40	**45.60**	**48**
112.05	76.95	81
132.80	91.20	96
103.75	71.25	**75**

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{\left(n_{ij} - n^*{}_{ij}\right)^2}{n^*{}_{ij}} = 113.5579$$

rifiutiamo l'ip. di indip.

$$\chi^2(3 \times 2)_{0.0001} = 27.8527$$

***p*-valore < 0.0001**

Esercizio di compito

La seguente tabella riporta la ripartizione per fasce d'età delle donne americane non sposate con figli nel 1986 (U.S. National Center for Health Statistics, *Vital Statistics of the United States*):

Fascia d'età	**%**
≤ 14	1.1
15-|19	32.0
20-|24	36.0
25-|29	18.9
≥ 30	12.0

Da un recente campione di 1000 nati da donne non sposate risulta che 42 delle madri hanno 14 anni o meno, 403 tra i 15 ed i 19 anni, 315 tra 20 e 24 anni, 150 tra 25 e 29 anni e 90 ne hanno almeno 30. Si può sostenere che la distribuzione sia variata?

Esercizio (soluzione)

La seguente tabella riporta la ripartizione per fasce d'età delle donne americane non sposate con figli nel 1986 (U.S. National Center for Health Statistics, *Vital Statistics of the United States*):

Fascia d'età	p_i (%)	$n^*{}_j(*)$	n_i
≤ 14	1.1	11	42
15-|19	32.0	320	403
20-|24	36.0	360	315
25-|29	18.9	189	150
≥ 30	12.0	120	90

$H_0 \;:\; n_j = n^*{}_j$ per ogni j

$H_1 \;:\; n_j \neq n^*{}_j$ per almeno un j

$$\sum_{i=1}^{c} \frac{(n_j - n^*{}_j)^2}{n^*{}_j} > \chi^2(c-1)_\alpha$$

$(*)\; n^*{}_j = 1000 \times p_i/100$

$$\frac{(42-11)^2}{11} + \frac{(403-320)^2}{320} + \frac{(315-360)^2}{360} + \frac{(150-189)^2}{189} + \frac{(90-120)^2}{120} = 130.06$$

Esercizio (soluzione)

La seguente tabella riporta la ripartizione per fasce d'età delle donne americane non sposate con figli nel 1986 (U.S. National Center for Health Statistics, *Vital Statistics of the United States*):

Fascia d'età	p_i (%)	$n^*_j(*)$	n_i
≤ 14	1.1	11	42
15-|19	32.0	320	403
20-|24	36.0	360	315
25-|29	18.9	189	150
≥ 30	12.0	120	90

$H_0 \; : \; n_j = n^*_j$ per ogni j

$H_1 \; : \; n_j \neq n^*_j$ per almeno un j

$$\sum_{i=1}^{c} \frac{(n_j - n^*_j)^2}{n^*_j} > \chi^2(c-1)_\alpha$$

$(*)\; n^*_j = 1000 \times p_i/100$

$$\frac{(42-11)^2}{11} + \frac{(403-320)^2}{320} + \frac{(315-360)^2}{360} + \frac{(150-189)^2}{189} + \frac{(90-120)^2}{120} = 130.06$$

Evidenza molto forte contro H_0 $\qquad > \chi^2(4)_{0.01} = 13.28$

Esercizio (soluzione)

La seguente tabella riporta la ripartizione per fasce d'età delle donne americane non sposate con figli nel 1986 (U.S. National Center for Health Statistics, *Vital Statistics of the United States*):

Fascia d'età	p_i (%)	$n^*_j(*)$	n_i
≤ 14	1.1	**11**	**42**
15-|19	32.0	**320**	**403**
20-|24	36.0	360	315
25-|29	18.9	189	150
≥ 30	12.0	120	90

$H_0 \; : \; n_j = n^*_j$ per ogni j

$H_1 \; : \; n_j \neq n^*_j$ per almeno un j

$$\sum_{i=1}^{c} \frac{(n_j - n^*_j)^2}{n^*_j} > \chi^2(c-1)_\alpha$$

$(*)\; n^*_j = 1000 \times p_i/100$

$$\frac{(42-11)^2}{11} + \frac{(403-320)^2}{320} + \frac{(315-360)^2}{360} + \frac{(150-189)^2}{189} + \frac{(90-120)^2}{120} = 130.06$$

Evidenza molto forte contro H_0 $\qquad > \chi^2(4)_{0.01} = 13.28$

Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'età al primo parto (in anni) di 10 donne:

14, 36, 22, 21, 28, 19, 32, 28, 30, 20

a. Di che tipo di variabile si tratta?
b. Rappresentare graficamente la distribuzione di frequenza dei dati.
c. Calcolare l'età media al primo parto, e la varianza dei dati.
d. Calcolare i quartili e disegnare il boxplot.

Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'**età** al primo parto (**in anni**) di 10 donne:

14, 36, 22, 21, 28, 19, 32, 28, 30, 20

a. Di che tipo di variabile si tratta? ***Quantitativa discreta*** **(anche se…)**

Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'età al primo parto (in anni) di 10 donne:

14, 19, 20, 21, 22, 28, 28, 30, 32,36

b. Rappresentare graficamente la distribuzione di frequenza dei dati.

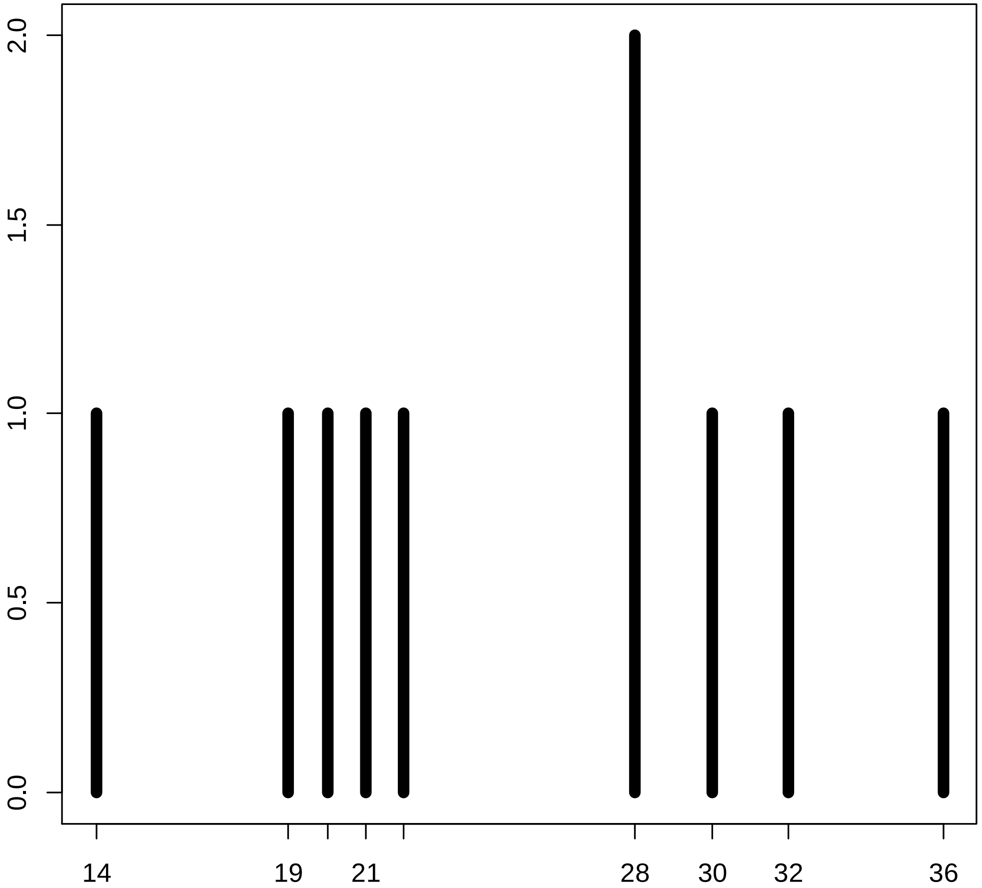


diagramma ad aste/bastoncini

Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'età al primo parto (in anni) di 10 donne:

14, 19, 20, 21, 22, 28, 28, 30, 32,36

b. Rappresentare graficamente la distribuzione di frequenza dei dati.

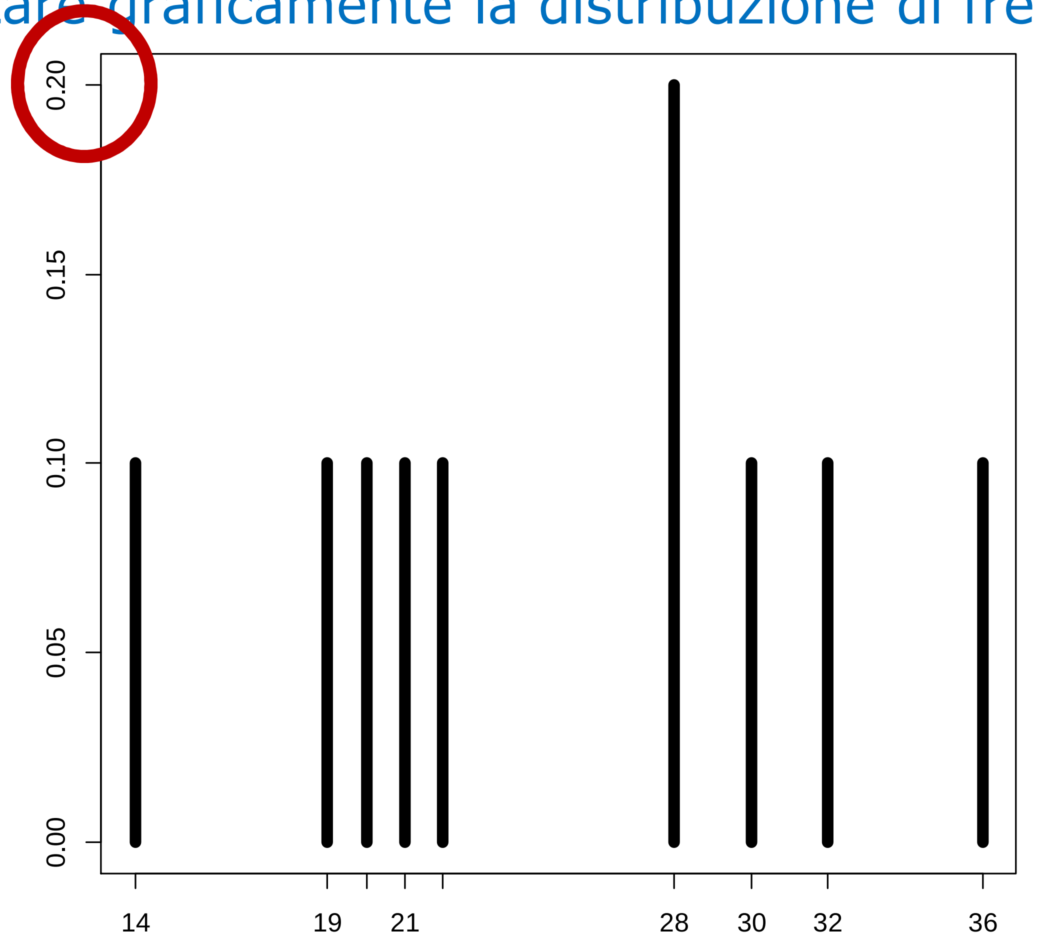


diagramma ad aste/bastoncini, con le freq. relative

Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'età al primo parto (in anni) di 10 donne:

14, 19, 20, 21, 22, 28, 28, 30, 32,36

c. Calcolare l'età media al primo parto, e la varianza dei dati.

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{14 + 19 + \cdots + 36}{10} = 25$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \left[\frac{1}{n}\sum_{i=1}^{n} {x_i}^2\right] - {\bar{x}_n}^2 = \frac{14^2 + 19^2 + \cdots + 36^2}{10} - 25^2 = 42$$

Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'età al primo parto (in anni) di 10 donne:

14, 19, 20, 21, 22, 28, 28, 30, 32,36

c. Calcolare l'età media al primo parto, e la varianza dei dati.

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{14 + 19 + \cdots + 36}{10} = 25$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \left[\frac{1}{n}\sum_{i=1}^{n} {x_i}^2\right] - {\bar{x}_n}^2 = \frac{14^2 + 19^2 + \cdots + 36^2}{10} - 25^2 = 42$$

Attenzione a *varianza dei dati* e *varianza* ***campionaria***

Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'età al primo parto (in anni) di 10 donne:

14, 19, 20, 21, 22, 28, 28, 30, 32,36

d. Calcolare i quartili e disegnare il boxplot.

$$Q_1 : \frac{n+1}{4} = \frac{11}{4} = 2.75 \Rightarrow \ Q_1 = \frac{19+20}{2} = 19.5$$

Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'età al primo parto (in anni) di 10 donne:

14, 19, 20, 21, 22, 28, 28, 30, 32,36

d. Calcolare i quartili e disegnare il boxplot.

$$Q_1 : \frac{n+1}{4} = \frac{11}{4} = 2.75 \Rightarrow Q_1 = \frac{19+20}{2} = 19.5$$

$$Q_3 : 3 \times \frac{n+1}{4} = 3 \times \frac{11}{4} = 8.25 \Rightarrow Q_3 = \frac{30+32}{2} = 31.0$$

$$Q_2 : \frac{n+1}{2} = \frac{11}{2} = 5.5 \Rightarrow Q_2 = \frac{22+28}{2} = 25.0$$

Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'età al primo parto (in anni) di 10 donne:

14, 19, 20, 21, 22, 28, 28, 30, 32,36

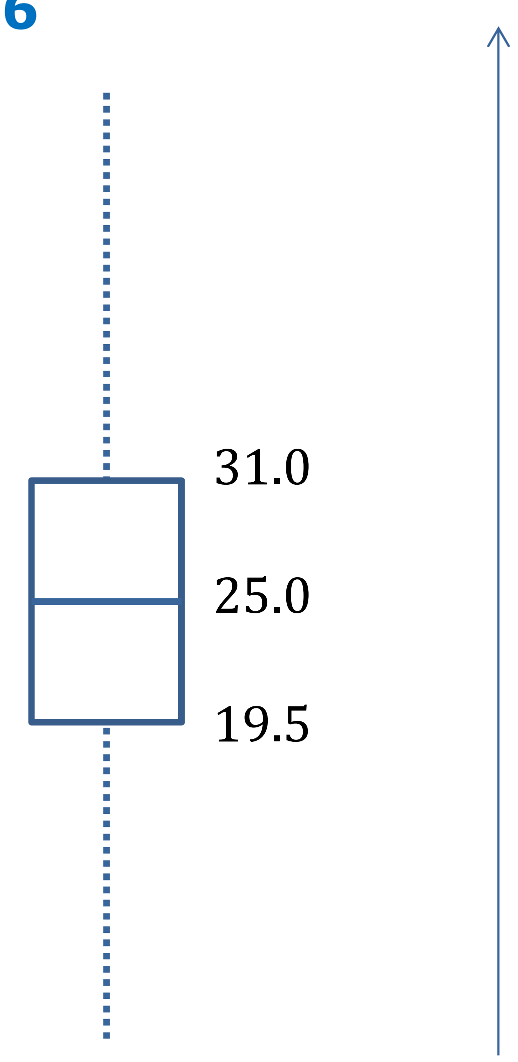
d. Calcolare i quartili e disegnare il boxplot.

$$Q_1 : \frac{n+1}{4} = \frac{11}{4} = 2.75 \Rightarrow \; Q_1 = \frac{19+20}{2} = 19.5$$

$$Q_3 : 3 \times \frac{n+1}{4} = 3 \times \frac{11}{4} = 8.25 \Rightarrow \; Q_3 = \frac{30+32}{2} = 31.0$$

$$Q_2 : \frac{n+1}{2} = \frac{11}{2} = 5.5 \Rightarrow \; Q_2 = \frac{22+28}{2} = 25.0$$

$$Baffi : \; 1.5 \times (Q_3 - Q_1) = 1.5 \times 11.5 = 17.25$$



Esercizio 2a (ripasso)

I seguenti dati si riferiscono all'età al primo parto (in anni) di 10 donne:

14, 19, 20, 21, 22, 28, 28, 30, 32,36

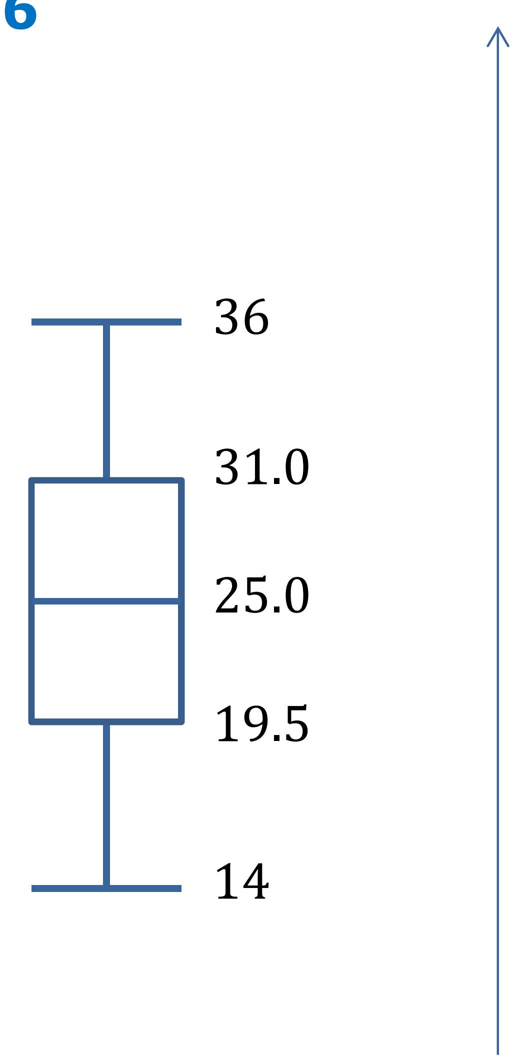
d. Calcolare i quartili e disegnare il boxplot.

$$Q_1 : \frac{n+1}{4} = \frac{11}{4} = 2.75 \Rightarrow \; Q_1 = \frac{19+20}{2} = 19.5$$

$$Q_3 : 3 \times \frac{n+1}{4} = 3 \times \frac{11}{4} = 8.25 \Rightarrow \; Q_3 = \frac{30+32}{2} = 31.0$$

$$Q_2 : \frac{n+1}{2} = \frac{11}{2} = 5.5 \Rightarrow \; Q_2 = \frac{22+28}{2} = 25.0$$

$$Baffi : \; 1.5 \times (Q_3 - Q_1) = 1.5 \times 11.5 = 17.25$$



Esercizio 2b (ripasso)

Da un recente campione di 1000 nati da donne non sposate risulta che 42 delle madri hanno 14 anni o meno, 403 tra i 15 ed i 19 anni, 315 tra 20 e 24 anni, 150 tra 25 e 29 anni e 90 ne hanno almeno 30.

Fascia d'età	n_i
< 15	42
$[15, 20)$	403
$[20, 25)$	315
$[25, 30)$	150
≥ 30	90

a. Rappresentare graficamente la distribuzione di frequenza dell'età al primo parto nel campione.
b. Calcolare l'età media al primo parto, e la varianza nel campione.

[(*) Per gli studenti presenti in aula e che hanno preso appunti: attenti che ho cambiato qualcosina… provate a capire cosa e perchè.]

Esercizio 2b (ripasso)

Da un recente campione di 1000 nati da donne non sposate risulta che 42 delle madri hanno 14 anni o meno, 403 tra i 15 ed i 19 anni, 315 tra 20 e 24 anni, 150 tra 25 e 29 anni e 90 ne hanno almeno 30.

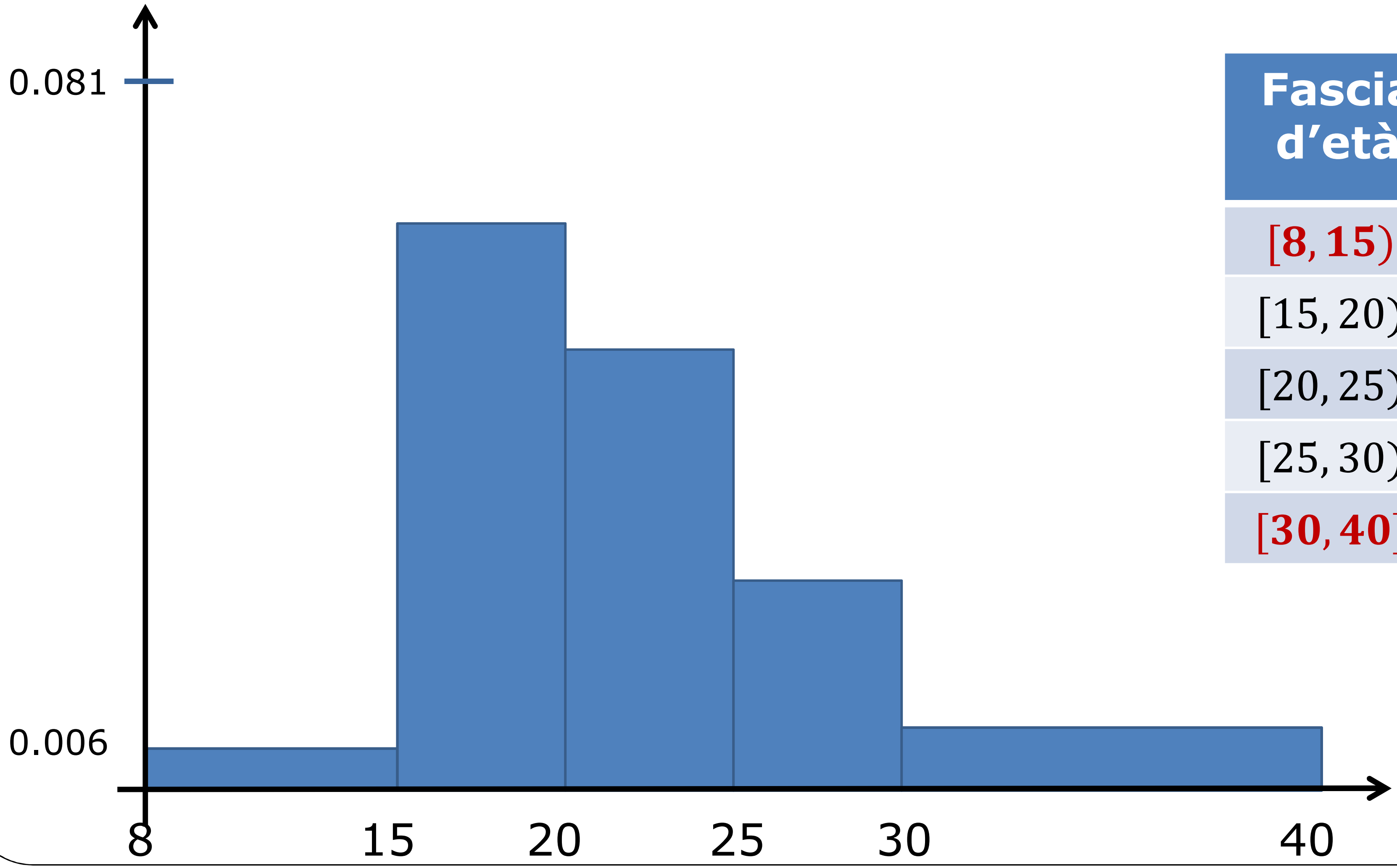
Fascia d'età	n_i
< 15	42
$[15, 20)$	403
$[20, 25)$	315
$[25, 30)$	150
≥ 30	90

a. Rappresentare graficamente la distribuzione di frequenza dell'età al primo parto nel campione. ⇒ **istogramma**

Fascia d'età	n_i	Freq. Rel. f_i	$l'_i = f_i / a_i$
$\mathbf{[8, 15)}$	42	0.042	0.006
$[15, 20)$	403	0.403	0.081
$[20, 25)$	315	0.315	0.063
$[25, 30)$	150	0.150	0.030
$\mathbf{[30, 40]}$	90	0.090	0.009

Esercizio 2b (ripasso)

Da un recente campione di 1000 nati da donne non sposate risulta che 42 delle madri hanno 14 anni o meno, 403 tra i 15 ed i 19 anni, 315 tra 20 e 24 anni, 150 tra 25 e 29 anni e 90 ne hanno almeno 30.



Fascia d'età	$l'_i = f_i/a_i$
$[8, 15)$	0.006
$[15, 20)$	0.081
$[20, 25)$	0.063
$[25, 30)$	0.030
$[30, 40]$	0.009

Esercizio 2b (ripasso)

Fascia d'età	x_i^*	Freq. Rel. f_i
$[8, 15)$	**11.5**	0.042
$[15, 20)$	**17.5**	0.403
$[20, 25)$	**22.5**	0.315
$[25, 30)$	**27.5**	0.150
$[30, 40]$	**35**	0.090

a. Calcolare l'età media al primo parto, e la varianza.

$$s_n{}^2 = \frac{1}{n-1}\sum (x_i{}^* - \bar{x})^2 n_i =$$

$$= \frac{n}{n-1}\sum (x_i{}^* - \bar{x})^2 f_i$$

$$\bar{x} = 11.5 \times 0.042 + 17.5 \times 0.403 + 22.5 \times 0.315 + 27.5 \times 0.15 + 35 \times 0.09 = 21.898$$

$$s_n{}^2 = \frac{1000}{999}[11.5^2 \times 0.042 + 17.5^2 \times 0.403 + \cdots + 35^2 \times 0.09 - 21.898^2]$$

$$= 32.607 \Longrightarrow s_n = \sqrt{32.607} = 5.710 \text{ anni}$$

Esercizio 4 (di compito)

Da un campione di 1200 intervistati nel 2016 risulta che 255 di essi sono a basso reddito.

a) Costruire un intervallo di confidenza del 98% per la vera proporzione di persone a basso reddito (BR) in Italia nel 2016.

b) C'è abbastanza evidenza nel campione per dimostrare con la significatività del 5% che la vera proporzione di persone BR nel 2016 supera il 20%?

c) Nel 1989 le persone BR rappresentavano il 16% del totale. C'è abbastanza evidenza nel campione per dimostrare che la vera proporzione di persone BR è aumentata dal 1989 ad oggi?

Esercizio 4 (soluzione)

Da un campione di 1200 intervistati nel 2016 risulta che 255 di essi sono a basso reddito.

a) Costruire un intervallo di confidenza del 98% per la vera proporzione di persone a basso reddito (BR) in Italia nel 2016.

$$\left(\hat{p}_n - z_{\alpha/2} \times \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{\alpha/2} \times \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right)$$

$$\alpha = 0.02$$

$$\hat{p}_n = \frac{255}{1200} = 0.2125$$

(sono soddisfatte le condizioni $np \geq 5$, $n(1-p) \geq 5$)

$$\left(0.2125 - 2.3263 \times \sqrt{\frac{0.2125(1-0.2125)}{1200}}, 0.2125 + 2.3263 \times \sqrt{\frac{0.2125(1-0.2125)}{1200}}\right)$$

$$(0.185\,, 0.240)$$

Esercizio 4 (soluzione)

Da un campione di 1200 intervistati nel 2016 risulta che 255 di essi sono a basso reddito.

b) C'è abbastanza evidenza nel campione per dimostrare con la significatività del 5% che la vera proporzione di persone BR nel 2016 **supera il 20%**?

$$\hat{p}_n = \frac{255}{1200} = 0.2125 \qquad H_0 : p = 0.20 \ , \quad \boldsymbol{H_1} \ : \boldsymbol{p > 0.20} \qquad \alpha = 0.05$$

$$\frac{\hat{p}_n - 0.20}{\sqrt{\frac{0.20 \times (1 - 0.20)}{1200}}} = \frac{0.2125 - 0.20}{\sqrt{\frac{0.20 \times (1 - 0.20)}{1200}}} = 1.082 \ ?>? \ z_{0.05} = 1.6449$$

Esercizio 4 (soluzione)

Da un campione di 1200 intervistati nel 2016 risulta che 255 di essi sono a basso reddito.

b) C'è abbastanza evidenza nel campione per dimostrare con la significatività del 5% che la vera proporzione di persone BR nel 2016 **supera il 20%**?

$$\hat{p}_n = \frac{255}{1200} = 0.2125 \qquad H_0: p = 0.20 \ , \quad \boldsymbol{H_1} \ : \boldsymbol{p > 0.20} \qquad \alpha = 0.05$$

$$\frac{\hat{p}_n - 0.20}{\sqrt{\frac{0.20 \times (1 - 0.20)}{1200}}} = \frac{0.2125 - 0.20}{\sqrt{\frac{0.20 \times (1 - 0.20)}{1200}}} = 1.082 \ ?\!>\!? \ z_{0.05} = 1.6449$$

NO, dato che non possiamo rifiutare l'ipotesi nulla.

Esercizio 4 (soluzione)

Da un campione di 1200 intervistati nel 2016 risulta che 255 di essi sono a basso reddito.

b) C'è abbastanza evidenza nel campione per dimostrare con la significatività del 5% che la vera proporzione di persone BR nel 2016 **supera il 20%**?

$$\hat{p}_n = \frac{255}{1200} = 0.2125$$

$$H_0 : p = 0.20 \ , \quad \boldsymbol{H_1} \ : \boldsymbol{p > 0.20} \qquad \alpha = 0.05$$

Non potevo usare l'intervallo di confidenza trovato in a) perchè l'intervallo di confidenza in a) corrisponde al test
$H_0 : p = 0.20$, $\boldsymbol{H_1} : \boldsymbol{p \neq 0.20}$ al livello del **2**%.

Esercizio 4 (soluzione)

Da un campione di 1200 intervistati nel 2016 risulta che 255 di essi sono a basso reddito.

c) Nel 1989 le persone BR rappresentavano il 16% del totale. C'è abbastanza evidenza nel campione per dimostrare che **la vera proporzione di persone BR è aumentata dal 1989 ad oggi**?

$$\hat{p}_n = \frac{255}{1200} = 0.2125 \qquad H_0 : p = 0.16 \quad , \quad H_1 : p > 0.16$$

$$\frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.2125 - 0.16}{\sqrt{0.16(1-0.16)/1200}} = 4.96 \Rightarrow p - value: < 0.001$$

SI', infatti possiamo rifiutare l'ipotesi nulla che la proporzione non sia cambiata a favore dell'alternativa che sia aumentata, con un livello di significatività minore di 0.001.

Esercizio 4 (soluzione)

Da un campione di 1200 intervistati nel 2016 risulta che 255 di essi sono a basso reddito.

c) Nel 1989 le persone BR rappresentavano il 16% del totale. C'è abbastanza evidenza nel campione per dimostrare che **la vera proporzione di persone BR è aumentata dal 1989 ad oggi**?

$$\hat{p}_n = \frac{255}{1200} = 0.2125 \qquad H_0 : p = 0.16 \quad , \quad H_1 : \; p > 0.16$$

$$\frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.2125 - 0.16}{\sqrt{0.16(1-0.16)/1200}} = 4.96 \Rightarrow p - value: < 0.001$$

SI', infatti possiamo rifiutare l'ipotesi nulla che la proporzione non sia cambiata a favore dell'alternativa che sia aumentata, con un livello di significatività minore di 0.001.

Corriere della Sera del 5/06/2016. La percentuale per il 1989 è esatta; il vero valore per il 2015 è 0.21. Dati Bankitalia.

Esercizio 5

Da un campione di 2000 famiglie per l'anno 1989 risulta che il consumo medio annuo di zucchero era di 19457.60 gr a famiglia, con una deviazione standard di 857 gr.

Da un campione di 750 famiglie per l'anno 2015 risulta un consumo medio annuo di zucchero di 16820.53 gr a famiglia, con una deviazione standard di 975.30 gr.

a) Possiamo affermare che il consumo familiare annuo di zucchero è diminuito dal 1989 ad oggi?

b) Se i dati del 1989 fossero riferiti all'intera popolazione per quell'anno, potremmo trarre le stesse conclusioni di cui al punto a)?

Esercizio 5

Da un campione di 2000 famiglie per l'anno 1989 risulta che il consumo medio annuo di zucchero era di 19457.60 gr a famiglia, con una deviazione standard di 857 gr.
Da un campione di 750 famiglie per l'anno 2015 risulta un consumo medio annuo di zucchero di 16820.53 gr a famiglia, con una deviazione standard di 975.30 gr.

a) Possiamo affermare che il consumo familiare annuo di zucchero è diminuito dal 1989 ad oggi?

Confronto delle medie in due campioni indipendenti:

$(X_1, \dots, X_{n_1})$ **i.i.d** $\sim N(\mu_1, \sigma^2)$,
(1989)

$(Y_1, \dots, Y_{n_2})$ **i.i.d** $\sim N(\mu_2, \sigma^2)$,
(2015)

σ^2 non nota

$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 > \mu_2$

$$\boxed{\frac{\bar{x}_1 - \bar{x}_2}{s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t(n_1 + n_2 - 2)_{\alpha}}$$

$$s^2 = \frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}$$

Esercizio 5

Da un campione di 2000 famiglie per l'anno 1989 risulta che il consumo medio annuo di zucchero era di 19457.60 gr a famiglia, con una deviazione standard di 857 gr.
Da un campione di 750 famiglie per l'anno 2015 risulta un consumo medio annuo di zucchero di 16820.53 gr a famiglia, con una deviazione standard di 975.30 gr.

a) Possiamo affermare che il consumo familiare annuo di zucchero è diminuito dal 1989 ad oggi?

Confronto delle medie in due campioni indipendenti:

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 > \mu_2$$

$$\frac{\bar{x}_1 - \bar{x}_2}{s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t(n_1 + n_2 - 2)_\alpha$$

$$s^2 = \frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2} = \frac{1999 \times 857^2 + 749 \times 975.30^2}{2748} = 793529.8 \rightarrow \sqrt{793529.8} = 890.8$$

Esercizio 5

Da un campione di 2000 famiglie per l’anno 1989 risulta che il consumo medio annuo di zucchero era di 19457.60 gr a famiglia, con una deviazione standard di 857 gr.
Da un campione di 750 famiglie per l’anno 2015 risulta un consumo medio annuo di zucchero di 16820.53 gr a famiglia, con una deviazione standard di 975.30 gr.

a) Possiamo affermare che il consumo familiare annuo di zucchero è diminuito dal 1989 ad oggi?

Confronto delle medie in due campioni indipendenti:

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 > \mu_2$$

$$\frac{\bar{x}_1 - \bar{x}_2}{s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{19457.60 - 16820.53}{890.8 \times \sqrt{\frac{1}{2000} + \frac{1}{750}}} = 69.14 > t(n_1 + n_2 - 2)_\alpha = \boldsymbol{z_\alpha} < \mathbf{5}!!$$

Esercizio 5

Da un campione di 2000 famiglie per l'anno 1989 risulta che il consumo medio annuo di zucchero era di 19457.60 gr a famiglia, con una deviazione standard di 857 gr.
Da un campione di 750 famiglie per l'anno 2015 risulta un consumo medio annuo di zucchero di 16820.53 gr a famiglia, con una deviazione standard di 975.30 gr.

b) Se i dati del 1989 fossero riferiti **all'intera popolazione** per quell'anno, potremmo trarre le stesse conclusioni di cui al punto a)?

Confronto delle medie in un campione con il valore nella popolazione

$(X_1, \dots, X_{n_1})$ **i.i.d** $\sim N(\mu, \sigma^2)$, **(2015)**

$\sigma = 857{,}00$ **gr (*)**

[(*) si ricordi la discussion fatta a lezione]

$H_0 : \mu = 19457.60 \qquad H_1 : \mu < 19457.60$

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$$

Confermato il risultato di cui al punto a)

$$\frac{16820.53 - 19457.60}{857.00/\sqrt{750}} = -84.27!!$$

Esercizio 5 *alternativa* (*)

Da un campione di 2000 famiglie per l'anno 1989 risulta che il consumo medio annuo di zucchero era di 19457.60 gr a famiglia, con una deviazione standard di 857 gr.
Da un campione di 750 famiglie per l'anno 2015 risulta un consumo medio annuo di zucchero di 16820.53 gr a famiglia, con una deviazione standard di 975.30 gr.

b) Se i dati del 1989 fossero riferiti **all'intera popolazione** per quell'anno, potremmo trarre le stesse conclusioni di cui al punto a)?

Confronto delle medie in un campione con il valore nella popolazione

$(X_1, \dots, X_{n_1})$ **i.i.d** $\sim N(\mu, \sigma^2)$, **(2015)**

σ **non nota gr (*)**

[(*) Per una questione legata alla stesura del testo, è ammessa anche questa soluzione]

$H_0: \mu = 19457.60 \qquad H_1: \mu < 19457.60$

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} < -t(n-1)_\alpha$$

Confermato il risultato di cui al punto a)

$$\frac{16820.53 - 19457.60}{975.30/\sqrt{750}} = -74.05!!$$

Esercizio 5

Da un campione di 2000 famiglie per l'anno 1989 risulta che il consumo medio annuo di zucchero era di 19457.60 gr a famiglia, con una deviazione standard di 857 gr.
Da un campione di 750 famiglie per l'anno 2015 risulta un consumo medio annuo di zucchero di 16820.53 gr a famigli,a con una deviazione standard di 975.30 gr.

b) Se i dati del 1989 fossero riferiti **all'intera popolazione** per quell'anno, potremmo trarre le stesse conclusioni di cui al punto a)?

Non era zucchero… Dal Corriere della Sera del 5/06/2016. **Reddito** medio di un lavoratore dipendente 1989: circa 20000€; reddito medio per il 2015, meno di 17000 € . Dati Bankitalia.

Confermato il risultato di cui al punto a)

$$\frac{16820.53 - 19457.60}{857.00/\sqrt{750}} = -84.27!!$$