Esercizi di Statistica, con soluzioni e non solo...

G. Marchetti

2016 ver. 1.9

Indice

1 Introduzione

- Introduzione alla statistica
- Fenomeni collettivi
- Distinzione tra unità e variabili
- Classificazione delle variabili qualitative/quantitative, discrete/continue
- Distribuzione di frequenza
- Diagramma stem-and-leaf
- Istogramma
- Distribuzioni doppie di frequenza

1.1

Considera 20 famiglie. Per ciascuna rileva il numero di componenti. Ecco i dati:

```
1 3 2 5 4 2 2 3 3 2 3 4 4 3 2 5 4 3 3 1
```

Costruisci la distribuzione di frequenza.

Soluzione

```
Componenti    1 2 3 4 5
Famiglie      2 5 7 4 2
```

1.2

Nell'esempio precedente dire

- qual è l'unità
- qual è la variabile
- La variabile è quantitativa o qualitativa?

Soluzione

L'unità è la famiglia, la variabile il numero di componenti, quantitativa discreta.

1.3

Per 20 giorni hai registrato i minuti di ritardo del treno per arrivare a Firenze.
Eccoli:

```
28  5  4 12 17 12 14  5  4  4 11  8  4 26 17  6  0 19  8 38
```

Fai un grafico stem-and-leaf prendendo come stelo la cifra delle decine. Costruisci una distribuzione di frequenza con classi (in minuti)

```
0-9  10-19 20-29 30-39
```

Soluzione

```
0 | 0444455688
1 | 1224779
2 | 68
3 | 8
```

1.4

Nell'esempio precedente dire

- qual è l'unità
- qual è la variabile
- La variabile è quantitativa o qualitativa?

Soluzione

L'unità è il viaggio (o il treno), la variabile è il ritardo in min. quantitativa, continua.

1.5

In un'ora una libreria fa 20 scontrini per i seguenti importi in Euro

```
10 13 13 18 18 18 19 19 20 20 20 20 22 22 23 24 24 25 26 27
```

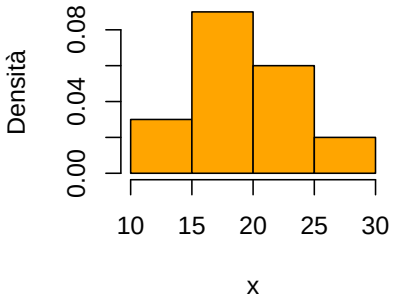
Fate un istogramma con classi

```
10 |- 15 , 15 |- 20, 20 |- 25, 25 |- 30
```

Soluzione

```
Importo    n. clienti
10 |- 15      3
15 |- 20      5
20 |- 25      9
```

```
25 |- 30       3
Totale         20
```



1.6

Ecco la distribuzione del salario mensile (in Euro) di un campione di 1000 lavoratori:

Classi di reddito	Frequenze	Altezze dei rettangoli
0 |- 500	100	0.2
500 |- 1000	200	0.4
1000 |- 2000	500	0.5
2000 |- 4000	600	0.3
4000 |- 8000	400	0.1

Disegnare l'istogramma e verificare che le aree dei rettangoli sono uguali alle frequenze. Come sono state calcolate le altezze dei rettangoli?

Soluzione

Le altezze sono = frequenze/ampiezza di classe.

2 Indici

- Moda
- Media
- Mediana
- Deviazione standard e varianza
- Quartili e quantili
- Box-plot
- Introduzione alla disuguaglianza di Chebyshev

2.1

Le temperature nella località X alle 12 sono state

`12.1 14.5 9.7 8.1 13.0 12.5 10.5`

Calcola la media e la mediana.

Soluzione

Media = 80.4/7 = 11.48 gradi centigradi.

Mediana = 12.1 gradi centigradi.

8.1 9.7 10.5 (12.1) 12.5 13.0 14.5

2.2

La distribuzione di 520 studenti per numero di esami superati è

Esami	Studenti
0	50
1	100
2	160
3	120
4	80
5	10

Calcolare le frequenze relative, la moda e il numero medio di esami superati. Calcolare la mediana.

Soluzione

Esami	Studenti	Freq. relative	Cumulate	Prodotto
0	50	0.10	0.10	0
1	100	0.19	0.29	100
2	160	0.31	0.60	320
3	120	0.23	0.83	360
4	80	0.15	0.98	320
5	10	0.02	1.00	50
Totale	520	1.00		1150

La moda è 2 esami superati. La media è 1150/520 = 2.21 esami. La mediana si trova notando che le due unità centrali sono la 260 e la 261-ma nella successione ordinata e dalla distribuzione si vede che stanno entrambe nella classe di 2 esami.

Quindi la mediana è 2 esami superati.

2.3

La popolazione delle prime 10 città americane in milioni è la seguente.

New York (New York)	9.21
Los Angeles (California)	4.05
Chicago (Illinois)	2.83
Houston (Texas)	2.01
Phoenix (Arizona)	1.55
Filadelfia	1.45
Dallas (Texas)	1.31
San Diego (California)	1.30
San Antonio (Texas)	1.24
San Jose (California)	0.94

Calcolare la popolazione media e la popolazione mediana. Quale indice è migliore?

Soluzione

La popolazione totale è 25.80 milioni. Quindi la media è 2.58 milioni. La mediana è la semisomma tra 1.45 e 1.55 cioè 1.5 milioni di abitanti.

È meglio la mediana perché non risente troppo dei valori anomali (come New York).

2.4

Ecco il voto di laurea di 5 studenti di Lettere

```
110 109 108 110 110
```

Ecco il voto di laurea di 5 studenti di Economia

```
90 98 110 105 102
```

C'è maggiore variabilità di voto a Economia o a Lettere? Giustificare calcolando le varianze del voto e le deviazioni standard.

Soluzione

- A Lettere il voto medio di laurea è 109.4.
- A Economia il voto medio è 101.
- La varianza del voto a Lettere è 0.8
- La varianza del voto a Economia è 57. (Ho usato il denominatore $n = 4$.

Le deviazioni standard sono perciò

- Lettere punti 0.8944
- Economia punti 7.5498.

Evidentemente la variabilità è minore a Lettere. Volendo usare il *coefficiente di variazione* per ottenere una misura relativa di variabilità si ottiene

- Lettere CV = 0.0082
- Economia CV = 0.0748

Quindi il coefficiente di variazione è minore a Lettere.

2.5

Considera la distribuzione di 1000 studenti secondo il voto di laurea

Voto	Frequenza
98	10
99	40
100	250
101	400
102	250
103	40
104	10

- Mostrate che la media la moda e la mediana sono uguali a 101. Verificate che la deviazione standard è di 1 punto.
- Calcolate la frequenza relativa di studenti che hanno preso un voto compreso tra 101 - 2 = 99 e 101 + 2 = 103. Secondo la disuguaglianza di Chebyshev questa frequenza relativa quanto dovrebbe essere?

Soluzione

La distribuzione è simmetrica e quindi media moda e mediana sono uguali. La moda e la mediana sono evidentemente uguali a 101 punti. La media è 101000/1000 = 101.

Per calcolare la varianza compiliamo la tabella seguente

```
Voto    Freq.   Voto^2    Voto^2 * freq.
98        10      9604         96040
99        40      9801        392040
100      250     10000       2500000
101      400     10201       4080400
102      250     10404       2601000
103       40     10609        424360
104       10     10816        108160
Totale  1000                10202000
```

Quindi la media dei voti al quadrato è

$$MQ = 10202000/1000 = 10202$$

e la varianza è uguale alla MQ meno la media al quadrato :

$$MQ - (media^2) = 10202 - 101^2 = 1.$$

Quindi anche la deviazione standard è 1.

La proporzione di studenti che hanno preso un voto tra 99 e 103 è $980/1000 = 98\%$.

Come vedremo la disuguaglianza di Chebychev asserisce che per forza questa proporzione deve essere maggiore di 1 - $1/4 = 75\%$. E infatti così avviene.

3 Indici di associazione

- Covarianza
- Coefficiente di correlazione
- Retta dei minimi quadrati

3.1

Su 4 famiglie di 2 componenti misuriamo il reddito di Febbraio X e le relative spese per l'alimentazione Y.

```
X:    1500 1700  1400 1600
Y:     200  350   150  300
```

Calcolare la covarianza e il coefficiente di correlazione e interpretarli.

Soluzione

media$(X) = 1550$, media$(Y) = 250$, var$(X) = 12500$, var$(Y) = 6250$.

$(X$ - media$(X))(Y$ - media$(Y)) : 2500, 15000, 15000, 2500$

cov$(X, Y) = 35000/4 = 8750$.

cor$(X, Y) = 8750/\sqrt{(12500)(6250)} = 0.9899$.

3.2

Provate a calcolare la covarianza con la formula equivalente

media(XY) - media(X) media(Y)

Soluzione

Poiché media$(XY) = (1500 \times 200 + 1700 \times 350 + 1400 \times 150 + 1600 \times 300)/4$ abbiamo

$\text{cov}(X, Y) = \text{media } (XY) - \text{ media } (X) \text{ media}(Y) = 396250 - (1550)(250) = 8750.$

3.3

Rappresentate la relazione tra spesa e reddito con la retta dei minimi quadrati e verificate che questa è: Spesa $= -835 + 0.7$ Reddito. Provate a interpretare il risultato.

Soluzione

Coefficiente angolare:

$$\text{cor}(X, Y)\sqrt{\text{var}(Y)/\text{var}(X)} = \text{cov}(X, Y)/\text{var}(X) = 8750/12500 = 0.7.$$

La retta deve passare per il punto (media(X), media(Y)) $= (1550, 250)$ e quindi ha equazione

$$y = 250 + 0.7(x - 1550) = -835 + 0.7x.$$

La pendenza ha l'interpretazione: per ogni Euro in più di reddito la spesa aumenta di 70 centesimi.

3.4

Vero o Falso? Se X e Y sono due variabili con $\text{var}(X) = 3.25$, $\text{var}(Y) = 5.8$, $\text{cov}(X, Y) = 14.703$

allora il coefficiente di correlazione è 0.78. Giustificare.

Soluzione

Falso. Perché $0.78 = \text{cov}(X, Y)/(\text{var}(X)\text{var}(Y))$ mentre il coefficiente di correlazione è

$$\text{cor}(X, Y) = \text{cov}(X, Y)/\sqrt{\text{var}(X)\text{var}(Y)}.$$

4 Probabilità

1. Esperimenti aleatori
2. Eventi elementari, eventi
3. Probabilità (classica e frequentista)
4. Combinazioni
5. Probabilità condizionata
6. Indipendenza
7. Tavole di probabilità
8. Formula di Bayes

- Un esperimento aleatorio è un esperimento il cui esito è incerto.
- Un evento è un insieme di esiti specificati prima di fare l'esperimento.
- Un evento si verifica se si verifica uno dei suoi componenti elementari.
- La probabilità è una misura $P(A)$ dell'incertezza associata al verificarsi di evento A.

Gli eventi si comportano e si combinano fra loro come insiemi.

Si distinguono:

- L'evento certo S che si verifica sempre
- L'evento impossibile $\emptyset$ che non si verifica mai

- L'evento $A \cup B$ unione di due eventi A e B che si verifica se si verifica l'uno, l'altro o entrambi
- L'evento $A \cap B$ intersezione di A e B che si verifica se si verificano entrambi
- L'evento complementare $\bar{A}$ di un evento A che si verifica se A non si verifica.
- Due eventi incompatibili o disgiunti per cui l'intersezione è impossibile: $A \cap B = \emptyset$
- Due eventi esustivi la cui unione è l'evento certo: $A \cup B = S$
- Se A è un evento incluso in un evento B allora se A se verifica anche B si verifica.
- Se A è un evento la sua probabilità è $0 \leq P(A) \leq 1$
- Se A è l'evento certo $P(A) = 1$
- Se A è l'evento impossibile $P(A) = 0$
- Se A è incluso in B $P(A) \leq P(B)$.
- Se A e B sono incompatibili $(A \cap B) = P(A) + P(B)$
- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Regola di de Morgan

$$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$$

(probabilità che non si verifichi né A né B)

- Interpretazione classica: se ogni esito (= caso, evento elementare) ha la stesse chances di verificarsi

$$P(A) = \frac{\text{n. esiti favorevoli}}{\text{n.esiti possibili}}.$$

- Combinazioni: Il numero di combinazioni di n oggetti di classe k è

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdots (n-k+1)}{n!}$$

- Il numero di tutti i possibili sottoinsiemi di k elementi di un insieme di $n > k$ elementi è $\binom{n}{k}$.
- *Probabilità condizionata*: è la probabilità $P(A \mid B)$ di un evento A supponendo che si sia verificato un evento B. Vale la relazione seguente

  $$P(A \mid B) = P(A \cap B)/P(B)$$

  Vale anche la relazione invertita

  $$P(A \cap B) = P(A \cap B)P(B).$$

- Due eventi si dicono indipendenti se $P(A \mid B) = P(A)$ ovvero se $P(A \cap B) = P(A)P(B)$.
- *Tavole di probabilità*: Se A e B sono due eventi si definisce la tabella seguente

	B	$\bar{B}$	
A	$P(A \cap B)$	$P(A \cap \bar{B})$	$P(A)$
$\bar{A}$	$P(\bar{A} \cap B)$	$P(\bar{A} \cap \bar{B})$	$P(\bar{A})$
	$P(B)$	$P(\bar{B})$	1

- Se A e B sono indipendenti

	B	$\bar{B}$	
A	$P(A)P(B)$	$P(A)P(\bar{B})$	$P(A)$
$\bar{A}$	$P(\bar{A})P(B)$	$P(\bar{A})P(\bar{B})$	$P(\bar{A})$
	$P(B)$	$P(\bar{B})$	1

- *Formula delle probabilità totali*: Sostituendo nella tavola $P(A \cap B) = P(A \mid B)P(B)$ etc.

	B	$\bar{B}$	
A	$P(A \mid B)P(B)$	$P(A \mid \bar{B})P(\bar{B})$	$P(A)$
$\bar{A}$	$P(\bar{A} \mid B)P(B)$	$P(\bar{A} \mid \bar{B})P(\bar{B})$	$P(\bar{A})$
	$P(B)$	$P(\bar{B})$	1

  e quindi

$$P(A) = P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})$$

- *Formula di Bayes*: Se B e $\bar{B}$ sono due ipotesi esaustive e disgiunte (sano, malato) e A è un sintomo (risulta positivo a un test diagnostico) la formula permette di calcolare la probabilità dell'ipotesi B dato A usando $P(A \mid B)$:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})}$$

  $P(B)$ è la probabilità iniziale di B, $P(B \mid A)$ è la probabilità finale aggiornata dopo aver visto il sintomo.

4.1

Una classe di studenti di Statistica è formata da 200 persone di cui 120 sono femmine. Ci sono 60 maschi che provengono da Ragioneria. Ci sono in totale 90 studenti che non hanno fatto Ragioneria.

Supponiamo di estrarre a caso uno studente. Trovate:

- la probabilità di selezionare una femmina
- la probabilità di selezionare un maschio
- la probabilità di selezionare una persona che non ha fatto Ragioneria
- la probabilità di selezionare una femmina che ha fatto Ragioneria
- la probabilità di selezionare o una femmina o una persona che ha fatto Ragioneria.

Suggerimento:

```
                        Sesso
Scuola           Maschio    Femmina   Totale
Ragioneria         60          *        *
Altra              *           *        90
Totale            120          200
```

Soluzione

Usando le informazioni possiamo completare la tabella:

```
                              Sesso
Scuola           Maschio    Femmina   Totale
Ragioneria         60          50        110
Altra              20          70         90
Totale             80         120        200
```

Quindi

- la probabilità di selezionare una femmina = 120/200 = 0.6
- la probabilità di selezionare un maschio = 0.4
- la probabilità di selezionare una persona che non ha fatto Ragioneria = 90/200 = 0.45
- la probabilità di selezionare una femmina che ha fatto Ragioneria = 50/200 = 0.25
- la probabilità di selezionare o una femmina o una persona che ha fatto Ragioneria = (50 + 60 + 70)/200 = 0.9.

4.2

Supponiamo che in una partita di calcio la probabilità che vinca la squadra di casa sia 0.5 e la probabilità che vinca la squadra ospite sia 0.2. Qual è la probabilità di pareggio?

Soluzione

P(pareggio) = P(non(Vinca 1 o Vinca 2)) = 1 - P(Vinca 1 o Vinca 2) = 1 - 0.5 - 0.2 = 0.3.

4.3

Da un mazzo di carte (da 52) si pescano a caso 2 carte. Calcolare con la regola classica la probabilità che siano due assi. Suggerimento: quanti sono i casi possibili = numero di eventi elementari? Quanti sono i casi favorevoli = numero di eventi elementari componenti l'evento "2 assi"?

NOTA: in quest'ultimo esempio occorre capire che l'evento elementare è la coppia non ordinata di carte. Il numero di coppie (non ordinate) possibili prese da N oggetti sono $N(N-1)/2$. Queste si chiamano anche combinazioni di N oggetti di classe 2. Per esempio il numero di coppie prese da tre oggetti $\{a, b, c\}$ sono $\{a, b\}, \{a, c\}$ e $\{b, c\}$. Infatti se $N = 3, N(N-1)/2 = 3$. Notare che $\{a, b\} = \{b, a\}$, l'ordine non conta.

Soluzione

I casi possibili sono $52 \cdot 51/2 = 1326$ (tutte le coppie possibili). Sono gli eventi elementari che ti possono capitare. I casi favorevoli all'evento sono quei casi in cui estrai due assi dei 4 esistenti. Sono $4 \cdot 3/2 = 6$:

```
     C     F    Q     P       C =  Asso di cuori, F = fiori, Q = quadri, P = picche
C    -     CF   CQ    CP
F           -   FQ    FP
Q                 -   QP
```

Quindi la probabilità di ottenere esattamente due assi in due carte è 6/1326.

4.4

La probabilità di estrarre una persona che fumi oppure che sia maschio è 0.7. Qual è la probabilità di estrarre una persona che sia una femmina e non fumi?

NOTA: Sia A = maschio e B = fuma. L'evento: "persona che fumi oppure che sia maschio" = A unione B.

L'evento: "persona che sia una femmina e non fumi" = (non A). intersezione (non B) = non(A unione B).

Questa è una legge logica. Se diciamo: "non voglio né questo né quello" è come dire: "(non voglio questo) e (non voglio quello)". Ma significa anche: "non voglio (questo oppure quello) = non voglio (questo solo, quello solo, entrambi)".

Soluzione

P(fuma o maschio) = 0.7

P(non fuma e non maschio) = P(non(fuma o maschio)) = 1 - P(fuma o maschio) = 1 - 0.7 = 0.3.

4.5

Un analista finanziario fornisce le stime dell'utile di un'azienda nel prossimo anno, considerando anche il tasso di interesse. Eccole.

	Utile		
Tasso di interesse	<8%	da 8% a 12%	> 12%
< 3%	0.09	0.15	0.16
da 4% a 5%	0.14	0.17	0.05
> 5%	0.16	0.07	0.01

Qual è la probabilità che l'azienda realizzi un utile di almeno l'8%?

Soluzione

La Probabilità che dell'evento A = l'azienda realizza un utile minore dell' 8% è

$$P(A) = 0.09 + 0.14 + 0.16 = 0.39.$$

(basta fare il totale della prima colonna). Quindi la probabilità cercata è

$$P(non\ A) = 1 - 0.39 = 0.61$$

4.6

La probabilità dell'intersezione tra due eventi A e B non può essere maggiore né della probabilità di A né della probabilità di B. Vero o falso?

Soluzione

Se un evento C è contenuto in un altro evento A, la probabilità di C deve per forza essere minore o uguale alla probabilità di A. Infatti tutti gli eventi elementari di C sono contenuti in A e quindi se C si verifica per forza si verifica anche A. Ora l'intersezione di A e B è per forza contenuta sia in A che in B. Quindi

$$P(A\&B) \leq P(A) \text{ e } P(A\&B) \leq P(B)$$

e l'affermazione è vera.

4.7

Una recente indagine ha rivelato che il 14% delle segretarie ha dolore al polso. Inoltre, il 6% delle segretarie intervistate ha dolore al polso e al tempo stesso assume regolarmente un farmaco antinfiammatorio. Qual è la probabilità che una segretaria che ha dolore al polso assuma regolarmente un farmaco antinfiammatorio?

Soluzione

Sia A = la segretaria ha il dolore al polso e B = la segretaria assume il farmaco. Perciò: $P(A) = 0.14$ e $P(A\&B) = 0.06$.

La probabilità richiesta è la probabilità condizionata $P(B \mid A)$.

Per definizione

$$P(B|A) = P(B\&A)/P(A) = 0.06/0.14 = 0.428.$$

4.8

Si abbia una popolazione di 10 oggetti. Si estraggano senza ripetizione tutti i possibili campioni non ordinati di dimensione 4. Quanti campioni fanno parte dello spazio campionario?

Soluzione

Sono le combinazioni di 10 oggetti di classe 4

$$\frac{(10)(9)(8)(7)}{(4)(3)(2)(1)} = 210.$$

4.9

La tabella seguente riporta le probabilità congiunte di un insieme di nuove aziende che operano nel settore del commercio via internet, classificate per regione di ubicazione e prospettiva di crescita.

Crescita	Nord-Est	Sud	Centro	Nord-Ovest
Bassa	0.04	0.12	0.14	0.19
Media	0.05	0.08	0.06	0.12
Alta	0.03	0.05	0.08	0.04

Se l'azienda ha una crescita attesa media o alta, qual è la probabilità che sia ubicata nel Nord-Ovest?

A) 0.16 B) 0.31 C) 0.27 D) 0.46

Soluzione

Sia M = l'azienda ha una crescita media e A = l'azienda ha una crescita alta.

Sia N = l'azienda è ubicata nel Nord-Ovest. Abbiamo

$$P(M) = 0.05 + 0.08 + 0.06 + 0.12 = 0.31$$
$$P(A) = 0.03 + 0.05 + 0.08 + 0.04 = 0.20$$

Quindi

$$P(M \cup A) = P(M) + P(A) = 0.31 + 0.20 = 0.51.$$

Per comodità chiamiamo $B = M \cup A$. Si chiede $Pr(N \mid M \cup A) = P(N|B)$.

Allora per definizione

$$\begin{aligned} P(N \mid B) &= P(N\&B)/P(B) \\ &= \frac{P(\text{ubicata a NW e a crescita medio/alta})}{P(\text{a crescita medio/alta})} \\ &= (0.12 + 0.04)/0.51 = 0.31. \end{aligned}$$

Quindi la soluzione è B).

4.10

Un negozio di computer ha ricevuto una fornitura di 14 computer, 5 dei quali con modem già istallato. Sfortunatamente sulle scatole mancano le etichette per distinguere i computer con modem dagli altri. Supponi di scegliere casualmente 4 computer. Qual è la probabilità che esattamente 2 di essi siano provvisti di modem?

A) 0.3012

B) 0.3704

C) 0.3596

D) 0.3288

Soluzione

Casi possibili = # combinazioni di 14 oggetti di classe 4

$$\binom{14}{4} = \frac{(14)(13)(12)(11)}{(4)(3)(2)(1)} = 1001.$$

Casi favorevoli = # di modi di appaiare 2 computer col modem e 2 senza. Siccome i 2 col modem sono presi dai 5 con e i 2 senza sono presi dai 9 senza, i casi favorevoli sono

$$\binom{5}{2}\binom{9}{2} = (5)(4)/2 \cdot (9)(8)/2 = 360.$$

Quindi la probabilità è $360/1001 = 0.3596$.

DIFFICILE!

4.11

In un recente sondaggio sul sindaco di un certa città, il 62% dei rispondenti ha fiducia nel sindaco. Le donne costituiscono il 53% del campione, e tra queste il 46% ha fiducia nel sindaco. Si seleziona a caso una persona tra quelle intervistate. Qual è la probabilità che la persona selezionata sia maschio ?

Soluzione

Eventi: F = un rispondente ha fiducia nel sindaco. D = un rispondente è donna. M = un rispondente è maschio.

Cosa sappiamo?

$$P(F) = 0.62, \quad P(D) = 0.53, \quad P(F \mid D) = 0.46$$

Chiede la probabilità $P(M) = P(\text{non } D) = 1 - P(D) = 1 - 0.53 = 0.47$.

L'esercizio spiazza un po' perché fornisce un dato che non serve.

4.12

Supponiamo di lanciare due dadi. Si consideri la somma dei due dadi: sia A l'evento *si osserva un numero pari* e B l'evento *si osserva un numero maggiore di 7*. Quale delle seguenti affermazioni è vera?

A) Gli eventi A e B sono mutuamente esclusivi.
B) L'intersezione tra A e B è l'insieme [6, 8, 10, 12].
C) Gli eventi A e B sono collettivamente esaustivi.
D) Nessuna delle precedenti.

Soluzione

$$A = \{2, 4, 6, 8, 10, 12\}, B = \{8, 9, 10, 11, 1, 2\}, A\&B = \{8, 10, 12\}.$$

A) FALSO (l'intersezione non è vuota)
B) FALSO (l'intersezione non è quella)
C) FALSO (A o B non comprende per esempio il 3)
D) VERO

4.13

La probabilità di passare lo scritto di Statistica è il 50%. La probabilitá di passare l'esame orale dato che si è superato lo scritto è il 98%. Qual è la probabilità di passare l'esame?

Soluzione

S = lo studente passa lo scritto. O = lo studente passa l'orale

$$P(S) = 0.5, \quad P(O|S) = 0.98$$

S e O = lo studente passa l'esame

$$P(S\&O) = P(S)P(O \mid S) = (0.5)(0.98) = 0.49.$$

4.14

In una popolazione ci sono il 50% di maschi e il 50% di femmine. Supponiamo che il 5% degli uomini e il 10% delle donne siano daltonici (non riconoscono i colori). Si sceglie a caso una persona daltonica. Qual è la probabilità che sia un maschio?

Soluzione

M = maschio, F = femmina, D = persona daltonica

Sappiamo che $P(M) = 0.5 = P(F)$. Inoltre $P(D \mid M) = 0.05$ e $P(D \mid F) = 0.10$.

Si chiede $P(M|D) = ?$

Abbiamo $P(M \mid D) = P(D\&M)/P(D)$. Inoltre:

$$\begin{aligned} P(D\&M) &= P(D \mid M)P(M) = (0.05)(0.5) = 0.025 \\ P(D) &= P(D\&M) + P(D\&F) \\ &= P(D \mid M)P(M) + P(D \mid F)P(F) \\ &= (0.05)(0.5) + (0.10)(0.5) = 0.025 + 0.05 = 0.075. \end{aligned}$$

Quindi $P(M \mid D) = 0.025/0.075 = 1/3$.

4.15

Se in una certa prova si ha che $P(A) = 0.7$ può accadere che $P(A \cup B) = 0.5$?

A) Solo se A e B sono eventi incompatibili
B) Solo se B=$\emptyset$
C) SI è possibile
D) NO mai

Soluzione A è incluso in $A \cup B$. Perciò $P(A) \leq P(A \cup B)$. Quindi non può succedere: risposta D).

5 Variabili casuali discrete

1. Funzione di probabilità
2. Funzione di ripartizione
3. Valore atteso
4. Varianza

5. Binomiale

- Una v.c. discreta X è definita da una lista di modalità x_i e da una lista di probabilità $p_i \geq 0$ tali che $\sum_{i=1}^{k} p_i = 1$. Le p_i definiscono la *funzione di massa di probabilità* di X detta spesso la *distribuzione di probabilità* di X.
- La *funzione di ripartizione* di X è la distribuzione delle probabilità cumulate: $F(x) = P(X \leq x)$.
- Il *valore atteso* di X è la media di X

$$E(X) = \mu_X = \sum_{i=1}^{k} x_i p_i$$

La *varianza* di X è

$$\text{var}(X) = \sigma_X^2 = \sum_{i=1}^{k} (x_i - \mu)^2 p_i = \sum_{i=1}^{k} x_i^2 p_i - \mu^2.$$

La *deviazione standard* è $\sigma_X = \sqrt{\text{var}(X)}$. Il coefficiente di variazione è $CV = \sigma/\mu$ (supposto $\mu > 0$).

- Il valore atteso e la varianza hanno le proprietà fondamentali:

$$E(a + bX)a + bE(X), \quad \text{var}(a + bX) = b^2 \text{var}(X).$$

- Una prova di Bernoulli è un esperimento che produce due soli possibili risultati: *successo* e *insuccesso*.
- La variabile casuale *Bernoulli* è una variabile X che vale 0 se il risultato di una prova di Bernoulli è un insuccesso e vale 1 se è un successo. Ha distribuzione

x	0	1	Totale
$p(x)$	$1-p$	p	1

dove p è la *probabilità di successo*.

- La v.c. di Bernoulli ha media e varianza

$$\mu = p, \qquad \sigma^2 = p(1-p) = pq.$$

- Una successioni di *prove di Bernoulli indipendenti e identiche* è una una successione di prove di Bernoulli indipendenti l'una dall'altra e con la stessa probabilità di successo.
- Il numero di successi S in una successione di prove di Bernoulli indipendenti e identiche è $S = X_1 + X_2 + \cdots + X_n$ dove le X_i sono Bernoulli indipendenti tutte con probabilità di successo p.
- *Distribuzione del numero di successi*: S è una variabile casuale detta *Binomiale* che assume valori $s = 0, 1, \ldots, n$ e probabilità

$$p(s) = \binom{n}{s} p^s (1-p)^{n-s}$$

Si scrive $S = Bin(n, p)$ dove n e p sono i parametri della distribuzione Binomiale.

- La media e la varianza della Binomiale sono

$$E(S) = np, \qquad \text{var}(S) = np(1-p).$$

5.1

Considera la seguente distribuzione di probabilità. Calcola le probabilità mancanti.

Modalità	Probabilità %	Prob. Cumulata %
1	4	4
2	60	64
3	16	80
4	*	92
5	4	96
6	*	100

Calcola il valore atteso.

Soluzione

$$P(X = 4) = P(X \leq 4) - P(X \leq 3) = 0.92 - 0.80 = 0.12.$$
$$P(X = 6) = 1 - P(X \leq 5) = 1 - 0.96 = 0.04.$$

Quindi

Modalità	Probabilità %	Prob. Cumulata %
1	4	4
2	60	64
3	16	80
4	12	92
5	4	96
6	4	100

Il valore atteso è

$$E(X) = 1 \cdot 0.04 + 2 \cdot 0.6 + 3 \cdot 0.16 + 4 \cdot 0.12 + 5 \cdot 0.04 + 6 \cdot 0.04 = 2.64.$$

5.2

Data la seguente distribuzione di probabilità

x	0	1	2	3	4	5	6	7
P(x)	0.05	0.16	0.19	0.24	0.18	0.11	0.03	0.04

Quali delle seguenti affermazioni è vera?

A) $P(X \geq 3) = 0.64$
B) $P(2 < X < 5) = 0.42$
C) $P(X > 6) = 0.07$
D) $P(X \leq 6) = 0.93$

Soluzione

A) $P(X \geq 3) = 0.24 + 0.18 + 0.11 + 0.03 + 0.04 = 0.6$. Falsa
B) $P(2 < X < 5) = 0.24 + 0.18 = 0.42$. Vera
C) $P(X > 6) = 0.04$. Falsa
D) $P(X \leq 6) = 1 - 0.04 = 0.96$. Falsa.

5.3

Il numero di volte che uno studente ripete l'esame di statistica è una variabile casuale X con distribuzione

x	1	2	3	4
P(x)	0.5	0.25	0.15	0.1

Calcolare la probabilità che uno studente:

a) ripeta l'esame più di una volta.
b) ripeta l'esame almeno 2 volte.
c) ripeta l'esame al massimo 2 volte.

Soluzione

$$a) = P(X > 1) = 1 - P(X = 1) = 1 - 0.5 = 0.5.$$
$$b) = P(X \geq 2) = P(X > 1) = 0.5$$
$$c) = P(X \leq 2) = P(X = 1 \cup X = 2) = P(X = 1) + P(X = 2) = 0.75.$$

5.4

Calcolare il valore atteso e la deviazione standard di X.

Soluzione

$$E(X) = 1 \cdot 0.5 + 2 \cdot 0.25 + 3 \cdot 0.15 + 4 \cdot 0.1 = 1.85.$$
$$E(X^2) = 1 \cdot 0.5 + 4 \cdot 0.25 + 9 \cdot 0.15 + 16 \cdot 0.1 = 4.45.$$
$$\text{var}(X) = E(X^2) - E(X)^2 = 4.45 - 1.85^2 = 1.0275.$$

Quindi $\sigma(X) = \sqrt{1.0275} = 1.014$.

5.5

Supponi di avere un urna con i numeri $\{1, 2, 3\}$ e di pescarne 2 SENZA ripetizione. Descrivi lo spazio campionario che contiene tutti i campioni di due elementi senza ripetizione.

Soluzione

$$S = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$$

5.6

Considera ora la variabile casuale X = somma dei numeri estratti. Determina la sua distribuzione di probabilità e calcolane il valore atteso.

Soluzione

x	3	4	5
P(x)	1/3	1/3	1/3

$E(X) = 3(1/3) + 4(1/3) + 5(1/3) = 4$.

5.7

Si tirano 4 monete. Qual è la probabilità che escano tutte teste?

Soluzione

Posto $p = P(1 \text{ successo in } 1 \text{ prova}) = 0.5$. Allora $P($ 4 successi su 4 prove indipendenti$) = p^4 = 0.5^4 = 0.0625$.

5.8

Si tirano 4 monete. Qual è la probabilità non esca mai croce?

Soluzione

$$P(0 \text{ successi su } 4 \text{ prove indipendenti}) = (1 - p)^4 = 0.5^4 = 0.0625.$$

5.9

Per andare da Piazza del Popolo a Piazza Italia ci sono 4 semafori indipendenti ognuno dei quali è verde con probabilità 0.3. Qual è la probabilità che guidando da PP a PI non si trovi mai un semaforo verde?

Soluzione

Successo = semaforo verde. $p = P($ 1 successo in una prova$) = 0.3$.

$$P(\text{0 successi su 4 prove indipendenti}) = (1-p)^4 = 0.7^4 = 0.2401.$$

5.10

Si lancia 2 volte una moneta truccata per cui $P(T) = 0.2$ e $P(C) = 0.8$. Considerate la variabile casuale X = numero di teste. Definite la sua distribuzione di probabilità e calcolate il valore atteso E(X) e la varianza.

Soluzione

```
x                    P(x)
0            0.8^2 = 0.64
1        2*0.8*0.2 = 0.32
2            0.2^2 = 0.04
                     1.00
```

$$E(X) = 0 \cdot 0.64 + 1 \cdot 0.32 + 2 \cdot 0.04 = 0.4$$
$$E(X^2) = 0 \cdot 0.64 + 1 \cdot 0.32 + 4 \cdot 0.04 = 0.48$$
$$\text{var}(X) = 0.48 - 0.4^2 = 0.32$$

5.11

Sia X una variabile casuale binomiale con $n = 12$ e $p = 0.4$. Allora

A) La X ha due mode in $X = 5$ e $X = 4$
B) La X ha due mode in $X = 5$ e $X = 6$.
C) La X ha una moda in $X = 5$.
D) La X ha una moda in $X = 6$.

Soluzione

A) La X ha due mode in $X = 5$ e $X = 4$ è falso perché la binomiale è sempre UNIMODALE

B) La X ha due mode in $X = 5$ e $X = 6$. Falso.

C) La X ha una moda in $X = 5$. Qui occorre vedere se la maggiore probabilità si ha per $X = 5$ o $X = 6$.

$$P(X = 5) = \binom{12}{5}(0.4)^5(0.6)^7 = 0.227$$
$$P(X = 6) = \binom{12}{6}(0.4)^6(0.6)^6 = 0.176.$$

Quindi VERO.

D) La X ha una moda in $X = 6$. Falso di conseguenza.

5.12

Una squadra di operai edili deve essere composta da due muratori e da quattro manovali, scelti da un totale di cinque muratori e di sei manovali. Le selezioni dei muratori e dei manovali sono indipendenti. Quante diverse combinazioni sono possibili?

Soluzione

Numero di modi con cui possiamo prendere 2 muratori da 5: $\binom{5}{2} = 10$

Numero di modi con cui possiamo prendere 4 manovali da 6: $\binom{6}{4} = 15$.

Quindi i diversi modi sono $10 \cdot 15 = 150$.

5.13

In una scatola contenente 16 cioccolatini, 4 sono con ripieno al cocco. Qual è la probabilità che scegliendo 4 cioccolatini, nessuno sia con ripieno al cocco?

A) 0.272
B) 0.264
C) 0.248
D) 0.236

Soluzione

Dimensione popolazione $N = 16$. Successi = cocco = 4, Insuccessi = 12.

$X = \#$ cioccolatini al cocco su un campione senza ripetizione di 4.

$$P(X = 0) = \binom{4}{0}\binom{12}{4}/\binom{16}{4} = 1 \cdot 495/1820 = 0.272.$$

Quindi è vera la A).

5.14

Un test a risposta multipla ha 5 domande, ognuna con 5 possibili risposte. Se rispondi sempre a caso, qual è la probabilità di rispondere correttamente a esattamente 3 domande?

A) 0.00032
B) 0.008
C) 0.0512
D) 0.0016

Soluzione

Sono $n = 5$ prove di Bernoulli indipendenti ciascuna con probabilità di successo $p = 1/5$. Se $X =$ numero successi, la probabilità di $X = 3$ è binomiale

$$P(X = 3) = \binom{5}{3}(0.2)^3(0.8)^2 = (10)(0.008)(0.64) = 0.0512.$$

5.15

La probabilità che una persona prenda il raffreddore durante l'inverno è 0.4. Si selezionano a caso 10 persone. Qual è la probabilità che esattamente 4 di loro prenderanno il raffreddore?

Soluzione

Sono $n = 10$ prove di Bernoulli indipendenti con probabilità di successo $p = P(\text{prendere il raffreddore}) = 0.4$. Notare che la selezione a caso è equivalente a un campione con ripetizione di 10 elementi da una popolazione molto grande.

Perché con ripetizione? perché non si dice la dimensione della popolazione supponendola infinita. Quindi si pone $X =$ numero di successi in 10 prove.

$$P(X = 4) = \binom{10}{4}(0.4)^4(0.6)^6 = 210 \cdot 0.0256 \cdot 0.046656 = 0.251.$$

5.16

In un laghetto ci sono 10 pesci di cui 2 sono rossi. Peschi a caso senza ripetizione 5 pesci. Qual è la probabilità di pescare 1 pesce rosso?

Soluzione

La popolazione dei pesci è finita e ci dice che $N =$ dimensione della popolazione $= 10$. Si estrae un campione SENZA ripetizione di 5 pesci. Se $X =$ numero di pesci rossi (successi) nel campione

$$P(X = 1) = P(1 \text{ pesce rosso nel campione}) = \binom{2}{1}\binom{8}{4}/\binom{10}{5} = (2)(70)/(252) = 0.55555.$$

5.17

In un laghetto ci sono 10 pesci di cui 2 sono rossi. Peschi a caso con ripetizione 5 pesci. Qual è la probabilità di pescare 1 pesce rosso?

Soluzione

Stavolta il campione è CON ripetizione. Quindi si usa la Binomiale con n = 5 e p = probabilità di estrarre un pesce rosso dal lago = 2/10 = 0.2.

$$P(X = 1) = \binom{5}{1}(0.2)^1(0.8)^4 = (5)(0.2)(0.4096) = 0.4096.$$

5.18

Una macchina produce pezzi difettosi con probabilità 0.2. Prendi un lotto di 5 pezzi: qual è la probabilità di trovare 1 pezzo difettoso?

Soluzione

La selezione di un lotto di 5 pezzi equivale a un campione con ripetizione perché la popolazione è infinita (nota che non si dà la dimensione della popolazione di pezzi che possono essere prodotti). Quindi si usa la binomiale con n = 5 = dimensione del campione e $p = P(\text{difettoso}) = 0.2$.

$$P(X = 1) = \binom{5}{1}(0.2)^1(0.8)^4 = 0.4096$$

5.19

Tiro 3 dadi. Qual è la probabilità che la somma sia 3? Qual è la probabilità di ottenere tre 1?

Soluzione

La somma può essere 1 solo se tutti e tre i dadi danno un 1.

$$P(\text{somma} = 3) = P(1 \text{ primo dado e } 1 \text{ secondo e } 1 \text{ terzo}) = P(1)P(1)P(1)$$

perché si suppone che i tre lanci diano risultati indipendenti.

Poiché $P(1) = 1/6$, la probabilità che la somma sia 3 è $(1/6)(1/6)(1/6) = 0.0046$.

5.20

Estraggo un campione casuale senza ripetizione di 100 elettori da una popolazione in cui vi è il 30% di favorevoli a Renzi. Qual è la probabilità che il campione contenga 35 persone favorevoli a Renzi?

Soluzione

La popolazione non si sa quanto sia grande e quindi si suppone che sia infinita o molto grande. Quindi non c'è differenza tra un campione senza e con ripetizione. Allora posso usare la binomiale con $n = 100, p = P(\text{estrarre un favorevole a Renzi}) = 0.3$. Se $X =$ numero di favorevoli a Renzi su 100, abbiamo

$$P(X = 35) = \binom{100}{35}(0.3)^{35}(0.7)^{65} = 0.04677968$$

Ma qui ci vuole un calcolatore perché ad esempio

$$\binom{100}{35} = 109506715318796288646116 5020$$

5.21

Quale dei seguenti è un esempio di variabile casuale discreta?

A) L'ammontantare di pioggia che cade in un intervallo temporale di 24 ore.
B) Il peso di un pacco all'ufficio postale.
C) La distanza che puoi percorrere con un pieno di benzina.
D) Il numero di vacche in una fattoria.

Soluzione

Solo D) Le altre sono misure e quindi variabili continue.

6 Variabili casuali doppie

1. Distribuzioni doppie
2. Funzione di probabilità congiunta, marginale, condizionata
3. Indipendenza
4. Covarianza e correlazione
5. Incorrelazione
6. L'incorrelazione non implica l'indipendenza
7. Valore atteso di una somma $X + Y$
8. Varianza di una somma $X + Y$

- *Distribuzione congiunta*: Due v.c. X e Y con modelità x_i e y_j hanno una distribuzione di probabilità congiunta definita da
$$p_{ij} = P(X = x_i, Y = y_j).$$
- *Distribuzioni marginali*: La distribuzione di probabilità di X e Y separate sono
$$p_{i+} = \sum_{j=1}^{J} p_{ij}, \qquad p_{+j} = \sum_{i=1}^{I} p_{ij}.$$
- *Indipendenza*: due v.c. X e Y sono indipendenti se
$$p_{ij} = p_{i+}p_{+j}, \quad \text{per ogni } i, j$$
- *Covarianza*: La covarianza tra X e Y è una misura di associazione lineare definita da
$$\text{cov}(X, Y) = \sigma_{XY} = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_i - \mu_X)(y_j - \mu_Y)p_{ij} = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_i \cdot y_j)\, p_{ij} - \mu_X\mu_Y$$
- Risulta sempre $\sigma_{XY}^2 \leq \sigma_X^2\sigma_Y^2$.
- Se $\sigma_{XY} > 0$ X e Y hanno una associazione lineare positiva.
- Se $\sigma_{XY} < 0$ X e Y hanno una associazione lineare negativa.
- Se $\sigma_{XY} = 0$ X e Y non una associazione lineare e si dicono *incorrelate*. Questa condizione è molto più debole dell'indipendenza.
- Se X e Y sono indipendenti allora sono per forza incorrelate.
- *Coefficiente di correlazione lineare*. È una misura di associazione lineare normalizzata che quindi misura la forza della associazione. Si indica con
$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X\sigma_Y}}.$$
- Risulta sempre $-1 \leq \rho_{XY} \leq 1$.
- Se $\rho_{XY} = 1$ X e Y sono perfettamente dipendenti in modo lineare e la retta ha una pendenza positiva. Se $\rho_{XY} = -1$ X e Y sono perfettamente dipendenti in modo lineare e la retta ha una pendenza negativa.
- *Somma di due variabili*: $T = X + Y$ è una variabile casuale. La sua media e la sua varianza sono
$$E_{X+Y} = \mu_X + \mu_Y, \qquad \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}.$$
- Se X e Y sono incorrelate la varianza di $T = X + Y$ è la somma delle varianze $\sigma_T^2 = \sigma_X^2 + \sigma_Y^2$.
- La deviazione standard di $T = X + Y$ è
$$\sigma_T = \sqrt{\sigma_X^2 + \sigma_Y^2} \neq \sigma_X + \sigma_Y.$$
- *Combinazione lineare di variabili*: $T = c_1X + c_2Y$ è una variabile casuale. Il suo valore atteso e la sua varianza sono
$$\mu_T = c_1\mu_X + c_2\mu_Y, \quad \sigma_T^2 = c_1^2\sigma_X^2 + c_2^2\sigma_Y^2 + 2c_1c_2\sigma_{XY}.$$

6.1

Se la distribuzione di X,Y ha varianze var$(X) = 10$, var$(Y) = 5$ e coefficiente di correlazione 0.4 qual è la varianza di $X - Y$?

Soluzione

$$\begin{aligned}\text{var}(X-Y) &= \text{var}(X+(-1)Y) \\ &= \text{var}(X) + \text{var}[(-1)Y] + 2\text{cov}(X,(-1)Y) \\ &= \text{var}(X) + (1)^2\text{var}(Y) + 2(-1)\text{cov}(X,Y) \\ &= \text{var}(X) + \text{var}(Y) - 2\text{cov}(X,Y) \\ &= 10 + 5 - 2(0.4)\sqrt{10}\sqrt{5} \\ &= 9.343.\end{aligned}$$

6.2

Data la distribuzione congiunta

	X	
Y	1	2
0	0.0	0.6
1	0.4	0.0

Calcolare

- La covarianza e il coefficiente di correlazione
- La media e la varianza di W = 2X - 4Y

Soluzione

X e Y sono legate linearmente perché $X = 2-Y$. È una relazione lineare decrescente. Quindi $\text{cor}(X,Y) = -1$. Infatti:

$$\begin{aligned}E(X) &= 1(0.4) + 2(0.6) = 0.4 + 1.2 = 1.6. \\ E(Y) &= 0(0.6) + 1(0.4) = 0.4. \\ E(XY) &= (0)(1)(0) + (0)(2)(0.6) + (1)(1)(0.4) + (1)(2)(0) = 0.4. \\ \text{cov}(X,Y) &= E(XY) - E(X)E(Y) = 0.4 - (1.6)(0.4) = -0.24. \\ E(X^2) &= (1)(0.4) + (4)(0.6) = 2.8 \\ E(Y^2) &= (0)(0.6) + (1)(0.4) = 0.4 \\ \text{var}(X) &= E(X^2) - E(X)^2 = 2.8 - 1.6^2 = 0.24 \\ \text{var}(Y) &= E(Y^2) - E(Y)^2 = 0.4 - 0.4^2 = 0.24\end{aligned}$$

Quindi $\text{cor}(X,Y) = -0.24/sqrt((0.24)(0.24)) = -1$.

La media e la varianza di W sono

$$\begin{aligned}E(W) &= E(2X-4Y) = 2E(X) - 4E(Y) = (2)(1.6) - (4)(0.4) = 1.6 \\ \text{var}(W) &= 4\text{var}(X) + 16\text{var}(Y) - 16\text{cov}(X,Y) \\ &= (4)(0.24) + (16)(0.24) - (16)(-0.24) = 8.64.\end{aligned}$$

6.3

Un portafoglio comprende 20 azioni ALFA e 30 azioni BETA. Il prezzo delle azioni ALFA è una variabile casuale con media 10 e varianza 9, il prezzo delle azioni BETA è una variabile casuale con media 25 e varianza

16. I prezzi delle due azioni sono correlati negativamente con un coefficiente di correlazione lineare pari a 0.4.

Calcolare il valore atteso e la varianza del valore del portafoglio.

Soluzione

Poiché $X = 20A + 30B$

$$\begin{aligned} E(X) &= 20E(A) + 30E(B) = (20)(10) + (30)(25) = 950. \\ \text{var}(X) &= 400\text{var}(A) + 900\text{var}(B) + (2)(20)(30)\text{cor}(A,B)\sqrt{\text{var}(A)\text{var}(B)} \\ &= (400)(9) + (900)(16) + (1200)(-0.4)(3)(4) = 12240. \end{aligned}$$

7 Variabili casuali continue

1. La distribuzione uniforme
2. La distribuzione normale
3. Quantili della Normale
4. Disuguaglianza di Chebyshev e regola empirica

- Una *variabile continua* X ha modalità x che appartengono all'insieme dei numeri reali $\mathbb{R} = (-\infty, +\infty)$.
- È definita da una *funzione di densità* positiva $f(x)$ tale che

$$\int_{\mathbb{R}} f(x)dx = 1.$$

- La probabilità che X appartenga a un intervallo $[a, b]$ è

$$\Pr(a \leq X \leq b) = \Pr(X \leq b) - \Pr(X \leq a) = \int_a^b f(x)dx.$$

- La *funzione di ripartizione* è $F(x) = \Pr(-\infty \leq X \leq x)$.
- Il *valore atteso* di X è $\mu_X = \int_{\mathbb{R}} xf(x)dx$.
- La *varianza* di X è $\sigma_X^2 = \int_{\mathbb{R}}(x - \mu_x)^2 f(x)dx$.
- La *variabile uniforme* su un intervallo $[a, b]$, $X \sim U(a, b)$ ha funzione di densità

$$f(x) = 1/(b - a)$$

  se $x \in [a, b]$ e $f(x) = 0$ altrove.
- La variabile uniforme ha valore atteso e varianza

$$\mu_X = (a + b)/2, \qquad \sigma_X^2 = (b - a)^2/12.$$

- La *variabile normale standard* $Z \sim N(0, 1)$ ha densità

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad z \in \mathbb{R}$$

- Il valore atteso e la varianza della normale standard sono rispettivamente 0 e 1.
- La *variabile normale* generale $X \sim N(\mu, sigma^2)$ si ottiene con la strasformazione di Z

$$X = \mu + \sigma Z.$$

- La $N(\mu, \sigma^2)$ ha valore atteso μ e varianza σ^2.
- La *funzione di ripartizione* della normale standard $F(z) = \Pr(Z \leq z)$ è tabulata in fondo al libro.
- I *quantili superiori* della normale standard sono i valori z_α tali che

$$\Pr(Z > z_\alpha) = \alpha$$

- La *disuguaglianza di Chebyshev* dice che *per qualunque variabile casuale* con media μ e deviazione standard σ, la probabilità che X differisca dalla media per k deviazioni standard, cioè che

$$\mu - k\sigma \leq X \leq \mu + k\sigma$$

  è sempre superiore al limite $1 - \frac{1}{k^2}$.
- Se X è normale la probabilità precedente si può calcolare esattamente usando le tavole.

7.1

Se X ha distribuzione uniforme tra 2 e 5, qual è la probabilità che X assuma valori tra 3 e 4?

Soluzione

$$P(3 < X < 4) = P(X < 4) - P(X < 3) = 2/3 - 1/3 = 1/3$$

7.2

Trova il valore $\Pr(Z \leq 0.67)$ se Z è normale standard.

Soluzione

Dalle tavole si vede

```
    z   F(z)
 0.66 0.7454
 0.67 0.7486
 0.68 0.7517
 0.69 0.7549
 0.70 0.7580
```

Quindi $\Pr(Z \leq 0.67) = F(0.67) = 0.7486$. Nota che sul libro è indicato con .7686 omettendo lo zero iniziale.

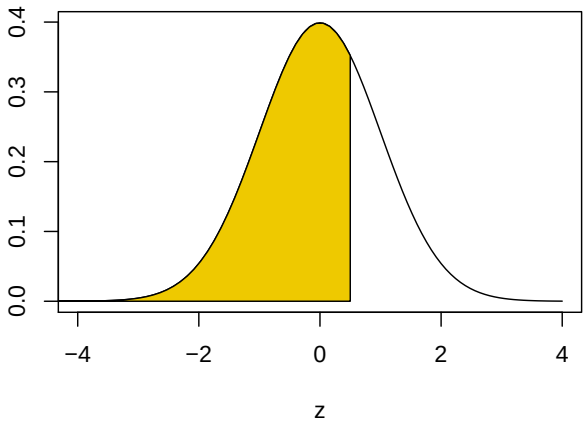
7.3

Se $Z \sim N(0, 1)$ calcola le probabilità (disegnando il grafico!)

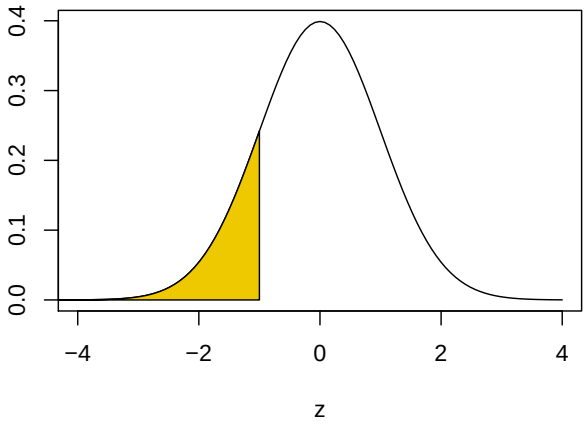
- $P(Z < 0.5)$
- $P(Z < -1)$
- $P(0 < Z < 0.5)$
- $P(1 < Z < 2)$
- $P(-0.5 < Z < 1)$
- $P(-2 < Z < 0)$
- $P(-2 < Z < -1)$
- $P(Z > 1)$
- $P(Z > -1)$

Soluzione

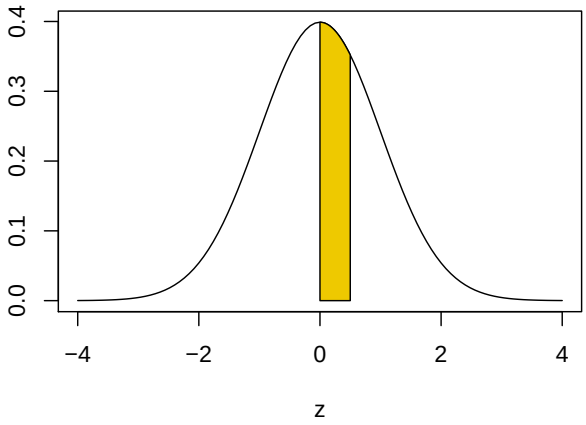
- $P(Z < 0.5) = 0.6915$ (vedi le tavole).



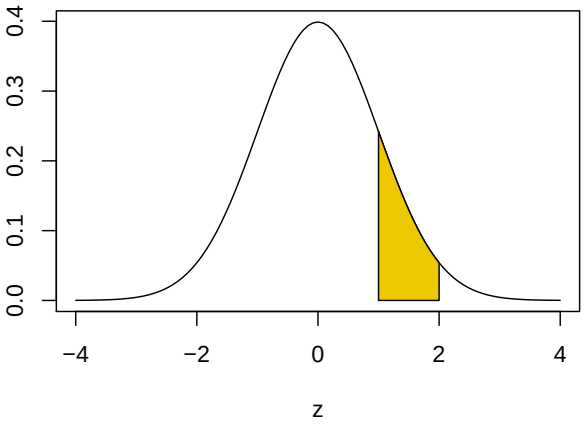
- $P(Z < -1) = 1 - P(Z < 1) = 1 - 0.8413 = 0.1587$



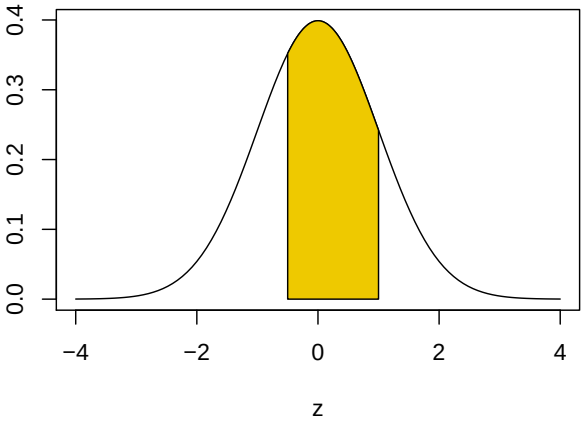
- $P(0 < Z < 0.5) = P(Z < 0.5) - P(Z < 0) = 0.6915 - 0.5 = 0.1915$



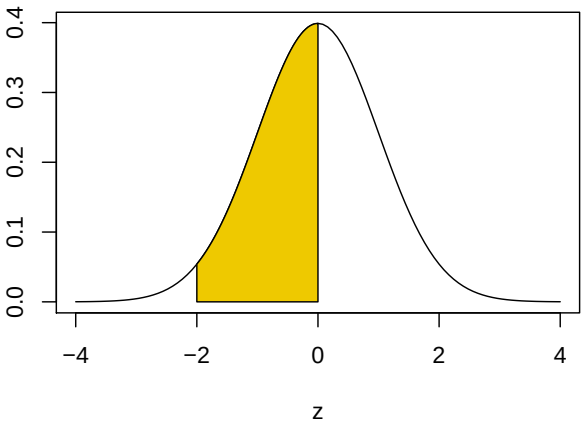
- $P(1 < Z < 2) = P(Z < 2) - P(Z < 1) = 0.9772 - 0.8413 = 0.1359$



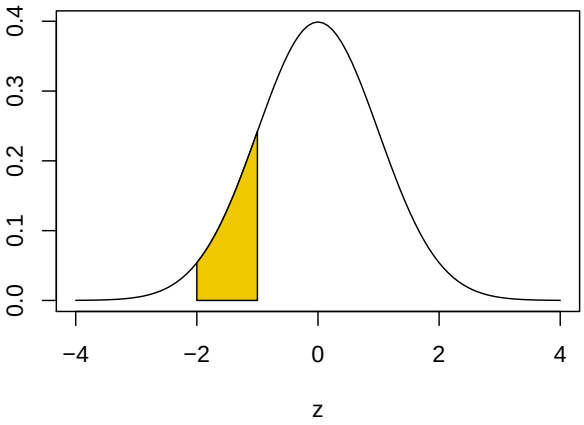
- $P(-0.5 < Z < 1) = P(Z < 1) - (1 - P(Z < 0.5)) = 0.8413 - (1 - 0.6915) = 0.5328$



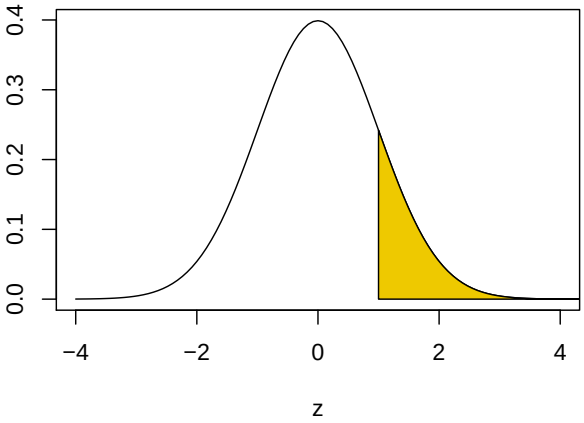
- $P(-2 < Z < 0) = P(0 < Z < 2) = P(Z < 2) - 0.5 = 0.9772 - 0.5 = 0.4772$



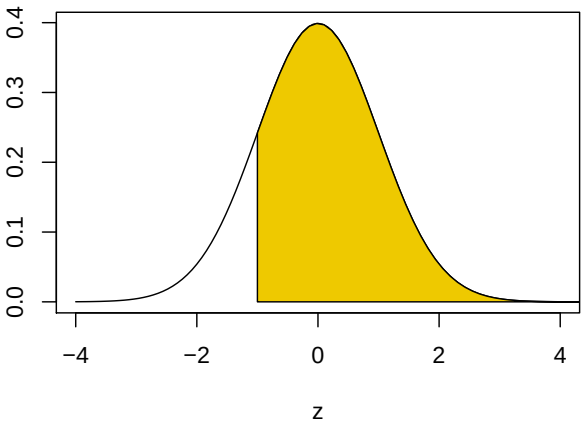
- $P(-2 < Z < -1) = P(1 < Z < 2) = 0.1359$



- $P(Z > 1) = 1 - P(Z < 1) = 0.1587$



- $P(Z > -1) = P(Z < 1) = 0.8413$.



7.4

Sia Z una normale standard. Trovare sulle tavole della Normale quel valore z^* tale che

$$P(X < z^*) = 0.75$$

NOTA: scegliere il valore z^* che corrisponde al valore più vicino possibile a 0.75.

NOTA: il valore risultante è il *terzo quartile* della Normale standard.

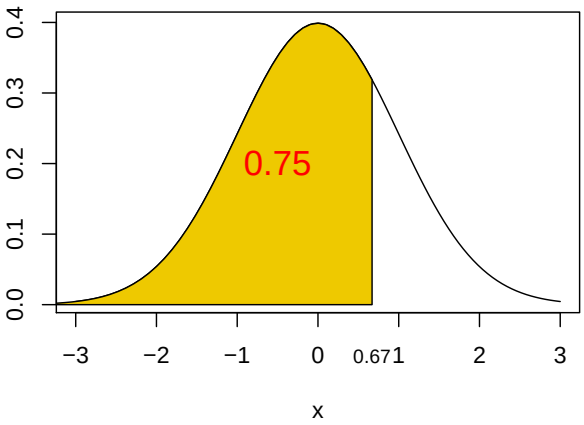
Soluzione

Dalle tavole si vede

```
    z   F(z)
 0.66 0.7454
 0.67 0.7486
 0.68 0.7517
 0.69 0.7549
 0.70 0.7580
```

Quindi $F(0.67) = 0.7486$ e $F(0.68) = 0.7517$. Quindi il valore più vicino a 0.75 è il primo. Quindi prendiamo $z^* = 0.67$,

$$P(Z < 0.67) \simeq 0.75 \text{ e quindi } Q_3 = 0.67.$$



7.5

Trova il *primo* quartile della normale standard.

Soluzione

Per simmetria $Q_1 = -0.67$

$$P(Z < -0.67) = 1 - P(Z < 0.67) = 1 - 0.75 = 0.25.$$

7.6

Trova lo scarto interquartile della normale standard.

Soluzione

$$SIQ = Q_3 - Q_1 = 0.67 - (-0.67) = 1.34$$

7.7

Sia $Z \sim N(0, 1)$. Qual è quel valore z^* che **è superato** dal 25% delle Z?

Soluzione

È il valore tale che $P(Z > z^*) = 0.25$ cioè tale che

$$P(Z < z^*) = 1 - 0.25 = 0.75$$

cioè $z^* = 0.67$.

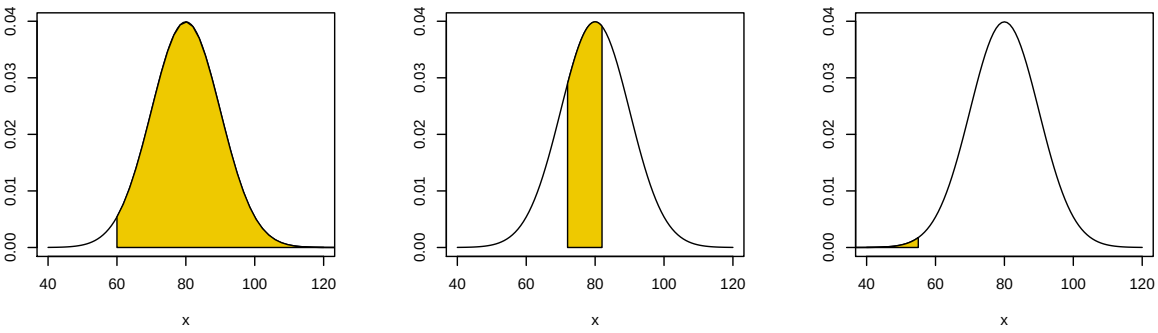
7.8

Se $X \sim N(80, \sigma^2 = 100)$ calcola

- $P(X > 60)$
- $P(72 < X < 82)$
- $P(X < 55)$

Soluzione

$$\begin{aligned} P(X > 60) &= P(Z > (60-80)/10) = P(Z > -2) = P(Z < 2) = 0.9772. \\ P(72 < X < 82) &= P(X < 82) - P(X < 72) \\ &= P(Z < (82-80)/10) - P(Z < (72-80)/10) \\ &= P(Z < 0.2) - P(Z < -0.8) = 0.5793 - (1 - P(Z < 0.8)) \\ &= 0.5793 - (1 - 0.7881) = 0.3674. \\ P(X < 55) &= P(Z < (55-80)/10) \\ &= P(Z < -2.5) = 1 - P(Z < 2.5) = 1 - 0.9938 = 0.0062. \end{aligned}$$



7.9

Sempre per $X \sim N(80, \sigma^2 = 100)$ determina il valore x^* tale che

$$P(X > x*) = 0.10$$

Soluzione

Prima cosa occorre trasformare il problema

$$P(X > x^*) = 0.10 \text{ è equivalente a } P(X < x*) = 0.90$$

Poi si standardizza:

$$P(Z < (x^* - 80)/10) = 0.90.$$

Quindi il quantile 0.9 della normale standard è 1.28 (dalle tavole). Cioè

$$(x^* - 80)/10 = 1.28$$

da cui risolvendo

$$x^* = 80 + (10)(1.28) = 92.8.$$

7.10

Determina l'intorno centrato nella media tale che la probabilità che X assuma valori all'*esterno* sia 0.05.

Soluzione

Equivale a dire: trova il valore k tale che la probabilità che X stia tra $80-k$ e $80+k$ sia 0.95. Cioè

$$P(80-k<X<80+k)=0.95.$$

Bisogna perciò risolvere

$$P(X<80+k)-P(X<80-k)=0.95$$

Siccome la normale è simmetrica rispetto alla media

$$P(X<80-k)=1-P(X<80+k)$$

Quindi

$$\begin{aligned} P(X<80+k)-(1-P(X<80+k))&=0.95\\ 2P(X<80+k)-1&=0.95\\ P(X<80+k)&=0.975 \end{aligned}$$

Quindi standardizzando

$$P(Z<(80+k-80)/10)=0.975$$

Il quantile 0.975 della normale standard è 1.96. Perciò

$$(80+k-80)/10=1.96$$

da cui risolvendo si ottiene $k=19.6$.

7.11

Le previsioni sulla domanda di un prodotto sono una variabile normale X con media 1200 e deviazione standard 100.

- Qual è la probabilità che le vendite superino 1000?
- Qual è la probabilità che le vendite stiano fra 1100 e 1300?
- Qual è il valore delle vendite x^* che ha probabilitå 0.10 di essere superato?

Soluzione

$X=$ vendite.

- $P(X>1000)=P(Z>(1000-1200)/100)=P(Z>-2)=P(Z<2)=0.9772.$
- $$\begin{aligned} P(1100<X<1300)&=P((1100-1200)/100<Z<(1300-1200)/100)\\ &=P(-1<Z<1)=P(Z<1)-P(Z<-1)\\ &=P(Z<1)-(1-P(Z<1))=2P(Z<1)-1=2(0.8413)-1=0.6826. \end{aligned}$$
- Qual è il numero di unità vendute che ha probabilità 0.1 di essere superato? Se $X\sim N(1200,1002)$ sono le vendite si deve calcolare il valore d tale che $P(X>d)=0.1$.

Devi usare le tavole all'inverso. Quindi si standardizzano ambo i membri:

$$P(Z>d1200100)=0.1$$

Sulle tavole trovo che $P(Z\leq 1.28)\simeq 0.9$ e quindi $P(Z>1.28)\simeq 0.1$. Perciò $d1200100=1.28$.

Dunque il valore d cercato è $d=1200+1.28\text{»}100=1328$.

7.12

Una variabile X ha una distribuzione con media 250 e e deviazione standard 20. Dare indicazioni sulla probabilità:

- $P(210 < X < 290)$
- $P(220 < X < 280)$

Determinare le stesse probabilità sapendo che X è normale $N(250, \sigma = 20)$.

Soluzione

Se la distribuzione di X è ignota si usa la disuguaglianza di Chebyshev. Entrambi gli intervalli hanno come punto centrale la media della distribuzione.

Per esempio, nel primo caso $(210, 290)4 ha punto centrale $(210 + 290)/2 = 250$.

Si trova il raggio dell'intorno di 250 che è $(290 - 210)/2 = 40$. Quindi l'intervallo è 250 ± 40.

Quindi si esprime il raggio come un multiplo della deviazione standard cioè $40 = k20$ da cui si ottiene $k = 2$ (il raggio 40 è il doppio della deviazione standard).

Infine si usa la disuguaglianza:

$$P(210 < X < 290) \geq 1 - 1/(k^2) = 1 - 1/4 = 0.75.$$

Nel secondo caso il raggio è $(280 - 220)/2 = 60/2 = 30$ e poiché $30 = k20$ implica $k = 1.5$. Quindi abbiamo

$$P(220 < X < 280) \geq 1 - 1/(k^2) = 1 - 1/(1.5^2) = 0.555$$

Sapendo che X è normale le probabilità si possono ottenere usando le tavole.

- $P(210 < X < 290) = P(-2 < Z < 2) = 2(0.9772) - 1 = 0.9544$
- $P(220 < X < 280) = P(-1.5 < Z < 1.5) = 2(0.9332) - 1 = 0.8664$

8 Stima e stimatori

1. Campioni casuali e distribuzioni campionarie
2. Stima
3. Stimatori corretti
4. Errore standard
5. Teorema centrale del limite
6. Intervalli di confidenza per la media (varianza nota)
7. Intervalli di confidenza per la media (varianza incognita)
8. Distribuzione t e tavole dei quantili
9. Intervalli di confidenza asintotici per una proporzione
10. Ampiezza dell'intervallo di confidenza
11. Scelta della dimensione campionaria

- Un campione casuale da una popolazione con una distribuzione X è un'n-upla di osservazioni $(x_1, \ldots, x_n)$ che si può equiparare alle realizzazioni di n variabili aleatorie $X_1, \ldots, X_n$ che siano
  - indipendenti
  - identicamente distribuite come X.

- I campioni casuali sono importanti per stimare una caratteristica della popolazione senza fare un censimento.
- La caratteristica della popolazione X si chiama *parametro*. Per esempio μ o σ^2 sono parametri di una popolazione continua. Invece una proporzione di successi p è un parametro di una popolazione binaria (composta di 0 e 1).
- Il parametro si stima con i dati campionari $x_1, \ldots, x_n$. Le stime più comuni sono
  - $\bar{x} = \sum_{i=1}^{n} x_i/n$ la media campionaria stima μ.
  - $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$ stima la varianza σ^2.
  - $\hat{p}$ = proporzione di successi nel campione, stima p.
- Uno *stimatore* è la stima nel campionamento ripetuto. Si descrive con una variabile aleatoria che esprime il variare della stima nell'*universo dei campioni*. Gli stimatori corrispondenti alle stime precedenti sono
  - Lo stimatore di μ: $\bar{X} = \sum_{i=1}^{n} X_i/n$
  - Lo stimatore di σ^2: $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$.
  - Lo stimatore di p: $\hat{P} = \sum_{i=1} X_i/n$.
- Uno stimatore è una variabile aleatoria con una distribuzione detta *distribuzione campionaria*.
- Dato un campione casuale iid $(X_1, \ldots, X_n)$ da una popolazione X continua con una media μ si dice *stimatore media campionaria* $\bar{X} = (X_1 + \cdots + X_n)/n$.
- Lo stimatore $\bar{X}$ è rappresentato dalla sua *distribuzione campionaria* cioè dalla distribuzione ottenuta calcolando la media in ogni campione dell'*universo dei campioni*.
- Lo stimatore va distinto dalla *stima* della media cioè il semplice dato $\bar{x} = (x_1 + \cdots + x_n)/n$ indicato con la lettera minuscola.
- Lo stimatore *media campionaria* $\bar{X} = (X_1 + \cdots + X_n)/n$ è *corretto* per μ cioè

  $$E(\bar{X}) = \mu$$

  qualunque sia la popolazione e qualunque sia μ.
- Dati due stimatori corretti di μ si dice che T_1 è più *efficiente* di T_2 se

  $$\text{var}(T_1) \leq \text{var}(T_2)$$

  Per esempio la media campionaria $\bar{X}$ è più efficiente della mediana campionaria in campioni provenienti da una distribuzione binomiale.
- Lo stimatore $\bar{X}$ ha varianza σ^2/n e la sua deviazione standard, chiamata *errore standard* nel contesto della stima, è

  $$ES = \frac{\sigma}{\sqrt{n}}.$$

- Se la popolazione è normale lo stimatore $\bar{X}$ ha distribuzione normale $N(\mu, \sigma^2/n)$ esattamente, qualunque sia la dimensione del campione.
- Se la popolazione non è normale, ma la dimensione del campione è sufficientemente grande (> 100) lo stimatore $\bar{X}$ ha distribuzione approssimata da $N(\mu, \sigma^2/n)$ (teorema centrale del limite).
- Data una popolazione dicotomica, cioè di 0 e 1, con una proporzione di 1 (i successi) pari a p, e un campione casuale iid $(X_1, \ldots, X_n)$ si dice *stimatore proporzione campionaria* $\hat{P} = (X_1 + \cdots + X_n)/n = \#successi/n$.
- Lo stimatore $\hat{P}$ è rappresentato dalla sua distribuzione campionaria che è esattamente Binomiale divisa per n.

- La proporzione campionaria $\hat{P}$ è uno stimatore corretto di p, cioè

$$E(\hat{P}) = p$$

  qualunque sia p e per ogni numerosità n.

- La varianza della proporzione $\hat{P}$ è $p(1-p)/n$.
- L'errore standard di $\hat{P}$ è

$$ES(\hat{P}) = \sqrt{p(1-p)/n}.$$

- la distribuzione campionaria di una proporzione $\hat{P}$ è approssimativamente normale se $np(1-p) > 9$

$$\hat{P} \approx N(p, p(1-p)/n).$$

- Gli errori standard dipendono dai parametri incogniti e quindi vengono stimati opportunamente.
  - $\hat{ES}(\bar{X}) = s/\sqrt{n}$
  - $\hat{ES}(\hat{P}) = \sqrt{\hat{p}(1-\hat{p})/n}$

Questo fornisce una misura calcolabile dell'errore di campionamento che si commette nella stima del parametro.

- La distribuzione di $\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}$ è normale standard.
- La distribuzione di $\frac{\bar{X}-\mu}{\sqrt{s^2/n}}$ dove

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  è *t di Student* con $n-1$ gradi di libertà. I suoi quantili sono tabulati sul libro. Quando $n > 100$ la t di Student si può approssimare con una normale standard.

- Un *intervallo di confidenza* per μ al livello c (tipicamente $c = 0.95$ o $c = 0.99$) è un intervallo di stima con estremi A e B tali nell'universo dei campioni una proporzione c di campioni produce intervalli che contengono μ:

$$P(A < \mu < B) = c.$$

- Un intervallo di confidenza di livello $1-\alpha$ per la media di una distribuzione normale *con varianza nota* è

$$\bar{X} \pm z \cdot \sigma/\sqrt{n}$$

  dove z è tale che $P(Z < z) = 1 - \alpha/2$ ossia $P(Z > z) = \alpha/2$ e Z è la normale standard.

- Un intervallo di confidenza di livello $1-\alpha$ per la media di una distribuzione normale è

$$\bar{X} \pm t \cdot \sqrt{s^2/n}$$

  dove $s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$. Il valore t è un quantile della t di Student, cioè è tale che $P(T < t) = 1 - \alpha/2$ ossia $P(T > t) = \alpha/2$ in cui T è la t di Student con $n-1$ gradi di libertà.

- Un intervallo di confidenza di livello approssimato $1-\alpha$ per la media di una distribuzione normale con un campione di dimensione elevata è

$$\bar{X} \pm z \cdot \sqrt{s^2/n}$$

  dove z è tale che $P(Z < z) = 1 - \alpha/2$ ossia $P(Z > z) = \alpha/2$ e Z è normale standard.

- Un intervallo di confidenza di livello approssimato $1-\alpha$ per la proporzione di una popolazione dicotomica in cui $n\hat{p}(1-\hat{p}) > 9$ è

$$\hat{p} \pm z \cdot \sqrt{\hat{p}(1-\hat{p})/n}$$

  dove z è tale che $P(Z < z) = 1 - \alpha/2$ ossia $P(Z > z) = \alpha/2$ e Z è normale standard.

8.1

Un'azienda produce un modello di auto la cui percorrenza X (in km con 1 litro di benzina) ha distribuzione normale, media 25 km/l e deviazione standard 2 km/l. Supponiamo di avere un campione casuale di 4 auto prodotte in serie.

- La percorrenza media campionaria che distribuzione ha?
- Qual è la probabilità che la percorrenza media sia superiore a 26 km/l?
- Ricalcolare la probabilità precedente con una dimensione campionaria di 25 auto.

Soluzione

$X \sim N(25, \sigma = 2)$. Campione $= (X_1, X_2, X_3, X_4)$ indipendenti e identicamente distribuiti come X.

- La percorrenza media è $\bar{X} = (X1 + X2 + X3 + X4)/4$ ed ha distribuzione normale $N(25, \sigma_{\bar{X}} = 2/\sqrt{4} = 1)$.
- $P(\bar{X} > 26) = P(Z > (26 - 25)/1) = P(Z > 1) = 1 - P(Z < 1) = 1 - 0.8413 = 0.1587$.
- $P(\bar{X} > 26) = P(Z > (26 - 25)/(2/5)) = P(Z > 2.5) = 1 - P(Z < 2.5) = 1 - 0.9938 = 0.0062$.

8.2

Una popolazione di studenti è composta dal 40% di femmine e dal 60 % di maschi. Se si estrae un campione casuale con ripetizione di 25 studenti qual è la distribuzione di probabilità della proporzione di femmine nel campione? Qual è la varianza della proporzione di femmine nel campione?

Soluzione

$X \sim$ *Bernoulli* con $p = 0.4, q = 0.6$. Perciò $\text{var}(X) = pq = 0.24$.

Campione di $n = 25$ elementi $(X_1, \ldots, X_{25})$.

\$P = \$ Proporzione di femmine = #femmine $/25 = (X_1 + ... + X_{25})/25$.

$P * 25$ è distribuita come una Binomiale$(25, p = 0.4)$. Sappiamo allora che

$$\text{var}(P) = \text{var}(X)/n = 0.24/25 = 0.0096.$$

8.3

Sia X la distribuzione dell'età di una popolazione con $E(X) = 50$ anni e $\sigma(X) = 10$ anni. Se seleziono un campione di $n = 4$ persone e calcolo la media:

- Si conosce la distribuzione campionaria dell'età media?
- Si conosce il valore atteso della distribuzione campionaria?
- Si conosce la varianza della distribuzione campionaria?

Giustificare.

Soluzione

$X =$ età $\sim$ incognita $?(\mu = 50, \sigma = 10)$.

Campione: (X_1, X_2, X_3, X_4) indipendenti e identicamente distribuiti come X.

$\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$.

- Si conosce la distribuzione campionaria della media? NO è incognita.
- Si conosce il valore atteso della distribuzione campionaria della media? SÌ, è $E(\bar{X}) = \mu = 50$ anni.
- Si conosce la varianza della distribuzione campionaria? SÌ, è $\text{var}(\bar{X}) = \sigma^2/n = 100/4 = 25$.

8.4

Rispondere all'esercizio precedente se n = 100.

Soluzione

- Si conosce la distribuzione campionaria della media? SÌ, poiché la dimensione del campione è grande ha approssimativamente distribuzione normale.
- Si conosce il valore atteso della distribuzione campionaria della media? SÌ, è $E(\bar{X}) = \mu = 50$ anni.
- Si conosce la varianza della distribuzione campionaria? SÌ, è $\text{var}(\bar{X}) = \sigma^2/n = 100/100 = 1$.

8.5

Il numero di televisori che escono ogni giorno da una certa linea di produzione si distribuisce come una variabile casuale con deviazione standard (nota) di 17.4. La media giornaliera della linea di produzione determinata su un campione di 20 giorni è 452.3. Quale dei seguenti intervalli rappresenta un intervallo di confidenza al 95% per la media della produzione in un giorno?

A) 453 ± 9.4
B) 452.3 ± 13.8
C) 452.3 ± 11.3
D) 452.3 ± 7.63

Soluzione

$X = \#$ televisioni $\sim N(\mu = ?, \sigma = 17.4)$.

NOTA: il sigma fornito dal testo è la deviazione standard **della popolazione**.

La stima di μ con un campione di $n = 20$ elementi è 452.3 con un errore standard $ES = \sigma/\sqrt{20} = 17.4/\sqrt{20} = 3.890758$. L'intervallo di confidenza (IC) al 95% è

$$452.3 \pm ME$$

con un margine di errore

$$ME = 1.96\, ES = (1.96)(3.890758) = 7.625886.$$

Quindi la risposta è (arrotondando a 2 decimali) la D).

8.6

L'errore di stima è la differenza tra il valore di una statistica determinata su un campione ed il corrispondente valore del parametro determinato nella popolazione. Vero o falso?

Soluzione

L'errore di stima è la differenza tra uno stimatore e il parametro. Il termine "statistica" è sinonimo di stimatore, e vuol dire un indice calcolato sul campione.

Quindi la risposta è: Vero.

8.7

Il tempo che gli studenti dedicano allo studio segue una distribuzione normale con deviazione standard di 8 ore. Si estrae un campione casuale di 4 studenti. La probabilità che la media campionaria **differisca dalla media della popolazione** per più di 4 ore è

A) 0.2987 B) 0.3080 C) 0.3174 D) 0.3085

Soluzione

Tempo = $X \sim N(\mu =?, \sigma = 8)$.

Se ho un campione di dimensione $n = 4$, la media campionaria $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4$ ha distribuzione

$$\bar{X} \sim N(\mu =?, \sigma_{\bar{X}} = 8/2 = 4).$$

La probabilità che $\bar{X}$ differisca da μ per più di 4 ore è

$$P(|\bar{X} - \mu| > 4) = 1 - P(-4 < \bar{X} - \mu < 4) = 1 - P(\mu - 4 < \bar{X} < mu + 4)$$

NOTA: Fate attenzione a queste disuguaglianze, studiatele con calma.

Quindi se si standardizza $\bar{X}$ rispetto alla sua media μ e alla sua deviazione standard = 4 si ha

$$P(|\bar{X} - \mu| > 4) = 1 - P(-1 < Z < 1) = 1 - (0.8413 - (1 - 0.8413)) = 0.3174$$

e quindi la risposta corretta è la C).

8.8

Uno stimatore è una variabile casuale calcolata su un campione casuale che fornisce la stima puntuale per il parametro della popolazione. Vero o falso?

Soluzione È esattamente così: la stima è un numero, mentre lo stimatore è una variabile casuale calcolata sul campione che fornisce una stima del parametro della popolazione. Quindi: Vero.

8.9

Un intervallo di confidenza al 95% per la media della popolazione μ è stimato da 65.48 a 76.52. Se ora viene stimato un intervallo di confidenza al 90% per μ sarà:

A) più ampio di quello al 95%.
B) lo stesso dell'intervallo al 95%.
C) più stretto di quello al 95%.
D) Non c'è abbastanza informazione per rispondere.

Soluzione

Un intervallo di confidenza è

$$\bar{X} \pm ME$$

dove ME, il margine di errore è

$$ME = z_{\alpha/2}\sigma/\sqrt{n}.$$

L'ampiezza dell'intervallo cresce o decresce con ME. Se il livello di confidenza cambia, ME cambia perché cambia $z_{\alpha/2}$. Allora, ad esempio abbiamo:

Livello	$1-\alpha$	$\alpha/2$	$z_{\alpha/2}$
	95%	2.5%	1.96
	90%	5%	1.64

Quindi se il livello è 90% il ME è più piccolo e l'intervallo di confidenza è meno ampio. La risposta giusta è C).

8.10

Un'agenzia turistica è interessata all'ammontare medio di denaro speso al giorno da un tipico studente universitario durante le vacanze estive. Un'indagine condotta su 30 studenti mette in luce che la somma media spesa è 63.57 Euro con una deviazione standard di 17.32 Euro. Determinare l'intervallo di confidenza al 95% per la spesa media nella popolazione.

Soluzione

Spesa giornaliera di uno studente = $X \sim N(\mu =?, \sigma =?)$. Da un campione di $n = 30$ studenti si sa che

$$\bar{X} = 63.57 \text{ Euros} = 17.32 \text{ Euro}$$

NOTA: la deviazione standard fornita è quella del campione NON quella della popolazione.

Quindi l'IC per μ è basato sulla t di Student con $n - 1 = 29$ gradi di libertà:

$$\bar{X} \pm t_{\alpha/2} s/\sqrt{n}$$

ossia

$$63.57 \pm (2.045)17.32/\sqrt{30}$$

Cioè IC = $(57.10333, 70.03667)$.

8.11

Da una popolazione infinita con media pari a 80 e deviazione standard 18, vengono selezionati campioni casuali di dimensione $n = 36$. La media e l'errore standard della relativa distribuzione campionaria della media sono rispettivamente:

A) 80 e 18. B) 80 e 3. C) 36 e 2. D) 80 e 2.

Soluzione

La media $\bar{X}$ di un campione iid di $n = 36$ elementi da una QUALSIASI distribuzione X è dale che

$$E(\bar{X}) = \mu, \quad \text{var}(\bar{X}) = \sigma^2(X)/n, \quad ES(\bar{X}) = \sigma(X)/\sqrt{n}$$

Quindi in questo caso

$$E(\bar{X}) = 80, \quad ES(\bar{X}) = 18/6 = 3.$$

Quindi la soluzione è B).

8.12

Se l'errore standard della distribuzione della proporzione campionaria è 0.0229 per campioni di dimensione 400, allora la vera proporzione nella popolazione deve essere:

A) 0.2 o 0.8.
B) 0.3 o 0.7.
C) 0.4 o 0.6.
D) 0.5 o 0.5.

Soluzione

Una proporzione campionaria P ha un errore standard

$$ES(P) = \sqrt{pq/n}$$

poichè questo deve essere 0.0229 basta provare nei casi A), B), C) , D) che cosa otteniamo:

A) $ES = \sqrt{(}0.2 \cdot 0.8/400) = 0.02$
B) $ES = \sqrt{(}0.3 \cdot 0.7/400) = 0.02291288$
C) $ES = \sqrt{(}0.4 \cdot 0.6/400) = 0.0244949$
D) $ES = \sqrt{(}0.5 \cdot 0.5/400) = 0.025$

Quindi, approssimando, la risposta giusta è B).

8.13

Nel costruire un intervallo di confidenza per la media della popolazione è stato utilizzato un campione di 40 osservazioni. La stima intervallare risultante è stata 28.76 ± 1.48. Se la numerosità campionaria fosse stata 160 invece che 40, la stima intervallare sarebbe stata:

A) 28.76 ± 0.74.
B) 28.76 ± 0.37.
C) 7.19 ± 0.37.
D) 7.19 ± 1.48.

Soluzione Siccome 160 è una numerosità quadrupla di quella di partenza di n = 40, l'ampiezza dell'intervallo di confidenza è la metà perché l'errore standard è

$$ES = \sigma/\sqrt{4n} = (1/2)\sigma/\sqrt{n}.$$

quindi il margine di errore con $n = 160$ è $ME = 1.48/2 = 0.74$. Quindi la risposta giusta è A).

NOTA: Naturalmente qui si suppone che sia nel caso $n = 40$ che nel caso $n = 160$ la media campionaria sia sempre la stessa.

8.14

Siano X_1, X_2, X_3 e X_4 le osservazioni di un campione casuale semplice estratto da una popolazione X con media μ e varianza σ^2. Si consideri il seguente stimatore di μ:

$$T = 0.15X_1 + 0.35X_2 + 0.20X_3 + 0.30X_4.$$

Qual è la varianza di T?

Soluzione La varianza di T è

$$\begin{aligned}
\text{var}(T) &= \text{var}(0.15X_1 + 0.35X_2 + 0.20X_3 + 0.30X_4) \\
&= 0.15^2\text{var}(X_1) + 0.35^2\text{var}(X_2) + 0.20^2\text{var}(X_3) + 0.30^2\text{var}(X_4) \\
&= (0.15^2 + 0.35^2 + 0.20^2 + 0.30^2)\text{var}(X) \\
&= 0.275\sigma^2.
\end{aligned}$$

8.15

La distribuzione campionaria della media avrà la stessa media della popolazione dalla quale sono stati estratti i campioni che l'hanno generata. Vero o falso?

Soluzione

È vero perché la media campionaria è uno stimatore corretto di μ:

$$E(\bar{X}) = \mu.$$

8.16

Trova il quantile della t di Student con 8 gradi di libertà che lascia a destra una probabilità di 0.025.

Soluzione

$$t(0.025) = 2.306$$

8.17

Si supponga che il tempo medio che un ragazzo passa su Facebook sia distribuito come una variabile normale con una deviazione standard di 1.5 ore. In un campione di 100 ragazzi è stata rilevata una media di 6.5 ore. Determinare l'intervallo di confidenza al 95% per il tempo medio passato su Facebook nella popolazione.

Soluzione

Tempo passato su Facebook = $X \sim N(\mu = ?, \sigma = 1.5)$. L'intervallo di confidenza è basato sulla normale perché la varianza della popolazione è nota.

$$IC = \bar{X} \pm ME$$

ossia

$$IC = 6.5 \pm 1.96(1.5)/10$$

da cui IC = $(6.206, 6.794)$.

8.18

Un ricercatore, incaricato di stimare la percentuale di famiglie italiane che hanno più di un computer, dopo aver rilevato che il 27% di un campione costituito da 492 famiglie ha dichiartato di possedere più di un computer, fornisce l'intervallo di confidenza $(0.2308; 0.3092)$, ma omette di dire il livello di confidenza. Qual è il livello di confidenza associato a questo intervallo?

Soluzione Un intervallo di confidenza per p (approssimato per grandi campioni) è

$$\hat{p} \pm z_{\alpha/2} ES \quad \text{dove} \quad ES = \sqrt{\hat{p}\hat{q}/n}.$$

Il margine di errore è $ME = z_{\alpha/2} ES$ e quindi $z_{\alpha/2} = ME/ES$.

Qui $\hat{p}$ è la proporzione stimata che ovviamente è il punto centrale dell'intervallo cioè

$$\hat{p} = (0.2308 + 0.3092)/2 = 0.27.$$

L'errore standard è

$$ES = \sqrt{0.27 * 0.73/492} = 0.02.$$

Il margine di errore è la lunghezza di mezzo intervallo e lo sappiamo:

$$ME = 0.3092 - 0.27 = 0.0392.$$

Perciò

$$z_{\alpha/2} = ME/ES = 0.0392/0.02 = 1.96$$

Siamo fortunati! A occhio sappiamo che il livello di confidenza è il 95%.

Ecco tutti i passaggi:

$$1 - \alpha/2 = P(Z < 1.96) = 0.975$$

Quindi $\alpha = 0.05$ e $1 - alpha$ = livello di confidenza = 0.95.

8.19

Determinare l'ampiezza campionaria necessaria per stimare la proporzione p nella popolazione se $ME = 0.05$ e il livello di confidenza è il 99%.

Soluzione

Deve essere $ME = 2.58 sqrt(p(1-p)/n)$.

Siccome p non è noto si prende il caso peggiore (variabilità massima) con $p = 0.5$.

Quindi

$$ME = 2.58 \cdot \sqrt{0.5^2/n}$$

da cui $0.05^2 = (2.58^2)(0.5^2)/n$ e quindi

$$n = (2.58^2)(0.5^2)/(0.05^2) = 665.64.$$

8.20

Si intervista un campione casuale di 220 famiglie e si rileva che il 58.7% legge la pubblicità postale. Trovare l'intervallo di confidenza al 99% per la proporzione di destinatari che legge la pubblicità postale nella popolazione.

Soluzione

Con 220 famiglie si usa l'approssimazione normale. Il quantile appropriato è $z_{\alpha/2} = 2.58$ dove $\alpha = 0.01$. Quindi

$$ME = 2.58 \cdot \sqrt{0.587 * (1 - 0.587)/220} = 0.0856$$

Quindi l'intervallo è 0.587 ± 0.0856 ovvero $(0.501, 0.673)$.

8.21

La quantità di stoffa usata per produrre poltrone è distribuita come una variabile casuale normale. Su un campione casuale di 15 poltrone, si è riscontrato che l'ammontare medio del materiale è 912 centimetri quadrati, con una deviazione standard di 64 centimetri quadrati. Quali dei seguenti intervalli rappresenta l'intervallo di confidenza al 99% per la media della quantità di materiale?

A) 912 ± 44.3
B) 912 ± 42.6
C) 912 ± 49.2
D) 912 ± 46.8

Campione di $n = 15$. $\bar{X} = 912$ ed $s = 64$ (deviazione standard campionaria).

L'intervallo è basato sulla t di Student con $n - 1 = 14$ gradi di libertà. Quindi con $\alpha = 0.01$ il quantile è $t_{\alpha/2} = 2.977$.

L'errore standard è $ES = s/\sqrt{15} = 64/\sqrt{15} = 16.52473$.

Quindi il margine di errore è

$$ME = 2.977 \cdot 16.52473 = 49.19 \simeq 49.2$$

Dunque la risposta è la C).

NOTA: l'errore tipico nei compiti è

$$ME = 2.58 \cdot 16.52473 = 42.6$$

dove 2.58 è il quantile della normale (che in questo caso è sbagliato).

8.22

Si supponga che il tempo trascorso dai clienti in un negozio sia distribuito in modo normale con media incognita e deviazione standard pari a 6 minuti. Si supponga di aver stimato il tempo medio della popolazione tramite un intervallo di confidenza al 95% e di aver ottenuto il seguente risultato: $(22.06, 27.94)$. Qual è stata la dimensione del campione necessaria ad ottenere il precedente intervallo di confidenza?

Soluzione Come visto il margine di errore per un intervallo di livello 95% è $ME = 1.96ES$ e qui

$$ME = \text{semiampiezza dell'intervallo} = 2.94.$$

e inoltre

$$ES = \sigma/\sqrt{n} = 6/\sqrt{n}$$

Quindi si imposta l'equazione

$$2.94 = 1.966/\sqrt{n}$$

e si risolve con

$$2.94^2 = (1.96^2 \cdot 6^2)/n$$

da cui

$$n = (1.96^2 \cdot 6^2)/(2.94^2) = 16.$$

9 Test delle ipotesi

- Introduzione ai test delle ipotesi
- Errore del I e II tipo
- test sulla media con varianza nota
- test sulla media con varianza incognita
- Test su una proporzione
- p-value
- Potenza del test

9.1

Se si rifiuta l'ipotesi nulla contro l'ipotesi alternativa ad un livello di significatività del 5% , allora, con gli stessi dati deve essere rifiutata anche ad un livello di significatività dell'1%. Vero o Falso?

Soluzione

Falso. Per esempio può capitare che in un test con la normale la statistica stia tra 1.96 e 2.56. In questo caso si rifiuta al 5% ma non all'1%.

9.2

Un idraulico afferma di poter completare l'installazione di un box doccia in meno di un'ora. Per un campione di 24 interventi, l'idraulico impiega una media di 63.2 minuti con una deviazione standard di 7.7 minuti. Qual è la statistica test osservata?

A) t = 1.79
B) t = 2.04
C) Z = 2.04
D) Z = 1.79

Soluzione

Ipotesi: $H_0 : \mu \leq 60$ contro $H_1 : \mu > 60$. Il parametro $\mu_0 = 60$. Abbiamo le statistiche $\bar{X} = 63.2$ e $s = 7.7$, con $n = 24$.

La statistica test è

$$t = \frac{\bar{x} - \mu 0}{s/\sqrt{n}} = \frac{63.2 - 60}{7.7/\sqrt{24}} = 2.03594$$

che arrotondato a 2 decimali è $t = 2.04$. La risposta è B) e non C) perché la statistica è una t di Student con 23 gradi di libertà.

9.3

Il valore atteso della media della popolazione è dato dalla media campionaria. Vero o Falso?

Soluzione

Falso. La verità è che $E(\bar{X}) = \mu$. Qui si dice invece che $E(\mu) = \bar{X}$ e questo non è assolutamente vero.

9.4

La ditta produttrice di un nuovo macchinario afferma che il suo macchinario incrementerà la produzione per macchina di almeno 29 unità di prodotto all'ora. Vengono acquistate 15 nuove macchine e si trova che l'incremento di produzione medio ottenuto è pari a 26 pezzi per macchina all'ora con una deviazione standard di 4.2. C'è evidenza empirica sufficiente per dubitare dell'affermazione fatta dal produttore dei nuovi macchinari considerando un livello di significatività $\alpha = 0.05$?

Soluzione

Affermazione: $\mu \geq 29$. Dato che c'è l'uguale (almeno...) questa è H0. Quindi

$$H_0 : \mu \geq 29 \text{ contro } H_1 : \mu < 29$$

Inoltre $X \sim N(\mu =?, \sigma =?)$ entrambi incogniti.

Il test è unilaterale. Al livello del 5% si rifiuta se

$$t = (\bar{x} - \mu 0)/(s/\sqrt{n}) < -1.761$$

distribuita come una t di Student con $n - 1 = 14$ gradi di libertà e probabilità a sinistra = 0.05.

Dati: $n = 15$; $\bar{x} = 26, s = 4.2$.

Quindi $t = (26 - 29)/(4.2/\sqrt{15}) = -2.766$ e si rifiuta perché $t < -1.761$. Quindi la risposta è SÌ c'è evidenza empirica per dubitare al livello del 5%.

9.5

Un professore sostiene che il punteggio medio in un certo test è stato almeno 83. Si assuma che il punteggio al test si distribuisca normalmente. Tu ritieni che invece il punteggio medio sia inferiore ad 83, per cui decidi di chiedere ad un campione casuale di studenti il loro voto e risulta:

$$82, 77, 85, 76, 81, 91, 70, 82.$$

Verifica che la media e la varianza corretta sono: 80.5 e $s^2 = 39.71429$.

Ritieni sia lecito dubitare dell'affermazione del professore ad un livello di significatività del 5%?

Soluzione

Si ha $n = 8$, la media è

$$\bar{x} = (82 + 77 + 85 + 76 + 81 + 91 + 70 + 82)/8 = 80.5.$$

La varianza campionaria corretta è

$$s^2 = ((82-80.5)^2+(77-80.5)^2+(85-80.5)^2+(76-80.5)^2+(81-80.5)^2+(91-80.5)^2+(70-80.5)^2+(82-80.5)^2)/(8-1) = 39.7142$$

Quindi la deviazione standard è $s = \sqrt{39.71429} = 6.3$.

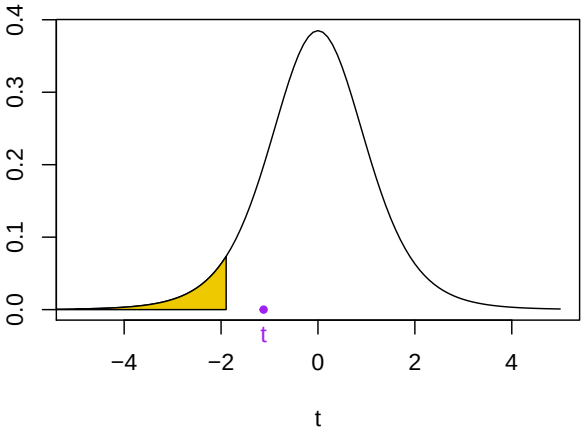
Il sistema di ipotesi è

$$H_0 : \mu \geq 83 \text{ contro } H_1 : \mu < 83$$

La regione critica unilaterale al livello del 5% è $t < -1.895$ (quantile inferiore t di Student con 7 gradi di libertà al 5%)

Statistica test $= (80.5 - 83)/(6.3/\sqrt{8}) = -1.122$.

Quindi non si rifiuta al livello del 5% perché $-1.122 > -1.895$.



Non si dubita dell'affermazione al livello del 5%.

9.6

Supponi di voler effettuare con un livello di significatività $\alpha = 0.10$ il seguente test sulla media di una popolazione:

$$H_0 : \mu = 277 \text{ contro } H_1 : \mu \neq 277.$$

Supponi inoltre di sapere che la deviazione standard della popolazione è $\sigma = 13.5$. Se selezioni un campione casuale di 20 osservazioni, per quale valore della media campionaria rifiuterai l'ipotesi nulla?

Soluzione

È un test per la media di una normale con deviazione standard nota = 13.5. Campione di dimensione $n = 20$. L'errore standard è

$$ES = \sigma/\sqrt{n} = 13.5/\sqrt{20} = 3.018692.$$

Sia $z^* = 1.645$ il valore tale che $P(Z > z^*) = 0.05$ (vedi le tavole della t di Student, ultima riga).

Si rifiuta se

$$(\bar{X} - \mu 0)/ES > 1.645 \quad \text{oppure} \quad (\bar{X} - \mu 0)/ES < -1.645$$

cioè se

$$\bar{X} > \mu 0 + 1.645ES \quad \text{oppure} \quad \bar{X} < \mu 0 - 1.645ES$$

ossia se

$$\bar{X} > 277 + 1.645(3.018692) = 281.9657$$

oppure

$$\bar{X} < 277 - 1.645(3.018692) = 272.0343$$

9.7

Si verifica un errore del I tipo quando viene rifiutata un'ipotesi nulla vera.Vero o Falso?

Vero. È la definizione: errore del I tipo = rifiutare H_0 quando H_0 è vera. Si tratta di sbagliare gli innocenti per colpevoli.

9.8

Un'azienda produttrice di caffè asserisce che ciascun lotto contiene almeno 50.1 kg di prodotto. Si assuma che la deviazione standard della quantità di caffè contenuta in ciascun lotto sia 1.2 kg. La regola di decisione adottata dall'azienda è di fermare le consegne se la media campionaria della quantità di caffé in un campione di 40 lotti è inferiore a 49.7. Qual è la probabilità di commettere un errore del primo tipo?

Soluzione

Ipotesi nulla $H_0 : \mu \geq 50.1$, alternativa $H_1 : \mu < 50.1$.

Siccome il testo dice "si assuma" vuol dire che $1.2 = \sigma$ la deviazione standard della popolazione. Si usano campioni di numerosità $n = 40$.

La regola di decisione definisce la regione critica: ferma le consegne se si rifiuta H_0 cioè se $\bar{X} < 49.7$.

$$P(I) = P(\text{Rifiutare } H_0, \text{ quando } \mu = 50.1) = P(\bar{X} < 49.7 \text{ quando } \mu = 50.1)$$

Quindi poiché sotto $H0$ $\bar{X} \sim N(\mu = 50.1, \sigma = 1.2/\sqrt{40} = 0.1897367)$ risulta

$$P(I) = P(\bar{X} < 49.7) = P(Z < (49.7 - 50.1)/0.1897367) = P(Z < -2.11) = 1 - P(Z < 2.11) = 0.0174.$$

Cioè il livello del test è 0.0174.

9.9

Associa al simbolo β la definizione opportuna.

A) La potenza del test.
B) La probabilità dell'errore di II tipo.
C) La probabilità dell'errore di I tipo.
D) La probabilità di rifiutare H_0.

Soluzione

È la probabilità di errore del II tipo cioè di accettare H_0 quando è falsa. Cioè la probabilità di scambiare un colpevole per innocente. Quindi la risposta giusta è B).

9.10

Aumentando il livello di significatività di un test, la probabilità dell'errore del II tipo aumenta. Vero o Falso?

Soluzione

Falso. Infatti il livello del test è α e la probabilità di errore di II tipo è β. Ma è noto che se α aumenta β diminuisce.

9.11

Quale delle seguenti frasi NON è vera?

A) La regione di rifiuto è l'insieme di tutti i valori della statistica test per cui l'ipotesi alternativa viene rifiutata.
B) Una statistica test è una funzione dei dati campionari sulla base della quale si decide se rifiutare o meno l'ipotesi nulla.
C) La regione di rifiuto è l'insieme di tutti i valori della statistica test per cui l'ipotesi nulla viene rifiutata.
D) Una buona procedura di test delle ipotesi deve comportare una probabilità dell'errore del I tipo e del II tipo piccola.

Soluzione

A) è falsa: La regione di rifiuto è l'insieme di tutti i valori della statistica test per cui l'ipotesi NULLA viene rifiutata.
B) è vera
C) è vera (vedi sopra)
D) Ovviamente è vera.

Quindi la risposta è A)

9.12

Quale deve essere la dimensione del campione necessaria per stimare la media di una popolazione distribuita normalmente se $ME = 5$, $\sigma = 40$, livello di confidenza = 99% ?

Soluzione

Deve essere $ME = 2.58\sigma/\sqrt{n}$ quindi

$$5 = 2.58 * 40/\sqrt{n}$$

ossia

$$25 = (2.58^2)(40^2)/n$$

che risolto dà

$$n = (2.58^2)(40^2)/25 = 426.$$

9.13

Il livello di significatività di un test è la probabilità che l'ipotesi nulla sia vera. Vero o Falso?

Soluzione

FALSO. Il livello è la probabilità di rifiutare H_0 quando è vera non la probabilità che H_0 sia vera.

9.14

Si supponga di voler effettuare un test su $H_0 : \mu \geq 0.54$ contro $H_1 : \mu < 0.54$ basato su un campione iid di $n = 25$ da $N(\mu, \sigma^2)$ sapendo che nel campione $s = 13.2$. Quale dovrebbe essere la statistica test? A) $(\bar{X} - 0.54)/2.64$ C) $(\bar{X} - 0.54)/0.528$ B) $(\bar{X} - 0.54)/34.848$ D) $(\bar{X} - 0.54)/0.2789$

Soluzione

Nel test t di Student per la media la statistica test è

$$t = (\bar{X} - \mu 0)/ES$$

dove $ES = s/\sqrt{n}$. Qui $\mu 0 = 0.54$ e $ES = 13.2/5 = 2.64$.

Quindi la soluzione è A).

9.15

Un professore asserisce che il punteggio medio conseguito ad un recente esame è stato 83. Si assuma che la variabile punteggio conseguito si distribuisca normalmente. Tu chiedi ad alcuni in classe quale punteggio abbiano conseguito ed ottieni le seguenti risposte: 82, 77, 85, 76, 81, 91, 70 e 82. Supponi di voler verificare se l'affermazione del professore è corretta contro un alternativa bilaterale.

Quale affermazione tra le seguenti è più appropriata per il p-value?

A) p-value < 0.10
B) p-value < 0.01
C) p-value < 0.05
D) p-value > 0.10

Soluzione

Il sistema di ipotesi è $H_0 : \mu = 83$ contro $H_1 : \mu \neq 83$. La dimensione campionaria è $n = 8$ e la media campionaria è 80.5 e la deviazione standard è $s = 6.3$. Si usa la statistica t di Student (dato che la varianza della popolazione incognita è stimata dai dati) è

$$t = (80.5 - 83)/(6.3/\sqrt{8}) = -1.12$$

(distribuzione t di Student con 7 gradi di libertà sotto H_0).

Il p-value è la probabilità che la statistica test assuma un valore più estremo di quello osservato sotto H_0. Cioè è

$$p = P(T < -1.12) + P(T > 1.12) = 1 - P(-1.12 < T < 1.12).$$

Questa probabilità non si può calcolare esattamente dalle tavole della t. Tuttavia si nota che le regioni critiche classiche sono

al livello dell'1% $t < -3.499, t > +3.499$ al livello dell'5% $t < -2.365, t > 2.365$ al livello del 10% $t < -1.895, t > 1.895$

quindi la probabilità di $t < -1.12$ o $t > 1.12$ deve essere per forza maggiore del 10%. Conclusione: la risposta giusta è la D).

9.16

Supponiamo di voler effettuare una procedura di verifica delle ipotesi su una porporzione p, e che la proporzione campionaria $\hat{p}$ sia approssimativamente normale. Se l'ipotesi alternativa è $H_1 : p \neq p_0$, allora la regione di rifiuto al livello $\alpha = 0.05$ è $Z < -1.96$ o $Z > 1.96$. Vero o Falso?

Soluzione La regione critica di livello $\alpha = 0.05$ è $Z > z_{\alpha/2}$ o $Z < -z_{\alpha/2}$. Dove $Z = (\hat{p} - p_0)/ES(\hat{p})$.
Quindi se $\alpha = 0.05$, $z_{\alpha/2} = 1.96$. Quindi è vero.

9.17

Associa al simbolo $1-\beta$ la definizione opportuna.

A) La probabilità di rifiutare H_0.
B) La probabilità dell'errore di I tipo.
C) La probabilità di corretto rifiuto di H_0.
D) La probabilità dell'errore di II tipo.

Soluzione Poiché β è la probabilità P(II) di errore di secondo tipo cioè la probabilità di accettare H_0 quando è falsa, $1-\beta$ è la probabilità di rifiutare H_0 quando è falsa, cioè la potenza del test. Soluzione C).

9.18

La Regione Veneto ha dichiarato che il reddito medio familiare annuo della regione è superiore a 37000 Euro. Si assuma che la distribuzione del reddito medio familiare della Regione Veneto sia distribuito come una variabile normale con deviazione standard di 5756 euro. Si supponga che in un campione di 25 famiglie si sia rilevato un reddito medio annuo pari a 36243 euro. Quale affermazione tra le seguenti è più appropriata per il p-value?

A) p-value < 0.01
B) p-value < 0.10
C) p-value < 0.05
D) p-value > 0.10

Soluzione

Il problema suppone che il reddito $X \sim N(\mu =?, sigma = 5756)$. L'ipotesi della Regione Veneto è che $\mu = E(X) = E(\text{reddito annuo}) > 37000$ Euro. Poiché non è specificata un'uguaglianza questa è l'ipotesi alternativa. L'ipotesi da verificare è

$$H_0 : \mu \leq 37000 \quad \text{contro} \quad H_1 : \mu > 37000 \text{ (l'ipotesi della Regione)}$$

La statistica test è (con $n = 25$)

$$z = (36243 - 37000)/(5756/\sqrt{25}) = -0.657.$$

Le regioni critiche classiche sono

- 1% $z < -2.58$ o $z > 2.58$
- 5% $z < -1.96$ o $z > 1.96$
- 10% $z < -1.64$ o $z > 1.64$

Quindi il p-value è sicuramente > 0.10. Infatti $z = -0.657$ porterebbe ad accettare H_0 al livello del 10%.

9.19

Un'ipotesi nulla è rifiutata a livello di significatività 0.025, ma non ad un livello di 0.01. Ciò significa che il p-value del test è compreso tra 0.01 e 0.025. Vero o Falso?

Soluzione Se un'ipotesi è rifiutata al livello del 2.5% vuol dire che il p-value è < del 2.5%.

Se un'ipotesi non è rifiutata al livello dell' 1% vuol dire che il p-value è > dell' 1%.

Il p-value p è $0.01 < p < 0.025$. Quindi è vero.

9.20

Un commercialista afferma di poter completare una dichiarazione dei redditi standard in meno di un'ora. Per un campione di 24 dichiarazioni, il commercialista impiega una media di 63.2 minuti con una deviazione standard di 7.7 minuti. Quale affermazione tra le seguenti è più appropriata per il p-value?

A) $0.025 <$ p-value < 0.05
B) p-value < 0.025
C) p-value > 0.05
D) p-value < 0.01

Soluzione

Il tempo X impiegato ha distribuzione $N(\mu =?, \sigma =?)$. L'affermazione è $\mu < 60$ min e quindi è l'ipotesi alternativa. L'ipotesi nulla (che contiene il segno di uguaglianza) è $H_0 : \mu \geq 60$.

Abbiamo un campione di dimensione $n = 24$, $\bar{x} = 63.2$, $s = 7.7$.

Quindi la statistica test è t di Student (con 23 gradi di libertà):

$$t = (63.2 - 60)/(7.7/\sqrt{24}) = 2.034.$$

Le regioni critiche standard sono unilaterali sinistre:

- 1% $t < -2.500$
- 2.5% $t < -2.069$
- 5% $t < -1.714$
- 10% $t < -1.319$

Quindi $t = 2.034 > -1.319$ e quindi il p-value deve essere $> 10\%$ e quindi anche maggiore del 5%.
La risposta quindi è C).

9.21

Aumentando il livello di significatività di un test, la probabilità dell'errore del II tipo aumenta. Vero o Falso?

Soluzione Se il livello α del test aumenta la probabilità di errore di secondo tipo β diminuisce a parità di altri elementi. Quindi è falso.

9.22

Associa al simbolo $1 - \alpha$ la definizione opportuna.

A) La probabilità dell'errore di II tipo.
B) La probabilità dell'errore di I tipo.
C) La probabilità di non rifiutare l'ipotesi nulla quando questa è vera.
D) La potenza del test.

Soluzione

α è la $P(I) =$ probabilità rifiutare H_0 quando H_0 è vera. Quindi $1 - \alpha$ è la probabilità di accettare H_0 quando è vera. Quindi la risposta è C).

9.23

Avete un campione da una normale la cui media può essere 10 o 12 (non si sa quale delle due) e deviazione standard 2. Con un campione di 4 elementi dalla popolazione dovete verificare $H_0 : \mu = 10$ contro $H_1 : \mu = 12$ e pensate di rifiutare H_0 quando la media campionaria è maggiore di 11.5.

- Qual è la probabilità di errore del I tipo?
- Qual è la probabilità di corretto rifiuto (ossia la potenza del test)?

Soluzione

La $P(I)$ = probabilità che la media campionaria sia > 11.5 quando H_0 è vera cioè se $\mu = 10$.

La potenza = probabilità che la media campionaria sia > 11.5 quando H_0 è falsa cioè se $\mu = 12$.

Quindi si possono calcolare sapendo che la media campionaria è

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n} = 2/2 = 1).$$

Abbiamo

$$P(I) = P(\bar{X} > 11.5, \bar{X} \sim N(10, 1)) = P(Z > (11.5 - 10)) = P(Z > 1.5) = 1 - P(Z < 1.5) = 0.0668.$$

Inoltre

$$\text{Potenza} = P(\bar{X} > 11.5, \bar{X} \sim N(12, 1)) = P(Z > (11.5 - 12)) = P(Z > -0.5) = P(Z < 0.5) = 0.6915.$$

9.24

Associa al simbolo β la definizione opportuna:

A) La probabilità dell'errore di I tipo.
B) La probabilità di non rifiutare H_0 vera.
C) La probabilità di accettare un'ipotesi nulla falsa.
D) La potenza del test.

Soluzione

È la probabilità di errore di II tipo e cioè la probabilità di accettare un'ipotesi nulla falsa. Risposta C).

9.25

L'errore del II tipo può essere definito come:

A) Rifiutare un'ipotesi alternativa vera.
B) Non rifiutare un'ipotesi alternativa falsa.
C) Non rifiutare un'ipotesi nulla falsa.
D) Rifiutare un'ipotesi nulla vera.

Soluzione

La definizione di errore del II tipo è accettare (ossia non rifiutare) H0 quando è falsa. Risposta C). È l'errore di scambiare un colpevole con un innocente.

9.26

L'azienda produttrice di sacchi di farina afferma che ciascun sacco contiene almeno 50.1 kg di farina. Si assuma che la deviazione standard della quantità di farina contenuta in ciascun sacco sia 1.21 kg. La regola di decisione adottata dall'azienda è di mettere in manutenzione una macchina riempitrice se la media campionaria della quantità di farina in un campione di 40 sacchi è inferiore a 49.7. Qual è la probabilità di commettere un errore del primo tipo?

Soluzione

- X = quantità di farina nel sacco = aleatoria $\sim N(\mu = ?, \sigma = 1.21)$.

- Test: $H_0 : \mu \geq 50.1, \quad H_1 : \mu < 50.1$
- Regione critica: $\bar{X}$ = media di 40 sacchi (campione casuale) < 49.7
- $P(I) = P(\bar{X} < 49.7$ quando $\mu = 50.1)$

Questa si calcola sapendo che se $\mu = 50.1$ allora

$$\bar{X} \sim N(50.1, \sigma_{\bar{X}} = 1.21/\sqrt{40} = 0.1913178)$$

Quindi

$$P(I) = P(Z < (49.7 - 50.1)/0.1913178) = P(Z < -2.09) = 1 - P(Z < 2.09) = 0.02.$$

9.27

Un produttore di lenti per occhiali sostiene che almeno l'80% degli oculisti preferisce il suo tipo di lenti per occhiali. Decidi di verificare la sua affermazione e, su un campione di 200 oculisti, trovi che il 74.1% preferisce quelle lenti. C'è sufficiente evidenza per dubitare dell'affermazione del produttore? Usa il livello $\alpha = 0.025$.

Soluzione

L'ipotesi nulla è $H_0 : p \geq 0.8$ contro $H_1 : p < 0.8$. In un campione di dimensione $n = 200$, $\hat{p} = 0.741$. La statistica test è

$$z = (0.741 - 0.8)/ES$$

con

$$ES = \sqrt{(}p_0(1 - p_0)/n) = \sqrt{(}0.8(1 - 0.8)/200) = 0.02828427$$

NOTA: l'errore standard viene calcolato SOTTO H_0. Pertanto $z = -2.085$ che sotto H_0 ha una distribuzione approssimata normale (la binomiale è approssimata bene da una normale se $n = 200$).

La regione critica unilaterale al livello 0.025 è (i valori della normale sono all'ultima riga della tavola della t di Student) $z < -1.96$.

Perciò con un valore osservato di $z = -2.085$ si rifiuta H_0. Quindi c'è sufficiente evidenza per dubitare dell'affermazione al livello del 2.5%.