Tracce e soluzioni degli esami di

PROBABILITÀ E STATISTICA [3231]

Corso di Studi: Laurea Triennale in Ingegneria Gestionale
Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Appelli a.a. 2023–2024

Gianluca Orlando

Indice

1 Tracce

Di seguito le tracce dell'a.a. 2023-2043.

Esame di Probabilità e Statistica [3231]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: giugno 2024 - turno 1
Matricola: _____ Data: 17/06/2024

È obbligatorio consegnare la traccia con cognome e nome. In caso contrario, l'esito sarà "RITIRATO". Questa è la traccia n. 120. Scrivere il numero di traccia sullo svolgimento del compito.

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Una studentessa di Probabilità e Statistica vuole determinare se esiste una relazione tra le ore di studio giornaliere nella preparazione dell'esame e il voto all'esame. Per farlo, ha intervistato alcuni studenti e ha ottenuto i seguenti dati:

ore di studio	0	1	2	3	4	5	6
voto	0	12	11	25	26	30	30

1. Rappresentare i dati in uno scatterplot.
2. Calcolare la retta di regressione lineare (derivando le formule) e disegnarla.
3. Calcolare il coefficiente di correlazione lineare e il coefficiente di determinazione R^2.

Esercizio 2. (8 punti) Sia (X, Y) un vettore aleatorio discreto con legge congiunta descritta dalla seguente tabella:

X \ Y	0	1	2
0	0	$\frac{1}{4}$	a
1	$\frac{1}{4}$	0	b
2	c	d	0

Si assuma che:

$$\bullet\ \mathbb{P}(\{X=2\}) = \frac{1}{8}\,, \quad \bullet\ \mathbb{P}(\{Y=1\}|\{X=2\}) = 1\,, \quad \bullet\ \mathbb{E}(XY) = \frac{1}{2}\,.$$

Dopo aver determinato i valori di a, b, c, d, rispondere ai seguenti quesiti:

1. Calcolare la covarianza tra X e Y e stabilire se X e Y sono indipendenti.
2. Calcolare $\mathrm{Var}(X+Y)$.
3. Si supponga di estrarre 14 realizzazioni indipendenti della variabile aleatoria Y. Calcolare la probabilità che l'evento $\{Y=0\}$ si realizzi almeno 3 volte (3 incluso).

4. Si estraggono tante realizzazioni indipendenti della variabile aleatoria Y. In media, qual è la prima volta in cui si realizza l'evento $\{Y = 0\}$?

Esercizio 3. (7 punti) Una catena di fast food analizza il tempo di servizio dei suoi clienti. I clienti possono ordinare due tipi di menu: il menu A e il menu B. Si assuma che

- Il tempo di servizio per il menu A sia distribuito con legge esponenziale con media 2 minuti.
- Il tempo di servizio per il menu B sia distribuito con legge esponenziale con deviazione standard 3 minuti.
- Il 20% dei clienti ordina il menu A e il 80% il menu B.

Si risponda alle seguenti domande:

1. Si consideri un cliente che ha ordinato il menu A. Qual è la probabilità che il suo tempo di servizio sia inferiore a 3 minuti?
2. Si consideri un cliente che ha ordinato il menu B. Qual è la probabilità che il servizio avvenga esattamente al minuto 1?
3. Si consideri un cliente che ha ordinato il menu B. Ha aspettato 2 minuti e non è ancora stato servito. Sapendo questo fatto, qual è la probabilità che debba aspettare in tutto almeno 5 minuti?
4. Un cliente ha fatto un ordine e ha aspettato un tempo compreso tra 1 e 4 minuti. Qual è la probabilità che abbia ordinato il menu A?

Esercizio 4. (7 punti) Una fabbrica di cereali vuole stimare la media del peso delle confezioni prodotte. Un controllo su un campione di alcune confezioni ha fornito i seguenti pesi in grammi:

$$500 \quad 506 \quad 499 \quad 507 \quad 502 \quad 500 \quad 496 \quad 501$$

Si supponga che il peso sia distribuito normalmente.

1. Calcolare sui dati un intervallo di confidenza bilaterale al 90% per la media del peso delle confezioni. (N.B.: derivare le formule)
2. Un intervallo di confidenza bilaterale al 91% calcolato sugli stessi dati è più grande o più piccolo di quello calcolato al punto 1? Perché?

Quesito teorico 1. (4 punti) Spiegare in che senso la legge di Poisson approssima la legge binomiale enunciando e dimostrando un teorema.

Quesito teorico 2. (2 punti) Calcolare media e varianza di $X \sim \mathrm{U}(a, b)$.

Esame di Probabilità e Statistica [3231]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: giugno 2024 - turno 2
Matricola: _____ Data: 17/06/2024

È obbligatorio consegnare la traccia con cognome e nome. In caso contrario, l'esito sarà "RITIRATO". Questa è la traccia n. 120. Scrivere il numero di traccia sullo svolgimento del compito.

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Si studia il tempo di attesa per il servizio clienti di una banca. I dati vengono raggruppati in classi nella seguente tabella:

intervalli (minuti)	frequenze assolute
$[0, 2)$	19
$[2, 5)$	18
$[5, 7)$	5
$[7, 11)$	16
$[11, 18)$	12
$[18, 100)$	10

1. Rappresentare un istogramma delle densità di frequenze relative.
2. Determinare la classe modale (o le classi modali, se più di una).
3. Calcolare un'approssimazione della media e della varianza dei dati.
4. Calcolare un'approssimazione del 55-esimo percentile.

Esercizio 2. (7 punti) Il numero di errori nelle soluzioni degli esercizi scritte dal docente di Probabilità e Statistica segue una distribuzione di Poisson con una media di 2 errori per soluzione. Si assuma che i numeri di errori in soluzioni di esercizi distinti siano indipendenti.

1. Qual è la probabilità che in una soluzione ci siano almeno 4 errori (inclusi)?
2. Qual è la deviazione standard del numero di errori in una soluzione?
3. Qual è la probabilità che in 5 soluzioni ci siano almeno 4 errori (inclusi)?
4. Consideriamo 5 soluzioni. Abbiamo letto le prime 3 soluzioni e abbiamo individuato almeno 4 errori (inclusi). Sapendo questo fatto, qual è la probabilità che nelle 5 soluzioni ci siano in tutto 6 errori?

5. Uno studente legge le soluzioni degli esercizi in sequenza e si blocca quando trova la prima soluzione con almeno 1 errore (incluso). Qual è la probabilità che lo studente si blocchi entro la lettura della terza soluzione (inclusa)?

Esercizio 3. (8 punti) In un centro di assistenza, il tempo necessario per completare un backup di un computer segue una distribuzione esponenziale con un tempo medio di 2 ore.

1. Qual è la probabilità che il backup di un computer duri meno di 1 ora?

2. Qual è la varianza del tempo necessario per completare il backup di un computer?

3. Al centro di assistenza arrivano 16 computer per i quali occorre un backup. Qual è la probabilità che per almeno 3 computer (inclusi) il backup duri più di 3 ore? Si assuma che i tempi di backup dei computer siano indipendenti.

4. Al centro di assistenza arrivano 2 computer per i quali occorre un backup. Il backup del secondo computer inizia non appena il backup del primo computer è completato. Qual è la media del tempo necessario per completare il backup totale dei 2 computer? E la varianza? Si assuma che i tempi di backup dei computer siano indipendenti.

5. Nella situazione del punto 4., qual è la probabilità che il backup totale dei 2 computer sia inferiore a 7 ore?

Esercizio 4. (7 punti) Un'azienda sostiene che la durata media giornaliera degli smartphone che produce è di 12 ore. Un'indagine condotta su alcuni smartphone è volta a mostrare che la durata è in realtà inferiore. Vengono rilevate le seguenti durate (in ore):

11.2	15.2	12.1	10.8	8.6	11.7	13.6	12.7	9.5	11.0
18.2	11.9	14.2	12.5	10.0	11.3	11.4	9.3	10.3	9.5
10.5	13.6	10.8	9.6	13.3	8.9	13.8	12.7	8.9	15.5
11.7	13.2	9.6	10.4	12.2	11.9	9.1	12.2	12.6	9.6

La media calcolata sui dati risulta essere 11.63 ore. È noto che la deviazione standard della popolazione è di 4 ore.

1. È possibile sostenere con significatività 1% che la durata media degli smartphone è in realtà inferiore a 12 ore? (N.B.: derivare le formule)

2. Calcolare il p-value del test.

Quesito teorico 1. (4 punti) Sia $Z \sim \mathcal{N}(0,1)$. Che legge ha Z^2? Motivare la risposta. Siano $Z_1, \ldots, Z_n \sim \mathcal{N}(0,1)$ indipendenti. Che legge ha $\sum_{i=1}^{n} Z_i^2$? Motivare la risposta.

Quesito teorico 2. (2 punti) Calcolare media e varianza di $X \sim \mathrm{B}(n,p)$.

Esame di Probabilità e Statistica [3231]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: luglio 2024 - turno 1
Matricola: _____ Data: 15/07/2024

È obbligatorio consegnare la traccia con cognome e nome. In caso contrario, l'esito sarà "RITIRATO". Questa è la traccia n. 1. Scrivere il numero di traccia sullo svolgimento del compito.

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Si studia il prezzo di affitto di appartamenti a Bari tramite annunci pubblicati su un servizio online. Vengono rilevati i seguenti dati (in euro):

750 700 550 650 900 780 1100 450 530 1650

1. Determinare i quartili (esclusivi) dei dati.
2. Determinare eventuali dati anomali e sospetti.
3. Tracciare un box plot.
4. Calcolare il 35-esimo percentile (esclusivo).

Esercizio 2. (7 punti) In una linea di produzione, il 5% dei prodotti è difettoso. Si assuma che prodotti diversi siano indipendenti tra loro.

Nei punti 1., 2., 3. vengono esaminati 10 prodotti.

1. Qual è la probabilità che (strettamente) più di 3 prodotti siano difettosi?
2. Qual è la probabilità che esattamente 3 prodotti siano difettosi?
3. Quali sono la media e la varianza del numero di prodotti difettosi?

Nel punto 4. vengono esaminati 100 prodotti.

4. Considerando che il numero di prodotti è elevato e la probabilità di difetto è piccola, qual è un'approssimazione adeguata della probabilità che almeno 4 (4 inclusi) prodotti siano difettosi? Motivare la risposta.

Nel punto 5. vengono esaminati prodotti in sequenza fino a trovare il primo difettoso.

5. Qual è la probabilità che il primo prodotto difettoso sia il quarto esaminato?

Esercizio 3. (8 punti) Alice e Bob generano due numeri casualmente. Alice genera un numero con una variabile casuale esponenziale con media 2, mentre Bob genera un numero in base al risultato ottenuto da Alice:

- Se il numero generato da Alice è minore di 1, Bob genera un numero con una variabile aleatoria uniforme $\mathrm{U}(0,1)$.
- Se il numero generato da Alice è maggiore di 1, Bob genera un numero con una variabile aleatoria uniforme $\mathrm{U}(1,2)$.

Si risponda alle seguenti domande:

1. Sapendo che Alice ha ottenuto un numero minore di 1, qual è la media del numero generato da Bob? E la deviazione standard?
2. Qual è la probabilità che il numero generato da Bob sia maggiore di $\frac{1}{2}$?
3. Bob ha generato un numero minore di $\frac{3}{2}$. Sapendo questo fatto, qual è la probabilità che il numero generato da Alice sia minore di 1?
4. Qual è la probabilità che il minimo tra il numero generato da Alice e quello generato da Bob sia minore di 1?

Esercizio 4. (7 punti) Una squadra di calcio lo scorso anno aveva una media di gol per partita pari a 2.5. Nel campionato di quest'anno, la squadra ha segnato un numero di gol per partita come descritto dai dati raccolti nella seguente tabella:

gol per partita	frequenza assoluta
0	5
1	9
2	13
3	7
4	4

Si assuma che la deviazione standard del numero di gol sia 2.

1. Si può stabilire con significatività del 5% che la media di gol per partita è diversa da quella dell'anno scorso? (N.B.: derivare le formule)
2. Stabilire in quali dei seguenti intervalli è collocato il p-value dei dati: $[0, 1\%)$, $[1\%, 2\%)$, $[2\%, 5\%)$, $[5\%, 10\%)$, $[10\%, 20\%)$, $[20\%, 100\%]$.

Quesito teorico 1. (4 punti) Siano $X_1, \ldots, X_n \sim \mathrm{Exp}(\lambda)$ indipendenti. Che distribuzione ha $X_1 + \cdots + X_n$? Enunciare e dimostrare il risultato.

Quesito teorico 2. (2 punti) Dimostrare il legame tra il coefficiente di determinazione R^2 e il coefficiente di correlazione lineare per un campione di dati bivariato.

Esame di Probabilità e Statistica [3231]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: luglio 2024 - turno 2
Matricola: _____ Data: 15/07/2024

È obbligatorio consegnare la traccia con cognome e nome. In caso contrario, l'esito sarà "RITIRATO". Questa è la traccia n. 1. Scrivere il numero di traccia sullo svolgimento del compito.

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) In una partita di pallacanestro viene misurata la distanza dei tiri effettuati da un giocatore. Vengono misurati i seguenti dati (in metri):

2.3 8.6 3.1 5.2 1.1 5.6 6.3 1.1 7.5 6.8 .

1. Determinare i quartili (esclusivi) dei dati.
2. Determinare eventuali dati anomali e sospetti.
3. Tracciare un box plot.
4. Calcolare il 65-esimo percentile (esclusivo).

Esercizio 2. (8 punti) Una compagnia aerea studia il numero di bagagli smarriti durante i voli intercontinentali su cui opera. Si osserva che il numero di bagagli smarriti in un mese è distribuito con legge di Poisson con una media di 4 bagagli smarriti al mese. Si assuma che i bagagli smarriti in mesi diversi siano indipendenti tra loro.

1. Qual è la probabilità che in un mese vengano smarriti almeno 3 bagagli (3 inclusi)?
2. Qual è la varianza del numero di bagagli smarriti in un mese?
3. Qual è la probabilità che in un anno vengano smarriti esattamente 50 bagagli?
4. L'azienda nota che a gennaio e febbraio sono stati smarriti in tutto 6 bagagli. Sapendo che si è verificato questo evento, qual è la probabilità che da gennaio ad aprile vengano smarriti in tutto al più 10 bagagli (10 inclusi)?
5. Applicare il Teorema del Limite Centrale per stimare la probabilità che in 3 anni vengano smarriti più di 150 bagagli.

Esercizio 3. (7 punti) Un call center studia la durata delle telefonate effettuate. Si osserva che:

- Se un cliente non abbandona la chiamata, la durata della telefonata è distribuita con legge uniforme nell'intervallo $[5, 10]$ minuti.
- Se un cliente abbandona la chiamata, la durata della telefonata è distribuita con legge uniforme nell'intervallo $[2, 5]$ minuti.
- La probabilità che un cliente abbandoni la chiamata è 80%.

Si risponda ai seguenti quesiti:

1. Si consideri un cliente che abbandona la chiamata. Qual è la probabilità che la durata della telefonata sia inferiore a 3 minuti?
2. Si consideri un cliente che abbandona la chiamata. Dopo 3 minuti non è ancora terminata la chiamata. Sapendo questo fatto, qual è la probabilità che l'intera chiamata duri almeno 4 minuti? C'è un teorema che si può applicare per calcolare questa probabilità?
3. Calcolare la media e la varianza delle durate delle telefonate per un cliente che non abbandona la chiamata.
4. Si consideri un cliente qualunque. Calcolare la probabilità che la durata della telefonata sia superiore a 3 minuti.

Esercizio 4. (7 punti) Il prezzo medio degli immobili venduti in una città nel 2023 era 2100€/mq. Vengono esaminati i prezzi di alcuni immobili nel 2024, osservando i seguenti dati (in €/mq):

2111 2410 1600 3900 1988 1875 2250

Si assuma che il prezzo a metro quadro degli immobili sia distribuito con legge normale.

1. È possibile affermare con significatività del 5% che il prezzo medio degli immobili nel 2024 è aumentato rispetto al 2023?
2. Stabilire in quali dei seguenti intervalli è collocato il p-value dei dati: $[0, 0.5\%]$, $[0.5\%, 1\%)$, $[1\%, 2.5\%)$, $[2.5\%, 5\%)$, $[5\%, 10\%)$, $[10\%, 100\%]$.

Quesito teorico 1. (4 punti) Siano X e Y due variabili aleatorie continue indipendenti con densità di probabilità f_X e f_Y, rispettivamente. Che densità ha $X + Y$? Dimostrare il risultato.

Quesito teorico 2. (2 punti) Spiegare il fenomeno della scimmia instancabile di Borel.

Esame di Probabilità e Statistica [3231]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: settembre 2024 - I
Matricola: _____ Data: 03/09/2024

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Vengono raccolti i risultati (in metri) nel salto in alto di un campione olimpionico ottenuti negli ultimi anni:

2.32 2.35 2.37 2.31 2.33 2.30 2.29 2.36 2.37 2.22

1. Calcolare i quartili (esclusivi) dei dati.
2. Determinare eventuali dati anomali e sospetti.
3. Disegnare il box-plot dei dati.
4. Calcolare il 35-esimo percentile (esclusivo).

Esercizio 2. (8 punti) Un piccolo bar in una località turistica pugliese vende pasticciotti. Il bar apre alle 08:00 e sforna ogni ora 4 pasticciotti. Si vuole capire se questa produzione è sufficiente a soddisfare la richiesta dei clienti: il numero di pasticciotti ordinati in un'ora è distribuito con una legge di Poisson con media 3. Si assuma che i numeri di ordini in ore diverse siano indipendenti. (Attenzione: nelle domande seguenti tenere conto del fatto che la richiesta in una fascia oraria può essere maggiore dei pasticciotti disponibili, lasciando alcuni ordini insoddisfatti!)

1. Mostrare che la probabilità che dalle 08:00 alle 09:00 vengano ordinati esattamente 2 pasticciotti è uguale alla probabilità che vengano ordinati 3 pasticciotti.
2. Qual è la probabilità che la richiesta di pasticciotti dalle 08:00 alle 09:00 sia soddisfatta?
3. Calcolare la probabilità che il numero di pasticciotti invenduti dalle 08:00 alle 09:00 sia uguale a k per $k = 0, 1, 2, 3, 4$.
4. Se nella fascia oraria dalle 08:00 alle 09:00 restano dei pasticciotti invenduti, questi vengono offerti nella fascia oraria successiva dalle 09:00–10:00, in aggiunta a quelli sfornati alle 09:00. Qual è la probabilità che la richiesta di pasticciotti della fascia oraria 09:00–10:00 sia soddisfatta? (Suggerimento: Sfruttare i risultati del punto 3.)

Esercizio 3. (8 punti) Un chiosco in uno stabilimento balneare vende gelati. Il tempo che impiega un cliente a scegliere il gelato è distribuito con legge uniforme tra 10 secondi e 60 secondi.

1. Qual è la probabilità che un cliente scelga il gelato in meno di 30 secondi?

2. Quali sono la media e la deviazione standard del tempo di scelta del gelato?

3. Arrivano due persone in coppia al chiosco. Iniziano a scegliere insieme il gelato indipendentemente. Il tempo in cui la loro ordinazione termina è il massimo tra i due tempi di scelta. Qual è la probabilità che l'ordinazione della coppia termini in più di 30 secondi?

4. Arrivano 10 clienti che scelgono i gelati indipendentemente. Qual è la probabilità che (strettamente) più di 4 di essi impieghino più di 30 secondi per scegliere il gelato?

5. In un momento della giornata vengono serviti 40 clienti indipendenti in sequenza (un cliente inizia a scegliere il gelato solo dopo che il cliente precedente ha terminato). Utilizzare il Teorema del Limite Centrale per stimare la probabilità che in totale il chiosco sia impegnato a servire i 40 clienti per più di 25 minuti.

Esercizio 4. (7 punti) Si vuole stimare la variabilità della temperatura in una località estiva di montagna. Vengono misurate le seguenti temperature (in gradi Celsius) in momenti diversi di una giornata:

23 27 30 31 26 23 22 21

Per risolvere l'esercizio, si assuma che la temperatura abbia distribuzione normale.

1. Calcolare sui dati un intervallo di confidenza bilaterale al 90% per la varianza della temperatura. (N.B.: derivare le formule)

2. Un intervallo di confidenza al 93% sarebbe più grande o più piccolo di quello calcolato nel punto precendente?

3. Le misurazioni vengono ripetute per vari giorni consecutivi e per ogni giorno viene calcolato l'intervallo di confidenza al 90% come nel punto 1. Qual è la probabilità che la prima volta in cui la varianza appartiene all'intervallo di confidenza sia il quinto giorno?

Quesito teorico 1. (3 punti) Siano $X, Y \sim \text{Exp}(\lambda)$ indipendenti. Calcolare la densità di $X + Y$.

Quesito teorico 2. (2 punti) Enunciare e dimostrare il Teorema di Bayes e spiegarne il significato.

Esame di Probabilità e Statistica [3231]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: settembre 2024 - II
Matricola: _____ Data: 17/09/2024

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Un'azienda produce un dispositivo elettronico da utilizzare in un intervallo di temperature molto ampio. L'azienda sa che l'aumento della temperatura riduce la durata di vita del dispositivo, per questo motivo viene effettuato uno studio in cui la durata di vita viene determinata in funzione della temperatura. Si trovano i seguenti dati:

Temperatura (°C)	10	20	30	40	50	60	70	80	90
Durata di vita (ore)	420	365	285	220	176	117	69	34	5

1. Rappresentare i dati in uno scatterplot.
2. Determinare e disegnare la retta di regressione lineare. (N.B.: derivare le formule)
3. Calcolare il coefficiente di correlazione lineare e il coefficiente di determinazione.

Esercizio 2. (7 punti) Un famoso duo di musicisti inglesi annuncia un tour in UK dopo tanti anni di assenza dalla scena musicale. Per via della grande richiesta, acquistare un biglietto per il concerto è difficile: collegandosi al sito ufficiale, si riesce a caricare la pagina per l'acquisto con probabilità 10%.

1. Una persona si collega al sito. Ogni volta che non riesce a collegarsi al sito, ricarica la pagina, finché finalmente non riesce a collegarsi. Qual è la probabilità che la persona riesca a collegarsi al sito entro il terzo tentativo (terzo incluso)?
2. Nella situazione del punto 1., qual è la media del tentativo in cui la persona riesce a collegarsi al sito per la prima volta? E la varianza?
3. Due persone si mettono d'accordo: si collegano al sito insieme e indipendentemente e continuano a provare (ricaricando la pagina in caso di fallimento) finché almeno una di loro non riesce a collegarsi. Derivare una formula per la probabilità che nessuna delle due persone riesca a collegarsi al sito entro il k-esimo (k incluso) tentativo.
4. Che legge segue la variabile aleatoria che rappresenta il primo tentativo in cui almeno una delle due persone riesce a collegarsi al sito? (Utilizzare il risultato del punto 3.)

Esercizio 3. (8 punti) Un macchinario è costituito da 10 componenti e funziona correttamente se almeno 7 di essi (7 incluso) sono funzionanti. Ogni componente funziona, da quando viene installato, per un tempo distribuito con legge esponenziale con media 2 anni, dopodiché si guasta (e non viene sostituito). Si supponga che i componenti siano indipendenti tra loro.

1. Si consideri un singolo componente. Dal momento in cui viene installato, qual è la probabilità che funzioni correttamente per almeno 1 anno?
2. Si consideri un singolo componente. Dopo un anno viene controllato e risulta ancora funzionante. Sapendo questo fatto, qual è la probabilità che, dal momento in cui è stato installato, funzioni correttamente per almeno 2 anni?
3. Qual è la probabilità che l'intero macchinario funzioni correttamente per almeno 1 anno?
4. Calcolare media e varianza del numero di componenti del macchinario funzionanti dopo 1 anno.
5. Si supponga ora un nuovo scenario. Non appena un componente si guasta, viene sostituito con uno nuovo funzionante (con durata di vita distribuita come sopra). Un componente che si guasta per la seconda volta non viene più sostituito e resta guasto. Qual è la probabilità che un componente funzioni correttamente per almeno 1 anno?

Esercizio 4. (7 punti) Si studia la variabilità dei prezzi di volo relativi a una certa tratta di una compagnia aerea. Lo scorso anno, la varianza era di $1000\,€^2$. Quest'anno vengono registrati i seguenti prezzi di volo (in €):

$$78 \quad 129 \quad 259 \quad 78 \quad 109 \quad 133 \quad 133 \quad 101 \quad 48 \quad 56\,.$$

Si assuma la distribuzione dei prezzi normale.

1. Si può stabilire con significatività 5% che la varianza di quest'anno sia maggiore di quella dello scorso anno? (N.B.: derivare le formule)
2. Indicare in quale di questi intervalli si colloca il $p-$value del test: $[0\%, 0.5\%]$, $[0.5\%, 1\%]$, $[1\%, 2.5\%]$, $[2.5\%, 5\%]$, $[5\%, 100\%]$.

Quesito teorico 1. (4 punti) Enunciare e dimostrare la legge dei grandi numeri.

Quesito teorico 2. (2 punti) Sia (X, Y) un vettore aleatorio assolutamente continuo con densità $f_{(X,Y)}$. Calcolare la densità di X in termini di $f_{(X,Y)}$.

Esame di Probabilità e Statistica [3231]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: novembre 2024
Matricola: _____ Data: 05/11/2024

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Vengono registrati i consumi bimestrali di energia elettrica di alcune utenze domestiche. I dati vengono raccolti in classi di consumo, come riportato nella tabella seguente:

intervallo di consumo (kWh)	frequenza assoluta
$[100, 120)$	10
$[120, 130)$	11
$[130, 140)$	13
$[140, 170)$	29
$[170, 200)$	6

1. Rappresentare un istogramma delle densità di frequenze relative.
2. Determinare la classe modale.
3. Calcolare un'approssimazione della media e della varianza dei consumi.
4. Calcolare un'approssimazione del 70-esimo percentile dei consumi.

Esercizio 2. (7 punti) È periodo di elezioni e si vuole prevedere l'afflusso alle urne in un seggio elettorale. Si assuma che il numero di votanti che si presentano ad un certo seggio nell'arco di 10 minuti sia una variabile aleatoria con distribuzione di Poisson con parametro $\lambda = 3$. Si supponga che il numero di votanti che si presentano in intervalli di tempo diversi siano indipendenti tra loro.

1. Calcolare la probabilità che nell'arco di 10 minuti si presentino almeno 4 votanti (4 incluso).
2. Calcolare la media e la varianza del numero di votanti che si presentano al seggio nell'arco di 10 minuti.
3. Con che legge è distribuito il numero di votanti che si presentano al seggio nell'arco di un'ora? Quali sono la media e la varianza del numero di votanti che si presentano al seggio nell'arco di un'ora?

4. In una giornata il seggio è aperto per 12 ore. Utilizzare il Teorema del Limite Centrale per approssimare la probabilità che nella giornata si presentino al seggio almeno 200 votanti. (Suggerimento: dividere le 12 ore in intervalli di 10 minuti)

Esercizio 3. (8 punti) Due genitori studiano il tempo necessario affinché la loro bambina di 6 mesi si addormenti. Da quando iniziano a cullarla per farla addormentare, la bambina si addormenta in media in 5 minuti. Si supponga che il tempo necessario affinché la bambina si addormenti sia distribuito in modo esponenziale.

1. Calcolare la probabilità che la bambina si addormenti entro i primi 2 minuti da quando iniziano a cullarla.

2. Calcolare la deviazione standard del tempo necessario affinché la bambina si addormenti da quando iniziano a cullarla.

3. Uno dei genitori sta cullando la bambina da 10 minuti, ma lei ancora gli occhi spalancati. Sapendo questo fatto, calcolare la probabilità che, da quando ha iniziato a cullarla, la bambina si addormenti entro 15 minuti.

Nelle prossime richieste si consideri il seguente scenario. Alla bambina sta per spuntare il primo dentino e questo può renderla irritabile e influenzare il tempo necessario affinché si addormenti. Se la bambina è irritata, il tempo necessario affinché si addormenti è distribuito in modo esponenziale con media 10 minuti. Altrimenti, come prima, la media è 5 minuti. Ogni volta che si prova a farla addormentare, la probabilità che la bambina sia irritata è 70%.

4. Si prova a far addormentare la bambina una volta. Calcolare la probabilità che si addormenti entro i primi 12 minuti.

5. Uno dei genitori sta cullando la bambina da 12 minuti, ma lei ancora gli occhi spalancati. Sapendo questo fatto, calcolare la probabilità che la bambina sia irritata per via del dentino che sta spuntando.

6. Nell'arco della giornata si prova a far addormentare la bambina 5 volte. Calcolare la probabilità che la bambina si addormenti almeno 3 volte entro i primi 12 minuti.

Esercizio 4. (7 punti) Vengono registrati i punteggi ottenuti da alcune persone nell'Esercizio 4 di Probabilità e Statistica. I dati sono i seguenti:

$$4 \quad 3 \quad 1 \quad 4 \quad 5 \quad 7 \quad 7 \quad 0 \quad 5 \quad 5 \quad 5 \quad 0$$

Si supponga che i punteggi siano distribuiti in modo normale.

1. Calcolare sui dati un intervallo di confidenza bilaterale al 95% per la varianza dei punteggi. (N.B.: Derivare le formule)

2. Calcolare sui dati un intervallo di confidenza bilaterale al 95% per la media dei punteggi. (Non è obbligatorio derivare le formule)

Quesito teorico 1. (2 punti) Spiegare il problema di Monty Hall e la sua soluzione.

Quesito teorico 2. (4 punti) Derivare la formula esplicita per il calcolo della somma della serie geometrica e applicarla per calcolare $\mathbb{P}(\{X > k\})$ dove $X \sim \text{Geo}(p)$.

Esame di Probabilità e Statistica [3231]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: gennaio 2025
Matricola: _____ Data: 16/01/2025

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Una società vuole studiare gli effetti dell'investimento pubblicitario (in migliaia di euro) sull'incremento percentuale delle vendite annuali. Sono stati raccolti i seguenti dati:

Investimento pubblicitario	10	20	30	40	50	60
Incremento percentuale delle vendite	2	4	5	7	9	10

1. Rappresentare i dati in uno scatterplot.
2. Determinare e disegnare la retta di regressione lineare. (N.B.: derivare le formule).
3. Calcolare il coefficiente di correlazione lineare e il coefficiente di determinazione.

Esercizio 2. (7 punti) Sia (X, Y) un vettore aleatorio con funzione di probabilità congiunta data dalla seguente tabella:

Y \ X	0	1
0	a	b
1	c	d

1. Fornire valori di a, b, c, d tali che X e Y siano indipendenti (verificando l'indipendenza).
2. Fornire valori di a, b, c, d tali che $\mathrm{Cov}(X, Y) = \frac{1}{4}$.
3. Fornire valori di a, b, c, d tali che X e Y non siano indipendenti.
4. Fornire valori di a, b, c, d tali che $\mathbb{P}(\{X = 1\}|\{Y = 1\}) = \frac{1}{3}$.
5. X viene generata in sequenza tante volte fino a che non si verifica $\{X = 1\}$. Fornire valori di a, b, c, d tali che, in media, la prima volta che si ottiene 1 sia la quarta.

Esercizio 3. (8 punti) Per l'implementazione di un progetto, occorre completare alcuni *task*. Ogni *task* viene completato in un tempo distribuito con legge esponenziale con valore atteso di 2 giorni. Si assumano i tempi di completamento dei *task* indipendenti tra loro.

1. Qual è la probabilità che un *task* venga completato in meno di 3 giorni?

2. Calcolare la varianza del tempo impiegato per completare un *task*.

3. Due *task* $T1.1$ e $T1.2$ partono contemporaneamente e devono essere completati entrambi per poter avviare la seconda fase del progetto. Qual è la probabilità che dopo 4 giorni non si possa ancora avviare la seconda fase?

4. Dire se è vero che il massimo tra i tempi impiegati per completare i due *task* $T1.1$ e $T1.2$ gode di assenza di memoria, motivando la risposta.

5. La seconda fase è composta da due *task* $T2.1$ e $T2.2$ che devono essere completati in serie, ovvero $T2.2$ può essere avviato solo dopo il completamento di $T2.1$. Qual è la probabilità che il tempo impiegato per completare la seconda fase sia maggiore di 4 giorni?

Esercizio 4. (7 punti) Una ditta sostiene che il peso medio delle sue confezioni di zucchero è di 1 kg. Un campione di 10 confezioni fornisce i seguenti pesi (in kg):

$$1.01 \quad 0.92 \quad 1.02 \quad 1.00 \quad 0.97 \quad 1.03 \quad 1.01 \quad 0.96 \quad 0.90 \quad 1.02.$$

Si supponga che i pesi siano distribuiti normalmente.

1. Si può accettare l'affermazione della ditta con significatività $\alpha = 0.05$? (N.B.: derivare le formule).

2. Stabilire in quali dei seguenti intervalli è contenuto il p-value: $[0.01, 0.02)$, $[0.02, 0.05)$, $[0.05, 0.10)$, $[0.1, 0.20)$, nessuno dei precedenti.

Quesito teorico 1. (2 punti) Dimostrare che la somma di due variabili aleatorie indipendenti con distribuzione di Poisson è ancora una variabile aleatoria di Poisson.

Quesito teorico 2. (4 punti) Spiegare l'approssimazione di una variabile aleatoria binomiale con il Teorema del Limite Centrale, fornendo anche un esempio.

2 Soluzioni

Di seguito le soluzioni relative alle tracce di sopra.

Esame di Probabilità e Statistica [3231]

Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: giugno 2024 - turno 1
Matricola: _____ Data: 16/04/2024

Viene usata come riferimento la traccia n. 120.

Esercizio 1. Una studentessa di Probabilità e Statistica vuole determinare se esiste una relazione tra le ore di studio giornaliere nella preparazione dell'esame e il voto all'esame. Per farlo, ha intervistato alcuni studenti e ha ottenuto i seguenti dati:

ore di studio	0	1	2	3	4	5	6
voto	0	12	11	25	26	30	30

1. Rappresentare i dati in uno scatterplot.
2. Calcolare la retta di regressione lineare (derivando le formule) e disegnarla.
3. Calcolare il coefficiente di correlazione lineare e il coefficiente di determinazione R^2.

Soluzione. 1. Segue lo scatterplot (con la retta di regressione lineare):

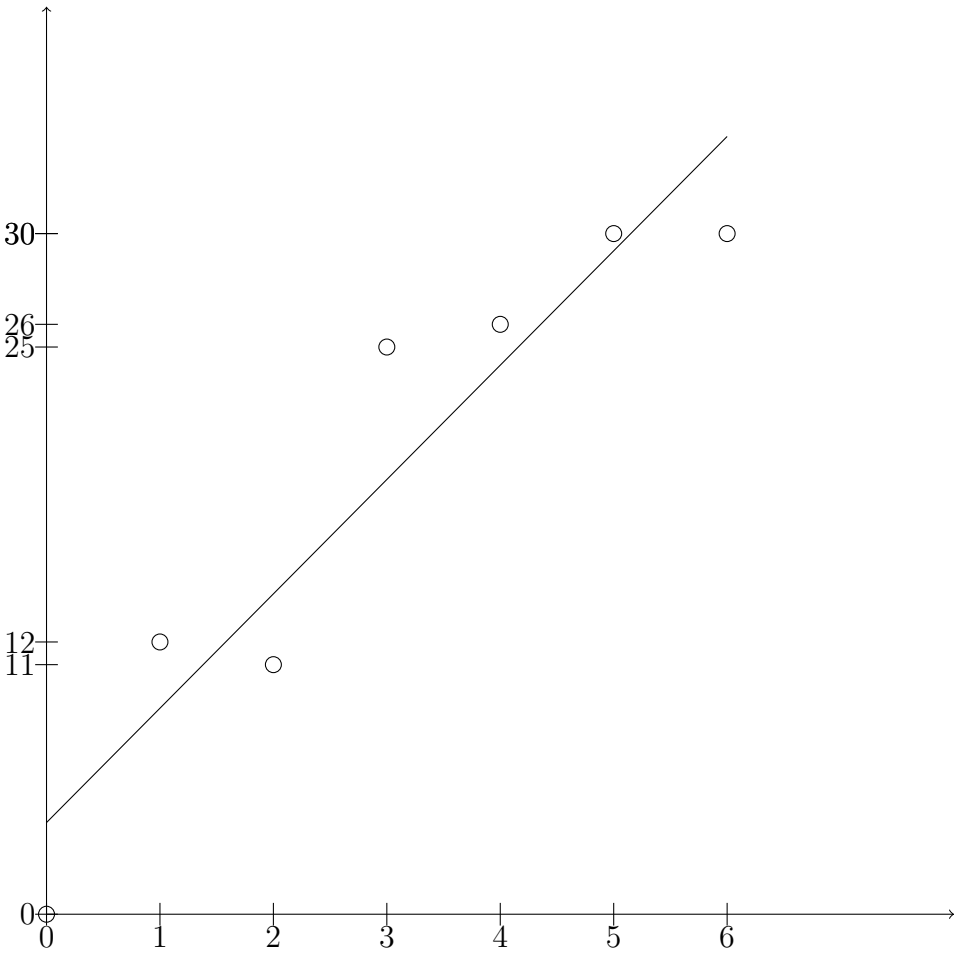


Figura 1: Scatterplot e retta di regressione lineare.

2. Denotiamo con $(x_1, y_1), \ldots, (x_n, y_n)$, $n = 6$, i dati del campione. Cerchiamo la retta di equazione

$$y = ax + b$$

che meglio approssima i dati, utilizzando il metodo dei minimi quadrati. Vogliamo minimizzare l'errore

$$e(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 .$$

Imponiamo che il gradiente rispetto ad (a, b) sia nullo, ovvero,

$$0 = \partial_a e(a, b) = -2 \sum_{i=1}^{n} (y_i - ax_i - b)x_i = -2 \sum_{i=1}^{n} (x_i y_i - ax_i^2 - bx_i) ,$$
$$0 = \partial_b e(a, b) = -2 \sum_{i=1}^{n} (y_i - ax_i - b)$$

Dalla seconda equazione segue che

$$nb = \sum_{i=1}^{n} (y_i - ax_i) \implies b = \overline{y} - a\overline{x} .$$

Sostituendo nella prima,

$$\sum_{i=1}^{n}(x_iy_i - ax_i^2 - bx_i) = 0 \implies \sum_{i=1}^{n}(x_iy_i - ax_i^2 - x_i\overline{y} + a\overline{x}x_i) = 0$$
$$\implies a\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big) = \sum_{i=1}^{n} x_iy_i - n\overline{x}\,\overline{y}$$
$$\implies a = \frac{\sum_{i=1}^{n} x_iy_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\,.$$

Completiamo la tabella con i valori necessari a calcolare a e b:

								somma
x_i	0	1	2	3	4	5	6	21
y_i	0	12	11	25	26	30	30	134
x_i^2	0	1	4	9	16	25	36	91
y_i^2	0	144	121	625	676	900	900	3366
x_iy_i	0	12	22	75	104	150	180	543

Pertanto $\overline{x} = 21/7 = 3$ e $\overline{y} = 134/7 = 19.14$. Segue che

$$a = \frac{543 - 7\cdot 3\cdot 19.14}{91 - 7\cdot 3^2} = \frac{141}{28} \simeq 5.04\,,$$
$$b = 19.14 - 5.04\cdot 3 = 4.02\,,$$

ovvero, la retta di regressione lineare ha equazione

$$y = 5.04x + 4.02\,.$$

3. Per calcolare il coefficiente di correlazione lineare usiamo la formula

$$\rho_{x,y} = \frac{\sum_{i=1}^{n} x_iy_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}} = \frac{543 - 7\cdot 3\cdot 19.14}{\sqrt{91 - 7\cdot 3^2}\sqrt{3366 - 7\cdot 19.14^2}}$$
$$= 0.9411\,.$$

Il coefficiente di determinazione è $R^2 = \rho_{x,y}^2 = 0.8857$.

Esercizio 2. (8 punti) Sia (X, Y) un vettore aleatorio discreto con legge congiunta descritta dalla seguente tabella:

X \ Y	0	1	2
0	0	$\frac{1}{4}$	a
1	$\frac{1}{4}$	0	b
2	c	d	0

Si assuma che:

$$\bullet\ \mathbb{P}(\{X = 2\}) = \frac{1}{8}\,, \quad \bullet\ \mathbb{P}(\{Y = 1\}|\{X = 2\}) = 1\,, \quad \bullet\ \mathbb{E}(XY) = \frac{1}{2}\,.$$

Dopo aver determinato i valori di a, b, c, d, rispondere ai seguenti quesiti:

1. Calcolare la covarianza tra X e Y e stabilire se X e Y sono indipendenti.

2. Calcolare $\mathrm{Var}(X+Y)$.

3. Si supponga di estrarre 14 realizzazioni indipendenti della variabile aleatoria Y. Calcolare la probabilità che l'evento $\{Y=0\}$ si realizzi almeno 3 volte (3 incluso).

4. Si estraggono tante realizzazioni indipendenti della variabile aleatoria Y. In media, qual è la prima volta in cui si realizza l'evento $\{Y=0\}$?

Soluzione. Per determinare i valori di a, b, c, d usiamo le informazioni fornite. Poiché le probabilità devono sommare a 1, otteniamo che

$$\frac{1}{4}+a+\frac{1}{4}+b+c+d=1 \implies a+b+c+d=\frac{1}{2}\,.$$

Dalla condizione $\mathbb{P}(\{X=2\})=1/8$ otteniamo che

$$c+d=\frac{1}{8}\,.$$

Dalla condizione $\mathbb{P}(\{Y=1\}|\{X=2\})=1$ otteniamo che

$$\frac{\mathbb{P}(\{Y=1\}\cap\{X=2\})}{\mathbb{P}(\{X=2\})}=1 \implies \frac{d}{1/8}=1 \implies d=\frac{1}{8}\,.$$

Dalla condizione $\mathbb{E}(XY)=1/2$ otteniamo che

$$\begin{aligned}\frac{1}{2}=\mathbb{E}(XY)&=\sum_{x,y}xy\mathbb{P}(\{X=x\}\cap\{Y=y\})\\&=0\cdot0\cdot0+1\cdot0\cdot\frac{1}{4}+2\cdot0\cdot c+0\cdot1\cdot\frac{1}{4}+1\cdot1\cdot0+2\cdot1\cdot d+0\cdot2\cdot a+1\cdot2\cdot b+2\cdot2\cdot0\\&=2d+2b \implies d+b=\frac{1}{4}\,.\end{aligned}$$

Risolviamo il sistema

$$\begin{cases}a+b+c+d=\frac{1}{2}\\c+d=\frac{1}{8}\\d=\frac{1}{8}\\d+b=\frac{1}{4}\end{cases}$$

ottenendo $a=1/4$, $b=1/8$, $c=0$, $d=1/8$.

La tabella completa è la seguente:

X \ Y	0	1	2
0	0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	0	$\frac{1}{8}$
2	0	$\frac{1}{8}$	0

1. Per calcolare la covarianza utilizziamo la formula

$$\mathrm{Cov}(X,Y)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)\,.$$

Calcoliamo

$$\mathbb{E}(X)=\sum_x x\mathbb{P}(\{X=x\}=0\cdot\left(0+\frac{1}{4}+\frac{1}{4}\right)+1\cdot\left(\frac{1}{4}+0+\frac{1}{8}\right)+2\cdot\left(0+\frac{1}{8}+0\right)=\frac{5}{8}\,,$$

e

$$\mathbb{E}(Y) = \sum_y y\mathbb{P}(\{Y = y\} = 0 \cdot \Big(0 + \frac{1}{4} + 0\Big) + 1 \cdot \Big(\frac{1}{4} + 0 + \frac{1}{8}\Big) + 2 \cdot \Big(\frac{1}{4} + \frac{1}{8} + 0\Big) = \frac{9}{8}\,.$$

Segue che

$$\mathrm{Cov}(X, Y) = \frac{1}{2} - \frac{5}{8} \cdot \frac{9}{8} = -\frac{13}{64}\,.$$

Poiché $\mathrm{Cov}(X, Y) \neq 0$, X e Y non sono indipendenti.

2. Calcoliamo $\mathrm{Var}(X + Y)$ utilizzando la formula

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)\,.$$

Utilizzando le formule $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ e $\mathrm{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$, occorre calcolare solo

$$\mathbb{E}(X^2) = \sum_x x^2\mathbb{P}(\{X = x\} = 0^2 \cdot \Big(0 + \frac{1}{4} + \frac{1}{4}\Big) + 1^2 \cdot \Big(\frac{1}{4} + 0 + \frac{1}{8}\Big) + 2^2 \cdot \Big(0 + \frac{1}{8} + 0\Big) = \frac{7}{8}\,,$$

e

$$E(Y^2) = \sum_y y^2\mathbb{P}(\{Y = y\} = 0^2 \cdot \Big(0 + \frac{1}{4} + 0\Big) + 1^2 \cdot \Big(\frac{1}{4} + 0 + \frac{1}{8}\Big) + 2^2 \cdot \Big(\frac{1}{4} + \frac{1}{8} + 0\Big) = \frac{15}{8}\,.$$

Quindi

$$\mathrm{Var}(X) = \frac{7}{8} - \Big(\frac{5}{8}\Big)^2 = \frac{31}{64}\,,$$
$$\mathrm{Var}(Y) = \frac{15}{8} - \Big(\frac{9}{8}\Big)^2 = \frac{39}{64}\,.$$

Concludiamo che

$$\mathrm{Var}(X + Y) = \frac{31}{64} + \frac{39}{64} + 2 \cdot \Big(-\frac{13}{64}\Big) = \frac{44}{64}\,.$$

3. Per calcolare la probabilità che l'evento $\{Y = 0\}$ si realizzi almeno 4 volte (4 incluso) in 14 estrazioni, possiamo utilizzare la distribuzione binomiale. La probabilità di successo è

$$\mathbb{P}(\{Y = 0\}) = \frac{1}{4}\,.$$

Consideriamo una variabile aleatoria

$$Z = \text{“numero di successi in 14 prove”} \sim \mathrm{B}(14, 1/4)\,.$$

Dobbiamo calcolare

$$\begin{aligned}\mathbb{P}(\{Z \geq 4\}) &= 1 - \mathbb{P}(\{Z < 4\}) = 1 - \mathbb{P}(\{Z = 0\}) - \mathbb{P}(\{Z = 1\}) - \mathbb{P}(\{Z = 2\}) - \mathbb{P}(\{Z = 3\}) \\ &= 1 - \Big(\frac{3}{4}\Big)^{14} - 14 \cdot \frac{1}{4} \cdot \Big(\frac{3}{4}\Big)^{13} - \binom{14}{2} \cdot \Big(\frac{1}{4}\Big)^2 \cdot \Big(\frac{3}{4}\Big)^{12} = 71.89\%\end{aligned}$$

4. Consideriamo la variabile aleatoria

$W =$ “prima volta nella successione di estrazioni in cui si verifica l'evento $\{Y = 0\}$” $\sim \mathrm{Geo}(1/4)$,

poiché

$$\mathbb{P}(\{Y = 0\}) = \frac{1}{4}\,.$$

Vogliamo calcolare

$$\mathbb{E}(W) = \frac{1}{1/4} = 4\,.$$

Esercizio 3. (7 punti) Una catena di fast food analizza il tempo di servizio dei suoi clienti. I clienti possono ordinare due tipi di menu: il menu A e il menu B. Si assuma che

- Il tempo di servizio per il menu A sia distribuito con legge esponenziale con media 2 minuti.
- Il tempo di servizio per il menu B sia distribuito con legge esponenziale con deviazione standard 3 minuti.
- Il 20% dei clienti ordina il menu A e il 80% il menu B.

Si risponda alle seguenti domande:

1. Si consideri un cliente che ha ordinato il menu A. Qual è la probabilità che il suo tempo di servizio sia inferiore a 3 minuti?
2. Si consideri un cliente che ha ordinato il menu B. Qual è la probabilità che il servizio avvenga esattamente al minuto 1?
3. Si consideri un cliente che ha ordinato il menu B. Ha aspettato 2 minuti e non è ancora stato servito. Sapendo questo fatto, qual è la probabilità che debba aspettare in tutto almeno 5 minuti?
4. Un cliente ha fatto un ordine e ha aspettato un tempo compreso tra 1 e 4 minuti. Qual è la probabilità che abbia ordinato il menu A?

Soluzione. Consideriamo le variabili aleatorie

$$X_A = \text{“tempo di servizio per il menu A”} \sim \text{Exp}(\lambda)$$

$$X_B = \text{“tempo di servizio per il menu B”} \sim \text{Exp}(\mu)$$

$$X = \text{“tempo di servizio (senza specificare il menu)”}\,.$$

Abbiamo che

$$2 = \mathbb{E}(X) = \frac{1}{\lambda} \implies \lambda = \frac{1}{2}\,,$$

$$3 = \sqrt{\text{Var}(X)} = \frac{1}{\mu} \implies \mu = \frac{1}{3}\,.$$

Quindi $X_A \sim \text{Exp}(1/2)$ e $X_B \sim \text{Exp}(1/3)$.

Consideriamo infine la variabile aleatoria

$$Y = \text{“1 se ordinato il menu A, 0 altrimenti”} \sim \text{Be}(0.2)\,.$$

1. Calcoliamo

$$\mathbb{P}(\{X_A < 3\}) = 1 - \mathbb{P}(\{X_A \geq 3\}) = 1 - e^{-3/2} = 77.69\%\,.$$

2. Poiché X_B è una variabile continua con densità, la probabilità che il servizio avvenga esattamente al minuto 1 è nulla:

$$\mathbb{P}(\{X_B = 1\}) = 0\,.$$

3. La probabilità che il cliente debba aspettare in tutto almeno 5 minuti, sapendo che ha già aspettato 2 minuti, si può calcolare utilizzando la proprietà di assenza di memoria della distribuzione esponenziale:

$$\mathbb{P}(\{X_B \geq 5\}|\{X_B \geq 2\}) = \mathbb{P}(\{X_B \geq 3\}) = e^{-3/3} = 36.79\% \,.$$

4. La probabilità che un cliente abbia ordinato il menu A, sapendo che ha aspettato un tempo compreso tra 1 e 4 minuti, si può calcolare utilizzando la formula di Bayes:

$$\begin{aligned}\mathbb{P}(\{Y = 1\}|\{1 \leq X \leq 4\}) &= \frac{\mathbb{P}(\{1 \leq X \leq 4\} \cap \{Y = 1\})}{\mathbb{P}(\{1 \leq X \leq 4\})} \\ &= \frac{\mathbb{P}(\{1 \leq X \leq 4\}|\{Y = 1\})\mathbb{P}(\{Y = 1\})}{\mathbb{P}(\{1 \leq X \leq 4\})} \\ &= \frac{\mathbb{P}(\{1 \leq X_A \leq 4\})\mathbb{P}(\{Y = 1\})}{\mathbb{P}(\{1 \leq X \leq 4\})} \,.\end{aligned}$$

Calcoliamo i tre termini:

$$\mathbb{P}(\{1 \leq X_A \leq 4\}) = \int_1^4 \frac{1}{2} e^{-x/2} \, \mathrm{d}x = e^{-1/2} - e^{-2} = 47.12\% \,.$$

$$\mathbb{P}(\{Y = 1\}) = 20\% \,.$$

Utilizzando il teorema della probabilità totale:

$$\begin{aligned}\mathbb{P}(\{1 \leq X \leq 4\}) &= \mathbb{P}(\{1 \leq X \leq 4\} \cap \{Y = 1\}) + \mathbb{P}(\{1 \leq X \leq 4\} \cap \{Y = 0\}) \\ &= \mathbb{P}(\{1 \leq X_A \leq 4\})\mathbb{P}(\{Y = 1\}) + \mathbb{P}(\{1 \leq X_B \leq 4\})\mathbb{P}(\{Y = 0\}) \\ &= (e^{-1/2} - e^{-2})20\% + (e^{-1/3} - e^{-4/3})80\% = 45.66\% \,.\end{aligned}$$

Concludiamo che

$$\mathbb{P}(\{Y = 1\}|\{1 \leq X \leq 4\}) = \frac{47.12\% \cdot 20\%}{45.66\%} = 20.64\% \,.$$

Esercizio 4.(7 punti) Una fabbrica di cereali vuole stimare la media del peso delle confezioni prodotte. Un controllo su un campione di alcune confezioni ha fornito i seguenti pesi in grammi:

500 506 499 507 502 500 496 501

Si supponga che il peso sia distribuito normalmente.

1. Calcolare sui dati un intervallo di confidenza bilaterale al 90% per la media del peso delle confezioni. (N.B.: derivare le formule)

2. Un intervallo di confidenza bilaterale al 91% calcolato sugli stessi dati è più grande o più piccolo di quello calcolato al punto 1? Perché?

Soluzione. 1. La popolazione è descritta da una variabile aleatoria $X \sim \mathcal{N}(\mu, \sigma^2)$. I parametri μ e σ^2 non sono noti. Dalla popolazione viene estratto un campione $X_1, \ldots, X_n$ di $n = 8$ osservazioni. Dalla definizione di IC si ha che

$$\beta = \mathbb{P}(\{U_n \leq \mu \leq V_n\}) \,.$$

Per stimare μ sfrutteremo lo stimatore media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Per stimare σ^2 sfrutteremo lo stimatore varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Poiché $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ sono indipendenti, $T_{n-1} = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$. Allora

$$\begin{aligned}\beta = \mathbb{P}(\{U_n \le \mu \le V_n\}) &= \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - V_n}{S_n/\sqrt{n}} \le \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \le \frac{\overline{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) \\ &= \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - V_n}{S_n/\sqrt{n}} \le T_{n-1} \le \frac{\overline{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) \\ &= 1 - \mathbb{P}\Big(\Big\{T_{n-1} < \frac{\overline{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big) - \mathbb{P}\Big(\Big\{T_{n-1} > \frac{\overline{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big)\,.\end{aligned}$$

Segue che

$$\mathbb{P}\Big(\Big\{T_{n-1} < \frac{\overline{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big) + \mathbb{P}\Big(\Big\{T_{n-1} > \frac{\overline{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) = 1 - \beta = \alpha\,.$$

Decidiamo di equipartire α:

$$\mathbb{P}\Big(\Big\{T_{n-1} < \frac{\overline{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big) = \mathbb{P}\Big(\Big\{T_{n-1} < \frac{\overline{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big) = \frac{\alpha}{2}\,.$$

Definiamo $t_{n-1,\alpha/2}$ come il punto tale che

$$\mathbb{P}(\{T_{n-1} \ge t_{n-1,\alpha/2}) = \frac{\alpha}{2}\,.$$

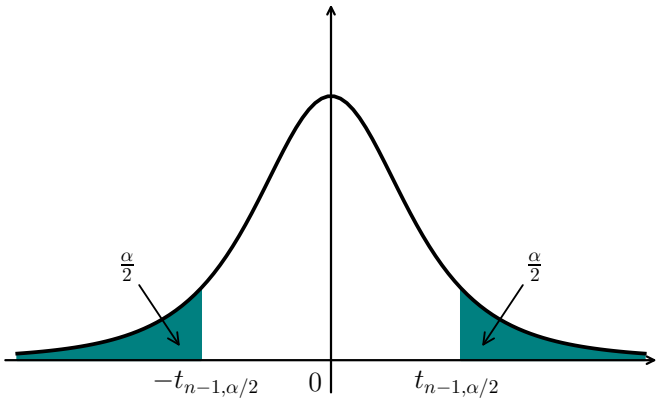


Figura 2: Definizione di $t_{n-1,\alpha/2}$.

Scegliendo

$$\begin{aligned}\frac{\overline{X}_n - U_n}{S_n/\sqrt{n}} = t_{n-1,\alpha/2} &\implies U_n = \overline{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,\alpha/2}\,, \\ \frac{\overline{X}_n - V_n}{S_n/\sqrt{n}} = -t_{n-1,\alpha/2} &\implies V_n = \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,\alpha/2}\,,\end{aligned}$$

si ottiene la condizione che definisce l'intervallo di confidenza. In conclusione

$$\Big[\overline{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,\alpha/2}, \overline{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,\alpha/2}\Big]$$

è un intervallo di confidenza bilaterale per μ con livello di confidenza $\beta = 1 - \alpha$.

Calcoliamo l'intervallo di confidenza sui dati. Per farlo, calcoliamo

$$\overline{x}_n = \frac{1}{8}(500 + 506 + 499 + 507 + 502 + 500 + 496 + 501) = 501.375\,,$$

$$s_n^2 = \frac{1}{7}(500^2 + 506^2 + 499^2 + 507^2 + 502^2 + 500^2 + 496^2 + 501^2 - 8 \cdot 501.375^2) = 13.125$$
$$\implies s_n = \sqrt{13.125} = 3.623\,.$$

Infine, per $\beta = 90\%$ abbiamo che $\alpha/2 = 0.05$. Dalla tabella della t di Student, $t_{7,0.05} = 1.895$. Pertanto l'intervallo di confidenza è

$$\left[501.375 - \frac{3.623}{\sqrt{8}} \cdot 1.895, 501.375 + \frac{3.623}{\sqrt{8}} \cdot 1.895\right] = [498.95, 503.80]\,.$$

Un intervallo di confidenza al 91% sarebbe più grande di quello al 90% poiché $\alpha/2$ sarebbe più piccolo e di conseguenza il quantile $t_{7,\alpha/2}$ sarebbe più grande.

Esame di Probabilità e Statistica [3231]

Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: giugno 2024 - turno 2
Matricola: _____ Data: 17/06/2024

Viene usata come riferimento la traccia n. 120.

Esercizio 1. (6 punti) Si studia il tempo di attesa per il servizio clienti di una banca. I dati vengono raggruppati in classi nella seguente tabella:

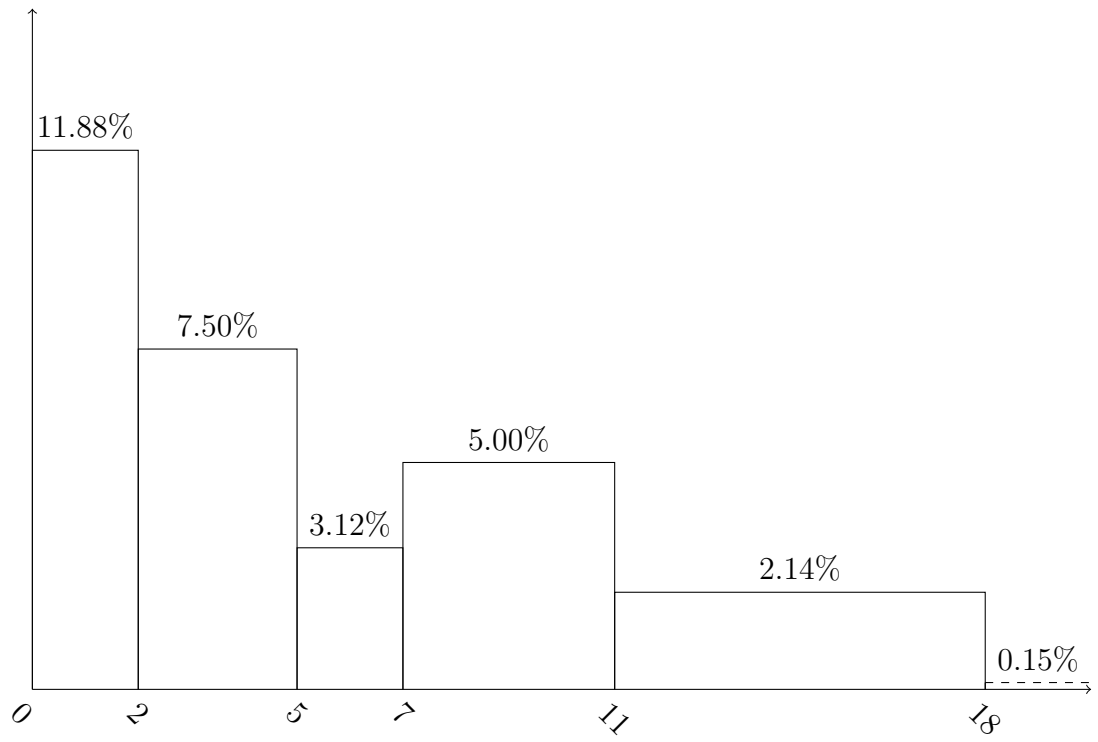
intervalli (minuti)	frequenze assolute
$[0, 2)$	19
$[2, 5)$	18
$[5, 7)$	5
$[7, 11)$	16
$[11, 18)$	12
$[18, 100)$	10

1. Rappresentare un istogramma delle densità di frequenze relative.
2. Determinare la classe modale (o le classi modali, se più di una).
3. Calcolare un'approssimazione della media e della varianza dei dati.
4. Calcolare un'approssimazione del 55-esimo percentile.

Soluzione. 1. Completiamo la tabella:

intervallo	freq. assolute	freq. relative	densità di freq. rel.	freq. cumulate
$[0, 2)$	19	23.75%	11.88%	19
$[2, 5)$	18	22.50%	7.50%	37
$[5, 7)$	5	6.25%	3.12%	42
$[7, 11)$	16	20.00%	5.00%	58
$[11, 18)$	12	15.00%	2.14%	70
$[18, 100)$	10	12.50%	0.15%	80

Rappresentiamo le densità di frequenze relative in un istogramma.



2. La classe modale è quella con maggiore densità di frequenza relativa, quindi è l'intervallo $[0, 2)$.

3. Per calcolare un'approssimazione della media utilizziamo le frequenze relative ottenute da $p_j = f_j/n$ dove $n = 80$ e i valori centrali $\tilde{v}_j$ degli intervalli

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \simeq \frac{1}{n}\sum_{j=1}^{k} f_j \tilde{v}_j = \sum_{j=1}^{k} p_j \tilde{v}_j$$
$$= 23.75\% \cdot 1 + 22.50\% \cdot 3.5 + 6.25\% \cdot 6 + 20.00\% \cdot 9 + 15.00\% \cdot 14.5 + 12.50\% \cdot 59 = 12.75\,.$$

Calcoliamo un'approssimazione della varianza

$$s^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big) \simeq \frac{1}{n-1}\Big(\sum_{j=1}^{n} f_j \tilde{v}_j^2 - n\overline{x}^2\Big) = \frac{n}{n-1}\Big(\sum_{j=1}^{n} p_j \tilde{v}_j^2 - \overline{x}^2\Big)$$
$$= \frac{80}{79}\Big(23.75\% \cdot 1^2 + 22.50\% \cdot 3.5^2 + 6.25\% \cdot 6^2 + 20.00\% \cdot 9^2 + 15.00\% \cdot 14.5^2 + 12.50\% \cdot 59^2$$
$$- 12.75^2\Big) = 329.66\,.$$

4. Per calcolare un'approssimazione del 55-esimo percentile dei dati, usiamo le frequenze cumulate. Troviamo l'intervallo I_j tale che $F_j \leq 55\%n = 44 < F_{j+1}$. Si tratta dell'intervallo $[7, 11)$. Approssimiamo la mediana con

$$Q_2 \simeq a_j + \lambda_j(b_j - a_j)$$

dove

$$\lambda_j = \frac{55\%n - F_{j-1}}{F_j - F_{j-1}} = \frac{44 - 42}{58 - 42} = 0.125\,.$$

Quindi

$$P_{55} \simeq 7 + 0.125 \cdot (11 - 7) = 7.5\,.$$

Esercizio 2. (7 punti) Il numero di errori nelle soluzioni degli esercizi scritte dal docente di Probabilità e Statistica segue una distribuzione di Poisson con una media di 2 errori per soluzione. Si assuma che i numeri di errori in soluzioni di esercizi distinti siano indipendenti.

1. Qual è la probabilità che in una soluzione ci siano almeno 4 errori (inclusi)?
2. Qual è la deviazione standard del numero di errori in una soluzione?
3. Qual è la probabilità che in 5 soluzioni ci siano almeno 4 errori (inclusi)?
4. Consideriamo 5 soluzioni. Abbiamo letto le prime 3 soluzioni e abbiamo individuato almeno 4 errori (inclusi). Sapendo questo fatto, qual è la probabilità che nelle 5 soluzioni ci siano in tutto 6 errori?
5. Uno studente legge le soluzioni degli esercizi in sequenza e si blocca quando trova la prima soluzione con almeno 1 errore (incluso). Qual è la probabilità che lo studente si blocchi entro la lettura della terza soluzione (inclusa)?

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{“numero di errori in una soluzione”} \sim \mathrm{P}(\lambda)\,.$$

Poiché

$$\lambda = \mathbb{E}(X) = 2 \implies X \sim \mathrm{P}(2)\,.$$

1. La probabilità che in una soluzione ci siano almeno 4 errori è

$$\begin{aligned}\mathbb{P}(\{X \geq 4\}) &= 1 - \mathbb{P}(X < 4) = 1 - \mathbb{P}(X \leq 3) \\ &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) - \mathbb{P}(X = 2) - \mathbb{P}(X = 3) \\ &= 1 - e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!}\right) = 14.30\%\,.\end{aligned}$$

2. La deviazione standard del numero di errori in una soluzione è

$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{\lambda} = \sqrt{2} = 1.41\,.$$

3. Per calcolare questa probabilità, consideriamo le variabili aleatorie X_1, X_2, X_3, X_4, X_5 indipendenti e identicamente distribuite con $X_i \sim \mathrm{P}(2)$. La variabile X_i rappresenta il numero di errori nella i-esima soluzione. Poiché la somma di variabili aleatorie di Poisson indipendenti è ancora una variabile aleatoria di Poisson, la variabile aleatoria $Y = X_1 + X_2 + X_3 + X_4 + X_5$ ha distribuzione $\mathrm{P}(5 \cdot 2) = \mathrm{P}(10)$. La probabilità che in 5 soluzioni ci siano almeno 4 errori è quindi

$$\begin{aligned}\mathbb{P}(\{Y \geq 4\}) &= 1 - \mathbb{P}(Y < 4) = 1 - \mathbb{P}(Y \leq 3) \\ &= 1 - \mathbb{P}(Y = 0) - \mathbb{P}(Y = 1) - \mathbb{P}(Y = 2) - \mathbb{P}(Y = 3) \\ &= 1 - e^{-10}\left(\frac{10^0}{0!} + \frac{10^1}{1!} + \frac{10^2}{2!} + \frac{10^3}{3!}\right) = 98.97\%\,.\end{aligned}$$

4. Consideriamo le variabili aleatorie

$$Y = X_1 + X_2 + X_3 \sim \mathrm{P}(6) \quad \text{e} \quad Z = X_4 + X_5 \sim \mathrm{P}(4)\,.$$

La probabilità che nelle 5 soluzioni ci siano in tutto 6 errori, sapendo che nelle prime 3 soluzioni ci sono almeno 4 errori, è, utilizzando l'indipendenza tra Y e Z,

$$\begin{aligned}
&\mathbb{P}(\{Y+Z=6\}|\{Y\geq 4\}) = \frac{\mathbb{P}(\{Y+Z=6\}\cap\{Y\geq 4\})}{\mathbb{P}(\{Y\geq 4\})} \\
&= \frac{\mathbb{P}(\{Y=4\}\cap\{Z=2\}) + \mathbb{P}(\{Y=5\}\cap\{Z=1\}) + \mathbb{P}(\{Y=6\}\cap\{Z=0\})}{1-\mathbb{P}(\{Y<4\})} \\
&= \frac{\mathbb{P}(\{Y=4\})\mathbb{P}(\{Z=2\}) + \mathbb{P}(\{Y=5\})\mathbb{P}(\{Z=1\}) + \mathbb{P}(\{Y=6\})\mathbb{P}(\{Z=0\})}{1-\mathbb{P}(\{Y=0\})-\mathbb{P}(\{Y=1\})-\mathbb{P}(\{Y=2\})-\mathbb{P}(\{Y=3\})} \\
&= \frac{e^{-6}\frac{6^4}{4!}e^{-4}\frac{4^2}{2!} + e^{-6}\frac{6^5}{5!}e^{-4}\frac{4^1}{1!} + e^{-6}\frac{6^6}{6!}e^{-4}\frac{4^0}{0!}}{1-e^{-6}\left(\frac{6^0}{0!}+\frac{6^1}{1!}+\frac{6^2}{2!}+\frac{6^3}{3!}\right)} = 4.04\%\,.
\end{aligned}$$

5. Consideriamo la variabile aleatoria

$$Y = \text{“prima soluzione con almeno 1 errore”} \sim \text{Geo}(p)\,,$$

dove p è la probabilità che una soluzione abbia almeno 1 errore, ovvero

$$p = \mathbb{P}(\{X\geq 1\}) = 1-\mathbb{P}(\{X=0\}) = 1-e^{-2} = 86.47\%\,.$$

La probabilità che lo studente si blocchi entro la lettura della terza soluzione è

$$\mathbb{P}(\{Y\leq 3\}) = 1-\mathbb{P}(\{Y>3\})1-(1-p)^3 = 1-(1-86.47\%)^3 = 99.75\%\,.$$

Esercizio 3. (8 punti) In un centro di assistenza, il tempo necessario per completare un backup di un computer segue una distribuzione esponenziale con un tempo medio di 2 ore.

1. Qual è la probabilità che il backup di un computer duri meno di 1 ora?
2. Qual è la varianza del tempo necessario per completare il backup di un computer?
3. Al centro di assistenza arrivano 16 computer per i quali occorre un backup. Qual è la probabilità che per almeno 3 computer (inclusi) il backup duri più di 3 ore? Si assuma che i tempi di backup dei computer siano indipendenti.
4. Al centro di assistenza arrivano 2 computer per i quali occorre un backup. Il backup del secondo computer inizia non appena il backup del primo computer è completato. Qual è la media del tempo necessario per completare il backup totale dei 2 computer? E la varianza? Si assuma che i tempi di backup dei computer siano indipendenti.
5. Nella situazione del punto 4., qual è la probabilità che il backup totale dei 2 computer sia inferiore a 7 ore?

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{“tempo necessario per completare un backup”} \sim \text{Exp}(\lambda)\,.$$

Poiché

$$\frac{1}{\lambda} = \mathbb{E}(X) = 2 \implies \lambda = \frac{1}{2} \implies X \sim \text{Exp}\left(\frac{1}{2}\right)\,.$$

1. La probabilità che il backup di un computer duri meno di 1 ora è

$$\mathbb{P}(\{X < 1\}) = 1 - \mathbb{P}(X \geq 1) = 1 - e^{-\frac{1}{2}} = 39.35\% \,.$$

2. La varianza del tempo necessario per completare il backup di un computer è

$$\mathrm{Var}(X) = \frac{1}{\lambda^2} = 2^2 = 4 \,.$$

3. Per calcolare questa probabilità, consideriamo la variabile aleatoria

$$Y = \text{“numero dei 16 computer per i quali il backup dura più di 3 ore”} \sim \mathrm{B}(16, p) \,,$$

dove

$$p = \mathbb{P}(\{X > 3\}) = e^{-\frac{3}{2}} = 22.31\% \,.$$

La probabilità che per almeno 3 computer il backup duri più di 3 ore è

$$\begin{aligned}
&\mathbb{P}(\{Y \geq 3\}) = 1 - \mathbb{P}(\{Y < 3\}) = 1 - \mathbb{P}(\{Y = 0\}) - \mathbb{P}(\{Y = 1\}) - \mathbb{P}(\{Y = 2\}) \\
&= 1 - \binom{16}{0}(22.31\%)^0(77.69\%)^{16} - \binom{16}{1}(22.31\%)^1(77.69\%)^{15} - \binom{16}{2}(22.31\%)^2(77.69\%)^{14} \\
&= 72.72\% \,.
\end{aligned}$$

4. Consideriamo le variabili aleatorie

$$X_1 = \text{“tempo necessario per completare il backup del primo computer”} \sim \mathrm{Exp}\left(\frac{1}{2}\right)$$

$$X_2 = \text{“tempo necessario per completare il backup del secondo computer”} \sim \mathrm{Exp}\left(\frac{1}{2}\right) \,.$$

Per calcolare il valore atteso, sfruttiamo la linearità del valore atteso per ottenere che

$$\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2) = 2 + 2 = 4 \,.$$

Per calcolare la varianza, sfruttiamo l'indipendenza delle variabili aleatorie per ottenere che

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 4 + 4 = 8 \,.$$

5. La variabile aleatoria $X_1 + X_2$ ha distribuzione $\mathrm{Gamma}(2, 1/2)$, quindi ha densità (ricordiamo che $\Gamma(2) = 1! = 1$)

$$f(x) = \begin{cases} \left(\frac{1}{2}\right)^2 x e^{-\frac{1}{2}x} & \text{per } x > 0 \\ 0 & \text{per } x \leq 0 \,. \end{cases}$$

La probabilità che il backup totale dei 2 computer sia inferiore a 7 ore è quindi, integrando per parti,

$$\begin{aligned}
\mathbb{P}(\{X_1 + X_2 < 7\}) &= \int_0^7 \left(\frac{1}{2}\right)^2 x e^{-\frac{1}{2}x} \,\mathrm{d}x = \Big[-\frac{1}{2} x e^{-\frac{1}{2}x} \Big]_0^7 + \int_0^7 \frac{1}{2} e^{-\frac{1}{2}x} \,\mathrm{d}x \\
&= -\frac{7}{2} e^{-\frac{7}{2}} + \Big[-e^{-\frac{1}{2}x} \Big]_0^7 = -\frac{7}{2} e^{-\frac{7}{2}} + 1 - e^{-\frac{7}{2}} \\
&= 1 - \frac{9}{2} e^{-\frac{7}{2}} = 86.41\% \,.
\end{aligned}$$

Esercizio 4. (7 punti) (7 punti) Un'azienda sostiene che la durata media giornaliera degli smartphone che produce è di 12 ore. Un'indagine condotta su alcuni smartphone è volta a mostrare che la durata è in realtà inferiore. Vengono rilevate le seguenti durate (in ore):

11.2	15.2	12.1	10.8	8.6	11.7	13.6	12.7	9.5	11.0
18.2	11.9	14.2	12.5	10.0	11.3	11.4	9.3	10.3	9.5
10.5	13.6	10.8	9.6	13.3	8.9	13.8	12.7	8.9	15.5
11.7	13.2	9.6	10.4	12.2	11.9	9.1	12.2	12.6	9.6

La media calcolata sui dati risulta essere 11.63 ore. È noto che la deviazione standard della popolazione è di 4 ore.

1. È possibile sostenere con significatività 1% che la durata media degli smartphone è in realtà inferiore a 12 ore? (N.B.: derivare le formule)

2. Calcolare il p-value del test.

Soluzione. Si deve impostare un test di ipotesi. La popolazione è descritta da una variabile aleatoria X con media $\mathbb{E}(X) = \mu$ e varianza $\mathrm{Var}(X) = \sigma^2 = 4^2$. La legge di X non è nota. Dalla popolazione viene estratto un campione $X_1, \dots, X_n$ con $n = 40 > 30$. Osserviamo che il campione è numeroso. Sia $\mu_0 = 12$. Il test di ipotesi è il seguente:

$$H_0 : \mu = \mu_0 \,, \quad H_1 : \mu < \mu_0 \,,$$

con livello di significatività $\alpha = 1\%$.

Poiché l'ipotesi alternativa è $H_1 : \mu < \mu_0$, i dati saranno significativi se la media è sufficiente più piccola di μ_0. La regione critica è allora della forma

$$R_c = \{(x_1, \dots, x_n) \in R(X_1, \dots, X_n) \ : \ \overline{x}_n < \mu_0 - \delta\} \,,$$

dove $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ è la media calcolata sulla realizzazione $x_1, \dots, x_n$ del campione casuale.

Assumiamo l'ipotesi nulla H_0 vera, cioè $\mu = \mu_0$, ovvero $\mathbb{E}(X_i) = \mu_0$. Consideriamo la media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ e utilizziamo la definizione di significatività per ottenere

$$\alpha = \mathbb{P}(\{(X_1, \dots, X_n) \in R_c\}) = \mathbb{P}(\{\overline{X}_n < \mu_0 - \delta\}) = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} < -\frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) \,.$$

Non conosciamo la distribuzione della popolazione, ma il campione è numeroso ($n \geq 30$). Per il Teorema del Limite Centrale, si ha che $\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \overset{n\to+\infty}{\to} Z$ in legge, dove $Z \sim \mathcal{N}(0,1)$. Quindi

$$\alpha = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} < -\frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) \simeq \mathbb{P}\Big(\Big\{Z < -\frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) = \mathbb{P}\Big(\Big\{Z > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) \,.$$

Introduciamo il valore z_α tale che

$$\mathbb{P}(\{Z > z_\alpha\}) = \alpha \,.$$

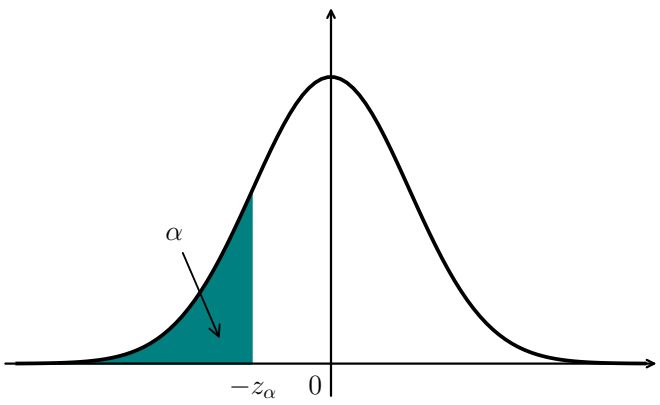


Figura 1: Coda sinistra della legge normale con probabilità α.

Allora, scegliendo

$$\frac{\delta}{\sigma/\sqrt{n}} = z_\alpha \implies \delta = \frac{\sigma}{\sqrt{n}} z_\alpha \,,$$

otteniamo la condizione desiderata sulla probabilità di errore del primo tipo.

In conclusione, la regione critica è

$$R_c = \left\{ (x_1, \dots, x_n) \in R(X_1, \dots, X_n) \ : \ \overline{x}_n < \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha \right\},$$

e decidiamo come segue:

- Se $\overline{x}_n < \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$, i dati sono sufficientemente significativi da rifiutare H_0. L'ipotesi nulla H_0 viene rifiutata (con livello di significatività α).
- Se $\overline{x}_n \geq \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$, i dati non sono sufficientemente significativi da rifiutare. L'ipotesi nulla H_0 non viene rifiutata (con livello di significatività α).

È anche possibile calcolare esplicitamente il p-value dei dati. Utilizzando il fatto che la funzione di distribuzione cumulativa della normale standard è strettamente crescente, otteniamo che

$$\begin{aligned} p\text{-value} &= \inf \left\{ \alpha \ : \ \overline{x}_n - \mu_0 < -\frac{\sigma}{\sqrt{n}} z_\alpha \right\} = \inf \left\{ \alpha \ : \ \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha \right\} \\ &= \inf \left\{ \alpha \ : \ \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Big) < \Phi(-z_\alpha) \right\} = \inf \left\{ \alpha \ : \ \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Big) < \alpha \right\} \\ &= \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Big) \,. \end{aligned}$$

Calcoliamo il p-value dei dati:

$$\Phi\Big(\frac{11.63 - 12}{4/\sqrt{40}} \Big) = \Phi(-0.58) = 1 - \Phi(0.58) = 1 - 71.90\% = 28.1\% \,.$$

Poiché $1\% < 28.1\%$, non possiamo rifiutare l'ipotesi nulla H_0 con significatività 1%.

Esame di Probabilità e Statistica [3231]

Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: luglio 2024 - turno 1
Matricola: _____ Data: 15/07/2024

Viene usata come riferimento la traccia n. 120.

Esercizio 1. (6 punti) Si studia il prezzo di affitto di appartamenti a Bari tramite annunci pubblicati su un servizio online. Vengono rilevati i seguenti dati (in euro):

750 700 550 650 900 780 1100 450 530 1650

1. Determinare i quartili (esclusivi) dei dati.
2. Determinare eventuali dati anomali e sospetti.
3. Tracciare un box plot.
4. Calcolare il 35-esimo percentile (esclusivo).

Soluzione. 1. Per prima cosa ordiniamo i dati:

$$450 \quad 530 \quad 550 \quad 650 \quad 700 \quad 750 \quad 780 \quad 900 \quad 1100 \quad 1650\,.$$

Abbiamo $n = 10$ dati.

Calcoliamo il primo quartile: $\frac{n+1}{4} = \frac{11}{4} = 2 + 0.75$. Allora

$$Q_1 = (1 - 0.75)x_2 + 0.75x_3 = 0.25 \cdot 530 + 0.75 \cdot 550 = 545\,.$$

Calcoliamo il secondo quartile: $(n+1)\frac{2}{4} = 11\frac{2}{4} = 5 + 0.5$. Allora

$$Q_2 = (1 - 0.5)x_5 + 0.5x_6 = 0.5 \cdot 700 + 0.5 \cdot 750 = 725\,.$$

Calcoliamo il terzo quartile: $(n+1)\frac{3}{4} = 11\frac{3}{4} = 8 + 0.25$. Allora

$$Q_3 = (1 - 0.25)x_8 + 0.25x_9 = 0.75 \cdot 900 + 0.25 \cdot 1100 = 950\,.$$

2. Per determinare i dati anomali e sospetti calcoliamo il range interquartile

$$IQR = Q_3 - Q_1 = 950 - 545 = 405\,.$$

I dati anomali apparterrebbero agli intervalli

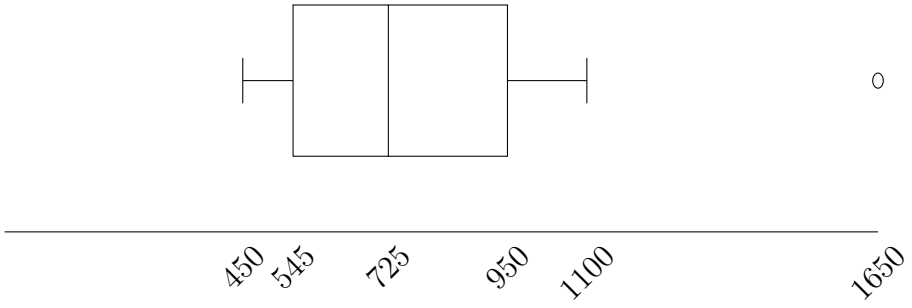
$$(-\infty, Q_1 - 3 \cdot IQR] \cup [Q_3 + 3 \cdot IQR, +\infty) = (-\infty, -670] \cup [2165, +\infty)\,,$$

quindi non ci sono dati anomali. I dati sospetti appartengono agli intervalli

$$(Q_1 - 3 \cdot IQR, Q_1 - 1.5 \cdot IQR] \cup [Q_3 + 1.5 \cdot IQR, Q_3 + 3 \cdot IQR) = (-670, -62.5] \cup [1557.5, 2165)\,,$$

quindi 1650 è un dato sospetto.

3. Segue il box-plot.



4. Calcoliamo il 35-esimo percentile. Osserviamo che $(n+1)35\% = 11 \cdot 35\% = 3 + 0.85$. Allora il 35-esimo percentile è

$$P_{35} = (1 - 0.85)x_3 + 0.85x_4 = 0.15 \cdot 550 + 0.85 \cdot 650 = 635\,.$$

Esercizio 2. (7 punti) In una linea di produzione, il 5% dei prodotti è difettoso. Si assuma che prodotti diversi siano indipendenti tra loro.

Nei punti 1., 2., 3. vengono esaminati 10 prodotti.

1. Qual è la probabilità che (strettamente) più di 3 prodotti siano difettosi?
2. Qual è la probabilità che esattamente 3 prodotti siano difettosi?
3. Quali sono la media e la varianza del numero di prodotti difettosi?

Nel punto 4. vengono esaminati 100 prodotti.

4. Considerando che il numero di prodotti è elevato e la probabilità di difetto è piccola, qual è un'approssimazione adeguata della probabilità che almeno 4 (4 inclusi) prodotti siano difettosi? Motivare la risposta.

Nel punto 5. vengono esaminati prodotti in sequenza fino a trovare il primo difettoso.

5. Qual è la probabilità che il primo prodotto difettoso sia il quarto esaminato?

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{numero di prodotti difettosi tra } 10 \sim \mathrm{B}(10, 5\%).$$

1. La probabilità che più di 3 prodotti siano difettosi è

$$\begin{aligned}\mathbb{P}(\{X > 3\}) &= 1 - \mathbb{P}(\{X \le 3\}) = 1 - \mathbb{P}(\{X = 0\}) - \mathbb{P}(\{X = 1\}) - \mathbb{P}(\{X = 2\}) - \mathbb{P}(\{X = 3\}) \\ &= 1 - \binom{10}{0} 0.05^0 0.95^{10} - \binom{10}{1} 0.05^1 0.95^9 - \binom{10}{2} 0.05^2 0.95^8 - \binom{10}{3} 0.05^3 0.95^7. \\ &\simeq 0.10\%.\end{aligned}$$

2. La probabilità che esattamente 3 prodotti siano difettosi è

$$\mathbb{P}(\{X = 3\}) = \binom{10}{3} 0.05^3 0.95^7 \simeq 1.05\%.$$

3. La media e la varianza di X sono

$$\mathbb{E}(X) = 10 \cdot 0.05 = 0.5,$$
$$\mathrm{Var}(X) = 10 \cdot 0.05 \cdot 0.95 = 0.475.$$

4. Una legge binomiale con un numero elevato di prove e probabilità di successo piccola può essere approssimata con una legge di Poisson con parametro uguale al prodotto del numero di prove e della probabilità di successo.

Consideriamo allora la variabile aleatoria

$$Y = \text{numero di prodotti difettosi tra } 100 \sim \mathrm{P}(100 \cdot 5\%) = \mathrm{P}(5).$$

Calcoliamo

$$\begin{aligned}\mathbb{P}(\{Y \geq 4\}) &= 1 - \mathbb{P}(\{Y < 4\}) = 1 - \mathbb{P}(\{Y = 0\}) - \mathbb{P}(\{Y = 1\}) - \mathbb{P}(\{Y = 2\}) - \mathbb{P}(\{Y = 3\}) \\ &= 1 - e^{-5}\left(1 + 5 + \frac{5^2}{2} + \frac{5^3}{6}\right) \simeq 73.50\%.\end{aligned}$$

5. Consideriamo la variabile aleatoria

$$Z = \text{numero di prodotti esaminati fino al primo difettoso} \sim \mathrm{Geo}(5\%).$$

La probabilità che il primo difettoso sia il quarto esaminato è

$$\mathbb{P}(\{Z = 4\}) = 0.95^3 \cdot 0.05 \simeq 4.29\%\,.$$

Esercizio 3. (8 punti) Alice e Bob generano due numeri casualmente. Alice genera un numero con una variabile casuale esponenziale con media 2, mentre Bob genera un numero in base al risultato ottenuto da Alice:

- Se il numero generato da Alice è minore di 1, Bob genera un numero con una variabile aleatoria uniforme $\mathrm{U}(0, 1)$.
- Se il numero generato da Alice è maggiore di 1, Bob genera un numero con una variabile aleatoria uniforme $\mathrm{U}(1, 2)$.

Si risponda alle seguenti domande:

1. Sapendo che Alice ha ottenuto un numero minore di 1, qual è la media del numero generato da Bob? E la deviazione standard?
2. Qual è la probabilità che il numero generato da Bob sia maggiore di $\frac{1}{2}$?
3. Bob ha generato un numero minore di $\frac{3}{2}$. Sapendo questo fatto, qual è la probabilità che il numero generato da Alice sia minore di 1?
4. Qual è la probabilità che il minimo tra il numero generato da Alice e quello generato da Bob sia minore di 1?

Soluzione. Consideriamo le variabili aleatorie

$$X = \text{numero generato da Alice} \sim \text{Exp}(\lambda),$$

$$Y = \text{numero generato da Bob}\,.$$

Per la X sappiamo che $\mathbb{E}(X) = 1/\lambda = 2$, quindi $\lambda = 1/2$, cioè $X \sim \text{Exp}(1/2)$.

Per la Y, conviene distinguere in base all'esito di X:

$$Y_{<1} = \text{numero generato da Bob se } X < 1 \sim \text{U}(0,1),$$

$$Y_{>1} = \text{numero generato da Bob se } X > 1 \sim \text{U}(1,2).$$

1. La media del numero generato da Bob sapendo che $X < 1$ è

$$\mathbb{E}(Y_{<1}) = \frac{0+1}{2}\,.$$

La deviazione standard è

$$\sqrt{\text{Var}(Y_{<1})} = \sqrt{\frac{(1-0)^2}{12}} = \frac{1}{\sqrt{12}}\,.$$

2. Per calcolare la probabilità che il numero generato da Bob sia maggiore di $1/2$ possiamo usare il teorema della probabilità totale:

$$\begin{aligned}\mathbb{P}(\{Y > 1/2\}) &= \mathbb{P}(\{Y > 1/2\}|\{X < 1\})\mathbb{P}(\{X < 1\}) + \mathbb{P}(\{Y > 1/2\}|\{X > 1\})\mathbb{P}(\{X > 1\}) \\ &= \mathbb{P}(\{Y_{<1} > 1/2\})\mathbb{P}(\{X < 1\}) + \mathbb{P}(\{Y_{>1} > 1/2\})\mathbb{P}(\{X > 1\})\,.\end{aligned}$$

Calcoliamo

$$\mathbb{P}(\{X > 1\}) = e^{-1/2} \simeq 60.65\%\,.$$

Inoltre

$$\mathbb{P}(\{Y_{<1} > 1/2\}) = \int_0^{1/2} \frac{1}{1-0}\,\mathrm{d}x = \frac{1}{2}\,,$$
$$\mathbb{P}(\{Y_{>1} > 1/2\}) = 1\,.$$

Quindi

$$\mathbb{P}(\{Y > 1/2\}) = \frac{1}{2}\cdot(1 - e^{-1/2}) + 1 \cdot e^{-1/2} = 80.33\%\,.$$

3. Utilizziamo il teorema di Bayes:

$$\begin{aligned}\mathbb{P}(\{X < 1\}|\{Y < 3/2\}) &= \frac{\mathbb{P}(\{Y < 3/2\}|\{X < 1\})\mathbb{P}(\{X < 1\})}{\mathbb{P}(\{Y < 3/2\})} \\ &= \frac{\mathbb{P}(\{Y_{<1} < 3/2\})\mathbb{P}(\{X < 1\})}{\mathbb{P}(\{Y_{<1} < 3/2\})\mathbb{P}(\{X < 1\}) + \mathbb{P}(\{Y_{>1} < 3/2\})\mathbb{P}(\{X > 1\})} \\ &= \frac{1\cdot(1-e^{-1/2})}{1\cdot(1-e^{-1/2}) + \frac{3/2-1}{1}\cdot e^{-1/2}} \simeq 56.47\%\,.\end{aligned}$$

4. Dobbiamo calcolare

$$\mathbb{P}(\{\min\{X,Y\} < 1\}) = 1 - \mathbb{P}(\{\min\{X,Y\} \geq 1\}) = 1 - \mathbb{P}(\{X \geq 1\} \cap \{Y \geq 1\})\,.$$

Le variabili aleatorie X e Y non sono indipendenti, quindi per calcolare la probabilità dell'intersezione dobbiamo utilizzare la formula della probabilità condizionata:

$$\begin{aligned}\mathbb{P}(\{X \geq 1\} \cap \{Y \geq 1\}) &= \mathbb{P}(\{Y \geq 1\}|\{X \geq 1\})\mathbb{P}(\{X \geq 1\}) \\ &= \mathbb{P}(\{Y_{>1} \geq 1\})\mathbb{P}(\{X \geq 1\}) \\ &= 1 \cdot e^{-1/2} \simeq 60.65\% \,.\end{aligned}$$

Quindi

$$\mathbb{P}(\{\min\{X, Y\} < 1\}) \simeq 39.35\% \,.$$

Esercizio 4. (7 punti) Una squadra di calcio lo scorso anno aveva una media di gol per partita pari a 2.5. Nel campionato di quest'anno, la squadra ha segnato un numero di gol per partita come descritto dai dati raccolti nella seguente tabella:

gol per partita	frequenza assoluta
0	5
1	9
2	13
3	7
4	4

Si assuma che la deviazione standard del numero di gol sia 2.

1. Si può stabilire con significatività del 5% che la media di gol per partita è diversa da quella dell'anno scorso? (N.B.: derivare le formule)
2. Stabilire in quali dei seguenti intervalli è collocato il p-value dei dati: $[0, 1\%)$, $[1\%, 2\%)$, $[2\%, 5\%)$, $[5\%, 10\%)$, $[10\%, 20\%)$, $[20\%, +\infty)$.

Soluzione. Si deve impostare un test di ipotesi. La popolazione è descritta da una variabile aleatoria X con media $\mathbb{E}(X) = \mu$ e varianza $\mathrm{Var}(X) = \sigma^2 = 2^2$. La legge di X non è nota. Dalla popolazione viene estratto un campione $X_1, \ldots, X_n$ con $n = 5+9+13+7+4 = 38 > 30$. Osserviamo che il campione è numeroso. Sia $\mu_0 = 2.5$. Il test di ipotesi è il seguente:

$$H_0 : \mu = \mu_0 \,, \quad H_1 : \mu \neq \mu_0 \,,$$

con livello di significatività $\alpha = 5\%$.

Poiché l'ipotesi alternativa è $H_1 : \mu \neq \mu_0$, i dati saranno significativi se la media è sufficientemente distante da μ_0. La regione critica è allora della forma

$$R_c = \{(x_1, \ldots, x_n) \in R(X_1, \ldots, X_n) \ : \ |\overline{x}_n - \mu_0| > \delta\} \,,$$

dove $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ è la media calcolata sulla realizzazione $x_1, \ldots, x_n$ del campione casuale.

Assumiamo l'ipotesi nulla H_0 vera, cioè $\mu = \mu_0$, ovvero $\mathbb{E}(X_i) = \mu_0$. Consideriamo la media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ e utilizziamo la definizione di significatività per ottenere

$$\alpha = \mathbb{P}(\{(X_1, \ldots, X_n) \in R_c\}) = \mathbb{P}(\{|\overline{X}_n - \mu_0| > \delta\}) = \mathbb{P}\Big(\Big\{\frac{|\overline{X}_n - \mu_0|}{\sigma/\sqrt{n}} > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) \,.$$

Non conosciamo la distribuzione della popolazione, ma il campione è numeroso ($n \geq 30$). Per il Teorema del Limite Centrale, si ha che $\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \overset{n\to+\infty}{\to} Z$ in legge, dove $Z \sim \mathcal{N}(0, 1)$. Quindi

$$\alpha = \mathbb{P}\Big(\Big\{\frac{|\overline{X}_n - \mu_0|}{\sigma/\sqrt{n}} > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) \simeq \mathbb{P}\Big(\Big\{|Z| > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) \,.$$

Per la simmetria della gaussiana

$$\alpha \simeq \mathbb{P}\Big(\Big\{|Z| > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) = \mathbb{P}\Big(\Big\{Z < -\frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) + \mathbb{P}\Big(\Big\{Z > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big)$$
$$= 2\mathbb{P}\Big(\Big\{Z > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big),$$

da cui

$$\mathbb{P}\Big(\Big\{Z > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) = \frac{\alpha}{2}.$$

Introduciamo il valore $z_{\alpha/2}$ tale che

$$\mathbb{P}(\{Z > z_{\alpha/2}\}) = \frac{\alpha}{2}.$$

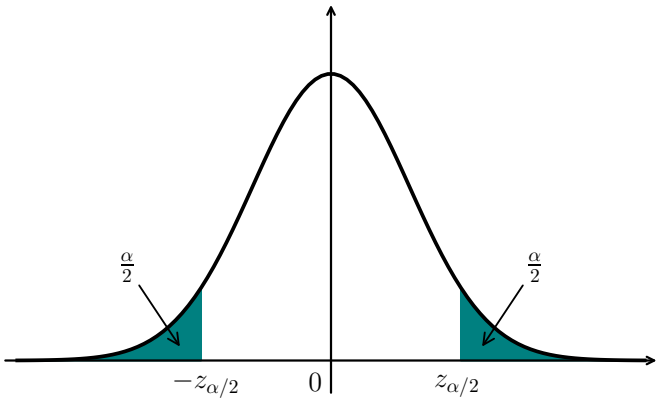


Figura 1: Code della legge normale con probabilità $\frac{\alpha}{2}$.

Allora, scegliendo

$$\frac{\delta}{\sigma/\sqrt{n}} = z_{\alpha/2} \implies \delta = \frac{\sigma}{\sqrt{n}} z_{\alpha/2},$$

otteniamo la condizione che definisce la significatività.

In conclusione, la regione critica è

$$R_c = \Big\{(x_1, \dots, x_n) \in R(X_1, \dots, X_n) \ : \ |\overline{x}_n - \mu_0| > \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\Big\},$$

e decidiamo come segue:

- Se $|\overline{x}_n - \mu_0| > \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$, i dati sono sufficientemente significativi da rifiutare H_0. L'ipotesi nulla H_0 viene rifiutata (con livello di significatività α).
- Se $|\overline{x}_n - \mu_0| \leq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$, i dati non sono sufficientemente significativi da rifiutare. L'ipotesi nulla H_0 non viene rifiutata (con livello di significatività α).

In questo caso è anche possibile calcolare esplicitamente il p-value dei dati. Utilizzando il fatto che la funzione di distribuzione cumulativa della normale standard è strettamente crescente, otteniamo che

$$p\text{-value} = \inf\Big\{\alpha \ : \ |\overline{x}_n - \mu_0| > \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\Big\} = \inf\Big\{\alpha \ : \ \frac{|\overline{x}_n - \mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}\Big\}$$
$$= \inf\Big\{\alpha \ : \ \Phi\Big(\frac{|\overline{x}_n - \mu_0|}{\sigma/\sqrt{n}}\Big) > \Phi(z_{\alpha/2})\Big\} = \inf\Big\{\alpha \ : \ \Phi\Big(\frac{|\overline{x}_n - \mu_0|}{\sigma/\sqrt{n}}\Big) > 1 - \frac{\alpha}{2}\Big\}$$
$$= \inf\Big\{\alpha \ : \ \alpha > 2 - 2\Phi\Big(\frac{|\overline{x}_n - \mu_0|}{\sigma/\sqrt{n}}\Big)\Big\} = 2\Big(1 - \Phi\Big(\frac{|\overline{x}_n - \mu_0|}{\sigma/\sqrt{n}}\Big)\Big).$$

Calcoliamo il p-value sui dati:

$$\overline{x}_n = \frac{1}{n}\sum_{j=1}^{k} f_j v_j = \frac{1}{38}(0\cdot 5 + 1\cdot 9 + 2\cdot 13 + 3\cdot 7 + 4\cdot 4) \simeq 1.89\,.$$

Lo z-score è quindi:

$$\frac{|\overline{x}_n - \mu_0|}{\sigma/\sqrt{n}} = \frac{|1.89 - 2.5|}{2/\sqrt{38}} = 1.88\,.$$

Quindi

$$p\text{-value} = 2\Big(1 - \Phi(1.88)\Big) \simeq 2(1 - 0.9699) = 6.02\%\,.$$

Concludiamo che:

1. L'ipotesi nulla non può essere rifiutata con significatività $5\% < 6.02\%$.
2. Il p-value è collocato nell'intervallo $[5\%, 10\%)$.

Esame di Probabilità e Statistica [3231]

Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: luglio 2024 - turno 2
Matricola: _____ Data: 15/07/2024

Viene usata come riferimento la traccia n. 1.

Esercizio 1. (6 punti) In una partita di pallacanestro viene misurata la distanza dei tiri effettuati da un giocatore. Vengono misurati i seguenti dati (in metri):

$$2.3 \quad 8.6 \quad 3.1 \quad 5.2 \quad 1.1 \quad 5.6 \quad 6.3 \quad 1.1 \quad 7.5 \quad 6.8\,.$$

1. Determinare i quartili (esclusivi) dei dati.
2. Determinare eventuali dati anomali e sospetti.
3. Tracciare un box plot.
4. Calcolare il 65-esimo percentile (esclusivo).

Soluzione. 1. Per prima cosa ordiniamo i dati:

$$1.1 \quad 1.1 \quad 2.3 \quad 3.1 \quad 5.2 \quad 5.6 \quad 6.3 \quad 6.8 \quad 7.5 \quad 8.6\,.$$

Abbiamo $n = 10$ dati.

Calcoliamo il primo quartile: $\frac{n+1}{4} = \frac{11}{4} = 2 + 0.75$. Allora

$$Q_1 = (1 - 0.75)x_2 + 0.75x_3 = 0.25 \cdot 1.1 + 0.75 \cdot 2.3 = 2\,.$$

Calcoliamo il secondo quartile: $(n+1)\frac{2}{4} = 11\frac{2}{4} = 5 + 0.5$. Allora

$$Q_2 = (1 - 0.5)x_5 + 0.5x_6 = 0.5 \cdot 5.2 + 0.5 \cdot 5.6 = 5.4\,.$$

Calcoliamo il terzo quartile: $(n+1)\frac{3}{4} = 11\frac{3}{4} = 8 + 0.25$. Allora

$$Q_3 = (1 - 0.25)x_8 + 0.25x_9 = 0.75 \cdot 6.8 + 0.25 \cdot 7.5 = 6.975\,.$$

2. Per determinare i dati anomali e sospetti calcoliamo il range interquartile

$$IQR = Q_3 - Q_1 = 6.975 - 2 = 4.975\,.$$

I dati anomali apparterrebbero agli intervalli

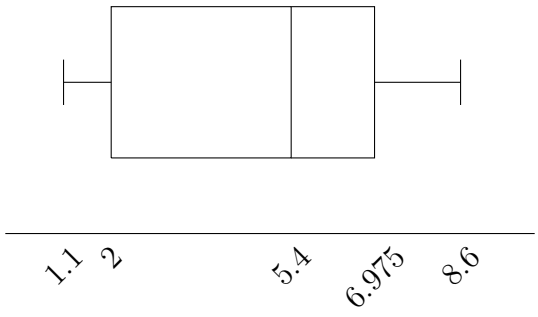
$$(-\infty, Q_1 - 3 \cdot IQR] \cup [Q_3 + 3 \cdot IQR, +\infty) = (-\infty, -12.925] \cup [21.9, +\infty)\,,$$

quindi non ci sono dati anomali. I dati sospetti appartengono agli intervalli

$$(Q_1-3\cdot IQR, Q_1-1.5\cdot IQR]\cup[Q_3+1.5\cdot IQR, Q_3+3\cdot IQR) = (-12.925, -5.4625]\cup[14.4375, 21.9)\,,$$

quindi non ci sono dati sospetti.

3. Segue il box-plot.



4. Calcoliamo il 65-esimo percentile. Osserviamo che $(n+1)65\% = 11\cdot 65\% = 7 + 0.15$. Allora il 65-esimo percentile è

$$P_{65} = (1-0.15)x_7 + 0.15x_8 = 0.85\cdot 6.3 + 0.15\cdot 6.8 = 6.375\,.$$

Esercizio 2. (8 punti) Una compagnia aerea studia il numero di bagagli smarriti durante i voli intercontinentali su cui opera. Si osserva che il numero di bagagli smarriti in un mese è distribuito con legge di Poisson con una media di 4 bagagli smarriti al mese. Si assuma che i bagagli smarriti in mesi diversi siano indipendenti tra loro.

1. Qual è la probabilità che in un mese vengano smarriti almeno 3 bagagli (3 inclusi)?
2. Qual è la varianza del numero di bagagli smarriti in un mese?
3. Qual è la probabilità che in un anno vengano smarriti esattamente 50 bagagli?
4. L'azienda nota che a gennaio e febbraio sono stati smarriti in tutto 6 bagagli. Sapendo che si è verificato questo evento, qual è la probabilità che da gennaio ad aprile vengano smarriti in tutto al più 10 bagagli (10 inclusi)?
5. Applicare il Teorema del Limite Centrale per stimare la probabilità che in 3 anni vengano smarriti più di 150 bagagli.

Soluzione. Sappiamo che la variabile aleatoria

$$X = \text{“numero di bagagli smarriti in un mese”}$$

è distribuita con legge $\mathrm{P}(\lambda)$. Inoltre

$$4 = \mathbb{E}(X) = \lambda \implies X \sim \mathrm{P}(4).$$

1. Ricordando che $\mathbb{P}(\{X=k\}) = e^{-\lambda}\frac{\lambda^k}{k!}$, otteniamo che

$$\begin{aligned}\mathbb{P}(\{X\geq 3\}) &= 1-\mathbb{P}(\{X<3\}) = 1-\left(e^{-\lambda}+e^{-\lambda}\lambda+e^{-\lambda}\frac{\lambda^2}{2}\right) = 1-e^{-4}\Big(1+4+8\Big)\\ &= 1-13e^{-4}\simeq 76.19\%\,.\end{aligned}$$

2. La varianza di X è $\mathrm{Var}(X) = \lambda = 4$.

3. Definiamo $X_1, \dots, X_{12} \sim \mathrm{P}(4)$ le variabili aleatorie che rappresentano il numero di bagagli smarriti nei 12 mesi dell'anno, cioè

$$X_i = \text{“numero di bagagli smarriti nel mese } i\text{”}\,.$$

Il numero di bagagli smarriti in un anno è la variabile aleatoria

$$Y = X_1 + \cdots + X_{12}\,.$$

Poiché le variabili X_i sono indipendenti tra loro, possiamo stabilire che $Y \sim \mathrm{P}(4 \cdot 12) = \mathrm{P}(48)$. Dunque

$$\mathbb{P}(\{Y = 50\}) = e^{-48}\frac{48^{20}}{20!} \simeq 5.4\%\,.$$

4. La variabile aleatoria che rappresenta il numero di bagagli smarriti nei mesi da gennaio ad aprile è

$$X_1 + X_2 + X_3 + X_4 \sim \mathrm{P}(4+4+4+4) = \mathrm{P}(16)\,.$$

La variabile aleatoria che rappresenta il numero di bagagli smarriti nei mesi da gennaio a febbraio è

$$X_1 + X_2 \sim \mathrm{P}(4+4) = \mathrm{P}(8)\,.$$

La variabile aleatoria che rappresenta il numero di bagagli smarriti nei mesi da marzo ad aprile è

$$X_3 + X_4 \sim \mathrm{P}(4+4) = \mathrm{P}(8)\,.$$

Dunque, la probabilità che da gennaio ad aprile vengano smarriti in tutto al più 10 bagagli sapendo che a gennaio e febbraio sono stati smarriti in tutto 6 bagagli è, utilizzando l'indipendenza,

$$\begin{aligned}
&\mathbb{P}(\{X_1 + X_2 + X_3 + X_4 \le 10\} | \{X_1 + X_2 = 6\}) \\
&= \frac{\mathbb{P}(\{X_1 + X_2 + X_3 + X_4 \le 10\} \cap \{X_1 + X_2 = 6\})}{\mathbb{P}(\{X_1 + X_2 = 6\})} \\
&= \frac{\mathbb{P}(\{X_3 + X_4 \le 4\} \cap \{X_1 + X_2 = 6\})}{\mathbb{P}(\{X_1 + X_2 = 6\})} \\
&= \frac{\mathbb{P}(\{X_3 + X_4 \le 4\})\mathbb{P}(\{X_1 + X_2 = 6\})}{\mathbb{P}(\{X_1 + X_2 = 6\})} \\
&= \mathbb{P}(\{X_3 + X_4 \le 4\}) \\
&= \mathbb{P}(\{X_3 + X_4 = 0\}) + \mathbb{P}(\{X_3 + X_4 = 1\}) + \mathbb{P}(\{X_3 + X_4 = 2\}) \\
&\quad + \mathbb{P}(\{X_3 + X_4 = 3\}) + \mathbb{P}(\{X_3 + X_4 = 4\}) \\
&= e^{-8}\frac{8^0}{0!} + e^{-8}\frac{8^1}{1!} + e^{-8}\frac{8^2}{2!} + e^{-8}\frac{8^3}{3!} + e^{-8}\frac{8^4}{4!} \simeq 9.96\%\,.
\end{aligned}$$

5. La variabile aleatoria che rappresenta il numero di bagagli smarriti in 3 anni è

$$X_1 + \cdots + X_{36} \sim \mathrm{P}(4 \cdot 12 \cdot 3) = \mathrm{P}(144)\,.$$

Applichiamo il Teorema del Limite Centrale alla media campionaria $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$, dove $X_1, \dots, X_n \sim \mathrm{P}(4)$ sono indipendenti e identicamente distribuite. La media e la varianza delle X_i sono rispettivamente

$$\mu = \mathbb{E}(X_i) = \lambda = 4 \quad \text{e} \quad \sigma^2 = \mathrm{Var}(X_i) = \lambda = 4\,.$$

Per il Teorema del Limite Centrale, la variabile aleatoria

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

è approssimata da una variabile aleatoria normale standard $Z \sim \mathcal{N}(0,1)$ per n grande. Nel nostro caso, $n = 36 > 30$.

Allora possiamo calcolare

$$\begin{aligned}
&\mathbb{P}(\{X_1 + \cdots + X_{36} > 150\}) = \mathbb{P}(\{X_1 + \cdots + X_{36} > 150.5\}) \\
&\simeq \mathbb{P}\Big(\Big\{\overline{X}_n \geq 4.18\Big\}\Big) = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \geq \frac{4.18 - 4}{2/\sqrt{36}}\Big\}\Big) = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \geq 0.54\Big\}\Big) \\
&\simeq \mathbb{P}(\{Z \geq 0.54\}) = 1 - \mathbb{P}(\{Z < 0.54\}) \simeq 1 - 70.54\% = 29.46\%\,,
\end{aligned}$$

dove abbiamo usato la tabella della distribuzione normale standard per calcolare $\mathbb{P}(\{Z < 0.54\}) \simeq 70.54\%$.

Esercizio 3. (7 punti) Un call center studia la durata delle telefonate effettuate. Si osserva che:

- Se un cliente non abbandona la chiamata, la durata della telefonata è distribuita con legge uniforme nell'intervallo $[5, 10]$ minuti.
- Se un cliente abbandona la chiamata, la durata della telefonata è distribuita con legge uniforme nell'intervallo $[2, 5]$ minuti.
- La probabilità che un cliente abbandoni la chiamata è 80%.

Si risponda ai seguenti quesiti:

1. Si consideri un cliente che abbandona la chiamata. Qual è la probabilità che la durata della telefonata sia inferiore a 3 minuti?
2. Si consideri un cliente che abbandona la chiamata. Dopo 3 minuti non è ancora terminata la chiamata. Sapendo questo fatto, qual è la probabilità che l'intera chiamata duri almeno 4 minuti? C'è un teorema che si può applicare per calcolare questa probabilità?
3. Calcolare la media e la varianza delle durate delle telefonate per un cliente che non abbandona la chiamata.
4. Si consideri un cliente qualunque. Calcolare la probabilità che la durata della telefonata sia superiore a 3 minuti.

Soluzione.
Definiamo le variabili aleatorie

$$X_0 = \text{“durata della telefonata se il cliente non abbandona”} \sim \mathrm{U}(5, 10)$$

e

$$X_1 = \text{“durata della telefonata se il cliente abbandona”} \sim \mathrm{U}(2, 5)\,.$$

Inoltre definiamo

$$Y = \text{“il cliente abbandona la chiamata”} \sim \mathrm{Be}(80\%)\,.$$

Infine definiamo la variabile aleatoria

$$X = \text{“durata della telefonata”}.$$

1. La probabilità che la durata della telefonata sia inferiore a 3 minuti sapendo che il cliente ha abbandonato la chiamata è

$$\mathbb{P}(\{X < 3\}|\{Y = 1\}) = \mathbb{P}(\{X_1 < 3\}) = \int_2^3 \frac{1}{5-2}\,\mathrm{d}x = \frac{1}{3}\,.$$

2. Attenzione: non si può usare l'assenza di memoria perché non si tratta di una legge esponenziale. Calcoliamo la probabilità richiesta utilizzando la definizione di probabilità condizionata:

$$\mathbb{P}(\{X_1 > 4\}|\{X_1 > 3\}) = \frac{\mathbb{P}(\{X_1 > 4\} \cap \{X_1 > 3\})}{\mathbb{P}(\{X_1 > 3\})} = \frac{\mathbb{P}(\{X_1 > 4\})}{\mathbb{P}(\{X_1 > 3\})} = \frac{\int_4^5 \frac{1}{5-2}\,\mathrm{d}x}{\int_3^5 \frac{1}{5-2}\,\mathrm{d}x} = \frac{1/3}{2/3} = \frac{1}{2}\,.$$

3. La media e la varianza della variabile aleatoria X_0 sono rispettivamente

$$\mathbb{E}(X_0) = \frac{5+10}{2} = 7.5 \quad \text{e} \quad \mathrm{Var}(X_0) = \frac{(10-5)^2}{12} = \frac{25}{12}\,.$$

4. Utilizzando il teorema della probabilità totale, la probabilità che la durata della telefonata sia superiore a 3 minuti è

$$\begin{aligned}\mathbb{P}(\{X > 3\}) &= \mathbb{P}(\{X > 3\}|\{Y = 0\})\mathbb{P}(\{Y = 0\}) + \mathbb{P}(\{X > 3\}|\{Y = 1\})\mathbb{P}(\{Y = 1\}) \\ &= \mathbb{P}(\{X_0 > 3\})\mathbb{P}(\{Y = 0\}) + \mathbb{P}(\{X_1 > 3\})\mathbb{P}(\{Y = 1\})\end{aligned}$$

Osserviamo che

$$\mathbb{P}(\{X_0 > 3\}) = 1\,, \quad \mathbb{P}(\{X_1 > 3\}) = 1 - \mathbb{P}(\{X_1 \le 3\}) = 1 - \frac{1}{3} = \frac{2}{3}\,,$$

quindi

$$\mathbb{P}(\{X > 3\}) = 1 \cdot 20\% + \frac{2}{3} \cdot 80\% = 73.33\%.$$

Esercizio 4. (7 punti) Il prezzo medio degli immobili venduti in una città nel 2023 era 2100€/mq. Vengono esaminati i prezzi di alcuni immobili nel 2024, osservando i seguenti dati (in €/mq):

2111 2410 1600 3900 1988 1875 2250

Si assuma che il prezzo a metro quadro degli immobili sia distribuito con legge normale.

1. È possibile affermare con significatività del 5% che il prezzo medio degli immobili nel 2024 è aumentato rispetto al 2023?

2. Stabilire in quali dei seguenti intervalli è collocato il p-value dei dati: $[0, 0.5\%]$, $[0.5\%, 1\%)$, $[1\%, 2.5\%)$, $[2.5\%, 5\%)$, $[5\%, 10\%)$, $[10\%, 100\%]$.

Soluzione. 1. Dobbiamo impostare un test di ipotesi. Sia $\mu_0 = 2100$. Consideriamo il seguente test di ipotesi

$$H_0 : \mu = \mu_0\,, \quad H_1 : \mu > \mu_0\,,$$

con livello di significatività $\alpha = 5\%$.

Poiché l'ipotesi alternativa è $H_1 : \mu > \mu_0$, i dati saranno significativi se con media calcolata sui dati sufficientemente più grande di μ_0. La regione critica è allora della forma

$$R_c = \{(x_1, \dots, x_n) \in R(X_1, \dots, X_n) \ : \ \overline{x}_n > \mu_0 + \delta\}\,,$$

dove $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ è la media calcolata sulla realizzazione $x_1, \dots, x_n$ del campione casuale.

Assumiamo l'ipotesi nulla H_0 vera, cioè $\mu = \mu_0$. Quindi $X_1, \dots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$. Non è nota σ^2, quindi verrà stimata dalla varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, dove $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ è la media campionaria. Dalla definizione di livello di significatività:

$$\alpha = \mathbb{P}(\{(X_1, \dots, X_n) \in R_c\}) = \mathbb{P}(\{\overline{X}_n > \mu_0 + \delta\}) = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big)\,.$$

Poiché $X_1, \dots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$ indipendenti, si ha che $T_{n-1} = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1)$. Quindi

$$\alpha = \mathbb{P}\Big(\Big\{T_{n-1} > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big)\,.$$

Introduciamo il valore $t_{n-1,\alpha}$ tale che

$$\mathbb{P}(\{T_{n-1} > t_{n-1,\alpha}\}) = \alpha\,.$$

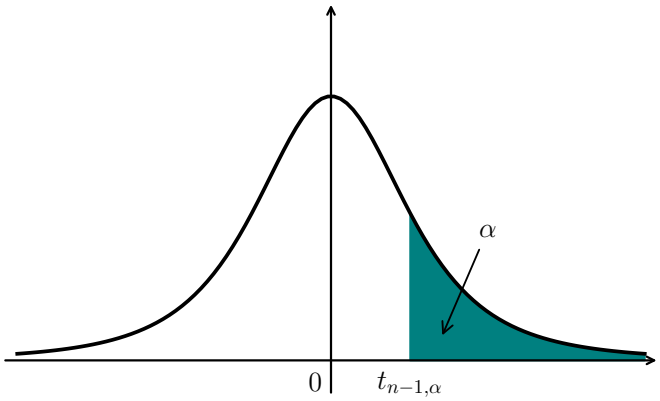


Figura 1: Coda destra della legge t-Student con probabilità α.

Allora, scegliendo

$$\frac{\delta}{S_n/\sqrt{n}} = t_{n-1,\alpha} \implies \delta = \frac{S_n}{\sqrt{n}} t_{n-1,\alpha}\,,$$

otteniamo la condizione che definisce la significatività.

In conclusione, la regione critica è

$$R_c = \Big\{(x_1, \dots, x_n) \in R(X_1, \dots, X_n) \ : \ \overline{x}_n > \mu_0 + \frac{s_n}{\sqrt{n}} t_{n-1,\alpha}\Big\}\,,$$

e decidiamo come segue:

- Se $\overline{x}_n > \mu_0 + \frac{s_n}{\sqrt{n}} t_{n-1,\alpha}$, i dati sono sufficientemente significativi da rifiutare H_0. L'ipotesi nulla H_0 viene rifiutata (con livello di significatività α).
- Se $\overline{x}_n \leq \mu_0 + \frac{s_n}{\sqrt{n}} t_{n-1,\alpha}$, i dati non sono sufficientemente significativi da rifiutare. L'ipotesi nulla H_0 non viene rifiutata (con livello di significatività α).

Calcoliamo sui dati:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{7}(2111 + 2410 + 1600 + 3900 + 1988 + 1875 + 2250) \simeq 2304.86\,.$$

$$s_n^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - \overline{x}_n^2\Big)$$
$$= \frac{1}{6}\Big(2111^2 + 2410^2 + 1600^2 + 3900^2 + 1988^2 + 1875^2 + 2250^2 - 7 \cdot 2304.86^2\Big) \simeq 563005.44\,,$$

da cui

$$s_n = \sqrt{563005.44} \simeq 750.34\,.$$

Dalle tavole otteniamo che

$$t_{n-1,\alpha} = t_{6,0.05} \simeq 1.943\,.$$

In conclusione

$$\mu_0 + \frac{s_n}{\sqrt{n}} t_{n-1,\alpha} = 2100 + \frac{750.34}{\sqrt{7}} \cdot 1.943 \simeq 2651.04\,.$$

Poiché $\overline{x}_n = 2304.86 < 2651.04$, non possiamo rifiutare l'ipotesi nulla.

2. Dalla tavola della distribuzione t-Student, non possiamo calcolare esplicitamente il p-value. Tuttavia, calcolando il t-score, abbiamo che

$$\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} = \frac{2304.86 - 2100}{750.34/\sqrt{7}} = 0.722\,.$$

Osserviamo che $0.722 < 1.440 = t_{6,0.1}$. Quindi il p-value dei dati è maggiore del 10%.

Esame di Probabilità e Statistica [3231]

Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: settembre 2024 - I
Matricola: _____ Data: 03/09/2024

Esercizio 1. (6 punti) Vengono raccolti i risultati (in metri) nel salto in alto di un campione olimpionico ottenuti negli ultimi anni:

2.32 2.35 2.37 2.31 2.33 2.30 2.29 2.36 2.37 2.22

1. Calcolare i quartili (esclusivi) dei dati.
2. Determinare eventuali dati anomali e sospetti.
3. Disegnare il box-plot dei dati.
4. Calcolare il 35-esimo percentile (esclusivo).

Soluzione. 1. Per prima cosa ordiniamo i dati:

2.22 2.29 2.30 2.31 2.32 2.33 2.35 2.36 2.37 2.37 .

Abbiamo $n = 10$ dati.

Calcoliamo il primo quartile: $\frac{n+1}{4} = \frac{11}{4} = 2 + 0.75$. Allora

$$Q_1 = (1 - 0.75)x_2 + 0.75x_3 = 0.25 \cdot 2.29 + 0.75 \cdot 2.30 = 2.2975 .$$

Calcoliamo il secondo quartile: $(n+1)\frac{2}{4} = 11\frac{2}{4} = 5 + 0.5$. Allora

$$Q_2 = (1 - 0.5)x_5 + 0.5x_6 = 0.5 \cdot 2.32 + 0.5 \cdot 2.33 = 2.325 .$$

Calcoliamo il terzo quartile: $(n+1)\frac{3}{4} = 11\frac{3}{4} = 8 + 0.25$. Allora

$$Q_3 = (1 - 0.25)x_8 + 0.25x_9 = 0.75 \cdot 2.36 + 0.25 \cdot 2.37 = 2.3625 .$$

2. Per determinare i dati anomali e sospetti calcoliamo il range interquartile

$$IQR = Q_3 - Q_1 = 2.3625 - 2.2975 = 0.065 .$$

I dati anomali apparterrebbero agli intervalli

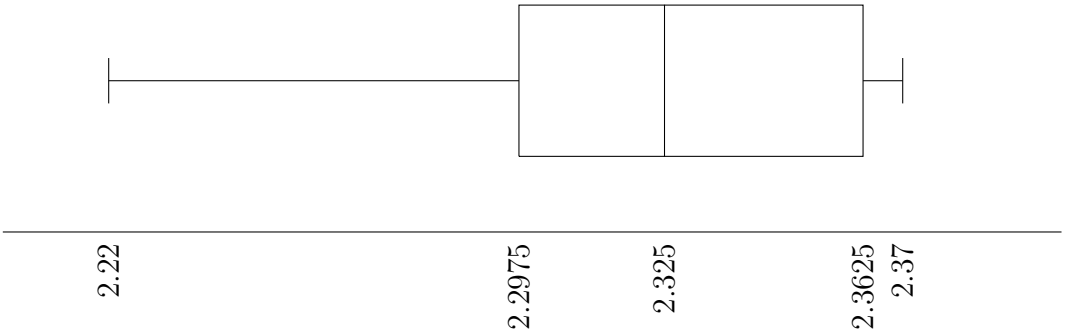
$$(-\infty, Q_1 - 3 \cdot IQR] \cup [Q_3 + 3 \cdot IQR, +\infty) = (-\infty, 2.1025] \cup [2.5575, +\infty) ,$$

quindi non ci sono dati anomali. I dati sospetti appartengono agli intervalli

$$(Q_1 - 3 \cdot IQR, Q_1 - 1.5 \cdot IQR] \cup [Q_3 + 1.5 \cdot IQR, Q_3 + 3 \cdot IQR) = (2.1025, 2.2] \cup [2.46, 2.5575)\,,$$

quindi non ci sono dati sospetti.

3. Segue il box-plot.



4. Calcoliamo il 35-esimo percentile. Osserviamo che $(n+1)35\% = 11 \cdot 35\% = 3.85$. Allora il 35-esimo percentile è

$$P_{35} = (1 - 0.85)x_3 + 0.85x_4 = 2.3085\,.$$

Esercizio 2. (8 punti) Un piccolo bar in una località turistica pugliese vende pasticciotti. Il bar apre alle 08:00 e sforna ogni ora 4 pasticciotti. Si vuole capire se questa produzione è sufficiente a soddisfare la richiesta dei clienti: il numero di pasticciotti ordinati in un'ora è distribuito con una legge di Poisson con media 3. Si assuma che i numeri di ordini in ore diverse siano indipendenti. (Attenzione: nelle domande seguenti tenere conto del fatto che la richiesta in una fascia oraria può essere maggiore dei pasticciotti disponibili, lasciando alcuni ordini insoddisfatti!)

1. Mostrare che la probabilità che dalle 08:00 alle 09:00 vengano ordinati esattamente 2 pasticciotti è uguale alla probabilità che vengano ordinati 3 pasticciotti.

2. Qual è la probabilità che la richiesta di pasticciotti dalle 08:00 alle 09:00 sia soddisfatta?

3. Calcolare la probabilità che il numero di pasticciotti invenduti dalle 08:00 alle 09:00 sia uguale a k per $k = 0, 1, 2, 3, 4$.

4. Se nella fascia oraria dalle 08:00 alle 09:00 restano dei pasticciotti invenduti, questi vengono offerti nella fascia oraria successiva dalle 09:00–10:00, in aggiunta a quelli sfornati alle 09:00. Qual è la probabilità che la richiesta di pasticciotti della fascia oraria 09:00–10:00 sia soddisfatta? (Suggerimento: Sfruttare i risultati del punto 3.)

Soluzione. 1. Definiamo la variabile aleatoria

$$X = \text{“numero di pasticciotti ordinati dalle 08:00 alle 09:00”} \sim \mathrm{P}(\lambda).$$

Poiché $\mathbb{E}(X) = \lambda$, otteniamo che $\lambda = 3$. Quindi

$$\mathbb{P}(\{X = 2\}) = \frac{e^{-3}3^2}{2!} \simeq 22.40\%\,,$$

$$\mathbb{P}(\{X=3\}) = \frac{e^{-3}3^3}{3!} = \frac{e^{-3}3^2}{2!} \simeq 22.40\%\,.$$

2. La richiesta di pasticciotti dalle 08:00 alle 09:00 è soddisfatta se il numero di pasticciotti ordinati è minore o uguale a 4 (ovvero i pasticciotti sfornati). Quindi

$$\begin{aligned}\mathbb{P}(\{X\leq 4\}) &= \mathbb{P}(\{X=0\})+\mathbb{P}(\{X=1\})+\mathbb{P}(\{X=2\})+\mathbb{P}(\{X=3\})+\mathbb{P}(\{X=4\})\\ &= \frac{e^{-3}3^0}{0!}+\frac{e^{-3}3^1}{1!}+\frac{e^{-3}3^2}{2!}+\frac{e^{-3}3^3}{3!}+\frac{e^{-3}3^4}{4!}\\ &\simeq 81.53\%\,.\end{aligned}$$

3. Definiamo la variabile aleatoria

$$Y = \text{“numero di pasticciotti invenduti dalle 08:00 alle 09:00”}$$

e calcoliamo la probabilità richiesta utilizzando la variabile aleatoria X:

$$\begin{aligned}\mathbb{P}(\{Y=0\}) &= \mathbb{P}(\{X\geq 4\}) = 1-\mathbb{P}(\{X\leq 3\})\\ &= 1-\Big(\mathbb{P}(\{X=0\})+\mathbb{P}(\{X=1\})+\mathbb{P}(\{X=2\})+\mathbb{P}(\{X=3\})\Big)\\ &= 1-\Big(\frac{e^{-3}3^0}{0!}+\frac{e^{-3}3^1}{1!}+\frac{e^{-3}3^2}{2!}+\frac{e^{-3}3^3}{3!}\Big)\\ &\simeq 35.28\%\,.\end{aligned}$$

Per le altre probabilità, si ha semplicemente che (utilizzando anche i risultati del punto 1.)

$$\begin{aligned}\mathbb{P}(\{Y=1\}) = \mathbb{P}(\{X=3\}) &= \frac{e^{-3}3^3}{3!} \simeq 22.40\%\,,\\ \mathbb{P}(\{Y=2\}) = \mathbb{P}(\{X=2\}) &= \frac{e^{-3}3^2}{2!} \simeq 22.40\%\,,\\ \mathbb{P}(\{Y=3\}) = \mathbb{P}(\{X=1\}) &= \frac{e^{-3}3^1}{1!} \simeq 14.94\%\,,\\ \mathbb{P}(\{Y=4\}) = \mathbb{P}(\{X=0\}) &= \frac{e^{-3}3^0}{0!} \simeq 4.98\%\,.\end{aligned}$$

4. Definiamo la variabile aleatoria

$$Z = \text{“numero di pasticciotti ordinati dalle 09:00 alle 10:00”} \sim \mathrm{P}(3)\,.$$

Utilizziamo il teorema della probabilità totale, considerando i vari possibili casi di pasticciotti invenduti dalle 08:00 alle 09:00 e utilizzando l'indipendenza tra Z e Y (conseguenza dell'indipendenza tra Z e X):

$$\begin{aligned}&\mathbb{P}(\text{“richiesta delle 09:00–10:00 soddisfatta”}) =\\ &= \mathbb{P}(\{Z\leq 4\}|\{Y=0\})\mathbb{P}(\{Y=0\})+\mathbb{P}(\{Z\leq 5\}|\{Y=1\})\mathbb{P}(\{Y=1\})\\ &\quad+\mathbb{P}(\{Z\leq 6\}|\{Y=2\})\mathbb{P}(\{Y=2\})+\mathbb{P}(\{Z\leq 7\}|\{Y=3\})\mathbb{P}(\{Y=3\})\\ &\quad+\mathbb{P}(\{Z\leq 8\}|\{Y=4\})\mathbb{P}(\{Y=4\})\\ &= \mathbb{P}(\{Z\leq 4\})\mathbb{P}(\{Y=0\})+\mathbb{P}(\{Z\leq 5\})\mathbb{P}(\{Y=1\})\\ &\quad+\mathbb{P}(\{Z\leq 6\})\mathbb{P}(\{Y=2\})+\mathbb{P}(\{Z\leq 7\})\mathbb{P}(\{Y=3\})\\ &\quad+\mathbb{P}(\{Z\leq 8\})\mathbb{P}(\{Y=4\})\,.\end{aligned}$$

Poiché Z e X sono identicamente distribuite, si ha che

$$\mathbb{P}(\{Z\leq 4\}) = \mathbb{P}(\{X\leq 4\}) = 81.53\%\,.$$

Quindi resta solo da calcolare

$$\mathbb{P}(\{Z \leq 5\}) = \mathbb{P}(\{Z \leq 4\}) + \mathbb{P}(\{Z = 5\}) = 81.53\% + \frac{e^{-3}3^5}{5!} \simeq 86.51\%\,,$$

$$\mathbb{P}(\{Z \leq 6\}) = \mathbb{P}(\{Z \leq 5\}) + \mathbb{P}(\{Z = 6\}) = 86.51\% + \frac{e^{-3}3^6}{6!} \simeq 91.55\%\,,$$

$$\mathbb{P}(\{Z \leq 7\}) = \mathbb{P}(\{Z \leq 6\}) + \mathbb{P}(\{Z = 7\}) = 91.55\% + \frac{e^{-3}3^7}{7!} \simeq 93.71\%\,,$$

$$\mathbb{P}(\{Z \leq 8\}) = \mathbb{P}(\{Z \leq 7\}) + \mathbb{P}(\{Z = 8\}) = 93.71\% + \frac{e^{-3}3^8}{8!} \simeq 94.52\%\,.$$

Concludiamo che

$$\begin{aligned}&\mathbb{P}(\text{“richiesta delle 09:00–10:00 soddisfatta”})\\ &\simeq 81.53\% \cdot 35.28\% + 86.51\% \cdot 22.40\% + 91.55\% \cdot 22.40\% + 93.71\% \cdot 14.94\% + 94.52\% \cdot 4.98\%\\ &= 87.36\%\,.\end{aligned}$$

Esercizio 3. (8 punti) Un chiosco in uno stabilimento balneare vende gelati. Il tempo che impiega un cliente a scegliere il gelato è distribuito con legge uniforme tra 10 secondi e 60 secondi.

1. Qual è la probabilità che un cliente scelga il gelato in meno di 30 secondi?
2. Quali sono la media e la deviazione standard del tempo di scelta del gelato?
3. Arrivano due persone in coppia al chiosco. Iniziano a scegliere insieme il gelato indipendentemente. Il tempo in cui la loro ordinazione termina è il massimo tra i due tempi di scelta. Qual è la probabilità che l'ordinazione della coppia termini in più di 30 secondi?
4. Arrivano 10 clienti che scelgono i gelati indipendentemente. Qual è la probabilità che (strettamente) più di 4 di essi impieghino più di 30 secondi per scegliere il gelato?
5. In un momento della giornata vengono serviti 40 clienti indipendenti in sequenza (un cliente inizia a scegliere il gelato solo dopo che il cliente precedente ha terminato). Utilizzare il Teorema del Limite Centrale per stimare la probabilità che in totale il chiosco sia impegnato a servire i 40 clienti per più di 25 minuti.

Soluzione. 1. Consideriamo

$$X = \text{“tempo di scelta del gelato”} \sim \mathrm{U}(10, 60)\,.$$

Allora

$$\mathbb{P}(\{X < 30\}) = \int_{10}^{30} \frac{1}{60 - 10}\,\mathrm{d}x = \frac{20}{50} = 40\%\,.$$

2. Poiché $X \sim \mathrm{U}(a, b)$, si ha che $\mathbb{E}(X) = \frac{a+b}{2}$ e $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$. Quindi

$$\mu = \mathbb{E}(X) = \frac{10 + 60}{2} = 35\,,$$

$$\mathrm{Var}(X) = \frac{(60 - 10)^2}{12} = 208.33 \implies \sigma = \sqrt{\mathrm{Var}(X)} \simeq 14.43\,.$$

3. Consideriamo le variabili aleatorie indipendenti X_1 e X_2 che rappresentano i tempi di scelta del gelato delle due persone, con $X_1, X_2 \sim \mathrm{U}(10, 60)$. Allora

$$\begin{aligned}\mathbb{P}(\{\max\{X_1, X_2\} > 30\}) &= 1 - \mathbb{P}(\{\max\{X_1, X_2\} \leq 30\}) = 1 - \mathbb{P}(\{X_1 \leq 30\} \cap \{X_2 \leq 30\}) \\ &= 1 - \mathbb{P}(\{X_1 \leq 30\})\mathbb{P}(\{X_2 \leq 30\}) = 1 - 0.4^2 = 84\%\,.\end{aligned}$$

4. Definiamo la variabile aleatoria Y che rappresenta il numero di clienti che impiegano più di 30 secondi per scegliere il gelato. Allora $Y \sim \mathrm{B}(n, p)$ dove $n = 10$ è il numero di tentativi e

$$p = \mathbb{P}(\{X > 30\}) = 1 - \mathbb{P}(\{X \leq 30\}) = 1 - 0.4 = 0.6\,.$$

Allora

$$\begin{aligned}\mathbb{P}(\{Y > 4\}) &= 1 - \mathbb{P}(\{Y = 0\}) - \mathbb{P}(\{Y = 1\}) - \mathbb{P}(\{Y = 2\}) - \mathbb{P}(\{Y = 3\}) - \mathbb{P}(\{Y = 4\}) \\ &= 1 - \binom{10}{0}0.6^0(1-0.6)^{10} - \binom{10}{1}0.6^1(1-0.6)^9 \\ &\quad - \binom{10}{2}0.6^2(1-0.6)^8 - \binom{10}{3}0.6^3(1-0.6)^7 - \binom{10}{4}0.6^4(1-0.6)^6 \\ &\simeq 83.38\%\,.\end{aligned}$$

5. Consideriamo le variabili aleatorie $X_1, \ldots, X_n \sim \mathrm{U}(10, 60)$ con $n = 40$. Il tempo totale impegnato a servire i clienti è $X_1 + \cdots + X_n$. Poiché n è grande ($n \geq 30$), grazie al Teorema del Limite Centrale possiamo approssimare la media campionaria standardizzata $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ con una variabile aleatoria normale standard $Z \sim \mathcal{N}(0, 1)$. Utilizzando i risultati del punto 2., otteniamo quindi che (25 minuti sono 1500 secondi)

$$\begin{aligned}\mathbb{P}(\{X_1 + \cdots + X_n > 1500\}) = \mathbb{P}(\{\overline{X}_n > 37.5\}) &= \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} > \frac{37.5 - \mu}{\sigma/\sqrt{n}}\Big\}\Big) \\ &\simeq \mathbb{P}\Big(\Big\{Z > \frac{45 - \mu}{\sigma/\sqrt{n}}\Big\}\Big) = \mathbb{P}\Big(\Big\{Z > \frac{37.5 - 35}{14.43/\sqrt{40}}\Big\}\Big) \\ &\simeq \mathbb{P}(\{Z > 1.095\}) = 86.32\%\,.\end{aligned}$$

dove abbiamo utilizzato le tavole per calcolare l'ultima probabilità.

Esercizio 4. (7 punti) Si vuole stimare la variabilità della temperatura in una località estiva di montagna. Vengono misurate le seguenti temperature (in gradi Celsius) in momenti diversi di una giornata:

23 27 30 31 26 23 22 21

Per risolvere l'esercizio, si assuma che la temperatura abbia distribuzione normale.

1. Calcolare sui dati un intervallo di confidenza bilaterale al 90% per la varianza della temperatura. (N.B.: derivare le formule)

2. Un intervallo di confidenza al 93% sarebbe più grande o più piccolo di quello calcolato nel punto precendente?

3. Le misurazioni vengono ripetute per vari giorni consecutivi e per ogni giorno viene calcolato l'intervallo di confidenza al 90% come nel punto 1. Qual è la probabilità che la prima volta in cui la varianza appartiene all'intervallo di confidenza sia il quinto giorno?

Soluzione. 1. Abbiamo a che fare con una popolazione $X \sim \mathcal{N}(\mu, \sigma^2)$ e un campione casuale $X_1, \dots, X_n$ estratto dalla popolazione con $n = 8$. Sia μ che σ^2 non sono note.

Dalla definizione di IC si ha che

$$90\% = \beta = \mathbb{P}(\{U_n \le \sigma^2 \le V_n\})\,.$$

Per stimare σ^2 sfrutteremo lo stimatore varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Poiché $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ sono indipendenti, $Q_{n-1} = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$. Allora

$$\begin{aligned}
\beta = \mathbb{P}(\{U_n \le \sigma^2 \le V_n\}) &= \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{V_n} \le \frac{(n-1)S_n^2}{\sigma^2} \le \frac{(n-1)S_n^2}{U_n}\Big\}\Big) \\
&= \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{V_n} \le Q_{n-1} \le \frac{(n-1)S_n^2}{U_n}\Big\}\Big) \\
&= 1 - \mathbb{P}\Big(\Big\{Q_{n-1} < \frac{(n-1)S_n^2}{V_n}\Big\}\Big) - \mathbb{P}\Big(\Big\{Q_{n-1} > \frac{(n-1)S_n^2}{U_n}\Big\}\Big)\,.
\end{aligned}$$

Segue che

$$\mathbb{P}\Big(\Big\{Q_{n-1} < \frac{(n-1)S_n^2}{V_n}\Big\}\Big) + \mathbb{P}\Big(\Big\{Q_{n-1} > \frac{(n-1)S_n^2}{U_n}\Big\}\Big) = 1 - \beta = \alpha = 10\%\,.$$

Decidiamo di equipartire α:

$$\mathbb{P}\Big(\Big\{Q_{n-1} < \frac{(n-1)S_n^2}{V_n}\Big\}\Big) = \mathbb{P}\Big(\Big\{Q_{n-1} > \frac{(n-1)S_n^2}{U_n}\Big\}\Big) = \frac{\alpha}{2}\,,$$

da cui

$$\mathbb{P}\Big(\Big\{Q_{n-1} > \frac{(n-1)S_n^2}{U_n}\Big\}\Big) = \frac{\alpha}{2}\,, \quad \mathbb{P}\Big(\Big\{Q_{n-1} \ge \frac{(n-1)S_n^2}{V_n}\Big\}\Big) = 1 - \frac{\alpha}{2}\,.$$

Definiamo $\chi^2_{n-1,\alpha/2}$ e $\chi^2_{n-1,1-\alpha/2}$ come i punti tali che

$$\mathbb{P}(\{Q_{n-1} \ge \chi^2_{n-1,\alpha/2}\}) = \frac{\alpha}{2}\,, \quad \mathbb{P}(\{Q_{n-1} \ge \chi^2_{n-1,1-\alpha/2}\}) = 1 - \frac{\alpha}{2}\,.$$

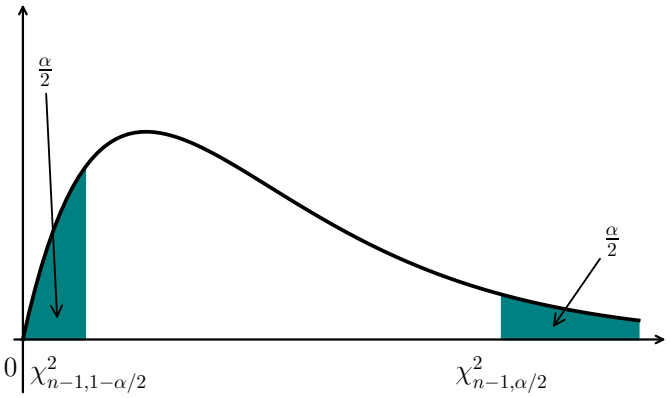


Figura 1: Definizione di $\chi^2_{n-1,\alpha/2}$ e $\chi^2_{n-1,1-\alpha/2}$.

Scegliendo

$$\frac{(n-1)S_n^2}{U_n} = \chi^2_{n-1,\alpha/2} \implies U_n = \frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}}\,,$$

$$\frac{(n-1)S_n^2}{V_n} = \chi^2_{n-1,1-\alpha/2} \implies V_n = \frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}}\,,$$

si ottiene la condizione che definisce l'intervallo di confidenza. In conclusione

$$\Big[\frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}}\Big]$$

è un intervallo di confidenza bilaterale per σ^2 con livello di confidenza $\beta = 1-\alpha$.

Calcoliamo la sua realizzazione sui dati:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{8}(23+27+30+31+26+23+22+21) = 25.375\,,$$

$$\begin{aligned} s_n^2 &= \frac{1}{n-1}\Big(\sum_{i=1}^n x_i^2 - n\overline{x}_n^2\Big) \\ &= \frac{1}{7}\Big(23^2+27^2+30^2+31^2+26^2+23^2+22^2+21^2 - 8\cdot 25.375^2\Big) \\ &\simeq 13.98\,, \end{aligned}$$

e dalle tavole

$$\chi^2_{n-1,\alpha/2} = \chi^2_{7,0.05} \simeq 14.067\,, \quad \chi^2_{n-1,1-\alpha/2} = \chi^2_{7,0.95} \simeq 2.167\,.$$

Quindi l'intervallo di confidenza al 90% per la varianza della temperatura calcolato sui dati è

$$\Big[\frac{7\cdot 13.98}{14.067}, \frac{7\cdot 13.98}{2.167}\Big] \simeq [6.96, 45.16]\,.$$

2. Un intervallo di confidenza al 93% sarebbe più grande di quello calcolato nel punto precedente, poiché il livello di confidenza è maggiore. Infatti, al crescere di β, decresce α, cresce quindi il quantile $\chi^2_{n-1,\alpha/2}$ e decresce $\chi^2_{n-1,1-\alpha/2}$. Poiché questi termini compaiono a denominatore nell'intervallo di confidenza, esso diventa più ampio.

3. Consideriamo la variabile aleatoria

$$Y = \text{“primo giorno in cui la varianza appartiene all'IC al 90\%”} \sim \mathrm{Geo}(p)\,,$$

dove

$$p = 90\%\,.$$

Allora

$$\mathbb{P}(\{Y=5\}) = (1-p)^4 p = (1-0.9)^4 \cdot 0.9 = 0.009\%\,.$$

Esame di Probabilità e Statistica [3231]

Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: settembre 2024 - II
Matricola: _____ Data: 17/09/2024

Esercizio 1. (6 punti) Un'azienda produce un dispositivo elettronico da utilizzare in un intervallo di temperature molto ampio. L'azienda sa che l'aumento della temperatura riduce la durata di vita del dispositivo, per questo motivo viene effettuato uno studio in cui la durata di vita viene determinata in funzione della temperatura. Si trovano i seguenti dati:

Temperatura (°C)	10	20	30	40	50	60	70	80	90
Durata di vita (ore)	420	365	285	220	176	117	69	34	5

1. Rappresentare i dati in uno scatterplot.
2. Determinare e disegnare la retta di regressione lineare. (N.B.: derivare le formule)
3. Calcolare il coefficiente di correlazione lineare e il coefficiente di determinazione.

Soluzione. 1. Segue lo scatterplot (con la retta di regressione lineare):

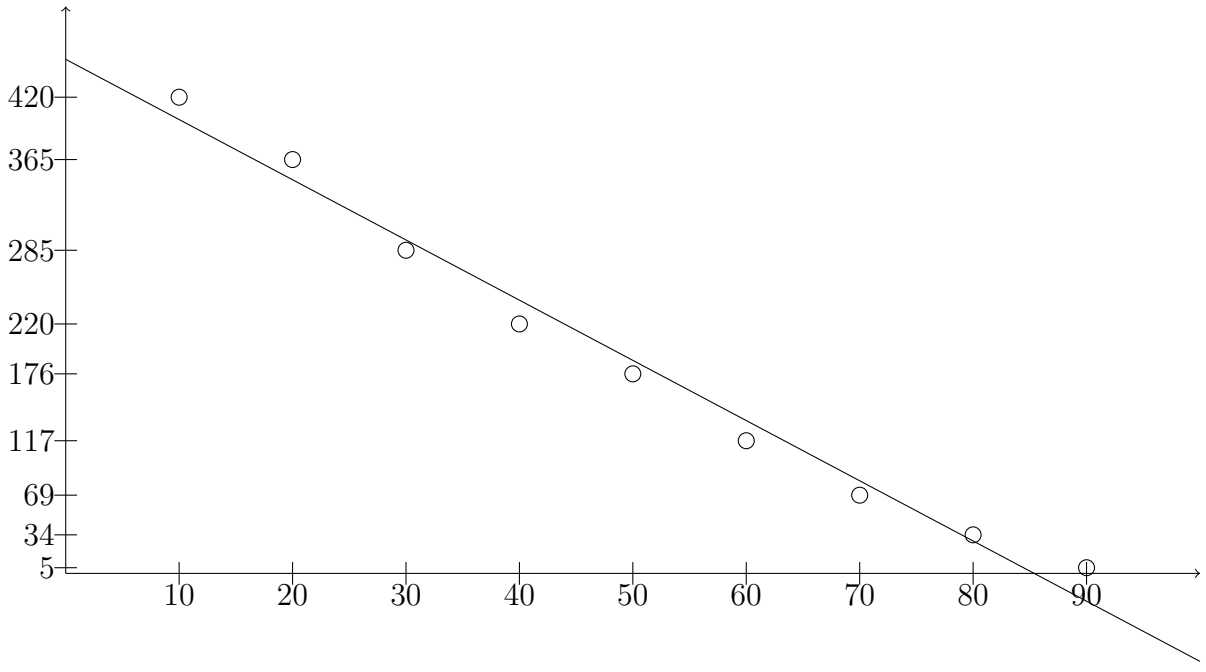


Figura 1: Scatterplot e retta di regressione lineare.

2. Denotiamo con $(x_1, y_1), \ldots, (x_n, y_n)$, $n = 9$, i dati del campione. Cerchiamo la retta di equazione

$$y = ax + b$$

che meglio approssima i dati, utilizzando il metodo dei minimi quadrati. Vogliamo minimizzare l'errore

$$e(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \,.$$

Imponiamo che il gradiente rispetto ad (a, b) sia nullo, ovvero,

$$0 = \partial_a e(a,b) = -2\sum_{i=1}^{n} (y_i - ax_i - b)x_i = -2\sum_{i=1}^{n} (x_i y_i - ax_i^2 - bx_i)\,,$$
$$0 = \partial_b e(a,b) = -2\sum_{i=1}^{n} (y_i - ax_i - b)$$

Dalla seconda equazione segue che

$$nb = \sum_{i=1}^{n} (y_i - ax_i) \implies b = \overline{y} - a\overline{x}\,.$$

Sostituendo nella prima,

$$\begin{aligned}
\sum_{i=1}^{n} (x_i y_i - ax_i^2 - bx_i) = 0 &\implies \sum_{i=1}^{n} (x_i y_i - ax_i^2 - x_i\overline{y} + a\overline{x}x_i) = 0 \\
&\implies a\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \\
&\implies a = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\,.
\end{aligned}$$

Completiamo la tabella con i valori necessari a calcolare a e b:

										somma
x_i	10	20	30	40	50	60	70	80	90	450
y_i	420	365	285	220	176	117	69	34	5	1691
x_i^2	100	400	900	1600	2500	3600	4900	6400	8100	28500
y_i^2	176400	133225	81225	48400	30976	13689	4761	1156	25	489857
$x_i y_i$	4200	7300	8550	8800	8800	7020	4830	2720	450	52670

Pertanto $\overline{x} = 450/9 = 50$ e $\overline{y} = 1691/9 \simeq 187.89$. Segue che

$$a = \frac{52670 - 9 \cdot 50 \cdot 187.89}{28500 - 9 \cdot 50^2} = \frac{52670 - 84550.5}{28500 - 22500} = \frac{-31880.5}{6000} \simeq -5.31\,,$$
$$b = 187.89 + 5.31 \cdot 50 = 453.39\,,$$

ovvero, la retta di regressione lineare ha equazione

$$y = -5.31x + 453.39\,.$$

3. Per calcolare il coefficiente di correlazione lineare usiamo la formula

$$\begin{aligned}
\rho_{x,y} &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}} = \frac{52670 - 9 \cdot 50 \cdot 187.89}{\sqrt{28500 - 9 \cdot 50^2}\sqrt{489857 - 9 \cdot 187.89^2}} \\
&= \frac{-31880.5}{\sqrt{6000}\sqrt{172133.1311}} \simeq -0.9920\,.
\end{aligned}$$

Il coefficiente di determinazione è $R^2 = \rho^2_{x,y} \simeq 0.9841$.

Esercizio 2. (7 punti) Un famoso duo di musicisti inglesi annuncia un tour in UK dopo tanti anni di assenza dalla scena musicale. Per via della grande richiesta, acquistare un biglietto per il concerto è difficile: collegandosi al sito ufficiale, si riesce a caricare la pagina per l'acquisto con probabilità 10%.

1. Una persona si collega al sito. Ogni volta che non riesce a collegarsi al sito, ricarica la pagina, finché finalmente non riesce a collegarsi. Qual è la probabilità che la persona riesca a collegarsi al sito entro il terzo tentativo (terzo incluso)?

2. Nella situazione del punto 1., qual è la media del tentativo in cui la persona riesce a collegarsi al sito per la prima volta? E la varianza?

3. Due persone si mettono d'accordo: si collegano al sito insieme e indipendentemente e continuano a provare (ricaricando la pagina in caso di fallimento) finché almeno una di loro non riesce a collegarsi. Derivare una formula per la probabilità che nessuna delle due persone riesca a collegarsi al sito entro il k-esimo (k incluso) tentativo.

4. Che legge segue la variabile aleatoria che rappresenta il primo tentativo in cui almeno una delle due persone riesce a collegarsi al sito? (Utilizzare il risultato del punto 3.)

Soluzione. 1. Consideriamo la variabile aleatoria

$$X = \text{“prima volta in cui la persona riesce a collegarsi al sito”} \sim \mathrm{Geo}(p)\,,$$

con $p = 10\%$.

La probabilità che la persona riesca a collegarsi al sito entro il terzo tentativo è

$$\mathbb{P}(\{X \leq 3\}) = 1 - \mathbb{P}(\{X > 3\}) = 1 - (1-0.1)^3 = 1 - 0.9^3 = 1 - 0.729 = 27.1\%\,.$$

2. In media, la persona riesce a collegarsi al sito per la prima volta al tentativo

$$\mathbb{E}(X) = \frac{1}{p} = \frac{1}{10\%} = 10\,.$$

La varianza è

$$\mathrm{Var}(X) = \frac{1-p}{p^2} = \frac{0.9}{0.01} = 90\,.$$

3. Consideriamo la variabile aleatoria

$$Y = \text{“prima volta in cui la seconda persona riesce a collegarsi al sito”} \sim \mathrm{Geo}(p)\,,$$

con $p = 10\%$. La probabilità che nessuna delle due persone riesca a collegarsi al sito dopo il k-esimo tentativo è, utilizzando l'indipendenza,

$$\begin{aligned} 1 - \mathbb{P}(\{X \leq k\} \cup \mathbb{P}(\{Y \leq k\})) &= \mathbb{P}(\{X > k\} \cap \{Y > k\}) = \mathbb{P}(\{X > k\} \cap \{Y > k\}) \\ &= \mathbb{P}(\{X > k\})\mathbb{P}(\{Y > k\}) = (1-p)^k(1-p)^k = (1-p)^{2k} \\ &= 0.9^{2k} = 0.81^k\,. \end{aligned}$$

4. La variabile aleatoria che rappresenta il primo tentativo a cui almeno una delle due persone riesce a collegarsi al sito è $\min\{X, Y\}$. Nel punto precedente abbiamo visto che

$$\mathbb{P}(\{\min\{X,Y\} > k\}) = \mathbb{P}(\{X > k\} \cap \{Y > k\}) = 0.81^k = (1-0.19)^k\,,$$

quindi $\min\{X, Y\} \sim \text{Geo}(0.19)$.

Esercizio 3. (8 punti) Un macchinario è costituito da 10 componenti e funziona correttamente se almeno 7 di essi (7 incluso) sono funzionanti. Ogni componente funziona, da quando viene installato, per un tempo distribuito con legge esponenziale con media 2 anni, dopodiché si guasta (e non viene sostituito). Si supponga che i componenti siano indipendenti tra loro.

1. Si consideri un singolo componente. Dal momento in cui viene installato, qual è la probabilità che funzioni correttamente per almeno 1 anno?
2. Si consideri un singolo componente. Dopo un anno viene controllato e risulta ancora funzionante. Sapendo questo fatto, qual è la probabilità che, dal momento in cui è stato installato, funzioni correttamente per almeno 2 anni?
3. Qual è la probabilità che l'intero macchinario funzioni correttamente per almeno 1 anno?
4. Calcolare media e varianza del numero di componenti del macchinario funzionanti dopo 1 anno.
5. Si supponga ora un nuovo scenario. Non appena un componente si guasta, viene sostituito con uno nuovo funzionante (con durata di vita distribuita come sopra). Un componente che si guasta per la seconda volta non viene più sostituito e resta guasto. Qual è la probabilità che un componente funzioni correttamente per almeno 1 anno?

Soluzione.

1. Consideriamo la variabile aleatoria

$$X = \text{“durata di vita di un componente in anni”} \sim \text{Exp}(\lambda)\,.$$

Abbiamo che

$$2 = \mathbb{E}(X) = \frac{1}{\lambda} = 1 \implies \lambda = \frac{1}{2}\,.$$

La probabilità che un componente funzioni correttamente per almeno 1 anno è

$$\mathbb{P}(\{X > 1\}) = e^{-1/2} \simeq 60.65\%\,.$$

2. Utilizziamo la proprietà di assenza di memoria della legge esponenziale per calcolare

$$\mathbb{P}(\{X > 2\}|\{X > 1\}) = \mathbb{P}(\{X > 1\}) = e^{-1/2} \simeq 60.65\%\,.$$

3. Per calcolare la probabilità che l'intero macchinario funzioni correttamente per almeno 1 anno, consideriamo la variabile aleatoria

$$Y = \text{“numero di componenti funzionanti dopo 1 anno”} \sim \text{B}(n, p)\,,$$

dove $n = 10$ e $p = \mathbb{P}(\{X > 1\}) = e^{-1/2}$. La probabilità che l'intero macchinario funzioni correttamente è

$$\begin{aligned}
&\mathbb{P}(\{Y \geq 7\}) = \mathbb{P}(\{Y = 7\}) + \mathbb{P}(\{Y = 8\}) + \mathbb{P}(\{Y = 9\}) + \mathbb{P}(\{Y = 10\}) \\
&= \binom{10}{7} p^7 (1-p)^3 + \binom{10}{8} p^8 (1-p)^2 + \binom{10}{9} p^9 (1-p)^1 + \binom{10}{10} p^{10} (1-p)^0 \\
&\simeq 22.07\% + 12.76\% + 4.37\% + 0.67\% = 39.87\%\,.
\end{aligned}$$

4. La media del numero di componenti funzionanti dopo 1 anno è

$$\mathbb{E}(Y) = np = 10 \cdot e^{-1/2} \simeq 6.065\,.$$

La varianza è

$$\mathrm{Var}(Y) = np(1-p) = 10 \cdot e^{-1/2} \cdot (1 - e^{-1/2}) \simeq 2.39\,.$$

5. Consideriamo le variabili aleatorie

$$X_1 = \text{“durata di vita di un componente nuovo”} \sim \mathrm{Exp}(1/2)\,,$$

$$X_2 = \text{“durata di vita del secondo componente”} \sim \mathrm{Exp}(1/2)\,.$$

Segue che la durata di vita di un componente che viene installato, si rompe, e viene sostituito una volta è $X_1 + X_2 \sim \mathrm{Gamma}(2, 1/2)$ e ha densità

$$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} = \frac{1}{4} x e^{-x/2}\,.$$

La probabilità che un componente funzioni correttamente per almeno 1 anno è, integrando per parti,

$$\begin{aligned}\mathbb{P}(\{X_1 + X_2 > 1\}) &= \int_1^{+\infty} \frac{1}{4} x e^{-x/2}\,\mathrm{d}x = \Big[-\frac{1}{2} x e^{-x/2} \Big]_1^{+\infty} + \int_1^{+\infty} \frac{1}{2} e^{-x/2}\,\mathrm{d}x \\ &= \frac{1}{2} e^{-1/2} + \Big[-e^{-x/2} \Big]_1^{+\infty} = \frac{1}{2} e^{-1/2} + e^{-1/2} = \frac{3}{2} e^{-1/2} \simeq 90.98\%\,.\end{aligned}$$

Esercizio 4. (7 punti) Si studia la variabilità dei prezzi di volo relativi a una certa tratta di una compagnia aerea. Lo scorso anno, la varianza era di 1000 €^{2}. Quest'anno vengono registrati i seguenti prezzi di volo (in €):

$$78 \quad 129 \quad 259 \quad 78 \quad 109 \quad 133 \quad 133 \quad 101 \quad 48 \quad 56\,.$$

Si assuma la distribuzione dei prezzi normale.

1. Si può stabilire con significatività 5% che la varianza di quest'anno sia maggiore di quella dello scorso anno? (N.B.: derivare le formule)

2. Indicare in quale di questi intervalli si colloca il $p-$value del test: $[0\%, 0.5\%]$, $[0.5\%, 1\%]$, $[1\%, 2.5\%]$, $[2.5\%, 5\%]$, $[5\%, 100\%]$.

Soluzione. 1. Abbiamo a che fare con un campione casuale $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma)$ con $n = 10$. Sia μ che σ non sono note.

Il problema è un test di ipotesi

$$H_0 : \sigma^2 = \sigma_0^2\,, \quad H_1 : \sigma^2 > \sigma_0^2\,,$$

con livello di significatività $\alpha = 5\%$. Qui $\sigma_0^2 = 1000$.

Poiché l'ipotesi alternativa è $H_1 : \sigma^2 > \sigma_0^2$, i dati saranno significativi se la varianza è sufficientemente più grande di σ_0^2. La regione critica è allora della forma

$$R_c = \{(x_1, \ldots, x_n) \in R(X_1, \ldots, X_n) \ : s_n^2 > \delta \sigma_0^2\}\,,$$

dove $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x}_n)^2$ e $\overline{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ sono la varianza e la media calcolata sulla realizzazione $x_1, \ldots, x_n$ del campione casuale.

Assumiamo l'ipotesi nulla H_0 vera, cioè $\sigma^2 = \sigma_0^2$. Quindi $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$. Dalla definizione di livello di significatività e utilizzando la varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$,

$$\begin{aligned}\alpha &= \mathbb{P}(\{(X_1, \dots, X_n) \in R_c\}) = \mathbb{P}(\{S_n^2 > \delta\sigma_0^2\}) \\ &= \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{\sigma^2} > (n-1)\delta\Big\}\Big)\,.\end{aligned}$$

Poiché $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ indipendenti, si ha che $Q_{n-1} = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$. Quindi

$$\alpha = \mathbb{P}(\{Q_{n-1} > (n-1)\delta\})\,.$$

Introduciamo il valore $\chi^2_{n-1,\alpha}$ tale che

$$\mathbb{P}(\{Q_{n-1} > \chi^2_{n-1,\alpha}\}) = \alpha\,.$$

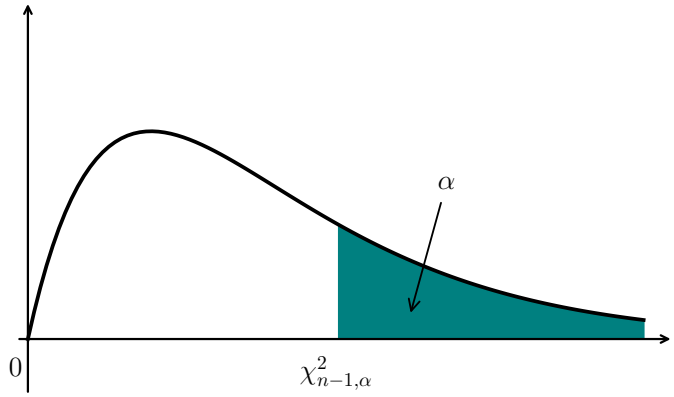


Figura 2: Coda destra della legge chi-quadro con probabilità α.

Allora, scegliendo

$$(n-1)\delta = \chi^2_{n-1,\alpha} \implies \delta = \frac{\chi^2_{n-1,\alpha}}{n-1}\,,$$

otteniamo la condizione che definisce la significatività.

In conclusione, la regione critica è

$$R_c = \Big\{(x_1, \dots, x_n) \in R(X_1, \dots, X_n) \;:\; s_n^2 > \frac{\chi^2_{n-1,\alpha}}{n-1}\sigma_0^2\Big\}\,,$$

e decidiamo come segue:

- Se $s_n^2 > \frac{\chi^2_{n-1,\alpha}}{n-1}\sigma_0^2$, i dati sono sufficientemente significativi da rifiutare H_0. L'ipotesi nulla H_0 viene rifiutata (con livello di significatività α).
- Se $s_n^2 \leq \frac{\chi^2_{n-1,\alpha}}{n-1}\sigma_0^2$, i dati non sono sufficientemente significativi da rifiutare. L'ipotesi nulla H_0 non viene rifiutata (con livello di significatività α).

Calcoliamo sui dati:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{10}(78 + 129 + 259 + 78 + 109 + 133 + 133 + 101 + 48 + 56) = 112.4\,,$$

$$
\begin{aligned}
s_n^2 &= \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}_n^2\Big) \\
&= \frac{1}{9}\Big(78^2 + 129^2 + 259^2 + 78^2 + 109^2 + 133^2 + 133^2 + 101^2 + 48^2 + 56^2 - 10 \cdot 112.4^2\Big) \\
&= \frac{1}{9}(158790 - 126337.6) = 3605.82\,.
\end{aligned}
$$

Dalle tavole

$$\chi^2_{n-1,\alpha} = \chi^2_{9,0.05} = 16.919\,.$$

$$\frac{\chi^2_{n-1,\alpha}}{n-1}\sigma_0^2 = \frac{16.919}{9} \cdot 1000 \simeq 1879.89 < 3605.82 = s_n^2\,.$$

Pertanto l'ipotesi nulla è rifiutata.

2. Il $p-$value è più piccolo di 5%, poiché l'ipotesi nulla è rifiutata con significatività 5%.

Calcoliamo le soglie per il test utilizzando le diverse significatività:

$$\frac{\chi^2_{9,0.25}}{9} \cdot 1000 = \frac{19.023}{9} \cdot 1000 = 2113.67 < s_n^2\,,$$

$$\frac{\chi^2_{9,0.01}}{9} \cdot 1000 = \frac{21.666}{9} \cdot 1000 = 2407.33 < s_n^2\,,$$

$$\frac{\chi^2_{9,0.01}}{9} \cdot 1000 = \frac{23.589}{9} \cdot 1000 = 2621 < s_n^2\,.$$

Segue che il $p-$value è compreso nell'intervallo $[0\%, 0.5\%]$.

Esame di Probabilità e Statistica [3231]

Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: novembre 2024
Matricola: _____ Data: 05/11/2024

Esercizio 1. (6 punti) Vengono registrati i consumi bimestrali di energia elettrica di alcune utenze domestiche. I dati vengono raccolti in classi di consumo, come riportato nella tabella seguente:

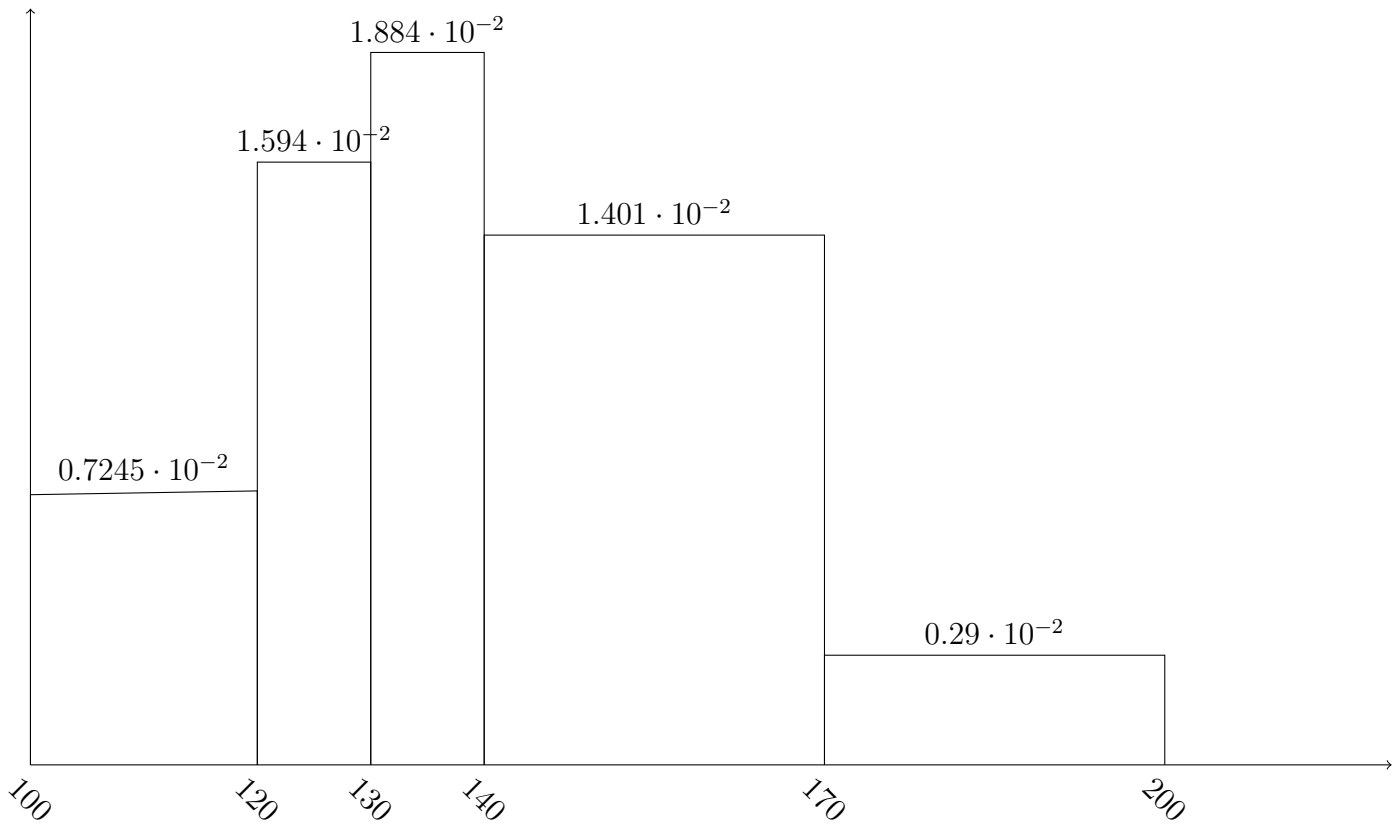
intervallo di consumo (kWh)	frequenza assoluta
$[100, 120)$	10
$[120, 130)$	11
$[130, 140)$	13
$[140, 170)$	29
$[170, 200)$	6

1. Rappresentare un istogramma delle densità di frequenze relative.
2. Determinare la classe modale.
3. Calcolare un'approssimazione della media e della varianza dei consumi.
4. Calcolare un'approssimazione del 70-esimo percentile dei consumi.

Soluzione. 1. Completiamo la tabella:

intervallo	freq. assolute	freq. relative	densità di freq. rel.	freq. cumulate
$[100, 120)$	10	14.49%	$0.7245 \cdot 10^{-2}$	10
$[120, 130)$	11	15.94%	$1.594 \cdot 10^{-2}$	21
$[130, 140)$	13	18.84%	$1.884 \cdot 10^{-2}$	34
$[140, 170)$	29	42.03 %	$1.401 \cdot 10^{-2}$	63
$[170, 200)$	6	8.70%	$0.29 \cdot 10^{-2}$	69

Rappresentiamo le densità di frequenze relative in un istogramma.



2. La classe modale è quella con maggiore densità di frequenza relativa, quindi è l'intervallo $[130, 140)$.

3. Per calcolare un'approssimazione della media utilizziamo le frequenze relative ottenute da $p_j = f_j/n$ dove $n = 69$ e i valori centrali $\tilde{v}_j$ degli intervalli

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \simeq \frac{1}{n}\sum_{j=1}^{k} f_j \tilde{v}_j = \sum_{j=1}^{k} p_j \tilde{v}_j \\ &= 14.49\% \cdot 110 + 15.94\% \cdot 125 + 18.84\% \cdot 135 + 42.03\% \cdot 155 + 8.70\% \cdot 185 \simeq 142.54\,.\end{aligned}$$

Calcoliamo un'approssimazione della varianza

$$\begin{aligned}s^2 &= \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big) \simeq \frac{1}{n-1}\Big(\sum_{j=1}^{n} f_j \tilde{v}_j^2 - n\overline{x}^2\Big) = \frac{n}{n-1}\Big(\sum_{j=1}^{n} p_j \tilde{v}_j^2 - \overline{x}^2\Big) \\ &= \frac{69}{68}\Big(14.49\% \cdot 110^2 + 15.94\% \cdot 125^2 + 18.84\% \cdot 135^2 + 42.03\% \cdot 155^2 + 8.70\% \cdot 185^2 - 142.54^2\Big) \\ &\simeq 441.53\,.\end{aligned}$$

4. Per calcolare un'approssimazione del 70-esimo percentile dei dati, usiamo le frequenze cumulate. Troviamo l'intervallo I_j tale che $F_j \leq 70\%n = 48.3 < F_{j+1}$. Si tratta dell'intervallo $[140, 170)$. Approssimiamo la mediana con

$$Q_2 \simeq a_j + \lambda_j(b_j - a_j)$$

dove

$$\lambda_j = \frac{70\%n - F_{j-1}}{F_j - F_{j-1}} = \frac{48.3 - 34}{63 - 34} \simeq 0.4931$$

Quindi

$$P_{70} \simeq 140 + 0.4931 \cdot (170 - 140) \simeq 154.79\,.$$

Esercizio 2. (7 punti) È periodo di elezioni e si vuole prevedere l'afflusso alle urne in un seggio elettorale. Si assuma che il numero di votanti che si presentano ad un certo seggio nell'arco di 10 minuti sia una variabile aleatoria con distribuzione di Poisson con parametro $\lambda = 3$. Si supponga che il numero di votanti che si presentano in intervalli di tempo diversi siano indipendenti tra loro.

1. Calcolare la probabilità che nell'arco di 10 minuti si presentino almeno 4 votanti (4 incluso).
2. Calcolare la media e la varianza del numero di votanti che si presentano al seggio nell'arco di 10 minuti.
3. Con che legge è distribuito il numero di votanti che si presentano al seggio nell'arco di un'ora? Quali sono la media e la varianza del numero di votanti che si presentano al seggio nell'arco di un'ora?
4. In una giornata il seggio è aperto per 12 ore. Utilizzare il Teorema del Limite Centrale per approssimare la probabilità che nella giornata si presentino al seggio almeno 200 votanti (200 inclusi). (Suggerimento: dividere le 12 ore in intervalli di 10 minuti)

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{numero di votanti che si presentano al seggio nell'arco di 10 minuti} \sim \mathrm{P}(3)\,.$$

1. Ricordando che per la legge di Poisson vale

$$\mathbb{P}(\{X = k\}) = e^{-\lambda}\frac{\lambda^k}{k!}\,,$$

la probabilità che si presentino almeno 4 votanti nell'arco di 10 minuti è

$$\begin{aligned}\mathbb{P}(\{X \geq 4\}) &= 1 - \mathbb{P}(\{X < 4\}) = 1 - \mathbb{P}(\{X = 0\}) - \mathbb{P}(\{X = 1\}) - \mathbb{P}(\{X = 2\}) - \mathbb{P}(\{X = 3\}) \\ &= 1 - e^{-3}\left(1 + 3 + \frac{3^2}{2} + \frac{3^3}{6}\right) \simeq 35.28\%\,.\end{aligned}$$

2. La media e la varianza di X sono rispettivamente

$$\mathbb{E}(X) = \lambda = 3 \quad \text{e} \quad \mathrm{Var}(X) = \lambda = 3\,.$$

3. Considerando le variabili aleatorie

$$X_1, X_2, \ldots, X_6 \sim \mathrm{P}(3) \quad \text{indipendenti}\,,$$

dove X_i è il numero di votanti che si presentano al seggio nell'arco di 10 minuti, la variabile aleatoria

$$Y = X_1 + X_2 + \ldots + X_6 \sim \mathrm{P}(3 \cdot 6) = \mathrm{P}(18)$$

rappresenta il numero di votanti che si presentano al seggio nell'arco di un'ora. La media e la varianza di Y sono rispettivamente

$$\mathbb{E}(Y) = \lambda = 18 \quad \text{e} \quad \mathrm{Var}(Y) = \lambda = 18\,.$$

4. Nell'arco di 12 ore si hanno $12 \cdot 6 = 72$ intervalli di 10 minuti. Consideriamo le variabili aleatorie

$$X_1, X_2, \ldots, X_{72} \sim \mathrm{P}(3) \quad \text{indipendenti}\,,$$

dove X_i è il numero di votanti che si presentano al seggio nell'i-esimo intervallo di 10 minuti. La variabile aleatoria

$$W = X_1 + X_2 + \ldots + X_{72} \sim \mathrm{P}(3 \cdot 72) = \mathrm{P}(216)$$

rappresenta il numero di votanti che si presentano al seggio nell'arco di 12 ore. Calcolare esplicitamente

$$\mathbb{P}(\{W \geq 200\}) = 1 - \sum_{k=0}^{199} e^{-216} \frac{216^k}{k!}$$

è computazionalmente oneroso. Utilizziamo allora il Teorema del Limite Centrale per approssimare la probabilità richiesta. Osserviamo che

$$\overline{X}_{72} = \frac{1}{72}(X_1 + X_2 + \ldots + X_{72}) = \frac{1}{72} W \,,$$

Per il Teorema del Limite Centrale, la variabile aleatoria $\frac{\overline{X}_{72} - \mu}{\sigma/\sqrt{72}}$ converge in legge ad una variabile aleatoria Z distribuita come una normale standard. Ricordiamo che

$$\mu = \mathbb{E}(X_i) = 3 \,, \quad \sigma^2 = \mathrm{Var}(X_i) = 3 \,.$$

Segue che, utilizzando la simmetria della normale standard e le tavole della normale standard,

$$\begin{aligned}
\mathbb{P}(\{W \geq 200\}) = \mathbb{P}(\{W > 199.5\}) = \mathbb{P}(\{\overline{X}_{72} > 2.77\}) &= \mathbb{P}\Big(\Big\{ \frac{\overline{X}_{72} - 3}{\sqrt{3}/\sqrt{72}} > \frac{2.77 - 3}{\sqrt{3}/\sqrt{72}} \Big\}\Big) \\
&\simeq \mathbb{P}\Big(\Big\{ Z > \frac{2.77 - 3}{\sqrt{3}/\sqrt{72}} \Big\}\Big) = \mathbb{P}(\{Z > -1.13\}) = \mathbb{P}(\{Z < 1.13\}) = \Phi(1.13) \\
&\simeq 87.08\% \,.
\end{aligned}$$

Esercizio 3. (8 punti) Due genitori studiano il tempo necessario affinché la loro bambina di 6 mesi si addormenti. Da quando iniziano a cullarla per farla addormentare, la bambina si addormenta in media in 5 minuti. Si supponga che il tempo necessario affinché la bambina si addormenti sia distribuito in modo esponenziale.

1. Calcolare la probabilità che la bambina si addormenti entro i primi 2 minuti da quando iniziano a cullarla.

2. Calcolare la deviazione standard del tempo necessario affinché la bambina si addormenti da quando iniziano a cullarla.

3. Uno dei genitori sta cullando la bambina da 10 minuti, ma lei ancora gli occhi spalancati. Sapendo questo fatto, calcolare la probabilità che, da quando ha iniziato a cullarla, la bambina si addormenti entro 15 minuti.

Nelle prossime richieste si consideri il seguente scenario. Alla bambina sta per spuntare il primo dentino e questo può renderla irritabile e influenzare il tempo necessario affinché si addormenti. Se la bambina è irritata, il tempo necessario affinché si addormenti è distribuito in modo esponenziale con media 10 minuti. Altrimenti, come prima, la media è 5 minuti. Ogni volta che si prova a farla addormentare, la probabilità che la bambina sia irritata è 70%.

4. Si prova a far addormentare la bambina una volta. Calcolare la probabilità che si addormenti entro i primi 12 minuti.

5. Uno dei genitori sta cullando la bambina da 12 minuti, ma lei ancora gli occhi spalancati. Sapendo questo fatto, calcolare la probabilità che la bambina sia irritata per via del dentino che sta spuntando.

6. Nell'arco della giornata si prova a far addormentare la bambina 5 volte. Calcolare la probabilità che la bambina si addormenti almeno 3 volte entro i primi 12 minuti.

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{tempo necessario affinché la bambina si addormenti} \sim \mathrm{Exp}(\lambda)\,.$$

Ricordiamo che $\mathbb{E}(X) = 1/\lambda$, quindi $\lambda = 1/5$. La densità di X è

$$f_X(x) = \lambda e^{-\lambda x} = \frac{1}{5} e^{-x/5} \quad \text{per } x \geq 0\,.$$

1. La probabilità che la bambina si addormenti entro i primi 2 minuti da quando iniziano a cullarla è

$$\mathbb{P}(\{X \leq 2\}) = \int_0^2 \frac{1}{5} e^{-x/5}\,\mathrm{d}x = 1 - e^{-2/5} \simeq 32.96\%\,.$$

2. La deviazione standard di X è

$$\sqrt{\mathrm{Var}(X)} = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda} = 5\,.$$

3. La probabilità che la bambina si addormenti entro 15 minuti da quando iniziano a cullarla, sapendo che è stata cullata per 10 minuti, è, per la proprietà di assenza di memoria della distribuzione esponenziale,

$$\begin{aligned}\mathbb{P}(\{X \leq 15 \,|\, X > 10\}) &= 1 - \mathbb{P}(\{X > 15 \,|\, X > 10\}) = 1 - \mathbb{P}(\{X > 5\}) = 1 - e^{-5/5} \\ &= 1 - e^{-1} \simeq 63.21\%\,.\end{aligned}$$

4. Consideriamo le variabili aleatorie

$$X = \text{tempo necessario affinché la bambina si addormenti se non irritata} \sim \mathrm{Exp}(5)\,,$$

$$Y = \text{tempo necessario affinché la bambina si addormenti se irritata} \sim \mathrm{Exp}(10)\,,$$

$$Z = \text{tempo necessario affinché la bambina si addormenti}\,,$$

$$W = \text{la bambina è irritata} \sim \mathrm{Be}(0.7)\,.$$

La probabilità che la bambina si addormenti entro i primi 12 minuti è, per il teorema della probabilità totale,

$$\begin{aligned}\mathbb{P}(\{Z \leq 12\}) &= \mathbb{P}(\{Z \leq 12\}|\{W = 0\})\mathbb{P}(\{W = 0\}) + \mathbb{P}(\{Z \leq 12\}|\{W = 1\})\mathbb{P}(\{W = 1\}) \\ &= \mathbb{P}(\{X \leq 12\})\mathbb{P}(\{W = 0\}) + \mathbb{P}(\{Y \leq 12\})\mathbb{P}(\{W = 1\}) \\ &= (1 - e^{-12/5}) \cdot 0.3 + (1 - e^{-12/10}) \cdot 0.7 \simeq 76.19\%\,.\end{aligned}$$

5. La probabilità che la bambina sia irritata per via del dentino che sta spuntando, sapendo che è stata cullata per 12 minuti, è, per il teorema di Bayes,

$$\begin{aligned}\mathbb{P}(\{W = 1\}|\{Z > 12\}) &= \frac{\mathbb{P}(\{Z > 12\}|\{W = 1\})\mathbb{P}(\{W = 1\})}{1 - \mathbb{P}(\{Z \leq 12\})} = \frac{\mathbb{P}(\{Y > 12\})\mathbb{P}(\{W = 1\})}{1 - \mathbb{P}(\{Z \leq 12\})} \\ &= \frac{e^{-12/10} \cdot 0.7}{1 - 0.7619} \simeq 88.55\%\,.\end{aligned}$$

6. Consideriamo la variabile aleatoria distribuita con legge binomiale

$S =$ numero di volte che la bambina si addormenta entro i primi 12 minuti $\sim \mathrm{B}(5,p)$,

dove $p = 76.19\%$. Calcoliamo

$$\begin{aligned}\mathbb{P}(\{S \geq 3\}) &= 1 - \mathbb{P}(\{S < 3\}) \\ &= 1 - \left(\binom{5}{0}p^0(1-p)^5 + \binom{5}{1}p^1(1-p)^4 + \binom{5}{2}p^2(1-p)^3\right) \\ &= 90.86\%\,.\end{aligned}$$

Esercizio 4. (7 punti) Vengono registrati i punteggi ottenuti da alcune persone nell'Esercizio 4 di Probabilità e Statistica. I dati sono i seguenti:

4 3 1 4 5 7 7 0 5 5 5 0

Si supponga che i punteggi siano distribuiti in modo normale.

1. Calcolare sui dati un intervallo di confidenza bilaterale al 95% per la varianza dei punteggi. (N.B.: Derivare le formule)
2. Calcolare sui dati un intervallo di confidenza bilaterale al 95% per la media dei punteggi. (Non è obbligatorio derivare le formule)

Soluzione. Abbiamo a che fare con una popolazione $X \sim \mathcal{N}(\mu,\sigma^2)$ e un campione casuale $X_1,\dots,X_n$ estratto dalla popolazione.

1. Dalla definizione di IC si ha che

$$\beta = \mathbb{P}(\{U_n \leq \sigma^2 \leq V_n\})\,.$$

Per stimare σ^2 sfrutteremo lo stimatore varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Poiché $X_1,\dots,X_n \sim \mathcal{N}(\mu,\sigma^2)$ sono indipendenti, $Q_{n-1} = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$. Allora

$$\begin{aligned}\beta = \mathbb{P}(\{U_n \leq \sigma^2 \leq V_n\}) &= \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{V_n} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \frac{(n-1)S_n^2}{U_n}\Big\}\Big) \\ &= \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{V_n} \leq Q_{n-1} \leq \frac{(n-1)S_n^2}{U_n}\Big\}\Big) \\ &= 1 - \mathbb{P}\Big(\Big\{Q_{n-1} < \frac{(n-1)S_n^2}{V_n}\Big\}\Big) - \mathbb{P}\Big(\Big\{Q_{n-1} > \frac{(n-1)S_n^2}{U_n}\Big\}\Big)\,.\end{aligned}$$

Segue che

$$\mathbb{P}\Big(\Big\{Q_{n-1} < \frac{(n-1)S_n^2}{V_n}\Big\}\Big) + \mathbb{P}\Big(\Big\{Q_{n-1} > \frac{(n-1)S_n^2}{U_n}\Big\}\Big) = 1 - \beta = \alpha\,.$$

Decidiamo di equipartire α:

$$\mathbb{P}\Big(\Big\{Q_{n-1} < \frac{(n-1)S_n^2}{V_n}\Big\}\Big) = \mathbb{P}\Big(\Big\{Q_{n-1} > \frac{(n-1)S_n^2}{U_n}\Big\}\Big) = \frac{\alpha}{2}\,,$$

da cui

$$\mathbb{P}\Big(\Big\{Q_{n-1} > \frac{(n-1)S_n^2}{U_n}\Big\}\Big) = \frac{\alpha}{2}\,, \quad \mathbb{P}\Big(\Big\{Q_{n-1} \geq \frac{(n-1)S_n^2}{V_n}\Big\}\Big) = 1 - \frac{\alpha}{2}\,.$$

Definiamo $\chi^2_{n-1,\alpha/2}$ e $\chi^2_{n-1,1-\alpha/2}$ come i punti tali che

$$\mathbb{P}(\{Q_{n-1} \geq \chi^2_{n-1,\alpha/2}\}) = \frac{\alpha}{2}\,, \quad \mathbb{P}(\{Q_{n-1} \geq \chi^2_{n-1,1-\alpha/2}\}) = 1 - \frac{\alpha}{2}\,.$$

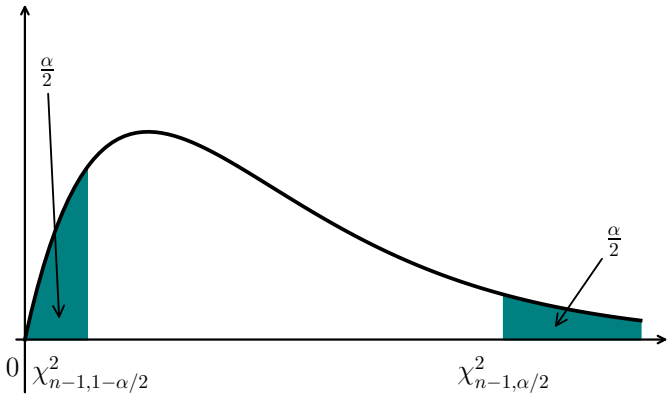


Figura 1: Definizione di $\chi^2_{n-1,\alpha/2}$ e $\chi^2_{n-1,1-\alpha/2}$.

Scegliendo

$$\frac{(n-1)S_n^2}{U_n} = \chi^2_{n-1,\alpha/2} \implies U_n = \frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}}\,,$$
$$\frac{(n-1)S_n^2}{V_n} = \chi^2_{n-1,1-\alpha/2} \implies V_n = \frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}}\,,$$

si ottiene la condizione che definisce l'intervallo di confidenza. In conclusione

$$\Big[\frac{(n-1)S_n^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S_n^2}{\chi^2_{n-1,1-\alpha/2}}\Big]$$

è un intervallo di confidenza bilaterale per σ^2 con livello di confidenza $\beta = 1-\alpha$.

Calcoliamolo sui dati. Per farlo, calcoliamo

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{12}(4+3+1+4+5+7+7+0+5+5+5+0) = 3.83\,,$$

$$s_n^2 = \frac{1}{n-1}\Big(\sum_{i=1}^n x_i^2 - n\overline{x}_n^2\Big) = \frac{1}{11}(4^2+3^2+1^2+4^2+5^2+7^2+7^2+0^2+5^2+5^2+5^2+0^2-12\cdot 3.83^2) = 5.82\,.$$

Dalle tavole della distribuzione χ^2 troviamo

$$\chi^2_{11,0.025} = 21.920\,, \quad \chi^2_{11,0.975} = 3.816\,.$$

Concludiamo che l'intervallo di confidenza bilaterale al 95% per la varianza calcolato sui dati è

$$\Big[\frac{(n-1)s_n^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s_n^2}{\chi^2_{n-1,1-\alpha/2}}\Big] = \Big[\frac{11\cdot 5.82}{21.920}, \frac{11\cdot 5.82}{3.816}\Big] \simeq [2.92, 16.78]\,.$$

2. Un intervallo di confidenza bilaterale al 95% per la media calcolato sui dati è

$$\Big[\overline{x}_n - \frac{s_n}{\sqrt{n}}t_{n-1,\alpha/2}, \overline{x}_n + \frac{s_n}{\sqrt{n}}t_{n-1,\alpha/2}\Big]\,.$$

Dalle tavole ricaviamo che

$$t_{11,0.025} = 2.201$$

e quindi l'intervallo di confidenza è

$$\Big[3.83 - \frac{\sqrt{5.82}}{\sqrt{12}}\cdot 2.201, 3.83 + \frac{\sqrt{5.82}}{\sqrt{12}}\cdot 2.201\Big] \simeq [2.30, 5.36]\,.$$

Esame di Probabilità e Statistica [3231]

Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: _____ Docente: Gianluca Orlando
Nome: _____ Appello: gennaio 2025
Matricola: _____ Data: 16/12/2025

Esercizio 1. (6 punti) Una società vuole studiare gli effetti dell'investimento pubblicitario (in migliaia di euro) sull'incremento percentuale delle vendite annuali. Sono stati raccolti i seguenti dati:

Investimento pubblicitario	10	20	30	40	50	60
Incremento percentuale delle vendite	2	4	5	7	9	10

1. Rappresentare i dati in uno scatterplot.
2. Determinare e disegnare la retta di regressione lineare. (N.B.: derivare le formule).
3. Calcolare il coefficiente di correlazione lineare e il coefficiente di determinazione.

Soluzione. 1. Segue lo scatterplot:

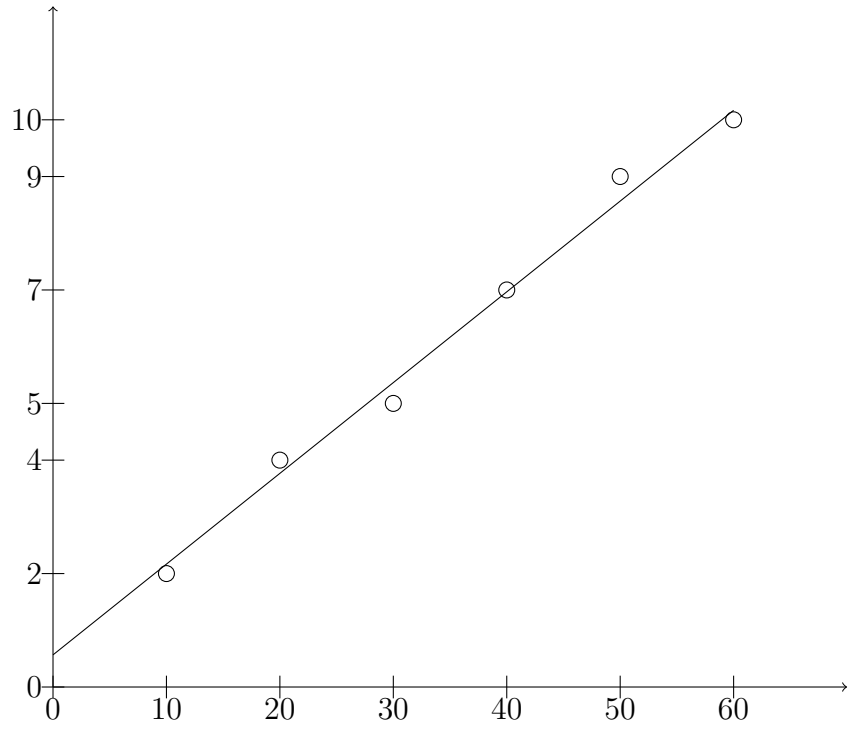


Figura 1: Scatterplot dei dati.

2. Determiniamo la retta di regressione lineare utilizzando il metodo dei minimi quadrati. Siano $(x_1, y_1), \ldots, (x_n, y_n)$ i dati del campione, con $n = 6$. Cerchiamo la retta di equazione

$$y = ax + b$$

che minimizza l'errore

$$e(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Imponendo che il gradiente sia nullo, otteniamo il sistema:

$$0 = \partial_a e(a, b) = -2 \sum_{i=1}^{n} (y_i - ax_i - b)x_i,$$

$$0 = \partial_b e(a, b) = -2 \sum_{i=1}^{n} (y_i - ax_i - b).$$

Dalla seconda equazione si ricava

$$b = \overline{y} - a\overline{x},$$

dove $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ e $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$. Sostituendo nella prima equazione:

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}.$$

Completiamo la tabella dei calcoli:

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	10	2	100	4	20
	20	4	400	16	80
	30	5	900	25	150
	40	7	1600	49	280
	50	9	2500	81	450
	60	10	3600	100	600
somma	**210**	**37**	**9100**	**275**	**1580**

Calcoliamo $\overline{x} = 35$, $\overline{y} = 6.1667$, e quindi:

$$a = \frac{1580 - 6 \cdot 35 \cdot 6.1667}{9100 - 6 \cdot 35^2} \approx 0.16, \quad b = 6.1667 - 0.16 \cdot 35 \approx 0.5667.$$

La retta di regressione è quindi

$$y = 0.16x + 0.5667.$$

3. Il coefficiente di correlazione lineare è dato da

$$\rho_{x,y} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}.$$

Sostituendo i valori:

$$\rho_{x,y} = \frac{1580 - 6 \cdot 35 \cdot 6.1667}{\sqrt{9100 - 6 \cdot 35^2} \sqrt{275 - 6 \cdot 6.1667^2}} \approx 0.9955.$$

Il coefficiente di determinazione è $R^2 = \rho_{x,y}^2 \approx 0.9910$.

Esercizio 2. (7 punti) Sia (X, Y) un vettore aleatorio con funzione di probabilità congiunta data dalla seguente tabella:

Y \ X	0	1
0	a	b
1	c	d

1. Fornire valori di a, b, c, d tali che X e Y siano indipendenti (verificando l'indipendenza).
2. Fornire valori di a, b, c, d tali che $\mathrm{Cov}(X, Y) = \frac{1}{4}$.
3. Fornire valori di a, b, c, d tali che X e Y non siano indipendenti.
4. Fornire valori di a, b, c, d tali che $\mathbb{P}(\{X = 1\}|\{Y = 1\}) = \frac{1}{3}$.
5. X viene generata in sequenza tante volte fino a che non si verifica $\{X = 1\}$. Fornire valori di a, b, c, d tali che, in media, la prima volta che si ottiene 1 sia la quarta.

Soluzione. Le soluzioni ai quesiti non hanno sempre una risposta unica. Viene spiegato il ragionamento per ciascun quesito e poi fatta una scelta di valori.

1. L'indipendenza di X e Y prevede che $\mathbb{P}(\{X = i\} \cap \{Y = j\}) = \mathbb{P}(\{X = i\})\mathbb{P}(\{Y = j\})$ per ogni i, j. Questo vuol dire che

$$a = \mathbb{P}(\{X = 0\} \cap \{Y = 0\}) = \mathbb{P}(\{X = 0\})\mathbb{P}(\{Y = 0\}) = (a + c)(a + b),$$

$$b = \mathbb{P}(\{X = 1\} \cap \{Y = 0\}) = \mathbb{P}(\{X = 1\})\mathbb{P}(\{Y = 0\}) = (b + d)(a + b),$$

$$c = \mathbb{P}(\{X = 0\} \cap \{Y = 1\}) = \mathbb{P}(\{X = 0\})\mathbb{P}(\{Y = 1\}) = (a + c)(c + d),$$

$$d = \mathbb{P}(\{X = 1\} \cap \{Y = 1\}) = \mathbb{P}(\{X = 1\})\mathbb{P}(\{Y = 1\}) = (b + d)(c + d).$$

Queste equazioni sono soddisfatte scegliendo, ad esempio, tutti i valori uguali $a = b = c = d = \frac{1}{4}$. Un'altra possibilità è cercare se ci sono valori ammissibili per $a = 0$. Si ottiene

$$0 = cb \implies c = 0 \quad \text{oppure} \quad b = 0.$$

Scegliamo $b = 0$. Allora

$$c = c(c + d) \implies 1 = c + d\,.$$

Si può scegliere, ad esempio, $c = \frac{1}{3}$ e $d = \frac{2}{3}$. Infatti, se

Y \ X	0	1
0	0	0
1	c	$1 - c$

le due variabili sono indipendenti, poiché

$$\mathbb{P}(\{X = 0\} \cap \{Y = 0\}) = 0 = \mathbb{P}(\{X = 0\})\mathbb{P}(\{Y = 0\}) = c \cdot 0,$$

$$\mathbb{P}(\{X = 1\} \cap \{Y = 0\}) = 0 = \mathbb{P}(\{X = 1\})\mathbb{P}(\{Y = 0\}) = (1 - c) \cdot 0,$$

$$\mathbb{P}(\{X = 0\} \cap \{Y = 1\}) = c = \mathbb{P}(\{X = 0\})\mathbb{P}(\{Y = 1\}) = c \cdot 1,$$

$$\mathbb{P}(\{X = 1\} \cap \{Y = 1\}) = 1 - c = \mathbb{P}(\{X = 1\})\mathbb{P}(\{Y = 1\}) = (1 - c) \cdot 1.$$

2. La covarianza è data da

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Calcoliamo i valori attesi:

$$\mathbb{E}(X) = b + d, \quad \mathbb{E}(Y) = c + d, \quad \mathbb{E}(XY) = d.$$

Sostituendo:

$$\mathrm{Cov}(X, Y) = d - (b + d)(c + d) = \frac{1}{4}.$$

Scegliamo $d = \frac{1}{2}$, $b = 0$, $c = 0$ e l'equazione è verificata. Resta da scegliere a, che si ottiene imponendo che la somma delle probabilità sia 1, cioè $a = \frac{1}{2}$. La tabella diventa

Y \ X	0	1
0	$\frac{1}{2}$	0
1	0	$\frac{1}{2}$

3. Poiché $\mathrm{Cov}(X, Y) = 0$ è una condizione necessaria per l'indipendenza, è sufficiente scegliere i valori del punto 2. per ottenere X e Y dipendenti.

4. La probabilità condizionata è data da

$$\mathbb{P}(\{X = 1\}|\{Y = 1\}) = \frac{\mathbb{P}(\{X = 1\} \cap \{Y = 1\})}{\mathbb{P}(\{Y = 1\})} = \frac{d}{c + d}.$$

Imponiamo

$$\frac{d}{c + d} = \frac{1}{3} \implies d = \frac{1}{3}c + \frac{1}{3}d \implies \frac{2}{3}d = \frac{1}{3}c \implies d = \frac{1}{2}c.$$

è sufficiente scegliere, ad esempio, $c = \frac{1}{2}$ e $d = \frac{1}{4}$. Occorre scegliere a e b in modo che la somma delle probabilità sia 1. Ad esempio, $a = \frac{1}{4}$ e $b = 0$. La tabella diventa

Y \ X	0	1
0	$\frac{1}{4}$	0
1	$\frac{1}{2}$	$\frac{1}{4}$

5. Consideriamo la variabile aleatoria

$$Z = \text{“prima volta nella sequenza in cui si verifica } \{X = 1\}\text{”} \sim \mathrm{Geo}(p),$$

dove $p = \mathbb{P}(\{X = 1\})$ è la probabilità del successo. Stiamo cercando p in modo che

$$\mathbb{E}(Z) = 4 \implies \frac{1}{p} = 4 \implies p = \frac{1}{4}.$$

Quindi

$$\frac{1}{4} = \mathbb{P}(\{X = 1\}) = b + d\,.$$

È sufficiente scegliere, ad esempio, $b = 0$ e $d = \frac{1}{4}$. Occorre poi scegliere a e c in modo che la somma delle probabilità sia 1. Ad esempio, $a = \frac{3}{4}$ e $c = 0$. La tabella diventa

Y \ X	0	1
0	$\frac{3}{4}$	0
1	0	$\frac{1}{4}$

Esercizio 3. (8 punti) Per l'implementazione di un progetto, occorre completare alcuni *task*. Ogni *task* viene completato in un tempo distribuito con legge esponenziale con valore atteso di 2 giorni. Si assumano i tempi di completamento dei *task* indipendenti tra loro.

1. Qual è la probabilità che un *task* venga completato in meno di 3 giorni?
2. Calcolare la varianza del tempo impiegato per completare un *task*.
3. Due *task* $T1.1$ e $T1.2$ partono contemporaneamente e devono essere completati entrambi per poter avviare la seconda fase del progetto. Qual è la probabilità che dopo 4 giorni non si possa ancora avviare la seconda fase?
4. Dire se è vero che il massimo tra i tempi impiegati per completare i due *task* $T1.1$ e $T1.2$ gode di assenza di memoria, motivando la risposta.
5. La seconda fase è composta da due *task* $T2.1$ e $T2.2$ che devono essere completati in serie, ovvero $T2.2$ può essere avviato solo dopo il completamento di $T2.1$. Qual è la probabilità che il tempo impiegato per completare la seconda fase sia maggiore di 4 giorni?

Soluzione. 1. Consideriamo la variabile aleatoria

$$X = \text{“tempo impiegato per completare un } \textit{task}\text{”} \sim \text{Exp}(\lambda),$$

Sappiamo che

$$\mathbb{E}(X) = \frac{1}{\lambda} = 2 \implies \lambda = \frac{1}{2}.$$

Possiamo allora calcolare

$$\mathbb{P}(\{X < 3\}) = \int_0^3 e^{-\lambda t}\,\mathrm{d}t = 1 - e^{-3/2} \approx 77.69\%.$$

2. La varianza di X è data da

$$\text{Var}(X) = \frac{1}{\lambda^2} = \frac{1}{\left(\frac{1}{2}\right)^2} = 4.$$

3. Consideriamo le variabili aleatorie

$$T1.1 = \text{“tempo impiegato per completare il } \textit{task}\ T1.1\text{”} \sim \text{Exp}(\lambda),$$

$$T1.2 = \text{“tempo impiegato per completare il } \textit{task}\ T1.2\text{”} \sim \text{Exp}(\lambda),$$

$$T1 = \max\{T1.1, T1.2\} = \text{“tempo massimo tra i due } \textit{task}\text{”}\ .$$

Dire che dopo 4 giorni non si possa ancora avviare la seconda fase è equivalente a dire che $\{T1 > 4\}$. Calcoliamo quindi

$$\begin{aligned}\mathbb{P}(\{T1 > 4\}) &= \mathbb{P}(\{\max\{T1.1, T1.2\} > 4\}) = \mathbb{P}(\{T1.1 > 4\} \cup \{T1.2 > 4\}) \\ &= \mathbb{P}(\{T1.1 > 4\}) + \mathbb{P}(\{T1.2 > 4\}) - \mathbb{P}(\{T1.1 > 4\} \cap \{T1.2 > 4\}) \\ &= e^{-2} + e^{-2} - e^{-4} = 2e^{-2} - e^{-4} \approx 25.24\%.\end{aligned}$$

4. La proprietà di assenza di memoria è soddisfatta se

$$\mathbb{P}(\{T1 > t + s\} | \{T1 > s\}) = \mathbb{P}(\{T1 > t\})$$

per ogni t, s, ovvero,

$$\mathbb{P}(\{T1 > t\}) = \frac{\mathbb{P}(\{T1 > t+s\} \cap \{T1 > s\})}{\mathbb{P}(\{T1 > s\})} = \frac{\mathbb{P}(\{T1 > t+s\})}{\mathbb{P}(\{T1 > s\})}\,.$$

Osserviamo che

$$\begin{aligned}\mathbb{P}(\{T1 > t+s\}) &= \mathbb{P}(\{\max\{T1.1, T1.2\} > t+s\}) = \mathbb{P}(\{T1.1 > t+s\} \cup \{T1.2 > t+s\}) \\ &= 2e^{-2(t+s)} - e^{-4(t+s)}\,,\end{aligned}$$

e

$$\begin{aligned}\mathbb{P}(\{T1 > s\}) &= 2e^{-2s} - e^{-4s}\,, \\ \mathbb{P}(\{T1 > t\}) &= 2e^{-2t} - e^{-4t}\,.\end{aligned}$$

Scegliamo $t = 2$ e $s = 2$. Allora

$$\mathbb{P}(\{T1 > t+s\}) = \mathbb{P}(\{T1 > 4\}) = 25.24\%,$$

$$\mathbb{P}(\{T1 > t\}) = \mathbb{P}(\{T1 > s\}) = \mathbb{P}(\{T1 > 2\}) = e^{-1} \approx 36.79\%.$$

Poiché

$$\frac{\mathbb{P}(\{T1 > 4\})}{\mathbb{P}(\{T1 > 2\})} = \frac{25.24\%}{36.79\%} \approx 68.60\% \neq 36.79\% \approx \mathbb{P}(\{T1 > 2\}),$$

concludiamo che il massimo tra i tempi impiegati per completare i due *task* $T1.1$ e $T1.2$ non gode di assenza di memoria.

5. Il tempo impiegato per completare la seconda fase è dato dalla somma dei tempi impiegati per completare i due *task*, $T2 = T2.1 + T2.2$. Pertanto $T2 \sim \text{Gamma}(2, 1/2)$, e la probabilità richiesta è

$$\begin{aligned}\mathbb{P}(\{T2 > 4\}) &= \int_4^{+\infty} \lambda^2 x e^{-\lambda x}\,\mathrm{d}x = \Big[-\lambda x e^{-\lambda x}\Big]_4^{+\infty} + \int_4^{+\infty} \lambda e^{-\lambda x}\,\mathrm{d}x = 4\lambda e^{-4\lambda} + e^{-4\lambda} \\ &= (4\lambda + 1)e^{-4\lambda} = 3e^{-2} \approx 40.60\%.\end{aligned}$$

Esercizio 4. (7 punti) Una ditta sostiene che il peso medio delle sue confezioni di zucchero è di 1 kg. Un campione di 10 confezioni fornisce i seguenti pesi (in kg):

$$1.01 \quad 0.92 \quad 1.02 \quad 1.00 \quad 0.97 \quad 1.03 \quad 1.01 \quad 0.96 \quad 0.90 \quad 1.02.$$

Si supponga che i pesi siano distribuiti normalmente.

1. Si può accettare l'affermazione della ditta con significatività $\alpha = 0.05$? (N.B.: derivare le formule).

2. Stabilire in quali dei seguenti intervalli è contenuto il p-value: $[0.01, 0.02)$, $[0.02, 0.05)$, $[0.05, 0.10)$, $[0.1, 0.20)$, nessuno dei precedenti.

Soluzione. Abbiamo a che fare con una popolazione $X \sim \mathcal{N}(\mu, \sigma^2)$ e un campione casuale $X_1, \ldots, X_n$ estratto dalla popolazione.

Sia $\mu_0 = 1$. Consideriamo il seguente test di ipotesi

$$H_0 : \mu = \mu_0\,, \quad H_1 : \mu \neq \mu_0\,,$$

con livello di significatività α.

Poiché l'ipotesi alternativa è $H_1 : \mu \neq \mu_0$, i dati saranno significativi se la media è sufficientemente distante da μ_0. La regione critica è allora della forma

$$R_c = \{(x_1, \dots, x_n) \in R(X_1, \dots, X_n) \ : \ |\overline{x}_n - \mu_0| > \delta\}\,,$$

dove $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ è la media calcolata sulla realizzazione $x_1, \dots, x_n$ del campione casuale.

Assumiamo l'ipotesi nulla H_0 vera, cioè $\mu = \mu_0$. Quindi $X_1, \dots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$. Non è nota σ^2, quindi verrà stimata dalla varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, dove $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ è la media campionaria. Dalla definizione di livello di significatività:

$$\alpha = \mathbb{P}(\{(X_1, \dots, X_n) \in R_c\}) = \mathbb{P}(\{|\overline{X}_n - \mu_0| > \delta\}) = \mathbb{P}\Big(\Big\{\frac{|\overline{X}_n - \mu_0|}{S_n/\sqrt{n}} > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big)\,.$$

Poiché $X_1, \dots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$ indipendenti, si ha che $T_{n-1} = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1)$. Quindi, per la simmetria della t-Student

$$\begin{aligned}\alpha &= \mathbb{P}\Big(\Big\{|T_{n-1}| > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) = \mathbb{P}\Big(\Big\{T_{n-1} < -\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) + \mathbb{P}\Big(\Big\{T > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) \\ &= 2\mathbb{P}\Big(\Big\{T_{n-1} > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big)\,,\end{aligned}$$

da cui

$$\mathbb{P}\Big(\Big\{T_{n-1} > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) = \frac{\alpha}{2}\,.$$

Introduciamo il valore $t_{n-1,\alpha/2}$ tale che

$$\mathbb{P}(\{T_{n-1} > t_{n-1,\alpha/2}\}) = \frac{\alpha}{2}\,.$$

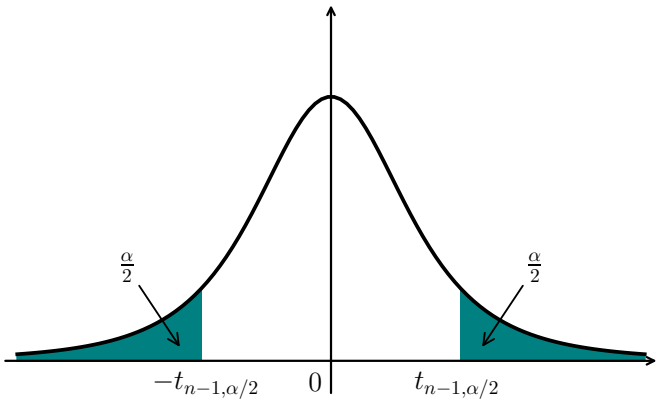


Figura 2: Code della legge t-Student con probabilità $\frac{\alpha}{2}$.

Allora, scegliendo

$$\frac{\delta}{S_n/\sqrt{n}} = t_{n-1,\alpha/2} \implies \delta = \frac{S_n}{\sqrt{n}} t_{n-1,\alpha/2}\,,$$

otteniamo la condizione desiderata.

In conclusione, la regione critica è

$$R_c = \Big\{(x_1, \dots, x_n) \in R(X_1, \dots, X_n) \ : \ |\overline{x}_n - \mu_0| > \frac{s_n}{\sqrt{n}} t_{n-1,\alpha/2}\Big\}\,,$$

e decidiamo come segue:

- Se $|\overline{x}_n - \mu_0| > \frac{s_n}{\sqrt{n}} t_{n-1,\alpha/2}$, i dati sono sufficientemente significativi da rifiutare H_0. L'ipotesi nulla H_0 viene rifiutata (con livello di significatività α).
- Se $|\overline{x}_n - \mu_0| \leq \frac{s_n}{\sqrt{n}} t_{n-1,\alpha/2}$, i dati non sono sufficientemente significativi da rifiutare. L'ipotesi nulla H_0 non viene rifiutata (con livello di significatività α).

Calcoliamo i valori sui dati.

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(1.01 + 0.92 + 1.02 + 1.00 + 0.97 + 1.03 + 1.01 + 0.96 + 0.90 + 1.02) = 0.984\,,$$

$$\begin{aligned} s_n^2 &= \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}_n^2\Big) \\ &= \frac{1}{9}\Big(1.01^2 + 0.92^2 + 1.02^2 + 1.00^2 + 0.97^2 + 1.03^2 + 1.01^2 + 0.96^2 + 0.90^2 + 1.02^2 - 10 \cdot 0.984^2\Big) \\ &= 2.03 \cdot 10^{-3}\,. \end{aligned}$$

Calcoliamo ora il valore $t_{9,0.025}$ dalla tavola:

$$t_{9,0.025} = 2.262\,.$$

Possiamo calcolare le soglie:

$$\mu_0 + \frac{s_n}{\sqrt{n}} t_{n-1,\alpha/2} = 1 + \frac{\sqrt{2.03 \cdot 10^{-3}}}{\sqrt{10}} \cdot 2.262 = 1.03\,,$$

$$\mu_0 - \frac{s_n}{\sqrt{n}} t_{n-1,\alpha/2} = 1 - \frac{\sqrt{2.03 \cdot 10^{-3}}}{\sqrt{10}} = 0.968\,.$$

Poiché $\overline{x}_n = 0.984$ è compreso tra 0.968 e 1.03, non possiamo rifiutare l'ipotesi nulla con significatività $\alpha = 0.05$.

2. Il p-value è sicuramente maggiore di 0.05, poiché non possiamo rifiutare l'ipotesi nulla con significatività $\alpha = 0.05$.

Calcoliamo le soglie per $\alpha = 0.10$ e $\alpha = 0.20$:

$$\mu_0 - \frac{s_n}{\sqrt{n}} t_{n-1,\alpha/2} = 1 - \frac{\sqrt{2.03 \cdot 10^{-3}}}{\sqrt{10}} 1.833 \approx 0.9738$$

$$\mu_0 - \frac{s_n}{\sqrt{n}} t_{n-1,\alpha/2} = 1 - \frac{\sqrt{2.03 \cdot 10^{-3}}}{\sqrt{10}} 1.383 \approx 0.9803\,.$$

Quindi il p-value è maggiore di 0.20 e non è contenuto in nessuno degli intervalli proposti.