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. 2021–2022

Gianluca Orlando

Indice

1 Tracce

Di seguito le tracce dell'a.a. 2021-2022.

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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: giugno 2022
Corso di studi: _____ Data: 20/06/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Si pensa che la massa di vapore utilizzato al mese da un impianto chimico sia correlato alla temperatura ambiente media di quel mese. L'utilizzo e la temperatura dell'ultimo anno sono riportati nella tabella seguente:

mese	temperatura ($°C$)	vapore ($kg/1000$)
gen.	-6	84.25
feb.	-4	97.26
mar.	0	130.62
apr.	8	192.67
mag.	10	206.15
giu.	15	244.45
lug.	20	288.88
ago.	23	306.14
set.	16	254.88
ott.	10	205.41
nov.	5	167.77
dic.	-1	124.25

1. Rappresentare i dati in un diagramma a dispersione.
2. Calcolare e rappresentare la retta di regressione lineare.
3. Calcolare il coefficiente di correlazione.

Esercizio 2. (7 punti) Una cioccolateria produce due varietà di cioccolatini (fondenti oppure al latte). Vende confezioni assortite composte da 10 cioccolatini. I cioccolatini possono essere indipendentemente fondenti o al latte e, in media, ci sono 6 cioccolatini al latte in una confezione.

1. Compri una confezione di cioccolatini. Qual è la probabilità di trovare almeno 8 cioccolatini fondenti?

2. Compri e ricompri confezioni di 10 cioccolatini (ogni acquisto è indipendente dal successivo) finché non hai una confezione con un ugual numero di cioccolatini fondenti e al latte. In media, quante confezioni devi acquistare prima di avere una confezione con con un ugual numero di cioccolatini fondenti e al latte?

3. Nella stessa situazione del punto 2., qual è la probabilità di dover acquistare più di 10 confezioni per avere una confezione con con un ugual numero di cioccolatini fondenti e al latte?

Esercizio 3. (7 punti) Sia (X_1, X_2) il vettore aleatorio con la seguente funzione di probabilità congiunta:

X_1 / X_2	-1	0	1
-1	a	$2a$	b
1	b	a	$2a$

1. Calcolare la varianza di X_2.

2. Determinare a e b tali che $\mathrm{Cov}(X_1, X_2) = 0$.

3. Per i valori a e b trovati nel punto 2., le variabili aleatorie X_1 e X_2 sono indipendenti?

Esercizio 4. (8 punti) Il contenuto di catrame (in mg) in sigarette prodotte da un'azienda si può supporre distribuito con legge normale. Dalle misurazione di 15 campioni di sigarette si ottengono i seguenti risultati:

$$6.9 \quad 7.4 \quad 7.3 \quad 6.6 \quad 7.0 \quad 6.7 \quad 7.1 \quad 6.2 \quad 7.2 \quad 6.6 \quad 6.9 \quad 6.5 \quad 7.2 \quad 7.7 \quad 7.5\,.$$

1. Determinare un intervallo di confidenza al 95% per la media del contenuto di catrame calcolata sui dati.

2. La realizzazione di un intervallo di confidenza al 97% sugli stessi dati (calcolata con lo stesso metodo del punto 1.) sarebbe più più o meno grande dell'intervallo trovato nel punto 1.? Motivare la risposta (N.B.: non è richiesto calcolare esplicitamente l'intervallo!)

Quesito teorico 1. (2 punti) Dimostrare che la covarianza di due variabili aleatorie indipendenti discrete (con valore atteso finito) è zero.

Quesito teorico 2. (4 punti) Siano $X \sim \mathrm{Gamma}(\alpha, \lambda)$ e $Y \sim \mathrm{Gamma}(\beta, \lambda)$ indipendenti. Dimostrare che $X + Y$ è distribuita con legge Gamma. Quali sono i parametri?

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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: luglio
Corso di studi: _____ Data: 18/07/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) I risultati dei test di adesione a trazione su 22 provini di lega U-700 mostrano i seguenti carichi di rottura (in megapascal):

23.1	10.1	15.4	18.5	11.4	14.1	19.5	8.8	14.9	7.5	7.9
12.7	15.4	15.4	11.9	11.4	17.6	16.7	15.8	13.6	11.9	11.4

1. Determinare i quartili dei dati.
2. Determinare eventuali dati anomali o sospetti.
3. Tracciare un box-plot.

Esercizio 2. (8 punti) Sei un ingegnere gestionale e fai parte di un gruppo di esperti selezionati per formare una commissione giudicatrice. Oltre a te ci sono: 2 ingegneri gestionali, 6 ingegneri elettrici, 4 ingegneri civili. Tra gli ingegneri elettrici ci sono tua sorella e tuo fratello.

1. Si deve formare una commissione composta da 2 ingegneri gestionali, 4 ingegneri elettrici, 2 ingegneri civili scegliendo in modo casuale (e uniformemente rispetto alle possibili commissioni realizzabili) tra i possibili esperti. Qual è la probabilità che tu non venga selezionato?
2. È stata selezionata la commissione come nel punto 1. Ti hanno detto che il tuo cognome (che è anche quello di tua sorella e tuo fratello) compare esattamente due volte, ma non sai di preciso chi di voi tre è stato selezionato. Qual è la probabilità che tu sia stato selezionato?

Esercizio 3. (7 punti) Devi aprire un conto alla poste. Prendi il biglietto e vedi che ci sono tre sportelli. A servire lo sportello 1 c'è una persona molto motivata ed efficiente, a servire lo sportello 2 una persona normale, a servire lo sportello 3 una persona evidentemente frustrata

e con poca voglia di lavorare. Il tempo che impiegherai per concludere l'apertura del conto è distribuito con una legge esponenziale, ma il tempo medio che impiegherai dipende da quale sportello ti capita: allo sportello 1 la media sarebbe 10 minuti; allo sportello 2, 15 minuti; allo sportello 3, 30 minuti. Nell'attesa ti accorgi che il 65% delle persone viene servito allo sportello 1, il 25% dallo sportello 2, il 10% dallo sportello 3.

1. Qual è la probabilità che impiegherai più di 30 minuti a concludere l'operazione?
2. Finita l'operazione, chiami una tua amica e le dici che hai impiegato più di 30 minuti ad aprire un conto! Lei è stata in quella filiale e conosce le tre persone addette agli sportelli (e le probabilità di essere serviti da questi). Con che probabilità è pronta a scommettere che sei stato servito dallo sportello 3?

Esercizio 4. (7 punti) Una riempitrice automatica viene utilizzata per riempire dosatori da 100 ml con gel igienizzante. Viene misurata la differenza tra il volume di riferimento 100 ml e il volume di riempimento effettivo in un campione casuale di dosatori:

$$-0.91 \quad 0.51 \quad 0.85 \quad 0.10 \quad -0.56 \quad 1.10 \quad 0.38 \quad 0.26 \quad 0.30 \quad -0.67 \quad 0.03 \quad -0.31 \quad 0.59$$

(ad esempio, il dato -0.91 indica che il dosatore è stato riempito con 100.91 ml di gel, il dato 0.51 indica che il dosatore è stato riempito con 99.49 ml di gel). Se la deviazione standard del volume di riempimento è superiore a 0.16 ml, la macchina riempitrice deve essere tarata nuovamente. I dati sono significativi all'1% per concludere che si deve tarare la macchina riempitrice? E al 5%? Si assuma che la popolazione sia distribuita con legge normale (Corretto in aula).
(N.B.: Ricavare le formule!)

Quesito teorico 1. (2 punti) Derivare la formula per la varianza di una variabile aleatoria distribuita con una legge di Poisson di parametro λ.

Quesito teorico 2. (4 punti) Ricavare la formula per la densità della somma di due variabili aleatorie assolutamente continue indipendenti X e Y in termini della densità di X e della densità di Y.

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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: luglio
Corso di studi: _____ Data: 18/07/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Un gruppo di topi di 5 settimane viene sottoposto a una dose di radiazione di 300 rad. La seguente tabella riporta le frequenze assolute dei giorni di vita dei topi suddivisi in intervalli di classi:

intervallo	frequenza
$[150, 200)$	6
$[200, 300)$	10
$[300, 500)$	12
$[500, 600)$	2
$[600, 900)$	3

1. Rappresentare un istogramma delle densità di frequenze relative.
2. Determinare la classe modale.
3. Calcolare un'approssimazione della media e della deviazione standard dei dati.
4. Calcolare un'approssimazione della mediana dei dati.

Esercizio 2. (7 punti) Un'azienda produce grandi numeri di mobili montabili. In un particolare mobile ci sono due pezzi (che chiameremo A e B) che possono risultare difettosi. In media vengono prodotti ogni giorno 1 pezzo A difettoso e (indipendentemente) 2 pezzi B difettosi. Per entrambi i tipi, il numero di pezzi difettosi è distribuito con una legge di Poisson.

1. Qual è la probabilità che vengano prodotti (strettamente) più di 4 pezzi A difettosi in 5 giorni? (Si assumano i difetti nei diversi giorni indipendenti)
2. Sappiamo che in 5 giorni sono stati prodotti in tutto 12 pezzi difettosi (contando sia tipo A che B). Qual è la probabilità che al più 3 pezzi A siano difettosi?

Esercizio 3. (8 punti) Sono le 8:30 e sei allo sportello della tua banca per sbrigare una pratica. In media la pratica dura 10 minuti, e la durata in minuti è distribuita come una legge esponenziale. Dopo aver sbrigato la pratica in banca devi andare a lavoro prendendo un bus, che però ha un tempo di arrivo alla fermata (proprio all'uscita della banca) incerto. L'orario di arrivo del bus ha distribuzione uniforme, in media arriva alle 8:40, con una deviazione standard di 10 min.

1. Con che probabilità si verifica il seguente evento: finirai la pratica dopo le 8:50 e il bus arriverà prima delle 8:50?

2. Hai finito la pratica, corri fuori e scopri che il bus è già passato. Non hai guardato l'orologio, quindi non sai quanto tempo è durata la pratica e a che ora è passato il bus. Constatando che non sei riuscito a prendere il bus, è più probabile che la pratica sia durata meno di 10 min oppure che il bus sia arrivato prima delle 8:40? Motivare la risposta. (N.B.: non è richiesto il calcolo esplicito delle due probabilità!)

Esercizio 4. (7 punti) Una riempitrice automatica viene utilizzata per riempire dosatori da 100 ml con gel igienizzante. Viene misurato il volume di riempimento effettivo in un campione casuale di 40 dosatori. La media campionaria e la deviazione standard campionaria calcolate sui dati risultano essere 99.90 ml e 0.55 ml rispettivamente. È possibile affermare con il 5% di significatività che in media la macchina immette meno di 100 ml di gel? Qual è il più piccolo livello di significatività per cui i dati permettono di affermare che la macchina immette meno di 100 ml di gel?
(N.B.: Ricavare le formule!)

Quesito teorico 1. (2 punti) Spiegare il fenomeno della scimmia di Borel. Come riferimento di testo da scrivere utilizzare il primo canto della Divina Commedia, composto da 4841 caratteri, utilizzando 52 caratteri (lettere, lettere accentate, punteggiatura).

Quesito teorico 2. (4 punti) Considerare una variabile aleatoria distribuita con una legge chi-quadro con n gradi di libertà (intesa come somma di quadrati di normali indipendenti) e dimostrare che ha una distribuzione Gamma. Con che parametri?

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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: settembre 2022 - I
Corso di studi: _____ Data: 07/09/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Un'azienda produce un dispositivo elettronico da utilizzare in un intervallo di temperatura molto ampio. L'azienda sa che l'aumento della temperatura riduce il tempo di vita del dispositivo, e quindi viene eseguito uno studio in cui il tempo di vita è determinato in funzione della temperatura. Si trovano i seguenti dati:

temperatura in $°C$	tempo di vita in ore
10	400
20	370
30	275
40	215
50	172
60	108
70	61
80	40
90	9

1. Rappresentare i dati in uno scatterplot.
2. Determinare (derivando le formule dei coefficienti) e rappresentare la retta di regressione lineare.
3. Calcolare il coefficiente di correlazione.

Esercizio 2. (7 punti) Una compagnia aerea ha osservato che su una certa tratta la probabilità che un passeggero che ha acquistato un biglietto non si presenti al momento dell'imbarco è del 5% (si supponga che i passeggeri siano indipendenti). L'aereo ha in tutto 96 posti, ma la compagnia prevede *overbooking* (sovrapprenotazione), quindi vende fino a 100 biglietti (supponiamo che la compagnia venda tutti i biglietti). Quindi non è detto che un posto a sedere sull'aereo sia garantito a tutti i passeggeri che hanno acquistato un biglietto e si presentano all'imbarco.

1. Qual è la probabilità che tutti i passeggeri che hanno acquistato il biglietto e si presentano all'imbarco abbiano un posto a sedere? (Suggerimento: considerare la variabile aleatoria che descrive il numero di passeggeri che hanno acquistato il biglietto e si presentano all'imbarco. Che distribuzione ha?)

2. La compagnia ricava 200€ da ogni biglietto acquistato, mentre deve pagare un risarcimento di 600€ ai passeggeri che si sono presentati all'imbarco ma per cui non erano disponibili posti. Qual è il guadagno atteso per questo volo considerando i risarcimenti dovuti? (Suggerimento: scrivere il guadagno in funzione della variabile aleatoria del punto 1.)

Esercizio 3. (7 punti) Sia X una variabile aleatoria assolutamente continua con la seguente densità

$$f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \quad x \in \mathbb{R}\,,$$

dove $\mu \in \mathbb{R}$ e $b > 0$ sono parametri da determinare.

1. Controllare che effettivamente $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$. (Suggerimenti: effettuare un cambio di variabile per traslazione, spezzare l'integrale in due e riscalare la variabile).

2. Determinare μ e b tali che $\mathbb{E}(X) = 0$ e $\mathrm{Var}(X) = 1$. (Suggerimenti: per $\mathbb{E}(X)$, effettuare un cambio di variabile per traslazione e utilizzare il punto 1.; per $\mathrm{Var}(X)$, utilizzare il valore di μ trovato, spezzare l'integrale in due, riscalare la variabile e integrare per parti)

3. Per i valori trovati, calcolare la probabilità che $X \leq 1$ sapendo che si è verificato l'evento $X \geq 0$.

Esercizio 4. (8 punti) Si sa che la percentuale di titanio in una lega utilizzata nelle fusioni aerospaziali è distribuita con legge normale. Nelle domande seguenti per "esperimento statistico" intendiamo la misurazione della percentuale di titanio in 20 campioni selezionati casualmente.

1. Si fa un esperimento statistico e la deviazione standard calcolata sul campione risulta essere 0.37. Calcolare sui dati un intervallo di confidenza unilaterale sinistro (ovvero un limite superiore di confidenza) al 95% per la varianza. N.B.: derivare le formule!

2. È vero o falso che la varianza della popolazione appartiene all'intervallo calcolato nel punto precedente con il 95% di probabilità? Motivare la risposta.

3. Si ripetono tanti esperimenti statistici indipendenti. In media, dopo quanti esperimenti accade per la prima volta che la varianza della popolazione è fuori dall'intervallo di confidenza unilaterale sinistro al 95%?

Quesito teorico 1. (3 punti) Siano X e Y due variabili aleatorie indipendenti distribuite con leggi di Poisson con parametri λ e μ rispettivamente. Dimostrare che $X + Y$ è una variabile aleatoria distribuita con una legge di Poisson. Con che parametro?

Quesito teorico 2. (3 punti) Enunciare e dimostrare la legge dei grandi numeri.

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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: settembre 2022 - II
Corso di studi: _____ Data: 20/09/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) In un'indagine sui consumi di nuove auto a benzina è stata osservata la distribuzione dei litri consumati per 100 km. I dati sono rappresentati raggruppati in intervalli di classi nella seguente tabella:

intervallo	frequenze assolute
$[4, 4.5)$	15
$[4.5, 5.5)$	40
$[5.5, 6)$	30
$[6, 7)$	25
$[7, 9)$	10

1. Rappresentare un istogramma delle densità di frequenze assolute.
2. Determinare la classe modale.
3. Calcolare un'approssimazione della media e della deviazione standard dei dati.
4. Calcolare un'approssimazione della mediana dei dati.

Esercizio 2. (8 punti) L'azienda per cui lavori offre ogni anno un corso di aggiornamento facoltativo. Il numero di persone che fa domanda per il corso è una variabile aleatoria distribuita con una legge di Poisson e, in media, 5 persone fanno domanda per seguire il corso. L'azienda deve decidere se offrire il corso in streaming oppure in presenza (e in tal caso deve organizzarsi per tempo per procurarsi un'aula). Se il numero di partecipanti è (strettamente) minore di 4, il corso è offerto in streaming, altrimenti il corso è offerto in presenza. (Chiarimento: "persone che fanno domanda" = "partecipanti")

1. Qual è la probabilità che il corso venga offerto in presenza?

2. L'azienda viene a conoscenza del numero dei primi iscritti e sa che il corso verrà offerto in presenza. Deve quindi prenotare un'aula. Se l'azienda vuole che la probabilità di far sedere tutti i partecipanti sia almeno del 90%, sono sufficienti 6 posti a sedere? Se no, quanti ne servono?

3. La seguente affermazione è vera oppure falsa? "Grazie all'assenza di memoria possiamo affermare che la probabilità che il numero di partecipanti sia maggiore di 15 sapendo che il numero di partecipanti è maggiore di 10 è uguale alla probabilità che il numero di partecipanti sia maggiore di 5." (N.B.: non sono richiesti i calcoli, ma si deve motivare la risposta)

Esercizio 3. (7 punti) Sia (X_1, X_2) un vettore aleatorio con funzione di probabilità congiunta descritta dalla seguente tabella:

X_2 \ X_1	1	2	3
-1	$1/6$	a	b
1	b	$1/6$	a

dove $a, b \geq 0$.

1. Determinare i valori di a e b per cui $\mathrm{Cov}(X_1, X_2) = 0$.

2. Per i valori di a e b determinati nel punto 2., si ha che X_1 e X_2 sono indipendenti?

3. Si vuole osservare una successione di realizzazioni indipendenti del vettore aleatorio (X_1, X_2). Qual è la probabilità di dover attendere (strettamente) più di 10 osservazioni perché si verifichi $X_1 = 1$?

Esercizio 4. (7 punti) Un produttore sostiene che le sue batterie abbiano una durata di almeno 100 ore. Si sa che la deviazione standard per questo tipo di batterie è di $\sigma = 10$ ore. Un cliente, insospettito dall'affermazione del produttore, fa una prova: acquista e testa 40 campioni, osservando una media campionaria di 96.5 ore.

1. L'osservazione del cliente è significativa al 5% per destare sospetti sull'effettiva qualità delle batterie?

2. Qual è il più piccolo livello di significatività per cui i dati osservati permettono di contestare l'affermazione del produttore?

Quesito teorico 1. (3 punti) Siano $X \sim B(n, p)$ e $Y \sim B(m, p)$ indipendenti. Mostrare che anche $X + Y$ ha distribuzione binomiale. Con che parametri?

Quesito teorico 2. (3 punti) Sia $X \sim \mathrm{Gamma}(\alpha, \lambda)$. Calcolare $\mathbb{E}(X)$ e $\mathrm{Var}(X)$.

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: _____ A.A.: 2022/2023
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: novembre 2022
Corso di studi: _____ Data: 11/11/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) I voti ottenuti dagli studenti e dalle studentesse a un appello dell'esame di Probabilità e Statistica[1] sono i seguenti:

26 8 28 5 27 30 26 18 28 26 25 20 22 30 20 21

1. Determinare i quartili.
2. Determinare eventuali dati anomali o sospetti.
3. Tracciare un box-plot.

Esercizio 2. (7 punti) Un ristorante studia i suoi clienti fissi e gli effetti della nuova campagna di marketing adottata dal ristorante tramite *ad* su un *social network*. Gli *ad* risultano efficaci su una certa proporzione p dei clienti. Lo studio porta a queste conclusioni:

- Per un cliente che è influenzato dagli *ad*, il numero di visite annue al ristorante è distribuito con una legge di Poisson con una media di 7 visite all'anno;
- Per un cliente che non è influenzato dagli *ad* (nella restante proporzione della popolazione $1 - p$), il numero di visite annue al ristorante è distribuito con una legge di Poisson con una media di 4 visite all'anno.

Rispondere ai seguenti quesiti:

1. Determinare (in funzione di p) la probabilità che il numero di visite annue di un cliente sia uguale a un dato numero $k = 0, 1, 2, \ldots$.
2. Un'analisi mostra che il 50% dei clienti visita il ristorante almeno 5 volte all'anno. (Si legga: la probabilità che il numero di visite di un cliente sia maggiore o uguale a 5 è il 50%.) Ricavare in questo caso la proporzione p dei clienti che viene influenzata dagli *ad*.

[1] I dati sono generati casualmente e non si riferiscono a fatti realmente accaduti.

Esercizio 3. (7 punti) L'ufficio informazioni di una compagnia ha due numeri verdi distinti. I tempi di attesa per parlare con gli operatori sono, per entrambi i numeri, variabili aleatorie distribuite con legge esponenziale con media 10 minuti. Inoltre i due tempi di attesa si possono considerare indipendenti. Avendo a disposizione due telefoni, decidi di chiamare contemporaneamente i due numeri.

1. Qual è la probabilità che qualcuno risponda dal primo numero dopo 5 minuti? E che qualcuno risponda dal secondo numero dopo 5 minuti?

2. Qual è la probabilità di attendere meno di 5 minuti fino alla risposta da uno dei due numeri (non importa quale dei due)?

3. (Domanda bonus con punteggio extra) Quanto tempo aspetterai in media fino alla risposta da uno dei due numeri?

Esercizio 4. (8 punti) Si vuole studiare l'effetto della delaminazione sulla frequenza naturale delle travi realizzate con laminati compositi. Si effettua il seguente esperimento statistico: si considera un campione di otto travi delaminate, le si sottopongono a carichi e se ne misurano le frequenze risultanti (in Hertz). Per un esperimento statistico si osservano i seguenti dati:

230.66 233.05 232.58 229.48 232.58 235.76 229.43 234.13

Si supponga che i dati provengano da una popolazione distribuita con legge normale.

1. Calcolare sui dati un intervallo di confidenza bilaterale al 90% sulla frequenza naturale media della popolazione.

2. Supponiamo di effettuare 20 esperimenti statistici indipendenti (ottenendo per ogni esperimento statistico nuovi valori) e di calcolare sui dati di ogni esperimento statistico un intervallo di confidenza bilaterale al 90% come nel punto precedente. Qual è la probabilità che almeno 4 volte la frequenza naturale media della popolazione sia fuori dall'intervallo di confidenza calcolato sui dati?

Quesito teorico 1. (3 punti) Calcolare valore atteso e varianza di una variabile aleatoria distribuita con una legge normale $\mathcal{N}(\mu, \sigma^2)$, motivando la risposta.

Quesito teorico 2. (3 punti) Enunciare e dimostrare la proprietà di assenza di memoria per variabili aleatorie distribuite con legge geometrica.

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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: gennaio 2023
Corso di studi: _____ Data: 26/01/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) I seguenti dati indicano la relazione tra velocità di lettura (parole al minuto) e il numero di settimane trascorse in un programma di lettura veloce per 10 studenti:

settimane	2	3	8	11	4	5	9	7	5	7
velocità di lettura	21	42	102	130	52	57	105	85	62	90

1. Rappresentare i dati in uno scatterplot.
2. Determinare (derivando le formule dei coefficienti) e rappresentare la retta di regressione lineare.
3. Calcolare il coefficiente di correlazione.

Esercizio 2. (8 punti) Si consideri un vettore aleatorio discreto (X_1, X_2) con funzione di probabilità congiunta data dalla seguente tabella:

X_1 / X_2	0	1	2	3
0	$a_{00}/8$	$a_{10}/8$	$a_{20}/8$	$a_{30}/8$
1	$a_{01}/8$	$a_{11}/8$	$a_{21}/8$	$a_{31}/8$

dove $a_{00}, a_{10}, a_{20}, a_{30}, a_{01}, a_{11}, a_{21}, a_{31} \geq 0$.

1. Trovare i valori espliciti di a_{ij} nella tabella sapendo che:
   - X_1 ha legge binomiale $\mathrm{B}(n, p)$ con parametri $n = 3$ e $p = \frac{1}{2}$.
   - X_2 ha legge di Bernoulli $\mathrm{Be}(q)$ con parametro $q = \frac{7}{8}$ (1 = successo).
   - $\mathbb{P}(\{X_1 = 3\}|\{X_2 = 1\}) = \frac{1}{14}$.

- $\mathbb{P}(\{X_1 + X_2 = 1\}) = \frac{1}{8}$.
- $\mathbb{P}(\{X_1 = 2\}, \{X_2 = 0\}) = \frac{1}{16}$.

(Suggerimento: usare le condizioni nell'ordine in cui sono fornite.)

2. Calcolare la covarianza di X_1 e X_2.
3. Le variabili aleatorie X_1 e X_2 sono indipendenti?

Esercizio 3. (7 punti) Un'azienda produce uno smartphone con una vita media di 4 anni, dopodiché si rompe. Assumiamo che la vita dello smartphone (misurata in anni) sia una variabile aleatoria con legge esponenziale.

1. Acquisti uno smartphone. Qual è la probabilità che funzioni per più di 6 anni?
2. Acquisti uno smartphone. Passano 3 anni e funziona ancora. Qual è la probabilità che funzioni in tutto per più di 6 anni, sapendo che è successo il fatto precedente?
3. Acquisti tre smartphone (assumiamo che le vite dei tre smartphone siano indipendenti). qual è la probabilità che almeno due dei tre funzionino per più di 6 anni?
4. (**Bonus**) Acquisti due smartphone (assumiamo che le vite dei due smartphone siano indipendenti). Ne usi solo uno, finché si rompe (assumiamo che intanto lo smartphone non utilizzato non perda anni di vita). Poi inizi a usare l'altro (che ha una vita media di 4 anni). Chiamiamo vita cumulata la somma delle vite dei due smartphone. Qual è la probabilità che la vita cumulata sia più di 12 anni?

Esercizio 4. (7 punti) Un'azienda produce una margarina dietetica per cui si sa, fino a prova contraria, che il livello di acidi grassi polinsaturi (in percentuale) ha una deviazione standard di 1.2. È stata proposta una nuova tecnica di produzione del prodotto, che tuttavia comporta un costo aggiuntivo. La direzione autorizzerà un cambiamento nella tecnica di produzione se si riesce a mostrare che la deviazione standard del livello di acidi grassi polinsaturi con il nuovo processo è significativamente inferiore a 1.2. Un campione del lotto ottenuto con il nuovo metodo ha prodotto le seguenti percentuali di livello di acidi grassi polinsaturi:

16.8 17.2 17.4 16.9 16.5 17.1 18.2 16.8 15.7 16.1

Si assuma che i dati siano distribuiti con legge normale.

1. I dati sono significativi al 5% per decidere di cambiare il metodo di produzione? (N.B.: Derivare le formule)
2. Siamo interessati al più piccolo livello di significatività per cui i dati porterebbero a decidere di cambiare il metodo di produzione. In quale di questi intervalli si colloca tale valore: $[0.5\%, 1\%)$, $[1\%, 2.5\%)$, $[2.5\%, 5\%)$, $[5\%, 10\%)$? (N.B.: Non è richiesto il calcolo esplicito del più piccolo livello di significatività)

Quesito teorico 1. (3 punti) Sia X una variabile aleatoria con legge geometrica di parametro p. Calcolarne media e varianza.

Quesito teorico 2. (3 punti) Enunciare il Teorema del Limite Centrale e discuterne un'applicazione.

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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: febbraio 2023
Corso di studi: _____ Data: 09/02/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Viene esaminata in un campione la resistenza alla compressione del calcestruzzo quando miscelato con la cenere volante (una miscela di silice, allumina, ossido di ferro, e altri elementi). Vengono riportati i seguenti dati in megapascal:

22.4 50.2 30.4 14.2 28.9 30.5 25.8 18.4 15.3 21.1

1. Calcolare i quartili dell'insieme dei dati.
2. Determinare eventuali dati anomali o sospetti.
3. Disegnare un box-plot.

Esercizio 2. (7 punti) Alice e Bob fanno un gioco. Alice lancia consecutivamente un dado a 6 facce non truccato tante volte (i lanci sono indipendenti). Se entro il quarto lancio (compreso) esce un 6, vince Bob, altrimenti vince Alice.

1. Alice e Bob giocano una partita. Con che probabilità vince Alice?

Alice però è molto brava nel calcolo delle probabilità e si rende conto che c'è qualcosa che va a suo sfavore nelle regole di questo gioco... Decide allora di truccare il dado, all'insaputa di Bob. Dopo aver giocato alcune partite, Bob si rende conto che Alice vince con il 70% di probabilità. Inoltre, il 6 esce per la prima volta, in media, troppo tardi. Quindi l'accusa di aver barato.

2. Cosa ha fatto esattamente Alice al dado?
3. In media, dopo quale lancio esce il 6 con questo dado truccato?

Esercizio 3. (7 punti) Il tempo necessario per un/a tecnico/a dell'assistenza per cambiare l'olio in un'auto è una variabile aleatoria. Se l'auto non presenta problemi, è distribuita uniformemente tra 10 e 20 minuti. Se l'auto presenta dei problemi, è distribuita uniformemente con media 20 minuti e varianza 12 min^2. In media, il 90% delle auto non presenta problemi.

1. Viene riparata un'auto. Sapendo che il tempo impiegato per cambiare l'olio è stato maggiore di 18 minuti, qual è la probabilità che l'auto avesse dei problemi?

2. Vengono riparate 2 auto (si assumano i tempi di cambio dell'olio per le due auto indipendenti). Qual è la probabilità che il cambio d'olio più rapido duri meno di 18 minuti?

Esercizio 4. (8 punti) Il/la docente di Probabilità e Statistica vuole un'evidenza significativa del fatto che in questo anno accademico l'esame sia stato più difficile per gli studenti e le studentesse.[1] Negli anni accademici precedenti, la media dei voti era 24. In un appello di questo anno accademico, invece, sono stati registrati i seguenti voti:

21	26	23	25	18	24	18	28	23	26	20	20	18	20	21	20
30	25	19	23	26	26	26	26	29	28	25	18	21	26	18	27

1. I voti assegnati nell'ultimo appello sono significativi al 5% per concludere che la media è effettivamente più bassa rispetto agli anni precedenti? (N.B.: ricavare le formule)

2. In questo anno accademico si svolgono 8 appelli, come programmato. Dopo ciascun appello si registrano i voti come nel punto precedente. A volte si conclude che la media è più bassa rispetto agli anni precedenti, altre volte no. Assumendo che la media di questo anno accademico sia rimasta 24, qual è la probabilità di commettere un errore (strettamente) meno di 6 volte arrivando a conclusioni con l'analisi precedente?

Quesito teorico 1. (3 punti) Sia X una variabile aleatoria distribuita con legge esponenziale. Calcolarne valore atteso e varianza.

Quesito teorico 2. (4 punti) Sia $X = (X_1, X_2)$ un vettore aleatorio con densità congiunta

$$f_{(X_1,X_2)}(x_1, x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}} .$$

Si consideri la variabile aleatoria $|X|^2$ data dalla norma al quadrato di X, ovvero $|X|^2 = (X_1)^2 + (X_2)^2$. Che legge ha $|X|^2$? Motivare la risposta.

[1] L'esercizio è inventato e ogni riferimento a persone o fatti realmente accaduti è puramente casuale.

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: aprile 2023
Matricola: _____ Data: 03/04/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) La tabella seguente mostra i dati sul consumo medio annuo di vino pro capite e sul numero di morti dovute a malattie cardiache in un campione casuale di 10 paesi:

consumo di vino (in litri)	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
morti	221	167	131	191	220	297	71	172	211	300

1. Rappresentare i dati in uno scatterplot.
2. Determinare (derivando le formule) la retta di regressione lineare e rappresentarla.
3. Determinare il coefficiente di correlazione lineare.

Esercizio 2. (7 punti) Un produttore di un componente elettronico sa che un componente prodotto è difettoso con una probabilità del 10% (si assumano i difetti dei componenti indipendenti tra loro).

1. Il produttore vende a un cliente una confezione con 20 componenti. Qual è la probabilità che la confezione contenga almeno 18 (18 incluso) componenti non difettose?

Il prezzo di vendita di una confezione da 20 pezzi è di 15€. Se il cliente riceve una confezione con almeno 18 componenti non difettose, non fa un reclamo. Altrimenti, il cliente fa un reclamo e chiede al produttore di inviare una nuova confezione (senza pagare nuovamente i 15€). Questa operazione si ripete finché il cliente non riceve una confezione con almeno 18 componenti sane.

2. In media, quante volte farà reclamo il cliente?

Per il produttore, il costo di produzione di una confezione da 20 pezzi è di 6€. Ogni volta che spedisce una confezione (la prima volta e per ogni eventuale reclamo), paga 2€ di costi di spedizione.

3. Qual è la probabilità che il produttore abbia una perdita per via dei ripetuti reclami dovuti ai difetti di una confezione?

Esercizio 3. (8 punti) Consideriamo una persona che sta svolgendo l'esame di Probabilità e Statistica. Se la persona ha studiato, il tempo (in minuti) che impiega a svolgere tutti gli esercizi del compito è distribuito con legge uniforme con media 90 min e varianza 12 min^2. Se la persona non ha studiato, il tempo (in minuti) che impiega a svolgere tutti gli esercizi del compito è distribuito con legge uniforme nell'intervallo $[90, 120]$. Il 70% delle persone che si presentano all'esame ha studiato.

1. Consideriamo una persona che sappiamo che non ha studiato. Con che probabilità impiegherà più di 100 minuti a svolgere il compito?

2. Consideriamo una persona che sappiamo che ha studiato. Con che probabilità impiegherà più di 90 minuti a svolgere il compito?

3. Consideriamo una persona che svolge l'esame (non sappiamo se ha studiato o se non ha studiato). Vediamo che ha terminato tutti gli esercizi del compito in meno di 95 minuti. Sapendo questo fatto, con che probabilità la persona ha studiato?

(I dati sono inventati.)

Esercizio 4. (7 punti) Uno studio statistico ha riferito che precedentemente gli/le adolescenti trascorrevano in media 3 ore al giorno con lo smartphone. Si vuole mostrare con un'evidenza statistica che la media è diventata più alta. Ad alcuni/e adolescenti scelti casualmente è stato chiesto quante ore al giorno trascorrono con lo smartphone. I dati (in ore) sono i seguenti:

$$3.4 \quad 2.8 \quad 4.9 \quad 3.5 \quad 4.8 \quad 4.1 \quad 4.0 \quad 3.2 \quad 5.5 \quad 3.2 \quad 4.4 \quad 5.3 \quad 5.3 \quad 4.7 \quad 4.3\,.$$

(I dati sono inventati.) Si assuma che la popolazione abbia una distribuzione normale.

1. I dati sono significativi al 10% per stabilire che la media è davvero più alta?

2. In quale dei seguenti intervalli si posiziona il più piccolo livello di significatività per cui i dati portano a stabilire che la media è davvero più alta? $[0\%, 0.5\%)$, $[0.5\%, 1\%)$, $[1\%, 2.5\%)$, $[2.5\%, 5\%)$, $[5\%, 10\%)$, $[10\%, 100\%]$?

Quesito teorico 1. (3 punti) Si enunci e dimostri il teorema di approssimazione della legge binomiale con legge di Poisson.

Quesito teorico 2. (3 punti) Siano $Z \sim \mathcal{N}(0, 1)$ e $X = Z^2$. Calcolare esplicitamente la densità di X. Che legge ha X?

2 Soluzioni

Di seguito le soluzioni relative alle tracce di sopra.

Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni 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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: giugno 2022
Corso di studi: _____ Data: 20/06/2022

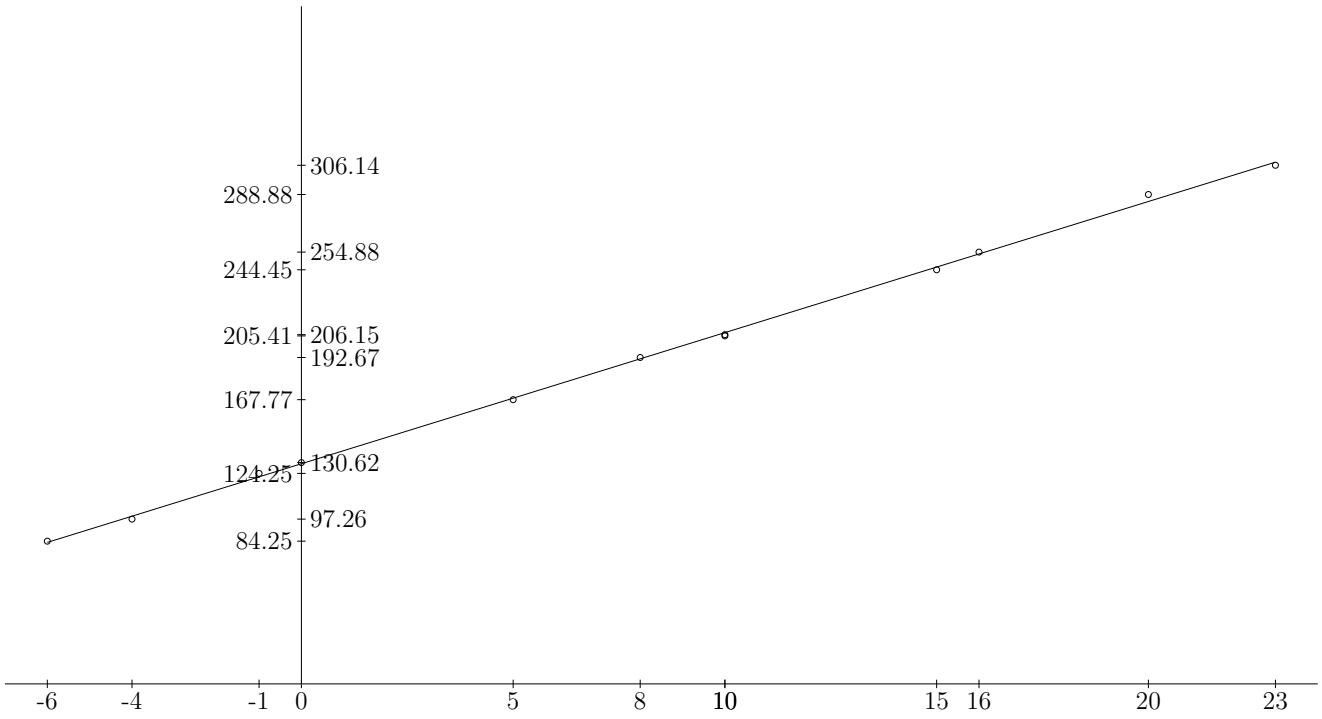
Tempo massimo: 2 ore.

Esercizio 1. Si pensa che la massa di vapore utilizzato al mese da un impianto chimico sia correlato alla temperatura ambiente media di quel mese. L'utilizzo e la temperatura dell'ultimo anno sono riportati nella tabella seguente:

mese	temperatura ($°C$)	vapore ($kg/1000$)
gen.	-6	84.25
feb.	-4	97.26
mar.	0	130.62
apr.	8	192.67
mag.	10	206.15
giu.	15	244.45
lug.	20	288.88
ago.	23	306.14
set.	16	254.88
ott.	10	205.41
nov.	5	167.77
dic.	-1	124.25

1. Rappresentare i dati in un diagramma a dispersione.
2. Calcolare e rappresentare la retta di regressione lineare.
3. Calcolare il coefficiente di correlazione.

Soluzione. 1. Segue il diagramma a dispersione.



2. Riportiamo i dati in una tabella:

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
-6	84.25	-505.50	36	7098.0625
-4	97.26	-389.04	16	9459.5076
0	130.62	0.00	0	17061.5844
8	192.67	1541.36	64	37121.7289
10	206.15	2061.50	100	42497.8225
15	244.45	3666.75	225	59755.8025
20	288.88	5777.60	400	83451.6544
23	306.14	7041.22	529	93721.6996
16	254.88	4078.08	256	64963.8144
10	205.41	2054.10	100	42193.2681
5	167.77	838.85	25	28146.7729
-1	124.25	-124.25	1	15438.0625

Utilizzando il metodo dei minimi quadrati si trovano i coefficienti α e β della retta di regressione lineare $y = \alpha x + \beta$. L'obiettivo è minimizzare:

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

Imponiamo che il gradiente rispetto ad a e b sia zero:

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

$$0 = \sum_{i=1}^{n} 2(y_i - ax_i - b)$$

da cui segue

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

Calcoliamo la media dei dati x_i e dei dati y_i:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{12}(-6 - 4 + \cdots + 5 - 1) = 8$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{1}{12}(84.25 + 97.26 + \cdots + 167.77 + 124.25) \simeq 191.89\,.$$

Otteniamo

$$a = \frac{(-505.50 - 389.04 + \cdots + 838.85 - 124.25) - 12 \cdot 8 \cdot 191.89}{(36 + 16 + \cdots + 25 + 1) - 12 \cdot 8^2} = \frac{26040.67 - 18421.44}{1752 - 768} \simeq 7.74$$

$$b = 129.95$$

quindi la retta di regressione lineare è

$$y = 7.74x + 129.95\,.$$

3. Per calcolare il coefficiente di correlazione possiamo usare le formule:

$$\rho = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}^2}} = a\frac{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}^2}}{\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}^2}} \simeq 7.74\frac{\sqrt{984}}{\sqrt{59029.33}} \simeq 0.9996703752\,.$$

Esercizio 2. Una cioccolateria produce due varietà di cioccolatini (fondenti oppure al latte). Vende confezioni assortite composte da 10 cioccolatini. I cioccolatini possono essere indipendentemente fondenti o al latte e, in media, ci sono 6 cioccolatini al latte in una confezione.

1. Compri una confezione di cioccolatini. Qual è la probabilità di trovare almeno 8 cioccolatini fondenti?

2. Compri e ricompri confezioni di 10 cioccolatini (ogni acquisto è indipendente dal successivo) finché non hai una confezione con un ugual numero di cioccolatini fondenti e al latte. In media, quante confezioni devi acquistare prima di avere una confezione con con un ugual numero di cioccolatini fondenti e al latte?

3. Nella stessa situazione del punto 2., qual è la probabilità di dover acquistare più di 10 confezioni per avere una confezione con con un ugual numero di cioccolatini fondenti e al latte?

Soluzione. Il numero di cioccolatini fondenti trovati in una confezione da 10 è distribuito come una variabile aleatoria binomiale $X \sim B(n,p)$ con parametri $n = 10$ e p tale che

$$np = \mathbb{E}(X) = \text{numero medio di cioccolatini fondenti} = 10 - 6 = 4\,,$$

da cui $p = 4/n = 4/10 = 2/5$. Quindi $X \sim B(10, 2/5)$.

1. Calcoliamo

$$\begin{aligned}\mathbb{P}(\{X \geq 8\}) &= \mathbb{P}(\{X = 8\}) + \mathbb{P}(\{X = 9\}) + \mathbb{P}(\{X = 10\}) \\ &= \binom{10}{8}\left(\frac{2}{5}\right)^8\left(\frac{3}{5}\right)^2 + \binom{10}{9}\left(\frac{2}{5}\right)^9\left(\frac{3}{5}\right) + \binom{10}{10}\left(\frac{2}{5}\right)^{10} \\ &= \frac{10!}{8!2!}\left(\frac{2}{5}\right)^8\left(\frac{3}{5}\right)^2 + \frac{10!}{9!1!}\left(\frac{2}{5}\right)^9\left(\frac{3}{5}\right) + \left(\frac{2}{5}\right)^{10} \simeq 1.23\%\,.\end{aligned}$$

2. Il numero di confezioni che si comprano fino al primo successo "è stata trovata una confezione con esattamente 5 cioccolatini fondenti" è una variabile aleatoria distribuita con una legge geometrica $Y \sim \mathrm{Geo}(q)$ con parametro q dato dalla probabilità di successo, ovvero $q = \mathbb{P}(\{X = 5\}) = \binom{10}{5}\left(\frac{2}{5}\right)^5\left(\frac{3}{5}\right)^5 \simeq 20.07\%$. Il numero medio di confezioni da comprare è dato dal valore atteso

$$\mathbb{E}(Y) = \frac{1}{q} \simeq 4.98$$

quindi circa 5 confezioni.
3. Utilizzando il punto precedente, dobbiamo calcolare

$$\begin{aligned}\mathbb{P}(\{Y > 10\}) &= 1 - \mathbb{P}(\{Y \leq 10\}) = 1 - \sum_{k=1}^{10}\mathbb{P}(\{Y = k\}) = 1 - \sum_{k=1}^{10}(1-q)^{k-1}q \\ &= 1 - 1 + (1-q)^{10} = (1-q)^{10} \simeq 10.64\%\,.\end{aligned}$$

Ricordiamo che, posto

$$s_M = \sum_{k=1}^{M}(1-q)^{k-1}q\,,$$

si ha

$$(1-q)s_M = \sum_{k=1}^{M}(1-q)^k q = \sum_{k=2}^{M+1}(1-q)^{k-1}q = s_M + (1-q)^M q - q$$

da cui

$$s_M = 1 - (1-q)^M\,.$$

Esercizio 3. Sia (X_1, X_2) il vettore aleatorio con la seguente funzione di probabilità congiunta:

X_1 / X_2	-1	0	1
-1	a	$2a$	b
1	b	a	$2a$

1. Calcolare la varianza di X_2.
2. Determinare a e b tali che $\mathrm{Cov}(X_1, X_2) = 0$.
3. Per i valori a e b trovati nel punto 2., le variabili aleatorie X_1 e X_2 sono indipendenti?

Soluzione. 1. La varianza è data da

$$\mathrm{Var}(X_2) = \mathbb{E}(X_2^2) - \mathbb{E}(X_2)^2.$$

Calcoliamo quindi

$$\mathbb{E}(X_2^2) = (-1)^2 \cdot \mathbb{P}(\{X_2 = -1\}) + 1^2 \cdot \mathbb{P}(\{X_2 = 1\}) = \mathbb{P}(\{X_2 \in \{-1, 1\}\}) = 1$$

e

$$\mathbb{E}(X_2) = -1 \cdot \mathbb{P}(\{X_2 = -1\}) + 1 \cdot \mathbb{P}(\{X_2 = 1\}) = -(3a + b) + (3a + b) = 0\,.$$

In conclusione $\mathrm{Var}(X_2) = 1$.

2. Ricordiamo che la covarianza è data da

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)\,.$$

Il range di X_1X_2 è l'insieme $\{-1, 0, 1\}$, quindi

$$\begin{aligned}\mathbb{E}(X_1X_2) &= -1 \cdot \mathbb{P}(\{X_1X_2 = -1\}) + 0 \cdot \mathbb{P}(\{X_1X_2 = 0\}) + 1 \cdot \mathbb{P}(\{X_1X_2 = 1\}) \\ &= -\big(\mathbb{P}(\{X_1 = -1, X_2 = 1\}) + \mathbb{P}(\{X_1 = 1, X_2 = -1\})\big) \\ &\quad + \big(\mathbb{P}(\{X_1 = 1, X_2 = 1\}) + \mathbb{P}(\{X_1 = -1, X_2 = -1\})\big) \\ &= -2b + 3a\,.\end{aligned}$$

Non è importante calcolare il valore atteso di X_1 poiché abbiamo già trovato nel punto 1. che $\mathbb{E}(X_2) = 0$.

Imponiamo

$$-2b + 3a = 0 \implies a = \frac{2}{3}b\,.$$

Sostituiamo nella relazione

$$6a + 2b = 1 \implies 4b + 2b = 1 \implies b = \frac{1}{6}\,.$$

Quindi $a = \frac{1}{9}$.

3. Calcoliamo, ad esempio,

$$\mathbb{P}(\{X_1 = 1\}) = b + 2a = \frac{1}{6} + \frac{2}{9} = \frac{7}{18}$$

$$\mathbb{P}(\{X_2 = 1\}) = b + 3a = \frac{1}{6} + \frac{3}{9} = \frac{7}{18} = \frac{1}{2}\,.$$

Poiché

$$\frac{2}{9} = \mathbb{P}(\{X_1 = 1, X_2 = 1\}) \neq \mathbb{P}(\{X_1 = 1\})\mathbb{P}(\{X_2 = 1\}) = \frac{7}{36}$$

le due variabili aleatorie non sono indipendenti.

Esercizio 4. Il contenuto di catrame (in mg) in sigarette prodotte da un'azienda si può supporre distribuito con legge normale. Dalle misurazione di 15 campioni di sigarette si ottengono i seguenti risultati:

6.9 7.4 7.3 6.6 7.0 6.7 7.1 6.2 7.2 6.6 6.9 6.5 7.2 7.7 7.5 .

1. Determinare un intervallo di confidenza al 95% per la media del contenuto di catrame calcolata sui dati.

2. La realizzazione di un intervallo di confidenza al 97% sugli stessi dati (calcolata con lo stesso metodo del punto 1.) sarebbe più più o meno grande dell'intervallo trovato nel punto 1.? Motivare la risposta (N.B.: non è richiesto calcolare esplicitamente l'intervallo!)

Soluzione. 1. Abbiamo un campione casuale $X_1, \dots, X_n$, dove ciascuna X_i ha distribuzione normale $X_i \sim \mathcal{N}(\mu, \sigma^2)$, ma i parametri μ e σ^2 non sono noti.
Sia $\beta = 95\%$. Un intervallo di confidenza di livello β è $[U_n, V_n]$ dove gli estremi U_n e V_n sono variabili aleatorie tali che

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

Utilizzando la media campionaria e la varianza campionaria

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i , \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_i)^2$$

otteniamo che

$$\beta = \mathbb{P}(\{U_n \le \mu \le V_n\}) = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - V_n}{S_n/\sqrt{n}} \le \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \le \frac{\overline{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) .$$

La statistica

$$T_{n-1} := \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

ha distribuzione t-Student con $n-1$ gradi di libertà, poiché le X_i hanno distribuzione normale. Segue che

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

da cui, ponendo $\alpha = 1 - \beta = 5\%$,

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

Decidiamo di equipartire α, ovvero

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

Queste uguaglianze si ottengono scegliendo

$$\frac{\overline{X}_n - U_n}{S_n/\sqrt{n}} = t_{n-1,\alpha/2} \quad \frac{\overline{X}_n - V_n}{S_n/\sqrt{n}} = -t_{n-1,\alpha/2} ,$$

dove $t_{n-1,\alpha/2}$ è il quantile della t-Student con $n-1$ gradi di libertà tale che

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

Possiamo calcolarlo nel nostro caso utilizzando le tavole:

$$t_{n-1,\alpha/2} = t_{14,0.025} \simeq 2.145 .$$

Quindi le variabili aleatorie

$$U_n = \overline{X}_n - \frac{S_n}{\sqrt{n}}2.145 , \quad V_n = \overline{X}_n + \frac{S_n}{\sqrt{n}}2.145$$

costituiscono gli estremi di un intervallo di confidenza al 95%.
Calcoliamo la realizzazione di questo intervallo di confidenza sui dati. La realizzazione della media campionaria è

$$\overline{x}_{15} = \frac{1}{15}(6.9 + \cdots + 7.5) \simeq 6.9866666667$$

mentre la realizzazione della deviazione standard campionaria è

$$s_{15} = \sqrt{\frac{1}{14}(6.9^2 + \cdots + 7.5^2 - 15 \cdot \overline{x}_{15}^2)} \simeq \sqrt{\frac{1}{14}(734.6 - 732.2026666737)} \simeq 0.4138092492\,.$$

La realizzazione degli estremi è

$$u_{15} = \overline{x}_{15} - \frac{s_{15}}{\sqrt{15}}2.145 \simeq 6.7574839514 \simeq 6.8\,, \quad v_{15} = \overline{x}_{15} + \frac{s_{15}}{\sqrt{15}}2.145 \simeq 7.215849382 \simeq 7.2\,.$$

2. L'unica differenza nel calcolo della realizzazione dell'intervallo di confidenza al 97% sarebbe nel quantile della t-Student da utilizzare, che sarebbe $t_{14,0.015}$, dove $0.015 = 1.5\% = 3\%/2 = (1 - 97\%)/2$. Si ha che

$$\mathbb{P}(\{T_{14} > t_{14,0.015}\}) = 1.5\% < 2.5\% = \mathbb{P}(\{T_{14} > t_{14,0.025}\})\,.$$

Osservando il grafico della t-Student, ci rendiamo conto del fatto che che $t_{14,0.015} > t_{14,0.025}$. Quindi l'intervallo di confidenza al 97% è più grande di quello al 95%. In effetti, se la probabilità che gli estremi dell'intervallo contengano il parametro è più grande, l'intervallo deve essere più ampio.

Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni 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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: luglio 2022
Corso di studi: _____ Data: 18/07/2022

Tempo massimo: 2 ore.

Esercizio 1. I risultati dei test di adesione a trazione su 22 provini di lega U-700 mostrano i seguenti carichi di rottura (in megapascal):

23.1	10.1	15.4	18.5	11.4	14.1	19.5	8.8	14.9	7.5	7.9
12.7	15.4	15.4	11.9	11.4	17.6	16.7	15.8	13.6	11.9	11.4

1. Determinare i quartili dei dati.
2. Determinare eventuali dati anomali o sospetti.
3. Tracciare un box-plot.

Soluzione. 1. Ordiniamo i dati:

7.5	7.9	8.8	10.1	11.4	11.4	11.4	11.9	11.9	12.7	13.6
14.1	14.9	15.4	15.4	15.4	15.8	16.7	17.6	18.5	19.5	23.1

Denotiamo con $x_1, \ldots, x_{22}$ i dati ordinati. L'ampiezza del campione è $n = 22$.
Calcolo di Q_1. Per trovare il primo quartile calcoliamo $\frac{n+1}{4} = \frac{23}{4} = 5 + \frac{3}{4} = 5 + 0.75$. Quindi

$$Q_1 = (1 - 0.75)x_5 + 0.75x_6 = 11.4\,.$$

Calcolo di Q_2. Per trovare il secondo quartile calcoliamo $(n+1)\frac{2}{4} = \frac{23}{2} = 11 + \frac{1}{2} = 11 + 0.5$. Quindi

$$Q_2 = 0.5x_{11} + 0.5x_{12} = 13.85\,.$$

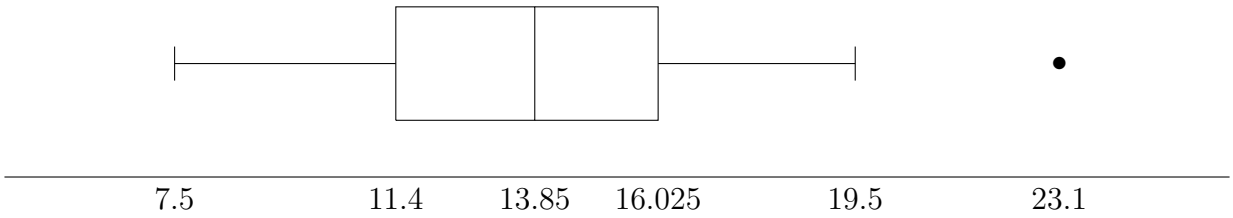
Calcolo di Q_3. Per trovare il terzo quartile calcoliamo $(n+1)\frac{3}{4} = 23\frac{3}{4} = 17 + \frac{1}{4} = 17 + 0.25$. Quindi

$$Q_3 = (1 - 0.25)x_{17} + 0.25x_{18} = 16.025\,.$$

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

$$IQR = Q_3 - Q_1 = 4.625\,.$$

I dati anomali sono più grandi di $Q_3 + 3IQR = 29.9$ o più piccoli di $Q_1 - 3IQR = -2.475$. Quindi non ci sono dati anomali.
I dati sospetti cadono tra $Q_3 + 1.5IQR = 22.9625$ e $Q_3 + 3IQR = 29.9$ oppure tra $Q_1 - 3IQR = -2.475$ e $Q_1 - 1.5IQR = 4.4625$. Quindi 23.1 è un dato anomalo.
3. Segue il box-plot:



Esercizio 2. Sei un ingegnere gestionale e fai parte di un gruppo di esperti selezionati per formare una commissione giudicatrice. Oltre a te ci sono: 2 ingegneri gestionali, 6 ingegneri elettrici, 4 ingegneri civili. Tra gli ingegneri elettrici ci sono tua sorella e tuo fratello.

1. Si deve formare una commissione composta da 2 ingegneri gestionali, 4 ingegneri elettrici, 2 ingegneri civili scegliendo in modo casuale (e uniformemente rispetto alle possibili commissioni realizzabili) tra i possibili esperti. Qual è la probabilità che tu non venga selezionato?

2. È stata selezionata la commissione come nel punto 1. Ti hanno detto che il tuo cognome (che è anche quello di tua sorella e tuo fratello) compare esattamente due volte, ma non sai di preciso chi di voi tre è stato selezionato. Qual è la probabilità che tu sia stato selezionato?

Soluzione. 1. Lo spazio degli eventi elementari è dato da

$$\begin{aligned}\Omega = \{(\{\omega_1, \omega_2\}, \{\omega_3, \omega_4, \omega_5, \omega_6\}, \{\omega_7, \omega_8\}) \ : \\ \omega_1, \omega_2 \in \{1,2,3\},\ \omega_3, \omega_4, \omega_5, \omega_6 \in \{4,5,6,7,8,9\},\ \omega_7, \omega_8 \in \{10,11,12,13\}, \\ \omega_i \neq \omega_j \text{ per } i \neq j\},\end{aligned}$$

dove abbiamo indicato con $\{1,2,3\}$ gestionali, $\{4,5,6,7,8,9\}$ elettrici, $\{10,11,12,13\}$ civili. Nella formazione di una commissione si scelgono 2 da 3 elementi (gestionali), 4 da 6 elementi (elettrici), 2 da 4 elementi (civili), quindi le possibilità sono

$$\#\Omega = \binom{3}{2}\binom{6}{4}\binom{4}{2}.$$

L'evento "vengo selezionato/a nella commissione" è, indicando con 1 l'elemento corrispondente a me,

$$\begin{aligned}A = \{(\{1, \omega_2\}, \{\omega_3, \omega_4, \omega_5, \omega_6\}, \{\omega_7, \omega_8\}) \ : \\ \omega_2 \in \{2,3\},\ \omega_3, \omega_4, \omega_5, \omega_6 \in \{4,5,6,7,8,9\},\ \omega_7, \omega_8 \in \{10,11,12,13\},\ \omega_i \neq \omega_j \text{ per } i \neq j\}\end{aligned}$$

che è composto da

$$\#A = \binom{2}{1}\binom{6}{4}\binom{4}{2}$$

elementi, poiché per i gestionali si sceglie solamente tra 2.
La probabilità di non essere selezionato è

$$\mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A) = 1 - \frac{\#A}{\#\Omega} = 1 - \frac{\binom{2}{1}}{\binom{3}{2}} = 1 - \frac{2}{3} = \frac{1}{3}\,.$$

2. Chiamiamo 4 mia sorella e 5 mio fratello. Consideriamo l'evento $B =$ "esattamente due dei tre della famiglia sono selezionati". Questo evento si spezza nell'unione dei tre eventi $B_1 =$ "io e mia sorella veniamo selezionati (mio fratello no)", $B_2 =$ "io e mio fratello veniamo selezionati (mia sorella no)", $B_3 =$ "mia sorella e mio fratello vengono selezionati (io no)". Utilizzando la notazione degli eventi elementari otteniamo

$$\begin{aligned}
B &= B_1 \cup B_2 \cup B_3 \\
&= \{(\{1,\omega_2\},\{4,\omega_4,\omega_5,\omega_6\},\{\omega_7,\omega_8\}) \ : \\
&\qquad \omega_2 \in \{2,3\},\ \omega_4,\omega_5,\omega_6 \in \{6,7,8,9\},\ \omega_7,\omega_8 \in \{10,11,12,13\},\ \omega_i \neq \omega_j \text{ per } i \neq j\} \cup \\
&\quad \cup \{(\{1,\omega_2\},\{5,\omega_4,\omega_5,\omega_6\},\{\omega_7,\omega_8\}) \ : \\
&\qquad \omega_2 \in \{2,3\},\ \omega_4,\omega_5,\omega_6 \in \{6,7,8,9\},\ \omega_7,\omega_8 \in \{10,11,12,13\},\ \omega_i \neq \omega_j \text{ per } i \neq j\} \cup \\
&\quad \cup \{(\{\omega_1,\omega_2\},\{4,5,\omega_5,\omega_6\},\{\omega_7,\omega_8\}) \ : \\
&\qquad \omega_1,\omega_2 \in \{2,3\},\ \omega_5,\omega_6 \in \{6,7,8,9\},\ \omega_7,\omega_8 \in \{10,11,12,13\},\ \omega_i \neq \omega_j \text{ per } i \neq j\}.
\end{aligned}$$

L'evento B_1 è composto da un numero di elementi dato dalle scelte di ω_2 per le scelte di $\omega_4,\omega_5,\omega_6$ per le scelte di ω_7,ω_8, quindi è composto da

$$\#B_1 = \binom{2}{1}\binom{4}{3}\binom{4}{2}$$

elementi. Analogamente contiamo il numero di elementi per gli altri eventi:

$$\#B = \#B_1 + \#B_2 + \#B_3 = \binom{2}{1}\binom{4}{3}\binom{4}{2} + \binom{2}{1}\binom{4}{3}\binom{4}{2} + \binom{2}{2}\binom{4}{2}\binom{4}{2}$$

elementi.
Dobbiamo calcolare

$$\begin{aligned}
\mathbb{P}(A|B) &= \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B_1)+\mathbb{P}(B_2)}{\mathbb{P}(B)} = \frac{\frac{\#B_1}{\#\Omega}+\frac{\#B_2}{\#\Omega}}{\frac{\#B}{\#\Omega}} = \frac{\#B_1+\#B_2}{\#B} \\
&= \frac{\binom{2}{1}\binom{4}{3}\binom{4}{2}+\binom{2}{1}\binom{4}{3}\binom{4}{2}}{\binom{2}{1}\binom{4}{3}\binom{4}{2}+\binom{2}{1}\binom{4}{3}\binom{4}{2}+\binom{2}{2}\binom{4}{2}\binom{4}{2}} = \frac{2\binom{2}{1}\binom{4}{3}}{2\binom{2}{1}\binom{4}{3}+\binom{2}{2}\binom{4}{2}} = \frac{2\cdot 2\cdot 4}{2\cdot 2\cdot 4+1\cdot 6} = \frac{16}{22} = \frac{8}{11} \\
&\simeq 72.72\%.
\end{aligned}$$

Osserviamo che è maggiore di $\frac{2}{3}$ (la probabilità di essere scelto senza sapere nulla).

Esercizio 3. Devi aprire un conto alla poste. Prendi il biglietto e vedi che ci sono tre sportelli. A servire lo sportello 1 c'è una persona molto motivata ed efficiente, a servire lo sportello 2 una persona normale, a servire lo sportello 3 una persona evidentemente frustrata e con poca voglia di lavorare. Il tempo che impiegherai per concludere l'apertura del conto è distribuito con una legge esponenziale, ma il tempo medio che impiegherai dipende da quale sportello ti capita: allo sportello 1 la media sarebbe 10 minuti; allo sportello 2, 15 minuti; allo sportello 3, 30 minuti. Nell'attesa ti accorgi che il 65% delle persone viene servito allo sportello 1, il 25% dallo sportello 2, il 10% dallo sportello 3.

1. Qual è la probabilità che impiegherai più di 30 minuti a concludere l'operazione?

2. Finita l'operazione, chiami una tua amica e le dici che hai impiegato più di 30 minuti ad aprire un conto! Lei è stata a quella filiale e conosce le tre persone addette agli sportelli. Con che probabilità è pronta a scommettere che sei stato servito dallo sportello 3?

Soluzione. 1. Consideriamo due variabili aleatorie: X con range $\{1,2,3\}$ che indica lo sportello che ci capita e Y con range $(0,+\infty)$ che indica il tempo in minuti necessario per l'operazione allo sportello. La variabile X è distribuita nel seguente modo:

$$\mathbb{P}(\{X=1\})=65\% \quad \mathbb{P}(\{X=2\})=25\% \quad \mathbb{P}(\{X=3\})=10\%.$$

La distribuzione di Y dipende dalla realizzazione di X. Ricordando che la media di una variabile aleatoria con legge esponenziale è il reciproco del parametro, otteniamo che

$$\mathbb{P}(\{Y\geq t\}|\{X=1\})=\int_t^{+\infty}\frac{1}{10}e^{-\frac{1}{10}y}\,\mathrm{d}y=\Big[-e^{-\frac{1}{10}x}\Big]_t^{+\infty}=e^{-\frac{1}{10}t},$$

$$\mathbb{P}(\{Y\geq t\}|\{X=2\})=\int_t^{+\infty}\frac{1}{15}e^{-\frac{1}{15}y}\,\mathrm{d}y=e^{-\frac{1}{15}t},$$

$$\mathbb{P}(\{Y\geq t\}|\{X=3\})=\int_t^{+\infty}\frac{1}{30}e^{-\frac{1}{30}y}\,\mathrm{d}y=e^{-\frac{1}{30}t}.$$

Possiamo utilizzare il teorema della probabilità totale per calcolare

$$\begin{aligned}\mathbb{P}(\{Y\geq 30\})&=\mathbb{P}(\{Y\geq 30\}|\{X=1\})\mathbb{P}(\{X=1\})+\mathbb{P}(\{Y\geq 30\}|\{X=2\})\mathbb{P}(\{X=2\})\\&\quad+\mathbb{P}(\{Y\geq 30\}|\{X=3\})\mathbb{P}(\{X=3\})\\&=e^{-\frac{30}{10}}65\%+e^{-\frac{30}{15}}25\%+e^{-\frac{30}{30}}10\%=10.29\%.\end{aligned}$$

2. Calcoliamo con il Teorema di Bayes

$$\mathbb{P}(\{X=3\}|\{Y\geq 30\})=\frac{\mathbb{P}(\{Y\geq 30\}|\{X=3\})\mathbb{P}(\{X=3\})}{\mathbb{P}(\{Y\geq 30\})}=\frac{e^{-\frac{30}{30}}10\%}{10.29\%}\sim 35.75\%$$

Esercizio 4. Una riempitrice automatica viene utilizzata per riempire dosatori da 100 ml con gel igienizzante. Viene misurata la differenza tra il volume di riferimento 100 ml e il volume di riempimento effettivo in un campione casuale di dosatori:

$$-0.91 \quad 0.51 \quad 0.85 \quad 0.10 \quad -0.56 \quad 1.10 \quad 0.38 \quad 0.26 \quad 0.30 \quad -0.67 \quad 0.03 \quad -0.31 \quad 0.59$$

(ad esempio, il dato -0.91 indica che il dosatore è stato riempito con 100.91 ml di gel, il dato 0.51 indica che il dosatore è stato riempito con 99.49 ml di gel). Se la deviazione standard del volume di riempimento è superiore a 0.16 ml, la macchina riempitrice deve essere tarata nuovamente. I dati sono significativi all'1% per concludere che si deve tarare la macchina riempitrice? E al 5%? Si assuma che la popolazione sia distribuita con legge normale.
(N.B.: Ricavare le formule!)

Soluzione. (Nota: la deviazione standard 0.16 ml è troppo piccola per un errore di trascrizione. Ad ogni modo i numeri non influiscono sullo svolgimento dell'esercizio.) La differenza tra 100 ml e il volume di riempimento effettivo è una variabile aleatoria con legge normale. Abbiamo

un campione casuale $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma)$ dove $n = 13$ e μ e σ non sono note. Il problema può essere posto come un test d'ipotesi (esattamente come fatto per l'Esercizio 10 risolto nella Lezione 23):

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

dove $\sigma_0 = 0.16$. Il test è impostato così perché la domanda è se i dati sono abbastanza significativi da rifiutare l'ipotesi nulla.
Ricordiamo che il livello di significatività α di un test è la probabilità di commettere un errore del I tipo, ovvero di rifiutare l'ipotesi H_0 quando questa è vera. Allora assumiamo che H_0 sia vera: $\sigma^2 \leq \sigma_0^2$. La regione critica per il test è della forma

$$R_C = \{(x_1, \dots, x_n) \in R(X_1, \dots, X_n) : s_n^2 > c\sigma_0^2\}$$

dove $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)^2$ è la varianza campionaria calcolata sui dati e $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ è la media campionaria calcolata sui dati. (N.B.: questo è coerente con la domanda del problema: la macchina va tarata se la varianza campionaria calcolata sui dati è esageratamente più grande di quella che vorremmo! Controllare sempre se la regione critica ha senso, altrimenti correggere il test d'ipotesi)
Osserviamo che se $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ è la varianza campionaria, allora non è detto che $\frac{(n-1)S_n^2}{\sigma_0^2}$ sia distribuita come una chi-quadro, poiché σ_0^2 non è la vera varianza della popolazione. Tuttavia, osserviamo che l'ipotesi nulla $\sigma^2 \leq \sigma_0^2$ implica che

$$S_n^2 > c\sigma_0^2 \implies S_n^2 > c\sigma^2$$

e quindi

$$\mathbb{P}(\{(X_1, \dots, X_n) \in R_C\}) = \mathbb{P}(\{S_n^2 > c\sigma_0^2\}) \leq \mathbb{P}(\{S_n^2 > c\sigma^2\}) = \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{\sigma^2} > (n-1)c\Big\}\Big)\,.$$

La statistica $\Xi_{n-1} = \frac{(n-1)S_n^2}{\sigma^2}$ è distribuita come una chi-quadro con $n-1$ gradi di libertà, poiché la popolazione è normale e σ^2 è la varianza della popolazione. Pertanto scegliendo $(n-1)c = \chi^2_{n-1,\alpha}$, dove $\chi^2_{n-1,\alpha}$ è il quantile α di una chi-quadro con $n-1$ gradi di libertà, si ha che

$$\mathbb{P}(\{S_n^2 > c\sigma_0^2\}) \leq \mathbb{P}(\{\Xi_{n-1} > \chi^2_{n-1,\alpha}\}) = \alpha\,,$$

cioè la probabilità di commettere un errore del I tipo è più piccola di α.
Utilizziamo le tavole per calcolare il quantile

$$\chi^2_{n-1,\alpha} = \chi^2_{12,0.01} = 26.217\,.$$

L'ipotesi H_0 viene rifiutata a favore di H_1 se i dati ricadono nella regione critica, ovvero

$$s_n^2 > c\sigma_0^2 = \frac{\chi^2_{n-1,\alpha}}{(n-1)}\sigma_0^2 = \frac{26.217}{12}0.0256 \simeq 0.0559.$$

Calcoliamo la realizzazione della varianza campionaria sui dati:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{13}(-0.91 + \dots + 0.59) \simeq 0.13\,.$$

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)^2 = \frac{1}{12}(0.91^2 + \dots + 0.59^2 - 13 \cdot 0.13^2) \simeq 0.36\,.$$

(N.B.: 0.36 è ben più grande di 0.16^2, quindi ha senso un test unilaterale con $H_1 : \sigma^2 > \sigma_0^2$.) Segue che i dati ricadono nella regione di rifiuto e pertanto l'ipotesi nulla è rifiutata con significatività 1%. A maggior ragione viene rifiutata con significatività 5% poiché se aumenta la probabilità di ricadere nella regione critica, la regione critica diventa più grande. In effetti,

$$\frac{\chi^2_{12,0.05}}{12}\sigma_0^2 \simeq \frac{21.026}{12}0.0256 \simeq 0.044\,.$$

Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni 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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: luglio 2022
Corso di studi: _____ Data: 18/07/2022

Tempo massimo: 2 ore.

Esercizio 1. Un gruppo di topi di 5 settimane viene sottoposto a una dose di radiazione di 300 rad. La seguente tabella riporta le frequenze assolute dei giorni di vita dei topi suddivisi in intervalli di classi:

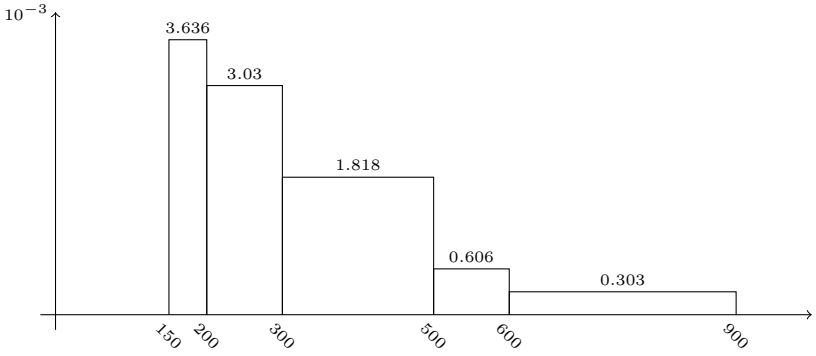
intervallo	frequenza
$[150, 200)$	6
$[200, 300)$	10
$[300, 500)$	12
$[500, 600)$	2
$[600, 900)$	3

1. Rappresentare un istogramma delle densità di frequenze relative.
2. Determinare la classe modale.
3. Calcolare un'approssimazione della media e della deviazione standard dei dati.
4. Calcolare un'approssimazione della mediana dei dati.

Soluzione. 1. Denotiamo con $I_1, \dots, I_5$ gli intervalli, $f_1, \dots, f_5$ le frequenze assolute. Abbiamo che $n = f_1 + \cdots + f_5 = 6 + 10 + 12 + 2 + 3 = 33$. Ricordiamo che le frequenze relative sono date da $p_j = f_j/n$ e le densità di frequenze relative da $d_j = p_j/|I_j|$ dove $|I_j| = b_j - a_j$ se $I_j = [a_j, b_j)$. Completiamo la tabella (scriviamo anche le frequenze assolute cumulate per il punto 4.):

intervallo	f. assolute	f. relative	densità f. rel.	f. cumulate
$[150, 200)$	6	18.18%	$3.636 \cdot 10^{-3}$	6
$[200, 300)$	10	30.30%	$3.03 \cdot 10^{-3}$	16
$[300, 500)$	12	36.36%	$1.818 \cdot 10^{-3}$	28
$[500, 600)$	2	6.06%	$6.06 \cdot 10^{-4}$	30
$[600, 900)$	3	9.1%	$3.03 \cdot 10^{-4}$	33

(nell'ultima percentuale abbiamo approssimato a 9.1 in modo che $18.18+30.30+36.36+6.06+9.1=100$.) Rappresentiamo l'istogramma:



2. La classe modale è l'intervallo $[150, 200)$ poiché è l'intervallo con la maggiore densità di frequenza relativa.
3. Ricordando che la media calcolata su un campione di dati $x_1, \dots, x_n$ è

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

possiamo riscrivere la formula in termini dei valori assunti $v_1, \dots, v_k$ utilizzando le frequenze assolute $f_1, \dots, f_k$ e le frequenze relative $p_1, \dots, p_k$

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{k} f_j v_j = \sum_{j=1}^{k} p_j v_j \,.$$

Per approssimare la media sostituiamo ai valori v_j i valori centrali $\tilde{v}_j$ degli intervalli e le frequenze relative:

$$\overline{x} \simeq 18.18\% \cdot 175 + 30.30\% \cdot 250 + 36.36\% \cdot 400 + 6.06\% \cdot 550 + 9.1\% \cdot 750 = 354.6\,.$$

Ricordiamo che la varianza calcolata su un campione di dati $x_1, \dots, x_n$ può essere calcolata mediante la formula:

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

Possiamo riscrivere la formula in termini dei valori assunti $v_1, \dots, v_k$ utilizzando le frequenze assolute $f_1, \dots, f_k$ e le frequenze relative $p_1, \dots, p_k$:

$$s^2 = \frac{1}{n-1}\Big(\sum_{j=1}^{k} f_j v_j^2 - n\overline{x}^2\Big) = \frac{n}{n-1}\Big(\sum_{j=1}^{k} p_j v_j^2 - \overline{x}^2\Big).$$

Per approssimare la varianza sostituiamo ai valori v_j i valori centrali $\tilde{v}_j$ degli intervalli e le frequenze relative e la media approssimata calcolata nel punto precedente:

$$s^2 \simeq \frac{33}{32}\Big(18.18\%\cdot 175^2 + 30.30\%\cdot 250^2 + 36.36\%\cdot 400^2 + 6.06\%\cdot 550^2 + 9.1\%\cdot 750^2 - 354.6^2\Big) = 27286$$

da cui segue che la deviazione standard è approssimata da

$$s \simeq 165.184.$$

4. Per calcolare un'approssimazione della mediana utilizziamo le frequenze cumulate $F_1, \dots, F_k$. La mediana divide l'insieme il campione di dati in due parti, quindi calcoliamo $\frac{n}{2} = \frac{33}{2} = 16.5$ e osserviamo che per l'intervallo $I_3 = [a_3, b_3) = [300, 500)$ si ha che

$$F_2 = 16 < 16.5 < 28 = F_3 \,.$$

La mediana è allora approssimata da

$$Q_2 \simeq a_3 + \lambda_3(b_3 - a_3), \quad \lambda_3 = \frac{\frac{n}{2} - F_2}{F_3 - F_2} = \frac{16.5 - 16}{12} = \frac{1}{24}$$

quindi

$$Q_2 \simeq 300 + \frac{1}{24}200 \simeq 308.33.$$

Esercizio 2. Un'azienda produce grandi numeri di mobili montabili. In un particolare mobile ci sono due pezzi (che chiameremo A e B) che possono risultare difettosi. In media vengono prodotti ogni giorno 1 pezzo A difettoso e (indipendentemente) 2 pezzi B difettosi. Per entrambi i tipi, il numero di pezzi difettosi è distribuito con una legge di Poisson.

1. Qual è la probabilità che vengano prodotti (strettamente) più di 4 pezzi A difettosi in 5 giorni? (Si assumano i difetti nei diversi giorni indipendenti)

2. Sappiamo che in 5 giorni sono stati prodotti in tutto 12 pezzi difettosi (contando sia tipo A che B). Qual è la probabilità che al più 3 pezzi A siano difettosi?

Soluzione. 1. Il numero di pezzi A difettosi in un giorno è distribuito come una variabile aleatoria $X_1 \sim P(\lambda)$. Poiché $\mathbb{E}(X_1) = \lambda$ e la traccia ci dice che il numero medio di pezzi A difettosi in un giorno è 1, abbiamo che $\lambda = 1$. Quindi

$$\mathbb{P}(\{X_1 = k\}) = \frac{e^{-1}}{k!}.$$

Il numero di pezzi A difettosi in 5 giorni è la somma di 5 variabili aleatorie di Poisson $X_i \sim P(1)$

$$X = X_1 + X_2 + X_3 + X_4 + X_5 \sim P(5\lambda) = P(5)\,,$$

poiché i difetti nei diversi giorni sono indipendenti e la somma di Poisson indipendenti è una Poisson con il parametro dato dalla somma dei parametri. Quindi

$$\mathbb{P}(\{X = k\}) = e^{-5}\frac{5^k}{k!}\,.$$

Possiamo calcolare

$$\begin{aligned}\mathbb{P}(\{X > 4\}) &= 1 - \mathbb{P}(\{X \le 4\}) \\ &= 1 - \mathbb{P}(\{X = 0\}) - \mathbb{P}(\{X = 1\}) - \mathbb{P}(\{X = 2\}) - \mathbb{P}(\{X = 3\}) - \mathbb{P}(\{X = 4\}) \\ &= 1 - e^{-5} - e^{-5}5 - e^{-5}\frac{5^2}{2} - e^{-5}\frac{5^3}{3!} - e^{-5}\frac{5^4}{4!} \simeq 55.95\%.\end{aligned}$$

2. Il numero di pezzi B difettosi in un giorno è distribuito come una variabile aleatoria $Y_1 \sim P(\mu)$. Poiché $\mathbb{E}(Y_1) = \mu$ e la traccia ci dice che il numero medio di pezzi B difettosi in un giorno

è 2, abbiamo che $\mu = 2$. Il numero di pezzi B difettosi in 5 giorni è la somma di 5 variabili aleatorie di Poisson $Y_i \sim P(2)$

$$Y = Y_1 + Y_2 + Y_3 + Y_4 + Y_5 \sim P(5\mu) = P(10)\,,$$

poiché i difetti nei diversi giorni sono indipendenti e la somma di Poisson indipendenti è una Poisson con il parametro dato dalla somma dei parametri. Quindi

$$\mathbb{P}(\{Y = k\}) = e^{-10}\frac{10^k}{k!}\,.$$

Il numero di pezzi difettosi sia di tipo A che B è la somma $X + Y$. Queste sono due variabili aleatorie di Poisson indipendenti (lo dice la traccia), quindi $X + Y \sim P(5 + 10) = P(15)$ e

$$\mathbb{P}(\{X + Y = k\}) = e^{-15}\frac{15^k}{k!}.$$

La traccia chiede di calcolare

$$\begin{aligned}\mathbb{P}(\{X \le 3\}|\{X + Y = 12\}) &= \frac{\mathbb{P}(\{X \le 3\} \cap \{X + Y = 12\})}{\mathbb{P}(\{X + Y = 12\})} \\ &= \frac{\mathbb{P}(\{X = 0\} \cap \{Y = 12\})}{\mathbb{P}(\{X + Y = 12\})} + \frac{\mathbb{P}(\{X = 1\} \cap \{Y = 11\})}{\mathbb{P}(\{X + Y = 12\})} \\ &\quad + \frac{\mathbb{P}(\{X = 2\} \cap \{Y = 10\})}{\mathbb{P}(\{X + Y = 12\})} + \frac{\mathbb{P}(\{X = 3\} \cap \{Y = 9\})}{\mathbb{P}(\{X + Y = 12\})}.\end{aligned}$$

Utilizzando l'indipendenza di X e Y:

$$\begin{aligned}&\mathbb{P}(\{X \le 3\}|\{X + Y = 12\}) \\ &= \frac{\mathbb{P}(\{X = 0\})\mathbb{P}(\{Y = 12\})}{\mathbb{P}(\{X + Y = 12\})} + \frac{\mathbb{P}(\{X = 1\})\mathbb{P}(\{Y = 11\})}{\mathbb{P}(\{X + Y = 12\})} \\ &\quad + \frac{\mathbb{P}(\{X = 2\})\mathbb{P}(\{Y = 10\})}{\mathbb{P}(\{X + Y = 12\})} + \frac{\mathbb{P}(\{X = 3\})\mathbb{P}(\{Y = 9\})}{\mathbb{P}(\{X + Y = 12\})} \\ &= \frac{e^{-5}e^{-10}\frac{10^{12}}{12!}}{e^{-15}\frac{15^{12}}{12!}} + \frac{e^{-5}5e^{-10}\frac{10^{11}}{11!}}{e^{-15}\frac{15^{12}}{12!}} + \frac{e^{-5}\frac{5^2}{2!}e^{-10}\frac{10^{10}}{10!}}{e^{-15}\frac{15^{12}}{12!}} + \frac{e^{-5}\frac{5^3}{3!}e^{-10}\frac{10^9}{9!}}{e^{-15}\frac{15^12}{12!}} \\ &= \frac{1}{\frac{15^{12}}{12!}}\Big(\frac{10^{12}}{12!} + 5\frac{10^{11}}{11!} + \frac{5^2}{2!}\frac{10^{10}}{10!} + \frac{5^3}{3!}\frac{10^9}{9!}\Big) \simeq 39.31\%.\end{aligned}$$

Esercizio 3. Sono le 8:30 e sei allo sportello della tua banca per sbrigare una pratica. In media la pratica dura 10 minuti, e la durata in minuti è distribuita come una legge esponenziale. Dopo aver sbrigato la pratica in banca devi andare a lavoro prendendo un bus, che però ha un tempo di arrivo alla fermata (proprio all'uscita della banca) incerto. L'orario di arrivo del bus ha distribuzione uniforme, in media arriva alle 8:40, con una deviazione standard di 10 min.

1. Con che probabilità si verifica il seguente evento: finirai la pratica dopo le 8:50 e il bus arriverà prima delle 8:50?

2. Hai finito la pratica, corri fuori e scopri che il bus è già passato. Non hai guardato l'orologio, quindi non sai quanto tempo è durata la pratica e a che ora è passato il bus. Constatando che non sei riuscito a prendere il bus, è più probabile che la pratica sia durata meno di 10 min oppure che il bus sia arrivato prima delle 8:40? Motivare la risposta. (N.B.: non è richiesto il calcolo esplicito delle due probabilità!)

Soluzione. 1. Denotiamo con $X \sim \mathrm{Exp}(\lambda)$ la durata in minuti della pratica allo sportello della banca. Sappiamo che $\mathbb{E}(X) = \frac{1}{\lambda} = 10$, quindi $\lambda = \frac{1}{10}$ e quindi

$$\mathbb{P}(\{X \geq t\}) = \int_t^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x} \, \mathrm{d}x = \Big[-e^{-\frac{1}{10}x} \Big]_t^{+\infty} = e^{-\frac{1}{10}t}.$$

Denotiamo con Y il minuto di arrivo del bus. Sappiamo che $Y \sim \mathrm{U}(a,b)$ per qualche intervallo $[a,b]$. Calcoliamo a, b utilizzando le informazioni sulla media e sulla varianza. Ricordiamo che

$$\mathbb{E}(Y) = \int_{-\infty}^{+\infty} y f_Y(y) \, \mathrm{d}y = \frac{1}{b-a} \int_a^b y \, \mathrm{d}y = \frac{1}{b-a} \Big[\frac{y^2}{2} \Big]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

da cui

$$\frac{b+a}{2} = 40.$$

Inoltre

$$\begin{aligned} \mathrm{Var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 &= \int_{-\infty}^{+\infty} y^2 f_Y(y) \, \mathrm{d}y - \Big(\frac{b+a}{2} \Big)^2 = \frac{1}{b-a} \int_a^b y^2 \, \mathrm{d}y - \Big(\frac{b+a}{2} \Big)^2 \\ &= \frac{1}{b-a} \Big[\frac{y^3}{3} \Big]_a^b - \Big(\frac{b+a}{2} \Big)^2 = \frac{b^3 - a^3}{3(b-a)} - \frac{b^2 + 2ab + a^2}{4} = \\ &= \frac{b^2 + ab + a^2}{3} - \frac{b^2 + 2ab + a^2}{4} = \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12} \end{aligned}$$

da cui

$$\sqrt{\frac{(b-a)^2}{12}} = 10 \implies \frac{b-a}{2} = 5\sqrt{12}.$$

Segue che

$$a = 40 - 5\sqrt{12}\,, \quad b = 40 + 5\sqrt{12}$$

Calcoliamo, utilizzando il fatto che la durata della pratica e il tempo di arrivo del bus sono indipendenti,

$$\begin{aligned} \mathbb{P}(\{X > 20\} \cap \{Y < 50\}) &= \mathbb{P}(\{X > 20\}) \mathbb{P}(\{Y < 50\}) \\ &= \Big(\int_{20}^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x} \, \mathrm{d}x \Big) \Big(\frac{1}{10\sqrt{12}} \int_{40-5\sqrt{12}}^{50} \mathrm{d}y \Big) \\ &= e^{-\frac{20}{10}} \frac{50 - 40 + 5\sqrt{12}}{10\sqrt{12}} = e^{-2} \frac{1 + \sqrt{3}}{2\sqrt{3}} \simeq 10.67\% \end{aligned}$$

2. L'evento "perdo il bus" è $\{30 + X > Y\}$. Osserviamo che se $X < 10$ e $X + 30 > Y$, necessariamente $Y < 40$. Quindi

$$\{X < 10\} \cap \{X + 30 > Y\} = \{X < 10\} \cap \{X + 30 > Y\} \cap \{Y < 40\} \subset \{X + 30 > Y\} \cap \{Y < 40\}.$$

Usando la monotonia della probabilità possiamo allora stimare che

$$\begin{aligned} \mathbb{P}(\{X < 10\} | \{X + 30 > Y\}) &= \frac{\mathbb{P}(\{X < 10\} \cap \{X + 30 > Y\})}{\mathbb{P}(\{X + 30 > Y\})} \\ &\leq \frac{\mathbb{P}(\{Y < 40\} \cap \{X + 30 > Y\})}{\mathbb{P}(\{X + 30 > Y\})} = \mathbb{P}(\{Y < 40\} | \{X + 30 > Y\}) \end{aligned}$$

ovvero, è più probabile che il bus sia arrivato prima delle 8:40.

Osserviamo che

$$\mathbb{P}(\{X < 10\}) = 1 - \mathbb{P}(\{X \geq 10\}) = 1 - e^{-1} \simeq 63.21\%$$

e

$$\mathbb{P}(\{Y < 40\}) = \frac{1}{2},$$

cioè

$$\mathbb{P}(\{X < 10\}) \geq \mathbb{P}(\{Y < 40\}),$$

quindi il fatto di sapere che abbiamo perso il bus cambia la relazione tra le due probabilità.

Esercizio 4. Una riempitrice automatica viene utilizzata per riempire dosatori da 100 ml con gel igienizzante. Viene misurato il volume di riempimento effettivo in un campione casuale di 40 dosatori. La media campionaria e la deviazione standard campionaria calcolate sui dati risultano essere 99.90 ml e 0.55 ml rispettivamente. È possibile affermare con il 5% di significatività che in media la macchina immette meno di 100 ml di gel? Qual è il più piccolo livello di significatività per cui i dati permettono di affermare che la macchina immette meno di 100 ml di gel?
(N.B.: Ricavare le formule!)

Soluzione. Si tratta di un campione casuale $X_1, \ldots, X_n$ con $n = 40$ di cui non è nota la distribuzione. I dati forniscono le realizzazioni della media campionaria e della deviazione standard campionaria

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

che sono data rispettivamente da 99.90 ml e 0.55 ml.
La domanda della traccia (se i dati sono abbastanza significativi per concludere che in media la macchina immette meno di 100ml di gel) può essere impostata come un test d'ipotesi sulla media μ della popolazione

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

dove $\mu_0 = 100$. (È accettata come soluzione corretta anche la formulazione con $H_0 : \mu \geq \mu_0$, ma in quel caso bisogna stare attenti a usare μ come media vera della popolazione per l'utilizzo della statistica corretta!).
Il livello di significatività del test α è la probabilità di commettere un errore del I tipo, ovvero di rifiutare l'ipotesi nulla quando questa è vera. Supponiamo H_0 vera, ovvero $\mu = \mu_0$. La regione di rifiuto è della forma

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

dove $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. Allora

$$\alpha = \mathbb{P}(\{\overline{X}_n < \mu_0 - \delta\}) = \mathbb{P}(\{\overline{X}_n - \mu_0 < -\delta\}) = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} < -\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big).$$

(Se l'ipotesi nulla è invece $H_0 : \mu \geq \mu_0$, allora si deve usare il seguente fatto: se $\overline{X}_n < \mu_0 - \delta$ allora a maggior ragione $\overline{X}_n < \mu - \delta$ e quindi $\mathbb{P}(\{\overline{X}_n < \mu_0 - \delta\}) \leq \mathbb{P}(\{\overline{X}_n < \mu - \delta\})$ e si può procedere con i calcoli utilizzando la media vera della popolazione μ al posto di μ_0.)
Non è detto che la statistica $\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}}$ sia una t-Student con $n-1$ gradi di libertà, poiché non conosciamo la distribuzione della popolazione. Tuttavia il campione è numeroso ($n = 40 > 30$)

e possiamo quindi sfruttare un teorema di approssimazione. Nella fattispecie utilizziamo il Teorema di Slutsky (N.B.: a denominatore c'è S_n, non σ) per affermare che $\frac{\overline{X}_n-\mu_0}{S_n/\sqrt{n}}$ è approssimata da una variabile aleatoria Z distribuita con legge normale standard. Quindi

$$\mathbb{P}\Big(\Big\{\frac{\overline{X}_n-\mu_0}{S_n/\sqrt{n}}<-\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big)\simeq\mathbb{P}\Big(\Big\{Z<-\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big).$$

Scegliamo $\frac{\delta}{S_n/\sqrt{n}}=z_\alpha$, dove z_α è il quantile gaussiano. In questo modo

$$\mathbb{P}\Big(\Big\{Z<-\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big)=\mathbb{P}(\{Z<-z_\alpha\})=\alpha$$

e la condizione sulla significatività $\mathbb{P}(\{\overline{X}_n<\mu_0-\delta\})=\alpha$ è effettivamente verificata. In conclusione, la regione critica è

$$R_C=\Big\{(x_1,\dots,x_n)\in R(X_1,\dots,X_n)\ :\ \overline{x}_n<\mu_0-\frac{s_n}{\sqrt{n}}z_\alpha\Big\},$$

dove $s_n^2=\frac{1}{n-1}\sum_{i=1}^n(x_i-\overline{x}_n)^2$ è la varianza campionaria calcolata sul campione dei dati.
Risolviamo direttamente il secondo quesito, poiché la risposta porterà a rispondere al primo quesito. Per definizione, il p-value è il più piccolo livello di significatività per cui i dati osservati portano a un rifiuto dell'ipotesi nulla, ovvero

$$\begin{aligned}p\text{-value}&=\inf_\alpha\Big\{\overline{x}_n<\mu_0-\frac{s_n}{\sqrt{n}}z_\alpha\Big\}=\inf_\alpha\Big\{\frac{\overline{x}_n-\mu_0}{s_n/\sqrt{n}}<-z_\alpha\Big\}=\inf_\alpha\Big\{\Phi\Big(\frac{\overline{x}_n-\mu_0}{s_n/\sqrt{n}}\Big)<\Phi(-z_\alpha)\Big\}\\&=\inf_\alpha\Big\{\Phi\Big(\frac{\overline{x}_n-\mu_0}{s_n/\sqrt{n}}\Big)<\Phi(-z_\alpha)\Big\}=\inf_\alpha\Big\{\Phi\Big(\frac{\overline{x}_n-\mu_0}{s_n/\sqrt{n}}\Big)<\alpha\Big\}=\Phi\Big(\frac{\overline{x}_n-\mu_0}{s_n/\sqrt{n}}\Big).\end{aligned}$$

dove abbiamo usato che

$$\Phi(-z_\alpha)=\mathbb{P}(\{Z<-z_\alpha\})=\alpha.$$

Quindi, utilizzando le tavole,

$$\begin{aligned}p\text{-value}&=\Phi\Big(\frac{\overline{x}_n-\mu_0}{s_n/\sqrt{n}}\Big)=\Phi\Big(\frac{99.90-100}{0.55/\sqrt{40}}\Big)=\Phi(-1.15)=1-\Phi(1.15)\\&\simeq 1-0.8749=0.1251=12.51\%.\end{aligned}$$

In particolare, $5\%<12.51\%$, quindi l'ipotesi nulla non è rifiutata con il 5% di significatività.
(Per completezza, mostriamo come controllarlo direttamente. La regione critica è costituita dai dati tali che

$$\overline{x}_n<\mu_0-\frac{s_n}{\sqrt{n}}z_\alpha.$$

Per calcolare z_α utilizziamo che $\alpha=\mathbb{P}(\{Z>z_\alpha\})$, quindi $0.95=1-\alpha=\mathbb{P}(\{Z\le z_\alpha\})$. Utilizzando le tavole:

$$z_\alpha=z_{0.05}\simeq 1.645.$$

Segue che

$$\mu_0-\frac{s_n}{\sqrt{n}}z_\alpha=100-\frac{0.55}{\sqrt{40}}1.645\simeq 99.86.$$

Poiché $\overline{x}_n=99.90>99.86$, non rifiutiamo l'ipotesi nulla.)

Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni 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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: settembre 2022 - I
Corso di studi: _____ Data: 07/09/2022

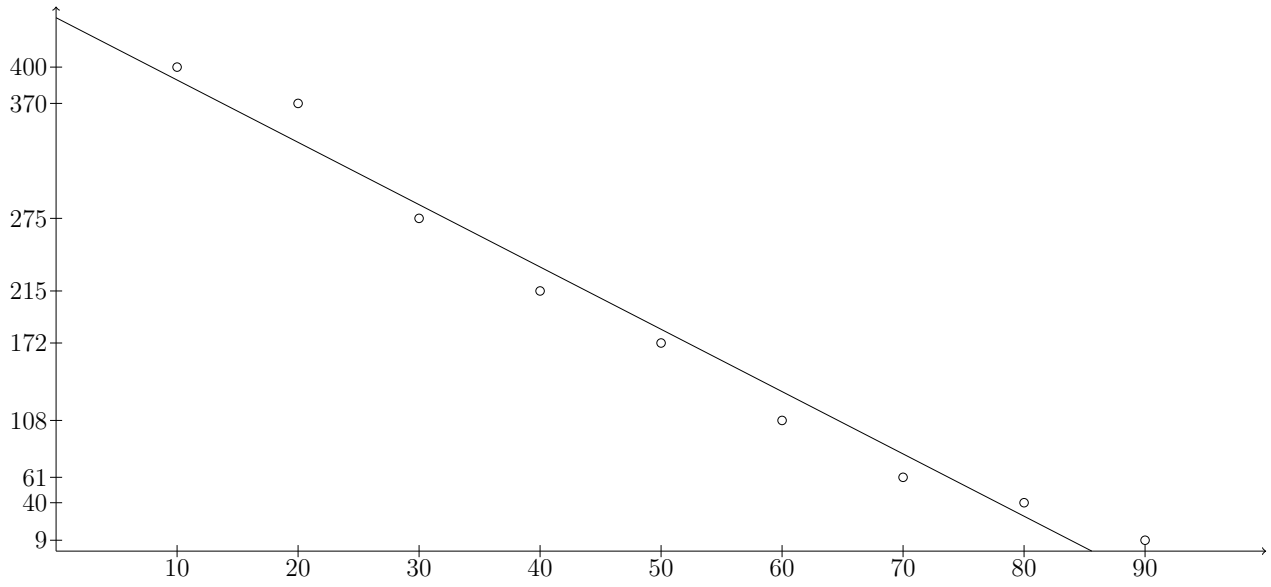
Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Un'azienda produce un dispositivo elettronico da utilizzare in un intervallo di temperatura molto ampio. L'azienda sa che l'aumento della temperatura riduce il tempo di vita del dispositivo, e quindi viene eseguito uno studio in cui il tempo di vita è determinato in funzione della temperatura. Si trovano i seguenti dati:

temperatura in $°C$	tempo di vita in ore
10	400
20	370
30	275
40	215
50	172
60	108
70	61
80	40
90	9

1. Rappresentare i dati in uno scatterplot.
2. Determinare (derivando le formule dei coefficienti) e rappresentare la retta di regressione lineare.
3. Calcolare il coefficiente di correlazione.

Soluzione. 1. Segue lo scatterplot.



2. Riportiamo i dati in una tabella:

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
10	400	4000	100	160000
20	370	7400	400	136900
30	275	8250	900	75625
40	215	8600	1600	46225
50	172	8600	2500	29584
60	108	6480	3600	11664
70	61	4270	4900	3721
80	40	3200	6400	1600
90	9	810	8100	81

Utilizzando il metodo dei minimi quadrati si trovano i coefficienti a e b della retta di regressione lineare $y = ax + b$. L'obiettivo è minimizzare:

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

Imponiamo che il gradiente rispetto ad a e b sia zero:

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

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

Dalla seconda condizione segue

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

che sostituita nella prima dà

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

Calcoliamo la media dei dati x_i e dei dati y_i:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{9}(10 + 20 + \cdots + 80 + 90) = 50$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{9}(400 + 370 + \cdots + 40 + 9) \simeq 183.33\,.$$

Otteniamo

$$a = \frac{(4000 + 7400 + \cdots + 3200 + 810) - 9 \cdot 50 \cdot 183.33}{(100 + 400 + \cdots + 6400 + 8100) - 9 \cdot 50^2} = \frac{51610 - 82498.5}{28500 - 22500} = -\frac{30888.5}{6000} \simeq -5.15$$

$$b = 183.33 + 5.15 \cdot 50 = 440.83$$

quindi la retta di regressione lineare è

$$y = -5.15x + 440.83\,.$$

Per disegnarla, determiniamo due punti per cui passa, ad esempio $(0, 440.83)$ e $(85.6, 0)$.
3. Per calcolare il coefficiente di correlazione possiamo usare le formule:

$$\rho = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}^2}} = a\frac{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}^2}}{\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}^2}} \simeq -5.15\frac{\sqrt{6000}}{\sqrt{162911}} \simeq -0.9883\,.$$

Esercizio 2. (7 punti) Una compagnia aerea ha osservato che su una certa tratta la probabilità che un passeggero che ha acquistato un biglietto non si presenti al momento dell'imbarco è del 5% (si supponga che i passeggeri siano indipendenti). L'aereo ha in tutto 96 posti, ma la compagnia prevede *overbooking* (sovrapprenotazione), quindi vende fino a 100 biglietti (supponiamo che la compagnia venda tutti i biglietti). Quindi non è detto che un posto a sedere sull'aereo sia garantito a tutti i passeggeri che hanno acquistato un biglietto e si presentano all'imbarco.

1. Qual è la probabilità che tutti i passeggeri che hanno acquistato il biglietto e si presentano all'imbarco abbiano un posto a sedere?

2. La compagnia ricava 200€ da ogni biglietto acquistato, mentre deve pagare un risarcimento di 600€ ai passeggeri che si sono presentati all'imbarco ma per cui non erano disponibili posti. Qual è il guadagno atteso per questo volo considerando risarcimenti dovuti?

Soluzione. 1. Consideriamo la seguente variabile aleatoria

$X =$ "numero di passeggeri che hanno acquistato il biglietto e si presentano all'imbarco" .

Osserviamo che X è una variabile aleatoria con legge binomiale con parametri $n = 100$ e $p = 0.95$, $X \sim \mathrm{B}(100, 0.95)$. Per convincerci di questo fatto, osserviamo che $X = X_1 + \cdots + X_{100}$ dove X_i sono le variabili aleatorie di Bernoulli indipendenti $X_i \sim \mathrm{Be}(p)$ tali che $X_i = 1$ se l'i-esimo passeggero che ha acquistato il biglietto si presenta all'imbarco e $X_i = 0$ se non si presenta all'imbarco.
In termini della variabile aleatoria X, l'evento "tutti i passeggeri che hanno acquistato il biglietto e si presentano all'imbarco hanno un posto a sedere" è $\{X \leq 96\}$ (cioè si presentano meno passeggeri dei posti disponibili). Allora calcoliamo

$$\begin{aligned}\mathbb{P}(X \leq 96) &= 1 - \mathbb{P}(X > 96) = 1 - \mathbb{P}(X = 97) - \mathbb{P}(X = 98) - \mathbb{P}(X = 99) - \mathbb{P}(X = 100)\\ &= 1 - \binom{100}{97}0.95^{97}0.05^3 - \binom{100}{98}0.95^{98}0.05^2 - \binom{100}{99}0.95^{99}0.05^1 - \binom{100}{100}0.95^{100}\\ &\simeq 74.21\%.\end{aligned}$$

2. Consideriamo la variabile aleatoria

$$Y = \text{“ricavo biglietti meno risarcimenti”} .$$

Possiamo scrivere Y in funzione della variabile aleatoria X. Se $X = x$, allora il guadagno dipende dal valore assunto x. Se $x \leq 96$, la compagnia non deve pagare alcun risarcimento, quindi ricava il massimo $200 \cdot 100$. Se invece $x \geq 97$, la compagnia deve pagare 600 a $x - 96$ passeggeri che non sono saliti sull'aereo, quindi guadagna $200 \cdot 100 - 600(x - 96)$. Quindi

$$Y = H(X) = \begin{cases} 200 \cdot 100 & \text{se } X \leq 96\,, \\ 200 \cdot 100 - 600(X - 96) & \text{se } X \geq 97\,. \end{cases}$$

Possiamo allora calcolare il valore atteso

$$\begin{aligned}
\mathbb{E}(Y) = \mathbb{E}(H(X)) &= \sum_{x=0}^{100} H(x)\mathbb{P}(\{X = x\}) = \sum_{x=0}^{96} H(x)\mathbb{P}(\{X = x\}) + \sum_{x=97}^{100} H(x)\mathbb{P}(\{X = x\}) \\
&= \sum_{x=0}^{96} 200 \cdot 100\ \mathbb{P}(\{X = x\}) + \sum_{x=97}^{100} (200 \cdot 100 - 600(x - 96))\mathbb{P}(\{X = x\}) \\
&= \sum_{x=0}^{100} 200 \cdot 100\ \mathbb{P}(\{X = x\}) - \sum_{x=97}^{100} 600(x - 96)\mathbb{P}(\{X = x\}) \\
&= 200 \cdot 100 \\
&\quad - 600\ \mathbb{P}(\{X = 97\}) - 600 \cdot 2\ \mathbb{P}(\{X = 98\}) - 600 \cdot 3\ \mathbb{P}(\{X = 99\}) - 600 \cdot 4\ \mathbb{P}(\{X = 100\}) \\
&= 20000 \\
&\quad - 600\Big(\binom{100}{97} 0.95^{97} 0.05^3 + 2\binom{100}{98} 0.95^{98} 0.05^2 + 3\binom{100}{99} 0.95^{99} 0.05^1 + 4\binom{100}{100} 0.95^{100}\Big) \\
&\simeq 20000 - 600 \cdot 41.91\% = 19748.54\,.
\end{aligned}$$

Esercizio 3. (7 punti) Sia X una variabile aleatoria assolutamente continua con la seguente densità

$$f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \quad x \in \mathbb{R}\,,$$

dove $\mu \in \mathbb{R}$ e $b > 0$ sono parametri da determinare.

1. Controllare che effettivamente $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$. (Suggerimenti: effettuare un cambio di variabile per traslazione, spezzare l'integrale in due e riscalare la variabile).

2. Determinare μ e b tali che $\mathbb{E}(X) = 0$ e $\mathrm{Var}(X) = 1$. (Suggerimenti: per $\mathbb{E}(X)$, effettuare un cambio di variabile per traslazione e utilizzare il punto 1.; per $\mathrm{Var}(X)$, utilizzare il valore di μ trovato, spezzare l'integrale in due, riscalare la variabile, integrare per parti)

3. Per i valori trovati, calcolare la probabilità che $X \leq 1$ sapendo che si è verificato l'evento $X \geq 0$.

Soluzione. 1. Calcoliamo l'integrale utilizzando i suggerimenti

$$\begin{aligned}
\int_{\mathbb{R}} f(x)\,\mathrm{d}x = \int_{\mathbb{R}} \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}\,\mathrm{d}x &\overset{y=x-\mu}{=} \int_{\mathbb{R}} \frac{1}{2b} e^{-\frac{|y|}{b}}\,\mathrm{d}y = \int_0^{+\infty} \frac{1}{2b} e^{-\frac{y}{b}}\,\mathrm{d}y + \int_{-\infty}^0 \frac{1}{2b} e^{\frac{y}{b}}\,\mathrm{d}y \overset{z=-y}{=} \\
&= \int_0^{+\infty} \frac{1}{2b} e^{-\frac{y}{b}}\,\mathrm{d}y - \int_{+\infty}^0 \frac{1}{2b} e^{-\frac{z}{b}}\,\mathrm{d}z = 2\int_0^{+\infty} \frac{1}{2b} e^{-\frac{y}{b}}\,\mathrm{d}y = \int_0^{+\infty} \frac{1}{b} e^{-\frac{y}{b}}\,\mathrm{d}y \overset{s=y/b}{=} \\
&= \int_0^{+\infty} e^{-s}\,\mathrm{d}s = \Big[-e^{-s} \Big]_0^{+\infty} = 1\,.
\end{aligned}$$

2. Calcoliamo il valore atteso

$$0 = \mathbb{E}(X) = \int_{\mathbb{R}} x f(x)\,\mathrm{d}x = \int_{\mathbb{R}} \frac{1}{2b} x e^{-\frac{|x-\mu|}{b}}\,\mathrm{d}x \overset{y=x-\mu}{=} \int_{\mathbb{R}} \frac{1}{2b}(y+\mu)e^{-\frac{|y|}{b}}\,\mathrm{d}y$$
$$= \int_{\mathbb{R}} \frac{1}{2b} y e^{-\frac{|y|}{b}}\,\mathrm{d}y + \int_{\mathbb{R}} \frac{1}{2b}\mu e^{-\frac{|y|}{b}}\,\mathrm{d}y \overset{ye^{-\frac{|y|}{b}} \text{ è dispari}}{=} \mu \int_{\mathbb{R}} \frac{1}{2b} e^{-\frac{|y|}{b}}\,\mathrm{d}y \overset{\text{punto 1.}}{=} \mu\,,$$

quindi $\mu = 0$. Calcoliamo la varianza, usando il fatto che $\mathbb{E}(X) = 0$,

$$1 = \mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X^2) = \int_{\mathbb{R}} x^2 f(x)\,\mathrm{d}x = \int_{\mathbb{R}} \frac{1}{2b} x^2 e^{-\frac{|x|}{b}}\,\mathrm{d}x$$
$$= \int_0^{+\infty} \frac{1}{2b} x^2 e^{-\frac{x}{b}}\,\mathrm{d}x + \int_{-\infty}^0 \frac{1}{2b} x^2 e^{-\frac{x}{b}}\,\mathrm{d}x \overset{\text{come in 1.}}{=}$$
$$= 2\int_0^{+\infty} \frac{1}{2b} x^2 e^{-\frac{x}{b}}\,\mathrm{d}x = \int_0^{+\infty} \frac{1}{b} x^2 e^{-\frac{x}{b}}\,\mathrm{d}x \overset{y=x/b}{=} \int_0^{+\infty} b^2 y^2 e^{-y}\,\mathrm{d}y \overset{\text{per parti}}{=}$$
$$= b^2\Big[-y^2 e^{-y}\Big]_0^{+\infty} + b^2 \int_0^{+\infty} 2y e^{-y}\,\mathrm{d}y = b^2\Big[-2y e^{-y}\Big]_0^{+\infty} + 2b^2 \int_0^{+\infty} e^{-y}\,\mathrm{d}y = 2b^2,$$

da cui segue $b = 1/\sqrt{2}$.
3. Utilizziamo la definizione di probabilità condizionata per calcolare

$$\mathbb{P}(\{X \le 1\} | \{X \ge 0\}) = \frac{\mathbb{P}(\{0 \le X \le 1\})}{\mathbb{P}(\{X \ge 0\})} = \frac{\int_0^1 f(x)\,\mathrm{d}x}{\int_0^{+\infty} f(x)\,\mathrm{d}x} = \frac{\int_0^1 \frac{1}{\sqrt{2}} e^{-\sqrt{2}x}\,\mathrm{d}x}{1/2} = 2\Big[-\frac{1}{2}e^{-\sqrt{2}x}\Big]_0^1$$
$$= 1 - e^{-\sqrt{2}}.$$

Esercizio 4. (8 punti) Si sa che la percentuale di titanio in una lega utilizzata nelle fusioni aerospaziali è distribuita con legge normale. Nelle domande seguenti per "esperimento statistico" intendiamo la misurazione della percentuale di titanio in 20 campioni selezionati casualmente.

1. Si fa un esperimento statistico e la deviazione standard calcolata sul campione risulta essere 0.37. Calcolare sui dati un intervallo di confidenza unilaterale sinistro (ovvero un limite superiore di confidenza) al 95% per la varianza. N.B.: derivare le formule.

2. È vero o falso che la varianza della popolazione appartiene all'intervallo calcolato nel punto precedente con il 95% di probabilità? Motivare la risposta.

3. Si ripetono tanti esperimenti statistici indipendenti. In media, dopo quanti esperimenti accade per la prima volta che la varianza della popolazione è fuori dall'intervallo di confidenza unilaterale sinistro al 95%?

Soluzione. L'esperimento statistico consiste nell'osservare un campione casuale $X_1, \ldots, X_n$ con $n = 20$, dove $X_i \sim \mathcal{N}(\mu, \sigma^2)$ e μ e σ sono incognite.
1. Un intervallo di confidenza unilaterale sinistro al 95% per la varianza della popolazione σ^2 è un intervallo della forma $(-\infty, V_n]$ dove V_n è una variabile aleatoria tale che

$$95\% = \mathbb{P}(\{\sigma^2 \le V_n\})\,.$$

Utilizzeremo la varianza campionaria

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Ricordiamo che, poiché la popolazione è distribuita con legge normale, si ha che $\Xi_{n-1} = (n-1)S_n^2/\sigma^2$ è distribuita come una chi-quadro con $(n-1) = 19$ gradi di libertà. Allora

$$\begin{aligned} 0.95 = \mathbb{P}(\{\sigma^2 \leq V_n\}) &= \mathbb{P}\Big(\Big\{\frac{1}{V_n} \leq \frac{1}{\sigma^2}\Big\}\Big) = \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{V_n} \leq \frac{(n-1)S_n^2}{\sigma^2}\Big\}\Big) \\ &= \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{V_n} \leq \Xi_{n-1}\Big\}\Big)\,. \end{aligned}$$

Scegliamo $\frac{(n-1)S_n^2}{V_n} = \chi^2_{19,0.95}$, dove $\chi^2_{19,0.95} = 10.117$ (ottenuto dalle tavole) è il quantile della distribuzione chi-quadro tale che

$$0.95 = \mathbb{P}(\chi^2_{19,0.95} \leq \Xi_{n-1})$$

in modo che valga la condizione di intervallo di confidenza. Segue che

$$\frac{(n-1)S_n^2}{V_n} = \chi^2_{19,0.95} \implies V_n = \frac{(n-1)S_n^2}{\chi^2_{19,0.95}}\,.$$

I dati del problema forniscono una realizzazione della deviazione standard campionaria, e quindi della varianza campionaria

$$s_n^2 = 0.37^2 = 0.1369\,.$$

Utilizziamo questa per calcolare la realizzazione di V_n sui dati.

$$v_n = \frac{19 \cdot 0.1369}{10.117} = 0.2571$$

quindi $(-\infty, 0.2571]$ è un intervallo di confidenza unilaterale sinistro al 95% calcolato sui dati.
2. Se si considera l'intervallo $(-\infty, 0.2571]$ calcolato sui dati, l'affermazione è falsa! Non ha nemmeno senso calcolare la probabilità che $\sigma^2 \leq 0.2571$ perché σ^2 è un parametro e 0.2571 è un numero, non sono variabili aleatorie. Quello che si può dire è che σ^2 appartiene all'intervallo $(-\infty, V_n]$ dove $V_n = \frac{(n-1)S_n^2}{\chi^2_{19,0.95}}$ è una variabile aleatoria (non una sua realizzazione).
3. L'estremo dell'intervallo di confidenza V_n è una variabile aleatoria, e per definizione definizione di intervallo di confidenza si ha che σ^2 è fuori dall'intervallo $(-\infty, V_n]$ con probabilità $1 - 0.95 = 0.05$. Consideriamo la variabile aleatoria

$$Y = \text{“prima volta in cui } \sigma^2 > V_n\text{”}.$$

Poiché questa variabile aleatoria rappresenta il primo successo in una successione di prove indipendenti, ha una distribuzione geometrica. Il parametro della distribuzione è la probabilità di successo, quindi è 0.05. Il valore atteso di una variabile aleatoria con distribuzione geometrica è il reciproco del parametro, quindi

$$\mathbb{E}(Y) = \frac{1}{0.05} = 20\,.$$

In media, alla ventesima prova accadrà che σ^2 è fuori dall'intervallo di confidenza.

Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni 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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: settembre 2022 - II
Corso di studi: _____ Data: 20/09/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) In un'indagine sui consumi di nuove auto a benzina è stata osservata la distribuzione dei litri consumati per 100 km. I dati sono rappresentati raggruppati in intervalli di classi nella seguente tabella:

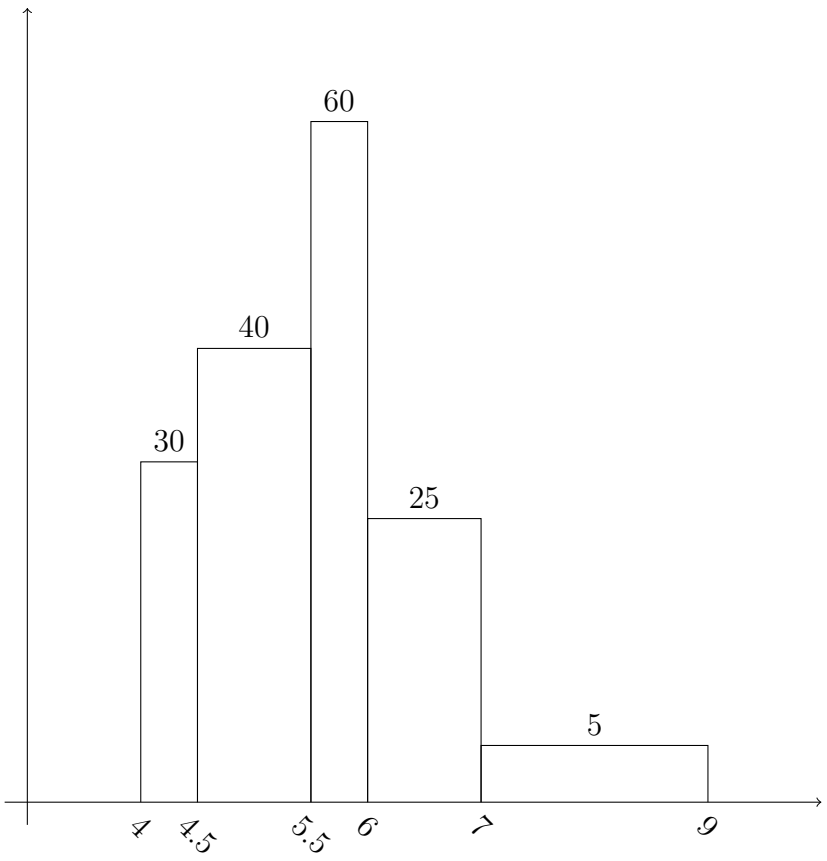
intervalli	frequenze assolute
$[4, 4.5)$	15
$[4.5, 5.5)$	40
$[5.5, 6)$	30
$[6, 7)$	25
$[7, 9)$	10

1. Rappresentare un istogramma delle densità di frequenze assolute.
2. Determinare la classe modale.
3. Calcolare un'approssimazione della media e della deviazione standard dei dati.
4. Calcolare un'approssimazione della mediana dei dati.

Soluzione. 1. Denotiamo con $I_1, \dots, I_5$ gli intervalli, $f_1, \dots, f_5$ le frequenze assolute. Abbiamo che $n = f_1 + \cdots + f_5 = 15 + 40 + 30 + 25 + 10 = 120$. Ricordiamo che le densità di frequenze assolute sono date da $d_j = f_j/|I_j|$ dove $|I_j| = b_j - a_j$ se $I_j = [a_j, b_j)$. Completiamo la tabella (scriviamo anche le frequenze relative per il punto 3. e le frequenze assolute cumulate per il punto 4.):

intervalli	f. assolute	densità f. ass.	f. relative	f. cumulate
$[4, 4.5)$	15	30	12.5%	15
$[4.5, 5.5)$	40	40	33.33%	55
$[5.5, 6)$	30	60	25%	85
$[6, 7)$	25	25	20.83%	110
$[7, 9)$	10	5	8.34%	120

Rappresentiamo l'istogramma:



2. La classe modale è l'intervallo $[5.5, 6)$ poiché è l'intervallo con la maggiore densità di frequenza relativa.
3. Ricordando che la media calcolata su un campione di dati $x_1, \ldots, x_n$ è

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

possiamo riscrivere la formula in termini dei valori assunti $v_1, \ldots, v_k$ utilizzando le frequenze assolute $f_1, \ldots, f_k$ e le frequenze relative $p_1, \ldots, p_k$

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{k} f_j v_j = \sum_{j=1}^{k} p_j v_j \,.$$

Per approssimare la media sostituiamo ai valori v_j i valori centrali $\tilde{v}_j$ degli intervalli e le frequenze relative:

$$\overline{x} \simeq 12.5\% \cdot 4.25 + 33.33\% \cdot 5 + 25\% \cdot 5.75 + 20.83\% \cdot 6.5 + 8.34\% \cdot 8 = 5.6564 \,.$$

Ricordiamo che la varianza calcolata su un campione di dati $x_1, \ldots, x_n$ può essere calcolata mediante la formula:

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

Possiamo riscrivere la formula in termini dei valori assunti $v_1, \ldots, v_k$ utilizzando le frequenze assolute $f_1, \ldots, f_k$ e le frequenze relative $p_1, \ldots, p_k$:

$$s^2 = \frac{1}{n-1} \Big(\sum_{j=1}^{k} f_j v_j^2 - n\overline{x}^2 \Big) = \frac{n}{n-1} \Big(\sum_{j=1}^{k} p_j v_j^2 - \overline{x}^2 \Big).$$

Per approssimare la varianza sostituiamo ai valori v_j i valori centrali $\tilde{v}_j$ degli intervalli e le frequenze relative e la media approssimata calcolata nel punto precedente:

$$s^2 \simeq \frac{120}{119}\Big(12.5\% \cdot 4.25^2 + 33.33\% \cdot 5^2 + 25\% \cdot 5.75^2 + 20.83\% \cdot 6.5^2 + 8.34\% \cdot 8^2 - 5.6564^2\Big) \simeq 1.0077$$

da cui segue che la deviazione standard è approssimata da

$$s \simeq 1.0038.$$

4. Per calcolare un'approssimazione della mediana utilizziamo le frequenze cumulate $F_1, \dots, F_k$. La mediana divide l'insieme il campione di dati in due parti, quindi calcoliamo $\frac{n}{2} = \frac{120}{2} = 60$ e osserviamo che per l'intervallo $I_3 = [a_3, b_3) = [5.5, 6)$ si ha che

$$F_2 = 55 < 60 < 85 = F_3\,.$$

La mediana è allora approssimata da

$$Q_2 \simeq a_3 + \lambda_3(b_3 - a_3), \quad \lambda_3 = \frac{\frac{n}{2} - F_2}{F_3 - F_2} = \frac{60 - 55}{30} = \frac{1}{6}$$

quindi

$$Q_2 \simeq 5.5 + \frac{1}{6} \cdot 0.5 \simeq 5.58.$$

Esercizio 2. (8 punti) L'azienda per cui lavori offre ogni anno un corso di aggiornamento facoltativo. Il numero di persone che fa domanda per il corso è una variabile aleatoria distribuita con una legge di Poisson e, in media, 5 persone fanno domanda per seguire il corso. L'azienda deve decidere se offrire il corso in streaming oppure in presenza (e in tal caso deve organizzarsi per tempo per procurarsi un'aula). Se il numero di persone partecipanti è almeno 4 (compreso), il corso è offerto in presenza, altrimenti il corso è offerto in streaming online.

1. Qual è la probabilità che il corso venga offerto in presenza?

2. L'azienda viene a conoscenza del numero delle prime persone iscritte e sa che il corso verrà offerto in presenza. Deve quindi prenotare un'aula. Se l'azienda vuole che la probabilità di far sedere tutti i partecipanti sia almeno del 90%, sono sufficienti 6 posti a sedere? Se no, quanti ne servono?

3. La seguente affermazione è vera oppure falsa? "Grazie all'assenza di memoria possiamo affermare che la probabilità che il numero di partecipanti sia maggiore di 15 sapendo che il numero di partecipanti è maggiore di 10 è uguale alla probabilità che il numero di partecipanti sia maggiore di 5." (N.B.: non sono richiesti i calcoli, ma si deve motivare la risposta)

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{"numero di partecipanti"} \sim \mathrm{P}(\lambda)\,.$$

Per determinare il parametro della legge di Poisson ricordiamo che $\mathbb{E}(X) = \lambda$ e la traccia spiega che $\mathbb{E}(X) = 5$, quindi $\lambda = 5$ e $X \sim \mathrm{P}(5)$, cioè

$$\mathbb{P}(\{X = k\}) = e^{-5}\frac{5^k}{k!}\,.$$

1. Ci viene chiesto di calcolare la probabilità che ci siano almeno 4 persone partecipanti

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

2. Poiché sappiamo che si è realizzato l'evento $\{X \geq 4\}$, ci viene chiesto di calcolare la probabilità condizionata

$$\mathbb{P}(\{X \leq 6\}|\{X \geq 4\}) = \frac{\mathbb{P}(\{4 \leq X \leq 6\})}{\mathbb{P}(\{X \geq 4\})} = \frac{e^{-5}\Big(\frac{5^4}{4!} + \frac{5^5}{5!} + \frac{5^6}{6!}\Big)}{73.50\%} \simeq 67.64\%.$$

Prenotare 6 posti non è sufficiente per ottenere il 90% di probabilità. Aggiungiamo un posto alla volta finché la probabilità non supera il 90%:

$$\mathbb{P}(\{X \leq 7\}|\{X \geq 4\}) = \frac{\mathbb{P}(\{4 \leq X \leq 7\})}{\mathbb{P}(\{X \geq 4\})} = \frac{e^{-5}\Big(\frac{5^4}{4!} + \frac{5^5}{5!} + \frac{5^6}{6!} + \frac{5^7}{7!}\Big)}{73.50\%} \simeq 81.85\%,$$

$$\mathbb{P}(\{X \leq 8\}|\{X \geq 4\}) = \frac{\mathbb{P}(\{4 \leq X \leq 8\})}{\mathbb{P}(\{X \geq 4\})} = \frac{e^{-5}\Big(\frac{5^4}{4!} + \frac{5^5}{5!} + \frac{5^6}{6!} + \frac{5^7}{7!} + \frac{5^8}{8!}\Big)}{73.50\%} \simeq 90.73\%,$$

quindi sono sufficienti 8 posti.
3. L'affermazione è falsa: le uniche variabili aleatorie discrete che godono dell'assenza di memoria sono quelle con distribuzione geometrica.

Esercizio 3. (7 punti) Sia (X_1, X_2) un vettore aleatorio con funzione di probabilità congiunta descritta dalla seguente tabella:

X_2 \ X_1	1	2	3
-1	$1/6$	a	b
1	b	$1/6$	a

dove $a, b \geq 0$.

1. Determinare i valori di a e b per cui $\mathrm{Cov}(X_1, X_2) = 0$.
2. Per i valori di a e b determinati nel punto 2., si ha che X_1 e X_2 sono indipendenti?
3. Si vuole osservare una successione di realizzazioni indipendenti del vettore aleatorio (X_1, X_2). Qual è la probabilità di dover attendere (strettamente) più di 10 osservazioni perché si verifichi $X_1 = 1$?

Soluzione. Ricordiamo che la somma dei valori assunti dalla funzione di probabilità congiunta deve essere 1:

$$1 = \frac{1}{6} + a + b + b + \frac{1}{6} + a \implies a + b + \frac{1}{6} = \frac{1}{2}.$$

1. Calcoliamo la covarianza, ricordando che $\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)$. Partiamo dai valori attesi di X_1 e X_2 e notiamo che

$$\mathbb{E}(X_2) = -1 \cdot \Big(\frac{1}{6} + a + b\Big) + 1 \cdot \Big(b + \frac{1}{6} + a\Big) = 0$$

quindi $\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2)$. Resta da calcolare

$$\mathbb{E}(X_1 X_2) = -1 \cdot \frac{1}{6} - 2 \cdot a - 3 \cdot b + 1 \cdot b + 2 \cdot \frac{1}{6} + 3 \cdot a = a - 2b + \frac{1}{6}.$$

Imponiamo che questa quantità si annulli come richiesto e mettiamo a sistema con la relazione trovata precedentemente:

$$\begin{cases} a + b + \frac{1}{6} = \frac{1}{2} \\ a - 2b + \frac{1}{6} = 0 \end{cases} \implies \begin{cases} a = \frac{1}{3} - b \\ 3b = \frac{1}{2} \end{cases} \implies \begin{cases} a = \frac{1}{6} \\ b = \frac{1}{6}. \end{cases}$$

2. Riscrivendo la tabella con i valori trovati

X_2 \ X_1	1	2	3
-1	1/6	1/6	1/6
1	1/6	1/6	1/6

ci rendiamo conto che (X_1, X_2) è distribuito in modo uniforme. Per avere l'indipendenza di X_1 e X_2 deve valere che

$$\mathbb{P}(\{X_1 = h, X_2 = k\}) = \mathbb{P}(\{X_1 = h\})\mathbb{P}(\{X_2 = k\}) \quad \text{per ogni } h \in \{1,2,3\}, k \in \{-1,1\}.$$

Da un lato abbiamo che, per qualunque h e k,

$$\mathbb{P}(\{X_1 = h, X_2 = k\}) = \frac{1}{6}\,.$$

D'altro canto, per qualunque h abbiamo che

$$\mathbb{P}(\{X_1 = h\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

mentre per qualunque k

$$\mathbb{P}(\{X_2 = k\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Segue che

$$\mathbb{P}(\{X_1 = h, X_2 = k\}) = \frac{1}{6} = \frac{1}{2}\frac{1}{3} = \mathbb{P}(\{X_1 = h\})\mathbb{P}(\{X_2 = k\}),$$

cioè X_1 e X_2 sono effettivamente indipendenti.
(Attenzione: per l'indipendenza è importante controllare la condizione su tutti i valori assunti, non è sufficiente controllare una sola coppia di valori, come $\mathbb{P}(\{X_1 = 1, X_2 = 1\}) = \mathbb{P}(\{X_1 = 1\})\mathbb{P}(\{X_2 = 1\})$.)
3. Consideriamo la variabile aleatoria

$$Y = \text{“prima volta che si osserva } X_1 = 1\text{”} \sim \mathrm{Geo}(p)$$

dove $p = \mathbb{P}(\{X_1 = 1\}) = \frac{1}{3}$. Ci viene chiesto di calcolare $\mathbb{P}(\{Y > 10\})$. Possiamo farlo direttamente, calcolando

$$\begin{aligned} \mathbb{P}(\{Y > 10\}) &= 1 - \mathbb{P}(\{Y \le 10\}) \\ &= 1 - \mathbb{P}(\{Y = 1\}) - \mathbb{P}(\{Y = 2\}) - \mathbb{P}(\{Y = 3\}) - \mathbb{P}(\{Y = 4\}) - \mathbb{P}(\{Y = 5\}) \\ &\quad - \mathbb{P}(\{Y = 6\}) - \mathbb{P}(\{Y = 7\}) - \mathbb{P}(\{Y = 8\}) - \mathbb{P}(\{Y = 9\}) - \mathbb{P}(\{Y = 10\}) \\ &= 1 - p - (1-p)p - (1-p)^2 p - (1-p)^3 p - (1-p)^4 p \\ &\quad - (1-p)^5 p - (1-p)^6 p - (1-p)^7 p - (1-p)^8 p - (1-p)^9 p\,, \end{aligned}$$

e sostituendo il valore di p, ma questo è noioso e può portare a errori di conto. È meglio utilizzare il trucco della somma geometrica

$$s_{10} = \sum_{i=1}^{10}(1-p)^{i-1}p \implies (1-p)s_{10} = \sum_{i=1}^{10}(1-p)^i p = \sum_{i=2}^{11}(1-p)^{i-1}p = s_{10} + (1-p)^{10}p - p$$
$$\implies s_{10} = 1-(1-p)^{10}$$

da cui segue che

$$\mathbb{P}(\{Y > 10\}) = 1 - s_{10} = (1-p)^{10} \simeq 1.73\%.$$

Possiamo anche pensarla così: $\{Y > 10\}$ vuol dire che ci sono stati 10 insuccessi consecutivi. Questo evento ha probabilità $(1-p)^{10}$.

Esercizio 4. (7 punti) Un produttore sostiene che le sue batterie abbiano una durata di almeno 100 ore. Si sa che la deviazione standard per questo tipo di batterie è di $\sigma = 10$ ore. Un cliente, insospettito dall'affermazione del produttore, fa una prova: acquista e testa 40 campioni, osservando una media campionaria di 96.5 ore.

1. L'osservazione del cliente è significativa al 5% per destare sospetti sull'effettiva qualità delle batterie?

2. Qual è il più piccolo livello di significatività per cui i dati osservati permettono di contestare l'affermazione del produttore?

Soluzione. Stiamo considerando un campione casuale $X_1, \dots, X_n$ con $n = 40 > 30$. Non conosciamo la distribuzione delle X_i, ma è noto che la deviazione standard della popolazione è $\sigma = 10$. La media della popolazione μ è incognita e si effettua un test d'ipotesi su μ.
1. Poiché ci viene chiesto se i dati sono abbastanza significativi da rifiutare l'affermazione del produttore, il test d'ipotesi in questione è

$$H_0 : \mu \geq \mu_0 \text{ (affermazione del produttore)} \quad H_1 : \mu < \mu_0,$$

dove $\mu_0 = 100$. La regione critica è pertanto 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 campionaria calcolata sui dati, realizzazione della media campionaria (variabile aleatoria) $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, stimatore corretto della media. (Brevemente: se i dati mostrano una media campionaria esageratamente più piccola di μ_0, l'ipotesi nulla va rifiutata).
Ricordiamo che il livello di significatività $\alpha = 5\%$ del test è la probabilità di commettere un errore del I tipo. Supponiamo quindi che l'ipotesi nulla sia vera, cioè $\mu \geq \mu_0$. Non possiamo dire molto sulla statistica

$$\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

poiché μ_0 non è la media della popolazione. Allora osserviamo che, essendo $\mu \geq \mu_0$, si ha che

$$\overline{X}_n \leq \mu_0 - \delta \implies \overline{X}_n \leq \mu - \delta$$

quindi

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

Per il Teorema del Limite Centrale e poiché il campione è numeroso ($n > 30$), possiamo approssimare $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ con una variabile aleatoria con legge normale standard

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \simeq Z \sim \mathcal{N}(0,1).$$

Consideriamo il quantile Gaussiano z_α definito da $\alpha = \mathbb{P}(\{Z \geq z_\alpha\})$. Scegliendo $\frac{\delta}{\sigma/\sqrt{n}} = z_\alpha$ abbiamo che

$$\mathbb{P}(\{(X_1, \ldots, X_n) \in R_C\}) \leq \mathbb{P}(\{Z < -z_\alpha\}) = \alpha,$$

cioè la probabilità di commettere un errore del I tipo è meno di α. Per calcolare la regione critica utilizziamo le tavole per calcolare il quantile Gaussiano

$$z_\alpha = z_{0.05} \simeq 1.645$$

da cui segue

$$\frac{\delta}{\sigma/\sqrt{n}} = z_\alpha \implies \delta = z_\alpha \frac{\sigma}{\sqrt{n}} = 1.645 \frac{10}{\sqrt{40}} \simeq 2.60.$$

Non resta che controllare se la realizzazione del campione è nella regione critica:

$$\overline{x}_n = 96.5$$

$$\mu_0 - \delta = 100 - 2.60 = 97.4$$

quindi $\overline{x}_n < \mu_0 - \delta$, cioè la realizzazione del campione è nella regione critica e l'ipotesi nulla va rifiutata.

2. Il più piccolo livello di significatività per cui i dati osservati permettono di rifiutare l'ipotesi nulla è il p-value del test

$$\begin{aligned} p - \text{value} &= \inf_\alpha \left\{ \overline{x}_n < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \right\} = \inf_\alpha \left\{ \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha \right\} = \inf_\alpha \left\{ \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Big) < \Phi(-z_\alpha) \right\} \\ &= \inf_\alpha \left\{ \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Big) < \alpha \right\} = \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Big) = \Phi\Big(\frac{96.5 - 100}{10/\sqrt{40}} \Big) \\ &\simeq \Phi(-2.21) = 1 - \Phi(2.21) \simeq 1 - 0.9864 = 0.0136 = 1.36\%. \end{aligned}$$

Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni 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: _____ A.A.: 2022/2023
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: novembre 2022
Corso di studi: _____ Data: 11/11/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) I voti ottenuti dagli studenti e dalle studentesse a un appello dell'esame di Probabilità e Statistica[1] sono i seguenti:

26 8 28 5 27 30 26 18 28 26 25 20 22 30 20 21

1. Determinare i quartili.
2. Determinare eventuali dati anomali o sospetti.
3. Tracciare un box-plot.

Soluzione. 1. Ordiniamo i dati:

5 8 18 20 20 21 22 25 26 26 26 27 28 28 30 30

Denotiamo con $x_1, \ldots, x_{16}$ i dati ordinati. L'ampiezza del campione è $n = 16$.

Calcolo di Q_1. Per trovare il primo quartile calcoliamo $\frac{n+1}{4} = \frac{17}{4} = 4 + \frac{1}{4} = 4 + 0.25$. Quindi

$$Q_1 = (1 - 0.25)x_4 + 0.25x_5 = 20\,.$$

Calcolo di Q_2. Per trovare il secondo quartile calcoliamo $(n+1)\frac{2}{4} = \frac{17}{2} = 8 + \frac{1}{2} = 8 + 0.5$. Quindi

$$Q_2 = 0.5x_8 + 0.5x_9 = 25.5\,.$$

Calcolo di Q_3. Per trovare il terzo quartile calcoliamo $(n+1)\frac{3}{4} = 17\frac{3}{4} = 12 + \frac{3}{4} = 12 + 0.75$. Quindi

$$Q_3 = (1 - 0.75)x_{12} + 0.75x_{13} = 27.75\,.$$

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

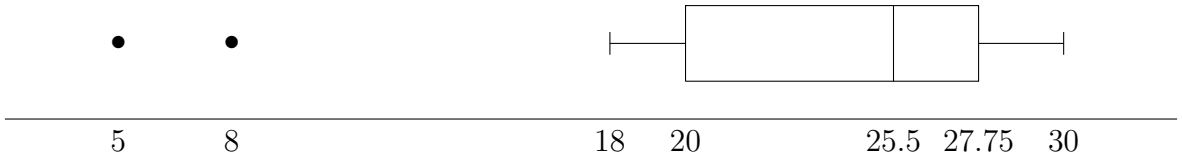
$$IQR = Q_3 - Q_1 = 27.75 - 20 = 7.75\,.$$

[1] I dati sono generati casualmente e non si riferiscono a fatti realmente accaduti.

I dati anomali sono più grandi di $Q_3 + 3IQR = 51$ o più piccoli di $Q_1 - 3IQR = -3.25$. Quindi non ci sono dati anomali.

I dati sospetti cadono tra $Q_3+1.5IQR = 39.375$ e $Q_3+3IQR = 51$ oppure tra $Q_1-3IQR = -3.25$ e $Q_1 - 1.5IQR = 8.375$. Quindi 5 e 8 sono dati sospetti.

3. Segue il box-plot:



Esercizio 2. (7 punti) Un ristorante studia i suoi clienti fissi e gli effetti della nuova campagna di marketing adottata dal ristorante tramite *ad* su un *social network*. Gli *ad* risultano efficaci su una certa proporzione p dei clienti. Lo studio porta a queste conclusioni:

- Per un cliente che è influenzato dagli *ad*, il numero di visite annue al ristorante è distribuito con una legge di Poisson con una media di 7 visite all'anno;
- Per un cliente che non è influenzato dagli *ad* (nella restante proporzione della popolazione $1 - p$), il numero di visite annue al ristorante è distribuito con una legge di Poisson con una media di 4 visite all'anno.

Rispondere ai seguenti quesiti:

1. Determinare (in funzione di p) la probabilità che il numero di visite annue di un cliente sia uguale a un dato numero $k = 0, 1, 2, \ldots$.
2. Un'analisi mostra che il 50% dei clienti visita il ristorante almeno 5 volte all'anno. (Si legga: la probabilità che il numero di visite di un cliente sia maggiore o uguale a 5 è il 50%.) Ricavare in questo caso la proporzione p dei clienti che viene influenzata dagli *ad*.

Soluzione. Lo svolgimento è analogo a quella dell'esercizio sui farmaci svolto nella Lezione 13 il 12/04/2022.

Consideriamo le seguenti variabili aleatorie

$$X = \begin{cases} 1 & \text{se gli } ad \text{ sono efficaci sul cliente} \\ 0 & \text{se gli } ad \text{ non sono efficaci sul cliente} \end{cases}$$

$$Y = \text{“numero di visite annuali del cliente”} .$$

La variabile X è distribuita con una legge di Bernoulli con parametro p. La variabile Y ha una distribuzione dipendente dal valore assunto da X. Dalla traccia abbiamo che

$$\mathbb{P}(\{Y = k\}|\{X = 1\}) = e^{-7}\frac{7^k}{k!}$$

perché questa è una legge di Poisson con media 7 (ricordiamo che il parametro della legge di Poisson è uguale al valore atteso) e

$$\mathbb{P}(\{Y = k\}|\{X = 0\}) = e^{-4}\frac{4^k}{k!}$$

perché questa è una legge di Poisson con media 4.

1. Dal teorema della probabilità totale abbiamo che

$$\begin{aligned}\mathbb{P}(\{Y=k\}) &= \mathbb{P}(\{Y=k\}|\{X=0\})\mathbb{P}(\{X=0\}) + \mathbb{P}(\{Y=k\}|\{X=1\})\mathbb{P}(\{X=1\}) \\ &= e^{-4}\frac{4^k}{k!}(1-p) + e^{-7}\frac{7^k}{k!}p\,.\end{aligned}$$

2. Utilizzando la formula ottenuta nel primo punto

$$\begin{aligned}&\mathbb{P}(\{Y\geq 5\}) \\ &= 1-\mathbb{P}(\{Y\leq 4\}) = 1-\sum_{k=0}^{4}\mathbb{P}(\{Y=k\}) \\ &= 1-(1-p)\sum_{k=0}^{4}e^{-4}\frac{4^k}{k!} + p\sum_{k=0}^{4}e^{-7}\frac{7^k}{k!} \\ &= 1-(1-p)e^{-4}\Big(1+4+\frac{4^2}{2!}+\frac{4^3}{3!}+\frac{4^4}{4!}\Big) - pe^{-7}\Big(1+7+\frac{7^2}{2!}+\frac{7^3}{3!}+\frac{7^4}{4!}\Big) \\ &\simeq 1-62.88\%(1-p)-17.30\%p = 1-62.88\%+62.88\%p-17.30\%p = 37.12\%+45.58\%p\end{aligned}$$

La traccia fornisce la seguente informazione

$$\mathbb{P}(\{Y\geq 5\}) = 50\%\,.$$

Possiamo allora imporre:

$$37.12\%+45.58\%p = 50\% \implies p\simeq 28.26\%\,.$$

Esercizio 3. (7 punti) L'ufficio informazioni di una compagnia ha due numeri verdi distinti. I tempi di attesa per parlare con gli operatori sono, per entrambi i numeri, variabili aleatorie distribuite con legge esponenziale con media 10 minuti. Inoltre i due tempi di attesa si possono considerare indipendenti. Avendo a disposizione due telefoni, decidi di chiamare contemporaneamente i due numeri.

1. Qual è la probabilità che qualcuno risponda dal primo numero dopo 5 minuti? E che qualcuno risponda dal secondo numero dopo 5 minuti?
2. Qual è la probabilità di attendere meno di 5 minuti fino alla risposta da uno dei due numeri (non importa quale dei due)?
3. (Domanda bonus: come se fosse un quesito teorico) Quanto tempo aspetterai in media fino alla risposta da uno dei due numeri?

Soluzione. Il tempo di attesa ai due sportelli è dato da due variabili aleatorie

$$X_1 = \text{“tempo di attesa al primo sportello”} \sim \mathrm{Exp}\Big(\frac{1}{10}\Big)\,,$$

$$X_2 = \text{“tempo di attesa al secondo sportello”} \sim \mathrm{Exp}\Big(\frac{1}{10}\Big)\,.$$

In entrambi i casi abbiamo utilizzato il fatto che per $X_1\sim\mathrm{Exp}(\lambda)$ si ha $\mathbb{E}(X_1)=\frac{1}{\lambda}$.

1. La probabilità che qualcuno risponda dal primo numero dopo 5 minuti è

$$\mathbb{P}(\{X_1 \geq 5\}) = \int_5^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x}\, \mathrm{d}x = \Big[-e^{-\frac{1}{10}x} \Big]_5^{+\infty} = e^{-\frac{1}{10}5} = e^{-\frac{1}{2}} \simeq 60.65\%\,.$$

La stessa probabilità è $\mathbb{P}(\{X_2 \geq 5\})$ poiché X_1 e X_2 sono identicamente distribuite.

2. La risposta da uno dei due numeri entro i 5 minuti se almeno uno tra X_1 e X_2 è più piccolo di 5, ovvero si verifica il complementare dell'evento: entrambi X_1 e X_2 sono maggiori di 5. Utilizzando l'indipendenza:

$$\begin{aligned}\mathbb{P}(\{X_1 < 5\} \cup \{X_2 < 5\}) &= 1 - \mathbb{P}(\{X_1 \geq 5\} \cap \{X_2 \geq 5\}) = 1 - \mathbb{P}(\{X_1 \geq 5\})\mathbb{P}(\{X_2 \geq 5\}) \\ &= 1 - 60.65\% \cdot 60.65\% \simeq 63.22\%\,.\end{aligned}$$

3. Per calcolare il tempo medio di attesa integriamo i valori assunti dal tempo di attesa (il più piccolo tra il valore assunto da X_1 e da X_2) contro la densità di probabilità congiunta, che è data dal prodotto delle densità di probabilità essendo le due variabili aleatorie indipendenti:

$$\begin{aligned}&\int_0^{+\infty}\int_0^{+\infty} \min\{x_1, x_2\} \frac{1}{10} e^{-\frac{1}{10}x_1} \frac{1}{10} e^{-\frac{1}{10}x_2}\, \mathrm{d}x_1\, \mathrm{d}x_2 = \\
&= \int_0^{+\infty}\int_0^{x_2} x_1 \frac{1}{10} e^{-\frac{1}{10}x_1} \frac{1}{10} e^{-\frac{1}{10}x_2}\, \mathrm{d}x_1\, \mathrm{d}x_2 + \int_0^{+\infty}\int_{x_2}^{+\infty} x_2 \frac{1}{10} e^{-\frac{1}{10}x_1} \frac{1}{10} e^{-\frac{1}{10}x_2}\, \mathrm{d}x_1\, \mathrm{d}x_2 \\
&= \int_0^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x_2} \int_0^{x_2} \frac{1}{10} x_1 e^{-\frac{1}{10}x_1}\, \mathrm{d}x_1\, \mathrm{d}x_2 + \int_0^{+\infty} \frac{1}{10} x_2 e^{-\frac{1}{10}x_2} \int_{x_2}^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x_1}\, \mathrm{d}x_1\, \mathrm{d}x_2 \\
&= \int_0^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x_2} \Big[-10 e^{-\frac{1}{10}x_1} - x_1 e^{-\frac{1}{10}x_1} \Big]_0^{x_2} \mathrm{d}x_2 + \int_0^{+\infty} \frac{1}{10} x_2 e^{-\frac{1}{10}x_2} \Big[-e^{-\frac{1}{10}x_1} \Big]_{x_2}^{+\infty} \mathrm{d}x_2 \\
&= \int_0^{+\infty} \Big(-e^{-\frac{1}{5}x_2} - \frac{1}{10} x_2 e^{-\frac{1}{5}x_2} + e^{-\frac{1}{10}x_2} \Big)\, \mathrm{d}x_2 + \int_0^{+\infty} \frac{1}{10} x_2 e^{-\frac{1}{5}x_2}\, \mathrm{d}x_2 \\
&= \int_0^{+\infty} \Big(-e^{-\frac{1}{5}x_2} + e^{-\frac{1}{10}x_2} \Big)\, \mathrm{d}x_2 = \Big[5e^{-\frac{1}{5}x_2} - 10 e^{-\frac{1}{10}x_2} \Big]_0^{+\infty} = -5 + 10 = 5\,.\end{aligned}$$

Esercizio 4. (8 punti) Si vuole studiare l'effetto della delaminazione sulla frequenza naturale delle travi realizzate con laminati compositi. Si effettua il seguente esperimento statistico: si considera un campione di otto travi delaminate, le si sottopongono a carichi e se ne misurano le frequenze risultanti (in Hertz). Per un esperimento statistico si osservano i seguenti dati:

230.66 233.05 232.58 229.48 232.58 235.76 229.43 234.13

Si supponga che i dati provengano da una popolazione distribuita con legge normale.

1. Calcolare sui dati un intervallo di confidenza bilaterale al 90% sulla frequenza naturale media della popolazione.

2. Supponiamo di effettuare 20 esperimenti statistici indipendenti (ottenendo per ogni esperimento statistico nuovi valori) e di calcolare sui dati di ogni esperimento statistico un intervallo di confidenza bilaterale al 90% come nel punto precedente. Qual è la probabilità che almeno 4 volte la frequenza naturale media della popolazione sia fuori dall'intervallo di confidenza calcolato sui dati?

Soluzione. 1. Stiamo considerando un campione casuale $X_1, \ldots, X_n$, $n = 8$, estratto da una popolazione normale. Quindi $X_1, \ldots, X_n$ sono indipendenti e tutte distribuite con legge

normale $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Osserviamo che la media della popolazione μ e la varianza σ^2 non sono note.

Un intervallo di confidenza bilaterale al $\beta = 90\%$ per la media della popolazione μ è un intervallo $[U_n, V_n]$ con estremi variabili aleatorie tali che

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

Consideriamo la media campionaria $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ e la varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_n - \bar{X}_n)^2$. Ricordiamo che, poiché $X_1, \ldots, X_n$ hanno distribuzione normale, la variabile aleatoria $T_{n-1} = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ è distribuita come una t-Student con $n-1 = 7$ gradi di libertà. Utilizziamo questo fatto per determinare gli estremi dell'intervallo di confidenza:

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

da cui segue

$$\mathbb{P}\Big(\Big\{T_{n-1} > \frac{\bar{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) + \mathbb{P}\Big(\Big\{T_{n-1} < \frac{\bar{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big) = 1 - \beta = \alpha = 10\%\,.$$

Decidiamo di equipartire la probabilità α, ovvero

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

Sia $t_{n-1,\alpha/2}$ il quantile della t-Student tale che

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

Allora scegliendo

$$\frac{\bar{X}_n - U_n}{S_n/\sqrt{n}} = t_{n-1,\alpha/2}\,, \quad \frac{\bar{X}_n - V_n}{S_n/\sqrt{n}} = -t_{n-1,\alpha/2}$$

abbiamo, per la simmetria della t-Student, che (1) è verificata. Quindi

$$U_n = \bar{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\,, \quad V_n = \bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\,.$$

Abbiamo tutti gli strumenti per calcolare la realizzazione degli estremi sui dati $x_1, \ldots, x_n$ osservati. La realizzazione della media campionaria sui dati è

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{8}(230.66+233.05+232.58+229.48+232.58+235.76+229.43+234.13) \simeq 232.21\,.$$

La realizzazione della deviazione standard sui dati è

$$\begin{aligned}
s_n &= \sqrt{\frac{1}{n-1}\Big(\sum_{i=1}^n x_i^2 - n\bar{x}_n^2\Big)} = \sqrt{\frac{1}{7}(431402.0807 - 431371.8728)} \\
&= \sqrt{4.3154142857142857143} \simeq 2.08\,.
\end{aligned}$$

Il quantile della t-Student è calcolato dalle tabelle

$$t_{n-1,\alpha/2} = t_{7,0.05} \simeq 1.895\,.$$

Concludiamo che la realizzazione degli estremi sui dati è

$$u_n = \bar{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}} \simeq 230.82\,, \quad v_n = \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}} \simeq 233.60$$

e l'intervallo di confidenza calcolato sui dati è $[230.82, 233.60]$.

2. Consideriamo l'esperimento $j = 1, \dots, 20$ e la variabile aleatoria Y_j

$$Y_j = \begin{cases} 1 & \text{se la media della popolazione è fuori dall'intervallo di confidenza}\,, \\ 0 & \text{se la media della popolazione è nell'intervallo di confidenza}\,. \end{cases}$$

Osserviamo che Y_j è distribuita come una Bernoulli con probabilità di successo $p = 10\% = \frac{1}{10}$, poiché la popolazione è fuori dall'intervallo di confidenza con il 10% di probabilità. Le $Y_1, \dots, Y_{20}$ sono indipendenti (lo dice la traccia), quindi la variabile aleatoria

$$Y = \text{“numero di successi in 20 esperimenti statistici”} = Y_1 + \cdots + Y_{20}$$

è distribuita con legge binomiale $B(20, \frac{1}{10})$. Ricordando che $\mathbb{P}(\{Y = k\}) = \binom{n}{k} p^k (1-p)^{n-k}$, possiamo calcolare la probabilità di osservare più di 4 volte un successo (media della popolazione è fuori dall'intervallo di confidenza):

$$\begin{aligned} &\mathbb{P}(\{Y \geq 4\}) = 1 - \mathbb{P}(\{Y \leq 3\}) = 1 - \mathbb{P}(\{Y = 0\}) - \mathbb{P}(\{Y = 1\}) - \mathbb{P}(\{Y = 2\}) - \mathbb{P}(\{Y = 3\}) \\ &= 1 - \binom{20}{0}\left(\frac{1}{10}\right)^0\left(\frac{9}{10}\right)^{20} - \binom{20}{1}\left(\frac{1}{10}\right)^1\left(\frac{9}{10}\right)^{19} - \binom{20}{2}\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^{18} - \binom{20}{3}\left(\frac{1}{10}\right)^3\left(\frac{9}{10}\right)^{17} \\ &\simeq 13.30\%\,. \end{aligned}$$

Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni 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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: gennaio 2023
Corso di studi: _____ Data: 26/01/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) I seguenti dati indicano la relazione tra velocità di lettura (parole al minuto) e il numero di settimane trascorse in un programma di lettura veloce per 10 studenti:

settimane	2	3	8	11	4	5	9	7	5	7
velocità di lettura	21	42	102	130	52	57	105	85	62	90

1. Rappresentare i dati in uno scatterplot.
2. Determinare (derivando le formule dei coefficienti) e rappresentare la retta di regressione lineare.
3. Calcolare il coefficiente di correlazione.

Soluzione. 1. Lo scatterplot è rappresentato in Figura 1.

2. Denotiamo con $(x_1, y_1), \dots, (x_n, y_n)$, $n = 10$, 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

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

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

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

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

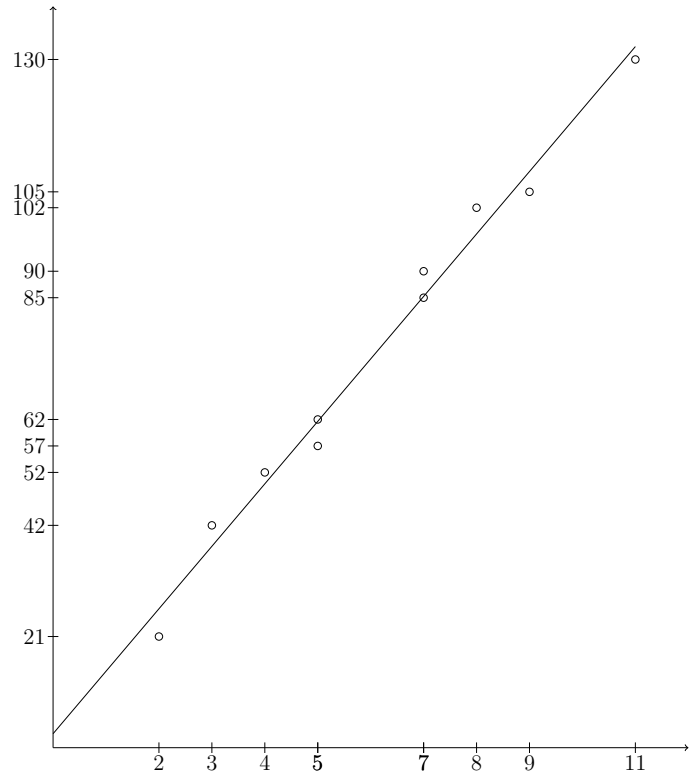


Figura 1: Scatterplot e retta di regressione lineare.

Dalla seconda equazione segue che

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

Sostituendo nella prima,

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

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

											somma
x_i	2	3	8	11	4	5	9	7	5	7	61
y_i	21	42	102	130	52	57	105	85	62	90	746
x_i^2	4	9	64	121	16	25	81	49	25	49	443
y_i^2	441	1764	10404	16900	2704	3249	11025	7225	3844	8100	65656
x_iy_i	42	126	816	1430	208	285	945	595	310	630	5387

Pertanto $\overline{x} = 61/10 = 6.1$ e $\overline{y} = 746/10 = 74.6$. Segue che

$$a = \frac{5387 - 10\cdot 6.1\cdot 74.6}{443 - 10\cdot 6.1^2} = \frac{836.4}{70.9} \simeq 11.80\,,$$

$$b = 74.6 - 11.80 \cdot 6.1 \simeq 2.64\,,$$

ovvero, la retta di regressione lineare ha equazione

$$y = 11.80x + 2.64\,.$$

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

$$\rho_{x,y} = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}^2}} = \frac{5387 - 10 \cdot 6.1 \cdot 74.6}{\sqrt{443 - 10 \cdot 6.1^2}\sqrt{65656 - 10 \cdot 74.6^2}}$$
$$= \frac{836.4}{\sqrt{70.9}\sqrt{10004.4}} \simeq 0.9931\,.$$

Esercizio 2. (8 punti) Si consideri un vettore aleatorio discreto (X_1, X_2) con funzione di probabilità congiunta data dalla seguente tabella:

X_2 \ X_1	0	1	2	3
0	$a_{00}/8$	$a_{10}/8$	$a_{20}/8$	$a_{30}/8$
1	$a_{01}/8$	$a_{11}/8$	$a_{21}/8$	$a_{31}/8$

dove $a_{00}, a_{10}, a_{20}, a_{30}, a_{01}, a_{11}, a_{21}, a_{31} \geq 0$.

1. Trovare i valori espliciti di a_{ij} nella tabella sapendo che:
   - X_1 ha legge binomiale $\mathrm{B}(n,p)$ con parametri $n = 3$ e $p = \frac{1}{2}$.
   - X_2 ha legge di Bernoulli $\mathrm{Be}(q)$ con parametro $q = \frac{7}{8}$ ($1 =$ successo).
   - $\mathbb{P}(\{X_1 = 3\}|\{X_2 = 1\}) = \frac{1}{14}$.
   - $\mathbb{P}(\{X_1 + X_2 = 1\}) = \frac{1}{8}$.
   - $\mathbb{P}(\{X_1 = 2\}, \{X_2 = 0\}) = \frac{1}{16}$.

   (Suggerimento: usare le condizioni nell'ordine in cui sono fornite.)
2. Calcolare la covarianza di X_1 e X_2.
3. Le variabili aleatorie X_1 e X_2 sono indipendenti?

Soluzione. 1. Poiché $X_1 \sim \mathrm{B}(n,p)$ con $n = 3$ e $p = \frac{1}{2}$, abbiamo che (scrivendo le probabilità marginali)

$$\frac{1}{8} = (1-p)^3 = \mathbb{P}(\{X_1 = 0\}) = \frac{1}{8}(a_{00} + a_{01})\,, \tag{1}$$
$$\frac{3}{8} = \binom{3}{1} p(1-p)^2 = \mathbb{P}(\{X_1 = 1\}) = \frac{1}{8}(a_{10} + a_{11})\,, \tag{2}$$
$$\frac{3}{8} = \binom{3}{2} p^2(1-p) = \mathbb{P}(\{X_1 = 2\}) = \frac{1}{8}(a_{20} + a_{21})\,, \tag{3}$$
$$\frac{1}{8} = p^3 = \mathbb{P}(\{X_1 = 3\}) = \frac{1}{8}(a_{30} + a_{31})\,. \tag{4}$$

Poiché $X_2 \sim \mathrm{Be}(\frac{7}{8})$ abbiamo che

$$\frac{1}{8} = \mathbb{P}(\{X_2 = 0\}) = \frac{1}{8}(a_{00} + a_{10} + a_{20} + a_{30})\,, \tag{5}$$
$$\frac{7}{8} = \mathbb{P}(\{X_2 = 1\}) = \frac{1}{8}(a_{01} + a_{11} + a_{21} + a_{31})\,. \tag{6}$$

Osservo che la condizione

$$\frac{1}{8}(a_{00} + a_{10} + a_{20} + a_{30} + a_{01} + a_{11} + a_{21} + a_{31}) = 1\,, \tag{7}$$

che si ottiene imponendo che la somma di tutte le probabilità sia 1, è superflua. Infatti si ottiene già sommando (1)–(4). Anche la (6) è superflua: segue automaticamente sottraendo (5) da (7). Abbiamo quindi il sistema

$$\begin{cases} a_{00} + a_{01} = 1\,, \\ a_{10} + a_{11} = 3\,, \\ a_{20} + a_{21} = 3\,, \\ a_{30} + a_{31} = 1\,, \\ a_{00} + a_{10} + a_{20} + a_{30} = 1\,. \end{cases} \tag{8}$$

Seguiamo il suggerimento e sfruttiamo una condizione alla volta. Da $\mathbb{P}(\{X_1 = 3 | X_2 = 1\}) = \frac{1}{14}$ segue, usando la formula della probabilità condizionata e (6), che

$$\frac{1}{14} = \mathbb{P}(\{X_1 = 3 | X_2 = 1\}) = \frac{\mathbb{P}(\{X_1 = 3, X_2 = 1\})}{\mathbb{P}(\{X_2 = 1\})} = \frac{a_{31}/8}{7/8} = \frac{a_{31}}{7} \implies a_{31} = \frac{1}{2}\,.$$

Sostituendo in (8):

$$\begin{cases} a_{00} + a_{01} = 1\,, \\ a_{10} + a_{11} = 3\,, \\ a_{20} + a_{21} = 3\,, \\ a_{30} = \frac{1}{2}\,, \\ a_{00} + a_{10} + a_{20} = \frac{1}{2}\,, \\ a_{31} = \frac{1}{2}\,. \end{cases} \tag{9}$$

Utilizziamo la condizione $\mathbb{P}(\{X_1 + X_2 = 1\}) = \frac{1}{8}$. Si ha che

$$\frac{1}{8} = \mathbb{P}(\{X_1 + X_2 = 1\}) = \mathbb{P}(\{X_1 = 0, X_2 = 1\}) + \mathbb{P}(\{X_1 = 1, X_2 = 0\}) = \frac{1}{8}(a_{01} + a_{10})$$

da cui

$$a_{01} + a_{10} = 1\,.$$

Sostituendo nella prima equazione nel sistema (9) otteniamo che

$$a_{01} + a_{10} = a_{00} + a_{01} \implies a_{10} = a_{00}\,.$$

Il sistema diventa quindi

$$\begin{cases} a_{00} = a_{10}\,, \\ a_{10} + a_{11} = 3\,, \\ a_{20} + a_{21} = 3\,, \\ a_{30} = \dfrac{1}{2}\,, \\ 2a_{00} + a_{20} = \dfrac{1}{2}\,, \\ a_{31} = \dfrac{1}{2}\,, \\ a_{01} + a_{10} = 1\,, \end{cases} \tag{10}$$

L'ultima condizione $\mathbb{P}(\{X_1 = 2\}, \{X_2 = 0\}) = \frac{1}{16}$ implica che $a_{20}/8 = \frac{1}{16}$, cioè $a_{20} = \frac{1}{2}$. Quindi

$$\begin{cases} a_{00} = a_{10}\,, \\ a_{10} + a_{11} = 3\,, \\ a_{20} + a_{21} = 3\,, \\ a_{30} = \dfrac{1}{2}\,, \\ 2a_{00} + a_{20} = \dfrac{1}{2}\,, \\ a_{31} = \dfrac{1}{2}\,, \\ a_{01} + a_{10} = 1\,, \\ a_{20} = \dfrac{1}{2}\,. \end{cases} \implies \begin{cases} a_{10} = 0\,, \\ a_{11} = 3\,, \\ a_{21} = \dfrac{5}{2}\,, \\ a_{30} = \dfrac{1}{2}\,, \\ a_{00} = 0\,, \\ a_{31} = \dfrac{1}{2}\,, \\ a_{01} = 1\,, \\ a_{20} = \dfrac{1}{2}\,. \end{cases} \tag{11}$$

La tabella completa è quindi

X_2 \ X_1	0	1	2	3
0	0	0	1/16	1/16
1	1/8	3/8	5/16	1/16

2. Per calcolare la covarianza osserviamo che $\mathbb{E}(X_1) = np = \frac{3}{2}$ mentre $\mathbb{E}(X_2) = q = \frac{7}{8}$, per come sono le leggi delle variabili aleatorie. Dobbiamo calcolare $\mathbb{E}(X_1 \cdot X_2)$. Il range di $X_1 \cdot X_2$ è dato da $\{0, 1, 2, 3\}$, quindi

$$\begin{aligned} &\mathbb{E}(X_1 \cdot X_2) \\ &= 0 \cdot \mathbb{P}(\{X_1 \cdot X_2 = 0\}) + 1 \cdot \mathbb{P}(\{X_1 \cdot X_2 = 1\}) + 2 \cdot \mathbb{P}(\{X_1 \cdot X_2 = 2\}) + 3 \cdot \mathbb{P}(\{X_1 \cdot X_2 = 3\}) \\ &= 1 \cdot \mathbb{P}(\{X_1 = 1, X_2 = 1\}) + 2 \cdot \mathbb{P}(\{X_1 = 2, X_2 = 1\}) + 3 \cdot \mathbb{P}(\{X_1 = 3, X_2 = 1\}) \\ &= \frac{3}{8} + 2\frac{5}{16} + 3\frac{1}{16} = \frac{19}{16}\,. \end{aligned}$$

Concludiamo che

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1 \cdot X_2) - \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) = \frac{19}{16} - \frac{3}{2} \cdot \frac{7}{8} = -\frac{1}{8}.$$

3. Le variabili non sono indipendenti: la covarianza non è nulla (il fatto che la covarianza si annulli è una condizione necessaria per l'indipendenza).

Esercizio 3. (7 punti) Un'azienda produce uno smartphone con una vita media di 4 anni, dopodiché si rompe. Assumiamo che la vita dello smartphone (misurata in anni) sia una variabile aleatoria con legge esponenziale.

1. Acquisti uno smartphone. Qual è la probabilità che funzioni per più di 6 anni?
2. Acquisti uno smartphone. Passano 3 anni e funziona ancora. Qual è la probabilità che funzioni in tutto per più di 6 anni, sapendo che è successo il fatto precedente?
3. Acquisti tre smartphone (assumiamo che le vite dei tre smartphone siano indipendenti). qual è la probabilità che almeno due dei tre funzionino per più di 6 anni?
4. (**Bonus**) Acquisti due smartphone (assumiamo che le vite dei due smartphone siano indipendenti). Ne usi solo uno, finché si rompe (assumiamo che intanto lo smartphone non utilizzato non perda anni di vita). Poi inizi a usare l'altro (che ha una vita media di 4 anni). Chiamiamo vita cumulata la somma delle vite dei due smartphone. Qual è la probabilità che la vita cumulata sia più di 12 anni?

Soluzione. 1. Sia $X \sim \mathrm{Exp}(\lambda)$ la vita (in anni) dello smartphone. Ricordiamo che $\mathbb{E}(X) = \frac{1}{\lambda}$, quindi $\frac{1}{\lambda} = 4$, cioè $\lambda = \frac{1}{4}$. La probabilità che lo smartphone funzioni più di 6 anni è

$$\mathbb{P}(\{X \geq 6\}) = \int_6^{+\infty} \lambda e^{-\lambda x}\, \mathrm{d}x = \Big[- e^{-\lambda x} \Big]_6^{+\infty} = e^{-\lambda \cdot 6} = e^{-3/2} \simeq 22.31\% \,.$$

2. La legge esponenziale gode della proprietà di assenza di memoria, quindi

$$\mathbb{P}(\{X \geq 6\} | \{X \geq 3\}) = \mathbb{P}(\{X \geq 3\}) = \int_3^{+\infty} \lambda e^{-\lambda x}\, \mathrm{d}x = e^{-3/4} = 47.24\% \,.$$

3. Introduciamo la variabile $Y \sim \mathrm{B}(3, p)$ con $p = \mathbb{P}(\{X \geq 6\})$ (durata più di 6 anni è un successo). Ci viene chiesta la probabilità di almeno due successi in 3 prove, ovvero

$$\begin{aligned} \mathbb{P}(\{Y \geq 2\}) &= \mathbb{P}(\{Y = 2\}) + \mathbb{P}(\{Y = 3\}) = \binom{3}{2} p^2 (1-p) + \binom{3}{3} p^3 \\ &= 3(e^{-3/2})^2 (1 - e^{-3/2}) + e^{-9/4} \simeq 12.71\% \,. \end{aligned}$$

Possiamo anche risolvere l'esercizio direttamente senza l'uso della binomiale. Denotiamo con $X_1, X_2, X_3 \sim \mathrm{Exp}(\frac{1}{4})$ le vite dei tre smartphone (indipendenti). Almeno due smartphone funzionano per più di 6 anni se si verificano almeno due (anche tre) delle disuguaglianze $X_1 \geq 6, X_2 \geq 6, X_3 \geq 6$, ovvero

$$\begin{aligned} &\mathbb{P}(\{X_1 \geq 6\} \cap \{X_2 \geq 6\} \cap \{X_3 < 6\}) \\ &+ \mathbb{P}(\{X_1 \geq 6\} \cap \{X_2 < 6\} \cap \{X_3 \geq 6\}) \\ &+ \mathbb{P}(\{X_1 < 6\} \cap \{X_2 \geq 6\} \cap \{X_3 \geq 6\}) \\ &+ \mathbb{P}(\{X_1 \geq 6\} \cap \{X_2 \geq 6\} \cap \{X_3 \geq 6\}) \,. \end{aligned}$$

Usando l'indipendenza delle tre variabili si ottiene lo stesso risultato.

4. Denotiamo con $X_1, X_2 \sim \mathrm{Exp}(\lambda)$ i tempi di vita dei due smartphone. Siamo interessati alla variabile aleatoria $X_1 + X_2$. Ricordiamo che $X_1, X_2 \sim \mathrm{Gamma}(1, \lambda)$ e sono indipendenti,

pertanto $X_1+X_2 \sim \text{Gamma}(2,\lambda)$. Integrando per parti e usando il fatto che $\Gamma(2) = (2-1)! = 1$, calcoliamo

$$\begin{aligned}\mathbb{P}(\{X_1+X_2 \geq 12\}) &= \int_{12}^{+\infty} \frac{\lambda^2}{\Gamma(2)} x e^{-\lambda x}\, \mathrm{d}x = \lambda\Big[-xe^{-\lambda x}\Big]_{12}^{+\infty} + \int_{12}^{+\infty} \lambda e^{-\lambda x}\, \mathrm{d}x \\ &= 12\lambda e^{-12\lambda} + \Big[-e^{-\lambda x}\Big]_{12}^{+\infty} = 12\lambda e^{-12\lambda} + e^{-12\lambda} = 4e^{-3} \simeq 19.91\%\,.\end{aligned}$$

In alternativa, se non si ricordano le proprietà della legge Gamma, si può usare la formula per la densità della somma di due variabili indipendenti:

$$f_{X_1+X_2}(x) = f_{X_1} * f_{X_2}(x) = \int_{\mathbb{R}} f_{X_1}(y) f_{X_2}(x-y)\, \mathrm{d}y = \int_0^x \lambda e^{-\lambda y} \lambda e^{-\lambda x+\lambda y}\, \mathrm{d}y = \lambda^2 x e^{-\lambda x}$$

e continuare con il conto di sopra.

Esercizio 4. (7 punti) Un'azienda produce una margarina dietetica per cui si sa, fino a prova contraria, che il livello di acidi grassi polinsaturi (in percentuale) ha una deviazione standard di 1.2. È stata proposta una nuova tecnica di produzione del prodotto, che tuttavia comporta un costo aggiuntivo. La direzione autorizzerà un cambiamento nella tecnica di produzione se si riesce a mostrare che la deviazione standard del livello di acidi grassi polinsaturi con il nuovo processo è significativamente inferiore a 1.2. Un campione del lotto ottenuto con il nuovo metodo ha prodotto le seguenti percentuali di livello di acidi grassi polinsaturi:

16.8 17.2 17.4 16.9 16.5 17.1 18.2 16.8 15.7 16.1

Si assuma che i dati siano distribuiti con legge normale.

1. I dati sono significativi al 5% per decidere di cambiare il metodo di produzione? (N.B.: Derivare le formule)

2. Siamo interessati al più piccolo livello di significatività per cui i dati porterebbero a decidere di cambiare il metodo di produzione. In quale di questi intervalli si colloca tale valore: $[0.5\%, 1\%)$, $[1\%, 2.5\%)$, $[2.5\%, 5\%)$, $[5\%, 10\%)$?

Soluzione. Stiamo considerando un campione casuale $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ con $n = 10$ e dove μ e σ sono incognite. Si deve impostare un test di ipotesi sulla varianza:

$$H_0 : \sigma^2 = \sigma_0^2 = 1.2^2 = 1.44\,, \quad H_1 : \sigma^2 < \sigma_0^2 = 1.44\,.$$

Infatti ci si sta chiedendo se i dati sono abbastanza significativi da rifiutare l'ipotesi che la deviazione standard sia uguale a 1.2, a favore dell'ipotesi alternativa (deviazione standard strettamente più piccola di 1.2).

Per svolgere il test di ipotesi, scegliamo la regione critica della forma

$$R_C = \{(x_1, \ldots, x_n) \in R(X_1, \ldots, X_n) \ : s_n^2 < c\sigma_0^2\}$$

dove $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$ è la varianza campionaria calcolata su valori $(x_1, \ldots, x_n)$. La costante c nella definizione della regione critica deve essere definita in termini del livello di significatività α. Questo è la probabilità di commettere un errore del I tipo. Supponiamo allora che l'ipotesi nulla sia vera, cioè $\sigma^2 = \sigma_0^2 = 1.44$. La probabilità di commettere un errore

del I tipo (cioè di rifiutare l'ipotesi nulla a favore dell'ipotesi alternativa) è, usando la varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ (stimatore corretto della varianza),

$$\alpha = \mathbb{P}(\{(X_1,\dots,X_n) \in R_C\}) = \mathbb{P}(\{S_n^2 < c\sigma_0^2\}) = \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{\sigma_0^2} < (n-1)c\Big\}\Big)\,.$$

Ricordiamo che, poiché $X_1,\dots,X_n$ sono distribuite con legge normale e $\sigma^2 = \sigma_0^2$ per l'ipotesi nulla, si ha che $\Xi_{n-1} = \frac{(n-1)S_n^2}{\sigma_0^2}$ è distribuita con una legge chi-quadro con $n-1 = 9$ gradi di libertà. Allora

$$\alpha = \mathbb{P}(\{\Xi_{n-1} < (n-1)c\}) = 1 - \mathbb{P}(\{\Xi_{n-1} \geq (n-1)c\}) \implies \mathbb{P}(\{\Xi_{n-1} \geq (n-1)c\}) = 1-\alpha\,.$$

Chiamiamo $\chi_{n-1,1-\alpha}$ il quantile della chi-quadro che verifica

$$\mathbb{P}(\{\Xi_{n-1} \geq \chi_{n-1,1-\alpha}\}) = 1-\alpha\,.$$

In questo modo, scegliendo $c = \frac{\chi_{n-1,1-\alpha}}{n-1}$, otteniamo la condizione imposta all'inizio.

Possiamo ora prendere una decisione sul test di ipotesi calcolando i valori sui dati. Consultando le tavole della chi-quadro con $\alpha = 5\%$ otteniamo che

$$\chi_{n-1,1-\alpha} = \chi_{9,0.95} \simeq 3.325\,.$$

Quindi

$$c\sigma_0^2 = \frac{\chi_{n-1,1-\alpha}}{n-1}\sigma_0^2 = \frac{3.325}{9}1.44 \simeq 0.53\,.$$

Calcoliamo la media e la varianza sul campione:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{10}(16.8+17.2+17.4+16.9+16.5+17.1+18.2+16.8+15.7+16.1) = 16.87\,,$$

$$\begin{aligned} s_n^2 &= \frac{1}{n-1}\Big(\sum_{i=1}^n x_i^2 - n\overline{x}^2\Big) \\ &= \frac{1}{9}(16.8^2+17.2^2+17.4^2+16.9^2+16.5^2+17.1^2+18.2^2+16.8^2+15.7^2+16.1^2-10\cdot 16.87^2) \\ &\simeq 0.48 \end{aligned}$$

Poiché $0.48 < 0.53$, rifiutiamo l'ipotesi nulla a favore dell'ipotesi alternativa.

2. Ci basta calcolare come cambia la soglia della regione critica a seconda di α:

$$\begin{aligned} \alpha = 5\% &\implies c\sigma_0^2 = \frac{\chi_{9,0.95}}{9}1.44 \simeq \frac{3.325}{9}1.44 \simeq 0.53\,, \\ \alpha = 2.5\% &\implies c\sigma_0^2 = \frac{\chi_{9,0.975}}{9}1.44 \simeq \frac{2.700}{9}1.44 \simeq 0.432\,. \end{aligned}$$

Possiamo terminare qui: con il 2.5% di significatività l'ipotesi nulla non può essere rifiutata. Quindi il più piccolo livello di significatività che porta a rifiutare l'ipotesi nulla è nell'intervallo $[2.5\%, 5\%)$.

Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni 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: _____ A.A.: 2021/2022
Nome: _____ Docente: Gianluca Orlando
Matricola: _____ Appello: febbraio 2023
Corso di studi: _____ Data: 09/02/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Viene esaminata in un campione la resistenza alla compressione del calcestruzzo quando miscelato con la cenere volante (una miscela di silice, allumina, ossido di ferro, e altri elementi). Vengono riportati i seguenti dati in megapascal:

22.4 50.2 30.4 14.2 28.9 30.5 25.8 18.4 15.3 21.1

1. Calcolare i quartili dell'insieme dei dati.
2. Determinare eventuali dati anomali o sospetti.
3. Disegnare un box-plot.

Soluzione. 1. Per prima cosa ordiniamo i dati:

14.2 15.3 18.4 21.1 22.4 25.8 28.9 30.4 30.5 50.2

Abbiamo $n = 10$ dati.

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

$$Q_1 = (1 - 0.75)x_2 + 0.75x_3 = 0.25 \cdot 15.3 + 0.75 \cdot 18.4 = 17.625\,.$$

Calcoliamo il secondo quartile: $(n+1)\frac{2}{4} = \frac{22}{4} = 5 + 0.5$. Allora

$$Q_2 = (1 - 0.5)x_5 + 0.5x_6 = 0.5 \cdot 22.4 + 0.5 \cdot 25.8 = 24.1\,.$$

Calcoliamo il terzo quartile: $(n+1)\frac{3}{4} = \frac{33}{4} = 8 + 0.25$. Allora

$$Q_3 = (1 - 0.25)x_8 + 0.25x_9 = 0.75 \cdot 30.4 + 0.25 \cdot 30.5 = 30.425\,.$$

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

$$IQR = Q_3 - Q_1 = 30.425 - 17.625 = 12.8\,.$$

I dati anomali apparterrebbero agli intervalli

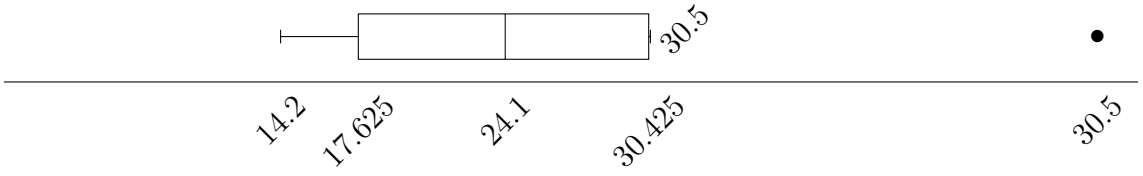
$$(-\infty, Q_1 - 3 \cdot IQR] \cup [Q_3 + 3 \cdot IQR, +\infty) = (-\infty, -20.775] \cup [68.825, +\infty)\,,$$

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

$$(Q_1 - 3 \cdot IQR, Q_1 - 1.5 \cdot IQR] \cup [Q_3 + 1.5 \cdot IQR, Q_3 + 3 \cdot IQR) = (-20.775, -1.575] \cup [49.625, 68.825)\,,$$

quindi 50.2 è un dato sospetto.

3. Segue il box-plot.



Esercizio 2. (7 punti) Alice e Bob fanno un gioco. Alice lancia consecutivamente un dado a 6 facce non truccato tante volte (i lanci sono indipendenti). Se entro il quarto lancio (compreso) esce un 6, vince Bob, altrimenti vince Alice.

1. Alice e Bob giocano una partita. Con che probabilità vince Alice?

Alice però è molto brava nel calcolo delle probabilità e si rende conto che c'è qualcosa che va a suo sfavore nelle regole di questo gioco... Decide allora di truccare il dado, all'insaputa di Bob. Dopo aver giocato alcune partite, Bob si rende conto che Alice vince con il 70% di probabilità. Inoltre, il 6 esce per la prima volta, in media, troppo tardi. Quindi l'accusa di aver barato.

2. Cosa ha fatto esattamente Alice al dado?

3. In media, dopo quale lancio esce il 6 con questo dado truccato?

Soluzione. 1. Il lancio di un dado non truccato è modellato da uno spazio di probabilità con spazio degli eventi elementari

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

in cui ciascun evento elementare ha probabilità $\frac{1}{6}$. La probabilità che esca un 6 in un lancio di un dado è quindi $\frac{1}{6}$. Chiamiamo "successo" l'evento "esce un 6".

Il gioco in questione è una sequenza di prove indipendenti, ciascuna con probabilità di successo $\frac{1}{6}$. La variabile aleatoria X (definita su uno spazio di probabilità differente) che fornisce la prima volta che viene osservato un successo in una sequenza di lanci indipendenti ha una distribuzione geometrica con parametro $\frac{1}{6}$, ovvero $X \sim \mathrm{Geo}(\frac{1}{6})$. Alice vince se il primo 6 non esce entro il quarto lancio (compreso), ovvero se esce dal quinto lancio in poi. La probabilità che si verifichi questo evento è (in questa formula $p = \frac{1}{6}$)

$$\begin{aligned}
\mathbb{P}(\{X \geq 5\}) &= \sum_{i=5}^{+\infty} \mathbb{P}(\{X = i\}) = \sum_{i=5}^{+\infty} (1-p)^{i-1} p = (1-p)^4 p \sum_{i=5}^{+\infty} (1-p)^{i-5} \\
&= (1-p)^4 p \sum_{j=0}^{+\infty} (1-p)^j = (1-p)^4 p \frac{1}{1-(1-p)} = (1-p)^4 \\
&= \left(1 - \frac{1}{6}\right)^4 = \left(\frac{5}{6}\right)^4 \simeq 48.23\%\,.
\end{aligned}$$

Si può anche ricordare direttamente la formula $(1-p)^4$, che va interpretata come: "I primi quattro risultati delle prove sono insuccessi".

In alternativa, poiché i tentativi interessanti sono pochi, si può calcolare direttamente la probabilità del complementare

$$\begin{aligned}\mathbb{P}(\{X \geq 5\}) &= 1 - \mathbb{P}(\{X \leq 4\}) = 1 - \mathbb{P}(\{X=1\}) - \mathbb{P}(\{X=2\}) - \mathbb{P}(\{X=3\}) - \mathbb{P}(\{X=4\}) \\ &= 1 - (1-p)^0 p - (1-p)^1 p - (1-p)^2 p - (1-p)^3 p \\ &= 1 - \frac{1}{6} - \left(\frac{5}{6}\right)\frac{1}{6} - \left(\frac{5}{6}\right)^2\frac{1}{6} - \left(\frac{5}{6}\right)^3\frac{1}{6} \simeq 48.23\%\,.\end{aligned}$$

Osserviamo che la probabilità di vittoria per Alice è leggermente meno del 50%, e questo è in accordo con la continuazione della traccia.

2. Alice ha truccato il dado in modo che il 6 esca con una probabilità diversa che chiamiamo p, diversa da $\frac{1}{6}$ e da determinare. Per capire esattamente con che probabilità esce il 6 con il dado truccato, usiamo l'informazione sulle partite vinte. Ci viene detto che Alice vince con il 70% di probabilità. Questo vuol dire, usando la formula del punto 1.,

$$0.7 = 70\% = \mathbb{P}(\{X \geq 5\}) = (1-p)^4 \implies p = 1 - (0.7)^{\frac{1}{4}} \simeq 8.53\,.$$

Osserviamo che questa probabilità è meno di $\frac{1}{6} \simeq 16.67$, infatti Alice vuole che il 6 esca con una probabilità più bassa. Il dado truccato ha quindi una distribuzione non uniforme, ad esempio può essere fatto in modo che per gli eventi elementari dello spazio Ω introdotto inizialmente si abbia che

$$\mathbb{P}(\{6\}) = p\,, \quad \mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \frac{1-p}{5}\,.$$

(Questa è solo una possibile configurazione, ce ne sono infinite per cui si ha la condizione richiesta.)

3. Ora la $X \sim \text{Geo}(p)$. Sappiamo che $\mathbb{E}(X) = \frac{1}{p} \simeq 11.72$, quindi in media il 6 esce all'incirca dopo l'undicesimo lancio (molto di più rispetto al dado non truccato).

Esercizio 3. (7 punti) Il tempo necessario per un/a tecnico/a dell'assistenza per cambiare l'olio in un'auto è una variabile aleatoria. Se l'auto non presenta problemi, è distribuita uniformemente tra 10 e 20 minuti. Se l'auto presenta dei problemi, è distribuita uniformemente con media 20 minuti e varianza 12 min^2. In media, il 90% delle auto non presenta problemi.

1. Viene riparata un'auto. Sapendo che il tempo impiegato per cambiare l'olio è stato maggiore di 18 minuti, qual è la probabilità che l'auto avesse dei problemi?

2. Vengono riparate 2 auto (si assumano i tempi di cambio dell'olio per le due auto indipendenti). Qual è la probabilità che il cambio d'olio più rapido duri meno di 18 minuti?

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{"tempo impiegato per cambio dell'olio"}$$

e la variabile aleatoria

$$Y = \text{"auto non presenta problemi"}\,.$$

Per quanto riguarda la seconda variabile, possiamo utilizzare una legge di Bernoulli $Y \sim \text{Be}(p)$ con $p = 90\%$ dove $\{Y = 1\}$ (successo) significa che l'auto non presenta problemi, mentre

$\{Y = 0\}$ (insuccesso) significa che l'auto presenta problemi. Infatti ricordiamo che $\mathbb{E}(Y) = p$ e ci viene detto che la media è 90%.

Per semplicità introduciamo le due variabili aleatorie ausiliarie

$$U_1 = \text{“tempo impiegato per cambio dell'olio se l'auto non presenta problemi”}$$

$$U_0 = \text{“tempo impiegato per cambio dell'olio se l'auto presenta problemi”}\,.$$

La traccia ci dice che $U_1 \sim \mathrm{U}(10, 20)$, ovvero, la densità è

$$f_{U_1}(x) = \begin{cases} \frac{1}{10}\,, & \text{se } 10 \le x \le 20\,, \\ 0\,, & \text{altrimenti}\,. \end{cases}$$

Per la seconda variabile si sa che $U_0 \sim \mathrm{U}(a, b)$, $\mathbb{E}(U_0) = 20$, $\mathrm{Var}(U_0) = 12$. Ricordiamo che per una variabile aleatoria con legge uniforme si ha che

$$\mathbb{E}(U_0) = \frac{a+b}{2}\,, \quad \mathrm{Var}(U_0) = \frac{(b-a)^2}{12}\,.$$

Imponiamo che

$$\begin{cases} \dfrac{a+b}{2} = 20 \\ \dfrac{(b-a)^2}{12} = 12 \end{cases} \implies \begin{cases} a+b = 40 \\ (b-a)^2 = 12^2 \end{cases} \implies \begin{cases} a+b = 40 \\ b-a = 12 \end{cases} \implies \begin{cases} 2b = 52 \\ 2a = 28 \end{cases} \implies \begin{cases} b = 26 \\ a = 14\,. \end{cases}$$

(Abbiamo usato che $b - a > 0$!) Quindi $X_2 \sim \mathrm{U}(14, 26)$, cioè

$$f_{U_0}(x) = \begin{cases} \frac{1}{12}\,, & \text{se } 14 \le x \le 26\,, \\ 0\,, & \text{altrimenti}\,. \end{cases}$$

Infine, sappiamo che

$$\mathbb{P}(\{X \ge x\}|\{Y = 1\}) = \mathbb{P}(\{U_1 \ge x\})\,, \quad \mathbb{P}(\{X \ge x\}|\{Y = 0\}) = \mathbb{P}(\{U_0 \ge x\})\,. \tag{1}$$

1. Ci viene chiesto di calcolare la probabilità condizionata

$$\mathbb{P}(\{Y = 0\}|\{X \ge 18\})\,.$$

Utilizziamo il Teorema di Bayes e (1):

$$\begin{aligned} \mathbb{P}(\{Y = 0\}|\{X \ge 18\}) &= \frac{\mathbb{P}(\{Y = 0\} \cap \{X \ge 18\})}{\mathbb{P}(\{X \ge 18\})} \\ &= \frac{\mathbb{P}(\{X \ge 18\}|\{Y = 0\})\mathbb{P}(\{Y = 0\})}{\mathbb{P}(\{X \ge 18\}|\{Y = 0\})\mathbb{P}(\{Y = 0\}) + \mathbb{P}(\{X \ge 18\}|\{Y = 1\})\mathbb{P}(\{Y = 1\})} \\ &= \frac{\mathbb{P}(\{U_0 \ge 18\})\mathbb{P}(\{Y = 0\})}{\mathbb{P}(\{U_0 \ge 18\})\mathbb{P}(\{Y = 0\}) + \mathbb{P}(\{U_1 \ge 18\})\mathbb{P}(\{Y = 1\})} \\ &= \frac{\int_{18}^{26} \frac{1}{12}\,\mathrm{d}x \cdot 10\%}{\int_{18}^{26} \frac{1}{12}\,\mathrm{d}x \cdot 10\% + \int_{18}^{20} \frac{1}{10}\,\mathrm{d}x \cdot 90\%} = \frac{8/12 \cdot 1/10}{8/12 \cdot 1/10 + 2/10 \cdot 9/10} \simeq 27.03\,. \end{aligned}$$

2. Abbiamo a che fare con due auto, quindi due tempi di cambio dell'olio X_1, X_2, variabili indipendenti e identicamente distribuite. Ci viene chiesto di calcolare la probabilità che il minimo tra X_1 e X_2 sia meno di 18 minuti. Possiamo calcolare la probabilità dell'evento

complementare, cioè che il minimo tra X_1 e X_2 sia maggiore di 18 minuti. Questo succede se sia X_1 è maggiore di 18 che X_2 è maggiore di 18. Quindi, usando l'indipendenza:

$$\begin{aligned}\mathbb{P}(\{\min\{X_1, X_2\} < 18\}) &= 1 - \mathbb{P}(\{\min\{X_1, X_2\} \geq 18\}) = 1 - \mathbb{P}(\{X_1 \geq 18\} \cap \{X_2 \geq 18\}) \\ &= 1 - \mathbb{P}(\{X_1 \geq 18\})\mathbb{P}(\{X_2 \geq 18\})\,.\end{aligned}$$

In realtà abbiamo già calcolato le ultime probabilità nel punto precedente. Si ottengono con la formula della probabilità totale (sono uguali perché le variabili aleatorie sono identicamente distribuite)

$$\begin{aligned}&\mathbb{P}(\{X_1 \geq 18\}) = \mathbb{P}(\{X_2 \geq 18\}) \\ &= \mathbb{P}(\{U_0 \geq 18\})\mathbb{P}(\{Y = 0\}) + \mathbb{P}(\{U_1 \geq 18\})\mathbb{P}(\{Y = 1\}) = 8/12 \cdot 1/10 + 2/10 \cdot 9/10 = 24.67\%\,.\end{aligned}$$

Quindi

$$\mathbb{P}(\{\min\{X_1, X_2\} < 18\}) = 1 - (24.67\%)^2 = 93.91\%\,.$$

Esercizio 4. (8 punti) Il/la docente di Probabilità e Statistica vuole un'evidenza significativa del fatto che in questo anno accademico l'esame sia stato più difficile per gli studenti e le studentesse.[1] Negli anni accademici precedenti, la media dei voti era 24. In un appello di questo anno accademico, invece, sono stati registrati i seguenti voti:

21	26	23	25	18	24	18	28	23	26	20	20	18	20	21	20
30	25	19	23	26	26	26	26	29	28	25	18	21	26	18	27

1. I voti assegnati nell'ultimo appello sono significativi al 5% per concludere che la media è effettivamente più bassa rispetto agli anni precedenti? (N.B.: ricavare le formule)

2. In questo anno accademico si svolgono 8 appelli, come programmato. Dopo ciascun appello si registrano i voti come nel punto precedente. A volte si conclude che la media è più bassa rispetto agli anni precedenti, altre volte no. Assumendo che la media di questo anno accademico sia rimasta 24, qual è la probabilità di commettere un errore (strettamente) meno di 6 volte arrivando a conclusioni con l'analisi precedente?

Soluzione. 1. Abbiamo a che fare con un campione casuale $X_1, \ldots, X_n$ di variabili aleatorie indipendenti e identicamente distribuite con $n = 32$. Non conosciamo la legge delle variabili aleatorie, non conosciamo la media della popolazione $\mathbb{E}(X_i) = \mu$ e non conosciamo la varianza della popolazione $\mathrm{Var}(X_i) = \sigma^2$.

Per rispondere alla domanda impostiamo un test d'ipotesi:

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

dove $\mu_0 = 24$. Infatti, fino a prova contraria, la media della popolazione non è cambiata ed è sempre 24. Se i dati sono abbastanza significativi (media molto più bassa di 24), rifiutiamo l'ipotesi nulla a favore dell'ipotesi alternativa, che ci permette di stabilire con un certo livello di significatività che la media della popolazione è effettivamente più bassa.

Definiamo una regione critica

$$R_C = \{(x_1, \ldots, x_n) \in R(X_1, \ldots, X_n) \ : \ \bar{x}_n < \mu_0 - \delta\}\,,$$

dove $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ è la media calcolata sul campione.

[1] L'esercizio è inventato e ogni riferimento a persone o fatti realmente accaduti è puramente casuale.

Il livello di significatività α è la probabilità di commettere un errore del I tipo, ovvero di rifiutare l'ipotesi nulla quando è vera. Supponiamo che l'ipotesi nulla sia vera, cioè $\mu = \mu_0 = 24$. Allora

$$\alpha = \mathbb{P}(\{(X_1, \dots, X_n) \in R_C\}) = \mathbb{P}(\{\bar{X}_n < \mu_0 - \delta\}) = \mathbb{P}(\{\bar{X}_n < \mu - \delta\}) = \mathbb{P}(\{\bar{X}_n - \mu < -\delta\})$$
$$= \mathbb{P}\Big(\Big\{\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < -\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big)\,.$$

Non avendo a disposizione la deviazione standard della popolazione, abbiamo usato lo stimatore $S_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, dove $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ è la media campionaria. Poiché il campione è numeroso ($n \geq 30$), possiamo utilizzare teoremi di approssimazione. In questo caso utilizziamo il Teorema di Slutsky, che ci garantisce che la statistica $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ ha approssimativamente una distribuzione normale standard (più precisamente: converge in legge a una variabile aleatoria con distribuzione normale standard per $n \to +\infty$). Quindi

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \simeq Z \sim \mathcal{N}(0,1)\,.$$

Possiamo allora utilizzare i quantili gaussiani per determinare δ. Chiamiamo z_α il punto della retta reale tale che

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

Osserviamo che per la simmetria della gaussiana

$$\mathbb{P}(\{Z < -z_\alpha\}) = \mathbb{P}(\{Z \geq z_\alpha\}) = \alpha$$

Allora scegliendo $\frac{\delta}{S_n/\sqrt{n}} = z_\alpha$ abbiamo che

$$\mathbb{P}\Big(\Big\{\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < -\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) \simeq \mathbb{P}(\{Z < -z_\alpha\}) = \alpha\,,$$

che è la condizione richiesta. Al posto di S_n, utilizzeremo la sua realizzazione sul campione dati $s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^n x_i^2 - n\bar{x}_n}$.

Abbiamo a disposizione tutti gli ingredienti e non ci resta che calcolare le quantità sui dati. Partiamo dal quantile gaussiano, per il quale utilizziamo la tabella della distribuzione normale standard:

$$\alpha = \mathbb{P}(\{Z \geq z_\alpha\}) \implies 1 - \alpha = \mathbb{P}(\{Z < z_\alpha\}) = \Phi(z_\alpha) \implies 0.95 = \Phi(z_{0.05}) \implies z_{0.05} \simeq 1.645\,.$$

Calcoliamo le realizzazioni della media campionaria e della deviazione standard campionaria sul campione di dati:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{32}(21 + 26 + \dots + 18 + 27) = 23.25\,.$$

$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^n x_i^2 - n\bar{x}_n} = \sqrt{\frac{1}{31}(21^2 + 26^2 + \dots + 18^2 + 27^2 - 32 \cdot 23.25^2)}$$
$$= \sqrt{\frac{1}{31}(17712 - 17298)} \simeq 3.65\,.$$

Segue che

$$\delta = s_n \frac{z_{0.05}}{\sqrt{32}} \simeq 3.65 \frac{1.645}{\sqrt{32}} \simeq 1.06\,.$$

Osserviamo che

$$\bar{x}_n = 23.25 > \mu_0 - \delta = 24 - 1.06 = 22.94\,,$$

quindi i dati non cadono nella regione critica e non sono abbastanza significativi da rifiutare l'ipotesi nulla. I dati non permettono di stabilire che l'esame è più difficile degli anni scorsi.

2. Abbiamo a che fare con 8 prove indipendenti. Consideriamo un "successo" un errore del I tipo, ovvero un rifiuto dell'ipotesi nulla assumendo che questa sia vera (ovvero che la media è rimasta 24). Abbiamo a che fare con una variabile aleatoria con legge binomiale $Y \sim \mathrm{B}(k, p)$ con $k = 8$ e $p = \alpha = 5\%$, perché la probabilità di commettere un errore del I tipo è esattamente la significatività del test. Ci viene chiesto di calcolare

$$\begin{aligned}\mathbb{P}(\{Y < 6\}) &= 1 - \mathbb{P}(\{Y \geq 6\}) = 1 - \mathbb{P}(\{Y = 6\}) - \mathbb{P}(\{Y = 7\}) - \mathbb{P}(\{Y = 8\}) \\ &= 1 - \binom{8}{6} 0.05^6 0.95^2 - \binom{8}{7} 0.05^7 0.95^1 - \binom{8}{8} 0.05^8 0.95^0 \simeq 99.99\%\,.\end{aligned}$$

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: aprile 2023
Matricola: _____ Data: 03/04/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) La tabella seguente mostra i dati sul consumo medio annuo di vino pro capite e sul numero di morti dovute a malattie cardiache in un campione casuale di 10 paesi:

consumo di vino (in litri)	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
morti	221	167	131	191	220	297	71	172	211	300

1. Rappresentare i dati in uno scatterplot.
2. Determinare (derivando le formule) la retta di regressione lineare e rappresentarla.
3. Determinare il coefficiente di correlazione lineare.

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

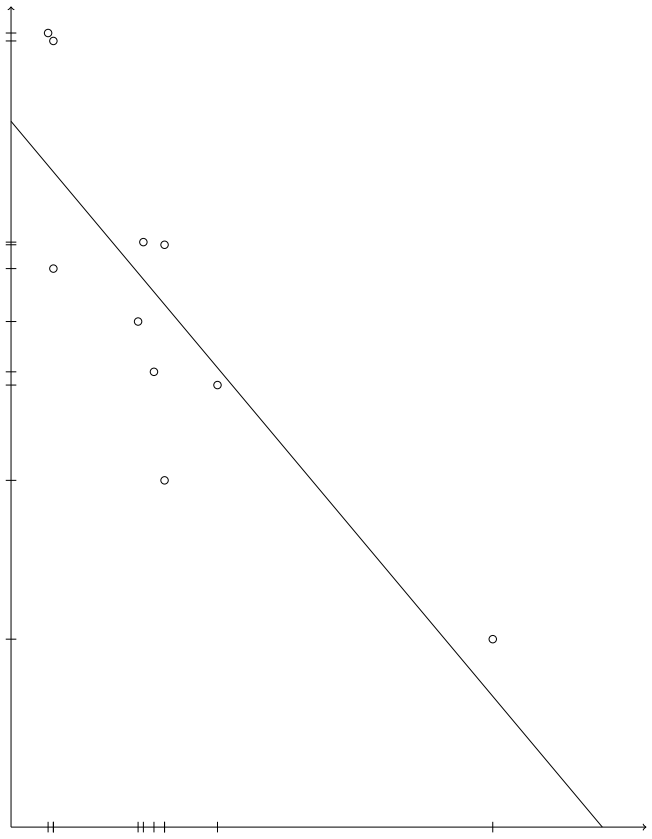


Figura 1: Scatterplot e retta di regressione lineare.

2. Denotiamo con $(x_1, y_1), \ldots, (x_n, y_n)$, $n = 10$, 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

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

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

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

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

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	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7	28.7
y_i	221	167	131	191	220	297	71	172	211	300	1981
x_i^2	6.25	15.21	8.41	5.76	8.41	0.64	82.81	7.29	0.64	0.49	135.91
y_i^2	48841	27889	17161	36481	48400	88209	5041	29584	44521	90000	436127
x_iy_i	552.5	651.3	379.9	458.4	638	237.6	646.1	464.4	168.8	210	4407

Pertanto $\overline{x} = 28.7/10 = 2.87$ e $\overline{y} = 1981/10 = 198.1$. Segue che

$$a = \frac{4407 - 10 \cdot 2.87 \cdot 198.1}{135.91 - 10 \cdot 2.87^2} = \frac{-1278.47}{53.541} \simeq -23.88\,,$$
$$b = 198.1 + 23.88 \cdot 2.87 \simeq 266.64\,,$$

ovvero, la retta di regressione lineare ha equazione

$$y = -23.88x + 266.64\,.$$

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

$$\rho_{x,y} = \frac{\sum_{i=1}^{n} x_iy_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}} = \frac{4407 - 10 \cdot 2.87 \cdot 198.1}{\sqrt{135.91 - 10 \cdot 2.87^2}\sqrt{436127 - 10 \cdot 198.1^2}}$$
$$= \frac{-1278.47}{\sqrt{53.541}\sqrt{43690.9}} \simeq -0.8359\,.$$

Esercizio 2. (7 punti) Un produttore di un componente elettronico sa che un componente prodotto è difettoso con una probabilità del 10% (si assumano i difetti dei componenti indipendenti tra loro).

1. Il produttore vende a un cliente una confezione con 20 componenti. Qual è la probabilità che la confezione contenga almeno 18 (18 incluso) componenti non difettose?

Il prezzo di vendita di una confezione da 20 pezzi è di 15€. Se il cliente riceve una confezione con almeno 18 componenti non difettose, non fa un reclamo. Altrimenti, il cliente fa un reclamo e chiede al produttore di inviare una nuova confezione (senza pagare nuovamente i 15€). Questa operazione si ripete finché il cliente non riceve una confezione con almeno 18 componenti sane.

2. In media, quante volte farà reclamo il cliente?

Per il produttore, il costo di produzione di una confezione da 20 pezzi è di 6€. Ogni volta che spedisce una confezione (la prima volta e per ogni eventuale reclamo), paga 2€ di costi di spedizione.

3. Qual è la probabilità che il produttore abbia una perdita per via dei ripetuti reclami dovuti ai difetti di una confezione?

Soluzione. 1. Identifichiamo come "successo" un componente non difettoso. Un successo ha probabilità 90%. Il numero di pezzi non difettosi in una confezione da 20 può essere modellato da una variabile aleatoria con legge binomiale $X \sim \mathrm{B}(n,p)$ con $n = 20$ e $p = 90\%$. Ci viene chiesto di calcolare

$$\begin{aligned}\mathbb{P}(\{X \geq 18\}) &= \mathbb{P}(\{X = 18\}) + \mathbb{P}(\{X = 19\}) + \mathbb{P}(\{X = 20\}) \\ &= \binom{20}{18}(90\%)^{18}(10\%)^2 + \binom{20}{19}(90\%)^{19}(10\%)^1 + \binom{20}{20}(90\%)^{20}(10\%)^0 \\ &= \frac{20\cdot 19}{2}(90\%)^{18}(10\%)^2 + 20(90\%)^{19}(10\%) + (90\%)^{20} = 67.69\%\,.\end{aligned}$$

2. Si sta effettuando una successione di prove in cui il "successo" è l'evento "la confezione contiene almeno 18 componenti sane", che ha probabilità 67.69%. Il primo successo (che corrisponde alla volta in cui il cliente smetterà di fare reclami) può essere modellato da una variabile aleatoria con legge geometrica $Y \sim \mathrm{Geo}(q)$ con $q = 67.69\%$. Il valore atteso di una variabile aleatoria con legge geometrica è

$$\mathbb{E}(Y) = \frac{1}{q} = \frac{1}{67.69\%} \simeq 1.48\,.$$

Poiché il numero di reclami è dato dal momento in cui osserviamo il successo meno 1, in media verranno fatti 0.48 reclami. (Se al primo tentativo arriva una confezione buona, ci saranno zero reclami, se al secondo tentativo arriva una confezione buona, ci sarà un reclamo, ecc.)

3. Consideriamo la variabile aleatoria Z che descrive il costo totale (costi di produzione e costi di spedizione). Poiché Y descritta nel punto 2. fornisce il numero di spedizioni effettuate, abbiamo che

$$Z = (6+2)\cdot Y = 8Y\,.$$

Si ottiene una perdita se il costo totale supera il ricavo, cioè $Z > 15$. Quindi dobbiamo calcolare

$$\begin{aligned}\mathbb{P}(\{Z > 15\}) = \mathbb{P}(\{8Y > 15\}) = \mathbb{P}(\{Y > 15/8\}) = \mathbb{P}(\{Y > 1.875\}) = \mathbb{P}(\{Y \geq 2\}) \\ = (1-q)^{2-1} = (1 - 67.69\%) = 32.31\%\,.\end{aligned}$$

Esercizio 3. (8 punti) Consideriamo una persona che sta svolgendo l'esame di Probabilità e Statistica. Se la persona ha studiato, il tempo (in minuti) che impiega a svolgere tutti gli esercizi del compito è distribuito con legge uniforme con media 90 min e varianza 12 min^2. Se la persona non ha studiato, il tempo (in minuti) che impiega a svolgere tutti gli esercizi del compito è distribuito con legge uniforme nell'intervallo $[90, 120]$. Il 70% delle persone che si presentano all'esame ha studiato.

1. Consideriamo una persona che sappiamo che non ha studiato. Con che probabilità impiegherà più di 100 minuti a svolgere il compito?

2. Consideriamo una persona che sappiamo che ha studiato. Con che probabilità impiegherà più di 90 minuti a svolgere il compito?

3. Consideriamo una persona che svolge l'esame (non sappiamo se ha studiato o se non ha studiato). Vediamo che ha terminato tutti gli esercizi del compito in meno di 95 minuti. Sapendo questo fatto, con che probabilità la persona ha studiato?

(I dati sono inventati.)

Soluzione. Consideriamo le seguenti variabili aleatorie:

$$X = \text{“tempo impiegato se la persona ha studiato”} \sim \mathrm{U}(a,b)$$

$$Y = \text{“tempo impiegato se la persona non ha studiato”} \sim \mathrm{U}(90,120)$$

$$T = \text{“tempo impiegato”}$$

$$S = \text{“la persona ha studiato”} \sim \mathrm{Be}(70\%)\,,$$

dove $S = 1$ (successo) quando la persona ha studiato e $S = 0$ quando la persona non ha studiato. Sappiamo che

$$\mathbb{P}(\{T \in E\}|\{S = 1\}) = \mathbb{P}(\{X \in E\})\,,$$
$$\mathbb{P}(\{T \in E\}|\{S = 0\}) = \mathbb{P}(\{Y \in E\})\,.$$

Per quanto riguarda la variabile aleatoria X utilizziamo il fatto che

$$\mathbb{E}(X) = \frac{a+b}{2}\,, \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12}\,.$$

Quindi, usando il fatto che $b - a > 0$,

$$\begin{cases} \dfrac{a+b}{2} = 90\,, \\ \dfrac{(b-a)^2}{12} = 12\,, \end{cases} \implies \begin{cases} a+b = 180\,, \\ b-a = 12\,, \end{cases} \implies \begin{cases} a = 84\,, \\ b = 96\,, \end{cases}$$

ovvero $X \sim \mathrm{U}(84,96)$.

1. Ci viene chiesto di calcolare

$$\mathbb{P}(\{Y \geq 100\}) = \int_{100}^{120} \frac{1}{120-90}\,\mathrm{d}x = \frac{120-100}{120-90} = \frac{20}{30} = \frac{2}{3}\,.$$

2. Ci viene chiesto di calcolare

$$\mathbb{P}(\{X \geq 90\}) = \int_{90}^{96} \frac{1}{96-84}\,\mathrm{d}x = \frac{96-90}{96-84} = \frac{6}{12} = \frac{1}{2}.$$

3. Ci viene chiesto di calcolare

$$\mathbb{P}(\{S = 1\}|\{T \leq 95\})\,.$$

Utilizziamo il Teorema di Bayes:

$$\begin{aligned} \mathbb{P}(\{S = 1\}|\{T \leq 95\}) &= \frac{\mathbb{P}(\{T \leq 95\}|\{S = 1\})\mathbb{P}(\{S = 1\})}{\mathbb{P}(\{T \leq 95\}|\{S = 1\})\mathbb{P}(\{S = 1\}) + \mathbb{P}(\{T \leq 95\}|\{S = 0\})\mathbb{P}(\{S = 0\})} \\ &= \frac{\mathbb{P}(\{X \leq 95\})\mathbb{P}(\{S = 1\})}{\mathbb{P}(\{X \leq 95\})\mathbb{P}(\{S = 1\}) + \mathbb{P}(\{Y \leq 95\})\mathbb{P}(\{S = 0\})} \\ &= \frac{\frac{95-84}{96-84}70\%}{\frac{95-84}{96-84}70\% + \frac{95-90}{120-90}30\%} \simeq 92.77\%\,. \end{aligned}$$

Esercizio 4. (7 punti) Uno studio statistico ha riferito che precedentemente gli/le adolescenti trascorrevano in media 3 ore al giorno con lo smartphone. Si vuole mostrare con

un'evidenza statistica che la media è diventata più alta. Ad alcuni/e adolescenti scelti casualmente è stato chiesto quante ore al giorno trascorrono con lo smartphone. I dati (in ore) sono i seguenti:

$$3.4 \quad 2.8 \quad 4.9 \quad 3.5 \quad 4.8 \quad 4.1 \quad 4.0 \quad 3.2 \quad 5.5 \quad 3.2 \quad 4.4 \quad 5.3 \quad 5.3 \quad 4.7 \quad 4.3\,.$$

(I dati sono inventati.) Si assuma che la popolazione abbia una distribuzione normale.

1. I dati sono significativi al 10% per stabilire che la media è davvero più alta?
2. In quale dei seguenti intervalli si posiziona il più piccolo livello di significatività per cui i dati portano a stabilire che la media è davvero più alta? $[0\%, 0.5\%)$, $[0.5\%, 1\%)$, $[1\%, 2.5\%)$, $[2.5\%, 5\%)$, $[5\%, 10\%)$, $[10\%, 100\%]$?

Soluzione. Si sta considerando un campione casuale $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ con $n = 15$. Sia la media μ che la varianza σ^2 della popolazione sono incognite. Fino a prova contraria, la media della popolazione è $\mu_0 = 3$. Grazie ai dati si può impostare un test di ipotesi unilaterale

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

ovvero ci si sta chiedendo se i dati sono abbastanza significativi per stabilire che la media della popolazione è, in realtà, maggiore di $\mu_0 = 3$.

Un livello di significatività α è la probabilità di commettere un errore del I tipo, ovvero di rifiutare l'ipotesi nulla quando questa è vera. Assumiamo allora che sia vera l'ipotesi nulla, ovvero che la media della popolazione sia $\mu = \mu_0 = 3$. Come regione critica per il rifiuto dell'ipotesi nulla considereremo un insieme della forma

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

dove $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ (ovvero, rifiutiamo l'ipotesi nulla se la realizzazione della media campionaria, stimatore corretto della media, sui dati del campione è sufficientemente lontana da μ_0).

Il livello di significatività α è allora, utilizzando la media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ e la varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ e il fatto che $\mu = \mu_0$,

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

Poiché $X_1, \dots, X_n$ hanno distribuzione normale, $T_{n-1} = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$ è distribuita come una t-Student con $n-1$ gradi di libertà. Scegliendo $\frac{\delta}{S_n/\sqrt{n}} = t_{n-1,\alpha}$, ovvero $\delta = \frac{S_n}{\sqrt{n}} t_{n-1,\alpha}$, dove $t_{n-1,\alpha}$ è il quantile della t-Student, si ha effettivamente che

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

Quindi rifiutiamo l'ipotesi nulla se la realizzazione della media campionaria e della varianza campionaria sui dati verificano che

$$\overline{x}_n > \mu_0 + \frac{s_n}{\sqrt{n}} t_{n-1,\alpha}\,.$$

Svolgiamo prima il punto 2. Svolgiamo il test di ipotesi con livello di significatività $\alpha = 0.5\%$. Dalla tavola della t-Student otteniamo che

$$t_{n-1,\alpha} = t_{14,0.005} = 2.977\,.$$

Calcoliamo anche

$$\overline{x}_n = \frac{1}{15}(3.4+2.8+4.9+3.5+4.8+4.1+4.0+3.2+5.5+3.2+4.4+5.3+5.3+4.7+4.3) \simeq 4.23\,.$$

$$s_n = \sqrt{\frac{1}{14}(3.4^2 + 2.8^2 + \cdots + 4.7^2 + 4.3^2 - 15 \cdot 4.23^2)} \simeq \sqrt{0.7119} \simeq 0.8437\,.$$

Si ha che

$$\overline{x}_n = 4.23 > \mu_0 + \frac{s_n}{\sqrt{n}} t_{n-1,\alpha} \simeq 3 + \frac{0.8437}{\sqrt{15}} 2.977 \simeq 3.65\,.$$

Questo vuol dire che con significatività $\alpha = 0.5\%$ viene rifiutata l'ipotesi nulla a favore di quella alternativa (cioè si ritiene che la media sia più alta di 3). Il più piccolo livello di significatività è nell'intervallo $[0\%, 0.5\%)$ (i dati sono molto significativi), rispondendo al punto 2. A maggior ragione, con un livello di significatività più alto ($\alpha = 10\%$) l'ipotesi nulla verrà rifiutata.