Politecnico di Milano
Esercizi di Statistica (2L) cod. 072900
Per gli allievi ING INF e TEL
Anno accademico 2008-2009[1]

Ilenia Epifani

26 aprile 2010

Questo materiale è stato elaborato durante gli Anni Accademici 2003-2007 per le esercitazioni lezioni e prove scritte del corso di Statistica (2L) per allievi di Ingegneria Informatica e delle Telecomunicazioni (2L). Gli esercizi sono organizzati tentando di seguire il programma del corso di Statistica (2L) AA 2008/2009. Per gli esercizi tratti da un libro sono forniti pagina, titolo, autore, casa editrice. Per gli esercizi tratti da prove d'esame sono forniti anno accademico e data dell'appello. Tutti i temi d'esame da cui sono tratti gli esercizi sono disponibili alla pagina
`http://www1.mate.polimi.it/~ileepi/stat-temi-esame/`

Milano, marzo 2009 — Ilenia Epifani

Indice

Capitolo 1

Calcolo delle probabilità

1.1 Ripasso: variabili aleatorie gaussiane

Esercizio 1.1.1 Siano $Z \sim \mathcal{N}(0,1)$ con f.d.r. Φ e $X \sim \mathcal{N}(\mu, \sigma^2)$ con f.d.r. F.

1. Usando le tavole della f.d.r. Φ calcolate: $P(Z \leq 1.51)$, $P(Z \leq -1.51)$, $P(Z > 0.31)$, $P(|Z| \leq 0.81)$, $P(0.31 < Z < 1.51)$
2. Usando le tavole della f.d.r. Φ calcolate i quantili di ordine 0.95 e 0.05 di Φ. cerca nella tabella partendo dai numeroni
3. Esprimete F in termini di Φ; esprimete il quantile di ordine a di F in termini del quantile di ordine a di Φ.
4. Sia $X \sim \mathcal{N}(20, 16)$. Calcolate i quantili di ordine 0.95 e 0.05 della f.d.r. di X e $P(X > 6)$.

Esercizio 1.1.2 Siano X, Y variabili aleatorie gaussiane indipendenti con $X \sim \mathcal{N}(\mu_X, 9)$ e $Y \sim \mathcal{N}(\mu_Y, 16)$.

1. Come si distribuisce $W = X - Y$?
2. Se $\mu_X = \mu_Y$, quanto vale $P(W > -2)$?

1.2 Ripasso: funzioni di variabili aleatorie, medie, varianze, teorema centrale del limite

Esercizio 1.2.1

1. Sia $Z \sim \mathcal{N}(0,1)$ e $Y = Z^2$. Determinate la densità di Y.
2. Sia $X \sim \mathcal{N}(\mu, \sigma^2)$ e $Y = \left(\frac{X-\mu}{\sigma}\right)^2$. Determinate la densità di Y.

Esercizio 1.2.2 La durata di una batteria espressa in ore è una variabile aleatoria continua con densità $f(x, \theta)$ data da

$$f(x,\theta) = \frac{1}{2\theta\sqrt{x}} \exp\left\{-\frac{\sqrt{x}}{\theta}\right\} \mathbf{1}_{(0,\infty)}(x), \qquad \theta > 0$$

1. Determinate in funzione di θ la probabilità che la batteria funzioni ancora dopo 11 ore dall'accensione.
2. Determinate la densità di $\sqrt{X}$ se $X \sim f(x, \theta)$.
3. Determinate la media di $X \sim f(x, \theta)$ in funzione di θ.

Esercizio 1.2.3 Sia X una variabile aleatoria discreta con densità data da

$$f(x, \theta) = \left(\frac{\theta}{2}\right)^{|x|} (1-\theta)^{1-|x|} \mathbf{1}_{\{-1,0,1\}}(x), \quad 0 < \theta < 1.$$

1. Quanto vale $P(X \geq 0)$? [`Risp:` $1 - \theta/2$]
2. Determinate media e varianza di X [`Risp:` $\mathrm{E}(X) = 0;\ \mathrm{Var}(X) = \theta$]
3. Determinate media e varianza di $5X - 1$ [`Risp:` $\mathrm{E} = -1;\ \mathrm{Var} = 25\theta$]
4. Determinate media e varianza di $Y = |X|$. [`Risp:` $\mathrm{E}(|X| = \theta;\ \mathrm{Var}(|X|) = \theta(1-\theta)$]
5. Qual è la densità della variabile aleatoria discreta Y?

Esercizio 1.2.4 Esercizio 1.2.5 (Tema d'esame 2.3 del 18/07/06 AA 05/06) La specifica per un certo tipo di chiodi stabilisce che devono avere una lunghezza nominale di 20mm; ma sono accettabili chiodi aventi lunghezza entro i limiti di tolleranza $[15, 24.6]$. L'azienda dichiara che le lunghezze reali dei chiodi sono variabili aleatorie uniformi di media 20mm e varianza 12mm^2.

1. Sia X la lunghezza reale di un chiodo. Se l'azienda dichiara il vero, quanto vale $\mathrm{E}(X^2)$?
2. Se l'azienda dichiara il vero, quanto valgono i parametri del modello uniforme delle lunghezze reali dei chiodi?
3. Se l'azienda dichiara il vero, qual è la percentuale attesa di chiodi troppo corti? Quale quella di chiodi troppo lunghi? E quella di chiodi che soddisfano la specifica?

Esercizio 1.2.6 Sia X una variabile aleatoria continua con densità di probabilità

$$f(x, \vartheta) = \frac{\vartheta}{x^{\vartheta+1}} \mathbf{1}_{(1,+\infty)}(x), \qquad \vartheta > 0.$$

1. Determinate $\mathrm{E}(X^{\theta-1})$. [`Risp:` θ]
2. Determinate la densità di $Y = 2\theta \ln X$; [`Risp:` $Y \sim \mathcal{E}(2)$]
3. Siano $X_1, \dots, X_n$ *i.i.d.* $\sim f(x, \vartheta)$ e sia $T_n = \dfrac{\ln(\prod_{i=1}^n X_i)}{n}$. Calcolate media e varianza di T_n. [`Risp:` $E(T_n) = 1/\theta;\ \mathrm{Var}(T_n) = (1/\theta^2)/n$]

Esercizio 1.2.7 Sia X una variabile aleatoria continua con densità

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} x^{\frac{1}{\theta}-1} & 0 < x < 1 \\ 0 & \text{altrove} \end{cases}, \quad \theta > 0$$

1. Determinare la densità di $Y = -\frac{1}{\theta} \log X$. [`Risp:` $Y \sim \mathcal{E}(1)$]

2. Siano $X_1, \dots, X_n$ *i.i.d.* $\sim f(x, \theta)$, e sia $T_n = -\dfrac{\log\left[\prod_{i=1}^n X_i\right]}{n}$. Calcolate media e varianza di T_n. [`Risp:` $\mathrm{E}(T_n) = \theta,\ \mathrm{Var}(T_n) = \theta^2/n$]

Esercizio 1.2.8 Sia X una variabile aleatoria continua di densità

$$f(x,\theta) = \frac{1}{2}\mathrm{e}^{-|x-\theta|}, \qquad x \in \mathbb{R},\ \theta \in \mathbb{R}$$

Calcolate $\mathrm{E}(X)$ e $\mathrm{Var}(X)$.

Esercizio 1.2.9 Il tempo espresso in minuti che impiego ogni mattina per raggiungere il Politecnico da casa mia può essere modellato con una variabile aleatoria continua di densità

$$f(x,\theta) = \begin{cases} \frac{\theta}{30^\theta}(x-30)^{\theta-1} & \text{se } 30 < x < 60 \\ 0 & \text{altrove} \end{cases}, \qquad \theta > 0.$$

1. Esprimete in termini di θ la probabilità che io impiegni meno di 35 minuti per raggiungere il Politecnico da casa mia. [`Risp:` $1/6^\theta$]

2. Se registro i tempi di viaggio per 10 giorni consecutivi, con quale modello probabilistico posso descrivere il numero di giorni su 10 in cui impiego meno di 35 minuti a raggiungere il Politecnico? Esprimete in termini di θ la probabilità che almeno 4 giorni su 10 io impieghi meno di 35 minuti per raggiungere il Politecnico. [`Risp:` $Bin(10, 1/6^\theta)$]

3. Se registro i tempi di viaggio per 100 giorni consecutivi, con quale modello probabilistico posso descrivere il numero di giorni su 100 in cui impiego meno di 35 minuti a raggiungere il Politecnico?

4. Esprimete in termini di θ un valore approssimato della probabilità che almeno 40 giorni su 100 io impieghi meno di 35 minuti per raggiungere il Politecnico. Quanto vale questa probabilità se $\theta = 0.6$? [`Risp:` $1 - \Phi\left((3.95 \times 6^\theta - 10)/\sqrt{6^\theta - 1}\right)$; 0.129]

Esercizio 1.2.10 Sia $X_1, \dots, X_n$ un campione casuale estratto da una popolazione di Poisson di media 2.
1. Come si distribuisce $X_1 + \cdots + X_n$, $n \geq 1$?
2. Calcolate $P(\bar{X} \leq 1.9)$ per $n = 1, n = 2, n = 100$. [`Risp:` $3\mathrm{e}^{-2}; 71\mathrm{e}^{-4}/3; \simeq 0.24$].

1.3 Distribuzioni campionarie

Esercizio 1.3.1 Sia $X_1, \dots, X_n$ un campione casuale da $\mathcal{N}(4.2, 4)$ e sia $\bar{X}$ la media campionaria.

1. Calcolate $P(|\bar{X} - 4.2| \leq 0.3)$, con $n = 4$ e $n = 25$ e confrontate i risultati;

2. Per quali valori di n $P(|\bar{X} - 4.2| \leq 0.3) \geq 0.8$?

Esercizio 1.3.2 Sia $X \sim \Gamma(4, 5)$ e $Y = 1/X$.

1. Determinate il quantile di ordine 5% della fdr di X. [`Risp:` $(5/2)2.733 = 6.8325$]

2. Determinate media e varianza di Y. [`Risp:` $\mathrm{E}(Y) = 1/15$, $\mathrm{E}(Y^2) = 1/150$, $\mathrm{Var}(Y) = 1/450$]

Esercizio 1.3.3 Sia $Z_1, \dots, Z_n$ un campione casuale estratto dalla popolazione di densità gaussiana standard e sia Sia $S_n = Z_1^2 + \ldots + Z_n^2$.

1. Calcolate media e varianza di S_n.
2. Sia $n = 18$. Calcolate $P(S_n \leq 9.39)$.
3. Sia $n = 300$. Calcolate $P(S_n \leq 312.98)$.

Esercizio 1.3.4 Abbiamo estratto il campione casuale $X_1, \ldots, X_n$ dalla popolazione di densità esponenziale di parametro θ: $f(x, \theta) = \frac{1}{\theta} \mathrm{e}^{-x/\theta} \mathbf{1}_{(0,\infty)}(x)$, $\theta > 0$.

1. Determinate la densità della media campionaria $\bar{X}$.
2. Determinate la densità della variabile aleatoria $\frac{2n\bar{X}}{\theta}$.
3. Sia $\alpha = 5\%$, $n = 3$ e $\theta = 2$: determinate k tale che $P(\bar{X} \leq k) = \alpha$ [`Risp:` $k \simeq 1.64/3$]
4. Sia $n = 3$ e $\theta = 1.49$: calcolate $P(\bar{X} > 1.64/3)$ [`Risp:` $P(\chi^2_6 > 2.2013) \simeq 0.9$]
5. Sia $n = 35$ e $\theta = 1.49$: calcolate $P(\bar{X} > 1.64/3)$ [`Risp:` $= \Phi(3.75) \simeq 0.9999$]
6. Determinate k dipendente da θ e da n tale che $P(\sum_{j=1}^n X_j > k) = 0.9$ [`Risp:` $k = \frac{\theta}{2} q_{\chi^2_{2n}}(0.1)$]

Esercizio 1.3.5 Il contenuto X di alcool metilico di una bottiglia di vino è una variabile aleatoria $\mathcal{N}(\mu, \sigma^2)$. Considerando un campione di 9 bottiglie,
a) calcolare $P(|\bar{X} - \mu| > S)$ dove S^2 è la varianza campionaria; [$2 \times 0.00853 \simeq 0.018$]
b) Se $\sigma^2 = 16$, determinare a tale che $P(S^2 \leq a) = 0.99$. [$a = 40.18$]

Esercizio 1.3.6 Sia $X_1, \ldots, X_n$ un campione casuale estratto dall popolazione gaussiana $\mathcal{N}(\mu, \sigma^2)$ e siano $\bar{X}, S^2$ rispettivamente media e varianza campionarie.

1. Determinate k t.c. $P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{S} \geq k\right) = \gamma$, in funzione di n, γ [`Risp:`$k = t_{n-1}(1 - \gamma)$]
2. Determinate k t.c. $P(\bar{X} - \mu \leq kS/\sqrt{n}) = 0.05$ per $n = 10$ e $n = 100$. [`Risp:`$k_{10} = -1.833$; $k_{100} = -1.645$]

Capitolo 2

Intervalli di confidenza per popolazione gaussiana

2.1 Intervalli di confidenza per media e varianza di popolazione gaussiana

Esercizio 2.1.1 Una ditta produce punte da trapano. Si provano n punte *dello stesso diametro* producendo n fori. Si indichi con $Y_1, \ldots, Y_n$ i diametri dei fori prodotti e si supponga che $Y_1, \ldots, Y_n$ siano normali con media μ incognita e varianza σ_0^2 nota. Rispondere alle seguenti domande giustificando *adeguatamente* le risposte.

1. Determinare un intervallo di confidenza bilatero per la media μ di livello γ.

2. Si supponga ora che $n = 100$, $\overline{Y}_n = 5\text{mm}$, $\sigma_0^2 = 10^{-2}\text{mm}^2$, $\gamma = 95\%$ ($\overline{Y}_n$ = media campionaria). Calcolare l'intervallo di confidenza del punto precedente con questi dati. [`Risp:` $(4.98, 5.02)$]

3. Se $\sigma_0^2 = 10^{-2}\text{mm}^2$, quanto grande occorre prendere il campione per essere sicuri al 95% che la nostra stima di μ sia precisa entro 10^{-2}mm ? [`Risp:` $n = 385$]

4. Abbiamo ora estratto un altro campione (totalmente nuovo) sempre dalla famiglia gaussiana $\mathcal{N}(\mu, 10^{-2})$ e sappiamo che per questo nuovo campione $(4.50, 4.54)$ è un IC (simmetrico) al 95% per μ. Quale è la media campionaria della popolazione e quale la dimensione di questo nuovo campione estratto? [`Risp:` $n = 97; \bar{x} = 4.52$]

Esercizio 2.1.2 (Tema d'esame 2.2 del 18/07/06 AA 05/06) Siano $\bar{X}$ e S^2 rispettivamente la media e la varianza campionarie di un campione casuale $X_1, \ldots, X_n$ estratto da una popolazione gaussiana di media 2θ e varianza σ^2 ($X_1, \ldots, X_n$ *i.i.d.* $\sim \mathcal{N}(2\theta, \sigma^2)$, $\theta \in \mathbb{R}, \sigma^2 > 0$ entrambi incogniti).

1. Se il campione è costituito da 25 osservazioni, quanto vale $P_{\theta,\sigma^2}(\bar{X} - 0.342S - 2\theta \leq 0)$?

Abbiamo misurato la pressione sistolica del sangue di 25 maschi sani e abbiamo ottenuto una media campionaria pari a 120.0mm di mercurio e un momento secondo campionario pari a 14616.00mm^2 di mercurio.

2. Sulla base di questi dati quanto siete confidenti che θ sia maggiore o uguale a 57.435?

Abbiamo raccolto ULTERIORI dati riguardanti 39 maschi sani e, per i nuovi 39 dati, abbiamo ottenuto una media campionaria pari a 110.0mm di mercurio e un momento secondo campionario pari a 12715.0mm^2

3. Aggiornate le stime puntuali di media e varianza sulla base di questi nuovi dati, usando l'intero campione di 64 misurazioni.

4. Determinate numericamente un intervallo di confidenza unilatero per il parametro θ di forma (c, ∞) e di confidenza 95%.

Esercizio 2.1.3 Quanto confidate che un intervallo IC simmetrico per μ, la cui lunghezza è minore di σ, contenga il vero valore del parametro μ, nel caso di un campione di 20 osservazioni da popolazione normale con varianza nota? [`Risp:` $\gamma \simeq 0.975$]

Esercizio 2.1.4 Da una popolazione di maschi adulti si estrae un campione di 25 unità su cui viene osservata la statura. Supponendo che la popolazione delle stature abbia distribuzione $\mathcal{N}(170m., 16m^2)$ e che la media campionaria osservata sia $\bar{X}_{25} = 173$

a) Si verifichi che l'intervallo di confidenza di livello 95% per la media che si costruisce a partire da tale stima non contiene il vero valore $\mu = 170$ della media. Si ragioni su questo fatto e si dia una giustificazione.

b) Entro quali valori si deve trovare una media campionaria affinchè la media $\mu = 170$ sia contenuta nell'intervallo di confidenza al 95%? [$\bar{X} \in [168.432, 171.568]$]

Esercizio 2.1.5 Vengono effettuate 20 misurazioni della concentrazione di un certo enzima nel sangue di diversi individui e si osservano una media campionaria $\bar{x} = 1.23$ ed una varianza campionaria $s^2 = 0.4$.

1. Supponendo che i valori di questa concentrazione seguano una distribuzione normale, qual è un intervallo di fiducia al livello 95% per la media della concentrazione? [[0.934; 1.523]]

2. Quale sarebbe l'intervallo di cui sopra se la concentrazione a cui si è interessati fosse distribuita come una normale di varianza nota $\sigma^2 = 0.4$? [[0.953; 1.507]]

3. Quale tra i due intervalli trovati in 1. ed in 2. è più ampio? Il risultato ottenuto sembra ragionevole? [E' più ampio l'intervallo trovato in 1., dove la varianza è incognita ed è necessario stimarla con S^2]

Esercizio 2.1.6 [1] Una ditta produce termometri di precisione per l'industria chimica. Un chimico acquista uno di questi termometri e misura 10 volte la temperatura di un bagno termico, ottenendo i seguenti valori $\{-0.06 \times 10^{-2}, 0.18 \times 10^{-2}, -1.24 \times 10^{-2}, 1.43 \times 10^{-2}, 0.23 \times 10^{-2}, -0.18 \times 10^{-2}, 0.01 \times 10^{-2}, 0.41 \times 10^{-2}, 0.27 \times 10^{-2}, -0.84 \times 10^{-2}\}$.

1. Calcolare un intervallo di confidenza bilaterale di livello $\gamma = 95\%$ per la varianza della variabile $X =$ "temperatura misurata", specificando le ipotesi statistiche che si stanno usando.

2. Calcolare un intervallo di confidenza unilaterale del tipo $(0, c)$ di livello $\gamma = 95\%$ per la varianza della variabile $X =$ "temperatura misurata", specificando le ipotesi statistiche che si stanno usando. [`Risp:` $(0, 1.396 \times 10^{-4})$]

3. Calcolare un intervallo di confidenza unilaterale di livello $\gamma = 95\%$ del tipo $(c, +\infty)$ per la deviazione standard della variabile $X =$ "temperatura misurata", specificando le ipotesi statistiche che si stanno usando.

[1] Tratto da un tema d'esame di MPSPS AA 97-99

Esercizio 2.1.7 [2] Una ditta produce termostati per scaldabagni; i termostati sono costruiti per aprire un circuito alla temperatura $\mu_0 = 60^0$C. Un addetto al controllo di qualità esamina uno di questi termostati nel modo seguente: immerge il termostato in un bagno termico a temperatura controllata da un termometro di precisione e fa aumentare la temperatura fino a quando il termostato apre il circuito, poi annota la temperatura alla quale il termostato ha aperto il circuito. Ripetendo questa procedura per 10 volte l'addetto ottiene il seguente campione (espresso in ^{0}C): 60.2, 59.6, 60.3, 59.5, 60.2, 60.4, 59.5, 60.4, 60.2, 59.0.

1. Calcolare un intervallo di confidenza unilaterale, di tipo $(0, c)$, di livello $\gamma = 95\%$ per la varianza della variabile X = "temperatura a cui il termostato apre il circuito", specificando le ipotesi statistiche sulla distribuzione di X che si stanno assumendo. [`Risp:` $(0, 0.556)$]

2. Calcolare un intervallo di confidenza bilatero, di livello $\gamma = 95\%$ per la varianza della variabile X = "temperatura a cui il termostato apre il circuito", specificando le ipotesi statistiche sulla distribuzione di X che si stanno assumendo. [`Risp:` $(2.19/20.483, 2.19/3.247)$]

Esercizio 2.1.8 Una ditta produce cinghie di trasmissione per automobili. È noto che la durata di tali cinghie, calcolata in *migliaia* di chilometri percorsi, ha fdr normale con media 50 e scarto quadratico medio 5. Denotiamo con X la durata di una cinghia. Successivamente la ditta progetta un congegno che –applicato al sistema di trasmissione– consente di allungare la "vita" di una cinghia di una quantità $Y \sim N(\mu, \sigma^2)$, con parametri μ e σ^2 entrambi incogniti. In altre parole, la durata delle cinghie in presenza del congegno è $Z = X + Y$ e assumiamo che X e Y siano indipendenti. Un test su 30 automobili provviste del nuovo congegno ha permesso di raccogliere dati relativi ad un campione $Z_1, \ldots, Z_{30}$ che ha fornito $\overline{Z}_{30} = 57$, $S_{30} = 6.5$, dove $\overline{Z}$ indica la media campionaria e S^2_{30} è la varianza campionaria (corretta) di $Z_1, \ldots, Z_{30}$.

1. Qual è la fdr di Z? [`Risp:` $Z \sim N(50 + \mu, 25 + \sigma^2]$

2. *Sulla base del campione* $Z_1, \ldots, Z_{30}$ fornire una stima per μ e σ^2. [`Risp:` $\hat{\mu} = \bar{Z} - 50 = 7$; $\hat{\sigma}^2 = S^2_Z - 25 = 17.25$;]

3. Determinare un intervallo di confidenza per μ di livello 0.95. [`Risp:` $(4.56, 9.43)$]

4. Determinare un intervallo di confidenza bilatero per la deviazione standard σ di livello 0.95. [`Risp:` $(1.341, 7.167)$]

[2] Tratto da un tema d'esame di MPSPS AA 97-99

Capitolo 3

Stima puntuale e intervallare

3.1 Metodi di costruzione degli stimatori, proprietà degli stimatori

Esercizio 3.1.1 Sia $X_1, \dots, X_n$ un campione casuale estratto dalla densità $\Gamma(\alpha, \beta)$ con entrambi i parametri incogniti. Determinate gli stimatori di α, β usando il metodo dei momenti.

Esercizio 3.1.2 (Esempio 7.2, (b) pag. 23 della dispensa di Teoria della stima puntuale) Sia $U_1, \dots, U_n$ un campione casuale estratto dalla popolazione di densità uniforme sull'intervallo (a, b). Determinate degli stimatori di a, b ottenuti con il metodo dei momenti.

Esercizio 3.1.3 (Esempi 7.2, (a) pag. 23 e 7.11 pag. 28 della dispensa di Teoria della stima puntuale) Siano $X_1, \dots, X_n$ un campione casuale estratto dalla densità uniforme $U(0, \theta)$, $\theta > 0$. Stimate θ con il metodo dei momenti e con il metodo di massima verosimiglianza.

Esercizio 3.1.4 (Tema d'esame 1.3 del 28/06/06 AA 05/06) Abbiamo raccolto 12 misurazioni di una certa grandezza Q di seguito riportate nell'ordine in cui sono state ottenute:

1.836 -0.400 1.168 -1.807 -1.475 1.661 -0.773 1.437 0.414 1.713 -0.860 -1.926

1. ...

Riteniamo sia sensato modellare una misurazione di Q come una variabile aleatoria X uniforme sull'intervallo $(a - b, a + b), b > 0$, cioè $f(x, a, b) = \dfrac{1}{2b}\mathbf{1}_{(a-b,a+b)}(x)$.

2. Determinate degli stimatori dei parametri a e b usando il metodo dei momenti e un campione casuale di n osservazioni $X_1, \dots, X_n$ estratto da $f(x, a, b)$. Fornitene poi le stime basate sul precedente set di 12 dati.

3. Supponete ora che il parametro a sia noto e pari a 0. Costruite lo stimatore di massima verosimiglianza di b basato solo sulla dodicesima misurazione $x_{12} = -1.926$.

Esercizio 3.1.5 (Tema d'esame 3.4 punto (2) del 19/09/05) Sappiamo che il tempo di esecuzione del programma xxx sul calcolatore yyy è modellabile come una variabile aleatoria X che ha densità

$$f(x, a, b) = \begin{cases} \frac{2(b-x)}{(b-a)^2} & a \le x \le b \\ 0 & \text{altrove} \qquad a, b > 0 \end{cases}$$

Noi non conosciamo né a né b; per questo vi chiediamo di determinare degli stimatori di a e b usando il metodo dei momenti (*Sugg: usare il fatto che* $\int_a^b 2x^2 \frac{b-x}{(b-a)^2}\,dx = \frac{(b-a)^2}{18} + \frac{(2a+b)^2}{3^2}$).

Esercizio 3.1.6 Sia $X_1, \dots, X_n$ un campione estratto dalla popolazione gaussiana $\mathcal{N}(\mu, \sigma^2)$.

1. Determinate lo stimatore di massima verosimiglianza di media e varianza.
2. Determinate lo stimatore di massima verosimiglianza della varianza quando la media è nota.
3. Determinate lo stimatore di massima verosimiglianza della varianza quando la media è incognita.

Esercizio 3.1.7 Sia $X_1, \dots, X_n$ un campione casuale estratto dalla densità di Bernoulli di parametro θ, $\theta \in (0,1)$. Determinate uno stimatore per θ con il metodo di massima verosimiglianza.

Esercizio 3.1.8 (Tema d'esame 2.2 del 15/07/04 AA 03/04) Il tempo di esecuzione del programma xxx sul calcolatore yyy è compreso fra 60 e 120 minuti primi. Idealmente, esso può essere modellato come una va X continua con densità

$$f(x,\theta) = \begin{cases} \frac{\theta}{60^\theta}(x-60)^{\theta-1} & 60 \le x \le 120 \\ 0 & \text{altrove} \end{cases}, \qquad \theta > 0$$

1. Determinate in funzione di θ la caratteristica $\kappa(\theta) =$ "Probabilità che il calcolatore impieghi più di 90 minuti per eseguire il programma".

Su ciascuno di n calcolatori tutti del tipo yyy (e che lavorano indipendentemente uno dall'altro) lanciamo il programma xxx e, allo scadere dei 90 minuti, controlliamo se il programma è stato eseguito o no. Sia Y_i la va che vale 1 se il programma lanciato sull'i-esimo calcolatore è eseguito (nei 90 minuti) e vale 0 se non lo è, per $i = 1, \dots, n$.

2. Quanto vale $P_\theta(Y_i = 1)$, $i = 1, \dots, n$?
3. Verificate che $\widehat{\kappa} = 1 - \bar{Y}$ è lo stimatore di massima verosimiglianza di $\kappa(\theta)$ basato sul campione casuale $Y_1, \dots, Y_n$.
4. (a) Verificate se $\widehat{\kappa}$ è stimatore non distorto e consistente in media quadratica per $\kappa(\theta)$ e (b) determinate la funzione di ripartizione asintotica di $\widehat{\kappa}$. Giustificate adeguatamente le risposte.
5. Alla luce di quanto ottenuto ai punti 1., 2. e 3., come stimereste θ sapendo che su $n = 15$ programmi lanciati 10 non sono stati eseguiti nei primi 90 minuti?

Esercizio 3.1.9 (Tema d'esame 4.2 del 15/02/05, AA 03/04) Il manufatto `aaa` è prodotto in un gran numero di stabilimenti. La proporzione X di manufatti difettosi (variabile da stabilimento a stabilimento) può essere modellata come una variabile aleatoria continua con densità

$$f(x,\theta) = \begin{cases} \frac{1}{\theta} x^{\frac{1}{\theta}-1} & 0 < x < 1 \\ 0 & \text{altrove} \end{cases}$$

dove θ è un parametro positivo incognito.
Per stimare θ gli addetti al controllo di qualità scelgono a caso n stabilimenti.

1. Determinate uno stimatore di θ usando il metodo dei momenti.
2. Determinate uno stimatore di θ usando il metodo di massima verosimiglianza.
3. Determinate la densità di $Y = -\log X$.

4. Discutete le proprietà dello stimatore di massima verosimiglianza individuato al punto 2. (non distorsione, consistenza, efficienza, ...).

5. Gli addetti al controllo di qualità decidono di visitare 4 stabilimenti e di ispezionare 30 manufatti in ognuno di essi. Se trovano 2 pezzi difettosi nel primo, 3 nel secondo, 3 nel terzo e 1 nel quarto, qual è la stima di θ basata sul metodo dei momenti? E qual è quella basata sul metodo di massima verosimiglianza?

Esercizio 3.1.10 (Tema d'esame 5.3 del 07/03/05 AA 03/04) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità

$$f(x, \theta) = \begin{cases} \frac{2}{\theta}\left(1 - \frac{x}{\theta}\right) & \text{se } 0 < x < \theta \\ 0 & \text{altrove} \end{cases}$$

dove θ è un parametro positivo incognito. Indichiamo con $\overline{X}$ la media campionaria di $X_1, \ldots, X_n$.

1. Calcolate $\mathrm{E}(\overline{X})$ e $\mathrm{Var}(\overline{X})$.

2. Costruite uno stimatore non distorto per θ (partendo da $\overline{X}$) e calcolatene l'errore quadratico medio (MSE).

Supponete ora di avere estratto una sola osservazione ($n = 1$).

3. Determinate lo stimatore di massima verosimiglianza di θ. *Potrebbe esservi utile disegnare il grafico della funzione di verosimiglianza (x fissato, θ variabile).*

4. Calcolate l'errore quadratico medio dello stimatore di massima verosimiglianza trovato al punto 3.

4bis Se $n = 1$, fra lo stimatore determinato al punto 2. e quello al punto 3. quale preferite?

Esercizio 3.1.11 (Tema d'esame 5.2 del 25/02/08 AA 06/07) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità

$$f(x, \theta) = \frac{1}{4} \mathrm{e}^{-\left|\frac{x}{2} - \theta\right|}, \qquad x \in \mathbb{R}$$

dove θ è un parametro reale incognito. Indichiamo con $\overline{X}$ la media campionaria di $X_1, \ldots, X_n$.

1. Calcolate $\mathrm{E}(\overline{X})$ e $\mathrm{Var}(\overline{X})$.

2. Costruite uno stimatore non distorto per θ (partendo da $\overline{X}$) e calcolatene l'errore quadratico medio (MSE).

Supponete ora di avere estratto una sola osservazione ($n = 1$).

3. Determinate lo stimatore di massima verosimiglianza di θ.
   Potrebbe esservi utile disegnare il grafico della funzione di verosimiglianza (x fissato, θ variabile).

Esercizio 3.1.12 Sia $X_1, \ldots, X_n$ un campione casuale da una popolazione con densità

$$f(x; \theta) = \frac{x^2}{2\theta^{3/2}} \mathrm{e}^{-\frac{x}{\sqrt{\theta}}} \mathbf{1}(x > 0)$$

dove $\theta > 0$ è un parametro incognito.

1. Determinare lo stimatore $\bar{\theta}_n$ di θ con il metodo dei momenti (*Suggerimento: osservare che* $X_1 \sim \Gamma(3, \sqrt{\theta})$).

2. Determinare gli stimatori di massima verosimiglianza $\hat{\theta}_n$ per θ, $\widehat{\kappa}_1$ per $\kappa_1(\theta) = \sqrt{\theta}$ e $\widehat{\kappa}_2$ per $\kappa_2(\theta) = \dfrac{1}{\sqrt{\theta}}$.

3. Verificare che $\widehat{\kappa}_1$ è stimatore non distorto ed efficiente di $\kappa_1(\theta)$.

4. Calcolare medie e varianze asintotiche di $\hat{\theta}_n$ e di $\widehat{\kappa}_2$.

Esercizio 3.1.13 Sia $X_1, \dots, X_n$ un campione casuale estratto da una popolazione di Poisson di parametro θ.

1. Determinate lo stimatore di massima verosimiglianza di θ e della caratteristica $\kappa(\theta) = P_\theta(X_1 > 0)$.

2. Verificate se la varianza dello stimatore di massima verosimiglianza di θ raggiunge il confine di Cramer-Rao.

Esercizio 3.1.14 (Tema d'esame 1.2 del 29/06/04) Sia $X_1, \dots, X_n$ un campione casuale estratto dalla funzione di densità discreta

$$p(x, \theta) = \begin{cases} \frac{e^{-\theta^2} \theta^{2x}}{x!} & x = 0, 1, \dots \\ 0 & \textit{altrove} \end{cases} \qquad \theta \in \mathbb{R} \setminus \{0\}$$

1. Determinate uno stimatore della caratteristica $\kappa(\theta) = \theta^2$ usando il metodo dei momenti.

Sia $\widehat{\kappa}$ lo stimatore individuato al punto 1.

2. Verificate se $\widehat{\kappa}$ è stimatore non distorto per $\kappa(\theta)$. Giustificate rigorosamente la risposta.
3. Calcolate la funzione di verosimiglianza del campione: $L_\theta(x_1, \dots, x_n)$.
4. Verificate se la varianza di $\widehat{\kappa}$ raggiunge il confine di Cramer-Rao.

Esercizio 3.1.15 (Esercitazione del 18/06/04 Dott.ssa Guatteri) Un macchinario produce oggetti a sezione circolare, (sferette per cuscinetti sfera) il cui raggio R è una v.a. gaussiana con media μ e varianza σ_0^2 nota. Sia $R_1, \dots, R_n$ un campione casuale da questa popolazione.

1. Determinate lo stimatore di massima verosimiglianza del raggio medio μ della sezione circolare degli oggetti.
2. Determinate lo stimatore di massima verosimiglianza T_1 dell'area media $\kappa(\mu) = \pi\mu^2$.
3. Come si può ottenere uno stimatore non distorto di $\kappa(\mu)$ a partire da T_1?
4. Discutete alcune proprietà dello stimatore di massima verosimiglianza T_1.
5. Si consideri un altro stimatore dell'area media $\kappa(\mu)$ dato da $T_2 = \dfrac{\pi}{n} \displaystyle\sum_{j=1}^{n} R_j^2 - \pi\sigma_0^2$. Discutete le proprietà dello stimatore T_2.

Esercizio 3.1.16 (Tema d'esame 4.1 del 29/09/06 AA 05/06) Nel tiro al bersaglio il punto colpito da un tiratore scelto ha coordinate X, Y che sono variabili aleatorie gaussiane indipendenti centrate in zero e di varianza σ^2 incognita. La precisione τ del tiratore è data dal reciproco della deviazione standard, cioè $\tau = 1/\sigma$. Per fare inferenza su τ, abbiamo chiesto al nostro tiratore scelto di effettuare n tiri a ogni tiro abbiamo registrato il quadrato della distanza (euclidea) fra il punto colpito (X, Y) e il bersaglio, cioè $Q = X^2 + Y^2$. Sia $Q_1, \dots, Q_n$ il campione casuale così ottenuto.

1. Qual è la densità di $Q = X^2 + Y^2$?

2. Determinate gli stimatori di massima verosimiglianza di σ^2 e di τ basati sul campione $Q_1, \dots, Q_n$.

3. Stabilite se lo stimatore di massima verosimiglianza di σ^2 è efficiente.

4. Provate ad argomentare perché lo stimatore di massima verosimiglianza della precisione τ non è efficiente.

Esercizio 3.1.17 (Esempio 6.13. pag. 20 della dispensa di Teoria della stima puntuale) Sia $X_1, \dots, X_n$ un campione casuale estratto dalla densità $f(x,\theta) = \dfrac{x+1}{\theta(\theta+1)}\mathrm{e}^{-x/\theta}\mathbf{1}_{(0,\infty)}(x),\ \theta > 0$. Determinate uno stimatore efficiente della caratteristica $\kappa(\theta) = (3+2\theta)(2+\theta)/(\theta+1)$

Esercizio 3.1.18 Una va continua X è detta *lognormale di parametri μ e σ* se il suo logaritmo naturale $\ln(X)$ ha densità $\mathcal{N}(\mu, \sigma^2)$; X ha media $\mathrm{E}(X) = \mathrm{e}^{\mu+\sigma^2/2}$ e varianza $\mathrm{Var}(X) = \mathrm{e}^{2\mu+\sigma^2}\left(\mathrm{e}^{\sigma^2} - 1\right)$.

Sia $X_1, \dots, X_n$ un campione casuale estratto dalla popolazione lognormale di parametri $\mu = \ln(\theta)$, $\theta > 0$ incognito e $\sigma = 2$.

1. Determinate uno stimatore T_1 di θ usando il metodo dei momenti e calcolatene il MSE. [`Risp:` $T_1 = \mathrm{e}^{-2}\bar{X}$; $MSE(T_1) = \mathrm{Var}(T_1) = \theta^2(\mathrm{e}^4 - 1)/n$]

2. Determinate uno stimatore $\hat{\theta}_{ML}$ di θ usando il metodo di massima verosimiglianza. [`Risp:` $\hat{\theta}_{ML} = \exp\left\{\frac{\sum_{j=1}^n \ln X_j}{n}\right\} = \prod_{j=1}^n X_j^{1/n}$]

3. Moltiplicate lo stimatore $\hat{\theta}_{ML}$ di θ per una opportuna costante in modo da ottenere uno stimatore T_2 per θ non distorto e calcolate $\mathrm{Var}(T_2)$. (*Sugg.: usate il fatto che la fgm di $X \sim N(m, \sigma^2)$ è $E(e^{tX}) = e^{tm+\frac{1}{2}t^2\sigma^2}\ \forall t \in \mathbb{R}$*) [`Risp:` $T_2 = \hat{\theta}_{ML}\mathrm{e}^{-\frac{2}{n}}$; $\mathrm{Var}(T_2) = \theta^2(\mathrm{e}^{4/n} - 1)$]

4. Quale tra i due stimatori T_1 e T_2 di θ è preferibile e perché? (*Sugg.: è utile usare la seguente diseguaglianza $e^{4/n} - 1 < (e^4 - 1)/n\ \forall n = 2, 3, \dots$*). [`Risp:` T_2]

3.2 Stimatori puntuali, metodo della quantità pivotale per intervalli di confidenza, intervalli di confidenza asintotici

Esercizio 3.2.1 (Tema d'esame 1.1 del 30/06/05 AA 04/05) Una grandezza incognita μ è stata misurata più volte usando due strumenti con precisioni diverse e note. Possiamo schematizzare la situazione dicendo che abbiamo due campioni casuali indipendenti $X_1, X_2, \dots, X_m$ da una popolazione gaussiana $\mathcal{N}(\mu; \sigma_X^2)$ ed $Y_1, Y_2, \dots, Y_n$ da una popolazione gaussiana $\mathcal{N}(\mu; \sigma_Y^2)$; le deviazioni standard $\sigma_X > 0$ e $\sigma_Y > 0$ sono note e μ è da stimare. Usiamo come stimatore di μ una statistica della forma $T = a\bar{X}_m + (1-a)\bar{Y}_n$, dove a è un opportuno numero reale.

1. Si calcolino la media e la varianza di T.
2. Si trovi a in modo da minimizzare l'errore quadratico medio di T.
3. Fissato per a il valore trovato al punto precedente, qual è la distribuzione di T?

Siano ora $m = 10, n = 20, \sigma_X^2 = 0.1, \sigma_Y^2 = 0.2$ e sia a come trovato al punto 2; se non siete stati in grado di trovarlo scegliete un valore ragionevole per a (diverso da 0 e da 1) in modo da poter proseguire.
L'esperimento ha fornito i seguenti valori per le due medie campionarie: $\bar{x}_m = 3.10, \bar{y}_m = 3.00$.

4. Si trovi un intervallo di confidenza (basato su T) al 95% per μ.

Esercizio 3.2.2 (Esercizio 2.4 pag. 11 delle dispense: Intervalli di confidenza) Sia $X_1, \dots, X_n$ un campione casuale dalla popolazione di densità $f(x,\theta) = (1/\theta)\mathrm{e}^{-x/\theta}\mathbf{1}_{(0,\infty)}(x)$, $\theta > 0$.

1. Determinate lo stimatore ML T della caratteristica $\kappa(\theta) = \mathrm{E}_\theta(X)$.
2. Determinate la densità di $2nT/\theta$, $\forall \theta > 0$.
3. Supponiamo ora $n = 10$ e $\bar{x} = 3$. Usando il risultato ottenuto al punto 2., proponete un intervallo di confidenza (unilatero) della forma $(0, u)$ di livello 95% per la caratteristica $\kappa(\theta) = \mathrm{E}_\theta(X)$.

Esercizio 3.2.3 (Tema d'esame 2.2 del 16/07/08 AA 07/08) Abbiamo estratto un campione casuale $X_1, \ldots, X_n$ da una popolazione con densità di parametro $\theta > 0$ data da

$$f(x, \theta) = \frac{3x^2}{\theta} \exp\left\{-\frac{x^3}{\theta}\right\} \mathbf{1}_{(0,\infty)}(x).$$

1. Determinate lo stimatore $\hat{\theta}_n$ della caratteristica θ usando il metodo di massima verosimiglianza.
2. Stabilite se lo stimatore $\hat{\theta}_n$ determinato al punto 1. è efficiente.
3. Determinate la densità di $Y = X^3$ e di $Q_n = \dfrac{2n\hat{\theta}}{\theta}$.
4. Fornite un intervallo di confidenza a due code per θ di livello 90% per $n = 10$ e $\hat{\theta} = 0.0387$.

Esercizio 3.2.4 Sia $X_1, \ldots X_n$ *i.i.d.* $\sim$ *Geometrica*(θ), $\theta \in (0, 1)$.
1. Determinate lo stimatore di massima verosimiglianza di θ; [`Risp:` $1/\bar{X}$]
2. Determinate $nI(\theta)$, con $I(\theta)$ informazione di Fisher del modello; [`Risp:` $nI(\theta) = \dfrac{n}{\theta^2(1-\theta)}$]
3. Determinate un intervallo di confidenza bilatero asintotico di livello approssimativamente γ per θ [`Risp:` $\dfrac{1}{\bar{X}} \pm z_{\frac{1+\gamma}{2}} \sqrt{\dfrac{\frac{1}{\bar{X}^2}\left(1 - \frac{1}{\bar{X}}\right)}{n}}$]

Esercizio 3.2.5 (Tema d'esame 5.2 (punti 1.-4.) del 08/03/06 AA 04/05) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità

$$f(x, \beta) = \frac{1}{2\beta} exp\left(-\frac{|x|}{\beta}\right), \quad \beta > 0$$

1. Determinate densità e media di $Y_i = |X_i|, i = 1, \ldots, n$.
2. Determinate lo stimatore di massima verosimiglianza $\hat{\beta}$ di β e verificate che è non distorto.
3. Determinate k tale che $P_\beta\left(\frac{\hat{\beta}}{\beta} \geq k\right) = \gamma$, in funzione del numero di osservazioni n e del quantile di ordine $1 - \gamma$ della funzione di ripartizione chiquadrato con $2n$ gradi di libertà.
4. Determinate un intervallo di confidenza a una coda inferiore per β di livello 90% in corrispondenza della realizzazione campionaria $-1.095, 0.836, -0.256, 1.996, 4.675, 0.971$.

Esercizio 3.2.6 (Tema d'esame 5.2 del 23/02/09 AA 07/08) Abbiamo estratto un campione casuale $X_1, \ldots, X_n$ dalla funzione di densità discreta

$$f(x, \theta) = \theta^{3|x|} \left(1 - 2\theta^3\right)^{1-|x|} \mathbf{1}_{\{-1,0,1\}}(x),$$

dove θ, con $0 < \theta < 1/\sqrt[3]{2}$, è un parametro incognito.

1. Determinate gli stimatori di massima verosimiglianza $\widehat{\theta}$ di θ e $\widehat{\kappa}$ di $\kappa = 2\theta^3$.
2. Determinate $\mathrm{E}(|X_1|)$, $\mathrm{Var}(|X_1|)$ e $\mathrm{E}(\widehat{\kappa})$, $\mathrm{Var}(\widehat{\kappa})$.
3. Discutete qualche proprietà (esatta e asintotica) dello stimatore $\widehat{\kappa}$.
4. Costruite un intervallo di confidenza bilatero asintotico per κ di livello $\gamma = 0.95$, per un campione formato da 30 osservazioni uguali a -1, 55 uguali a 0 e 15 uguali a 1. Quindi deducetene uno (bilatero, asintotico, di livello $\gamma = 0.95$) per θ.

Esercizio 3.2.7 (Tema d'esame 3.3 del 12/09/06 AA 05/06) Il tempo di funzionamento in mesi di un componente prima di guastarsi può essere modellato come una variabile aleatoria T continua con densità

$$f(t, \theta) = \frac{1}{6^{\frac{1}{\theta}}} \times \frac{1}{2\theta} t^{\frac{1}{2\theta} - 1} \mathbf{1}_{(0,36)}(t), \quad \theta > 0$$

Il parametro θ è incognito.

1. Determinate in funzione di θ la caratteristica κ data dalla probabilità che un componente non si guasti nei primi 6 mesi di vita.

In un esperimento vengono messi in prova simultaneamente 169 componenti e $T_1, \ldots, T_{169}$ sono i tempi di funzionamento rilevati.

2. Determinate gli stimatori di massima verosimiglianza di θ e di κ.

Sia $\hat{\theta}$ lo stimatore di massima verosimiglianza individuato al punto precedente.

3. Determinate media, varianza e distribuzione asintotica di $\hat{\theta}$. (Per comodità vi forniamo i seguenti risultati che è necessario usare: $\mathrm{E}_\theta[\ln(\sqrt{T})] = \dfrac{\mathrm{E}_\theta(\ln T)}{2} = \ln 6 - \theta$ e $\mathrm{E}_\theta[(\ln(\sqrt{T}))^2] = \dfrac{\mathrm{E}_\theta((\ln T)^2)}{4} = (\ln 6 - \theta)^2 + \theta^2$.)
4. Costruite un intervallo di confidenza asintotico bilatero per θ di livello 90% nel caso che $\displaystyle\sum_{j=1}^{169} \ln \sqrt{Y_j} = \frac{\sum_{j=1}^{169} \ln Y_j}{2} = -35.49$. Quindi deducetene uno per κ sempre di livello 90%.

Supponete ora di aver cancellato per distrazione l'unico file contenente i dati e che la sola informazione preservata sia la seguente: esattamente 52 componenti su 169 hanno funzionato per più di 6 mesi.

5. Stimate θ e κ sulla base di quest'ultima informazione. Quindi usate questa stima di κ per costruire un nuovo intervallo di confidenza asintotico bilatero per κ approssimativamente di livello 90%.

Esercizio 3.2.8 (Tratto in parte da Tema d'esame 3.2 del 05/09/2008 AA 07/08) Il tempo di vita T, espresso in anni, dei sacchetti di plastica biodegradabili del supermercato ZZ ha densità

$$f(t, b) = \begin{cases} \frac{b}{2}\left(1 - \frac{t}{2}\right)^{b-1} & \text{se } 0 < t < 2 \\ 0 & \text{altrimenti ,} \end{cases} \qquad \text{con } b > 0.$$

Il parametro positivo b è incognito e per stimarlo si registrano le durate di un campione di 75 sacchetti, $T_1, \ldots, T_{75}$, e si ottiene che $\displaystyle\sum_{j=1}^{75} \ln\left(1 - \frac{t_j}{2}\right) = \ln\left[\prod_{j=1}^{75}\left(1 - \frac{t_j}{2}\right)\right] \simeq -28.98$.

1. Calcolate la probabilità che un sacchetto estratto a caso dalla popolazione considerata duri meno di 15 mesi e la probabilità che un sacchetto estratto a caso dalla popolazione considerata duri un periodo compreso fra i 12 e i 18 mesi.

2. Usando il metodo di massima verosimiglianza stimate il parametro incognito b

3. Verificate che l'Informazione di Fisher del modello statistico $\{f(t,b), b > 0\}$ è $I(b) = 1/b^2$.
(*Suggerimento*: $\int_0^2 \left[\ln\left(1-\frac{t}{2}\right)\right]^k \frac{b}{2}\left(1-\frac{t}{2}\right)^{b-1} dt = \frac{(-1)^k k!}{b^k}, \forall k = 1, 2, \ldots$)

4. Determinate un intervallo di confidenza asintotico unilatero del tipo $(0, c)$ di livello 95% per il parametro b, per la probabilità che un sacchetto duri meno di 15 mesi e per la probabilità che un sacchetto duri fra i 12 e i 18 mesi.

Esercizio 3.2.9 (Tema d'esame 1.2 del 25/06/07 AA 06/07) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione di densità
$f(x,\theta) = \frac{1}{\theta}(1+x)^{-\frac{1}{\theta}-1}\mathbf{1}_{(0,\infty)}(x)$, con θ parametro positivo incognito.

1. Determinate uno stimatore di θ usando il metodo di massima verosimiglianza.

2. Stabilite se lo stimatore di massima verosimiglianza di θ trovato al punto 1. sia efficiente.

3. Determinate la distribuzione di $Y_j = \ln(1 + X_j)$ per ogni $j = 1, \ldots, n$ (ln indica il logaritmo in base naturale). Quindi deducete la distribuzione dello stimatore di massima verosimiglianza di θ trovato al punto 1.

4. Usate i risultati trovati ai punti 1. e 3. per calcolare un intervallo di confidenza esatto per θ unilatero della forma $(0, c)$ di livello $\gamma = 90\%$ per il campione di quattro osservazioni 0.11, 0.94, 0.48, 1.23 estratte da $f(x, \theta)$.

Esercizio 3.2.10 (Tema d'esame 1.2 del 01/07/08 AA 07/08) Una variabile aleatoria continua X è detta *lognormale di parametri* μ *e* σ se il suo logaritmo naturale $\log(X)$ ha densità $\mathcal{N}(\mu, \sigma^2)$. Comunque, qualora ne dobbiate aver bisogno, trovate l'espressione di una densità $f(x; \mu, \sigma^2)$ lognormale di parametri μ e σ a pagina 1 del Formulario.

Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione lognormale di parametri μ nota e uguale a 0, e $\sigma > 0$ incognito.

1. Determinate uno stimatore $\hat{\sigma}^2$ di σ^2 usando il metodo di massima verosimiglianza.

2. Discutete tutte le proprietà dello stimatore $\hat{\sigma}^2$ che conoscete, esatte e asintotiche.

3. Determinate un intervallo di confidenza bilatero di livello 95% per σ^2 per un campione di 55 osservazioni per il quale abbiamo ottenuto $\sum_{j=1}^{55}(\log X_j)^2 = 3644.0$.

4. ...

Capitolo 4

Verifica di ipotesi

4.1 Test di Neyman Pearson e del rapporto di verosimiglianza

Esercizio 4.1.1 (Tema d'esame 1.1 del 01/07/08 AA 07/08) Sia X_1 un'unica osservazione estratta da una popolazione di densità

$$f(x;\theta) = 2\theta x(1-x^2)^{\theta-1}\mathbf{1}_{(0,1)}(x)$$

con θ parametro positivo incognito.

1. Costruite il test più potente di livello $\alpha = 4\%$ per verificare $H_0 : \theta = 1$ contro $H_1 : \theta = 10$.
2. Calcolate la probabilità di errore di secondo tipo del test costruito al punto 1.
3. Qual è la probabilità di prendere la corretta decisione, se effettivamente $\theta = 10$?
4. Se abbiamo osservato $X_1 = 0.5$, quanto vale il p-value? Cosa si può concludere?

Esercizio 4.1.2 (Tema d'esame 1.1 del 29/06/04 AA 03/04) Un informatico smemorato ha scritto un codice per la generazione di numeri casuali dalla densità continua

$$f(x,\theta) = \begin{cases} \theta(\theta+1)x^{\theta-1}(1-x) & x \in (0,1) \\ 0 & \textit{altrove} \end{cases}, \qquad \theta > 0$$

ma non ricorda se il valore di θ sia 1 o 10. Pigro come pochi, genera un solo numero casuale x_1 e vi chiede di aiutarlo a decidere fra l'ipotesi nulla $H_0 : \theta = 1$ e l'ipotesi alternativa $H_1 : \theta = 10$.

1. Costruite il test più potente di livello α per il precedente problema di ipotesi, sulla base di una sola osservazione x_1. Fornite esplicitamente la regione critica del test.

2. Se $\alpha = 2.5\%$ e $x_1 = 0.88$ cosa deciderà l'informatico sulla base del test costruito al punto 1.?

3. Sia $\alpha = 2.5\%$. Calcolate la probabilità di errore di secondo tipo (β) del test costruito al punto 1.

Esercizio 4.1.3 (Tema d'esame 4.1 del 13/02/06 AA 04/05) Vi chiedo di aiutarmi a decidere se un'unica osservazione continua X abbia densità

$$f_a(x) = 2xe^{-x^2}\mathbf{1}_{(0,\infty)}(x) \text{ oppure } f_b(x) = 20xe^{-10x^2}\mathbf{1}_{(0,\infty)}(x)$$

A tal fine dovete:

1. calcolare la funzione di ripartizione della densità f_a e quella della densità f_b, in un generico punto $x > 0$;

2. costruire il test uniformemente più potente di livello $\alpha \in (0, 1)$ per verificare l'ipotesi nulla $H_0 : f = f_a$ contro l'ipotesi alternativa $H_1 : f = f_b$;

3. calcolare il p-value del test al punto 2. quando $X = 0.15$ e il p-value del test quando $X = 0.49$. Se $\alpha = 5\%$ e $X = 0.15$ cosa deciderò? E se $X = 0.49$?

Vi chiedo ora di scambiare le ipotesi e di verificare l'ipotesi nulla $H_0 : f = f_b$ contro l'ipotesi alternativa $H_1 : f = f_a$.

4. Se per verificare l'ipotesi nulla $H_0 : f = f_b$ contro l'ipotesi alternativa $H_1 : f = f_a$ usiamo un test di regione critica $X \geq k_2$, per quale valore di k_2 questo test ha ampiezza $\alpha \in (0, 1)$?

5. Calcolate la potenza del test costruito al punto 4., cioè del test per verificare l'ipotesi nulla $H_0 : f = f_b$ contro l'ipotesi alternativa $H_1 : f = f_a$, quando $\alpha = 5\%$.

6. Se $\alpha = 5\%$ e $X = 0.49$, in base al secondo test del punto 4., accettate o rifiutate f_b? Più in generale, se $\alpha = 5\%$, per quali valori dell'unica osservazione X i test di ipotesi dei punti 2. e 4. portano alla stessa decisione?

Esercizio 4.1.4 (cfr Esempio . . . , Sezione 10.3 in Trivedi (2002)) [1] La variabile aleatoria X rappresenta la frazione di memoria principale allocabile di un server richiesta da un *job* qualsiasi. Si assuma che X abbia la seguente densità:

$$f(x, \theta) = \begin{cases} \frac{1}{\theta} x^{\frac{1}{\theta}-1} & 0 < x < 1 \\ 0 & \text{altrove} \end{cases}$$

dove θ è un parametro positivo incognito. Un valore basso di θ implica la preponderanza di "grossi" jobs. Se $\theta = 1$ invece la distribuzione di richieste di memoria è uniforme. Si considerino le seguenti ipotesi: $H_0 : \theta = 2$ contro $H_1 : \theta = 0.2$. Sia ora $X_1, \dots, X_n$ un campione casuale estratto da $f(x, \theta)$.

1. Determinate la densità di $-\sum_{j=1}^n \log(X_j)$ nell'ipotesi H_0 e nell'ipotesi H_1.
[`Risp:` $-\sum_{j=1}^n \log(X_j) \sim \chi^2_{2n}$ se $\theta = 2$ e $\Gamma(n, 0.2)$ se $\theta = 0.2$]

2. Costruite il test più potente di livello $\alpha \in (0, 1)$. [`Risp:` $\mathcal{G} = \{-\sum_{j=1}^n \log(x_j) \leq \chi^2_{2n}(\alpha)\}$]

3. Calcolate la potenza del test di livello $\alpha = 5\%$ per un campione di 3 osservazioni. Come varia la potenza del test in funzione di α? [`Risp:` $\pi(0.2) \in (0.975, 0.99)$; π `è funzione crescente di` α]

4. Siano $n = 3$ e $x_1 \times x_2 \times x_3 = 0.13$. Che decisione prendete ai livelli $\alpha = 5\%$, $\alpha = 1\%$ e $\alpha = 10\%$? [`Risp: per` $1\%, 5\%, 10\%$ `accetto, accetto, rifiuto` H_0`, rispettivamente`]

Esercizio 4.1.5 (Tema d'esame 2.1 del 18/07/06 AA 05/06) La famosa azienda OISAC, leader nella produzione di batterie per computer portatili, dichiara che la sua nuova batteria ha un'autonomia di ben 8 ore, contro l'autonomia di quella di vecchia generazione pari a 3.92 (cioè 3 ore e 55 minuti circa).

[1] Trivedi. K. S. (2002) *Probability and Statistics with Reliability, Queueing, and Computer Science Applications*, 2nd Edition, Wiley

Ora noi vogliamo verificare quanto dichiarato da OISAC nell'ipotesi che la durata della nuova batteria sia una variabile aleatoria continua con densità $f(x,\theta)$ data da

$$f(x,\theta) = \frac{1}{2\theta\sqrt{x}} \exp\left\{-\frac{\sqrt{x}}{\theta}\right\} \mathbf{1}_{(0,\infty)}(x), \qquad \theta > 0$$

con θ incognito. Per questo usiamo un campione casuale $X_1, \ldots, X_n$ estratto da $\{f(x,\theta), \theta > 0\}$.

A tal fine, rispondete alle seguenti domande:

1. ...
2. determinate la densità di $\sqrt{X}$ se $X \sim f(x,\theta)$ e di $\frac{2}{\theta}\sum_{j=1}^{n}\sqrt{X_j}$;
3. determinate la media di $X \sim f(x,\theta)$ in funzione di θ; quindi traducete le ipotesi H_0 :"$\mu = 3.92$" e H_1 :"$\mu = 8$" in ipotesi su θ;
4. verificate che il test di Neyman-Pearson per verificare l'ipotesi nulla H_0:"$\mu = 3.92$" contro l'alternativa H_1:"$\mu = 8$" ha regione critica della forma $\left\{(x_1, \ldots, x_n) : \sum_{j=1}^{n}\sqrt{x_j} \geq k\right\}$;
5. determinate k nel caso di $n = 6$ osservazioni e di una significatività $\alpha = 10\%$;
6. calcolate la potenza del test al punto 4., o almeno indicate un intervallo in cui tale valore cade.

Esercizio 4.1.6 Sia X_1 un'osservazione di densità

$$f(x,\theta) = \begin{cases} \frac{2}{\theta}\left(1 - \frac{x}{\theta}\right) & \text{se } 0 < x < \theta \\ 0 & \text{altrove} \end{cases}$$

dove θ è un parametro positivo incognito.

a) Determinate lo stimatore di massima verosimiglianza di θ. [`Risp:` $\hat{\theta}_{MLE} = 2X_1$]

b) Costruite il test del rapporto di verosimiglianza di livello $\alpha = 5\%$ per il seguente problema di verifica di ipotesi: $H_0 : \theta = 1$ *versus* $H_0 : \theta \neq 1$. [`Risp:` $X_1(1-X_1) \leq 1-(1-\alpha)^2/4 \simeq 0.024$]

c) Se $X_1 = 0.06$, accettate o rifiutate H_0? [`Risp:` Accettiamo H_0]

Esercizio 4.1.7 (cfr Example 5 page 152,153 in Pestman 1998) La va X assume solo i valori $-1, 0, 1$. Sulla base di una sola osservazione X, vogliamo verificare $H_0 : P(X = -1) = P(X = 0) = P(X = 1) = 1/3$ contro $H_1 : P(X = -1) = P(X = 1) = 1/4$ e $P(X = 0) = 1/2$

1. Costruite il test più potente di livello $\alpha = 1/3$ [`Risp:` Rifiuto H_0 se $X_1 = 0$]
2. Se $X = 0$ cosa decidete?
3. Quanto vale la potenza del test? [1/2]

4.2 Test su media e varianza di popolazione gaussiana

Esercizio 4.2.1 [2] La ditta ACME vuole dotare i suoi dipendenti di laptop. A tal fine l'addetto hardware della ACME ordina 30 macchine identiche alla ditta Busy. La ditta Busy dichiara

[2] Tratto da un tema d'esame di MPSPS del 14 giugno 2001

sulle istruzioni che le batterie dei suoi laptop hanno una durata media di **almeno** 400 cicli di carica/scarica con uno scarto del 10% (significa che $\sigma_0 = 40$). Per verificare la veridicità dell'affermazione della ditta Busy l'addetto hardware della ACME verifica le batterie dei 30 laptop riscontrando una media campionaria di 392.5 cicli carica/scarica prima del deterioramento della batteria.

1. Verificare, sulla base del campione analizzato, l'ipotesi H_0: "la ditta Busy afferma il vero" contro H_1: "la ditta Busy afferma il falso" al 5% di significatività, indicando esplicitamente la regione critica e la regola che porta al rifiuto di H_0. [`Risp:` $\mathcal{G} = \{\bar{x} \leq 400 - 40 \times 1.645/\sqrt{30}\}$]

2. Con riferimento al punto precedente, indicare per quali valori di α va rifiutata l'ipotesi nulla con i dati a disposizione. [`Risp:` $\alpha \geq \Phi(-1.03) \simeq 0.1515$]

3. Fornite analiticamente e rappresentate graficamente la funzione di potenza del test di livello 5% costruito al punto 1. [`Risp:` $\pi(\mu) = \Phi(\sqrt{30}(387.99 - \mu)/40)$, $\mu < 400$]

4. Se la durata media effettiva delle batterie del laptop è pari a 380 cicli di carica/scarica, quante volte (percentualmente) si concluderebbe che la ditta afferma il falso? [`Risp:` $\pi(380) \simeq 86\%$]

Esercizio 4.2.2 (Tema d'esame 3.1 del 19/09/05, AA 04/05) Un chimico ha sviluppato un nuovo tipo di batterie per calcolatrici che, egli sostiene, dura "significativamente più a lungo" rispetto a quelle correntemente sul mercato.
Un'azienda intenzionata a produrle misura le durate di un campione di n batterie del nuovo tipo. L'azienda sa che la durata delle batterie per calcolatrici correntemente sul mercato è modellabile mediante una variabile aleatoria gaussiana di media 100.3 min e deviazione standard 6.25 min; inoltre, da verifiche preliminari, l'azienda può assumere che anche la durata del nuovo tipo di batterie sia gaussiana con deviazione standard 6.25 min. Infine, l'azienda è disposta a commettere un errore di prima specie di probabilità al più 5% di produrre batterie del nuovo tipo effettivamente *non migliori di quelle correnti*.

1. Impostate un opportuno test d'ipotesi per aiutare l'azienda a decidere se produrre o meno le nuove batterie; in particolare dovrete specificare: ipotesi nulla, ipotesi alternativa e regione critica.

2. Sono state misurate le durate di 16 batterie del nuovo tipo: se la durata media è risultata 104.6 minuti, cosa deciderà l'azienda usando il vostro test?

2bis Calcolate il p–value del test se $n = 16$ e $\bar{x} = 104.6$. [`Risp:` p–value=0.00298 $\simeq 0.3\%$]

3. Sempre nell'ipotesi che si misurino le durate di 16 batterie del nuovo tipo, qual è la potenza del vostro test se effettivamente le nuove batterie avessero durata media di 105 minuti?

4. Determinate il numero minimo n di durate da rilevare affinchè, se effettivamente le nuove batterie avessero durata media di 105 minuti, allora almeno il 95% delle volte si concluderebbe correttamente che le nuove batterie sono migliori delle correnti.

Esercizio 4.2.3 L'ammontare settimanale del caffè prodotto da una piccola azienda e venduto in una settimana può descriversi con un modello normale di media μ e varianza σ^2 entrambe incognite. Poichè l'azienda deve essere venduta, per fissarne il valore di mercato, si procede alla determinazione della vendita media di caffè [per settimana]. Il vecchio proprietario afferma che la quantità media di caffè venduta per settimana supera i 30 quintali. Un possibile acquirente dell'azienda, non convinto di ciò, analizza i dati delle ultime 9 settimane. Da questi risulta una vendita settimanale media di 28.6 quintali con una deviazione standard campionaria di 1.5 quintali. Sulla base di ciò l'acquirente sostiene che la vendita media settimanale non supera i 30 quintali.

1. Sulla base dei dati forniti, quale fra le posizioni del vecchio proprietario e dell'acquirente

è plausibile? Scegliete opportunamente ipotesi nulla e alternativa, individuate un test e fissate come livello di significatività $\alpha = 1\%, 10\%$ e 25%.

2. A parità di dati e livello di significatività, la risposta al punto precedente dipende dalla scelta delle ipotesi nulla e alternativa? Per esempio, la risposta cambia se scambiate la nulla con l'alternativa?

Esercizio 4.2.4 (Tema d'esame 5.1 del 08/03/06, AA 04/05) Sia

$$10.42,\ 6.69,\ 11.18,\ 7.94,\ 9.65,\ 8.38,\ 11.80,\ 8.28,\ 10.84,\ 10.72$$

una realizzazione del campione casuale $X_1, \ldots, X_{10}$ estratto dalla densità gaussiana di varianza 2.25 e media incognita μ.

1. Determinate un intervallo di confidenza (IC) simmetrico di livello 90% per μ.

Siete disposti ad allargare il vostro campione per fare inferenza su μ.

2. Qual è il numero minimo di dati necessari per essere fiduciosi almeno al 90% che l'errore che commettete stimando μ con la media campionaria non sia più grande di 0.26?

Sia n^* la risposta al punto 2.. Sulla base di quanto discusso al punto 2., decidete di raccogliere ulteriori $n^* - 10$ osservazioni $X_{11}, \ldots, X_{n^*}$. Sapete che la media campionaria delle nuove $n^* - 10$ osservazioni vale 8.96.

3. Aggiornate la stima puntuale di μ usando tutte le informazioni provenienti dal campione allargato $X_1, \ldots, X_{n^*}$; quindi determinate numericamente un IC simmetrico di livello 90% per μ.

Vogliamo ora studiare il problema di ipotesi $H_0 : \mu = 10$ *versus* $H_1 : \mu \neq 10$ e per questo estraiamo un campione di ampiezza $n = 20$ (totalmente nuovo) sempre dalla famiglia gaussiana $(\mu, 2.25)$. In corrispondenza di questo nuovo campione, $(9.307, 10.410)$ è l'IC simmetrico di μ di livello 0.95.

4. Sapere che $(9.307, 10.410)$ è l'IC simmetrico di μ di livello 0.95 basato su $n = 20$ osservazioni è sufficiente per scegliere quale ipotesi rifiutare a livello di significatività $\alpha = 1\%$? Se sì, quale decisione prendete?

Esercizio 4.2.5 (Tema d'esame 1.4 del 29/06/04, AA 03/04) Una compagnia di assicurazioni deve eseguire uno studio per stimare gli indennizzi pagati a seguito di incidenti automobilistici senza lesioni alle persone. Da studi precedenti è emerso che si può assumere che tali importi abbiano densità gaussiana con media μ incognita e deviazione standard nota e pari a 900 euro. Su un nuovo campione casuale di 100 incidenti del suddetto tipo è stato osservato un indennizzo medio pari a 5562 euro.

1. Determinate un intervallo di confidenza di livello 94% per il parametro μ.
2. Verificate l'ipotesi $H_0 : \mu = 5500$ contro $H_1 : \mu \neq 5500$ al livello $\alpha = 6\%$.
3. In realtà la compagnia assicurativa ritiene che l'intervallo di confidenza costruito al punto 1. non sia sufficientemente preciso. Decide quindi di condurre uno studio più vasto, cioè su un campione più numeroso. Determinate il numero minimo di casi da esaminare affinchè la lunghezza dell'intervallo di confidenza per μ non superi i 300 euro.

Esercizio 4.2.6 (Tema d'esame 4.3 del 15/02/05, AA 03/04) Un segnale di valore μ trasmesso dalla sorgente A viene raccolto dal ricevente B con un rumore additivo gaussiano di media nulla e varianza $\sigma^2 = 16$. Per ridurre l'errore, lo stesso segnale viene trasmesso 9 volte da A a B.

1. Quale fiducia avete che il segnale trasmesso da A fosse compreso fra 6.38 e 11.62?

2. B ha motivo di supporre che il segnale inviato dovesse essere 12. Verificate l'ipotesi nulla $H_0 : \mu = 12$ contro l'alternativa $H_1 : \mu \neq 12$, a livello di significatività $\alpha = 10\%$.

3. Determinate il p-value dei dati del test per l'ipotesi nulla $H_0 : \mu = 12$ contro l'alternativa $H_1 : \mu \neq 12$.

Esercizio 4.2.7 (Tema d'esame 3.2 del 12/09/06 AA 05/06) È noto che il peso di una donna italiana di 25 anni è una variabile aleatoria gaussiana X di media 62 Kg e varianza 16 Kg^2. Il programma dimagrante `WW` consente di ridurre il peso di una quantità $Y \sim \mathcal{N}(\mu, 20)$, con μ parametro incognito. In altre parole, alla fine della dieta il peso di una donna che ha seguito il programma è $Z = X - Y$ e assumiamo che X, Y siano indipendenti.

Nel depliant pubblicitario del programma dimagrante `WW` si sostiene che chi segue scrupolosamente il programma perde mediamente 5 Kg, e noi, per stabilire se il messaggio è ingannevole o no, abbiamo deciso di verificare il programma `WW` su 64 donne; $Z_1, \dots, Z_{64}$ è il campione casuale ottenuto.

1. Qual è la distribuzione di Z?

2. Costruite un test su μ tale che sia al più pari al 4% la probabilità di commettere l'errore di prima specie di ritenere ingannevole il messaggio quando in realtà seguendo il programma `WW` si perdono effettivamente (almeno) 5 Kg.

3. Fornite analiticamente e rappresentate graficamente la funzione di potenza del test di livello 4% costruito al punto 2.

4. Se il campione dei dati ha fornito $\sum_{j=1}^{64} Z_j = 3712.0$, qual è il p-value del test?

Esercizio 4.2.8 (Tema d'esame 2.2 del 10/07/07, AA 06/07) Una compagnia di assicurazioni deve eseguire uno studio per stimare gli indennizzi pagati a seguito di "*danni provocati dai figli minori per uso di giocattoli*". Sospetta infatti che mediamente questi importi siano aumentati rispetto al triennio precedente in cui l'indennizzo medio era stato di 3500 euro. Per questo motivo viene analizzato un nuovo campione casuale di 16 incidenti del suddetto tipo per il quale si osserva una media campionaria degli indennizzi pari a 3525.438 euro. Inoltre, da studi precedenti è emerso che si può assumere che tali importi abbiano densità gaussiana con deviazione standard nota e pari a 50 euro.

1. Aiutate la compagnia di assicurazioni a decidere se il suo sospetto sia fondato o meno usando un opportuno test di ipotesi al livello $\alpha = 6\%$. Sulla base dei dati forniti, il sospetto è fondato?

2. Sulla base dei dati a disposizione, quanto siete confidenti che attualmente l'indennizzo medio abbia superato i 3506 euro?

3. Con riferimento al test di ipotesi costruito al punto 1., indicate per quali livelli α ritenete il sospetto fondato, con i dati a disposizione.

4. Determinate per quali valori dell'indennizzo medio la probabilità di errore di secondo tipo è inferiore o uguale al livello $\alpha = 6\%$ del test al punto 1.

Esercizio 4.2.9 (Tratto da tema d'esame 2.1 del 15/07/04 AA 03/04) La ditta Baltic Sea produce macchine per l'inscatolamento di caviale. A causa delle fluttuazioni casuali la quantità di caviale dosata dalla macchine è una variabile aleatoria X gaussiana con media nota $\mu = 30$ grammi e varianza incognita σ^2. Prima di procedere all'acquisto di una di queste macchine controllo le quantità x_i (misurate in grammi) di caviale contenute in un campione casuale di 100 scatolette ottenendo:

$$\sum_{i=1}^{100} x_i = 3006 \qquad\qquad \sum_{i=1}^{100} x_i^2 = 90711$$

Inoltre, sono disposto a commettere un errore di prima specie di probabilità al più pari ad α di acquistare una macchina che abbia un'imprecisione (=deviazione standard di X) effettiva maggiore o uguale di 2 grammi.

1. Fornite una stima puntuale della varianza σ^2.
2. Impostate un opportuno test sulla varianza, specificando: ipotesi nulla, ipotesi alternativa e una regione critica di ampiezza $\alpha = 2.5\%$.
3. Fornite analiticamente e rappresentate graficamente la funzione di potenza del test di ampiezza $\alpha = 2.5\%$ costruito al punto 2.
4. Fornite un valore approssimato della probabilità di errore di secondo tipo in $\sigma^2 = 3$.
5. Fornite un valore approssimato del p-value.0

Esercizio 4.2.10 (Tema d'esame 5.1 del 07/03/05, AA 03/04) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione gaussiana di densità

$$f(x, \theta_1, \theta_2) = \sqrt{\frac{\theta_1}{\pi}} \mathrm{e}^{-\theta_1 (x-\theta_2)^2} \qquad \theta_1 > 0 \text{ e } \theta_2 \in \mathbb{R}$$

Entrambi i parametri θ_1, θ_2 sono incogniti.

1. Determinate uno stimatore di θ_1 usando il metodo di massima verosimiglianza.
2. Determinate un intervallo di confidenza a due code per θ_1 di livello $\gamma = 95\%$.
3. Avete ora a disposizione il campione di quattro osservazioni: $x_1 = -0.17$, $x_2 = 0.71$, $x_3 = 2.17$ e $x_4 = 1.00$ e dovete scegliere fra l'ipotesi nulla $H_0 : \theta_1 = 0.5$ e l'alternativa $H_1 : \theta_1 \neq 0.5$. Quale decisione prendete al livello $\alpha = 5\%$? (Giustificate rigorosamente la risposta).

Esercizio 4.2.11 (Tema d'esame 2.1 del 14/07/05, AA 04/05) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione gaussiana di densità

$$f(x, \theta) = \frac{\theta^{3/2}}{\sqrt{3\pi}} e^{-\frac{\theta^3 x^2}{3}}, \qquad \theta > 0$$

dove θ è un parametro incognito.

1. ...

2. Costruite un test per verificare l'ipotesi nulla $H_0 : \theta = 1$ contro l'alternativa $H_1 : \theta \neq 1$ a livello $\alpha = 5\%$. Se avete rilevato il campione di quattro osservazioni: $-0.17,\ 0.71,\ 2.17,\ 1.00$, accettate o rifiutate H_0?

3. Se X_1, X_2, X_3, X_4 sono *i.i.d.* $\sim \mathcal{N}(0, 12)$, qual è la distribuzione di $\frac{\sum_{j=1}^4 X_j^2}{12}$? Scrivete l'espressione della densità e determinate esplicitamente la funzione di ripartizione.

4. Calcolate la potenza del test costruito al punto 2. in $\theta = 1/2$. (*Sugg: è utile quanto trovato al punto precedente.*)

Esercizio 4.2.12 (Tema d'esame 5.1 del 23/02/09 AA 07/08) Un macchinario che produce Compact Disk (CD) è tarato in modo tale che il raggio di ogni CD, espresso in cm, sia una variabile aleatoria gaussiana R di media nominale $\mu_0 = 12$cm e varianza incognita σ^2. Inoltre, il produttore di questo macchinario dichiara che non più del 2% dei CD prodotti con questo macchinario ha lunghezza del raggio distante almeno 0.308cm dalla media nominale.

La ditta `xxx`, produttrice di CD e intenzionata ad acquistare questo macchinario, teme che il produttore del macchinario non dichiari il vero. Per questo motivo controlla i raggi r_i di un campione casuale di 25 CD, ottenendo:

$$\sum_{i=1}^{25} r_i = 299.0\text{cm} \qquad \sum_{i=1}^{25} r_i^2 = 3576.3\text{cm}^2$$

Sia ora κ la percentuale, sull'intera popolazione di CD, di quelli che hanno lunghezza del raggio distante almeno 0.308cm dalla media nominale μ_0.

1. Determinate κ in funzione di σ^2. Quindi, stabilite per quali valori di σ^2 la percentuale κ è minore o uguale al 2%.

2. Fornite una stima puntuale di σ^2 e di κ.

3. Sulla base di questi dati, la ditta `xxx` può accusare il produttore del macchinario di non dichiarare il vero? Per rispondere impostate un opportuno test sulla varianza, specificando: ipotesi nulla, ipotesi alternativa e una regione critica di ampiezza 5%.

4. Fornite analiticamente e rappresentate graficamente la funzione di potenza del test costruito al punto 3.

Esercizio 4.2.13 In un processo industriale una macchina miscela due sostanze versandone quantità opportune in tanti barattoli. Le proporzioni desiderate sono 1:2 (cioè ogni barattolo dovrebbe contenere 1/3 della sostanza A e 2/3 della sostanza B). In pratica le quantità X_A e X_B versate in ogni barattolo sono variabili aleatorie indipendenti e $X_A \sim N(\mu, \sigma^2)$ e $X_B \sim N(2\mu, \sigma^2)$.

1. Qual è il valore massimo ammissibile per σ^2 affinchè la probabilità che la differenza tra $2X_A$ e X_B non superi in valore assoluto 0.1 sia maggiore o uguale a 0.9.

2. Supponendo ora che $X_1^A, \dots X_n^A$ siano n v.a. che rappresentano la quantità di sostanza A versate in n barattoli e $X_1^B, \dots X_n^B$ l'analogo per la quantità B, proponete una regione di rifiuto dell'ipotesi nulla $\sigma^2 \leq (0.027)^2$ contro l'ipotesi alternativa $\sigma^2 > (0.027)^2$, basata sul campione $Z_1, \dots, Z_n$, con $Z_i := \dfrac{2X_i^A - X_i^B}{\sqrt{5}}$ per $i = 1, \dots, n$.

Esercizio 4.2.14 (Tema d'esame 1.2 del 28/06/06 AA 05/06) Una macchina imbottigliatrice è impiegata per riempire flaconi di bagnoschiuma. A causa di fluttuazioni casuali, la quantità di bagnoschiuma per flacone è una variabile aleatoria X gaussiana di media e varianza entrambe incognite. Se la varianza del volume riempito supera 25ml^2, una frazione non accettabile dei flaconi sarà sotto-riempita o sovra-riempita. Se la varianza non supera 25ml^2, la macchina imbottigliatrice è considerata precisa. Per controllare la precisione della macchina imbottigliatrice, abbiamo misurato la quantità di bagnoschiuma presente in 46 flaconi (espressa in ml), e abbiamo ottenuto che la funzione di ripartizione empirica $\hat{F}_{46}$ è

x	242.1	246.5	248.5	251.0	253.5	255.6
$\hat{F}_{46}(x)$	8/46	21/46	27/46	33/46	43/46	1

1. Calcolate la media e la varianza campionarie.

2. Costruite un intervallo di confidenza al 95% per la varianza, unilatero del tipo (c, ∞).

3. Costruite un test sulla varianza tale che sia al più pari a 5% la probabilità di commettere l'errore di prima specie di *ritenere imprecisa* una macchina *effettivamente precisa*.

4. Calcolate la probabilità di errore di secondo tipo del test costruito al punto 3. se la varianza vale effettivamente 42ml^2, o indicate un intervallo dove tale probabilità cade.

Esercizio 4.2.15 (Tema d'esame 5.4 del 06/03/07, AA 05/06) Il regolamento della Comunità Europea n. 975/98 ha stabilito che la moneta da 1 euro ha un peso di 7.50gr. In realtà, le monete coniate presentano pesi differenti dal valore prestabilito di 7.50gr e gli addetti ai lavori sospettano che la macchina che le conia non sia sufficientemente precisa. In particolare, gli addetti sospettano che la deviazione standard del peso di una moneta da 1 euro sia maggiore o uguale a 0.6gr. Per questo implementano il seguente esperimento: si pesano a caso 50 monete ottenendo una somma dei pesi pari a 379.77gr e una somma dei quadrati dei pesi pari a 2894.80gr^2.

Mettendovi sotto ipotesi di gaussianità del peso delle monete da 1 euro, rispondete alle seguenti domande.

1. Costruite un intervallo di confidenza 97.5% unilatero della forma $(0, c)$ per la deviazione standard del peso di una moneta da 1 euro.

2. Impostate un opportuno test di ipotesi tale che sia al più pari a $\alpha = 2.5\%$ la probabilità di commettere l'errore di prima specie di ritenere precisa una macchina che ha effettivamente deviazione standard maggiore o uguale a 0.6gr. Per rispondere dovete specificare: ipotesi nulla, ipotesi alternativa e regione critica di livello $\alpha = 2.5\%$.

3. Fornite analiticamente e rappresentate graficamente la funzione di potenza del test di livello 2.5% costruito al punto 2.

4. Quale decisione prendete a livello $\alpha = 0.1\%$? La vostra decisione cambia per $\alpha = 10\%$?

4.3 Verifica di ipotesi per grandi campioni

Esercizio 4.3.1 (Tema d'esame 3.2 del 16/09/04, AA 03/04) Vogliamo fare inferenza sulla proporzione ϑ degli individui di tipo A presenti in una certa popolazione. Perciò procediamo con il seguente esperimento: effettuiamo n estrazioni casuali con reimmissione da questa popolazione e registriamo il numero di individui di tipo A ottenuti nelle n estrazioni. Sia X_n la variabile aleatoria definita da "numero di individui di tipo A ottenuti nelle n estrazioni".

1. Qual è la densità, la media e la varianza di X_n?

In particolare, siamo interessati a verificare l'ipotesi $H_0 : \vartheta = \vartheta_0$ contro $H_1 : \vartheta = \vartheta_1$ con $\vartheta_0 < \vartheta_1$ ($\vartheta_0, \vartheta_1 \in (0, 1)$).

2. Verificate che la regione critica per il test determinato dal lemma di Neyman-Pearson e basato su X_n abbia la forma: $G = \left\{ \frac{X_n}{n} \geq t \right\}$.

Fissiamo ora i seguenti valori: $\vartheta_0 = 0.5$, $\vartheta_1 = 0.7$ e $\alpha = 5\%$.

3. Se $n = 30$, qual è la f.d.r. approssimata di $\dfrac{X_n}{n}$ sotto H_0? E sotto H_1?

4. Assumete $n = 30$ e determinate t tale che "approssimativamente" la regione critica G abbia ampiezza 5%. Se $X_n = 18$, che decisione prendete?

5. Assumete $n = 30$ e calcolate un valore approssimato probabilità d'errore di seconda specie β.

6. Supponiamo ora che n non sia dato, mentre ϑ_0, ϑ_1 e α rimangono sempre gli stessi. Quale sarà il minimo numero di estrazioni necessario per avere $\beta \leq 0.20$?

Esercizio 4.3.2 (Tema d'esame 4.1 del 15/02/05, AA 03/04) Sospettiamo che due dadi perfettamente identici siano stati entrambi truccati in modo tale che, lanciandoli in coppia, la probabilità di ottenere come somma delle facce superiori il valore 7 sia pari a 1/11. Denotiamo con p la probabilità che la somma delle facce superiori di due dadi uguali e lanciati simultaneamente sia 7.

1. Se i due dadi sono regolari quanto vale p?

Sia p_0 il valore determinato al punto 1. Vogliamo verificare l'ipotesi nulla $H_0 : p = p_0$ contro l'alternativa $H_1 : p = 1/11$. Decidiamo di eseguire questa verifica nel seguente modo: lanciamo 144 volte la coppia di dadi e rifiutiamo H_0 se la somma dei due dadi è 7 al più per 14 volte.

2. Calcolate un valore "approssimato" del livello di significatività α del test.
3. Calcolate un valore "approssimato" della potenza π del test.
4. Calcolate un valore "approssimato" della probabilità di errore di seconda specie β del test.

Esercizio 4.3.3 Vogliamo verificare l'ipotesi nulla che una moneta sia regolare contro l'alternativa che sia stata "truccata" in modo tale che testa sia meno probabile di croce. Decidiamo di rifiutare l'ipotesi di assenza di trucco se, lanciata la moneta 10 volte, otteniamo testa al più 4 volte su 10.

1. Qual è il livello di significatività di questo test di ipotesi? [0.3769]

2. Quanto vale la probabilità di errore di seconda specie se la probabilità di ottenere testa è la metà di quella di ottenere croce? [0.2131]

3. Come cambiano i risultati ai punti precedenti nel caso che decidiamo di rifiutare l'ipotesi di assenza di trucco se, lanciata la moneta 100 volte, otteniamo testa al più 40 volte su 100?

Esercizio 4.3.4 Un'urna contiene 400 palline rosse e nere e la percentuale di rosse $\vartheta \in (0, 1)$ è incognita. Per fare inferenza su θ si sono estratte a caso con reimmissione n palline e si è registrato il numero X_n delle rosse ottenute.

1. Verificate con un test asintotico l'ipotesi nulla H_0:"*L'urna contiene 64 palline rosse*" contro l'alternativa H_1:"*L'urna contiene più di 64 palline rosse*" a livello 6%. Scrivete esplicitamente la regione critica se avete effettuato n estrazioni con reimmissione.

2. Determinate il minimo numero n di estrazioni con reimmissione necessarie perché sia al più pari a 0.10 la probabilità di errore di seconda specie del test quando le palline rosse dell'urna sono effettivamente 120.

Esercizio 4.3.5 (Tema d'esame 2.1 del 16/07/08, AA 07/08) Per fare inferenza statistica sulla probabilità p che un allievo regolarmente iscritto a un appello d'esame non si presenti a sostenere l'esame, sono state controllate iscrizioni e presenze dell'appello del 25/06/07 di Statistica: esattamente 27 dei 128 allievi regolarmente iscritti sono risultati assenti.

1. Verificate l'ipotesi nulla $H_0 : p \leq 0.25$ contro l'alternativa $H_1 : p > 0.25$, con un test asintotico di livello $\alpha = 3\%$.
2. Determinate la potenza del test al punto 1. quando il vero valore di p è 0.35.
3. Costruite un intervallo di confidenza bilatero asintotico per p di confidenza $\gamma = 90\%$. Nell'approssimazione fermatevi alla seconda cifra decimale.

Vengono preparate le copie della traccia del compito, ovviamente una per ogni allievo, alla chiusura delle iscrizioni all'appello del 18 luglio[3], a cui risultano iscritti 100 allievi.

4. Sulla base dei risultati al punto 3., quanto deve essere fiducioso il docente che il numero medio di copie preparate inutilmente sia compreso fra 15 e 27?

Esercizio 4.3.6 (Tema d'esame 1.2 del 30/06/05, AA 04/05) La prova d'esame del corso di CPS è composta da un test di dieci domande a risposta multipla e una parte a esercizi e si è ammessi alla parte a esercizi solo se si risponde correttamente ad almeno cinque domande. L'80% degli allievi che sostengono la prova supera il test, ma il 40% (di questi ultimi) NON raggiunge la sufficienza negli esercizi.

Il docente di CPS ha il dubbio che il test non filtri adeguatamente la preparazione degli allievi. Per questo motivo, in un appello "affollato" da 120 allievi, a sorpresa il docente salta il test e ammette d'ufficio tutti alla parte a esercizi. In quell'appello risultano promossi 78 allievi su 120.

1. Determinate una stima puntuale della percentuale $\theta 100\%$ degli allievi che raggiungono la sufficienza negli esercizi, in una prova che non prevede test. [`Risp:` 0.65]
2. Verificate l'ipotesi nulla $H_0 : \theta = 64\%$ contro l'alternativa $H_1 : \theta > 64\%$ a livello approssimativamente $\alpha = 5\%$, proponendo un opportuno test e calcolate la funzione di potenza del test proposto in $\theta = 66\%$ [`Risp:` $\pi(0.66) = 0.11$]
3. Esprimete in funzione di θ la caratteristica della popolazione $\kappa = \kappa(\theta)$ data dalla percentuale $\kappa 100\%$ degli allievi bocciati al test che, se ammessi agli esercizi, raggiungerebbe la sufficienza e fornitene una stima puntuale. [`Risp:` $\kappa = 5\theta - 2.4; 0.85$]
4. Quanto valgono $\kappa(64\%)$ e $\kappa(66\%)$? [`Risp:` 80%, 90%]
5. Verificate l'ipotesi nulla $H_0 : \kappa = 80\%$ contro l'alternativa $H_1 : \kappa > 80\%$ ad un livello approssimato $\alpha = 5\%$ e calcolate la funzione di potenza in $\kappa = 90\%$. [`Risp:` Stesso test di prima, stessa potenza].

Esercizio 4.3.7 [4] Un'azienda che produce pannelli solari sottopone i suoi prodotti a un controllo di qualità. Il 75% dei pannelli supera il controllo e viene messo in vendita. Comunque, il 2% di essi si guasterà e dovrà essere sostituito prima dello scadere della garanzia sul prodotto. Il $\theta 100\%$ di quelli che non superano il controllo di qualità sono in realtà perfettamente funzionanti e lo sarebbero fino alla scadenza della garanzia. Il parametro θ è incognito.

Per stimare θ non è possibile utilizzare i pannelli che non hanno superato il controllo, in quanto essi vengono immediatamente disassemblati. Quindi si utilizza un campione di 200 pannelli che non sono stati sottoposti al controllo, ottenendo che 165 di essi funzionano per tutto il tempo garantito. Determinare una stima puntuale e un intervallo di confidenza di livello 0.9 per θ. *Sugg.: calcolare la probabilità che un pannello non controllato funzioni per tutto il tempo garantito.* [`Risp:` $\hat{\theta} = 0.36$; $IC(\theta) = (0.1832, 0.5368)$]

[3] Data fittizia

[4] Tratto da un tema d'esame di MPSPS del 14/07/99

Esercizio 4.3.8 (Tema d'esame 4.2 del 13/02/06, AA 04/05) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione di densità discreta

$$f(x, \theta) = \frac{(\ln \theta)^x}{\theta x!} \mathbf{1}_{\{0,1,2,\ldots\}}(x), \;\; \theta > 1$$

dove θ è un parametro incognito ed "ln" indica il logaritmo in base naturale.

1. Determinate lo stimatore $\hat{\theta}$ di θ usando il metodo di massima verosimiglianza.
2. Calcolate l'informazione di Fisher $I(\theta)$ del modello statistico $\{f(x, \theta), \theta > 1\}$ e determinate lo stimatore di massima verosimiglianza di $I(\theta)$.
3. Qual è la funzione di ripartizione asintotica di $\hat{\theta}$? (Fornite non solo la forma della fdr approssimata, ma determinatene anche i parametri esplicitamente).
4. Costruite un intervallo di confidenza a due code asintotico per θ di livello γ usando i punti 1. e 3. precedenti. Quindi determinate numericamente l'intervallo quando $\gamma = 0.95, n = 100$ e $\sum_{j=1}^{100} x_j = 150.0$.
5. Siano $n = 100$ e $\sum_{j=1}^{100} x_j = 150.0$. Verificate asintoticamente l'ipotesi nulla $H_0 : \theta = 4$ contro l'alternativa $H_1 : \theta \neq 4$, a un livello $\alpha = 5\%$.

Esercizio 4.3.9 Sia $X_1, \ldots, X_n$ un campione casuale estratto da una popolazione di Poisson di parametro θ incognito. Per verificare $H_0 : \theta = 3$ versus $H_1 : \theta = 1$, usiamo la regione critica $\{(x_1, \ldots, x_n) : \; \bar{x} \leq 1.4\}$.

1. Determinate significatività e potenza del test per $n = 1$, $n = 4$ e $n = 50$.

Esercizio 4.3.10 (Tema d'esame 2.2 del 14/07/05) Nella roulette francese regolare (primi 18 numeri pari neri, primi 18 numeri dispari rossi e lo zero verde), la probabilità θ di vincere puntando sul rosso è 18/37. Sospettiamo che la roulette dell'amico con cui giochiamo sempre sotto Natale sia truccata. Per questo motivo, una sera che si giocava, abbiamo contato il numero di "giochi" necessari perché uscisse rosso la prima volta e abbiamo annotato questo valore chiamandolo x_1. Poi abbiamo contato quante altre partite, dopo le prime x_1, abbiamo dovuto aspettare per vedere rosso per la seconda volta e abbiamo annotato questo numero x_2 e così via. Quella sera è uscito rosso n volte: la prima volta alla x_1-esima partita, la seconda volta alla partita numero $x_1 + x_2$, $\ldots$, l'n-esima volta alla partita numero $x_1 + \ldots + x_n$. Abbiamo quindi i dati di un campione casuale di n osservazioni $X_1, \ldots, X_n$ provenienti da una popolazione *geometrica di parametro* θ con $\theta \in (0, 1)$.

1. Determinate gli stimatori di massima verosimiglianza del parametro θ, della media e della varianza della densità geometrica (di parametro θ).

2. $\ldots$ 3. $\ldots$

Quella sera abbiamo registrato i seguenti dati:

$$(x_1, \ldots, x_{36}) = (1, 4, 1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 4, 1, 1, 2, 3, 1, 1, 3, 2, 4, 1, 1, 2, 5, 2, 1, 6, 1, 1, 6, 3, 4)$$

4. Sulla base di questi dati, stimate il numero medio di giocate necessarie per ottenere rosso.
5. Verificate l'ipotesi nulla $H_0 : \theta = 18/37$ contro l'alternativa $H_0 : \theta \neq 18/37$, usando un test per grandi campioni.

4.4 Miscellanea

Esercizio 4.4.1 (Tema d'esame 3.1 del 12/09/06 AA 05/06) Abbiamo due campioni casuali indipendenti $X_1, \ldots, X_m$ da una popolazione gamma[5] $\Gamma(\alpha, \theta)$ e $Y_1, \ldots, Y_n$ da una popolazione gamma $\Gamma(\beta, \theta)$, con $\theta > 0$ parametro incognito e α, β parametri noti e diversi. Usiamo come stimatore di θ una statistica della forma $T = c \left(\sum_{j=1}^{m} X_j + \sum_{j=1}^{n} Y_j \right)$, dove c è un opportuno numero strettamente positivo.

1. Determinate la distribuzione di T in funzione di c.

2. Determinate c tale che T sia stimatore non distorto di θ.

Siano ora $m = 20$, $n = 10$, $\alpha = 0.1$, $\beta = 0.3$ e sia c come trovato al punto precedente; se non siete stati in grado di trovarlo, scegliete voi un valore ragionevole per c (strettamente positivo) in modo da poter proseguire.

3. Usando T come statistica test, costruite una regione critica di ampiezza $\alpha = 5\%$ per verificare l'ipotesi nulla $H_0 : \theta \leq 2.5$ contro l'alternativa $H_1 : \theta > 2.5$. Se $T = 6.36$ cosa deciderete con il vostro test di livello 5%?

4. Calcolate la potenza del test del punto 3. in $\theta = 4.05$, oppure indicate un intervallo in cui tale potenza cade.

Esercizio 4.4.2 (Tema d'esame 5.2 del 08/03/06, AA 04/05) Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità

$$f(x, \beta) = \frac{1}{2\beta} \exp\left(-\frac{|x|}{\beta} \right), \qquad \beta > 0$$

1. Determinate densità e media di $Y_i = |X_i|$, $i = 1, \ldots .n$.

2. Determinate lo stimatore $\hat{\beta}$ di massima verosimiglianza di β e verificate che è non distorto.

3. Determinate k tale che $P_\beta \left(\frac{\hat{\beta}}{\beta} \geq k \right) = \gamma$, in funzione del numero di osservazioni n e del quantile di ordine $1 - \gamma$ della funzione di ripartizione chiquadrato con $2n$ gradi di libertà.

4. Determinate un intervallo di confidenza a una coda inferiore per β di livello 90%, in corrispondenza della realizzazione campionaria -1.095, 0.836, -0.256, 1.996, 4.675, 0.971.

5. Proponete un test (sensato) di ampiezza 10% basato sullo stimatore $\hat{\beta}$ per verificare l'ipotesi nulla $H_0 : \beta \geq 4$ contro l'ipotesi alternativa $H_1 : \beta < 4$.

6. Sia $n = 6$. Calcolate un valore approssimato della potenza del test del punto 5. in $\beta = 1$, o, almeno, stabilite un intervallo dove tale potenza cade.

Esercizio 4.4.3 (Tema d'esame 5.2 del 06/03/07 AA 05/06) Una grande azienda tessile italiana produce tappeti. Sul mercato vengono immessi solo i tappeti di prima e seconda scelta, mentre gli scarti sono venduti nell'outlet interno. Un tappeto è classificato di prima scelta se non presenta difetti, è di seconda se presenta uno o due difetti, ed è scarto se presenta più di due

[5] La densità $\Gamma(\alpha, \theta)$ è $\dfrac{x^{\alpha-1} e^{-x/\theta}}{\theta^\alpha \Gamma(\alpha)} \mathbf{1}_{(0,\infty)}(x)$, $\alpha > 0, \theta > 0$.

difetti. È poi noto che il numero di difetti di ogni tappeto può essere modellato mediante una variabile aleatoria discreta X avente densità di Poisson di parametro $\theta > 0$, cioè $X \sim f(x, \theta)$ con

$$f(x,\theta) = \frac{\mathrm{e}^{-\theta}\theta^x}{x!}\mathbf{1}_{\{0,1,2,\dots\}}(x),$$

dove $\theta > 0$ è un parametro incognito. Per fare inferenza su θ e sulle percentuali (attese) di tappeti di prima e seconda scelta è stato analizzato un campione casuale $X_1, \dots, X_n$ del numero di difetti di n tappeti.

1. Determinate lo stimatore di massima verosimiglianza del parametro θ e delle percentuali di tappeti di prima e seconda scelta prodotti dall'azienda e fornitene il valore numerico nel caso che siano stati contati in totale 25 difetti su un campione di 100 tappeti.

2. Discutete le proprietà asintotiche degli stimatori di massima verosimiglianza individuati al punto 1.

3. Chiamiamo p_1 la percentuale (attesa) dei tappeti di prima scelta prodotti dall'azienda. Verificate l'ipotesi nulla $H_0 : p_1 \leq 70\%$ contro l'alternativa $H_1 : p_1 > 70\%$ con un test asintotico di livello $\alpha = 10\%$, sempre nel caso che siano stati contati in totale 25 difetti su un campione di 100 tappeti.

4. Chiamiamo p_2 la percentuale (attesa) dei tappeti di seconda scelta prodotti dall'azienda e $\hat{p}_2$ il suo stimatore di massima verosimiglianza. Calcolate un intervallo di confidenza bilatero asintotico per p_2 (basato su $\hat{p}_2$) di livello approssimativamente pari a 95%, sempre nel caso che siano stati contati in totale 25 difetti su un campione di 100 tappeti.

5. Chiamiamo $\hat{p_1}$ lo stimatore di massima verosimiglianza di p_1. Qual è (approssimativamente) il numero minimo di tappeti da controllare affinché sia maggiore o uguale a 90% la probabilità che l'errore che commettete stimando p_1 con $\widetilde{p_1}$ non superi 0.26?

Capitolo 5

Inferenza non parametrica

5.1 Test di buon adattamento

Esercizio 5.1.1 (Tema d'esame 5.3 del 23/02/09, AA 07/08) Sia U una variabile aleatoria esponenziale di media $\lambda > 0$.

1. Determinate la funzione di ripartizione di $T = U^2$.
2. Determinate λ tale che la media di T sia pari a 128.

Vengono registrati i tempi di guasto, espressi in centinaia di ore, di 6 recipienti a pressione, ottenendo i seguenti risultati:

8.821 8.585 38.938 6.708 65.286 17.808.

3. Determinate la funzione di ripartizione empirica $\hat{F}_6$ associata al campione dei 6 recipienti a pressione.
4. Determinate una stima della probabilità che un recipiente a pressione non si guasti prima di 890 ore.
5. Usate un opportuno test di ipotesi per stabilire se il campione casuale dei tempi di guasto dei 6 recipienti a pressione provenga dalla distribuzione di T ottenuta al punto 1. con λ pari al valore determinato al punto 2.
6. Costruite le bande di confidenza 90% per la fdr dei tempi di guasto e ricavatene un IC per la probabilità che un recipiente a pressione non si guasti prima di 890 ore

Esercizio 5.1.2 (Tema d'esame 4.4 del 15/02/05, AA 03/04) I valori che seguono rappresentano i giorni di sopravvivenza di un campione di 6 topi affetti da cancro e curati con una terapia sperimentale:

29, 700, 1, 335, 15, 160

1. Determinate la funzione di ripartizione empirica $\hat{F}_6$ associata al campione dei 6 topi.
2. Determinate una stima della probabilità che un topo affetto da cancro sottoposto alla terapia viva più di 15 giorni.

Si pensa che la sopravvivenza dei topi malati di cancro e sottoposti alla terapia sperimentale possa essere modellata come una variabile aleatoria X continua che ha densità di Weibull:

$$f_0(x) = \frac{1}{20\sqrt{x}} \mathrm{e}^{-\frac{\sqrt{x}}{10}} \mathbf{1}_{(0,\infty)}(x)$$

3. Usate un opportuno test con il 5% di livello di significatività, per stabilire se i dati forniti sui topi possano provenire dalla densità di Weibull $f_0(x) = \frac{1}{20\sqrt{x}} \mathrm{e}^{-\frac{\sqrt{x}}{10}} \mathbf{1}_{(0,\infty)}(x)$ ipotizzata.

Esercizio 5.1.3 (Tema d'esame 3.3 del 19/09/05, AA 04/05) Sei settimane fa si è cominciato a quotare il titolo azionario `xxx`. Il suo prezzo iniziale Y_0 era di 1 euro, mentre il suo prezzo Y_k alla fine della k-esima settimana è stato

$$1.0304545,\ 0.9417645,\ 0.8693582,\ 0.8780954,\ 1.040811,\ 1.030454$$

È uso descrivere la dinamica dei prezzi azionari in termini dei rapporti $R_k = \frac{Y_k}{Y_{k-1}}$, $k \geq 1$, e un modello molto utilizzato suppone che gli $R_k, k = 1, \ldots$ costituiscano un campione di variabili aleatorie indipendenti e identicamente distribuite con densità "*lognormale*": una variabile aleatoria R è detta *lognormale di parametri* μ *e* σ se $\log R$ è gaussiana di media μ e varianza σ^2.

1. Stimate i parametri μ e σ della distribuzione lognormale dei rapporti fra i prezzi $R_k, k = 1, \ldots, 6$. Fornite una qualche giustificazione statistica delle stime proposte.

2. Determinate la funzione di ripartizione empirica dei logaritmi naturali dei rapporti $R_k, k = 1, \ldots, 6$.

3. Determinate una stima della probabilità che alla fine di una settimana il prezzo cresca rispetto alla settimana precedente.

In realtà, nutriamo qualche dubbio che il modello dei rapporti settimanali dei prezzi azionari `xxx` sia lognormale:

4. Implementate un opportuno test di significatività 5%, per verificare la bontà dell'adattamento del modello lognormale di parametri $\mu = 0.02$ e $\sigma = 0.073$ al campione dei rapporti settimanali dei prezzi R_k.

Esercizio 5.1.4 (Tema d'esame 3.3 del 13/09/07, AA 06/07) Una macchina dovrebbe tagliare del filo metallico in pezzi che hanno una lunghezza che può essere rappresentata da una variabile aleatoria X gaussiana con media 10.5 cm e deviazione standard 0.15 cm. Per verificare il corretto funzionamento del macchinario sono stati scelti a caso 16 pezzi da un lotto numeroso e le misure ottenute, espresse in cm ed ordinate, sono le seguenti:

$$10.1, 10.2, 10.2, 10.3, 10.3, 10.4, 10.4, 10.5, 10.5, 10.5, 10.6, 10.6, 10.7, 10.7, 10.8, 10.9.$$

Sulla base di questi dati, ci si propone di stabilire se c'è evidenza sperimentale che la macchina non funzioni correttamente.

1. Specificate ipotesi nulla e ipotesi alternativa.

2. Ricavate la funzione di ripartizione empirica sulla base del campione osservato.

3. Costruite un opportuno test per verificare le ipotesi specificate al punto 1, fissando un livello di significatività $\alpha = 10\%$. Cosa decidete sulla base del campione osservato?

4. Ricavate un intervallo di confidenza di livello 90% per $F_X(10.5)$.

Esercizio 5.1.5 (Tema d'esame 5.3 del 25/02/08 AA 06/07) Alcuni dei risultati di uno studio statistico sulla distribuzione dei redditi dei lavoratori dipendenti del quartiere `xxx` di Milano, condotto su un campione di 500 persone, sono sintetizzati nella seguente tabella:

A_k	$(0, 11588.0]$	$(11588.0, 13360.0]$	$(13360.0, 15287.0]$	$(15287.0, 18248.0]$	$(18248.0, \infty)$
N_k	175	105	95	65	60

dove A_k indica la classe di reddito annuo netto in euro e N_k il numero di dipendenti con reddito appartenente alla classe A_k. Inoltre, il reddito medio campionario è pari a 13381.1 euro e la varianza campionaria è 16×10^6.

1. 2.

Una variabile aleatoria continua X è detta *lognormale di parametri* μ *e* σ se il suo logaritmo naturale $\ln(X)$ ha densità $\mathcal{N}(\mu, \sigma^2)$.

3. Riportate le classi di reddito A_k in scala logaritmica.
4. Verificate con un opportuno test se il reddito dei lavoratori dipendenti del quartiere `xxx` di Milano può essere modellato con una variabile aleatoria X lognormale di parametri $\mu = 9.5$ e $\sigma = 0.11$.

Esercizio 5.1.6 (Tema d'esame 5.4 del 07/03/05) [1]
Il numero π scritto in forma decimale contiene nelle prime 10002 posizioni dopo il punto decimale le cifre 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 rispettivamente

$$968,\ 1026,\ 1021,\ 974,\ 1014,\ 1046,\ 1021,\ 970,\ 948,\ 1014$$

volte.

1. Sulla base di questi dati, ritenete che nella rappresentazione decimale di π, le cifre $0, 1, \dots, 9$ dopo il punto decimale siano uniformemente distribuite? Per rispondere alla domanda usate un opportuno test e scegliete come livello di significatività del test $\alpha = 5\%$.
2. Determinate in modo approssimato il p-value dei dati del test, o piuttosto indicate un intervallo dove tale p-value cade.

Esercizio 5.1.7 Un provider per il collegamento internet afferma che i suoi abbonati riescono a connettersi immediatamente nel 50% dei casi, con un'attesa maggiore di 0 minuti ma inferiore ad 1 minuto nel 40% dei casi e con un'attesa maggiore di 1 minuto nel 10% dei casi.

Un utente decide di verificare l'affermazione del provider, annota i tempi di attesa da lui sperimentati in 70 tentativi di collegamento e ottiene i seguenti risultati:

	linea libera	attesa ≤ 1 min.	attesa > 1 min.	Totale
Numero di tentativi	37	22	11	70

Sottoponete a test, al livello dell'1%, l'affermazione H_0: "*il provider ha ragione*". Determinate il valore del p-value dei dati.

[`Risp:` $Q_3 \simeq 3.6857$, $\chi^2_2 = \mathcal{E}(2)$. `Quindi` $\chi^2_2(0.99) \simeq 9.21$, p-`value`$=0.158365$, ...]

Esercizio 5.1.8 [2] Verificate se il numero quotidiano di interruzioni di corrente elettrica da maggio a settembre in una città italiana ha distribuzione di Poisson, basandovi sulle seguenti rilevazioni:

[1] Estratto da Bickel, P. J., and Doksum, K. A. (1977), Mathematical statistics: Basic ideas and selected topics, Holden-Day Inc, (San Francisco)

[2] cfr. Esempio 2.14 della dispensa di inferenza non parametrica

# interruzioni=	0	1	2	3	4	5	6	7	8	9	≥ 10
Numero di giorni=	0	5	22	23	32	22	19	13	6	4	4

Esercizio 5.1.9 (Tema d'esame 4.3 del 13/02/06, AA 04/05) Avete raccolto 136 misurazioni ripetute di una grandezza. Ma avete salvato solo le seguenti informazioni: esattamente 71 misurazioni sono minori o uguali della mediana della funzione di ripartizione (fdr) gaussiana di media 20 e varianza 16 –da questo momento indicata come fdr $\mathcal{N}(20, 16)$– e di queste 32 sono minori o uguali del quantile di ordine 0.25 della fdr $\mathcal{N}(20, 16)$; poi, 31 misurazioni sono maggiori del quantile di ordine 0.75 della fdr $\mathcal{N}(20, 16)$. Infine, 34 misurazioni hanno valore compreso fra la mediana e il quantile di ordine 0.75 della fdr $\mathcal{N}(20, 16)$.

1. Esprimete la fdr $\mathcal{N}(\mu, \sigma^2)$ in termini della fdr $\mathcal{N}(0, 1)$; esprimete il quantile di ordine a della fdr $\mathcal{N}(\mu, \sigma^2)$ in termini del quantile di ordine a della fdr $\mathcal{N}(0, 1)$.

2. Ritenete che le vostre misurazioni provengano da un modello $\mathcal{N}(20, 16)$? Per rispondere alla domanda costruite un opportuno test di livello $\alpha = 5\%$.

3. Calcolate un valore approssimato del p-value o, almeno, stabilite un intervallo dove tale p-value cade.

Esercizio 5.1.10 (MPSPS 21/06/01) I pezzi meccanici prodotti da una certa linea di produzione devono avere una lunghezza nominale di 20cm; un pezzo è classificato come 1^a *scelta* se la sua lunghezza reale -chiamiamola L- appartiene all'intervallo [19.705,20.295], è classificato come 2^a *scelta* se $L \in (19.6, 19.705) \cup (20.295, 20.4)$; infine, il pezzo è classificato come *scarto* se $L \notin (19.6, 20.4)$. L'azienda dichiara che le lunghezze reali dei pezzi prodotti sono distribuite normalmente attorno a una media di 20cm con varianza 0.018(cm^2).

1. Nell'ipotesi che l'affermazione dell'azienda sia vera, quali sono le percentuali di pezzi prodotti di 1^a, 2^a scelta e del tipo scarti? [`Risp:` $\theta_{01} = 97.21\%$, $\theta_{02} = 2.52\%$, $\theta_{03} = 0.27\%$]

Per convincere un acquirente della veridicità della sua affermazione, l'azienda fornisce i seguenti dati: su un lotto di 2000 pezzi controllati, 1938 sono di 1^a scelta, 52 di 2^a e 10 sono stati scartati.

2. Nei panni dell'acquirente, sulla base dei precedenti dati, voi cosa concludereste? [`Risp: impostiamo il test` χ^2 `per` $H_0 : \theta_i = \theta_{0i} \forall i = 1, 2, 3$. $Q_3 \simeq 3.9891$ `e` p`-value` $\simeq 0.1361$`, quindi...`]

Esercizio 5.1.11 (Tema d'esame 1.3 del 01/07/08, AA 07/08) Una delle densità tradizionalmente usate per modellare la distribuzione dei redditi è la densità di Pareto di parametri α, β data da

$$f(x; \alpha, \beta) = \begin{cases} \frac{\alpha\beta^\alpha}{x^{\alpha+1}} & \text{se } x > \beta \\ 0 & \text{altrove} \end{cases}, \qquad \alpha > 2, \beta > 0.$$

Nel modello di Pareto di parametri α, β con $\alpha > 2$ e $\beta > 0$, la media e il momento secondo sono

$$\mathrm{E}(X) = \frac{\alpha\beta}{(\alpha - 1)}, \qquad \mathrm{E}(X^2) = \frac{\alpha\beta^2}{\alpha - 2}.$$

Per verificare se un modello di Pareto ben si adatta ai redditi dei lavoratori dipendenti del quartiere `OQ` di Milano, abbiamo campionato i redditi di 500 lavoratori dipendenti che abitano nel quartiere e i dati ottenuti (espressi in migliaia di euro) sono sintetizzati nella seguente tabella:

A_k	$(0, 11.6]$	$(11.6, 13.4]$	$(13.4, 15.3]$	$(15.3, 18.3]$	$(18.3, 22.2]$	$(22.2, \infty)$
N_k	175	105	95	65	60	0

,

dove A_k indica la classe di reddito annuo netto, espresso in migliaia di euro, e N_k il numero di dipendenti con reddito appartenente alla classe A_k. Inoltre, il reddito medio campionario (espresso in migliaia di euro) e il momento secondo campionario dei dati raggruppati della precedente tabella valgono rispettivamente 12.00 e 169.64.

1. Determinate gli stimatori dei momenti di α e β sulla base dei valori di reddito medio campionario e momento secondo campionario dei dati raggruppati che vi sono stati forniti.
2. Determinate $P(x_1 < X \leq x_2)$ quando X ha densità di Pareto $f(x; \alpha, \beta)$.
3. Valutate con un opportuno test la bontà di adattamento del modello di Pareto ai dati sui redditi dei lavoratori dipendenti del quartiere `OQ` di Milano. (*Se non siete riusciti a risolvere il punto 1., scegliete voi dei valori per i parametri α e β ed eseguite un opportuno test.*)

Esercizio 5.1.12 [3] Una densità esponenziale si adatta ai seguenti dati raggruppati?

classi	N_i
$(0, 3]$	40
$(3, 4]$	25
$(4, 7]$	20
$(7, 10]$	15

Esercizio 5.1.13 Abbiamo misurato i diametri di 200 punte da trapano simili e abbiamo aggregato i dati (espressi in mm) nella seguente tabella delle numerosità:

classi	$(-\infty, 3.92)$	$[3.92, 4.34]$	$(4.34, 4.45]$	$(4.45, 4.56]$	$(4.56, 4.68]$	$(4.68, 5.06]$	$(5.06, \infty)$
N_i	0	44	35	41	39	41	0

Costruite un opportuno test per verificare se i diametri delle punte da trapano siano va gaussiane.

Esercizio 5.1.14 (Tema d'esame 3.4 del 16/09/04, AA 03/04) Il tempo di esecuzione del programma *xxx* sul calcolatore *yyy* è compreso fra 60 e 120 minuti primi. Vogliamo verificare se tale tempo possa essere modellato come una variabile aleatoria X continua di densità

$$f(x) = \frac{1}{1800}(x - 60)\mathbf{1}_{(60,120)}(x)$$

A tale fine, su ciascuno di 75 calcolatori tutti del tipo *yyy* e che lavorano indipendentemente uno dall'altro, viene lanciato il programma *xxx* e si registrano i tempi di esecuzione. Seguono i risultati sperimentali ottenuti:

intervalli di tempo A_k	# di programmi con tempo di esecuzione $\in A_k$
$A_1 = (60, 75)$	6
$A_2 = [75, 80)$	8
$A_3 = [80, 95)$	20
$A_4 = [95, 105)$	17
$A_5 = [105, 110)$	10
$A_6 = [110, 120)$	14

Sulla base dei dati raccolti, verificate con un opportuno test se la densità f fornisce un buon modello probabilistico per il tempo di esecuzione del programma *xxx* su un calcolatore del tipo *yyy*.

[3] cfr. Esempio 2.13 della dispensa di inferenza non parametrica

5.2 Test di omogeneità

Esercizio 5.2.1 (Esercizio 13 pag. 504 in Ross 2003) [4] Per uno studio sulla sicurezza stradale vengono selezionate 14 città di dimensioni molto simili. Un campione casuale di 8 di esse viene scelto per una campagna giornalistica di informazione sulla sicurezza stradale della durata di un mese. Alla fine di tale periodo, per un altro mese, si registra il numero di incidenti stradali in ciascuna delle 14 città. I dati osservati sono questi:

Gruppo di trattamento	19	31	39	45	47	66	74	81
Gruppo di controllo	28	36	44	49	52	60		

Calcolate il p-value nel verificare l'ipotesi che gli articoli non abbiano sortito alcun effetto apprezzabile.

Esercizio 5.2.2 (Tema d'esame 5.5 del 07/03/05, AA 03/04) Si sono registrati i minuti di funzionamento prima di rovinarsi di due tipi di isolanti elettrici A e B sottoposti a una forte differenza di potenziale ottenendo i seguenti risultati:

Tipo A:	162	88.5	122.3	125	132	66	211.9		
Tipo B:	34.6	54	116.4	49	77.3	121.3	127.8	120.2	49.8

Verificate l'ipotesi che i due campioni casuali provengano dalla stessa fdr contro l'alternativa che l'isolante elettrico di tipo B smetta di funzionare prima di quello di tipo A.
Scegliete come livello di significatività del test 2.5% e supponete che i dati siano tutti generati da modelli continui. La vostra risposta cambia se scegliete un livello di α superiore a 2.5%?

Esercizio 5.2.3 (Tema d'esame 4.3 del 29/09/06, AA 05/06) Un'azienda possiede due stabilimenti A e B per la produzione di transistor. Per verificare che i prodotti dei due stabilimenti siano equivalenti, si esaminano due campioni indipendenti di 8 e 10 transistor. Denotiamo con $X_1, \dots, X_8$ i tempi di vita degli 8 transistor campionati dallo stabilimento A e con $Y_1, \dots, Y_{10}$ i tempi di vita dei 10 provenienti dallo stabilimento B.

La classifica dei 18 transitor esaminati, ordinati dal meno al più durevole è la seguente

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	b	a	b	b	b	b	a	b	b	b	a	b	a	a	b	a	a	a

(a=stabilimento A, b=stabilimento B).

1. Impostate un test di verifica dell'ipotesi nulla H_0 :"*i prodotti dei due stabilimenti sono equivalenti*" contro l'alternativa H_1 :"*Il transistor dello stabilimento A dura più di quello dello stabilimento B*", di livello $\alpha = 5\%$. Sulla base dei dati cosa decidete?

In realtà, noi possiamo supporre che i transistor A e B abbiano tempi di vita esponenziali di medie θ_A, θ_B, entrambe incognite. Inoltre, conosciamo anche i valori delle medie campionarie: $\overline{X}_8 = 25.1$ e $\overline{Y}_{10} = 22.0$.

2. Determinate le distribuzioni di $\dfrac{16\overline{X}_8}{\theta_A}$, $\dfrac{20\overline{Y}_{10}}{\theta_B}$ e del rapporto $\overline{X}_8 / \overline{Y}_{10}$ nell'ipotesi $\theta_A = \theta_B$.

3. Traducete le ipotesi H_0 :"*i prodotti dei due stabilimenti sono equivalenti*" contro H_1 :"*Il transistor dello stabilimento A dura più di quello dello stabilimento B*" in ipotesi sui parametri θ_A, θ_B. Quindi, costruite un nuovo test (questa volta parametrico) di livello $\alpha = 5\%$ basato sulla statistica test $\overline{X}_8 / \overline{Y}_{10}$. Quale decisione prendete se $\overline{X}_8 = 25.1$ e $\overline{Y}_{10} = 22.0$?

[4] Ross, Sheldon, M. (2003) Probabilità e statistica per l'ingegneria e le scienze Apogeo.

5.2.1 Test di omogeneità per dati gaussiani

Esercizio 5.2.4 [5] Un gruppo di 10 agenti di commercio segue un corso commerciale. Di ciascuno di loro sono misurate le vendite medie mensili prima e dopo il corso. I dati sono i seguenti

membro del corso	1	2	3	4	5	6	7	8	9	10
vendite prima	20.0	18.2	22.7	25.0	19.0	24.1	21.9	24.3	25.2	34.2
vendite dopo	30.3	17.3	33.9	37	20.6	35.9	20.1	29.7	24.7	26.8

Sulla base di questi numeri, cosa pensate circa l'utilità del corso? Verificate l'ipotesi nulla che il corso non sia utile e calcolate il p-value. Risolvete il problema nel caso di dati bivariati gaussiani.

Esercizio 5.2.5 (Tema d'esame 4.4 del 30/06/05, AA 04/05) È stato condotto un esperimento per verificare se l'attività fisica diurna modifica il livello dell'ormone della crescita prodotto nelle ore di sonno notturno. L'esperimento ha coinvolto sei maschi sani e adulti, ai quali, durante le ore notturne di sonno, sono stati prelevati due campioni di sangue per mezzo di un catetere in vena. Il primo campione di sangue è stato prelevato dopo una giornata in assenza di attività fisica e il secondo dopo una giornata di faticosa attività fisica. Per ogni soggetto e per ogni campione di sangue è stato misurato (in mg/mL) il livello dell'ormone della crescita.

1. Se in 4 soggetti su 6 il livello di ormone della crescita nel primo campione di sangue è superiore al livello nel secondo, ritenete che l'attività fisica riduca il livello dell'ormone della crescita? Impostate ed eseguite un test d'ipotesi di livello 5%. Indicate le eventuali ipotesi necessarie alla conduzione del test.

Chiamiamo ora (x_j, y_j) la coppia che fornisce il livello dell'ormone della crescita dopo una giornata senza attività fisica e dopo una giornata di intensa attività fisica del soggetto j-esimo e assumiamo che il campione dei 6 dati accoppiati (x_j, y_j), $j = 1, ..., 6$, provenga da una popolazione bivariata gaussiana. Inoltre, delle 6 coppie di dati conosciamo le seguenti sintesi:

$$\sum_{i=1}^{6} x_i = 116.1, \quad \sum_{i=1}^{6} y_i = 101.9, \quad \sum_{i=1}^{6} x_i y_i = 2130.82, \quad \sum_{i=1}^{6} x_i^2 = 2992.81, \quad \sum_{i=1}^{6} y_i^2 = 1782.99 \tag{5.1}$$

2. Calcolate media e varianza campionarie del campione delle differenze $d_j = x_j - y_j$, $j = 1, \dots, 6$.

3. Eseguite un secondo test di ipotesi che usi i dati forniti in (5.1) sempre al fine di stabilire se l'attività fisica riduca il livello dell'ormone della crescita. Cosa concludete al livello 5%?

Esercizio 5.2.6 (Tema d'esame 3.3 del 16/09/04, AA 03/04) Due macchine A e B producono filo di rame, il cui diametro si è stabilito abbia un certo valore (assegnato) μ_0. Per controllare la qualità del processo vengono ispezionati 10 fili prodotti dalla macchina A e 15 prodotti dalla macchina B e per ogni filo viene registrato l'errore nella lunghezza del diametro (errore=valore misurato $-\mu_0$). Dalle misurazioni effettuate si ottiene: la somma dei quadrati degli errori nella lunghezza dei diametri è 0.017 mm^2 per i 10 fili prodotti dalla macchina A e 0.095 mm^2 per i 15 prodotti dalla macchina B.

Assumendo che gli errori nella lunghezza del diametro siano variabili aleatorie indipendenti, gaussiane e a media nota e uguale a zero:

1. fornite una stima puntuale della varianza σ_A^2 dell'errore nella lunghezza del diametro dei fili prodotti da A e della varianza σ_B^2 dell'errore nella lunghezza del diametro dei fili prodotti da B;

[5] Estratto da Pestman, Wiebe R. (1998), Mathematical Statistics de Gruyter (Example 2 pag. 229).

2. impostando un opportuno test di verifica di ipotesi di ampiezza $\alpha = 10\%$, potete ritenere che le due macchine abbiano la stessa precisione?

Esercizio 5.2.7 [6] I pesi in libbre dei campioni di neonati appartenenti a due contee x e y adiacenti nella Western Pennsylvania hanno fornito i seguenti valori:

$$m = 44,\ \bar{X}_m = 7.2.,\ S^2_X = 4.9, \qquad n = 53,\ \bar{Y}_n = 6.9,\ S^2_Y = 5.2$$

In quanto segue assumete che i dati siano gaussiani e con varianza uguale

1. Proponete uno stimatore non distorto della comune varianza e calcolatene il valore in corrispondenza dei dati forniti. [`Risp:` S^2_p`;` $s^2_p = 5.064211$]

2. Costruite un intervallo di confidenza bilatero della comune varianza di significatività $\gamma = 90\%$ [`Risp:` $(4.088, 6.651)$]

3. Costruite un test per verificare l'ipotesi che il peso medio dei neonati delle due contee sia lo stesso contro l'alternativa che i neonati della contea x pesino in media più di quelli della contea y. Quanto vale il p-value dei dati? [`Risp:` p`-value` $\simeq 25.7\%$]

Esercizio 5.2.8 (Tema d'esame 2.4 del 18/07/06 AA 05/06) La solita azienda OISAC ha proposto una pila di ultima generazione `yyy` più durevole del tipo `xxx` di vecchia generazione. Per confrontare la durata delle due pile `xxx` e `yyy` abbiamo a disposizione i seguenti due campioni indipendenti di dati continui (espressi in ore):

$$x_i: \quad 7.26, \quad 2.04, \quad 0.94, \quad 1.76, \quad 11.08, \quad 0.60, \quad 9.04$$

(=durate di 7 pile `xxx`)

$$y_i: \quad 0.80, \quad 1.71, \quad 4.10, \quad 6.10, \quad 7.89, \quad 24.10$$

(=durate di 6 pile `yyy`).

1. Proponete a OISAC un test per verificare se le pile `yyy` durano più di quelle `xxx`, che funzioni anche quando non si ha nessun'altra informazione sul modello statistico generatore dei dati. Sulla base dei dati forniti che decisione prendete a livello $\alpha = 10\%$?

In realtà, successivamente, la nostra conoscenza sul modello statistico generatore dei dati è aumentata. Infatti, ora riteniamo plausibile modellare le durate delle pile, sia di nuova che di vecchia generazione, come variabili aleatorie gaussiane.

2. Avendo quest'ulteriore informazione, la risposta alla domanda 1. cambia o no? Argomentate la risposta impostando una opportuna metodologia statistica.

   [Per risparmiare tempo, nell'eventualità vi occorrano, vi abbiamo già calcolato qualche statistica per i due campioni: $\sum_{j=1}^7 x_j = 32.72$, $\sum_{j=1}^6 y_j = 44.7$ $\sum_{j=1}^7 x_j^2 = 265.7$, $\sum_{j=1}^6 y_j^2 = 700.65$].

Esercizio 5.2.9 (Tema d'esame 2.3 del 14/07/05, AA 04/05) In un esperimento di laboratorio viene misurata la temperatura ritenuta "più piacevole" da 8 donne e 10 uomini. Dalla classifica dei 18 esaminati (fatta in base alla temperatura per ciascuno di loro più piacevole) scopriamo che la persona meno freddolosa è un uomo (nella classifica occupa la prima posizione) e la più freddolosa è una donna (in ultima posizione c'è appunto una donna). Andando nel dettaglio, la tabella delle posizioni occupate da uomini e donne nella classifica è la seguente

[6] Esercizio 37 pag. 334 in Ross 2003

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	u	d	u	u	u	u	d	u	u	u	d	u	d	d	u	d	d	d

(u=uomo, d=donna).

1. Impostate un test di verifica dell'ipotesi nulla H_0 :"*la temperatura più piacevole è la stessa per uomini e donne*" contro l'alternativa H_1 :"*Alle donne piace una temperatura più alta che agli uomini*", di livello 5%.

In realtà, i ricercatori del laboratorio ci hanno fornito anche i seguenti dati sui valori di temperatura (in gradi Celsius):

$$\sum_{j=1}^{8} x_j = 198.8, \qquad \sum_{j=1}^{8} x_j^2 = 4949.08, \qquad \sum_{j=1}^{10} y_j = 239.3, \qquad \sum_{j=1}^{10} y_j^2 = 5735.65$$

dove x_j è la temperatura più piacevole per la donna j-esima e y_j quella più piacevole per l'uomo j-esimo e ci hanno detto che i due campioni delle temperature possono essere assunti entrambi gaussiani.

2. Verificate con un test di significatività 5% se i due campioni provengano da popolazioni con la stessa varianza.

3. Costruite un secondo test di ipotesi di livello 10%, che usi le informazioni fornite dai ricercatori sui valori di temperature e le assunzioni di gaussianità, sempre per verificare H_0 :"*la temperatura più piacevole è la stessa per uomini e donne*" contro l'alternativa H_1 :"*Alle donne piace una temperatura più alta che agli uomini*".

5.2.2 Miscellanea su test di omogeneità

Esercizio 5.2.10 (Tema d'esame 2.5 del 15/07/04, AA 03/04) Un laboratorio informatico ha elaborato un nuovo protocollo P_{new} per la trasmissione di dati. Per confrontare P_{new} con il vecchio protocollo P_{old} si procede a inviare un certo file per 7 volte da un server ad un altro usando il protocollo P_{new} e 6 volte usando il protocollo P_{old} e si misurano i tempi (in secondi) intercorrenti tra l'invio e la ricezione. I risultati ottenuti per P_{new} sono:

$$x_i: \quad 1.49, \quad 1.50, \quad 1.96, \quad 2.33, \quad 1.45, \quad 1.71, \quad 2.83$$

e quelli per P_{old} sono

$$y_i: \quad 1.85, \quad 3.47, \quad 4.44, \quad 1.75, \quad 2.16, \quad 3.93$$

Vi si chiede ora di usare questi dati al fine di stabilire se il nuovo protocollo P_{new} sia migliore del vecchio P_{old} dove, in modo naturale, un protocollo è ritenuto "migliore" di un altro se trasferisce i dati in meno tempo. Quindi:

- Costruite una opportuna strategia statistica (che usi i dati precedenti) per affrontare il problema ipotetico del confronto fra i protocolli P_{new} e P_{old}. In particolare abbiate cura di specificare a) le ipotesi statistiche da verificare, b) le regioni critiche e, se necessario, c) le condizioni che il modello statistico generatore dei dati deve soddisfare perché la vostra procedura trovi ragionevoli giustificazioni nella teoria dei test (parametrici e/o non parametrici) vista durante il corso.

Esercizio 5.2.11 (Tema d'esame 4.4 del 13/02/06, AA 04/05) Abbiamo estratto a caso un campione di sette coppie sposate e abbiamo chiesto a ciascun marito e a ciascuna moglie *a*) quanto ha speso in euro per la spesa l'ultima volta che vi si è recato da solo (sola) e *b*) di esprimere con un numero compreso fra 0 e 1 la probabilità che l'altra(o) demandi a lui (lei) l'incombenza di fare la spesa. Le risposte sono state le seguenti:

coppia	1	2	3	4	5	6	7
marito	$(25.1, 0.356)$	$(21.5, 0.035)$	$(38.0, 0.506)$	$(64.3, 0.614)$	$(52.0, 0.330)$	$(16.1, 0.171)$	$(26.1, 0.447)$
moglie	$(16.0, 0.701)$	$(42.2, 0.903)$	$(56.5, 0.315)$	$(41.1, 0.220)$	$(19.0, 0.998)$	$(26.2, 0.355)$	$(24.5, 0.622)$

(dove, per esempio, per la prima coppia, alla riga dei mariti il dato accoppiato $(25.1, 0.356)$ indica che l'ultima volta che ha fatto la spesa da solo il marito della prima coppia ha speso 25.1 euro e che lui reputa che la probabilità che la moglie gravi lui dell'incombenza della spesa sia 0.356.[7])

1. Costruite un test di livello $\alpha = 5\%$ per verificare se i mariti siano oculati almeno quanto le mogli quando fanno la spesa contro l'alternativa che le mogli siano più oculate dei mariti, sotto ipotesi di gaussianità delle spese di mogli e mariti.

2. Costruite un test di livello per verificare se sia più frequente che i mariti demandino alle mogli l'onere di fare la spesa (almeno nella percezione delle mogli), contro l'alternativa che sia più frequente che le mogli demandino questa incombenza ai mariti (almeno nella percezione dei mariti). Fornite una stima del p-vale.

5.3 Test di indipendenza

Esercizio 5.3.1 (Tema d'esame 2.4 del 14/07/05, AA 04/05) Recita la guida all'immatricolazione in rete: "*La prova d'ammissione per i corsi di studio delle Facoltà di Ingegneria ha carattere orientativo: gli studenti sono invitati a valutarne seriamente l'esito e l'indice attitudinale che ne consegue.* $\cdots$ *La prova d'ammissione rappresenta la verifica di un'adeguata preparazione iniziale. Se il risultato segnala carenze formative in una o più aree -i cosiddetti debiti formativi- gli studenti dovranno provvedere a colmarle*".

Abbiamo i dati riguardanti un campione di 1366 studenti provenienti dal liceo scientifico, 118 dal liceo classico e 371 dagli istituti tecnici. Hanno raggiunto il punteggio minimo di sufficiente preparazione iniziale 805 studenti del liceo scientifico, 47 del liceo classico e 68 degli istituti tecnici.

1. Sulla base di questi dati, l'adeguatezza della preparazione iniziale è legata al tipo di scuola di provenienza? Rispondete costruendo un opportuno test di ipotesi.

Supponiamo ora che comunque sia andata la prova d'ammissione, tutti gli studenti di quel campione abbiano deciso di iscriversi a un corso di ingegneria.

2. Dopo aver fornito una stima puntuale della percentuale $\theta 100$ di matricole del prossimo AA 2005/2006 provenienti dagli istituti tecnici, determinate un intervallo asintotico di confidenza 90% per θ.

Esercizio 5.3.2 (Tema d'esame 2.4 del 15/07/04, AA 03/04) I dati disponibili sul sito del Ministero dell'Istruzione, dell'Università e della Ricerca rivelano che nell'anno solare 2002, i

[7]Attenzione alla lettura dei dati sulle probabilità di essere demandato a fare la spesa: per ogni coppia avete due numeri compresi fra 0 e 1. Questi due numeri sono valutazioni soggettive di probabilità fatte da due persone diverse (marito e moglie); quindi, per qualche coppia potrebbero sommare a un numero maggiore di 1, oppure potrebbero essere entrambe "piccole", perché per esempio quella coppia usa andare a far la spesa spesso insieme.

laureati e diplomati presso le facoltà di ingegneria del Politecnico di Milano sono stati 3502 di cui 2967 uomini. Fra le donne, 176 erano fuori corso da un anno, 72 da due anni e 61 da tre anni. Invece, fra gli uomini, i fuori corso da un anno erano 733, quelli da due anni 531 e quelli da tre anni 279. Tutti gli altri erano fuori corso da quattro anni o più; nel 2002 non ci sono stati laureati in corso. Verificate sulla base di questi dati se uomini e donne impiegano (più o meno) lo stesso tempo per laurearsi in ingegneria.

5.3.1 Test di indipendenza per dati gaussiani

Esercizio 5.3.3 (Tema d'esame 1.4 del 28/06/06, AA 05/06) Un campione di 1000 laureati della facoltà di `xxx`, selezionato casualmente, ha accettato di sottoporsi a un test di verifica della preparazione scientifica maturata nel corso di studi. Il test è stato gestito da una società esterna all'università e il punteggio ottenuto era espresso in trentesimi da 0 a 30. I dati ottenuti sono stati raggruppati e riportati nella tabella seguente, dove si è proceduto a raggruppare i voti di laurea nelle classi: $[66, 90], [91, 99], [100, 110]$ e i punteggi al test nelle classi $[0, 17], [18, 20], [21, 24], [25, 30]$. (La tabella deve essere completata).

laurea\test	[0,17]	[18,20]	[21,24]	[25,30]	
[66,90]	80	115	157	38	
[91,99]	65	130		65	
[100,110]	20	15	43	22	

1. Cosa concludete sull'affidabilità del voto di laurea come indicatore della preparazione conseguita nel corso di studi? Costruite un opportuno test di ipotesi.

In realtà siamo riusciti a recuperare anche le seguenti statistiche:

$$\sum_{i=1}^{1000} l_i = 92120, \ \sum_{i=1}^{1000} t_i = 20990, \ \sum_{i=1}^{1000} l_i t_i = 1935996, \ \sum_{i=1}^{1000} l_i^2 = 8518217, \ \sum_{i=1}^{1000} t_i^2 = 449233.5 \tag{5.2}$$

dove l_i e t_i rappresentano rispettivamente il voto di laurea e il punteggio totalizzato al test dal laureato i-esimo.

2. Alla luce dei nuovi dati in (5.2), la risposta alla domanda 1. cambia o no? Per rispondere assumete l'ipotesi di gaussianità dei dati accoppiati $(l_i, t_i), i = 1, \dots 1000$.

Esercizio 5.3.4 Abbiamo un campione di 100 dati accoppiati (x_j, y_j), $j = 1, ..., 100$, estratti da una popolazione bivariata gaussiana. Essi forniscono, rispettivamente, la media dei voti al primo e al secondo anno del corso di laurea in `aaa`. Conosciamo le seguenti sintesi:

$$\sum_{i=1}^{100} x_i = 2371.2, \quad \sum_{i=1}^{100} y_i = 2456.9, \quad \sum_{i=1}^{100} x_i y_i = 59601.8, \quad \sum_{i=1}^{100} x_i^2 = 57682.1, \quad \sum_{i=1}^{100} y_i^2 = 62261.8 \tag{5.3}$$

1. Eseguite un test di ipotesi che usi i dati forniti in (5.3) per stabilire se la media dei voti attesa al primo anno sia uguale a quella del secondo contro l'alternativa di una media al secondo maggiore del primo.

2. Eseguite un test di ipotesi che usi i dati forniti in (5.3) per stabilire se le medie al primo e secondo anno sono indipendenti. Cosa concludete al livello 5%?

5.4 Test di casualità di Kendall

Esercizio 5.4.1 (Tema d'esame 3.4 del 19/09/05, AA 04/05) Abbiamo eseguito dieci volte il programma `xxx` sul calcolatore `yyy`. I tempi di esecuzione espressi in minuti e riportati nell'ordine di esecuzione sono i seguenti:

81.8 79.9 91.5 67.3 114.9 82.9 63.6 88.4 68.7 81.2

1. Si può concludere sulla base di questi dati che ci sia una sorta di "effetto memoria" per cui i tempi di esecuzione non sono indipendenti? Per rispondere usate un opportuno test con livello di significatività 10%.

Esercizio 5.4.2 (Tema d'esame 4.5 del 15/02/05, AA 03/04) Quindici misure ripetute eseguite con lo stesso strumento e in modo indipendente hanno dato i seguenti risultati, riportati nell'ordine in cui sono stati ottenuti:

0.30 1.27 -0.25 -1.28 -1.20 -1.74 2.18 0.23 -1.10 1.08 0.69 1.69 1.84 0.97 2.00

Lo sperimentatore sospetta però che ci sia stato un deterioramento dello strumento nel corso dell'esperimento che potrebbe aver distrutto l'indipendenza fra le misure.

1. Verificate l'ipotesi d'indipendenza delle misure al livello del 10%.
2. Determinate in modo approssimato il p-value del test, o piuttosto indicate un intervallo dove tale p-value cade.

Esercizio 5.4.3 (Tema d'esame 5.4 del 08/03/06, AA 04/05) Abbiamo raccolto 11 dati giornalieri sul peso medio delle uova grandi covate da galline vecchie allevate a terra nell'azienda avicola `aaa`. Di seguito sono riportati i pesi medi X_i, $i = 1, \dots, 11$ delle uova (espressi in grammi) nell'ordine del giorno di rilevazione:

65.0, 70.3, 71.2, 63.0, 64.6, 67.0, 62.8, 68.9, 71.3, 60.9, 66.4

1. Impostate un test di ampiezza 5% per verificare se i pesi delle uova in giorni diversi siano indipendenti.
2. Impostate un test di ampiezza 5% sulla varianza del peso medio giornaliero delle uova, specificando ipotesi nulla e alternativa in modo tale che costituisca errore di seconda specie acquistare uova il cui peso abbia varianza minore di 9.0 grammi^2. Assumete l'ipotesi di gaussianità dei dati.

5.5 Miscellanea

Esercizio 5.5.1 (Tema d'esame 4.4 del 29/09/06 AA 05/06) Gli allievi del primo anno di un corso di laurea sono stati suddivisi in tre sezioni A, B, C nelle quali il corso di `PCCP` è tenuto da tre diversi docenti. Un controllo dei verbali delle prove d'esame di `PCCP` degli ultimi due anni ha fornito i dati sul numero di appelli che 1000 allievi hanno provato per superare l'esame. Questi dati sono ripartiti per sezione e riportati nella tabella seguente:

Sezione \ # *appelli provati*	1	2	3	4 o più	
A	202	100	28	20	
B	200	90	30	30	
C	172	78	20	30	

(Per esempio 202 è il numero di studenti della sezione A che sono stati promossi al primo appello provato).

1. Stabilite se i criteri di valutazione dei tre docenti delle tre sezioni siano omogenei con un opportuno test di ipotesi.

2. Usando un opportuno test, stabilite se il numero di appelli necessari per essere promossi sia una variabile aleatoria geometrica.

3. Costruite un intervallo di confidenza approssimativamente 90% della probabilità di non essere promossi prima del terzo appello provato.

Esercizio 5.5.2 (Tema d'esame 5.4 del 06/03/07 AA 05/06) Dobbiamo fare inferenza sui tempi di guasto di un sistema formato da due componenti 1 e 2 connessi in parallelo, cosicché il sistema funziona quando almeno uno dei due componenti funziona.

Per esempio vogliamo stabilire se i due componenti sono indipendenti, se funzionano mediamente lo stesso tempo e, ancora, vorremmo stimare la durata dell'intero sistema.

Per questo, abbiamo acquistato 60 componenti di tipo 1 e 60 di tipo 2 e abbiamo costruito 60 sistemi in parallelo.

Allo scadere della 11-esima ora dall'attivazione, abbiamo controllato quanti sistemi funzionavano ancora venendo a scoprire che: dopo 11 ore dall'attivazione, 16 sistemi funzionavano ancora sebbene il componente 1 fosse guasto, 6 funzionavano ancora sebbene il componente 2 fosse guasto e infine 10 non funzionavano più.

1. Alla luce di queste informazioni, possiamo concludere che i due componenti funzionano in modo indipendente? Rispondete con un opportuno test di ipotesi e determinate un valore approssimato del p-value.

2. Verificate l'ipotesi nulla che il tempo di guasto di un intero sistema abbia densità esponenziale di media 9 ore. Usate i precedenti dati e costruite un opportuno test.

Non avendo fretta, abbiamo deciso di proseguire l'esperimento e abbiamo continuato a tenere attivi fino a rottura solo i sistemi costituiti da due componenti che funzionavano entrambi dopo 11 ore. In questa seconda fase abbiamo ottenuto che esattamente 15 fra questi sistemi hanno smesso di funzionare perché si è rotto per primo il componente 2.

3 Integrando le prime informazioni con questo nuovo dato, verificate a livello $\alpha = 5\%$ se i tempi di guasto dei due componenti 1 e 2 sono omogenei contro l'alternativa che il componente 1 si guasta dopo il 2.

Esercizio 5.5.3 (Tema d'esame 1.3 del 30/06/05, AA 04/05) Si sono registrati i tempi di vita in ore di 30 pile AAA prodotte dall'azienda `bbb`. Questi dati costituiscono un campione casuale $X_1, \dots, X_{30}$ la cui funzione di ripartizione empirica $\hat{F}_{30}$ è data da

x	1	2	4	6	8	11	13	27	29	42
$\hat{F}_{30}(x)$	7/30	12/30	16 /30	20/30	23/30	26/30	27/30	28/30	29/30	1

Più precisamente, X_j rappresenta la durata della j-esima pila.

1. Determinate uno stimatore T_1 della probabilità che una pila AAA prodotta dall'azienda `bbb` duri più di 9 ore.

2. Costruite un intervallo di confidenza asintotico di livello 95% per la probabilità che una pila AAA prodotta da `bbb` duri più di 9 ore.

3. Calcolate la durata media campionaria delle 30 pile AAA prodotte da `bbb`.

In realtà, da precedenti analisi statistiche fatte su pile AAA ma prodotte da altre aziende, sappiamo che la durata X di una pila AAA ha distribuzione esponenziale di media $\theta > 0$. In altri termini, da questo momento i dati assegnati costituiscono la realizzazione di un campione casuale $X_1, \dots, X_{30}$ estratto dalla densità esponenziale di media θ incognita.

4. Determinate lo stimatore di massima verosimiglianza della probabilità che una pila AAA prodotta da `bbb` duri più di 9 ore, assumendo l'ipotesi che il modello sia esponenziale.

Indichiamo ora con $p = p(\theta)$ la probabilità che una pila AAA prodotta da `bbb` duri più di 9 ore. Vogliamo costuire un test di livello α per verificare l'ipotesi nulla $H_0 : p = 0.32$ contro l'alternativa $H_1 : p = 0.16$.

5. Dopo aver controllato che il problema di verifica di ipotesi $H_0 : p = 0.32$ contro $H_1 : p = 0.16$ è equivalente al seguente: $H_0 : \theta = 7.8986$ contro $H_1 : \theta = 4.9111$, proponete un test per verificare l'ipotesi nulla $H_0 : p = 0.32$ contro l'alternativa $H_1 : p = 0.16$. Se $\alpha = 1\%$, cosa decidete sulla base del test d'ipotesi costruito?

Esercizio 5.5.4 (Tema d'esame 5.3 del 08/03/06) Nella Repubblica di `AILATI` oggi è giorno di elezioni. I partiti in lizza sono tre: *Beige*, *Crema* ed *Ecrù*. Vince chi prende più voti. Fuori da un seggio si sta svolgendo uno strano exit-poll: all'uscita a un campione di votanti uomini scelti a caso si chiede per chi hanno votato e vengono pesati. Le informazioni raccolte sono le seguenti: 246 intervistati hanno votato *Beige*, 200 *Crema* e i rimanenti *Ecrù*. Invece, per quanto riguarda il peso: 120 pesano meno di 72.0 Kg, 133 pesano più 90.0 Kg, 266 hanno poi un peso compreso fra 72.0 Kg e 79.0 Kg e 301 fra 79.0 Kg e 90.0 Kg. Inoltre, fra gli intervistati più magri di 72.0 Kg, 37 hanno votato *Beige* e 21 *Crema*; invece fra quelli di peso compreso fra 72.0 Kg e 79.0 Kg, 73 hanno votato *Beige* e 129 *Ecrù*. Infine, fra i votanti più grossi di 90.0 Kg, 32 hanno votato *Crema* e 63 *Ecrù*.

1. Stabilite se la caratteristica peso influenzi le preferenze politiche degli elettori.
2. Costruite un intervallo di confidenza bilatero asintotico di livello $\gamma = 0.95$ della percentuale di elettori del partito *Ecrù* più grassi di 79.0 Kg.
3. Verificate ad un livello $\alpha = 5\%$ l'ipotesi che la metà dei maschi adulti della Repubblica di `AILATI` siano più grassi di 79.0 Kg contro l'alternativa che i più grassi di 79.0 Kg siano più della metà.

Esercizio 5.5.5 (tema d'esame 2.4 del 10/07/07 AA 06/07) Si vuole stabilire l'efficacia di un nuovo farmaco per aumentare il livello di sideremia (cioè della concentrazione di ferro nel plasma sanguigno). Per questo motivo, 75 donne iposideremiche sono state sottoposte a una cura di 30 giorni col nuovo farmaco. Quindi, sono stati misurati i livelli di sideremia, in microgrammi per decilitro (mcg/dl), prima e dopo la cura e di questi valori si sono calcolate le seguenti utili statistiche:

$$\sum_{j=1}^{75} x_j = 4394.294, \ \sum_{j=1}^{75} x_j^2 = 303548, \ \sum_{j=1}^{75} y_j = 2959.761, \ \sum_{j=1}^{75} y_j^2 = 133417,$$
$$\sum_{j=1}^{75} x_j y_j = 187947.7, \ \sum_{j=1}^{75} (x_j - y_j - (\bar{x} - \bar{y}))^2 = 33631.13.$$

dove y_j è il livello della sideremia prima della cura e x_j quello dopo la cura. Inoltre, si può supporre che i dati accoppiati $\{(x_j, y_j), j = 1, \dots, 75\}$ siano gaussiani.

2. Usando i dati forniti, verificate al livello 5% l'ipotesi nulla che il farmaco non sia efficace contro l'alternativa che invece lo sia.

3. Eseguite un test d'ipotesi di livello 1% per stabilire se c'è indipendenza tra i livelli di sideremia prima e dopo la cura, sulla base dei dati forniti.

Esercizio 5.5.6 (Tema d'esame 3.3 del 05/09/08, AA 07/08) Abbiamo raccolto dei dati su 200 allievi del corso di laurea in `xxx` che hanno superato l'esame di `LIN` e `SIN`, entrambi obbligatori, e li abbiamo sintetizzati come segue.

Tabella 5.1: # appelli sostenuti per superare `LIN` e `SIN`

`LIN\ SIN`	1	2	3 o più	
1	60	10	10	
2	30	15	15	
3 o più	10	20	30	

In Tabella 5.1[8] troviamo il numero di appelli sostenuti dagli allievi per superare gli esami `LIN` e `SIN`; per esempio, 20 è il numero di allievi che in `SIN` sono stati promossi al secondo appello sostenuto e che in `LIN` hanno dovuto faticare almeno 3 appelli. Poi, in (5.4) abbiamo le statistiche sui voti finali registrati, espressi in trentesimi:

$$\sum_{j=1}^{200} x_j = 4600, \sum_{j=1}^{200} x_j^2 = 107600, \sum_{j=1}^{200} y_j = 4900, \sum_{j=1}^{200} y_j^2 = 120850, \sum_{j=1}^{200} x_j y_j = 113704, \tag{5.4}$$

dove $\{x_j\}$ sono i voti di `LIN` e $\{y_j\}$ quelli di `SIN`.

1. Usate un opportuno test di livello $\alpha = 5\%$ per stabilire se ci sia dipendenza fra il numero di appelli necessari per superare gli esami `LIN` e `SIN`.

2. Stabilite se il voto medio registrato di `LIN` sia più basso di quello medio registrato per `SIN`. A tal fine, costruite un opportuno test tale che sia al più pari ad $\alpha = 5\%$ la probabilità di commettere l'errore di prima specie di ritenere il voto in `LIN` minore di quello in `SIN`, quando effettivamente è maggiore o uguale. Ipotizzate la normalità dei voti.

3. Stabilite se i voti riportati in `LIN` e `SIN` siano o no indipendenti, a livello $\alpha = 5\%$. Ipotizzate la normalità dei voti.

[8] Dati del tutto fittizi, generati in R.