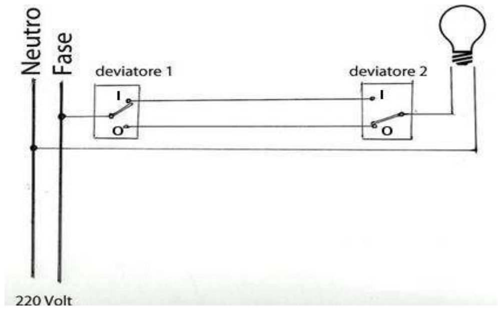
Corso di Statistica - Prof. Fabio Zucca	IV Appello - 5 febbraio 2015		
Nome e cognome:		**Matricola:**	

 8994

Esercizio 1

Si consideri il circuito in figura che consiste in due deviatori che assumono indipendentemente le due posizioni I e O. Il primo deviatore assume la posizione I con probabilità p mentre il secondo assume la posizione I con probabilità $q \neq 1/2$. Sia X la variabile che controlla lo stato della lampadina (1="accesa" e 0="spenta").



1. Descrivere la legge di X.
2. Calcolare, in funzione di p e q il valore atteso $\mathbb{E}[X]$ di X.
3. È possibile ricavare p in funzione di q ed $\mathbb{E}[X]$? In caso affermativo scrivere l'espressione.
4. Si inizializzi il circuito n volte e supponiamo che le posizioni dei due deviatori in tutti questi n esperimenti siano indipendenti. Siano $\{X_i\}_{i=1}^n$ gli stati della lampadina. Esibire uno stimatore non distorto per $\mathbb{E}[X]$.
5. Nelle ipotesi del punto precedente e supponendo che q sia noto, esibire uno stimatore non distorto per p (funzione della famiglia $\{X_i\}_{i=1}^n$). Giustificare la risposta.

Soluzione. Siano P e Q le posizioni dei due deviatori e si osservi che la lampadina è accesa se e solo se $P = Q$.

1. X assume solo due valori con probabilità $\mathbb{P}(P = Q)$ e $\mathbb{P}(P \neq Q) = 1 - \mathbb{P}(P = Q)$. Associando 1 allo stato "lampadina accesa" e 0 allo stato "lampadina spenta" la variabile X diviene una variabile di Bernoulli di parametro $\mathbb{P}(P = Q)$. Utilizzando la formula delle probabilità totali e l'indipendenza di P e Q otteniamo

$$\begin{aligned}\mathbb{P}(P = Q) &= \mathbb{P}(P = Q = I) + \mathbb{P}(P = Q = O) = pq + (1-p)(1-q)\\ &= 2pq + 1 - p - q = p(2q-1) + 1 - q.\end{aligned}$$

   Formalmente si può scrivere $X = \mathbb{1}_{\{P=Q\}}$.

2. Il valore atteso $\mathbb{E}(X)$ coincide con il parametro della legge di Bernoulli $p(2q-1) + 1 - q$

3. Essendo $q \neq 1/2$ si ha $p = (\mathbb{E}(X) + q - 1)/(2q-1)$. Si noti che nel caso $q = 1/2$ si avrebbe $\mathbb{E}(X) = 1 - q$ che, non essendo più iniettiva in p, non sarebbe invertibile.

4. Chiamiamo $\{P_i\}_{i=1}^n$, $\{Q_i\}_{i=1}^n$ e $\{X_i\}_{i=1}^n$ le variabili in gioco negli n esperimenti. Siano $\{P_i, Q_i\}_{i=1}^n$ indipendenti, da cui, essendo $X_i = \mathbb{1}_{\{P_i=Q_i\}}$, si ha che $\{X_i\}_{i=1}^n$ è una famiglia di variabili indipendenti (e chiaramente "identicamente distribuite"). Uno stimatore non distorto per $\mathbb{E}(X)$ è la media campionaria $\overline{X}_n := n^{-1}\sum_{i=1}^n X_i$.

5. Dai punti precedenti, essendo $p = (\mathbb{E}(X) + q - 1)/(2q-1)$ (è una trasformazione affine), allora $T_n := (\overline{X}_n + q - 1)/(2q-1)$ risulta uno stimatore non distorto per p. Infatti

$$\begin{aligned}\mathbb{E}(T_n) &= \mathbb{E}((\overline{X}_n + q - 1)/(2q-1)) = (\mathbb{E}(\overline{X}_n) + q - 1)/(2q-1)\\ &= (\mathbb{E}(X) + q - 1)/(2q-1) = q.\end{aligned}$$

Esercizio 2 Si consideri la seguente funzione reale di variabile reale (dipendente da un parametro $c \in \mathbb{R}$)

$$f_c(x) := \begin{cases} c^2/4 & x \in [-1,1] \\ (2c-1)/6 & x \in (1,2] \\ (2c-1)/3 & x = [-2,-1) \\ 0 & \text{altrove.} \end{cases}$$

1. Calcolare tutti i valori di c affinché f_c sia una densità di una variabile assolutamente continua.

2. Sia X una variabile aleatoria di densità f_c (per i valori di c calcolati al punto precedente): si calcoli la media e la varianza di X. Quanto vale $\mathbb{P}(X \geq 1/2)$? E quanto $\mathbb{P}(X \geq 0 | X > 1/2)$?

3. Siano $X_1, \ldots, X_{100}$ variabili aleatorie i.i.d. con distribuzione calcolata al punto (1). Calcolare (approssimativamente)
$$\mathbb{P}\left(\frac{1}{100}\sum_{i=1}^{100} X_i > 0\right).$$

Soluzione.

1. Le condizioni
$$\begin{cases} f_c(x) \geq 0 & \forall x \in \mathbb{R} \\ \int_{\mathbb{R}} f_c(x)\mathrm{d}x = 1 \end{cases}$$
sono equivalenti a
$$\begin{cases} c \geq 1/2 \\ \frac{c^2}{2} + \frac{2c-1}{3} + \frac{2c-1}{6} = 1 \end{cases}$$
cioè
$$\begin{cases} c \geq 1/2 \\ c^2 + 2c - 3 = 0 \end{cases}$$
che diviene
$$\begin{cases} c \geq 1/2 \\ c = 1 \text{ oppure } c = -3 \end{cases}$$
da cui si ha l'unica soluzione $c = 1$. La densità è
$$f_1(x) := \begin{cases} 1/4 & x \in [-1, 1] \\ 1/6 & x \in (1, 2] \\ 1/3 & x = [-2, -1) \\ 0 & \text{altrove.} \end{cases}$$

2.
$$\begin{aligned}
\mathbb{E}(X) &= \int_{\mathbb{R}} x f_1(x)\mathrm{d}x = x^2/8|_{-1}^{1} + x^2/12|_1^2 + x^2/6|_{-2}^{-1} = -x^2/12|_1^2 = -1/4; \\
\mathrm{var}(X) &= \int_{\mathbb{R}} x^2 f_1(x)\mathrm{d}x - \frac{1}{4^2} \\
&= x^3/12|_{-1}^{1} + x^3/18|_1^2 + x^3/9|_{-2}^{-1} - \frac{1}{4^2} = \frac{1}{6} + \frac{7}{18} + \frac{7}{9} - \frac{1}{16} \\
&= \frac{61}{48} =: \sigma^2; \\
\mathbb{P}(X \geq 1/2) &= \int_{[1/2,+\infty)} f_1(x)\mathrm{d}x = \int_{[1/2,1)} f_1(x)\mathrm{d}x + \int_{[1,2)} f_1(x)\mathrm{d}x + \int_{[2,+\infty)} f_1(x)\mathrm{d}x \\
&= \frac{1}{8} + \frac{1}{6} = \frac{7}{24}; \\
\mathbb{P}(X > 1/2) &= \mathbb{P}(X \geq 1/2) > 0.
\end{aligned}$$
Quindi, essendo $\{X \geq 0\} \supseteq \{X > 1/2\}$ si ha che $\mathbb{P}(X \geq 0 | X > 1/2) = 1$.

3. Dal Teorema Centrale del Limite
$$\frac{1}{100}\sum_{i=1}^{100} X_i \approx Y \sim \mathcal{N}\left(-\frac{1}{4}, \frac{\sigma^2}{100}\right)$$
pertanto, essendo $\sigma = \sqrt{61/48} \approx 1.127312438$,
$$\begin{aligned}
\mathbb{P}\left(\frac{1}{100}\sum_{i=1}^{100} X_i > 0\right) &\approx \mathbb{P}(Y > 0) = \mathbb{P}\left(\frac{Y + 1/4}{\sigma/10} > \frac{1/4}{\sigma/10}\right) \\
&= 1 - \Phi\left(2.217663813\right) \approx 1 - 0.9867 = 0.0133.
\end{aligned}$$

Esercizio 3 Il guadagno annuale della società ferroviaria *DustyRail*, espresso in milioni di euro, viene registrato per 10 anni consecutivi ottenendo $\sum_{i=1}^{10} x_i = -22.5$ e $\sum_{i=1}^{10} x_i^2 = 127.25$.

1. Fornire un estremo superiore (intervallo unilatero) per la varianza σ^2 del guadagno al 90%.

2. Fornire un intervallo bilatero di confidenza per il valore atteso μ del guadagno al 95%.

3. Ci sono evidenze statistiche per concludere che il guadagno annuale medio sia negativo al 5%? Stimare il P-value del test.

4. Ci sono evidenze statistiche per concludere che le perdite annuali medie siano superiori a 2 milioni di euro?

Soluzione. Si calcolano immediatamente la media e la varianza campionarie come $\overline{x}_{10} = -2.25$ mentre $s_{10}^2 = (127.25 - 10 \cdot (-2.25)^2)/9 \approx 8.513889$.

1. L'intervallo di confidenza cercato ha la forma
$$\left[0, \frac{(n-1)s_n^2}{\chi^2_{1-\alpha}(n-1)}\right]$$
dove α è il livello di confidenza. In questo caso $\chi^2_{0.1}(9) = 4.168156$ da cui
$$\left[0, \frac{9 \cdot 8.513889}{4.168156}\right] = \left[0, 18.383429\right].$$

2. L'intervallo di confidenza per la media a varianza incognita ha come estremi $\bar{x}_n \pm t_{(1+\alpha)/2}(n-1)s_n/\sqrt{n}$. Essendo $t_{0.975}(7) = 2.228139$ e $\sqrt{s^2_{10}/10} = 0.922707$ si ricavano gli estremi $-2.25 \pm 2.228139 \cdot \sqrt{8.513889/10} = -2.15 \pm 2.05592$. L'intervallo è quindi
$$[-4.30592, -0.19408].$$

3. Per stabilire se μ è negativo tramite una conclusione forte scegliamo $H_0 : \mu \geq 0$ e $H_1 : \mu_X < 0$. Utilizziamo la seguente regione di rifiuto a livello α
$$T = \frac{\overline{X}_{10} - \mu_0}{S_{10}/\sqrt{n}} < t_\alpha(9)$$
dove la stima di T è $t = -2.342334$ (essendo $\mu_0 = 0$).

   Stimiamo il P-value del test come segue: $-2.342334 = t_{\overline{\alpha}}(9) \equiv -t_{1-\overline{\alpha}}(9)$ da cui $2.342334 = t_{1-\overline{\alpha}}(9)$. Essendo
$$t_{0.975}(9) = 2.262159 < 2.342334 < 2.821434 = t_{0.99}(9)$$
si ha $t_{0.975}(9) < t_{1-\overline{\alpha}}(9) < t_{0.99}(9)$ da cui $0.025 > \overline{\alpha} > 0.01$ (anche se di fatto $t_{0.975}(9) \approx t_{1-\overline{\alpha}}(9)$ da cui $0.025 \approx \overline{\alpha}$). Quindi al 5% concludiamo che il guadagno medio è negativo.

4. Si deve studiare il test $H_0 : \mu_X \geq -2$ e $H_1 : \mu_X < -2$. Si utilizza pertanto la seguente regione di rifiuto a livello α
$$T = \frac{\overline{X}_{10} - \mu_0}{S_{10}/\sqrt{n}} < t_\alpha(9)$$
dove la stima di T è $t = -0.26026$ (essendo $\mu_0 = -2$).

   Similmente al caso precedente, per il secondo test si ottiene la stima $\overline{\alpha} \in (0.2, 0.5)$ (ricordiamo che $t_\alpha(n) < 0$ (risp. > 0) se e solo se $\alpha < 1/2$ (risp. $> 1/2$)). In questo caso non ci sono evidenze statistiche per concludere che le perdite medie annuali siano superiori a 2 milioni di euro.

Esercizio 4 È stata messa a punto una nuova segnaletica per ridurre il numero di incidenti mensili sulle autostrade. Si sono calcolate, per ciascuna delle 10 autostrade controllate il numero di incidenti nel mese antecedente all'introduzione della nuova segnaletica (siano x_i) e nel mese successivo (siano y_i). Sappiamo che $\sum_{i=1}^{10} x_i = 100$, $\sum_{i=1}^{10} y_i = 76$, $\sum_{i=1}^{10} x_i^2 = 1200$, $\sum_{i=1}^{10} y_i^2 = 698$ e $\sum_{i=1}^{10} x_i y_i = 901$.

1. Calcolare la covarianza campionaria della coppia (X, Y). Calcolare la media e la varianza campionaria della variabile $Z = Y - X$.

2. Si può ritenere davvero efficace la nuova segnaletica al 10%?

3. Si può ritenere che la media del numero di incidenti per autostrada si sia ridotta di 1 unità al 10%?

4. Stimare il P-value di ciascun test.

Soluzione.

1. L'ampiezza del campione è $n = 10$. Dati $\bar{x}_{10} = \frac{1}{n}\sum_{i=1}^n x_i = 100/10 = 10$ e $\bar{y}_{10} = \frac{1}{n}\sum_{i=1}^n y_i = 76/10 = 7.6$, la covarianza si calcola come $\frac{1}{n-1}\sum_{i=1}^n x_i y_i - \frac{n}{n-1}\bar{x}\bar{y} = \frac{10}{9}(901/10 - 10\cdot 7.6) = 47/3 = 15.66667$. La media campionaria di Z è $\bar{z}_{10} = \bar{y}_{10} - \bar{x}_{10} = -2.4$. La varianza invece è

$$\begin{aligned} s^2_{10} &= \frac{1}{n-1}\sum_{i=1}^n (x_i - y_i)^2 - \frac{n}{n-1}\bar{z}^2_{10} \\ &= \frac{1}{n-1}\Big(\sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2 - 2\sum_{i=1}^n x_i y_i\Big) - \frac{n}{n-1}\bar{z}^2_{10} \\ &= \frac{1}{9}(1200 + 698 - 1802) - 10\cdot 5.76/9 = 192/45 \approx 4.266667. \end{aligned}$$

2. Per stabilire l'efficacia della nuova segnaletica tramite una conclusione forte (cioè per avere una forte evidenza della sua efficacia) scegliamo $H_0 : \mu_Y \geq \mu_X$ e $H_1 : \mu_Y < \mu_X$. Questo significa studiare il test $H_0 : \mu_Y - \mu_X \geq 0$ e $H_1 : \mu_Y - \mu_X < 0$.

   Se $z_i := y_i - x_i$ allora $\overline{z}_{10} = -2.4$ mentre $s^2_{10} \approx 4.266667$. Utilizziamo la seguente regione di rifiuto a livello α

   $$T = \frac{\overline{Z}_{10} - \delta}{S_{10}/\sqrt{n}} < t_\alpha(9)$$

   dove la stima di T è $t = -3.674235$ (essendo $\delta = 0$). Dalle tavole $t_{0.10}(9) = -t_{0.9}(9) = -1.383029$ quindi rifiuto H_0 al livello 0.1.

3. Per la seconda parte si deve studiare il test $H_0 : \mu_Y - \mu_X \geq -1$ e $H_1 : \mu_Y - \mu_X < -1$. Si utilizza pertanto la seguente regione di rifiuto a livello α

   $$T = \frac{\overline{Z}_{10} - \delta}{S_{10}/\sqrt{n}} < t_\alpha(9)$$

   dove stavolta la stima di T è $t = -2.143304$ (essendo $\delta = -1$) che ci consente ancora di rifiutare H_0 a livello 0.1.

4. Stimiamo il P-value del primo test come segue: $-3.674235 = t_{\overline{\alpha}}(9) \equiv -t_{1-\overline{\alpha}}(9)$ da cui $3.674235 = t_{1-\overline{\alpha}}(9)$. Essendo

   $$t_{0.995}(9) = 3.249843 < 3.674235 < 3.689638 = t_{0.9975}(9)$$

   si ha $t_{0.995}(9) < t_{1-\overline{\alpha}}(9) < t_{0.9975}(9)$ da cui $0.005 > \overline{\alpha} > 0.0025$.

   Similmente al caso precedente, per il secondo test si ottiene la stima $\overline{\alpha} \in (0.025, 0.05)$.