Cognome : **Nome :** **Matricola:**

Modulo: Laboratorio di Metodi Matematici e Statistici, M-Z

Appello del 3 Febbraio 2017

Esercizio 1.

È stato calcolato che il 40% dei visitatori di un museo ha diritto a un biglietto di ingresso ridotto.
a) Calcolare la probabilità che, scegliendo a caso 8 visitatori, esattamente 2 abbiano la riduzione.
b) Calcolare la probabilità che, scegliendo a caso 8 visitatori, almeno 3 abbiano la riduzione.
c) Calcolare la probabilità che scegliendo a caso 5 visitatori non ci sia nessuno con riduzione.
d) Calcolare in modo approssimato la probabilità che, su 800 visitatori scelti a caso, coloro che hanno la riduzione siano in numero compreso tra 200 e 350 (giustificare le approssimazioni introdotte).

Soluzione. In tutti i casi seguenti facciamo riferimento a estrazioni con reimmissione, pertanto stiamo facendo riferimento a uno schema di prove ripetute indipendenti, con "successo" se il visitatore ha la riduzione e "insuccesso" in caso contrario. La probabilità di successo è $p = 0.4$.

a) Il numero di visitatori che ha riduzione, su un totale di 8, è una variabile aleatoria X con distribuzione binomiale di parametri $n = 8$, $p = 0.4$ e si trova

$$P(X = 2) = \binom{8}{2} 0.4^2 0.6^6 = 0.2090.$$

b) Occorre calcolare

$$P(X \geq 3) = 1 - P(X = 0) - P(X = 1) - P(X = 2) =$$

$$1 - \binom{8}{0} 0.4^0 0.6^8 - \binom{8}{1} 0.4^1 0.6^7 - \binom{8}{2} 0.4^2 0.6^6 = 0.6846.$$

c) Ora X ha parametri $n = 5$, $p = 0.4$, l'evento considerato capita precisamente quando $X = 0$ e si trova

$$P(X = 0) = (1 - p)^5 = 0.6^5 = 0.0778.$$

d) Ora X ha parametri $n = 800$, $p = 0.4$ e poiché $np = 320 > 5$, $n(1 - p) = 480 > 5$ si può usare l'approssimazione normale cioè considerare la variabile

$$Z = \frac{X - np}{\sqrt{np(1 - p)}}$$

come una normale standard. Poiché $\sqrt{np(1 - p)} = 13.8564$,

$$P(200 \leq X \leq 350) = P\left(\frac{200 - 320}{13.8564} \leq Z \leq \frac{350 - 320}{13.8564}\right) = P\left(-8.66 < Z < 2.17\right)$$

$$= P(Z < 2.17) - P(Z < -8.66) \simeq P(Z < 2.17) = 0.9850.$$

Esercizio 2.
La notizia di un inquinamento delle acque ha spinto l'associazione degli albergatori della costa a intervistare i visitatori dell'anno precedente, chiedendo loro se hanno intenzione di cambiare meta per le vacanze estive. Su un campione di 2500 intervistati scelti a caso, 178 hanno deciso di non tornare.
a) Fornire una stima della frazione di visitatori che hanno deciso di cambiare meta per le vacanze.
b) Calcolare un intervallo di confidenza approssimato al livello del 95% per la percentuale di visitatori che hanno deciso di cambiare meta (giustificare le approssimazioni introdotte).
c) L'associazione degli albergatori afferma che tali dati dimostrano che la frazione di visitatori che ha cambiato meta è superiore al 5%. I dati consentono di giungere a tale affermazione, con un livello di significatività del 5%?
d) Senza fare calcoli o consultare tabelle, potremmo giungere alla stessa conclusione con un livello di significatività del 2%? Giustificare la risposta.

Soluzione.

Sia p la proporzione di visitatori che cambiano meta, $n = 2500$ la numerosità del campione osservato, $s = 178$ il numero di coloro che cambiano meta nel campione osservato.

a) Si stima p con $\hat{p} = \frac{s}{n} = \frac{178}{2500} = 0.0712$.

b) L'intervallo di confidenza al livello $1 - \alpha = 0.95$ si ottiene calcolando il valore critico $z_{\alpha/2} = z_{0.025} = 1.96$ e applicando la formula

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0712 \pm 1.96\sqrt{\frac{0.0712(1-0.0712)}{2500}} = 0.0712 \pm 0.0101$$

che dà luogo all'intervallo $[0.0611, 0.0813]$, cioè tra il 6.11% e l'8.13%.

L'approssimazione è accettabile perché risulta

$$n\hat{p} = 178 > 5, \qquad n(1-\hat{p}) = 2322 > 5.$$

c) Si esegue un test di livello $\alpha = 0.05$ per l'ipotesi $H_0 : p = p_0 = 0.05$ contro l'alternativa $H_1 : p > 0.05$. Si calcola

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = 4.8636$$

e si rifiuta H_0 se risulta $z > z_\alpha = z_{0.05} = 1.6449$ Poiché ciò accade, i dati consentono di rifiutare H_0 e quindi di confermare l'affermazione degli albergatori. Si noti che non possiamo usare il risultato ottenuto al punto b) perché il test corretto da fare è unilatero.

d) In generale, senza ulteriori conti o verifiche non potremmo rispondere definitivamente, in quanto il nuovo valore critico potrebbe anche essere superiore al valore della statistica test. Tuttavia, in questo caso è possibile proprio guardando al valore della statistica test, che è molto prossimo a 5, e ricordando la regola del "3-sigma" e del "5-sigma" possiamo concludere che si rifiuta l'ipotesi nulla a qualunque livello di significatività "sensato".

Esercizio 3.
Il gestore di una centrale telefonica afferma che, mediamente, quando si esegue una chiamata alla centrale si trova la linea libera nel 50% dei casi, linea occupata con attesa minore o uguale a un minuto nel 40% dei casi, linea occupata con attesa maggiore di un minuto nel restante 10% dei casi. Sono state effettuate 70 chiamate di prova, e i risultati sono riportati nella seguente tabella di frequenze:

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

a) Verificare, tramite un opportuno test, l'affermazione del gestore al livello dell'1%.
b) Fornire un'approssimazione del p-valore del test eseguito al punto a), giustificando la risposta.
c) Si è anche presa nota della fascia oraria in cui le 70 telefonate sono state fatte:

	Esito della chiamata		
Fascia oraria	linea libera	attesa ≤ 1	attesa > 1
$8-12$	16	0	11
$12-16$	1	11	0
$16-20$	20	11	0

In che modo è possibile quantificare l'eventuale associazione tra l'esito e la fascia oraria della chiamata?
d) Senza fare conti, è possibile sostenere che le due variabili al punto c) sono indipendenti? Si giustifichi la risposta.

Soluzione. a) Si esegue un test χ^2 di buon adattamento. Poiché, in base alle affermazioni del gestore, i tre eventi considerati capitano rispettivamente nel 50%, 40%, 10% dei casi, le probabilità attese sono

	linea libera	attesa ≤ 1 min.	attesa > 1 min.	Totale
Probabilità attese p_i	0.5	0.4	0.1	1

La tabella delle frequenze (per $k = 3$ classi) è allora

	linea libera	attesa ≤ 1 min.	attesa > 1 min.	Totale
Frequenze osservate n_j	37	22	11	$n = 70$
Frequenze attese $n_j^\star = np_i$	35	28	7	$n = 70$

Si calcola

$$\sum_{j=1}^{k} \frac{(n_j^\star - n_j)^2}{n_j^\star} = \frac{(35-37)^2}{35} + \frac{(28-22)^2}{28} + \frac{(7-11)^2}{7} = 3.6857$$

e si rifiuta l'ipotesi che la distribuzione sia quella indicata dal gestore se tale valore supera $\chi^2_\alpha(k-1) = \chi^2_{0.01}(2) = 9.21035$. Poichè ciò non accade, i dati non consentono di confutare l'affermazione del gestore.
b) Con riferimento alla tavola dei valori critici della distribuzione $\chi^2(2)$ si può affermare che il p-valore del test, area sotto la curva di densità a destra del valore della statistica test (3.6857), è compreso tra 0.10 e 0.20.
c) L'associazione tra le due variabili si può valutare, descrittivamente, calcolando l'indice del χ^2 sulla tabella a doppia entrata, che in questo vaso vale 45.528 (si noti che in questo problema il calcolo dell'indice non è esplicitamente richiesto). L'eventuale indicazione di associazione emersa dal calcolo dell'indice andrebbe poi sottoposta a verifica d'ipotesi di livello α, confrontando il valore dell'indice con il valore critico $\chi^2(2 \times 2)_{1-\alpha}$ oppure tramite il calcolo del p-value: si rifiuta l'ipotesi nulla se il valore critico viene superato o, alternativamente, se il p-valore è inferiore a 0.05. Avendo fatto il calcolo dell'indice ed avendo trovato che questo supera di gran lunga il valore critico più elevato riportato sulle tavole per la densità $\chi^2(2 \times 2)$, si conclude che si può rifiutare l'ipotesi nulla di indipendenza a favore di una associazione significativa a qualunque livello di significatività .
d) Nella tabella a doppia entrata sono presenti dei valori pari a 0, che sono una chiara indicazione di associazione (o dipendenza) tra le variabili: infatti, per esempio, il valore 0 nell'ultima riga della tabella permette di escludere di aver atteso più di 1 minuto qualora la chiamata sia stata fatta nella fascia oraria $16-20$. Le due variabili, pertanto, non possono essere indipendenti, come appunto ottenuto nel punto c).