Nome e cognome: _____

indirizzo email: _____

Esame di Statistica – Esercitazione del 19 maggio 2017

1. Secondo un'indagine recente, negli Stati Uniti al 17% delle donne di età superiore a 65 anni è stato diagnosticato un tumore. Determinare la probabilità che, in un campione di 220 individui estratto da quella popolazione (donne di età over 65 negli Stati Uniti), oltre il 20% abbia avuto una diagnosi di tumore.

---

argomentare tutti i passaggi in modo adeguato

Per ciascun elemento del campione indichiamo con X_i la presenza o assenza di diagnosi di tumore. Quindi, per $i = 1, \cdots, n$ le variabili X_i hanno tutte distribuzione Bernulliana con $p = 0.17$ e sono indipendenti l'una dall'altra.

Sia X il numero totale di individui nel campione a cui è stato diagnosticato un tumore. Chiaramente $X = \sum_{i=1}^{n} X_i$ ha distribuzione Binomiale di parametri $(n = 220, p = 0.17)$.

Possiamo approssimare la probabilità richiesta utilizzando il Teorema del Limite Centrale: se il campione è grande (in questo caso lo è perché $n = 220$), la distribuzione di X è approssimativamente normale con parametri $\mu = np = 37.4$ e $\sigma^2 = np(1-p) = 31.042$.

Di conseguenza

$$P(X \geq 44) = P\left(\frac{X-\mu}{\sigma} \geq \frac{44 - 37.4}{5.57}\right) \simeq P(Z \geq 1.18) = 1 - \Phi(1.18) = 0.119.$$

In alternativa, invece della distribuzione di X si sarebbe potuta approssimare la distribuzione di $\hat{p} = X/n$ con una normale di media $p = 0.17$ e varianza $p(1-p)/n = 0.0006413636$ ottenendo

$$P(\hat{p} \geq 0.2) = P\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \geq \frac{0.2 - 0.17}{0.0253}\right) \simeq P(Z \geq 1.18) = 1 - \Phi(1.18) = 0.119.$$

2. Alcuni studi suggeriscono che un moderato consumo di alcol possa ridurre il rischio di infarto e che il vino rosso offra dei benefici particolari incrementando i polifenoli nel sangue.
In un esperimento alcuni uomini adulti in buona salute sono stati divisi a caso in due gruppi: quelli del primo gruppo hanno bevuto mezza bottiglia di vino rosso al giorno per due settimane, quelli del secondo gruppo hanno bevuto vino bianco con le stesse modalità.
Per ciascun elemento dei due gruppi, sono state osservate le variazioni percentuali di polifenoli nel sangue ottenendo i risultati seguenti:

Vino rosso	3.5	8.1	7.4	4.0	0.7	4.9	8.4	7.0	5.5
Vino bianco	3.1	0.5	-3.8	4.1	-0.6	2.7	1.9	-5.9	0.1

Ipotizzando che la variazione percentuale di polifenoli segua un andamento normale e che le varianze relative alle due popolazioni possano considerarsi uguali,

(a) costruire un intervallo di confidenza bilaterale al 90% per la differenza tra l'incremento percentuale medio ottenuto bevendo vino rosso e quello ottenuto bevendo vino bianco;

(b) utilizzando il risultato ottenuto al punto precedente, stabilire con livello di significatività al 5% se i dati forniscono evidenza in favore dell'ipotesi che il vino rosso possa far aumentare i polifenoli nel sangue mediamente più del vino bianco.

---

argomentare tutti i passaggi in modo adeguato

Indichiamo con X_R l'incremento percentuale di polifenoli dopo il "trattamento" con il vino rosso e con X_B quello relativo al vino bianco (useremo il pedice R anche per tutte le quantità (media, varianza, etc) relative al campione trattato con vino rosso e il pedice B per quelle relative al campione trattato con vino bianco).

Le ipotesi del problema permettono di modellare utilizzando delle distribuzioni normali con la stessa varianza, quindi

$$X_R \sim N(\mu_R, \sigma^2) \qquad e \qquad X_B \sim N(\mu_B, \sigma^2).$$

(a) Per calcolare l'intervallo di confidenza per $\mu_R - \mu_B$ ricordiamo che $\bar{X}_R - \bar{X}_B \sim N(\mu_R - \mu_B, \frac{\sigma^2}{n_R} + \frac{\sigma^2}{n_B})$.

Dal momento che la varianza (uguale per i due gruppi) è incognita, dobbiamo far riferimento alla quantità

$$\frac{\bar{X}_R - \bar{X}_B - (\mu_R - \mu_B)}{S_p\sqrt{\frac{1}{n_R} + \frac{1}{n_B}}} \sim T_{n_R+n_B-2}$$

dove $S_p^2 = \frac{(n_R-1)S_R^2+(n_B-1)S_B^2}{n_R+n_B-2}$. L'intervallo di confidenza quindi è

$$\bar{x}_R - \bar{x}_B \pm t_{\alpha/2, n_R+n_B-2} \cdot s_p\sqrt{\frac{1}{n_R} + \frac{1}{n_B}}.$$

Dal momento che $\bar{x}_R = 5.5$, $s_R^2 = 6.33$, $n_R = 9$, $\bar{x}_B = 0.23$, $s_B^2 = 10.84$, $n_B = 9$ e $s_p = 2.93$, considerando che $t_{0.05,16} = 1.746$, l'intervallo bilaterale al 90% è

$$5.5 - 0.23 \pm 1.746 \cdot 2.93\sqrt{2/9} = (2.86, 7.68)$$

(b) Si vuole testare $H_0 : \mu_R - \mu_B \leq 0$ contro $H_1 : \mu_R - \mu_B > 0$.

Dal punto precedente deduciamo immediatamente che l'intervallo di confidenza unilaterale destro al 95% per il parametro $\mu_R - \mu_B$ è $(2.86, +\infty)$. L'ipotesi nulla "di confine" $\mu_R - \mu_B = 0$ non è contenuta nell'intervallo e quindi va rifiutata.

In conclusione, i dati forniscono evidenza in favore dell'ipotesi che il vino rosso possa far aumentare i polifenoli nel sangue mediamente più del vino bianco.

3. Gli astrologi sostengono che la posizione dei pianeti al momento della nascita di un individuo possa caratterizzarne la personalità. In California è stato effettuato un esperimento per stabilire se effettivamente le informazioni sui dati di nascita possano aiutare un astrologo a determinare la personalità di un individuo.

- Tre volontari scelti a caso hanno compilato un questionario che ne evidenzia i tratti caratteristici delle personalità.
- Ad un astrologo sono stati consegnati i tre questionari compilati insieme ai dati di nascita (data e orario esatti) di uno solo dei tre individui.
- All'astrologo è stato chiesto di stabilire quale dei tre questionari corrispondesse ai dati di nascita.

L'operazione è stata ripetuta con $n = 116$ terne di volontari, l'astrologo ha associato correttamente dati di nascita e tratti caratteriali 40 volte.

(a) A che tipo di distribuzione si può fare riferimento?

(b) Esprimere il seguente sistema di ipotesi in termini di condizioni sui parametri:

H_0 : l'astrologo non ha nessun potere di prevedere la personalità basandosi sui dati di nascita (in sostanza associa a caso)

H_1 : l'astrologo ha dei poteri previsivi.

(c) Quanto vale il p-value (p-dei-dati) di questo test? Che conclusioni possiamo trarre?

---

argomentare tutti i passaggi in modo adeguato

(a) Per ogni terna di volontari consideriamo come successo il fatto che l'astrologo riesca ad associare correttamente la data di nascita ai tratti caratteriali. Siamo quindi in presenza di una distribuzione di Bernoulli il cui parametro p è la probabilità di associare correttamente.

(b) Nel caso in cui l'associazione viene fatta a caso, il valore di p sarebbe $1/3$. Se invece l'astrologo ha veramente delle capacità previsive allora $p > 1/3$. Il sistema di ipotesi quindi è

$$H_0 : p = p_0 = \frac{1}{3} \qquad \textit{contro} \qquad H_1 : p > \frac{1}{3}.$$

(c) Il campione è grande abbastanza da poter utilizzare l'approssimazione normale alla distribuzione di $\hat{p}$ fornita dal Teorema del Limite Centrale. Possiamo quindi supporre $\hat{p} \sim N(p, p(1-p)/n)$.

Adottiamo la statistica test

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

che, in caso di validità dell'ipotesi nulla di confine $p = p_0$, ha distribuzione normale standard.

Il test porterà al rifiuto dell'ipotesi nulla per valori alti di T (quindi, per un livello di significatività $\alpha = 5\%$, per $T \geq 1.64$).

Nel campione osservato si ha $\hat{p} = 40/116 = 0.345$ e $t = 0.26$.

Il p-value corrispondente è

$$P_{H_0}(T \geq t) = 1 - \Phi(0.26) \simeq 0.4.$$

Il p-value è decisamente molto alto, il che porta a NON RIFIUTARE l'ipotesi nulla e quindi concludere che non c'è nessuna evidenza che l'astrologo riesca a prevedere meglio di uno che decide a caso.