Corso di Statistica - Prof. Fabio Zucca	Appello 2 - 18 luglio 2018		
Nome e cognome:		**Matricola:**	

 12957

Esercizio 1 Supponiamo di avere $K+N$ monete di cui K equilibrate (una testa ed una croce, equiprobabili) e N con due teste ciascuna. Prendo una moneta a caso (ciascuna ha probabilità $1/(K+N)$ di essere scelta.

1. Qual è la probabilità che esca testa?
2. Sapendo che è uscita testa, qual è la probabilità di aver scelto una di quelle non equilibrate?
3. Se $K = 2$ e $N = 3$, qual è la probabilità che in 100 pescate/lanci escano almeno 60 teste (ogni volta che si pesca una moneta, questa viene rimessa insieme alle altre dopo il lancio; le pescate sono indipendenti).

Chi non riuscisse a svolgere l'esercizio in forma astratta, svolga tutto nel caso specifico $K = 2$ e $N = 3$.

Soluzione. Si considerino gli eventi E :="pesco una moneta equilibrata" e T :="esce testa". Dai dati si ha $\mathbb{P}(E) = K/(K+N)$, $\mathbb{P}(T|E) = 1/2$ e $\mathbb{P}(T|E^c) = 1$.

1. Dalla formula delle probabilità totali

$$\mathbb{P}(T) = \mathbb{P}(T|E)\mathbb{P}(E) + \mathbb{P}(T|E^c)\mathbb{P}(E^c) = \frac{1}{2}\frac{K}{K+N} + \frac{N}{K+N} = \frac{1}{2}\Big(1 + \frac{N}{K+N}\Big).$$

   Nel caso specifico 4/5.

2. $$\mathbb{P}(E^c|T) = \frac{\mathbb{P}(T|E^c)\mathbb{P}(E^c)}{\mathbb{P}(T)} = \frac{N/(N+K)}{(K+2N)/(2K+2N)} = \frac{2N}{K+2N}.$$

   Nel caso esplicito 3/4.

3. Siano $\{Z_i\}_{i=1}^{100}$ i risultati dei lanci (1 se esce testa e 0 se esce croce), sono tutte variabili Bernoulliane indipendenti di parametro 4/5. Pertanto, essendo $100 \cdot 4/5 > 100/5 > 5$ posso utilizzare l'approssimazione gaussiana da cui

$$\mathbb{P}\Big(\sum_i Z_i \geq 60\Big) = \mathbb{P}\Big((\sum_i Z_i - 100 \cdot 4/5)/\sqrt{100 \cdot 4/25} \geq (59.5 - 100 \cdot 4/5)/\sqrt{100 \cdot 4/25}\Big) \approx 1 - \Phi(-20.5/4) \approx 1.$$

Esercizio 2 Un atleta, specialista dello sprint sui 100m, ha una falcata costante di 2.5 metri di lunghezza. Il tempo necessario per completarla è una variabile aleatoria di valore atteso $\mu = 0.241$ e varianza $\sigma^2 = 10^{-5}\,\mathrm{s}^2$ ed i passi sono indipendenti. Sia X il tempo necessario a coprire i 100m (40 falcate).

1. Determinare il valore medio, la deviazione standard di X.
2. Se i tempi delle gare sono indipendenti, qual è approssimativamente la probabilità che entro le prossime 10 gare riesca a battere l'attuale record del mondo (9.58 s)?

Soluzione.

1. Poichè ci vogliono 40 passi per completare la gara, se $\{T_i\}_{i=1}^{40}$ sono i tempi necessari per ciascuno dei passi allora $X = \sum_{i=1}^{40} T_i$ da cui $\mathbb{E}(X) = 9.64$ e $\sqrt{\mathrm{var}(X)} = 0.02$ (in secondi).

2. Poiché X è una somma di 40 variabili i.i.d. possiamo approssimare la legge di X con quella di una normale $\mathcal{N}(9.64, 4 \cdot 10^{-4})$. Sia $\{X_i\}_{i=1}^{n}$ la successione dei tempi nelle prossime n gare. Se $n = 10$,

$$\begin{aligned}\mathbb{P}(\cup_{i=1}^{10}\{X_i < 9.58\}) &= 1 - \mathbb{P}(\cap_{i=1}^{10}\{X_i \geq 9.58\}) = 1 - \mathbb{P}(X \geq 9.58)^{10} \\ &= 1 - \mathbb{P}((X - 9.64)/0.02 \geq (9.58 - 9.64)/0.02)^{10} \\ &\approx 1 - (1 - \Phi((9.58 - 9.64)/0.02))^{10} = 1 - (1 - \Phi(-3))^{10} \\ &= 1 - \Phi(3)^{10} = 1 - 0.9987^{10} = 1 - 0.98708 = 0.01292.\end{aligned}$$

 12957

Esercizio 3 All'oratorio di quartiere organizzano tutte le sere una tombola per anziani. Si ricordi che i numeri della tombola sono 90. Mia nonna, assidua frequentatrice, ha notato che il numero 90 non esce da un po' di giorni. Pensiamo di assistere ad n estrazioni con reimmissione e supponiamo che siano tutte indipendenti ed equidistribuite tra i possibili esiti.

1. Scegliamo un test adeguato. Quanto deve essere ampio il campione di osservazioni per ritenere ragionevole che si siano persi un numero?
   Considereremo ragionevole supporre che si siano persi il numero 90 se $\bar{\alpha} \leq 0.001$ ovvero se il p-value sarà sufficientemente basso da poter rifutare H_0 per tutti i valori "ragionevoli" di α.

2. Supponiamo che la probabilità, a priori, di perdere il numero 90 sia 1/100. Qual è la probabilità di averlo perso se non è uscito dopo 300 estrazioni con reimmissione?

Soluzione.

1. Ci è richiesto di verificare con opportuno test, l'adattamento dei dati ad una distribuzione discreta uniforme sui primi $m = 90$ interi $\{1, 2, ..., 89, 90\}$.
   Il problema che vogliamo evidenziare è che non esce mai il numero 90. Possiamo quindi raggruppare i dati in due classi: $\{1, \ldots, 89\}$ e $\{90\}$.
   Sia quindi $N_c = 2$ il numero delle classi.
   Considerando le estrazioni equiprobabili avremo le frequenze relative teoriche: $p_1 = \frac{89}{90}$ e $p_2 = \frac{1}{90}$.
   Poiché il numero 90 non esce mai avremo disponibili anche le frequenze relative osservate senza bisogno di fare calcoli perché tutte le osservazioni finiranno nella prima classe che pertanto avrà frequenza relativa pari a 1 indipendentemente da n: $f_1 = 1$ e $f_2 = 0$.
   Per prima cosa verifichiamo le condizioni di applicabilità del teorema centrale del limite: $np_i \geq 5$ e $n(1 - p_i) \geq 5$. Esse saranno soddisfatte per $n \geq 450$.
   La statistica test è $Q(n) = n \sum_{i=1}^{N_c} \frac{(p_i - f_i)^2}{p_i} = n \left[\frac{\left(\frac{89}{90} - 1\right)^2}{\frac{89}{90}} + \frac{\left(\frac{1}{90} - 0\right)^2}{\frac{1}{90}} \right] = \frac{n}{89}$
   Ora si tratta di determinare il più piccolo valore di n per cui $\bar{\alpha} \leq 0.01$ ovvero $1 - \bar{\alpha} \geq 0.999$ ovvero $\chi^2_{1-\bar{\alpha}}(1) \geq \chi^2_{0.999}(1)$ ovvero $Q(n) \geq 10.82736$ essendo, per definizione di p-value, $Q(n) = \chi^2_{1-\bar{\alpha}}(k - 1)$ e, dalle tavole dei quantili della distribuzione Chi-quadrato, $\chi^2_{0.999}(1) = 10.82736$.
   Pertanto se $\frac{n}{89} \geq 10.82736$ ovvero se $n \geq \lceil 89 \cdot 10.82736 \rceil = 964$ il p-value sarà sufficientemente basso da poter considerare H_0 rifiutata per quasi tutti i valori di α.
   Quindi già con una successione di 964 estrazioni tra le quali manchi sistematicamente il numero 90 possiamo ragionevolmente concludere che si sono persi un numero.

   Naturalmente il metodo è inefficace in quanto si farebbe prima a far passare sistematicamente tutti i numeri. Si noti inoltre che un test su un campione bernoulliano non possibile in quanto la condizione $n\bar{p}_n \geq 5$ non può essere verificata, visto che la frequenza $\bar{p}_n$ di uscite del numero 90 è nulla.

2. Per la seconda parte si considerino gli eventi P="il numero 90 è stato perso" e N="il numero 90 non esce in 300 estrazioni"; sappiamo che $\mathbb{P}(P) = 1/100$ mentre $\mathbb{P}(N|P) = 1$ e $\mathbb{P}(N|P^c) = (89/90)^{300}$. Da cui
$$\mathbb{P}(P|N) = \frac{\mathbb{P}(N|P)\mathbb{P}(P)}{\mathbb{P}(N|P)\mathbb{P}(P) + \mathbb{P}(N|P^c)\mathbb{P}(P^c)} = \frac{0.01}{0.044664453} = 0.221975559.$$

 12957

Esercizio 4 Il frinosoma *Phrynosoma mcalli*, un rettile iguanide, ha molte caratteristiche insolite, a partire dalla sua testa che è circondata da una corona di aculei simili a corna. Gli erpetologi hanno recentemente sottoposto a verifica statistica l'ipotesi che i lunghi aculei contribuiscano a proteggere questi animali dalla predazione. Per fare ciò alcuni ricercatori hanno misurato la lunghezza delle corna di due gruppi di frinosomi distinguendo tra frinosomi vivi e frinosomi uccisi da predatori ottenendo i seguenti dati sintetici:

Gruppo di frinosomi	Media campionaria	Deviazione standard campionaria	Ampiezza del campione
Vivi	$\bar{x}_V = 24.28$ mm	$s_V = 2.63$ mm	$n_V = 154$
Morti	$\bar{x}_M = 21.99$ mm	$s_M = 2.71$ mm	$n_M = 30$

Sulla base di tali dati, supponendo che le popolazioni $\{V_i\}_{i=1}^{n_V}$ e $\{M_i\}_{i=1}^{n_M}$ di provenienza siano normali e in situazione di omoschedasticità, verificare tramite l'analisi del p-value se ci sono evidenze statistiche per affermare che i veri valori μ_V e μ_M delle medie delle due popolazioni sono diversi tra loro.

Soluzione. Ci è richiesto di condurre un test di ipotesi per il confronto tra le medie di due popolazioni normali indipendenti in situazione di omoschedasticità ovvero con varianze entrambe incognite ma che si presuppone siano uguali.
L'ipotesi da sottoporre a test è $H : \mu_V \neq \mu_M$. Vogliamo accettare H solo di fronte ad una evidenza statistica quindi $H_0 : \mu_V = \mu_M$ contro $H_1 : \mu_V \neq \mu_M$.

La statistica test è $\hat{t} = \frac{\bar{x}_V - \bar{x}_M}{s\sqrt{\frac{1}{n_V}+\frac{1}{n_M}}} = \frac{24.28-21.99}{2.6429\sqrt{\frac{1}{154}+\frac{1}{30}}} \approx 4.3418$ dove $s = \sqrt{\frac{(n_V-1)s_V^2+(n_M-1)s_M^2}{n_V+n_M-2}} = \sqrt{\frac{153\cdot 2.63+30\cdot 2.71}{182}} \approx \sqrt{6.9849} \approx 2.6429$ è la deviazione standard combinata.

Il p-value è quel valore $\bar{\alpha} : t_{1-\frac{\bar{\alpha}}{2}}(k) = |\hat{t}|$ dove $t_\beta(k)$ è il quantile di livello β di una distribuzione t di student con k gradi libertà . Dalle tavole dei quantili ricaviamo che $t_{1-\frac{\bar{\alpha}}{2}}(182) = 4.3418 \gg 3.159512 = t_{0.999}(120) \approx t_{0.999}(182)$ pertanto, passando ai livelli del quantile, otteniamo $0.999 < 1 - \frac{\bar{\alpha}}{2} < 1$ ovvero $0 < \bar{\alpha} < 0.001$ (analogo risultato si sarebbe ottenuto approssimando $t_\beta(182) \approx z_\beta$ o con calcolo diretto $\bar{\alpha} \approx 2 - 2\Phi(|\hat{t}|)$). Possiamo concludere che H_0 è rifiutata per quasi tutti i valori di α e che dunque ci sono (forti) evidenze statistiche per affermare che i veri valori delle medie delle due popolazioni sono diversi tra loro.