Corso di Statistica - Prof. Fabio Zucca	II Prova in itinere - 3 luglio 2015		
Nome e cognome:		**Matricola:**	

 12957

Esercizio 1 La I prova in itinere di Statistica si è svolta nelle aule A, B e C con 78, 74 e 97 studenti rispettivamente. La temperatura delle prime due aule era paragonabile mentre quella dell'aula C era sensibilmente più alta. I dati riguardanti i voti del primo esercizio sono sintetizzati nella tabella seguente

	A	B	C
Ampiezza campione	78	74	97
Somma dei voti	620.6375	588.8375	744.485
Somma quadrati dei voti	5223.5042	4981.6383	6167.5135

Si supponga di essere in regime di omoschedasticità.

1. Calcolare medie e varianze campionarie dei campioni.
2. Trovare l'intervallo di confidenza bilatero al 95% per la media dei voti dell'aula A.
3. Vi sono evidenze statistiche per affermare al 5% che il voto atteso di uno studente dell'aula A è differente da quello di uno studente dell'aula B? Descrivere il test utilizzato e stimarne il P-value.
4. Vi sono evidenze statistiche per affermare all'1% che la temperatura abbia penalizzato gli studenti dell'aula C rispetto a quelli della A? Descrivere il test utilizzato e stimarne il P-value.
5. **(Domanda bonus)** Il docente non vuole correre il rischio di sottovalutare gli effetti negativi della temperatura in aula C rispetto alla A. Proporre un test adeguato, svolgerlo al 10% e stimarne il P-value.

Soluzione.

1. La media campionaria è $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$. La varianza è $s_n^2 = (n-1)^{-1}\sum_{i=1}^n x_i^2 - n/(n-1)\bar{x}_n^2$. I dati richiesti sono riassunti nella seguente tabella

AULA	A	B	C
Media campionaria	7.9569	7.9573	7.6751
Varianza campionaria	3.70337	4.0562	4.7241

2. Si sa che gli estremi dell'intervallo sono

$$\begin{aligned} x^{\pm} &= \bar{x}_{78} \pm t_{(1+0.95)/2}(77)\cdot s_{78}/\sqrt{78} = 7.9569 \pm 1.99125\cdot\sqrt{3.70337/78} \\ &= 7.9569 \pm 0.43389 \\ &= \begin{cases} 7.523 \\ 8.39078 \end{cases} \end{aligned}$$

3. Chiamiamo X e Y le variabili relative alle aule A e B rispettivamente. Osserviamo che i dati sono sufficientemente numerosi per applicare i test per le variabili normali. Per avere evidenze statistiche che $\mu_X \neq \mu_Y$ si utilizza un test di confronto tra due medie a varianze incognite che si suppongono uguali. Si sceglie dunque $H_0 : \mu_X = \mu_Y$ contro $H_1 : \mu_X \neq \mu_Y$. Si osservi che i gradi di libertà $n+m-2 = n_X + n_Y - 2 = 150 > 120$ pertanto si possono approssimare i quantili della t-student con quelli di una variabile normale standard.

   La statistica per il test è
$$t = \frac{\bar{x}_n - \bar{y}_m}{s\sqrt{1/n+1/m}} = -0.001166$$
   essendo la varianza combinata $s^2 = ((n-1)s_X^2 + (m-1)s_Y^2)/(n+m-2) = 3.87508$ ($n = 78$ ed $m = 74$). Il P-value si calcola, in maniera interpolata, come $\bar{\alpha} = 2(1-\Phi(|t|)) = 0.99907$ essendo $\Phi(|t|) = 0.500465$. La zona di rifiuto a livello $\alpha = 0.05$ è $|t| > t_{1-\alpha/2}(n+m-2) = t_{0.975}(150) \approx q_{0.975} = 1.95996$. Sia utilizzando la regione di rifiuto, che semplicemente osservando che $\bar{\alpha} > 0.05$ si conclude che non vi sono evidenze statistiche per affermare che i due valori attesi siano differenti.

   Chiamiamo ora X e Y le variabili relative alle aule A e C rispettivamente. Osserviamo ancora che i dati sono sufficientemente numerosi per applicare i test per le variabili normali. Per avere evidenze statistiche che $\mu_X > \mu_Y$ si utilizza un test di confronto tra due medie a varianze incognite che si suppongono uguali. Si sceglie dunque $H_0 : \mu_X \leq \mu_Y$ contro $H_1 : \mu_X > \mu_Y$. Si osservi che, come in precedenza, i gradi di libertà $n+m-2 = n_X + n_Y - 2 = 173 > 120$ pertanto si possono approssimare i quantili della t-student con quelli di una variabile normale standard.

   La statistica per il test è sempre
$$t = \frac{\bar{x}_n - \bar{y}_m}{s\sqrt{1/n+1/m}} = 0.89667$$
   essendo la varianza combinata $s^2 = ((n-1)s_X^2 + (m-1)s_Y^2)/(n+m-2) = 4.26979$ ($n = 78$ ed $m = 97$). Il P-value si calcola, in maniera interpolata, come $\bar{\alpha} = 1-\Phi(t) = 0.184947$ essendo $\Phi(t) = 0.81505284$. La zona di rifiuto a livello $\alpha = 0.01$ è $t > t_{1-\alpha}(n+m-2) = t_{0.99}(173) \approx q_{0.99} = 2.32635$. Sia utilizzando la regione di rifiuto, che semplicemente osservando che $\bar{\alpha} > 0.05$ si conclude che non vi sono evidenze statistiche per affermare che l'aula C sia stata penalizzata.

4. Se il docente non vuole correre il rischio di penalizzare l'aula C nel caso la temperatura abbia avuto effetti negativi significativi, vuol dire che vorrebbe evitare di credere che la temperatura non abbia avuto effetti negativi nel caso gli effetti ci siano stati. In altre parole si vuole vedere se ci sono evidenze statistiche per concludere che la temperatura non abbia penalizzato gli studenti dell'aula C. Il test é analogo al caso precedente, ma con ipotesi invertite: $H_0 : \mu_X \geq \mu_Y$ contro $H_1 : \mu_X < \mu_Y$. La statistica è identica al caso precedente. Il P-value si calcola, in maniera interpolata, come $\bar{\alpha} = \Phi(t) = 0.81505284$. La zona di rifiuto a livello $\alpha = 0.1$ è $t < t_\alpha(n+m-2) = t_{0.1}(730) \approx -q_{0.9} = -1.2815516$. Sia utilizzando la regione di rifiuto, che semplicemente osservando che $\bar{\alpha} > 0.1$ si conclude che non vi sono evidenze statistiche per affermare che l'aula C NON sia stata penalizzata.

 12957

Esercizio 2 Un dado viene lanciato 100 volte ottenendo i seguenti risultati

n	1	2	3	4	5	6
$f_{ass}(n)$	12	20	17	15	17	19

1. È plausibile che il dado sia truccato ad un livello di significatività del 10%?
2. Stimare (specificando un intervallo) il P-value del precedente test.
3. Supponiamo ora che il dado venga lanciato 1000 volte e si ottengano gli stessi risultati del caso precedente moltiplicati per 10. È plausibile che il dado sia truccato ad un livello di significatività del 5%?
4. Stimare il P-value del test precedente.
5. Calcolare un intervallo di confidenza al 99% per la probabilità che esca 1 nel caso dei 1000 lanci descritto nei due punti precedenti. Quanti lanci occorrono come minimo per essere certi che l'intervallo di confidenza abbia ampiezza non superiore a 10^{-2}?

Soluzione. La regione di rifiuto del test di adattamento a livello α è $Q > \chi^2_{1-\alpha}(k-1)$ dove k è il numero di classi coinvolte e

$$Q = n \sum_{i^1}^{k} \frac{(p_i - f_{ass}(i)/n)^2}{p_i} = n \sum_{i^1}^{k} \frac{(p_i - f_{rel}(i))^2}{p_i}$$

(p_i probabilità teorica di appartenenza di un dato alla classe i-esima ed n ampiezza del campione). Nel caso specifico $p_i = 1/6$, $k = 6$ mentre $n = 100$ nel primo caso e $n = 1000$ nel secondo. Il P-value del test risolve l'equazione $Q = \chi^2_{1-\bar{\alpha}}(k-1)$. Il tast è applicabile in quanto $np_i \geq 5$.

1. Utilizzando i dati a disposizione si ottiene $q = 2.48$. Dalle tabelle $\chi^2_{0.9}(5) = 9.236349 > q$ quindi non si può rifiutare l'ipotesi nulla; pertanto non ci sono evidenze statistiche per affermare che il dado sia truccato.

2. Dalle disuguaglianze $\chi^2_{0.9}(5) = 9.236349 > \chi^2_{1-\bar{\alpha}}(5) \equiv 2.48 > 1.610309 = \chi^2_{0.1}(5) = 9.236349$ e dalla monotonia del quantile si ha $0.9 > 1 - \bar{\alpha} > 0.1$ cioè $0.9 > \bar{\alpha} > 0.1$. Il calcolatore restituirebbe un valore $\bar{\alpha} \simeq 0.779504$.

3. Osserviamo che $f_{rel}(i) = f_{ass}(i)/n$ sono identiche al caso precedente (tutte le grandezze assolute sono state moltiplicate per 10. Rimane il fattore moltiplicativo n prima della somma che definisce Q; questo comporta che la stima q risulta pari a quella precedente moltiplicata per 10, quindi $q = 24.8$. Essendo stavolta $\chi^2_{0.9}(5) = 9.236349 < q$ ci sono ragioni sufficienti per affermare (conclusione forte) che il dado è truccato.

4. Dalla disuguaglianza $\chi^2_{0.999}(5) = 20.51465 < 24.8 \equiv \chi^2_{1-\bar{\alpha}}(5)$ si ottiene $0.999 < 1 - \bar{\alpha}$ cioè $\bar{\alpha} < 0.001$. È un P-value molto basso, indice di una ipotesi nulla poco plausibile. Il calcolatore restituirebbe un valore $\bar{\alpha} \simeq 0.0001523$.

5. Il numero di volte che esce 1 è descritto da una Bernoulli di parametro p incognito. La stima di p è $\bar{p}_{1000} = 120/1000$. Si osservi che $n\bar{p}_{1000} \geq 5$ e $n(1 - \bar{p}_{1000}) \geq 5$ pertanto l'approssimazione gaussiana è buona. Un'intervallo di confidenza a livello $\alpha = 0.99$, ha come estremi

$$\begin{aligned} p^{\pm} &= \bar{p}_{1000} \pm z_{(1+0.99)/2} \cdot \sqrt{\frac{\bar{p}_{1000}(1 - \bar{p}_{1000})}{1000}} = 0.12 \pm 2.5758293035 \cdot \sqrt{0.0001056} \\ &= 0.12 \pm 2.5758293035 \cdot 0.0102761861 = \begin{cases} 0.1464697012 \\ 0.0935302988 \end{cases} \end{aligned}$$

L'ampiezza massima di un intervallo di confidenza a livello α di una Bernoulli si calcola come

$$2 \cdot z_{(1+\alpha)/2} \sqrt{\bar{p}(1 - \bar{p})/n} \leq 2 \cdot z_{(1+\alpha)/2} \sqrt{1/4n} = z_{(1+\alpha)/2} / \sqrt{n}$$

da cui $n \geq (z_{0.975}/10^{-2})^2 = 66348.9660102121$ cioè $n \geq 66349$.