Confronto fra varianze campionarie nel campionamento da popolazione normale

Sia $(X_{11}, X_{12}, \cdots, X_{1n_1})$ un campione casuale da una v.c. normale di aspettativa μ_1 e varianza σ_1^2. Sia $(X_{21}, X_{22}, \cdots, X_{2n_2})$ un campione casuale da una v.c. normale di aspettativa μ_2 e varianza σ_2^2. Siano fra loro indipendenti (cioè provengano da popolazioni differenti).

Si consideri il seguente problema di verifica d'ipotesi.

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

E' noto che

$$\frac{(n_i - 1)S_i^2}{\sigma_i^2} \quad \chi^2(n_i - 1) \qquad i = 1,2.$$

Sotto H_0, indicato con σ^2 il valore (non noto) comune della varianza delle due popolazioni, vale:

$$V = \frac{\frac{(n_1-1)S_1^2}{\sigma^2} \Big/ (n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma^2} \Big/ (n_2-1)} = \frac{S_1^2}{S_2^2} \quad F[(n_1-1);(n_2-1)].$$

Appare ragionevole rifiutare H_0 se

$$\frac{S_1^2}{S_2^2} \notin \left\{ F_{\frac{\alpha}{2}}[(n_1-1);(n_2-1)]; F_{1-\frac{\alpha}{2}}[(n_1-1);(n_2-1)] \right\}.$$

Osservazione

Conviene considerare come primo campione quello con varianza campionaria maggiore. *(Segue esempio)*

Distribuzione campionaria della differenza fra due frequenze relative campionarie

Sia $(X_{11}, X_{12}, \cdots, X_{1n})$ un campione casuale da una v.c. indicatore di parametro p_1. Sia $(X_{21}, X_{22}, \cdots, X_{2m})$ un campione casuale da una v.c. indicatore di parametro p_2. Siano fra loro indipendenti (cioè provengano da popolazioni differenti).
Le frequenze campionarie saranno:

$$\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i} \qquad \hat{p}_2 = \frac{1}{m}\sum_{i=1}^{m} X_{2i}\,.$$

$$D = \hat{p}_1 - \hat{p}_2 \qquad E(D) = p_1 - p_2$$

$$Var(D) = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$$

Al divergere di *m* e *n*,

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}}$$

tende in distribuzione alla v.c. normale standardizzata. Pertanto l'intervallo di confidenza asintotico al $100(1-\alpha)\%$ per $(p_1 - p_2)$ ha per estremi i valori

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{m}}$$

Sotto l'ipotesi statistica $H_0 : p_1 = p_2$ la statistica

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{m}}}$$

tende a distribuirsi, al divergere di m e n, come una v.c. normale standardizzata. Pertanto la regione critica del test è data da

$$\frac{\left|\hat{p}_1 - \hat{p}_2\right|}{\sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{m}}} > z_{1-\frac{\alpha}{2}}$$

Si noti che sotto $H_0 : p_1 = p_2 = p$ la stima della varianza di D è pari a

$$\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n} + \frac{1}{m}\right),$$

dove $\hat{p} = \frac{n\hat{p}_1 + m\hat{p}_2}{n+m} = \frac{\sum_{i=1}^{n} x_{1i} + \sum_{i=1}^{m} x_{2i}}{n+m}$

Pertanto

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n}+\frac{\hat{p}_2(1-\hat{p}_2)}{m}}}=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n}+\frac{1}{m}\right)}}$$

Le approssimazioni si ritengono soddisfacenti qualora ciascuno dei due campioni abbia almeno 100 osservazioni.

Esempio

Ad un concorso pubblico consegnano la prova scritta 1258 uomini e 1142 donne. Per farsi un'idea preliminare, si correggono, scegliendoli a caso, 112 scritti svolti da uomini, dei quali 53 sufficienti; si correggono, poi, 103 scritti, scegliendoli a caso, svolti da donne, dei quali 51 sufficienti. Si può ritenere che la frequenza relativa dei sufficienti sia la medesima per gli uomini e per le donne?

Soluzione

$$\hat{p}_u = \frac{53}{112} = 0{,}4732 \qquad \hat{p}_d = \frac{51}{103} = 0{,}4951$$

$$\frac{|0{,}4732 - 0{,}4951|}{\sqrt{\dfrac{0{,}4732 \cdot 0{,}5268}{112} + \dfrac{0{,}4951 \cdot 0{,}5049}{103}}} = \frac{0{,}0219}{0{,}0682} =$$

$= 0{,}3211 = z_{0,6255} \quad \Rightarrow \quad$ non si respinge l'ipotesi che la frequenza relativa di sufficienti sia la medesima per uomini e donne.

Esercizio 1

Il numero potenziale di utenti di un nuovo servizio di una struttura locale viene stimato pari a un migliaio di individui. Fra essi se ne scelgono a caso 78, ai quali viene presentato il nuovo servizio. Di essi 16 dichiarano di gradire il servizio.

a) Si può ritenere che la frazione della popolazione (dei potenziali utenti) che gradisce il servizio superi il 20%?

b) Si determini l'intervallo di confidenza al 97% per la proporzione di cui al punto a)

Soluzione

a) $\begin{array}{l} H_0 : p \geq 0{,}2 \\ H_1 : p < 0{,}2 \end{array}$

$$\hat{p} = \frac{13}{78} = 0{,}1\overline{6}.$$

Si rifiuta H_0 se

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} < -z_{1-\alpha}.$$

$$\frac{0{,}1\overline{6} - 0{,}2}{\sqrt{0{,}2 \cdot 0{,}8/78}} = -0{,}736 < -z_{1-\alpha}$$

$$0{,}74 = z_{0{,}7704} \Rightarrow \alpha^* = 0{,}2296$$

Per $\alpha \leq 0{,}2296$ accetto H_0.

b) *Per l'intervallo di confidenza su una proporzione si ritiene soddisfacente l'approssimazione alla normale per* $n \geq 50$.

L'intervallo di confidenza ha per estremi i valori

$$\hat{p} \mp z_{0,985}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0,1\overline{6} \mp 2,17\sqrt{\frac{0,1\overline{6}\cdot 0,8\overline{3}}{78}}$$

$[0,075;0,2582]$.

Esercizio 2

Su 35 lanci di una moneta 14 volte esce testa. Si può ritenere regolare la moneta?

$$H_0 : p = 0,5$$

$$H_1 : p \neq 0,5$$

$$\hat{p} = \frac{14}{35} = 0,4$$

Si rifiuta H_0 se

$$\frac{|\hat{p} - p_0|}{\sqrt{p_0(1-p_0)/n}} > z_{1-\frac{\alpha}{2}}$$

$$\frac{|0,4 - 0,5|}{\sqrt{0,5\cdot 0,5/35}} = 1,18 = z_{0,8810} \Rightarrow \alpha^* = 2\cdot 0,1190$$

Per $\alpha \leq 0,238$ si accetta H_0. La moneta non può ritenersi, in base ai (pochi) dati, non regolare.

Esercizio 3

Si sono rilevati i tempi (in secondi) di produzione di alcune viti speciali scelte a caso prodotte da due impianti, *A* e *B*, ottenendo le seguenti sintesi:

$$\sum_{i=1}^{16} x_{Ai} = 232 \quad \sum_{i=1}^{16} x_{Ai}^2 = 3716$$

$$\sum_{i=1}^{13} x_{Bi} = 200{,}2 \quad \sum_{i=1}^{13} x_{Bi}^2 = 3408{,}08.$$

a) Si verifichi l'ipotesi che le varianze dei tempi di produzione per i due impianti siano uguali, con probabilità d'errore di prima specie del 10%

b) Si verifichi l'ipotesi che i tempi medi di produzione coincidano per i due impianti, con probabilità d'errore di prima specie pari al 5%

Soluzione

a) $$H_0 : \sigma_A^2 = \sigma_B^2$$
$$H_1 : \sigma_A^2 \neq \sigma_B^2$$

$\alpha = 0{,}1$

$$\bar{x}_A = \frac{232}{16} = 14{,}5$$

$$\sum\left(x_{Ai} - \bar{x}_A\right)^2 = \sum x_{Ai}^2 - n_A \bar{x}_A^2 = 3716 - 16 \cdot 14{,}5^2 =$$

$$= 352 \Rightarrow S_A^2 = \frac{352}{16-1} = 23{,}4\bar{6}$$

$$\bar{x}_B = \frac{200{,}2}{13} = 15{,}4$$

$$\sum\left(x_{Bi} - \bar{x}_B\right)^2 = \sum x_{Bi}^2 - n_B \bar{x}_B^2 = 3408{,}08 - 13 \cdot 15{,}4^2 =$$

$$= 325 \Rightarrow \frac{325}{13-1} = 27{,}08\bar{3}$$

Rifiuto H_0 se

$$1{,}1541 = \frac{27{,}08\bar{3}}{23{,}4\bar{6}} > F_{0,95}\left(12{,}15\right) = 2{,}48.$$

Accetto l’ipotesi di uguaglianza delle varianze dei tempi di produzione dei due impianti.

b) Si rifiuta H_0 se

$$\frac{\left|\bar{X}_1 - \bar{X}_2\right|}{S\sqrt{\dfrac{n_1 + n_2}{n_1 \cdot n_2}}} > t_{1-\frac{\alpha}{2}}\left(n_1 + n_2 - 2\right)$$

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{352 + 325}{16 + 13 - 2} = 25,\overline{074}$$

$$S = \sqrt{25,\overline{074}} = 5,0074$$

Poiché

$$\frac{|14,5 - 15,4|}{5,0074\sqrt{\frac{29}{13 \cdot 16}}} = 0,4814 < t_{0,975}(27)$$

non rifiuto l'ipotesi di uguagliaza dei tempi medi di produzione.

Esercizio 4

In tabella è data la distribuzione del reddito annuo lordo di un campione casuale di 200 lavoratori dipendenti di un settore, per nell'area Nielsen del nord-est (Italia):

Reddito annuo lordo dichiarato	*Frequenze*
Meno di 10	17
$10 \vdash 20$	53
$20 \vdash 25$	46
$25 \vdash 35$	34
$35 \vdash 55$	31
$55 \vdash 75$	14
Almeno 75	5

Per un campione casuale di 150 individui occupati nello stesso settore nell'area Nielsen sud e isole, si sono rilevate le seguenti sintesi:

$\sum x_{Si} = 4101 \qquad \sum x_{Si}^2 = 120971.$

a) Si può ritenere che il reddito medio delle due aree geografiche sia il medesimo?

b) Si determini l'intervallo di confidenza al 93% per la differenza dei redditi medi delle due aree.

Soluzione

x_{c_i}	n_i	$x_{c_i} n_i$	$x_{c_i}^2 n_i$
7,5	17	127,5	956,25
15	53	795	11925
22,5	46	1035	23287,5
30	34	1020	30600
45	31	1395	62775
65	14	910	59150
87,5	5	437,5	38281,25
	200	6515	226975

$$\bar{x}_{Ne} = \frac{6515}{200} = 32,575$$

$$S_{Ne}^2 = \frac{226975 - 200(32,575)^2}{199} = \frac{14748,875}{199} = 74,1150$$

$$\bar{x}_S = \frac{4101}{150} = 27,34$$

$$S_S^2 = \frac{120971 - 150(27,34)^2}{149} = \frac{8849,66}{149} = 59,3937$$

$$S^2 = \frac{14748,875 + 8849,66}{200 + 150 - 2} = 67,8119 \Rightarrow S = 8,2348$$

Si rifiuta l'ipotesi nulla se vale

$$\frac{|\bar{X}_1 - \bar{X}_2|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_{1-\frac{\alpha}{2}}$$

$$\frac{|32,575 - 27,34|}{8,2348\sqrt{\frac{1}{200} + \frac{1}{150}}} = 5,8856$$

Si rifiuta l'ipotesi di uguaglianza dei redditi medi nelle due aree (per qualunque alfa).

Il test di Pizzetti-Pearson

Si considerino k eventi incompatibili ed esaustivi $A_1, A_2, \cdots, A_k$. L'ipotesi nulla H_0 assegna a detti eventi le probabilità $\pi_1, \pi_2, \cdots, \pi_k$: $\pi_i > 0 \quad i = 1, \cdots, k$ e $\sum_{i=1}^{k} \pi_i = 1$.

Si estrae un campione casuale di ampiezza n, del quale si rileva il numero di casi n_i con cui si verifica l'evento A_i. E' chiaro che vale $\sum_{i=1}^{k} n_i = n$.

Si consideri la seguente misura dovuta a Pizzetti-Pearson:

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i} \equiv \sum_{i=1}^{k} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}.$$

E' intuitivo che più grande è il valore di X^2, più si accumula evidenza empirica contro l'ipotesi nulla H_0.

Teorema

Sotto l'ipotesi

$H_0 : (p_1 = \pi_1) \cap \cdots (p_k = \pi_k),$

la statistica

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}$$

al divergere di n tende a distribuirsi secondo una v.c. chi-quadrato con $(k-1)$ g.d.l..

Pertanto, un test ragionevole prescrive di rifiutare H_0 se

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i} > \chi^2_{1-\alpha}(k-1).$$

Osservazione importante

Si ha una buona approssimazione alla distribuzione chi-quadrato se ogni frequenza assoluta osservata n_i risulta almeno pari a 5. Ciò significa che, qualora ciò non avvenisse per qualche i, è necessario unire l'evento A_i in modo da ottenere che tutte le frequenze osservate siano non minori di 5.

L'ipotesi nulla specifica il tipo di modello ma non i valori dei parametri (l'ipotesi nulla è composta)

Teorema

Si supponga che l'ipotesi H_0 preveda un modello per il quale bisogna stimare s parametri. Allora la v.c. $X^2 = \sum_{i=1}^{k} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$,

Al divergere di n, tende a distribuirsi come una v.c. chi-quadrato con $(k-1-s)$ g.d.l..

Stima dei parametri

- Modello normale:

$\hat{\mu} = \bar{X} \qquad \hat{\sigma}^2 = S^2$

- Modello di Poisson:

$\hat{\lambda} = \bar{X}$

- Modello esponenziale:

$\hat{\vartheta} = \frac{1}{\bar{X}}$

- Per altri modelli, gli stimatori saranno suggeriti nel testo dell'esercizio.

Esempio 1

Si è osservato un macchinario in un impianto industriale per 40 ore di lavoro. Per ogni ora di lavoro si sono registrati i seguenti valori:

Numero guasti x_i	*Frequenza* n_i
0	0
1	6
2	8
3	11
4	7
5	4
6	3
7	1
Totale	40

Si verifichi l'ipotesi che il numero *X* di guasti calcolato nelle 40 ore si distribuisca secondo la v.c. di Poisson a livello $\alpha = 0{,}05$.

Soluzione

x_i	n_i	$x_i \cdot n_i$	$\hat{\pi}_i$	$\hat{n}_i = n\hat{\pi}_i$
0	0	0	0,0408	1,6
1	6	6	0,1304	5,2
2	8	16	0,2087	8,3
3	11	33	0,2226	8,9
4	7	28	0,1781	7,1
5	4	20	0,1140	4,6
6	3	18	0,0608	2,4
7	1	7	0,0278	1,1
	40	128		39,9

$$\hat{\lambda} = \overline{X} = \frac{128}{40} = 3{,}2$$

x_i	n_i	$n_i - \hat{n}_i$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
0	0	-	-
1	6	$6-6{,}8$	0,094
2	8	$8-8{,}3$	0,011
3	11	$11-8{,}9$	0,496
4	7	$7-7{,}1$	0,001
5 e oltre	4	$8-8{,}8$	0,073
	40		0,675

$$\hat{\pi}(5\ e\ oltre) = 1 - \sum_{j=0}^{4} \hat{\pi}_j$$

Poichè $X^2 = 0{,}675$ e
$\chi^2_{0,95}(5-1-1) = \chi^2_{0,95}(3) = 7{,}81,$
l'ipotesi nulla non può essere rifiutata.

Esempio 2 Da un esperimento, sono stati osservati i seguenti valori:

x_i	1	2	3	4	5	6	7	8	≥ 9
n_i	382	262	143	84	53	28	20	12	16

Tenendo conto che la somma dei valori osservati nella classe *9 e oltre* è 160, si valuti ($\alpha = 0{,}1$) la bontà di adattamento del seguente modello (di Pascal):

$$p(x;\vartheta) = \vartheta(1-\vartheta)^{x-1} \qquad x = 1, 2, \cdots$$

(Si tenga conto che $E(X) = \dfrac{1}{\vartheta}$).

Soluzione

x_i	n_i	$x_i \cdot n_i$	$\hat{\pi}_i$	$\hat{n}_i = n\hat{\pi}_i$
1	382	382	0,4	400
2	262	524	0,24	240
3	143	429	0,144	144
4	84	336	0,087	87
5	53	265	0,052	52
6	28	168	0,031	31
7	20	140	0,019	19
8	12	96	0,011	11
≥ 9	16	160	0,016	16
	1000	2500		

$$\hat{\vartheta} = \frac{1}{\overline{X}} = \frac{1}{2500/1000} = \frac{1}{2{,}5} = 0{,}4.$$

x_i	n_i	$\hat{n}_i = n\hat{\pi}_i$	$n_i - \hat{n}_i$	$\dfrac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
1	382	400	-18	0,81
2	262	240	+22	$2{,}01\overline{6}$
3	143	144	-1	0,0069
4	84	87	-3	0,1034
5	53	52	+1	0,0192
6	28	31	-3	0,2903
7	20	19	+1	0,0526
8	12	11	+1	0,0909
≥ 9	16	16	0	0
	1000	1000		3,3900

$$\chi^2_{0,9}(9-1-1) = 12{,}0$$

Non rifiuto l'ipotesi nulla.

Impiego della statistica X^2 per la verifica dell'indipendenza fra due caratteri

Si consideri un campione di *n* unità statistiche classificate secondo due caratteri *X* e *Y*, di qualunque natura.

$Y \mid X$	x_1	x_2	x_j	x_c	
y_1					$n_{1.}$
y_2					$n_{2.}$
y_i			n_{ij}		$n_{i.}$
y_r					$n_{r.}$
	$n_{.1}$	$n_{.2}$	$n_{.j}$	$n_{.c}$	n

Sia $\pi_i(Y)$ la probabilità che la v.c. *Y* assuma modalità y_i (nel caso di carattere riclassificato, che la v.c. *Y* assuma valore nell'intervallo *i*-esimo).

Sia $\pi_j(X)$ la probabilità che la v.c. *X* assuma modalità x_j (nel caso di carattere riclassificato, che la v.c. *X* assuma valore nell'intervallo *j*-esimo).

L'ipotesi nulla H_0 preveda che i due caratteri siano indipendenti, cioè che:

$$P\{X = x_j \cap Y = y_i\} = \pi_j(X) \cdot \pi_i(Y)$$

$$i = 1,2,\cdots,r; \quad j = 1,2,\cdots,c.$$

Le probabilità marginali sono stimate attraverso le frequenze relative campionarie. In formule:

$$\hat{\pi}_j(X) = \frac{n_{.j}}{n} \qquad j = 1,2,\cdots,c$$

$$\hat{\pi}_i(Y) = \frac{n_{i.}}{n} \qquad i = 1,2,\cdots,r.$$

Sotto l'ipotesi nulla le probabilità congiunte stimate sono date da:

$$\hat{\pi}_j(X) \cdot \hat{\pi}_i(Y) = \frac{n_{.j}}{n} \cdot \frac{n_{i.}}{n}.$$

Pertanto, le frequenze assolute teoriche in ipotesi di indipendenza sono stimate da:

$$\hat{n}_{ij} = n\hat{\pi}_j(X) \cdot \hat{\pi}_i(Y) = n\frac{n_{.j}}{n} \cdot \frac{n_{i.}}{n} = \frac{n_{.j} \cdot n_{i.}}{n}$$

Per verificare l'indipendenza fra i due caratteri, si utilizza la statistica

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-\hat{n}_{ij}\right)^2}{\hat{n}_{ij}} = \left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{\hat{n}_{ij}}\right) - n.$$

Teorema

Sotto l'ipotesi nulla d'indipendenza distributiva, X^2, al divergere di n, tende a distribuirsi secondo una v.c. chi-quadrato con $(r-1)\cdot(c-1)$ g.d.l..

Esempio 1

Il Dipartimento di Statistica di una Facoltà che vanta un elevato numero di iscritti desidera stabilire se vi sia o meno un rapporto fra interesse degli studenti per la statistica (carattere X) e abilità per la matematica (carattere Y). Viene scelto un campione casuale di 200 studenti a cui viene chiesto se il loro interesse per la statistica e l'abilità per la matematica sia alto, medio o basso. I risultati sono stati riassunti nella seguente tabella:

$X \mid Y$	Bassa	*Media*	*Alta*	*Totale*
Basso	60	15	15	90
Medio	15	45	10	70
Alto	5	10	25	40
Totale	80	70	50	200

Ad un livello di significatività del 5%, si può affermare che fra i due caratteri rilevati vi sia indipendenza in probabilità?

Soluzione

$\hat{n}_{ij}$	B	*M*	*A*
B	36	31,5	22,5
M	28	24,5	17,5
A	16	14	10

$n_{ij} - \hat{n}_{ij}$	B	*M*	*A*
B	+24	-16,5	-7,5
M	-13	+20,5	-7,5
A	-11	-4	+15

$\frac{\left(n_{ij}-\hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$	B	*M*	*A*
B *M* A	16 6,03 7,56	8,64 17,15 1,14	2,5 3,21 22,5

$$X^2 = \sum\sum \frac{\left(n_{ij}-\hat{n}_{ij}\right)^2}{\hat{n}_{ij}} = 84{,}73 < \chi^2_{0,95}(2) = 9{,}49$$

I due caratteri possono considerarsi indipendenti.

Esempio 2

Il numero di confezioni vendute da un negozio alimentare sono riportati nella tabella che segue, classificati secondo la tipologia *X* e il paese di provenienza *Y*:

Y | X	Cereali	Carne	Latte e *Derivati*
Francia *Germania* *Italia* *Spagna*	79 74 117 71	107 102 84 83	92 88 77 102

Ad un livello di significatività del 5% si può affermare che tra i due caratteri rilevati vi sia indipendenza in distribuzione?

Soluzione.

	Cereali	Carne	Latte e Derivati	*Totale*
	79 74 117 71	107 102 84 83	92 88 77 102	278 264 278 256
Totale	*341*	376	359	1076

$\hat{n}_{ij}$	Cereali	Carne	Latte e Derivati
	88,10 83,67 88,10 81,13	97,14 92,25 97,14 89,46	92,75 88,08 92,75 85,41

$\hat{n}_{ij} - \hat{n}_{ij}$	Cereali	Carne	Latte e Derivati	
	-9,1 -9,67 +28,9 -10,13	+9,86 +9,75 -13,14 -6,46	-0,75 -0,08 -15,75 +16,59	0,01 0 0,01 0
	0	0,01	0,01	

$\frac{\left(\hat{n}_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$			
	0,9400 1,1176 9,4802 1,2648	1,0008 1,0305 1,7774 0,4665	0,0060 0,0001 2,6745 3,2224

$$X^2 = \sum\sum \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}} = 22{,}9808 > \chi^2_{0,95}(6) = 14{,}4$$

Si rifiuta l'ipotesi d'indipendenza.

Esempio 3

Si è osservato il seguente fenomeno:

Classi	*Frequenze*
0 ⊢— 0,5	24
0,5 ⊢— 1	16
1 ⊢— 1,5	13
1,5 ⊢— 2	14
2 ⊢— 3	19
3 ⊢— 4	13
4 ⊢— 5	9
5 e oltre	17

Sapendo che la somma dei valori della classe *5 e oltre* è pari a 102, si valuti se il fenomeno è ben descritto dalla distribuzione esponenziale, con probabilità d'errore di prima specie pari all'1%.

Soluzione

x_{ci}	n_i	$x_i \cdot n_i$
0,25	24	6
0,75	16	12
1,25	13	16,25
1,75	14	24,5
2,5	19	47,5

3,5	13	45,5
4,5	9	40,5
-	17	102
	125	294,25

$$\overline{X} = \frac{294{,}25}{125} = 2{,}354$$

$$\hat{\vartheta} = \frac{1}{\overline{X}} = 0{,}4248$$

$$P\{X \in [a;b)\} = e^{-\hat{\vartheta}a} - e^{-\hat{\vartheta}b} = e^{-0{,}4248a} - e^{-0{,}4248b}$$

Classi	*Frequenze*	$\hat{\pi}_i$	$\hat{n}_i = n\hat{\pi}_i$
0 $\vert$— 0,5	24	0,1914	23,925
0,5 $\vert$— 1	16	0,1547	19,3375
1 $\vert$— 1,5	13	0,1252	15,65
1,5 $\vert$— 2	14	0,1011	12,6375
2 $\vert$— 3	19	0,1480	18,5
3 $\vert$— 4	13	0,0968	12,1
4 $\vert$— 5	9	0,0633	7,9125
5 e oltre	17	0,1195	14,9375
	125	1,0000	

n_i	$\hat{n}_i$	$n_i - \hat{n}_i$	$\frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$
24	23,925	+0,075	0,0002
16	19,3375	-3,3375	0,5760
13	15,65	-2,65	0,4872
14	12,6375	+1,3625	0,1469
19	18,5	+0,5	0,0135
13	12,1	+0,9	0,0669
9	7,9125	+1,0875	0,1495
17	14,9375	+2,0625	0,2848
	125		1,725

$\chi^2_{0,99}(8-1-1) = 16,8$

Non rifiuto l'ipotesi nulla: il modello esponenziale si adatta bene ai dati.

Esempio 1

Nella tabella che segue sono riportati i pesi in grammi di 15 confezioni di latte prodotte da tre macchinari posti su tre linee di confezionamento parallele:

Linea1	*Linea2*	*Linea3*
1045	990	1012
990	1045	979
1012	1111	1012
1023	1034	1023
1001		1034
1034		

a) Verificare l'ipotesi che le macchine producano confezioni con il medesimo peso medio, con un livello di significatività del 5%. Specificare sotto quali ipotesi sia possibile tale verifica.

b) Verificare l'ipotesi che le varianze dei pesi delle confezioni prodotte dalle linee 1 e 3 siano uguali contro l'alternativa che differiscano, volendo commettere l'errore di prima specie con probabilità del 5%.

Soluzione.

a)

1045 990 1012 1023 1001 1034	990 1045 1111 1034	1012 979 1012 1023 1034
6105	4180	5060

$$\alpha = 0{,}05 \quad H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \sum\sum\left|\mu_i - \mu_j\right| > 0$$

Devianza fra i gruppi:

$$D.F. = \sum_{j=1}^{3}\left(\overline{X}_j - \overline{X}\right)^2 \cdot n_j = (1017{,}5 - 1023)^2 \cdot 6 +$$

$$+ (1045 - 1023)^2 \cdot 4 + (1012 - 1023)^2 \cdot 5 =$$

$$= 181{,}5 + 1936 + 605 = 2722{,}5$$

Devianza nei gruppi:

$$\sum_{j=1}^{3}\sum_{i=1}^{n_j}\left(X_{ji}-\overline{X}_j\right)^2 = 2177,5+7502+1694=11373,5$$

$$V=\frac{D.F./(k-1)}{D.N./n-k}=\frac{2722,5/2}{11373,5/12}=1,436$$

$$F_{0,95}(2,12)=3,89$$

Essendo $V < F$ si accetta l'ipotesi di uguaglianza del peso medio.

b) $\alpha = 0,05$

$$H_0 : \sigma_1 = \sigma_3$$

$$H_1 : \sigma_1 \neq \sigma_3$$

$$V=\frac{S_1^2}{S_3^2}=\frac{2157,5}{1694}=1,27$$

$$F_{1-\frac{\alpha}{2}}(n_1-1, n_3-1)=F_{0,975}(5,4)=9,36.$$

Poiché $V < F$, si accetta l'ipotesi di uguaglianza delle varianze.

Esempio 2

Un'indagine sui consumi mensili di un certo prodotto X (in migliaia di lire) delle famiglie di una certa zona, classificate per titolo di studio del capofamiglia, ha dato luogo ai seguenti risultati:

	n_j	$\sum_{i=1}^{n_j} x_{ji}$	$\sum_{i=1}^{n_j} x_{ji}^2$
Lic. Elem.	7	37	198
Lic. Media	8	50	322
Diploma	10	64	411
Laurea	9	61	416

Dopo aver indicato le opportune ipotesi, si chiede di:

a) verificare, a livello di significatività del 5%, l'ipotesi che I consumi medi delle 4 tipologie di famiglie siano uguali e commentare il risultato;

b) Verificare, sempre a livello di significatività del 5%, l'ipotesi che I consumi medi delle famiglie dei diplomati e dei laureati siano uguali e commentare il risultato.

Soluzione

a)

$$k = 4; \quad n = 7 + 8 + 10 + 9 = 34$$

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{j1} = \frac{37}{7} = 5{,}286$$

$$\overline{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{j2} = \frac{50}{8} = 6{,}25$$

$$\overline{X}_3 = \frac{1}{n_3}\sum_{i=1}^{n_3} X_{j3} = \frac{64}{10} = 6{,}4$$

$$\overline{X}_4 = \frac{1}{n_4}\sum_{i=1}^{n_4} X_{j4} = \frac{61}{9} = 6{,}\overline{7}$$

$$\overline{X} = \frac{37 + 50 + 64 + 61}{34} = 6{,}235$$

$$D.F. = \sum_{j=1}^{k}\left(\overline{X}_j - \overline{X}\right)^2 = 6{,}304 + 0{,}002 + 0{,}272 + 2{,}651 = 9{,}229$$

$$D..N. = \sum_{j=1}^{k}\left(n_j - 1\right)S_j^2 = \sum_{j=1}^{k}\left(\sum_{i=1}^{n_j} x_{ji}^2 - n_j\overline{X}_j^2\right)$$

$$6 \cdot S_1^2 = 198 - 7 \cdot 5{,}286^2 = 2{,}429$$

$$7 \cdot S_2^2 = 322 - 8 \cdot 6{,}25^2 = 9{,}5$$
$$9 \cdot S_3^2 = 411 - 10 \cdot 6{,}4^2 = 1{,}4$$
$$8 \cdot S_4^2 = 416 - 9 \cdot 6{,}\overline{7} = 2{,}\overline{5}.$$
$$D.N. = 2{,}429 + 9{,}5 + 1{,}4 + 2{,}\overline{5} = 15{,}885$$
$$V = \frac{D.F./(k-1)}{D.N./(n-k)} = \frac{9{,}229/3}{15{,}885/30} = \frac{3{,}076}{0{,}530} = 5{,}804$$
$$F_{0,95}(4-1;34-4) = 2{,}92$$
Si rifiuta l'ipotesi di eguaglianza delle medie.

b)

Si rifiuta H_0 se
$$\frac{\left|\overline{X}_3 - \overline{X}_4\right|}{S\sqrt{\frac{n_3+n_4}{n_3 \cdot n_4}}} > t_{1-\frac{\alpha}{2}}(n_3+n_4-2)$$
$$S^2 = \frac{1{,}4+2{,}5}{10+9-2} = 0{,}233$$
$$\frac{\left|6{,}4-6{,}\overline{7}\right|}{\sqrt{0{,}233}\sqrt{\frac{19}{90}}} = 1{,}705 < t_{0,975}(17) = 2{,}11$$
Si accetta l'ipotesi di eguaglianza delle medie del terzo e del quarto gruppo.

Esempio 3

In una ditta di pneumatici viene gestita la produzione secondo tre turni. Il direttore della produzione intende verificare se esiste una differenza, per quanto concerne la duratamedia dei prodotti, tra le lavorazioni dei tre turni. A tale scopo viene selezionato un campione casuale di 15 pneumatici e ne viene misurata la durata in migliaia di km, dando luogo ai seguenti risultati:

Durate		
Turno di giorno	Turno di sera	Turno di notte
38,87	37,65	32,18
40,33	35,08	35,72
36,20	37,81	31,78
38,39	36,60	33,15
	34,75	32,82
	38,78	

a) si verifichi l'ipotesi che la durata dei pneumatici delle tre produzioni sia la medesima, con un livello di significatività dell'1%.

b) Si precisino le assunzioni che debbono essere poste sulle tre popolazioni dei pneumatici da cui sono stati estratti I tre campioni.

Soluzione

$$\overline{X}_1 = \frac{153{,}79}{4} = 38{,}4475$$

$$\overline{X}_2 = \frac{220{,}67}{6} = 36{,}778\overline{3}$$

$$\overline{X}_3 = \frac{165{,}65}{5} = 33{,}13$$

$$\overline{X} = \frac{153{,}79 + 220{,}67 + 165{,}65}{4 + 6 + 5} = 36{,}007\overline{3}$$

$(x_{1i}-\bar{x}_1)^2$	$(x_{2i}-\bar{x}_2)^2$	$(x_{3i}-\bar{x}_3)^2$
0,1785 3,5438 5,0513 0,0033	0,7569 2,89 1,0609 0,0324 4,1209 4	0,9025 6,7081 1,8225 0,0004 0,0961
8,7769	12,8611	9,5292

$$D.N. = 31{,}1672$$

$$D.F. = 23{,}8177 + 3{,}5666 + 41{,}3943 = 68{,}7786$$

$$V = \frac{D.F./(k-1)}{D.N./(n-k)} = \frac{68{,}7786/2}{31{,}1672/12} = \frac{34{,}2404}{2{,}5973} = 13{,}2404$$

$$F_{0,99}(2{,}12) = 6{,}93$$

Rifiuto l'ipotesi che le durate medie delle tre produzioni siano uguali.

b) *(a voce)*