ANALISI DELLA VARIANZA A UN CRITERIO DI CLASSIFICAZIONE

Consideriamo k popolazioni normali di uguale varianza σ^2 (ignota) e medie μ_j.
Vogliamo verificare l'ipotesi

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$$

contro l'alternativa

$$H_1 : \sum_{i=1}^{k} \sum_{j=1}^{k} \left| \mu_i - \mu_j \right| > 0$$

(che vi siano almeno due medie non uguali).
A tal fine estraiamo da ciascuna popolazione campioni indipendenti di numerosità n_j, $j = 1,...,k$, per un totale di

$$n = \sum_{j=1}^{k} n_j \text{ estrazioni.}$$

I dati campionari sono allora

$$\left(X_{11},\ldots,X_{1i},\ldots,X_{1n_1}\right)$$
$$\vdots$$
$$\left(X_{j1},\ldots,X_{ji},\ldots,X_{jn_j}\right)$$
$$\vdots$$
$$\left(X_{k1},\ldots,X_{ki},\ldots,X_{kn_k}\right)$$

Di ciascun campione calcoliamo le medie (tra loro indipendenti) $\overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ji}$ e le varianze campionarie corrette (tra loro indipendenti)

$$S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(X_{ji} - \overline{X}_j\right)^2 \quad \left(j = 1,\ldots,k\right).$$

Poiché i campioni sono estratti da popolazioni normali, vi è indipendenza anche tra le v.c. $\overline{X}_j$ e le v.c. S_j^2.

E' noto che

$$E\left(\overline{X}_j\right) = \mu_j, \quad Var\left(\overline{X}_j\right) = \frac{\sigma^2}{n_j},$$

$$E\left(S_j^2\right)=\sigma^2.$$

Distribuzioni campionarie delle devianze nei gruppi e fra i gruppi

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

$$E\left(D.N.\right)=\sum_{j=1}^{k}\left(n_j-1\right)E\left(S_j^2\right)=$$

$$=\sum_{j=1}^{k}\left(n_j-1\right)\sigma^2=\left(n-k\right)\sigma^2,$$

quindi sia sotto H_0 sia sotto H_1

$$E\left(\frac{D.N.}{n-k}\right)=\sigma^2.$$

$$\boxed{D.F.=\sum_{j=1}^{k}\left(\overline{X}_j-\overline{X}\right)^2 n_j}$$

dove

$$\overline{X}=\frac{1}{n}\sum_{j=1}^{k}\overline{X}_j\cdot n_j.$$

Allora

$$E\left(\overline{X}\right)=\frac{1}{n}\sum_{j=1}^{k}n_jE\left(\overline{X}_j\right)=\frac{1}{n}\sum_{j=1}^{k}\mu_jn_j=\overline{\mu}$$

dove con $\overline{\mu}$ indichiamo la media aritmetica ponderata delle medie parziali μ_j con pesi le numerosità campionarie. Inoltre, sia sotto H_0 sia sotto H_1,

$$Var\left(\overline{X}\right)=Var\left(\frac{1}{n}\sum_{j=1}^{k}\overline{X}_j\cdot n_j\right)=$$

$$=\frac{1}{n^2}\sum_{j=1}^{k}n_j^2\ Var\left(\overline{X}_j\right)=\frac{1}{n^2}\sum_{j=1}^{k}n_j^2\ \frac{\sigma^2}{n_j}=$$

$$=\frac{\sigma^2}{n^2}\sum_{j=1}^{k}n_j=\frac{\sigma^2}{n}.$$

E' nota la scomposizione

$$\sum_{j=1}^{k}\left(\overline{X}_j-\overline{X}\right)^2 n_j=\sum_{j=1}^{k}\left(\overline{X}_j-\overline{\mu}+\overline{\mu}-\overline{X}\right)^2 n_j=$$

$$=\sum_{j=1}^{k}\left(\overline{X}_j-\overline{\mu}\right)^2 n_j+\sum_{j=1}^{k}\left(\overline{\mu}-\overline{X}\right)^2 n_j+$$

$$+2\sum_{j=1}^{k}\left(\overline{X}_j-\overline{\mu}\right)\left(\overline{\mu}-\overline{X}\right)n_j=$$

$$=\sum_{j=1}^{k}\left(\overline{X}_j-\overline{\mu}\right)^2 n_j+\left(\overline{\mu}-\overline{X}\right)^2\sum_{j=1}^{k}n_j+$$

$$+2\left(\overline{\mu}-\overline{X}\right)\sum_{j=1}^{k}\left(\overline{X}_j-\overline{\mu}\right)n_j=$$

$$=\sum_{j=1}^{k}\left(\overline{X}_j-\overline{\mu}\right)^2 n_j+n\left(\overline{\mu}-\overline{X}\right)^2+$$

$$+2\left(\overline{\mu}-\overline{X}\right)\sum_{j=1}^{k}\left(\overline{X}_j n_j-\overline{\mu}n_j\right)=$$

$$=\sum_{j=1}^{k}\left(\overline{X}_j-\overline{\mu}\right)^2 n_j-n\left(\overline{\mu}-\overline{X}\right)^2=$$

$$=\sum_{j=1}^{k}\left(\overline{X}_j-\mu_j+\mu_j-\overline{\mu}\right)^2 n_j-n\left(\overline{\mu}-\overline{X}\right)^2=$$

$$= \sum_{j=1}^{k} \left(\overline{X}_j - \mu_j \right)^2 n_j + \sum_{j=1}^{k} \left(\mu_j - \overline{\mu} \right)^2 n_j +$$

$$+2\sum_{j=1}^{k} \left(\overline{X}_j - \mu_j \right)\left(\mu_j - \overline{\mu} \right) n_j - n\left(\overline{\mu} - \overline{X} \right)^2 .$$

Dunque

$$E\left(\sum_{j=1}^{k} \left(\overline{X}_j - \overline{X} \right)^2 n_j \right) =$$

$$= \sum_{j=1}^{k} n_j E\left(\overline{X}_j - \mu_j \right)^2 + \sum_{j=1}^{k} \left(\mu_j - \overline{\mu} \right)^2 n_j +$$

$$+2\sum_{j=1}^{k} \left(\mu_j - \overline{\mu} \right) E\left(\overline{X}_j - \mu_j \right) n_j - nE\left(\overline{\mu} - \overline{X} \right)^2 =$$

$$= \sum_{j=1}^{k} n_j Var\left(\overline{X}_j \right) + \sum_{j=1}^{k} \left(\mu_j - \overline{\mu} \right)^2 n_j +$$

$$+2\sum_{j=1}^{k} \left(\mu_j - \overline{\mu} \right) \cdot 0 \cdot n_j - nVar\left(\overline{X} \right) =$$

$$= \sum_{j=1}^{k} n_j \frac{\sigma^2}{n_j} + \sum_{j=1}^{k} \left(\mu_j - \overline{\mu}\right)^2 n_j - n\frac{\sigma^2}{n} =$$

$$= k\sigma^2 + \sum_{j=1}^{k} \left(\mu_j - \overline{\mu}\right)^2 n_j - \sigma^2 =$$

$$= (k-1)\sigma^2 + \sum_{j=1}^{k} \left(\mu_j - \overline{\mu}\right)^2 n_j .$$

Pertanto

$$E\left(\frac{1}{k-1} \sum_{j=1}^{k} \left(\overline{X}_j - \overline{X}\right)^2 n_j \right) =$$

$$= \sigma^2 + \frac{1}{k-1} \sum_{j=1}^{k} \left(\mu_j - \overline{\mu}\right)^2 n_j .$$

Osserviamo che

$$E\left(\frac{D.F.}{k-1} \right) = E\left(\frac{1}{k-1} \sum_{j=1}^{k} \left(\overline{X}_j - \overline{X}\right)^2 n_j \right) =$$

$$= \begin{cases} \sigma^2 & \text{sotto } H_0 \\ \sigma^2 + \dfrac{1}{k-1} \displaystyle\sum_{j=1}^{k} \left(\mu_j - \overline{\mu}\right)^2 n_j & \text{sotto } H_1 . \end{cases}$$

Consideriamo ora il rapporto

$$V = \frac{D.F. \big/ (k-1)}{D.N. \big/ (n-k)} = \frac{\dfrac{1}{k-1}\displaystyle\sum_{j=1}^{k}\left(\overline{X}_j - \overline{X}\right)^2 n_j}{\dfrac{1}{n-k}\displaystyle\sum_{j=1}^{k}\left(n_j - 1\right) S_j^2}.$$

Poiché

$$E\left(\frac{1}{n-k}\sum_{j=1}^{k}\left(n_j - 1\right) S_j^2\right) = \sigma^2$$

sia sotto H_0 sia sotto H_1, la v.c. V tenderà ad assumere valori più elevati sotto H_0 rispetto ad H_1. Utilizziamo allora V come variabile test, e pertanto cerchiamone la distribuzione.

Dato che i campioni provengono da popolazioni normali, la v.c.

$$\frac{\displaystyle\sum_{i=1}^{n_j}\left(X_{ji} - \overline{X}_j\right)^2}{\sigma^2} = \frac{\left(n_j - 1\right) S_j^2}{\sigma^2}$$

si distribuisce come una v.c. χ_1^2.

Per la proprietà additiva della v.c. chi-quadrato la v.c.

$$W=\frac{D.N.}{\sigma^2}=\sum_{j=1}^{k}\frac{\left(n_j-1\right)S_j^2}{\sigma^2}\sim\chi^2_{n-k}$$

Per quanto riguarda il numeratore di *V*, *s*otto l'ipotesi H_0 scomponiamo la D.F. come segue:

$$\sum_{j=1}^{k}\left(\overline{X}_j-\overline{X}\right)^2 n_j=\sum_{j=1}^{k}\left(\overline{X}_j-\mu\right)^2 n_j-n\left(\overline{X}-\mu\right)^2.$$

Da cui, dividendo ambo i membri per σ^2,

$$\sum_{j=1}^{k}\left(\frac{\overline{X}_j-\mu}{\frac{\sigma}{\sqrt{n_j}}}\right)^2=\sum_{j=1}^{k}\left(\frac{\overline{X}_j-\overline{X}}{\frac{\sigma}{\sqrt{n_j}}}\right)^2+\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2.$$

Sotto l'ipotesi H_0 la v.c.

$$\frac{\overline{X}_j-\mu}{\frac{\sigma}{\sqrt{n_j}}}\sim N\left(0;1\right)$$

quindi

$$\left(\frac{\overline{X}_j - \mu}{\frac{\sigma}{\sqrt{n_j}}} \right)^2 \sim \chi_1^2 .$$

Per la proprietà additiva della v.c. chi-quadrato, la sommatoria

$$\sum_{i=1}^{k} \left(\frac{\overline{X}_j - \mu}{\frac{\sigma}{\sqrt{n_j}}} \right)^2 = \sum_{i=1}^{k} \frac{\left(\overline{X}_j - \mu\right)^2}{\sigma^2} n_j$$

si distribuisce sotto H_0 come una v.c. chi-quadrato con k gradi di libertà.

Analogamente

$$\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)^2 = n \frac{\left(\overline{X} - \mu\right)^2}{\sigma^2}$$

si distribuisce sotto H_0 come una v.c. chi-quadrato con 1 grado di libertà.

Deriva che la v.c.

$$Y = \frac{D.F.}{\sigma^2} = \sum_{j=1}^{k} \frac{\left(\overline{X}_j - \overline{X}\right)^2}{\sigma^2} n_j \sim \chi^2_{k-1}$$

Distribuzione campionaria del rapporto V

Dato che le medie $\overline{X}_j$ sono indipendenti dalle varianze S_j^2, deriva che la devianza fra i gruppi è indipendente dalla devianza nei gruppi e quindi che

$$V = \frac{D.F./(k-1)}{D.N./(n-k)} = \frac{Y/(k-1)}{W/(n-k)} = \frac{\frac{1}{k-1}\sum_{j=1}^{k}\left(\overline{X}_j - \overline{X}\right)^2 n_j}{\frac{1}{n-k}\sum_{j=1}^{k}\left(n_j - 1\right) S_j^2}$$

si distribuisce, sotto H_0, come una v.c. di Fisher con *(k-1)* e *(n-k)* g.d.l.
Pertanto respingiamo H_0 per $V > F_{1-\alpha}$.

Esempio 1

Riportiamo di seguito i pesi di 15 bielle prodotte in un'officina da 3 presse simili.

Pressa 1	*Pressa 2*	*Pressa 3*
90	92	95
95	89	90
101	92	92
94	93	93
	94	91
		94

Possiamo ipotizzare che i pesi delle bielle provengano da 3 popolazioni $N\left(\mu_j;\sigma^2\right)$.

Vogliamo verificare a livello $\alpha = 0,05$ l'ipotesi $H_0 : \mu_1 = \mu_2 = \mu_3$.

Le medie campionarie risultano $\bar{x}_1 = 95$, $\bar{x}_2 = 92$, $\bar{x}_3 = 92,5$, quindi $\bar{x} = 93$.

La devianza nei gruppi è

$$\sum_{j=1}^{3}\left(n_j - 1\right)S_j^2 = 62 + 14 + 17,5 = 93,5$$

e la devianza fra i gruppi è

$$\sum_{j=1}^{3}\left(\bar{x}_j - \bar{x}\right)^2 n_j = 22,5.$$

La v.c. test risulta

$$\frac{(D.F.)\big/(k-1)}{(D.N.)\big/(n-k)}=\frac{22,5\big/(3-1)}{93,5\big/(15-3)}=1,444$$

Dalle tavole della distribuzione F con 2 e 12 g.d.l. si ricava $F_{0,95}$=3,89 > 1,444. Accettiamo allora l'ipotesi che le presse producano bielle con lo stesso peso medio.

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

$$\frac{\left|\overline{X}_3 - \overline{X}_4\right|}{S\sqrt{\dfrac{n_3 + n_4}{n_3 \cdot n_4}}} > t_{1-\frac{\alpha}{2}}\left(n_3 + n_4 - 2\right)$$

$$S^2 = \frac{1{,}4 + 2{,}5}{10 + 9 - 2} = 0{,}233$$

$$\frac{|6{,}4-6{,}\overline{7}|}{\sqrt{0{,}233}\sqrt{\frac{19}{90}}}=1{,}705<t_{0{,}975}(17)=2{,}11$$

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

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

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

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

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

$(x_{1i} - \overline{x}_1)^2$	$(x_{2i} - \overline{x}_2)^2$	$(x_{3i} - \overline{x}_3)^2$
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)*