Introduzione all'analisi di varianza: varianza entro e tra gruppi

La procedura dell'analisi della varianza sfrutta il fatto che la varianza della popolazione da cui, in base all'ipotesi H_0, provengono i campioni può essere stimata in due modi:

1. Utilizzando la media delle varianze calcolate in ogni campione (calcolo della cosiddetta ***varianza entro i gruppi***).
2. Utilizzando la relazione che lega la varianza della popolazione delle medie campionarie (che è il quadrato dell'errore standard della media)* e la varianza della popolazione (calcolo della cosiddetta ***varianza tra i gruppi***).

• Se tutti i campioni fossero tratti dalla stessa popolazione (nell'esempio se nessuna dieta avesse effetto) i due metodi usati per la stima dovrebbero dare approssimativamente lo stesso risultato. Quando ciò accade, possiamo concludere che, probabilmente, i campioni sono tratti da un'unica popolazione; in caso contrario rifiuteremo questa ipotesi e concluderemo che almeno uno dei campioni è stato tratto da una popolazione diversa (nel caso del nostro esperimento concluderemo che la dieta modifica la gittata cardiaca).

* ricordando che $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, si ricava che $\sigma^2 = n\sigma_{\bar{x}}^2$

Analisi della varianza: calcolo della statistica F

• La stima della varianza della popolazione sulla base delle varianze campionarie (***varianza entro gruppi***) è:

$$\boxed{S^2{}_{entro} = \frac{1}{4}\left(S^2{}_{cont} + S^2{}_{spag} + S^2{}_{bisc} + S^2{}_{frut}\right)}$$

• La stima della varianza della popolazione sulla base delle medie campionarie (***varianza tra i gruppi***) è ricavata dalla relazione $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ ed è:

$$\boxed{S^2{}_{tra} = nS_{\bar{x}}{}^2}$$

dove $S_{\bar{x}}$ è la stima della deviazione standard delle medie campionarie.

• Se l'ipotesi che i gruppi appartengono alla stessa popolazione fosse vera, la *varianza entro* e la *varianza tra* gruppi rappresenterebbero buone stime della varianza della popolazione di origine ed il rapporto

$$\boxed{F = \frac{S^2{}_{tra}}{S^2{}_{entro}}}$$

definito come ***statistica del test F***, dovrebbe essere pressoché uguale a 1.

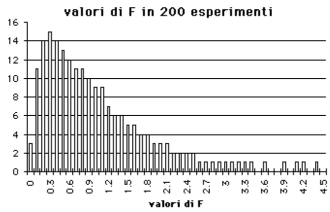
Valori della statistica F

• Se il valore di *F* è “elevato”, la variabilità tra le medie campionarie è maggiore di quella attesa sulla base della variabilità all’interno dei singoli campioni, pertanto rifiutiamo l’ipotesi che tutti i campioni appartengano alla stessa popolazione.

à ***Il problema è ora: quanto deve essere elevato F per poter rifiutare l’ipotesi iniziale ?*** à

- Supponendo di ripetere l’esperimento 200 volte, anche se la dieta non influisce sulla gittata cardiaca, otteniamo ogni volta valori diversi di *F* (vedi la distribuzione sotto).

- Poiché, quando i campioni sono tutti estratti dalla stessa popolazione, il valore di F risulta maggiore di 3 solo 10 volte su 200 (il 5% delle volte), abbiamo una stima sperimentale del base alla quale definirlo “elevato” e respingere l’ipotesi H_0 che i trattamenti si equivalgano. In questo caso 3 rappresenta il valore soglia (*valore critico*) oltre il quale rifiutiamo H_0 con un rischio di commettere un errore di tipo I° (*falso positivo*) del 5% ($p=<0,05$, $\alpha=0.05$).



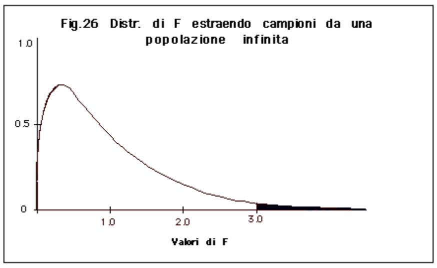
- $P<0,05$ significa dunque che, se H_0 fosse vera, avremmo meno del *5%* di probabilità di trovare in un singolo esperimento, un valore di *F* maggiore di 3.

La forma della distribuzione di F

Naturalmente il valore critico di F dovrebbe essere scelto non sulla base di 200 esperimenti, ma considerando tutti gli esperimenti possibili (praticamente possiamo considerarlo un numero infinito).

La forma esatta della distribuzione dei valori di F (e quindi i valori critici associati ai livelli di significatività prestabiliti) dipendono dal numero dei campioni estratti e dalla numerosità di ciascun campione.

La figura mostra la distribuzione di F che si otterrebbe estraendo tutte le possibili combinazioni di 4 campioni di ampiezza 7 dalla popolazione presa in considerazione nel nostro esempio.



In questo caso il valore limite che permette di considerare F “elevato” è quel valore di F che individua, nell’area delimitata dalla curva, un’area pari al 5% del totale (area ombreggiata), caratterizzata dai valori più estremi di F.

Analogamente a quanto fatto per le altre distribuzioni, i matematici hanno costruito tabelle che mostrano i valori critici di F in funzione del numero di campioni e della loro dimensione.

Condizioni di applicabilità dell'analisi di varianza

Per costruire queste tabelle i matematici hanno assunto che devono essere almeno approssimativamente soddisfatte le seguenti condizioni:

- ***ogni campione deve essere indipendente dagli altri;***
- ***ogni campione deve essere scelto con procedura casuale;***
- ***la popolazione da cui i campioni sono estratti deve essere distribuita normalmente.***
- ***Le varianze di ciascuna popolazione devono essere pressoché uguali.***

Quando i dati suggeriscono che queste condizioni non sono soddisfatte, non è lecito ricorrere all'analisi parametrica della varianza*.

* In questo caso si può ricorrere ai test non parametrici.

Tabella dei valori critici di F e gradi di libertà

Valori critici di F corrispondenti a P<=0.05 (caratteri chiari) e P<=0.01 (caratteri in grassetto)

ν_n

ν_d	1	2	3	4	5	6	7	8	9	10	11	12
10	4.69	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91
	10.04	7.56	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79
	9.65	7.20	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**
12	4.75	3.88	3.49	3.26	3.11	3.00	2.97	2.85	2.80	2.76	2.72	2.69
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**
100	3.94	3,09	2,70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.84
	6.85	**4.79**	**3.95**	**3.48**	**3.17**	**2.96**	**2.79**	**2.66**	**2.56**	**2.47**	**2.40**	**2.34**
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75
	6.63	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**

NOTA: nel caso dello studio sulle diete e la gittata cardiaca il valore critico corrispondente a P<=0.05 con ν_n=3 e ν_d=24 gradi di libertà. è F=3.01.

- La dipendenza dei valori critici dalla numerosità di ciascun gruppo e dal numero dei gruppi a confronto è espressa da due parametri detti ***gradi di libertà ν***.
- Per questo genere di analisi il numero di gradi di libertà ν_n *"tra gruppi"* (chiamati gradi di libertà del numeratore, poiché la varianza "tra gruppi" è il numeratore di F) è:

$$\nu_n = m - 1$$

con m = numero dei gruppi

- Il numero di gradi di libertà *"entro gruppi"* (chiamati gradi di libertà del denominatore, poiché la varianza "entro gruppi" è il denominatore di F) è:

$$\nu_d = m(n-1)$$

con m = numero dei gruppi
ed n = numerosità di ciascun gruppo

* Esercizio Anova *

• Studio per accertare l'eventuale relazione fra la frequenza di eventi febbrili e la pratica del jogging

	N. Medio Eventi febbrili/anno	Deviazione standard
Controlli	11.5	1.3
Joggers (5-30Km/settimana)	10.1	2.1
Runners (>30 km/settimana)	9.1	2.4

***Ampiezze campionarie pari a 26 unità**

QUESITO:
esiste una relazione fra la frequenza degli eventi febbrili e la pratica della corsa?

* Esercizio (soluzione) *

• Stima della varianza sulla base della media delle varianze entro i gruppi:

$$S^2_{entro} = \frac{1}{3}(S^2_{con} + S^2_{jog} + S^2_{run}) = \frac{1}{3}(1.3^2 + 2.1^2 + 2.4^2) = 3.95$$

• Stima della varianza a partire dalla variabilità fra le medie campionarie:

$$\bar{x} = \frac{1}{3}(\bar{x}_{con} + \bar{x}_{jog} + \bar{x}_{run}) = \frac{1}{3}(11.5 + 10.1 + 9.1) = 10.2$$

$$S_{\bar{x}} = \sqrt{\frac{(\bar{x}_{con} - \bar{x})^2 + (\bar{x}_{jog} - \bar{x})^2 + (\bar{x}_{run} - \bar{x})^2}{m-1}} = \sqrt{\frac{(11.5 - 10.2)^2 + (10.1 - 10.2)^2 + (9.1 - 10.2)^2}{3-1}} = 1.2$$

$$S^2_{tra} = nS^2_{\bar{x}} = 26(1.2)^2 = 37.4$$

• Calcolo di F:

$$F = \frac{S^2_{tra}}{S^2_{entro}} = \frac{37.44}{3.95} = 9.48 \qquad \nu_n = m - 1 = 2 \qquad \nu_d = m(n-1) = 3(26-1) = 75$$

• Interpolando nella tabella dei valori critici troviamo che, se l'ipotesi H_0 è vera, F sarà maggiore di 4.9 solo nell' 1% dei casi. Pertanto possiamo concludere che l'attività sportiva influisce significativamente sulla frequenza degli eventi febbrili.

Confronti multipli post-hoc

• Dopo un' analisi della varianza che eventualmente abbia portato alla decisione di rifiutare l' ipotesi di appartenenza dei gruppi alla stessa popolazione, è necessario procedere con tutti possibili ***confronti multipli (detti post-hoc)*** a coppie.
• In ogni confronto si dovrà tener conto della *disuguaglianza di Bonferroni* che prevede, come valore critico di riferimento, quello associato ad una frazione (pari al numero totale di confronti da effettuare) del livello globale di *significatività*.
• Ne consegue che, quando il numero dei confronti da effettuare è elevato, l' eventuale *t-test di Student* classico risulta troppo conservativo (il valore critico t_c richiesto per concludere che esiste una differenza diventa molto più "grande" del necessario) * ed è quindi necessario utilizzare altri tests.

*** Infatti, come avevamo osservato commentando la tabella delle aree sottese dalle distribuzioni *t di Student*, i *valori critici di t* necessari al rifiuto di H_0 "crescono" al diminuire del *livello di significatività* α.**

Il test t di Bonferroni

• Per i confronti post-hoc il test generalmente più utilizzato è quello *di Bonferroni*. Esso è meno conservativo del classico test *t di Student* perché, nel calcolo di *t*, utilizza la ***stima della varianza entro i gruppi*** (vedi analisi di varianza) in luogo della ***stima combinata della varianza***.
• In questo modo i *gradi di libertà* risultano più numerosi* di quelli di un semplice test ***t***. Il fatto che il *valore critico* di *t* decresce all' aumentare dei *gradi di libertà*, compensa lo svantaggio derivato dall' applicazione della disuguaglianza di Bonferroni e permette di rilevare una differenza, con grado prefissato di significatività, anche in presenza di minori differenze assolute fra le medie.
• La formula per il calcolo di *t* è:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2_{entro}}{n_1} + \frac{s^2_{entro}}{n_2}}}$$ oppure $$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2\frac{s^2_{entro}}{n}}}$$ nel caso di campioni di uguale ampiezza

* Essi sono $\nu = m(n-1)$ e non $\nu = 2(n-1)$

Il test t di Bonferroni (esempio)

• **Nell'esempio già visto sui disturbi febbrili nella pratica della corsa, in quale particolare coppia di gruppi risiede la differenza?**

Per confrontare il gruppo di controllo con quello delle dilettanti abbiamo:

$$t = \frac{\bar{x}_{con} - \bar{x}_{dil}}{\sqrt{2s^2{}_{entro}/n}} = \frac{11{,}5 - 10{,}1}{\sqrt{2(3{,}95)/26}} = 2{,}54$$

Per confrontare il gruppo di controllo con quello delle professioniste abbiamo:

$$t = \frac{\bar{x}_{con} - \bar{x}_{prof}}{\sqrt{2s^2{}_{entro}/n}} = \frac{11{,}5 - 9{,}1}{\sqrt{2(3{,}95)/26}} = 4{,}35$$

Per confrontare il gruppo delle dilettanti con quello delle professioniste abbiamo:

$$t = \frac{\bar{x}_{dil} - \bar{x}_{prof}}{\sqrt{2s^2{}_{entro}/n}} = \frac{10{,}1 - 9{,}1}{\sqrt{2(3{,}95)/26}} = 1{,}81$$

- Per mantenere la probabilità globale di errore sotto il 5% utilizziamo, in ogni confronto, il valore critico associato ad un livello di significatività $\alpha=(5/3)\%=1,6\%$. Il numero di gradi di libertà è $\nu=m(n-1)=3(26-1)=75$.

- Il valore critico associato a 75 gradi di libertà per α=0,016 è $t=2,47$. Perciò possiamo concludere che sia la corsa per dilettanti che per professioniste diminuiscono significativamente la frequenza degli eventi febbrili, mentre la corsa competitiva non aumenta gli eventi in misura significativa rispetto alla semplice corsa per dilettanti.