

4.2. IL TEST F DI FISHER O ANALISI DELLA VARIANZA (ANOVA)

L'analisi della varianza è un metodo sviluppato da Fisher, che è fondamentale per l'interpretazione statistica di molti dati biologici ed è alla base di molti disegni sperimentali. L'analisi della varianza (in inglese: Analysis of variance, abbreviata con l'acronimo ANOVA) è utilizzata per testare le differenze tra medie campionarie e per fare questo si prendono in considerazione le rispettive varianze. Il principio alla base di questo test è quello di stabilire se due o più medie campionarie possono derivare da popolazioni che hanno la stessa media parametrica. Quando le medie sono solamente due è indifferente usare questo test od il t-test, mentre dobbiamo necessariamente utilizzare l'ANOVA quando le medie sono più di due, o quando vogliamo suddividere la variabile di raggruppamento in più variabili per eliminare eventuali fonti di variazione oltre a quella prodotta dal fattore di cui vogliamo valutarne l'effetto. Esempio, se io voglio valutare l'effetto di un determinato farmaco, oltre alla variabile di raggruppamento che mi divide il campione in animali trattati e non trattati (ovvero animali a cui è stato somministrato il farmaco e altri che fungono da controllo) posso anche introdurre altre variabili di raggruppamento (maschi e femmine, oppure giovani, adulti e vecchi) per vedere se altri fattori possono essere responsabili della eventuale differenza osservata tra i due gruppi principali.

Per capire la logica che sta alla base di questo metodo di analisi si può partire da un esempio puramente teorico che, ovviamente, non ha nessun riscontro realistico. Ammettiamo infatti di poter misurare tutti gli individui di una popolazione e quindi di conoscere la vera media parametrica (μ) e la vera varianza parametrica (σ^2). Ovviamente se così fosse non ci sarebbe la necessità di utilizzare alcuna procedura statistica.

Population						
41	40	48	42	39	47	50
48	45	50	49	39	51	42
49	49	47	45	48	46	48
45	50	49	44	46	49	38
49	39	49	41	44	48	41
42	49	42	50	42	44	43
49	46	44	41	54	42	44
38	47	39	42	39	39	53
44	50	50	48	46	42	44
48	44	54	46	46	41	48
49	40	48	52	51	50	51
51	50	51	48	49	43	41
40	48	41	42	45	47	46
45	49	48	47	41	40	40
42	40	44	51	46	41	46
mean $\mu =$ 45,5		variance $\sigma^2 =$ 16,43		$n = 105$		

Ora passiamo ad una situazione più realistica, ovvero prendiamo dei campioni della popolazione e cerchiamo di estrapolare da essi i parametri di popolazione. In questo esempio consideriamo 7 **gruppi** (campioni), **a = 7**, ciascuno costituito di **n = 5 items** estratti a caso dalla popolazione.

a	b	c	d	e	f	g
41	48	40	40	49	40	41
44	49	50	39	41	48	46
48	49	44	46	50	51	54
43	49	48	46	39	47	44
42	45	50	41	42	51	42

Otteniamo così sette stime della media e della varianza parametrica

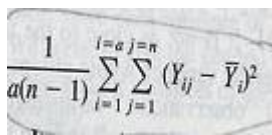
	a	b	c	d	e	f	g
Media	43,6	48	46,4	42,4	44,2	47,4	45,4
Devianza	29,2	12	75,2	45,2	98,8	81,2	107,2
Varianza, s^2	7,3	3	18,8	11,3	24,7	20,3	26,8

Come si vede otteniamo stime (s^2) alquanto approssimative della varianza parametrica ($\sigma^2 = 16,43$). Possiamo però ottenere una stima migliore se si fa una media di queste varianze stimate indipendentemente l'una dall'altra. Per fare questo si fa una media ponderata¹ di tutte le varianze, ovvero si sommano tutte le devianze e si divide per i la somma dei gradi di libertà ($n-1$):

$$s^2 = (29,2 + \dots + 107,2) / (4 + \dots + 4) = 448,8 / 28 = \mathbf{16,029}$$

La stima così ottenuta si avvicina di più a quella parametrica ed è chiamata **varianza in gruppi**.

La formula generale di questa operazione è la seguente:



Con $i = 1$ consideriamo il primo gruppo (a) e con j da 1 ad n sommiamo tutti gli scarti quadratici dalla media di quel gruppo, ovvero ne calcoliamo la devianza

Con i valori di i da 1 ad a sommiamo le devianze ottenute da tutti i gruppi. Dividendo per i gradi di libertà [$a*(n-1)$] otteniamo la varianza in gruppi

Un'altra possibilità è quella di considerare le 35 osservazioni come se fossero un unico campione:

Gruppo unico	n	$df (n-1)$	Media	SS, devianza	MS, Varianza
	35	34	45,343	575,886	16,938

Anche in questo modo otteniamo una stima della varianza parametrica, che è detta **varianza totale**:

$$s^2 = [(41 - 45,343)^2 + \dots + (42 - 45,343)^2] / [(7*5) - 1] = 575,886 / 34 = \mathbf{16,93782}$$

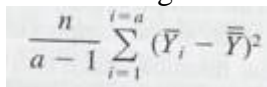
Infine potremmo considerare direttamente le medie dei 7 gruppi e calcolare la **varianza tra medie**:

$$s^2_{\bar{Y}} = [((43,6 - 45,343)^2 + \dots + (45,4 - 45,343)^2) / (7-1)] = 4,236$$

La **varianza tra medie** si ottiene considerando gli a gruppi come altrettante osservazioni, quindi si sommano tutte le differenze quadratiche delle a medie dalla grande media, si divide per i gradi di libertà, che questa volta sono $a-1$ (6).

Moltiplicando per n , ovvero per il numero di osservazioni su cui sono state ottenute le medie, la varianza tra medie otteniamo la **varianza tra gruppi**: $5 * 4,236 = \mathbf{21,181}$

La formula generale di questa operazione è la seguente:



Con i valori di i da 1 ad a sommiamo le differenze quadratiche delle medie di tutti i gruppi rispetto alla grande media. Dividendo per $a-1$ otteniamo la varianza tra medie e moltiplicando per n otteniamo la varianza tra gruppi.

Ricordiamo che la varianza non è una grandezza additiva, mentre lo sono le devianze. Avremmo quindi potuto stimare la varianza tra gruppi anche nel seguente modo:

$$[(SS_{Tot} - SS_{In}) / (df_{Tot} - df_{In})] \text{ Ovvero: } (575,886 - 448,8) / (34 - 28) = \mathbf{21,181}$$

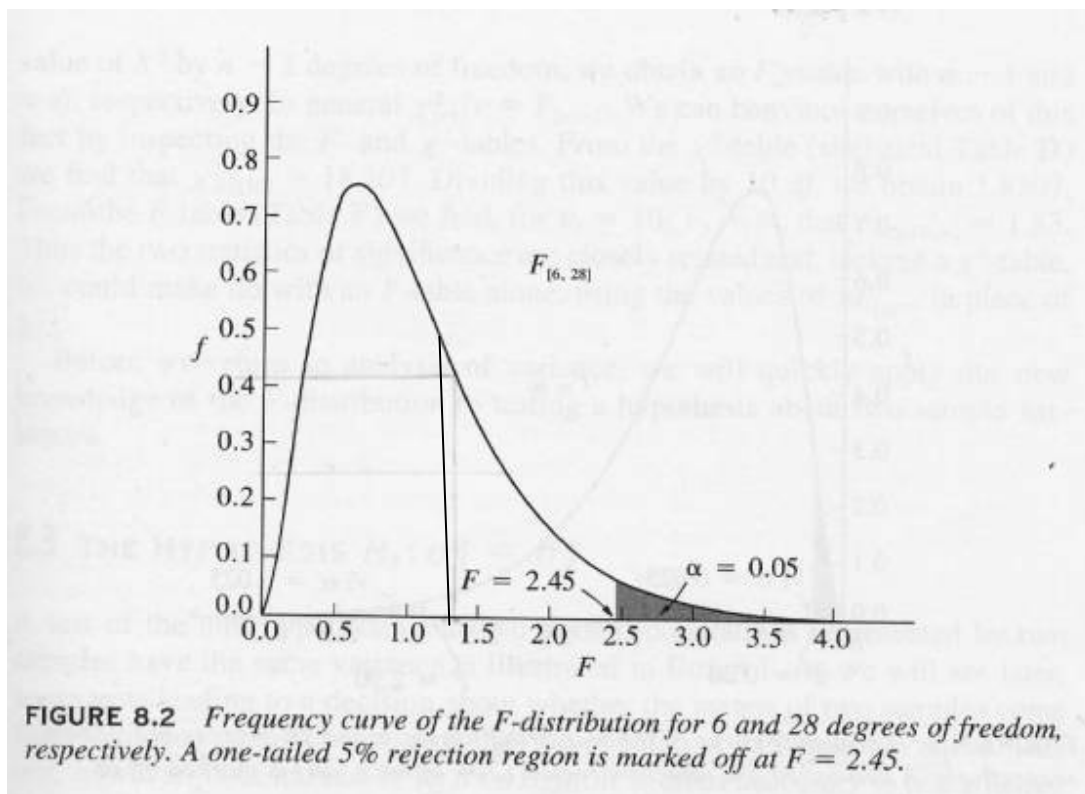
Infatti la variabilità totale, ovvero quella osservata in un campione di 35 osservazioni, è dovuta alla variabilità dentro i gruppi e quella tra gruppi.

¹ In questo caso, dato che tutti i gruppi hanno la stessa numerosità (quindi gli stessi gdl), si poteva fare una semplice media aritmetica delle varianze.

Il campione di sette gruppi ciascun con 5 animali fornisce quindi due stime della varianza parametrica
 Se si considera il **rapporto $F_s = s^2 \text{ tra gruppi} / s^2 \text{ in gruppi}$** , essendo stime di una stessa varianza parametrica dovrebbe essere, in teoria, uguale ad 1. Se il valore al numeratore è grande rispetto al denominatore si ottengono rapporti >1 . Esiste, per ogni combinazione di gradi di libertà del numeratore e del denominatore, e per ogni livello di probabilità, una particolare curva statistica (**distribuzione di F**) che ci consente di stabilire se il rapporto ottenuto è superiore ad un certo valore soglia per cui è poco probabile (e noi dobbiamo scegliere il livello di probabilità che vogliamo tenere in considerazione) ottenere quei valori quando si considerino gruppi random di una stessa popolazione. E' su questa distribuzione di probabilità, studiata da Fisher, che si basa l'**F- test** utilizzato nella **ANOVA**.

ANALISI VARIANZA ad un fattore

Origine variazione	SS	gdl	MS	F-test F	Valore significatività	F crit
Tra gruppi	127,086	6	21,181	1,321	0,281	2,445
In gruppi	448,8	28	16,029			
Totale	575,886	34	16,938			



Nel nostro caso circa il 50% di tutti i rapporti (basati su 6 e 28 gradi di libertà) avrebbero valori superiori a quello osservato di 1,32 (21,181 / 16,029) e quindi le due stime ottenute sono **stime di una stessa varianza parametrica, ovvero i gruppi appartengono ad una stessa popolazione**. Rapporti con valori $>2,45$ possono ancora stimare una stessa varianza ma si osservano solo in percentuali $<5\%$ e quindi abbiamo una probabilità $<5\%$ (0.05) di commettere un errore di tipo I nel rigettare l'ipotesi nulla, nel caso fosse vera.

		Ipotesi nulla	
		Accettata	Respinta
Ipotesi nulla	Vera	OK	Type I error
	Falsa	Type II error	OK

La nostra ipotesi H_0 è che i campioni siano rappresentativi di una stessa popolazione, per cui le due varianze, ovvero quella tra gruppi e quella dentro i gruppi sarebbero stime di una stessa varianza

parametrica. L'ipotesi alternativa è che i campioni non siano rappresentativi di una stessa popolazione per cui è verosimile che la varianza tra gruppi sia significativamente più grande di quella osservata tra osservazioni di uno stesso gruppo. Questa ipotesi alternativa restrittiva porta di conseguenza che l'F-test è un test ad una coda.

Consideriamo ora il fatto che i sette gruppi **non siano gruppi omogenei** ma che derivino da sette diversi trattamenti i cui effetti sono:

a	b	c	d	e	f	g	$\bar{\alpha} = 0$
$\alpha_1 = -5$	$\alpha_2 = -2$	$\alpha_3 = 0$	$\alpha_4 = +1$	$\alpha_5 = +1$	$\alpha_6 = +5$	$\alpha_7 = 0$	
41 -5 = 36	48 -2 = 46	40	40 +1 = 41	49 +1 = 50	40 +5 = 45	41	
44 -5 = 39	49 -2 = 47	50	39 +1 = 40	41 +1 = 42	48 +5 = 53	46	
48 -5 = 43	49 -2 = 47	44	46 +1 = 47	50 +1 = 51	51 +5 = 56	54	
43 -5 = 38	49 -2 = 47	48	46 +1 = 47	39 +1 = 40	47 +5 = 52	44	
42 -5 = 37	45 -2 = 43	50	41 +1 = 42	42 +1 = 43	51 +5 = 56	42	

Gruppi	a	b	c	d	e	f	g
Media	43,6 - 5 = 38,6	48 - 2 = 46	46,4 - 0 = 46,4	42,4 + 1 = 43,4	44,2 + 1 = 45,2	47,4 + 5 = 52,4	45,4 + 0 = 45,4
devianza	29,2	12	75,2	45,2	98,8	81,2	107,2
Varianza	7,3	3	18,8	11,3	24,7	20,3	26,8

Ovviamente questo esempio è solo a scopo illustrativo, in quanto noi non possiamo quantizzare a priori l'effetto di un trattamento, così come non possiamo sapere quale sarebbe stata la media di un determinato gruppo se non fosse stato applicato alcun trattamento.

Ricordando che la **varianza tra medie** si calcolava con $\frac{1}{a-1} \sum_{i=1}^{i=a} (\bar{Y}_i - \bar{Y})^2$

ora abbiamo:

$$\frac{1}{a-1} \sum_{i=1}^{i=a} [(\bar{Y}_i + \alpha_i) - (\bar{Y} + \bar{\alpha})]^2 \rightarrow \frac{1}{a-1} \sum_{i=1}^{i=a} [(\bar{Y}_i - \bar{Y}) + (\alpha_i - \bar{\alpha})]^2$$

Sviluppando il quadrato del binomio tra parentesi quadre:

$$\frac{1}{a-1} \sum_{i=1}^{i=a} (\bar{Y}_i - \bar{Y})^2 + \frac{1}{a-1} \sum_{i=1}^{i=a} (\alpha_i - \bar{\alpha})^2 + \frac{2}{a-1} \sum_{i=1}^{i=a} (\bar{Y}_i - \bar{Y})(\alpha_i - \bar{\alpha})$$

Il primo termine è la **varianza tra medie** che viene indicata come $S^2_{\bar{Y}}$

Il secondo è qualcosa di simile ad una **varianza dovuta all'effetto dei trattamenti**. Non sarebbe una vera varianza perché **i trattamenti sono fissati dallo sperimentatore**, quindi sotto controllo (cambiando dosaggi o il tipo di trattamento possiamo far variare questo termine a nostro piacere). In questo caso si parla di un modello **ANOVA di tipo I** e di **componente aggiunta dovuta all'effetto del trattamento**.

Il terzo termine è una nuova grandezza detta **covarianza** (tra medie dei gruppi ed effetti dei trattamenti) di cui parleremo in seguito. In questo caso è uguale a "0" perché la misura dell'effetto dovuto al trattamento non dipende dalla media del gruppo a cui è stato applicato.

Poiché a scopo dimostrativo, abbiamo codificato l'effetto dei trattamenti con dei valori la cui media è 0

possiamo riscrivere il **secondo termine** come:

$$\frac{1}{a-1} \sum_{i=1}^{i=a} (\alpha_i - \bar{\alpha})^2 = \frac{1}{a-1} \sum_{i=1}^{i=a} \alpha_i^2 = \frac{1}{a-1} \sum_{i=1}^a \alpha_i^2$$

(**componente aggiunta** dovuta all'effetto del trattamento)

e non considerare il terzo termine in quanto uguale a 0.

Ricordando che per ottenere **la varianza tra gruppi** si moltiplicava per n la varianza tra medie abbiamo:

$$n \left(s_{\bar{Y}}^2 + \frac{1}{a-1} \sum \alpha^2 \right) = s^2 + \frac{n}{a-1} \sum \alpha^2$$

$$n * \boxed{s_{\bar{Y}}^2} = s^2 \quad \text{Varianza tra gruppi}$$

Così vediamo che la stima della varianza parametrica ottenuta dalla **varianza tra gruppi è aumentata di** una quantità che è **n volte la componente aggiunta dovuta al trattamento.**

La varianza in gruppi, che è la somma di tutte le devianze diviso la somma di tutti i gradi di libertà, è rimasta inalterata (come ci aspettavamo in quanto a ciascun individuo di un gruppo è stato aggiunto un valore costante). Infatti ricordando che

$$\frac{1}{a(n-1)} \sum_{i=1}^a \sum_{j=1}^n (Y_{ij} - \bar{Y}_i)^2$$

Ora abbiamo

$$\frac{1}{a(n-1)} \sum_{i=1}^a \sum_{j=1}^n [(Y_{ij} + \alpha_i) - (\bar{Y}_i + \alpha_i)]^2$$

ma se sviluppiamo i termini dentro la parentesi quadra vediamo che i due alfa si elidono e la formula ritorna quella originaria.

ANALISI VARIANZA ad un fattore ANOVA type I

Origine variazione	SS	gdl	MS	F	Valore di significatività	F crit	MS atteso
Tra gruppi	503,086	6	83,848	5,231	0,001	2,445	$\sigma^2 + \frac{n}{a-1} \sum \alpha^2$
In gruppi	448,800	28	16,029				σ^2
Totale	951,886	34	27,997				

Nella precedente analisi: **MS tra gruppi = 21,181**; **MS in gruppi (o di errore) 16,029**; MS totale 16,938

$$\frac{n}{a-1} \sum \alpha^2$$

(added component due to treatment effects)

La varianza tra gruppi è ora composta dalla **varianza parametrica** (σ^2) + n volte il valore della **componente aggiunta** (dovuta al trattamento) che può essere schematizzata con K_A^2 da cui: $\sigma^2 + nK_A^2$

F-Test

Il rapporto $F_s = s^2 \text{ tra gruppi} / s^2 \text{ in gruppi}$, essendo stime di una stessa varianza parametrica, dovrebbe essere uguale ad 1 ma, nel nostro caso, il **numeratore**, ovvero **s^2 tra gruppi**, è **sensibile alla presenza della componente aggiunta** dovuta all'effetto del trattamento, per cui il valore che si ottiene

$$\frac{\sigma^2 + \frac{n}{a-1} \sum \alpha^2}{\sigma^2}$$

(83,848 / 16,029 = **5,231**) risulta superiore al valore critico (2,45), avendo una probabilità di essere ottenuto, a partire da gruppi omogenei, di **0,001**. Ne risulta che i trattamenti hanno un effetto tale da far variare le medie in modo significativo così che non possono più essere considerate campioni di una

stessa popolazione. In altri termini i trattamenti usati (α_i) producono effetti significativamente diversi.

In questo tipo di analisi (**ANOVA tipo I**, ovvero con **trattamenti fissi**) non siamo interessati a stimare

il valore totale della componente aggiunta dovuta ai trattamenti, ma **la grandezza dei valori** di α_i ,

ovvero, se esistono differenze significative tra i vari α_i (rilevabili dalla presenza di una componente aggiunta)

e successivamente quali degli α_i differiscono significativamente tra loro (in modo da produrre gruppi con

medie significativamente diverse), cosa che verrà fatta con test appropriati (contrast, confronti appaiati,

confronti multipli).

Nota. In effetti l'F-Test è una verifica dell'ipotesi nulla (nessun effetto)

$$\frac{n}{a-1} \sum \alpha^2 = 0$$

Che si ottiene valutando quanto il rapporto di due grandezze che stimano una stessa varianza parametrica si discosta dal valore ideale (1). Se questa discordanza supera un certo valore soglia allora è verosimile che ciò sia dovuto alla presenza di componenti addizionali dovute all'effetto del trattamento, quindi si rigetta l'ipotesi nulla.

Ricordiamo che quando i **gruppi** da confrontare sono solo **due** e siamo nel caso di un **modello tipo I** si può eseguire un altro test: il **t-test delle differenze tra due medie**. I due metodi si equivalgono e sono legati dalla relazione: $t^2 = F$

Consideriamo ora un altro possibile modello di analisi della varianza, **ANOVA di tipo II**, dove non ci sono trattamenti fissi ma **effetti casuali**, che non possiamo controllare. In questo caso l'algebra utilizzata per calcolare le due stime (varianza tra ed in gruppi) della varianza parametrica è la stessa, ma utilizziamo **lettere latine maiuscole** per indicare che siamo in presenza di effetti casuali. Pertanto, la formula sarà:

varianza tra medie componente aggiunta della varianza tra gruppi

$$\frac{1}{a-1} \sum_{i=1}^{i=a} (\bar{Y}_i - \bar{Y})^2 + \frac{1}{a-1} \sum_{i=1}^{i=a} (A_i - \bar{A})^2 + \frac{2}{a-1} \sum_{i=1}^{i=a} (\bar{Y}_i - \bar{Y})(A_i - \bar{A})$$

Per il primo ed il terzo termine vale quanto detto precedentemente (il primo = varianza tra medie, il terzo la covarianza) mentre il **termine intermedio** deve considerarsi una vera e propria varianza in quanto dovuta ad una variabile casuale, qui detta **componente aggiunta della varianza tra gruppi** ed indicata con s_A^2 .

Moltiplicando tutto per n , otteniamo che il **primo termine** diventa la **varianza tra gruppi** :
 $n * \boxed{s^2_{\bar{Y}}} = s^2$

il **secondo** la **varianza aggiunta** ns_A^2

Così vediamo che la stima della varianza parametrica ottenuta dalla **varianza tra gruppi è aumentata di** una quantità che è la **varianza aggiunta** (ovvero, n volte la componente aggiunta della varianza tra gruppi):

$s^2 + ns_A^2$ (added variance component among groups)

Essendo questa ancora una stima della varianza parametrica possiamo scriverla come

varianza parametrica + varianza parametrica della componente aggiunta: $\sigma^2 + n\sigma_A^2$

L'esistenza di questa componente aggiunta della varianza è testata ancora dal **F-test**:

$\frac{s^2 + ns_A^2}{s^2}$ ovvero $\frac{s^2 \text{ tra gruppi}}{s^2 \text{ in gruppi}}$ da cui, estendendo alla popolazione, $\frac{\sigma^2 + n\sigma_A^2}{\sigma^2}$

(In questo caso l'ipotesi nulla è: $n\sigma_A^2 = 0$)

Nel modello **ANOVA di tipo II** non siamo interessati ai singoli effetti né alla differenza tra essi, in quanto, essendo dovuti a fattori casuali, non avremmo gli strumenti per interpretarli. Siamo invece interessati a **quantificare il peso di questa componente della varianza aggiunta** ed eventualmente metterla relazione con il valore della varianza all'interno dei gruppi.

Se, per esempio, i dati della seconda tabella fossero misure morfometriche prese su 5 topolini, progenie di 7 madri diverse, non abbiamo nessuna ragione di prevedere a priori che i figli di "f" avrebbero avuto dimensioni maggiori dei topolini di altre covate, né sappiamo spiegarci il perché, salvo dire che i geni che controllano la crescita sono distribuiti in un certo modo (sconosciuto) nella popolazione e che il nostro campione di 7 femmine è un campione casuale, non omogeneo, di questi fattori e, quindi, possiamo solo **quantificare il peso di questa variabilità casuale** (s_A^2).

Se non ci fosse nessuna componente aggiunta il rapporto di F (s^2 tra gruppi / s^2 in gruppi) sarebbe uguale ad 1:
 $s^2 + ns^2_A / s^2 = 1$ da cui: $s^2 + ns^2_A = 1*s^2$

da cui* $1/n (s^2 + ns^2_A - s^2) = 1/n (ns^2_A) = s^2_A$

$s^2_A = 1/n$ (varianza tra gruppi - varianza in gruppi)

(*trasportando $1s^2$ al primo membro e dividendo per n)

Nel nostro esempio la componente aggiunta della varianza tra gruppi, $s^2_A = 1/5 (83,848 - 16,029) = 13,564$

Esprimendo la **componente aggiunta della varianza tra gruppi in %** e rispetto a $(s^2 + s^2_A)$,
 $100*s^2_A / (s^2 + s^2_A)$, abbiamo:

% componente aggiunta della varianza tra gruppi

$$\frac{13,56 * 100}{16,029 + 13,56} = 45,84$$

% varianza in gruppi

$$\frac{16,029 * 100}{16,029 + 13,56} = 54,16$$

RIASSUNTO

Nell'analisi della **varianza ad un fattore** entrano tre tipi di somme di quadrati (SS) e le rispettive varianze che vengono dette medie quadratiche (MS). Utilizzando i simboli "**a**" per il numero di gruppi ed "**n**" per il numero di osservazioni all'interno di ciascun gruppo abbiamo:

1) MS totale: è una misura statistica di dispersione di tutte le osservazioni (an) con **an-1** gradi di libertà attorno alla loro media (gran media) ovvero $Y - \bar{Y}$

Descrive la varianza dell'intero campione dovuta alle cause più svariate ed è una stima della varianza parametrica σ^2 (dell'intera popolazione), quando non ci sono effetti aggiuntivi dovuti ai trattamenti o a componenti della varianza tra gruppi.

2) MS in gruppi: è una misura di dispersione degli n individui attorno alla media del loro gruppo $Y - \bar{Y}$ con **a*(n-1)** gradi di libertà. Essa fa parte della "natura delle cose" è per questo è anche detta individuale o intragruppo o, più comunemente, **errore** della media quadratica. Se i gruppi sono campioni casuali è anch'essa una stima della varianza parametrica, non sensibile ad eventuali effetti o componenti aggiunte della varianza. Inoltre, è una misura di variazione che deve essere tenuta in considerazione quando si voglia stimare la significatività delle differenze tra medie di vari gruppi. In un anova monofattoriale, quale che sia il modello, rappresenta sempre il denominatore del test F.

3) MS tra gruppi: si basa sulla varianza delle medie dei gruppi, che descrive la dispersione degli a gruppi attorno alla gran media: $\bar{Y} - \bar{Y}$

Se i gruppi sono campioni random di una popolazione omogenea la varianza delle medie (attesa) è σ^2/n , per cui è possibile ottenere una stima della varianza parametrica tra gruppi moltiplicandola per **n**, e questa, in assenza di effetti additivi o di componenti aggiunte della varianza, è un'ulteriore stima della varianza parametrica.

Diversamente, è una stima della componente aggiunta dovuta all'effetto del trattamento, quando siamo in presenza di trattamenti fissi, ovvero siamo in un modello di **ANOVA di tipo I**.

$$\sigma^2 + \frac{n}{a-1} \sum \alpha^2$$

Per indicare questa quantità si può usare il simbolo $\sum \alpha^2 / a - 1$ (gdl) ma essendo effetti fissi

Controllati dallo sperimentatore, e quindi non essendo una vera e propria varianza, si preferisce usare il simbolo K^2_A , per cui le formula può essere scritta come $\sigma^2 + n K^2_A$

Nel caso che siamo in un modello di **ANOVA di tipo II** (dove c'è una componente aggiunta della varianza che è realmente una varianza in quanto è casuale, quindi non controllabile) al posto della formula estesa sommatoria degli scarti quadratici diviso i gdl, si utilizza il simbolo σ^2_A , per cui la formula diventa

$$\sigma^2 + n\sigma^2_A$$

ANOVA tipo I: assume che le differenze tra medie di diversi gruppi, se ci sono, siano dovute ad effetti prodotti da "trattamenti" controllati dallo sperimentatore.

I modelli di ANOVA possono essere espressi da una **funzione lineare**. In effetti i programmi di software statistico lavorano utilizzando una procedura "GLM" (General Linear Model) e solo per comodità interpretativa l'output risulta come comparazioni tra MS.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

; ovvero il valore dell'osservazione j esima del gruppo i esimo (*),

dipende dal valore della media parametrica μ (media di pop.), da una deviazione fissa α_i della media del proprio gruppo i dalla media di popolazione (dovuta al trattamento fisso subito dal gruppo) e da una deviazione casuale ε_{ij} dell'individuo j del gruppo i (variabilità individuale) dal valore atteso ($\mu + \alpha_i$). Si noti che α_i ed ε_{ij} possono assumere anche valori negativi, la media attesa di tutti gli ε_{ij} è 0, la loro varianza è una misura della varianza parametrica α^2 e, importantissimo essendo una relazione lineare, la distribuzione degli ε_{ij} (errori) deve essere normale, principale presupposto per fare l'analisi della varianza.

In questo modello di ANOVA siamo interessati a testare se esistono **differenze significative tra i vari α_i** rilevabili dalla presenza di una componente aggiunta dovuta ai trattamenti e, successivamente **quali** degli α_i **differiscono** significativamente **tra loro**, in modo da produrre gruppi con medie significativamente diverse.

$$Y_{11} = \mu + \alpha_1 + [Y - (\mu + \alpha_1)] \quad 36 = 45,5 + (-5) + [36 - (45,5 + (-5))]$$

$$36 = 40,5 + (-4,5)$$

(*) Nota: gli indici in statistica sono invertiti rispetto alle annotazioni matematiche

ANOVA tipo II: assume che le differenze tra medie di diversi gruppi, se ci sono, siano dovute ad effetti

casuali, non controllabili. Anche in questo caso c'è una relazione lineare $Y_{ij} = \mu + A_i + \varepsilon_{ij}$

per cui valgono le precedenti considerazioni salvo che la deviazione fissa A_i è dovuta al caso. Qui non siamo interessati a valutare le differenze tra le medie dei vari gruppi, ma a stimare la **varianza generale** tra questi gruppi e qual'è il **contributo** (in termini percentuali) dovuto alla presenza di **effetti casuali**.

A questo livello (ANOVA **monofattoriale**) la differenza tra i due modelli non implica nessuna differenza nel modo di condurre l'analisi, salvo indicare quali sono i successivi test da utilizzare, (confronti tra le medie dei gruppi con vari tests, nel tipo I, oppure analisi delle componenti della varianza nel tipo II). A partire da analisi più complesse (bifattoriali e multifattoriali) questa **distinzione** risulterà **importantissima per individuare nel F-test il giusto denominatore** da utilizzare per valutare la significatività dei vari fattori, in quanto la varianza tra gruppi, che individua la presenza di componenti o di varianze additive, non sarà testata sempre sulla base della varianza d'errore (MS in gruppi).

Un'altra implicazione nel considerare i due tipi di ANOVA sta nel fatto che **le conclusioni di un ANOVA di tipo I, non sono generalizzabili, mentre lo sono quelle di tipo II.** Per esempio, se in uno studio sulla biodiversità delle aree pascolive di un determinato parco queste risultano avere valori di ricchezza di specie significativamente diversi, nel caso che le aree siano state scelte secondo un determinato criterio stabilito dallo sperimentatore (ANOVA I), posso dire che quelle aree e solo quelle sono diverse. Diversamente, se le aree erano state scelte a caso tra le possibili aree esistenti, posso generalizzare le conclusioni e dire che le

_____ , anche se poi non so risalire al perché, in quanto non c'è stato un criterio specifico con cui ho identificato le aree utilizzate per l'indagine.

A SSUNZIONI DELL' ANALISI DELLA VARIANZA (ANOVA)

- A) L'ANOVA, come del resto altri metodi statistici, richiede che gli **elementi che costituiscono i vari gruppi** (items) non siano oggetto di una particolare selezione, ma siano stati **assegnati a caso (random)**. A volte, questa assunzione non viene rispettata anche se a noi potrebbe sembrare che l'assegnazione degli individui ad un gruppo o ad un altro sia avvenuta casualmente. Per esempio se vogliamo testare l'efficacia di un trattamento utilizzando due gruppi di 5 topi, uno che verrà sottoposto al trattamento e l'altro che funge da controllo, la cattura di questi animali potrebbe rappresentare un metodo selettivo in quanto la facilità con cui possiamo catturare un animale potrebbe essere indice di una sua particolare condizione fisiologica e/o patologica. Assegnando pertanto i primi 5 topi catturati ad un gruppo ed i successivi ad un altro gruppo potremmo, senza volere, aver creato due suddivisioni non casuali, che quindi andranno ad interferire sui risultati della sperimentazione. Assegnando invece alternativamente un animale ad un gruppo e quello catturato successivamente ad un altro gruppo, oppure numerando i vari animali ed estraendo a sorte quelli che dovranno costituire i due gruppi, è verosimile che si crei una suddivisione più casuale.
- B) **I campioni devono essere tra loro indipendenti**, ovvero i dati osservati in campione non devono essere influenzati da quelli osservati in un altro campione. Spesso quando si usano serie spaziali o temporali questa assunzione risulta violata. Ad esempio se si usano delle particelle di un terreno per verificare gli effetti di alcuni fertilizzanti è verosimile che particelle adiacenti siano più simili tra loro che non quelle distanti. A questo punto potrebbe essere importante non assegnare particelle adiacenti ad un unico trattamento ma distribuirli casualmente tra le varie particelle. Lo stesso dicasi per serie temporali dove i dati monitorati al tempo t_1 sono influenzati da quelli osservati al tempo t_0 . In questo caso occorre utilizzare analisi statistiche appropriate (analisi longitudinale dei dati) che tengono in considerazione questi effetti di **"autocorrelazione"** tra i dati monitorati nei diversi campioni spaziali o temporali.
- C) Prima di eseguire il test di F occorre **verificare che le varianze dei vari gruppi siano omogenee** (omoscedasticità delle varianze).

Tra i **vari tests di omogeneità**, Bartlett, Scheffè, Hartley, Levene, ecc.. riportiamo quello di **Cochran** che, se i **dati** sono **bilanciati** (gruppi di uguale numerosità), permette una verifica rapida:

$$R_{n,p} = s_{max}^2 / (s_1^2 + s_2^2 + \dots + s_p^2)$$

Nel nostro esempio: $R_{5,7} = 0,2388592$ vs $0,4783$ il valore critico, per n (5) ripetizioni e p (7) gruppi con un livello di probabilità di **0,05**. Pertanto, essendo R **inferiore** al valore critico, l'ipotesi della omogeneità (omoscedasticità) delle varianze viene accettata ed è lecito andare avanti con l'ANOVA.

Per **dati non bilanciati**, ovvero per gruppi che hanno numerosità diverse, possiamo utilizzare il test di Bartlett.

Facciamo ora un esempio pratico considerando 8 classi di età di giovani alpaca, le quali contengono un differente numero di animali (n) su cui è stato misurato il diametro medio delle fibre (in μm) che costituiscono il vello di questi animali

età	Mean	S. D.	n	variance	df	Weight sum of variance	Ln variance	Weight sum of Ln variance
0	17,88	4,46	401	19,89	400	7954,25	2,99	1196,00
15	17,05	4,61	400	21,21	399	8463,19	3,05	1218,75
30	18,07	5,39	1201	29,00	1200	34802,1249	3,37	4040,83
60	17,90	5,34	2201	28,52	2200	62737,20	3,35	7371,09
90	18,87	5,82	2600	33,85	2599	87966,2765	3,52	9153,23
120	19,96	5,60	2000	31,37	1999	62707,262	3,45	6888,21
150	19,54	5,87	1199	34,51	1198	41342,0234	3,54	4242,39
180	19,02	5,61	400	31,44	399	12543,84	3,45	1375,76
Total	18,76	5,60	10402	229,78	10394	318516,17	26,72	35486,27

$$\text{Weight variance} = 30,64 \quad \begin{array}{|c|} \hline \text{Ln} \\ \hline \end{array} \quad \begin{array}{|c|} \hline \text{quantity 3} \\ \hline \end{array}$$

$$(\text{quantity 2} \times \text{quantity 3}) - \text{quantity 4} = \text{CHI}^2 = 86,62 \quad \text{Sign. } 6,1E-16$$

con $a-1$ df
= 7

Innanzitutto occorre sommare le varianze ed ovviamente, essendo queste ottenute da gruppi con diversa numerosità, occorre fare una somma pesata (ovvero moltiplicare le singole varianze per i gradi di libertà²), poi dividiamo questa somma per il totale dei gradi di libertà (quantity 2 = 10.394). Così facendo otteniamo una media ponderata della varianza relativa a tutti gli 8 gruppi di età considerati (weight variance = 30,64 e ne calcoliamo il logaritmo naturale (quantity 3 = 3,42). Calcoliamo anche il logaritmo naturale di ciascuna varianza e poi facciamo una somma pesata di questi valori (quantity 4 = 35.486,27). Possiamo ora applicare il test di Bartlett svolgendo la seguente formula:

$$(\text{quantity 2} \times \text{quantity 3}) - \text{quantity 4} = (10.394 \times 3,42) - 35.486,27 = 86,62$$

Confrontiamo ora il valore ottenuto con il corrispettivo valore critico di Chi^2 , per 8-1 gradi di libertà che è 14,0667. Dato che il valore ottenuto con il test di Bartlett è molto superiore a quello critico possiamo scartare l'ipotesi nulla e dire che le varianze non sono omogenee tra loro, per cui non è soddisfatta l'assunzione dell'omogeneità delle varianze necessaria per eseguire l'ANOVA.

² Così facendo in effetti otteniamo le devianze. Quindi potevamo anche dire che occorre sommare le devianze.