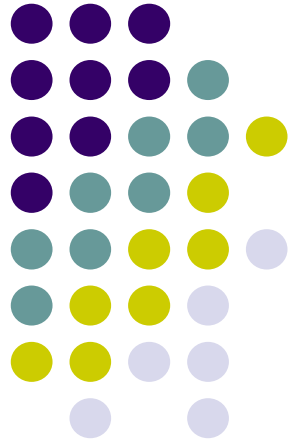
Levine, Krehbiel, Berenson

Statistica II ed.

Casa editrice: Pearson

Capitolo 10

Test basati su due campioni e ANOVA a una via



Insegnamento: Statistica
Corsi di Laurea Triennale in Economia

Dipartimento di Economia e Management, Università di Ferrara
Docenti: Prof. Stefano Bonnini, Dott.ssa Angela Grassi

Argomenti

- Confronto tra le medie di due popolazioni indipendenti
- Confronto tra le medie di due popolazioni non indipendenti
- Confronto tra le proporzioni di due popolazioni
- Test *F* per la differenza tra due varianze
- Analisi della varianza (ANOVA) ad una via

Confronto tra medie di due pop. indipendenti

- Consideriamo due popolazioni indipendenti e supponiamo di estrarre un campione di ampiezza n_1 dalla prima popolazione di ampiezza n_2 dalla seconda popolazione
- Siano μ_1 e μ_2 le medie che caratterizzano rispettiva-mente la prima e la seconda popolazione e si assumano i due scarti quadratici medi σ_1 e σ_2 come noti
- Si vuole verificare l'ipotesi nulla che le medie delle due popolazioni (indipendenti) sono uguali tra loro

$$H_0: \mu_1 = \mu_2 \ (\mu_1 - \mu_2 = 0)$$

contro l'ipotesi alternativa

$$H_1: \mu_1 \neq \mu_2 \ (\mu_1 - \mu_2 \neq 0)$$

- A questo scopo viene definita la **statistica test *Z* per la differenza tra le due medie**

Confronto tra medie di due pop. indipendenti

Test Z per la differenza fra due medie

La statistica test in questo caso è:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (7.1)$$

dove:

$\bar{X}_1$ = media degli elementi del campione estratto dalla popolazione 1
μ_1 = media della popolazione 1
σ_1^2 = varianza della popolazione 1
n_1 = ampiezza del campione estratto dalla popolazione 1
$\bar{X}_2$ = media degli elementi del campione estratto dalla popolazione 2
μ_2 = media della popolazione 2
σ_2^2 = varianza della popolazione 2
n_2 = ampiezza del campione estratto dalla popolazione 2

Confronto tra medie di due pop. indipendenti

- Se si assume che i due campioni siano estratti casualmente ed indipendentemente da due popolazioni normali la statistica *Z* ha distribuzione normale
- Se le due popolazioni non hanno distribuzione normale il test *Z* può essere utilizzato con ampiezza campionarie sufficientemente elevate (in virtù del teorema del limite centrale)
- In molti casi le varianze delle due popolazioni non sono note. Se si assume l'ipotesi di omogeneità della varianze ($\sigma^2_1=\sigma^2_2$), per verificare se c'è una differenza significativa tra le medie delle due popolazioni è possibile utilizzare il **test *t* basato sulle varianze campionarie combinate**
- Il test *t* è appropriato se le popolazioni hanno distribuzione normale oppure, in caso contrario, se le ampiezze campionarie sono sufficientemente elevate

Confronto tra medie di due pop. indipendenti

Test *t* per la differenza fra due medie basato sulle varianze campionarie ponderate

La statistica test in questo caso è:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (7.2)$$

dove

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

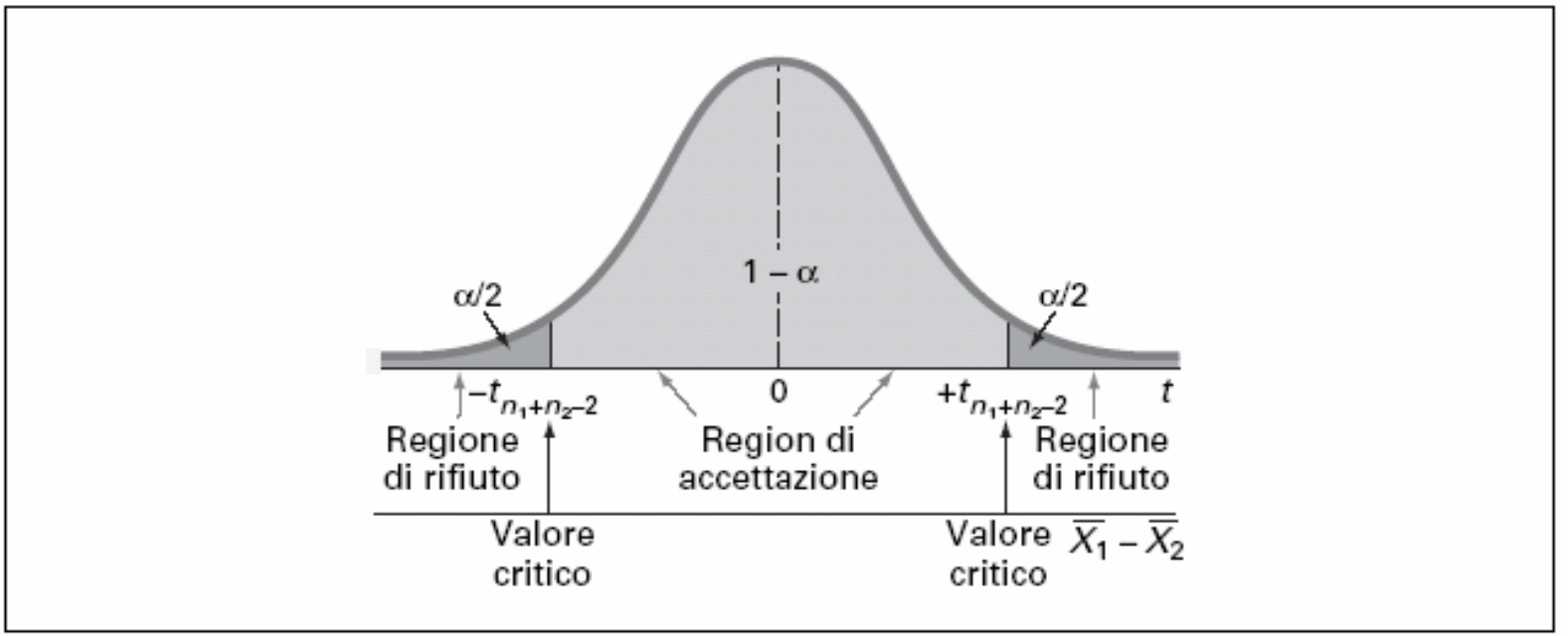
e

S_p^2 = varianza ponderata
$\bar{X}_1$ = media degli elementi del campione estratto dalla popolazione 1
S_1^2 = varianza degli elementi del campione estratto dalla popolazione 1
n_1 = ampiezza del campione estratto dalla popolazione 1
$\bar{X}_2$ = media degli elementi del campione estratto dalla popolazione 2
S_2^2 = varianza degli elementi del campione estratto dalla popolazione 2
n_2 = ampiezza del campione estratto dalla popolazione 2

Si dimostra che la statistica test t sotto l'ipotesi nulla si distribuisce secondo una t di Student con $n_1 + n_2 - 2$ gradi di libertà.

Confronto tra medie di due pop. indipendenti

Regione di rifiuto e di accettazione per la differenza tra due medie utilizzando la statistica test *t* basata sulle varianze combinate (test a due code)

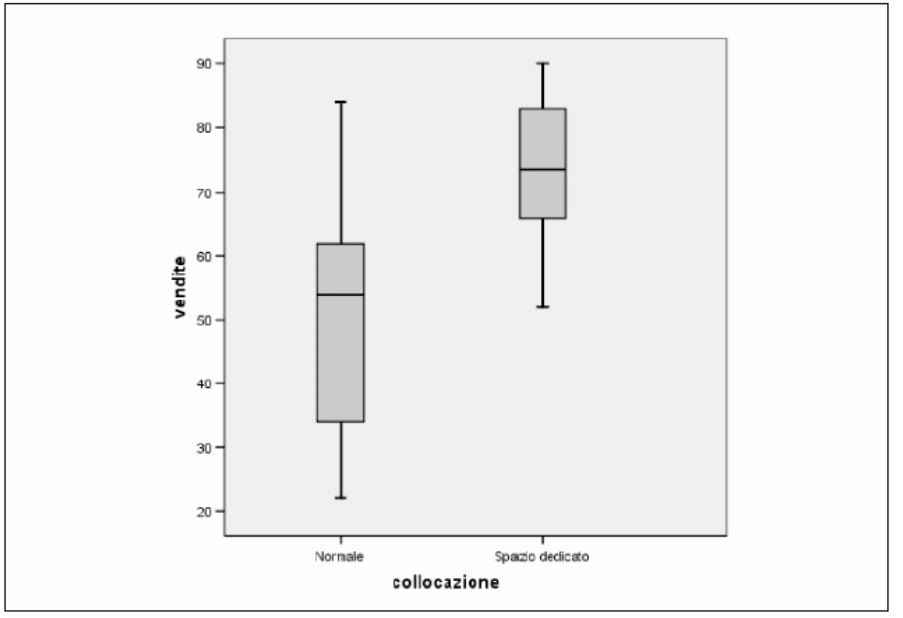


Quando l'assunzione dell'omogeneità delle varianze non è plausibile occorre fare riferimento al **test *t* con varianze diverse** (ricorrendo all'Excel o ad altri software statistici)

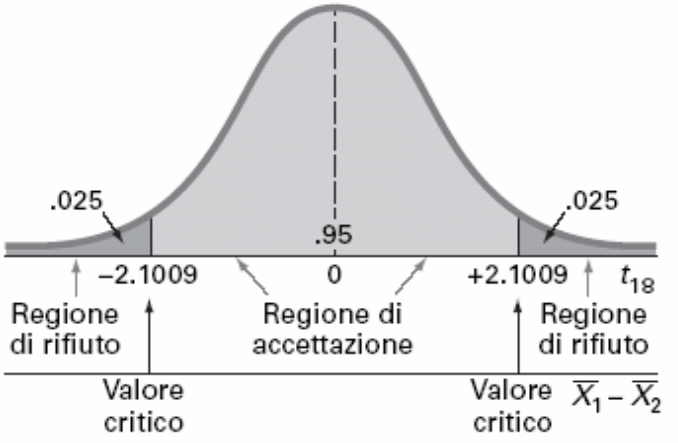
Confronto tra medie di due pop. indipendenti

Esempio: confronto tra le vendite settimanali (numero di pezzi venduti) della BLK cola in due gruppi di supermercati, dove il primo adotta la collocazione a scaffale mentre il secondo utilizza uno spazio dedicato

Collocazione									
Scaffale					Spazio dedicato				
22	34	52	62	30	52	71	76	54	67
40	64	84	56	59	83	66	90	77	84



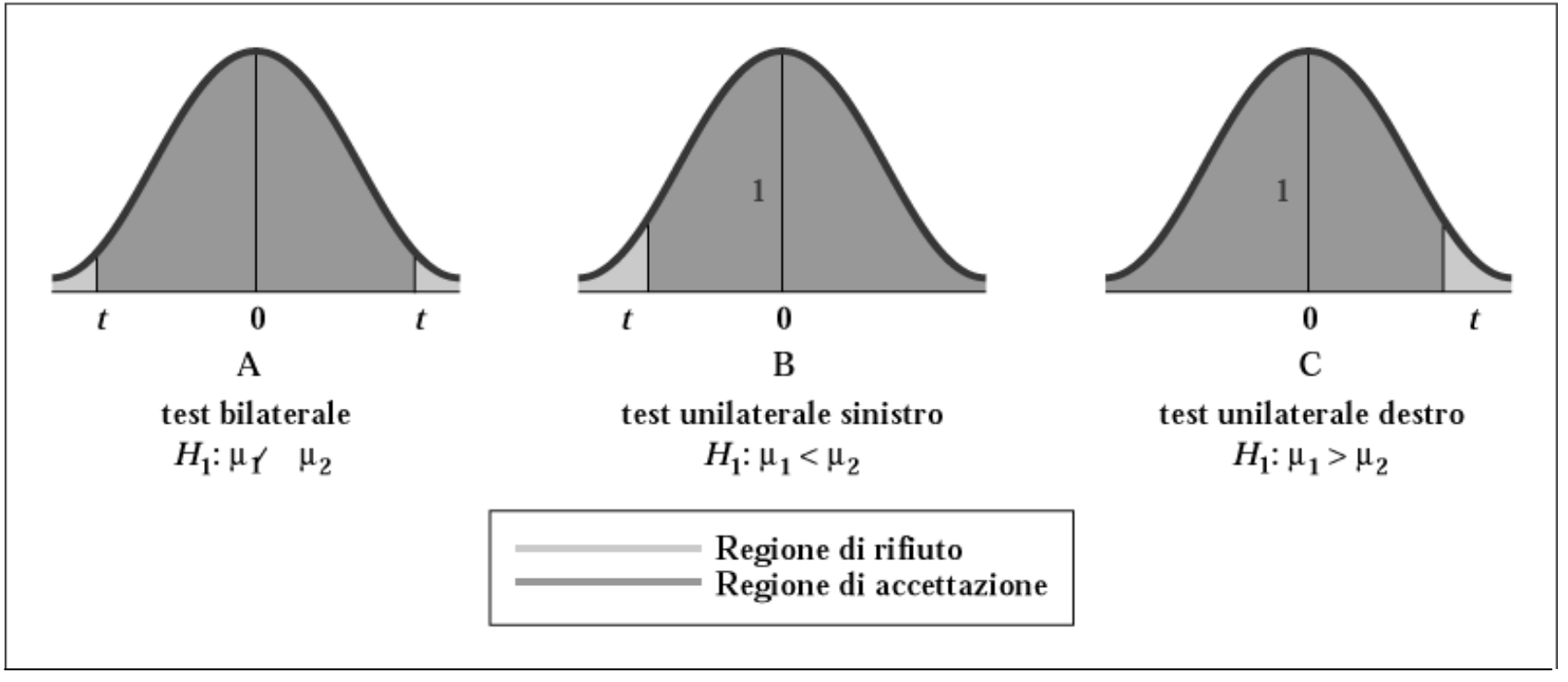
Confronto tra medie di due pop. indipendenti



	A	B	C
1	**Test t: due campioni assumendo uguale varianza**		
2			
3		*Scaffale*	*Spazio dedicato*
4	**Media**	**50,3**	**72**
5	**Varianza**	**350,6778**	**157,3333**
6	**Osservazioni**	**10**	**10**
7	**Varianza complessiva**	**254,0056**	
8	**Differenza ipotizzata per le r**	**0**	
9	**gdl**	**18**	
10	**Stat t**	**-3,0446**	
11	**P(T<=t) una coda**	**0,0035**	
12	**t critico una coda**	**1,7341**	
13	**P(T<=t) due code**	**0,0070**	
14	**t critico due code**	**2,1009**	

Confronto tra medie di due pop. indipendenti

In base al fatto che l'ipotesi alternativa sia nella forma A: H_1:$\mu_1 \neq \mu_2$ oppure B: H_1:$\mu_1 < \mu_2$ o C: H_1:$\mu_1 > \mu_2$ si parla di test ad una coda e test a due code



Intervallo di confidenza per la differenza tra le medie di due pop. indipendenti

Anziché (o oltre a) sottoporre a verifica l'ipotesi nulla secondo la quale due medie sono uguali, possiamo utilizzare l'equazione (10.3) per ottenere un intervallo di confidenza per la differenza tra le medie μ_1 e μ_2 delle due popolazioni:

Intervallo di confidenza per la differenza ($\mu_1 - \mu_2$)

$$(\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;\alpha/2} \cdot \sqrt{S_p^2 \left(1/n_1 + 1/n_2\right)} \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2;\alpha/2} \sqrt{S_p^2 \left(1/n_1 + 1/n_2\right)} \quad \textbf{(10.3)}$$

dove $t_{n_1+n_2-2;\alpha/2}$ è il valore critico a cui corrisponde un'area cumulata pari a (1−α/2) della distribuzione t di Student con (n_1+n_2−2) gradi di libertà.

Confronto tra medie di 2 pop. non indipendenti

Ci sono situazioni in cui le due popolazioni poste a confronto non sono indipendenti di modo che il campione estratto dalla prima popolazione non è indipendente dal campione estratto dalla seconda:

1. **campioni appaiati** (individui o casi che condividono una stessa caratteristica)
2. **misurazioni ripetute** (stesso insieme di individui o casi)

L'attenzione si sposta sulla differenze tra i valori nei due campioni:

	Gruppo		
Valore	1	2	Differenza
1	X_{11}	X_{21}	$D_1 = X_{11} - X_{21}$
2	X_{12}	X_{22}	$D_2 = X_{12} - X_{22}$
.	.	.	.
.	.	.	.
.	.	.	.
i	X_{1i}	X_{2i}	$D_i = X_{1i} - X_{2i}$
.	.	.	.
.	.	.	.
.	.	.	.
n	X_{1n}	X_{2n}	$D_n = X_{1n} - X_{2n}$

Confronto tra medie di 2 pop. non indipendenti

Quindi verificare l'ipotesi di uguaglianza delle medie μ_1 e μ_2 di due popolazioni non indipendenti equivale a verificare ipotesi di uguaglia a zero della media della differenza *D* tra le due popolazioni, cioè H_0: $\mu_D=0$. Se lo scarto quadratico medio della popolazione delle differenze σ_D è noto, allora il test di riferimento è basato sulla statistica *Z*. In caso σ_D sia ignoto si può fare ricorso al **test *t* su campioni appaiati**.

Statistica test *Z* per la media delle differenze

$$Z = \frac{\bar{D} - \mu_D}{\sigma_D / \sqrt{n}}, \text{con } \bar{D} = 1/n \sum_{i=1}^{n} D_i \qquad \textbf{(10.4)}$$

Statistica test *t* per la media delle differenze

$$t = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}}, \text{con } \bar{D} = \frac{\sum_{i=1}^{n} D_i}{n} \text{ e } S_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{(n-1)}} \qquad \textbf{(10.5)}$$

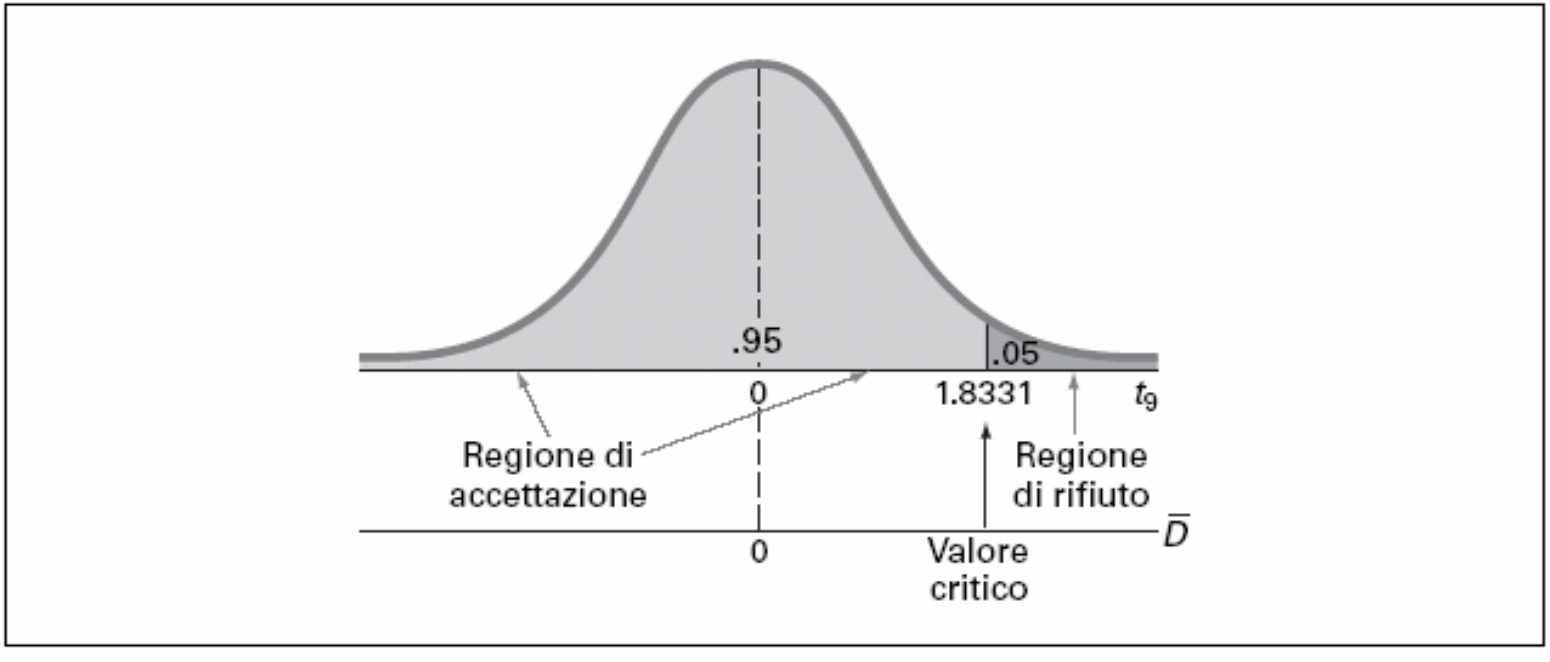
Confronto tra medie di 2 pop. non indipendenti

Esempio: Misurazioni ripetute del tempo (in secondi) di elaborazione di un progetto utilizzando due diversi software

Tempo di elaborazione (in secondi)			
Progetto	Software sul mercato	Nuovo software	Differenza (D_i)
1	9.98	9.88	+0.10
2	9.88	9.86	+0.02
3	9.84	9.75	+0.09
4	9.99	9.80	+0.19
5	9.94	9.87	+0.07
6	9.84	9.84	0.00
7	9.86	9.87	–0.01
8	10.12	9.86	+0.26
9	9.90	9.83	+0.07
10	9.91	9.86	+0.05
			+0.84

Confronto tra medie di 2 pop. non indipendenti

Test *t* a una coda per la differenza tra le medie di due popolazioni non indipendenti a un livello di significatività pari a 0.05 e con 9 gradi di libertà



Intervallo di confidenza per la differenza tra le medie di due pop. non indipendenti

Anziché (o oltre a) sottoporre a verifica l'ipotesi nulla secondo la quale due medie sono uguali, possiamo utilizzare l'equazione (10.6) per ottenere un intervallo di confidenza per la differenza μ_D:

Intervallo di confidenza per la differenza tra le medie di due popolazioni non indipendenti

$$\bar{D} - t_{n-1;\alpha/2} S_D / \sqrt{n} \leq \mu_D \leq \bar{D} + t_{n-1;\alpha/2} S_D / \sqrt{n} \qquad \textbf{(10.6)}$$

dove $t_{n-1;\alpha/2}$ è il valore critico a cui corrisponde un'area cumulata pari a $(1-\alpha/2)$ della distribuzione t di Student con $(n-1)$ gradi di libertà

Confronto tra le proporzioni di due popolazioni

- Spesso si è interessati a effettuare confronti e ad analizzare differenze tra due popolazioni con riferimento alla proporzione di casi con una certa caratteristica
- Per confrontare due proporzioni sulla base dei risultati di due campioni si può ricorrere al **test *Z* per la differenza tra due proporzioni**, la cui statistica test ha distribuzione approssimativamente normale quando le ampiezza campionarie sono sufficientemente elevate

Statistica *Z* per la differenza tra due proporzioni (10.7)

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ con } \bar{p} = \frac{X_1 + X_2}{n_1 + n_2},\ p_1 = \frac{X_1}{n_1},\ p_2 = \frac{X_2}{n_2}$$

Confronto tra le proporzioni di due popolazioni

- A seconda di come è formulata l'ipotesi alternativa avremo un test a due code (H_1: $\pi_1 \neq \pi_2$ ($\pi_1 - \pi_2 \neq 0$)) o un test a una coda (ipotesi direzionali: H_1: $\pi_1 > \pi_2$ ($\pi_1 - \pi_2 > 0$) oppure H_1: $\pi_1 < \pi_2$ ($\pi_1 - \pi_2 < 0$))
- Esempio

  La catena di alberghi *TC Resort* è interessata a valutare se esiste differenza tra la proporzione di clienti che intendono visitare nuovamente due dei suoi alberghi. Vengono campionati 227 clienti nel primo albergo e 262 dal secondo di cui 163 si dicono disposti a ritornare nel primo campione, 154 nel secondo.

  Adottando un livello di significatività pari a 0.05 si può affermare che nei due alberghi esiste una differenza tra la proporzione di coloro che sono disposti a ritornare?

Confronto tra le proporzioni di due popolazioni

Le ipotesi da verificare sono:

H_0: $\pi_1 = \pi_2$ oppure $\pi_1 - \pi_2 = 0$

H_1: $\pi_1 \neq \pi_2$ oppure $\pi_1 - \pi_2 \neq 0$

Al livello di significatività 0.05, i valori critici della normale standardizzata sono -1.96 e 1.96, e la regola decisionale:

Rifiuta H_0 se $Z > +1.96$

oppure $Z < -1.96$;

altrimenti accetta H_0.

$$p_{s_1} = \frac{X_1}{n_1} = \frac{163}{227} = 0.718 \qquad p_{s_2} = \frac{X_2}{n_2} = \frac{154}{262} = 0.588$$

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{163 + 154}{227 + 262} = \frac{317}{489} = 0.648$$

$$Z = \frac{(0.718 - 0.588) - 0}{\sqrt{(0.648)(0.352)\left(\frac{1}{227} + \frac{1}{262}\right)}} = \frac{0.13}{\sqrt{(0.228)(0.0082)}}$$

$$= \frac{0.13}{\sqrt{0.00187}} = \frac{0.13}{0.0432} = +3.01$$

Z= + 3,01 > +1,96 perciò si rifiuta H_0 concludendo che le percentuali sono diverse

Intervallo di confidenza per la differenza tra due proporzioni

Anziché (o oltre a) sottoporre a verifica l'ipotesi nulla secondo la quale due proporzioni sono uguali, possiamo utilizzare l'equazione (10.8) per ottenere un intervallo di confidenza per la differenza tra le due proporzioni

Intervallo di confidenza per la differenza tra due proporzioni

$$(p_1 - p_2) - Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \leq (\pi_1 - \pi_2) \leq$$

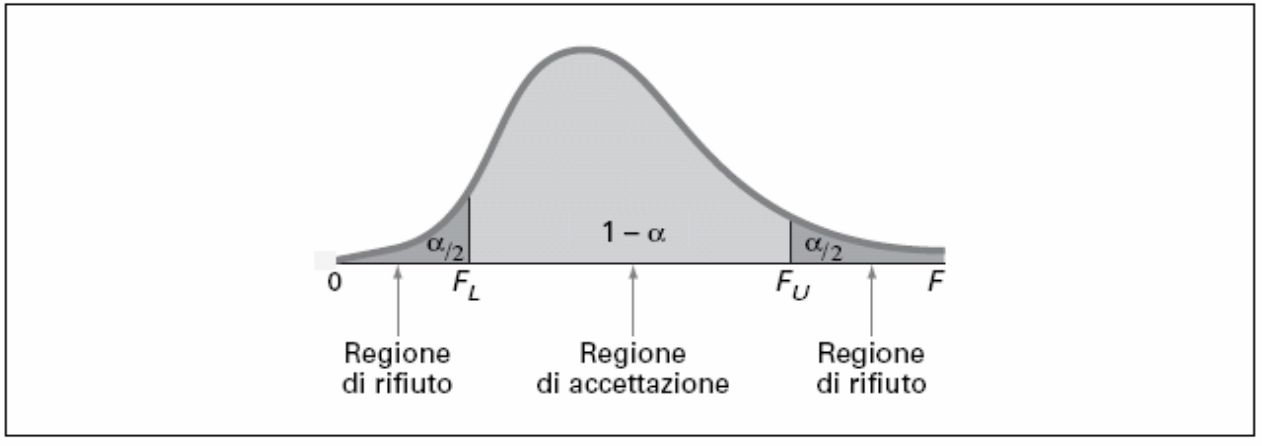
$$\leq (p_1 - p_2) + Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \qquad \textbf{(10.8)}$$

Test *F* per la differenza tra due varianze

- Talvolta si pone il problema di valutare l'ipotesi di omogeneità delle varianze e a questo scopo è possibile considerare un test statistico per verificare H_0: $\sigma^2_1 = \sigma^2_2$ contro l'ipotesi alternativa H_1: $\sigma^2_1 \neq \sigma^2_2$. Questo test è basato sul rapporto delle due varianze campionarie

$$F = S^2_1 / S^2_2 \qquad \textbf{(10.9)}$$

- La statistica test *F* segue una distribuzione *F* con (n_1-1) e (n_2-1) gradi di libertà rispettivamente a numeratore e a denominatore



Test *F* per la differenza tra due varianze

Esempio: determinazione del valore critico superiore F_U di una distribuzione F con 9 e 9 gradi di libertà corrispondente a un'area nella coda destra pari a 0.025.

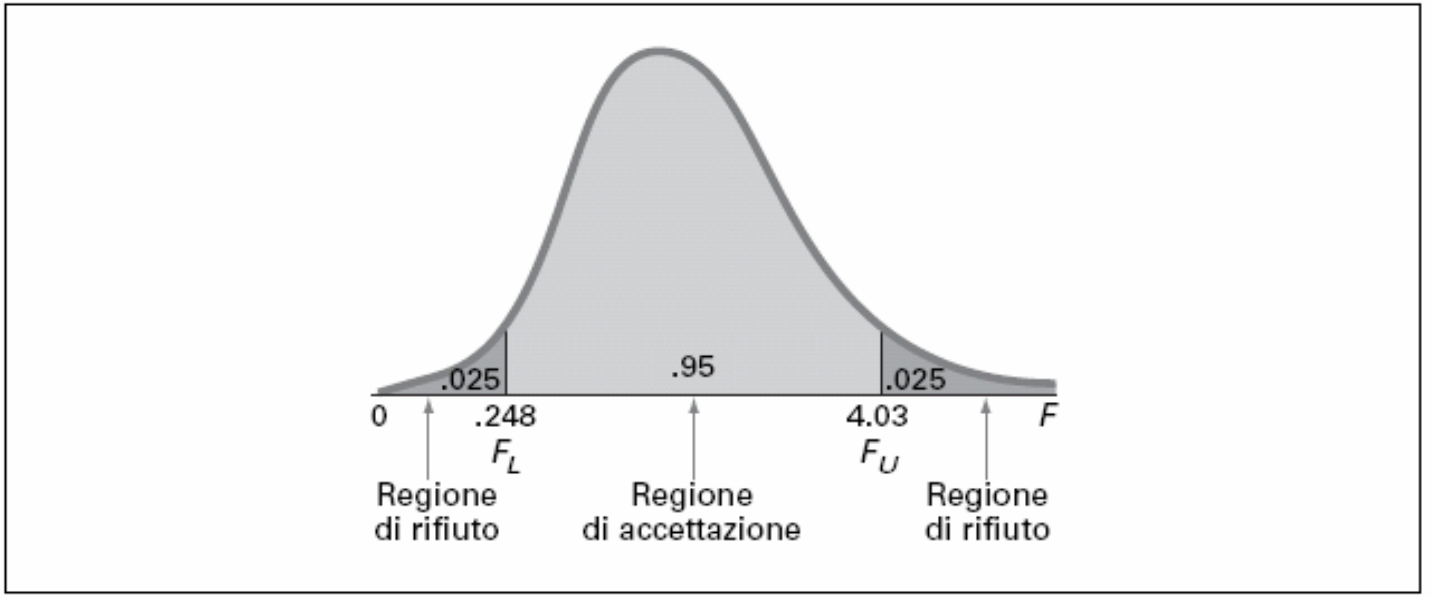
Denominatore	Numeratore df_1						
df_2	1	2	3	...	7	8	9
1	647.80	799.50	864.20	...	948.20	956.70	963.30
2	38.51	39.00	39.17	...	39.36	39.37	39.39
3	17.44	16.04	15.44	...	14.62	14.54	14.47
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
7	8.07	6.54	5.89	...	4.99	4.90	4.82
8	7.57	6.06	5.42	...	4.53	4.43	4.36
9	7.21	5.71	5.08	...	4.20	4.10	4.03

Fonte: estratto dalla Tavola E.5

Esiste un modo molto semplice per determinare il valore critico inferiore F_L: $F_L = 1/F_U^*$, dove F_U^* è il valore critico superiore delle distribuzione F con gradi di libertà invertiti, cioè (n_2-1) a numeratore e (n_1-1) a denominatore

Test *F* per la differenza tra due varianze

Regioni di rifiuto e di accettazione per un test *F* a due code sull'uguaglianza tra due varianze a un livello di significatività pari a 0.05, con 9 e 9 gradi di libertà



Nella verifica di ipotesi sulla omogeneità delle varianze si ipotizza che le due popolazioni siano normali. La statistica *F* non è robusta rispetto a violazioni di questa assunzione

Analisi della varianza (ANOVA) ad una via

- Finora abbiamo descritto test di ipotesi finalizzati alla verifica di ipotesi sulla differenza tra parametri di due popolazioni
- Spesso si presenta la necessità di prendere in considerazione esperimenti od osservazioni relative a più di due **gruppi** individuati sulla base di un **fattore** di interesse
- I gruppi sono quindi formati secondo i **livelli** assunti da un fattore, ad esempio
  - la temperatura di cottura di un oggetto in ceramica che assume diversi *livelli numerici* come 300°, 350°,400°,450° oppure
  - il fornitore che serve una azienda può assumere diversi *livelli qualitativi* come Fornitore 1, Fornitore 2, Fornitore 3, Fornitore 4

Analisi della varianza (ANOVA) ad una via

- L'**analisi della varianza** (o **ANOVA**, **AN**alysis **O**f **VA**riance) è una tecnica che consente di confrontare da un punto di vista inferenziale le medie di più di due gruppi (popolazioni)
- Quando i gruppi sono definiti sulla base di un singolo fattore si parla di **analisi della varianza a un fattore** o **a una via**
- Questa procedura, basata su un test *F*, è una estensione a più gruppi del test *t* per verificare l'ipotesi sulla differenza tra le medie di due popolazioni indipendenti
- Anche se si parla di analisi della varianza in realtà l'oggetto di interesse sono le differenze tra medie nei diversi gruppi e proprio tramite l'analisi della variabilità all'interno dei gruppi e tra gruppi che siamo in grado di trarre delle conclusioni sulla differenza delle medie

Analisi della varianza (ANOVA) ad una via

- **La variabilità all'interno dei gruppi** è considerata un **errore casuale**, mentre la **variabilità tra i gruppi** è attribuibile alle differenza tra i gruppi, ed è anche chiamata **effetto del trattamento**
- Ipotizziamo che *c* gruppi rappresentino popolazioni con distribuzione normale, caratterizzate tutte dalla stessa varianza e che le osservazioni campionarie siano estratte casualmente ed indipendentemente dai *c* gruppi
- In questo contesto l'ipotesi nulla che si è interessati a verificare è che le medie di tutti gruppi siano uguali tra loro

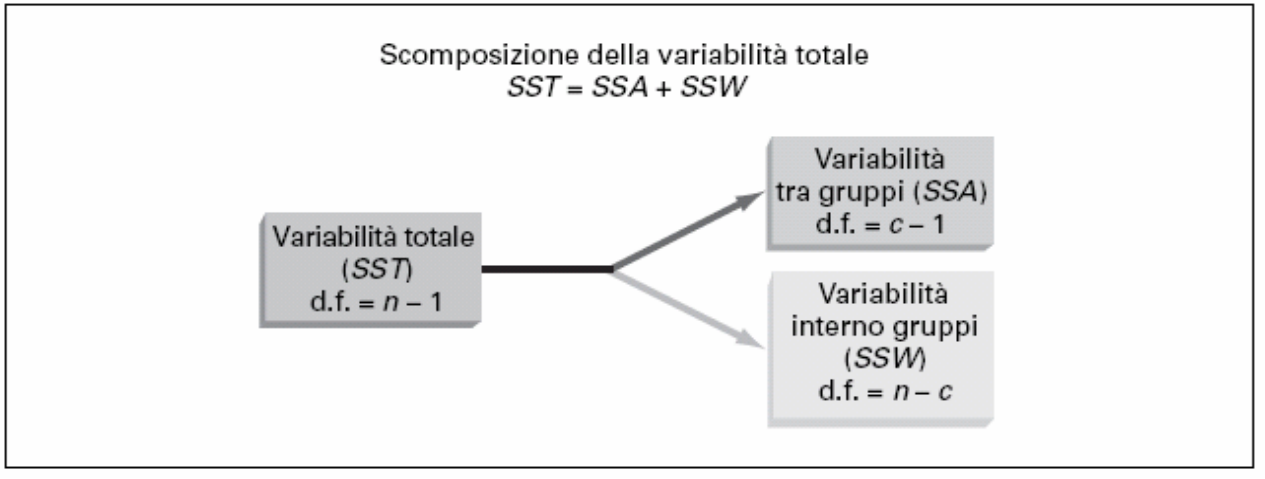
$$H_0: \mu_1 = \mu_2 = \ldots = \mu_c$$

contro l'ipotesi alternativa

H_1: non tutte le μ_j sono uguali tra loro (con j=1,…,c)

Analisi della varianza (ANOVA) ad una via

- Per verificare le due ipotesi considerate, la variabilità totale (misurata dalla **somma dei quadrati totale – *SST***) viene scomposta in due componenti: una componente attribuibile alla differenza tra i gruppi (misurata dalla **somma dei quadrati tra i gruppi – *SSA***) e una seconda componente che si riferisce alle differenze riscontrare all'interno del gruppi (misurata dalla **somma dei quadrati all'interno dei gruppi – *SSW***)



Analisi della varianza (ANOVA) ad una via

- Poiché sotto l'ipotesi nulla si assume che le medie dei gruppi siano tutti uguali, la variabilità totale *SST* si ottiene sommando le differenze al quadrato di ciascuna osservazione e la media complessiva, indicata con $\overline{\overline{X}}$

Variabilità totale nell'ANOVA a una via

$$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{\overline{X}}\right)^2 \qquad \textbf{(10.11)}$$

dove $\overline{\overline{X}} = \dfrac{\sum_{j=1}^{c}\sum_{i=1}^{n_j} X_{ij}}{n} = \text{media complessiva}, n = \sum_{j=1}^{c} n_j$

- *SST* è caratterizzata da (n−1) gradi di libertà poiché ciascuna osservazione X_{ij} viene confrontata con la media campionaria complessiva $\overline{\overline{X}}$

Analisi della varianza (ANOVA) ad una via

- La variabilità tra gruppi *SSA* si ottiene sommando le differenze al quadrato tra le medie campionarie di ciascun gruppo, $\bar{X}_j$, e la media complessiva, $\bar{\bar{X}}$, dove ogni differenza è ponderata con l'ampiezza campionaria n_j del gruppo a cui è riferita

Variabilità tra gruppi nell'ANOVA a una via

$$SSA = \sum_{j=1}^{c} n_j \left(\bar{X}_j - \bar{\bar{X}} \right)^2 \qquad \textbf{(10.12)}$$

dove $\bar{X}_j = \dfrac{\sum_{i=1}^{n_j} X_{ij}}{n_j}$ media campionaria nel *j*-esimo campione

- Poiché si tratta di confrontare *c* gruppi, *SSA* sarà caratterizzata da ($c-1$) gradi di libertà

Analisi della varianza (ANOVA) ad una via

- Infine, la variabilità nei gruppi *SSW* si ottiene sommando le differenze al quadrato tra ciascuna osservazione e la media campionaria del gruppo a cui appartiene

Variabilità all'interno dei gruppi nell'ANOVA a una via

$$SSW = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(X_{ij} - \bar{X}_j \right)^2 \qquad \textbf{(10.13)}$$

- Poiché ciascuno dei c gruppi contribuisce con $(n_j - 1)$ gradi di libertà, *SSW* avrà complessivamente $(n - c) = \sum(n_j - 1)$ gradi di libertà
- Dividendo ciascuna somma dei quadrati per i rispettivi gradi di libertà, si ottengono tre varianze, o **medie dei quadrati – *MSA*** (la media dei quadrati tra gruppi), ***MSW*** (la media dei quadrati all'interno dei gruppi) e ***MST*** (la media dei quadrati totale)

Analisi della varianza (ANOVA) ad una via

- Se l'ipotesi nulla è vera e non ci sono differenze significative tra le medie dei gruppi, le tre medie dei quadrati – *MSA*, *MSW* e *MST,* che sono esse stesse delle stime di varianze e rappresentano tutte stime della varianza globale della popolazione sottostante
- Quindi per verificare l'ipotesi nulla contro l'alternativa si fa riferimento alla **statistica test *F* per l'ANOVA a una via**, ottenuta come rapporto tra *MSA* e *MSW*

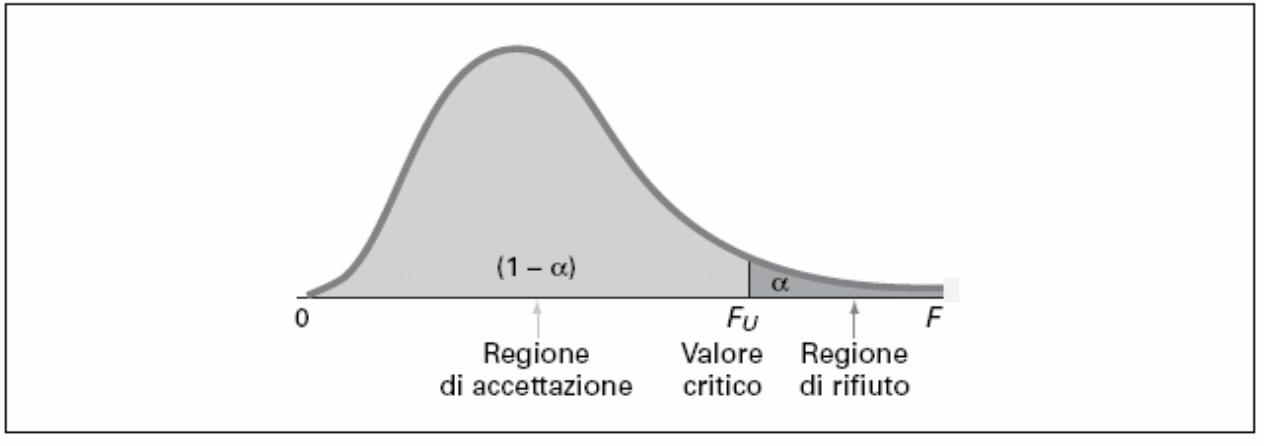
Statistica test *F* per l'ANOVA a una via

$$F = \frac{SSA/(n-c)}{SSW/(c-1)} = \frac{MSA}{MSW} \qquad \textbf{(10.15)}$$

- Se l'ipotesi nulla è vera, la realizzazione della statistica *F* dovrebbe essere approssimativamente 1, mentre se H_0 è falsa ci aspettiamo valori significativ. superiori all'unità

Analisi della varianza (ANOVA) ad una via

- La statistica *F* ha distribuzione *F* con (*c*−1) gradi di libertà al numeratore e (*n*−*c*) gradi di libertà al denominatore
- Quindi, fissato il livello di significatività α, l'ipotesi nulla dovrà essere rifiutata se il valore osservato della statistica test è maggiore del valore critico F_U di una distribuzione *F* con (*c*−1) e (*n*−*c*) gradi di libertà



Analisi della varianza (ANOVA) ad una via

- I risultati del test *F* per l'ANOVA vengono solitamente riportati nella cosiddetta **tabella dell'ANOVA**

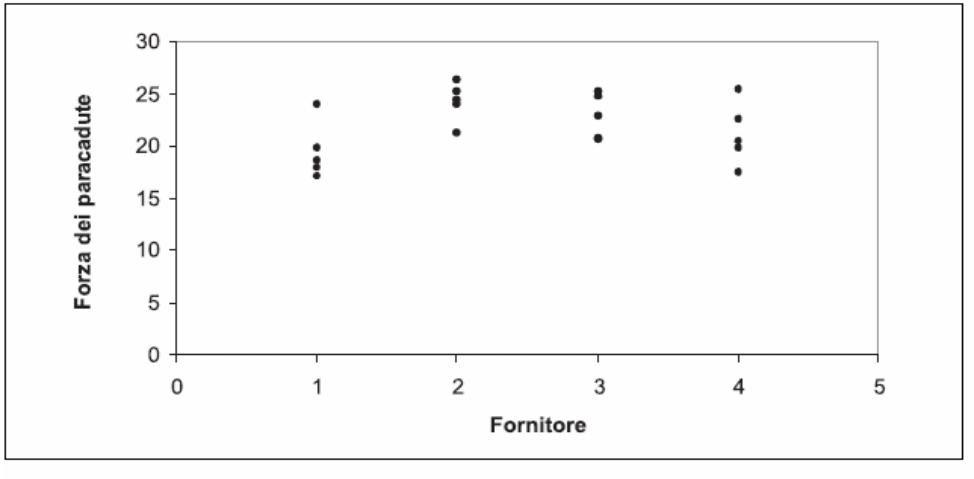
Fonte	Gradi di libertà	Somme dei quadrati	Medie dei quadrati (varianze)	*F*
Tra i gruppi	$c-1$	SSA	$MSA = \frac{SSA}{c-1}$	$F = \frac{MSA}{MSW}$
All'interno dei gruppi	$n-c$	SSW	$MSW = \frac{SSW}{n-c}$	

- Nella tabella dell'ANOVA viene solitamente riportato anche il *p*-value, cioè la probabilità di osservare un valore di *F* maggiore o uguale a quello osservato, nel caso l'ipotesi nulla sia vera. Come usuale, l'ipotesi nulla di uguaglianza tra le medie dei gruppi deve essere rifiutata quando il *p*-value è inferiore al livello di significatività scelto

Analisi della varianza (ANOVA) ad una via

- Esempio: una azienda produttrice di paracadute, vuole confrontare la resistenza dei paracadute prodotti con fibre sintetiche acquistate da quattro diversi fornitori

	A	B	C	D	E	F
1			Fornitore 1	Fornitore 2	Fornitore 3	Fornitore 4
2			18,5	26,3	20,6	25,4
3			24,0	25,3	25,2	19,9
4			17,2	24,0	20,8	22,6
5			19,9	21,2	24,7	17,5
6			18,0	24,5	22,9	20,4
7		Media	19,52	24,26	22,84	21,16
8		Scarto quadratico medio	2,69	1,92	2,13	2,98

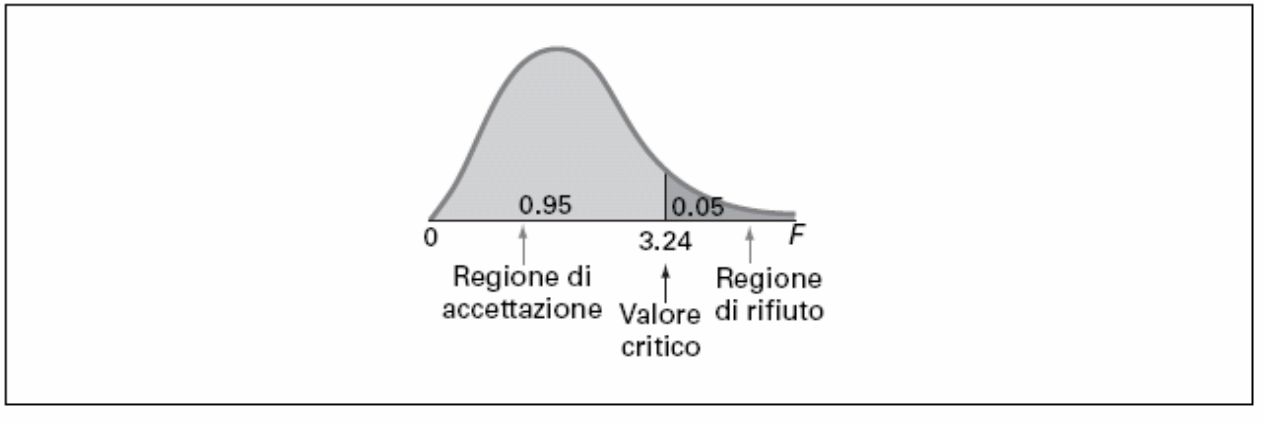


Analisi della varianza (ANOVA) ad una via

- Fissiamo α=0.05 e identifichiamo nelle tavole il valore critico di interesse

Numeratore df_1									
Denominatore df_2	1	2	3	4	5	6	7	8	9
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54

Fonte: estratto dalla Tavola E.5.



Analisi della varianza (ANOVA) ad una via

- Poiché il valore osservato della statistica test è F=3.46< 3.24=F_U l'ipotesi nulla deve essere rifiutata e si conclude che la resistenza media dei paracadute varia in modo significativo a seconda del fornitore

	A	B	C	D	E	F	G
1	Analisi varianza: ad un fattore						
2							
3	RIEPILOGO						
4	*Gruppi*	*Conteggio*	*Somma*	*Media*	*Varianza*		
5	Fornitore 1	5	97,6	19,52	7,237		
6	Fornitore 2	5	121,3	24,26	3,683		
7	Fornitore 3	5	114,2	22,84	4,553		
8	Fornitore 4	5	105,8	21,16	8,903		
9							
10							
11	ANALISI VARIANZA						
12	*Origine della variazione*	*SQ*	*gdl*	*MQ*	*F*	*Valore di significatività*	*F crit*
13	Tra gruppi	63,2855	3	21,0952	3,4616	0,0414	3,2389
14	In gruppi	97,504	16	6,094			
15							
16	Totale	160,7895	19				

Analisi della varianza (ANOVA) ad una via

- **Procedura di Tukey-Cramer**
  Quando si rifiuta l'ipotesi nulla del *F* per l'ANOVA, viene stabilito che ci sono almeno due medie significativa-mente diverse tra loro.
  Per identificare quali sono i gruppi che effettivamente differiscono tra loro si deve utilizzare una ulteriore procedura che rientra nei cosiddetti metodi dei **confronti multipli**
  Tra questi metodi, la procedura di Tukey-Cramer consente di effettuare simultaneamente confronti a due a due tra tutti i gruppi. A questo scopo si deve innanzi tutto calcolare $c\times(c-1)/2$ differenze tra le medie campionarie di tutti i gruppi ($\bar{X}_j - \bar{X}_{j'}$ con $j \neq j'$), quindi calcolare il **range critico** (*ampiezza critica*) della procedura di Tukey-Cramer

Analisi della varianza (ANOVA) ad una via

- **Procedura di Tukey-Cramer**
  Se la differenza tra due medie campionarie è superiore al range critico, le corrispondenti medie dei gruppi (popolazioni) sono dichiarate significativamente diverse a livello di significatività α

Calcolo del range critico per la procedura di Tukey-Cramer

$$\text{Range critico} = Q_U \sqrt{\frac{MSW}{2}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)} \qquad \textbf{(10.16)}$$

dove Q_U è il valore critico **superiore della distribuzione del range studentizzato** con c gradi di libertà al numeratore e $n-c$ gradi di libertà al denominatore

Analisi della varianza (ANOVA) ad una via

- Esempio della procedura di Tukey-Cramer per il caso dei paracadute

Numeratore								
Denominatore	2	3	4	5	6	7	8	9
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13
15	3.01	3.67	4.08	4.37	4.60	4.78	4.94	5.08
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03

Fonte: estratto dalla Tavola E.8.

	A	B	C	D	E	F	G	H	I	J
1	**Procedura di Tukey-Kramer per confronti multipli**									
2										
3		**Media**	**Ampiezza**			Differenza	Std. Error	Range		
4	**Gruppo**	**Campionaria**	**Campionaria**		Confronto	(valore assoluto)	della differenza	critico	**Risultati**	
5	**1**	19,52	5		**Gruppo 1 e Gruppo 2**	4,74	1,1040	4,4712	**Le medie sono diverse**	
6	**2**	24,26	5		**Gruppo 1 e Gruppo 3**	3,32	1,1040	4,4712	**Le medie non sono diverse**	
7	**3**	22,84	5		**Gruppo 1 e Gruppo 4**	1,64	1,1040	4,4712	**Le medie non sono diverse**	
8	**4**	21,16	5		**Gruppo 2 e Gruppo 3**	1,42	1,1040	4,4712	**Le medie non sono diverse**	
9					**Gruppo 2 e Gruppo 4**	3,1	1,1040	4,4712	**Le medie non sono diverse**	
10	Altri dati				**Gruppo 3 e Gruppo 4**	1,68	1,1040	4,4712	**Le medie non sono diverse**	
11	Livello di significatività	0,05								
12	Numeratore: d.f.	4								
13	Denominatore: d.f.	16								
14	MSW	6,094								
15	Statistica Q	4,05								

Analisi della varianza (ANOVA) ad una via

- Procedura di Tukey-Cramer

Numeratore								
Denominatore	2	3	4	5	6	7	8	9
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13
15	3.01	3.67	4.08	4.37	4.60	4.78	4.94	5.08
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03

Fonte: estratto dalla Tavola E.8.

	A	B	C	D	E	F	G	H	I	J
1	**Procedura di Tukey-Kramer per confronti multipli**									
2										
3		**Media**	**Ampiezza**			Differenza	Std. Error	Range		
4	**Gruppo**	**Campionaria**	**Campionaria**		**Confronto**	(valore assoluto)	della differenza	critico	**Risultati**	
5	**1**	19,52	5		**Gruppo 1 e Gruppo 2**	4,74	1,1040	4,4712	**Le medie sono diverse**	
6	**2**	24,26	5		**Gruppo 1 e Gruppo 3**	3,32	1,1040	4,4712	**Le medie non sono diverse**	
7	**3**	22,84	5		**Gruppo 1 e Gruppo 4**	1,64	1,1040	4,4712	**Le medie non sono diverse**	
8	**4**	21,16	5		**Gruppo 2 e Gruppo 3**	1,42	1,1040	4,4712	**Le medie non sono diverse**	
9					**Gruppo 2 e Gruppo 4**	3,1	1,1040	4,4712	**Le medie non sono diverse**	
10	Altri dati				**Gruppo 3 e Gruppo 4**	1,68	1,1040	4,4712	**Le medie non sono diverse**	
11	Livello di significatività	0,05								
12	Numeratore: d.f.	4								
13	Denominatore: d.f.	16								
14	MSW	6,094								
15	Statistica Q	4,05								

Analisi della varianza (ANOVA) ad una via

- **Assunzioni alla base del test *F* per l'ANOVA a una via**
  Prima di applicare un test di ipotesi è sempre necessario valutare se le assunzioni di base del test possono o mene essere ragionevolmente soddisfatte. Le ipotesi alla base del test *F* per l'ANOVA a una via sono essenzialmente tre:
  - casualità e indipendenza
  - normalità
  - omogeneità delle varianze

  L'ultima ipotesi stabilisce che le varianze nei gruppi sono tra loro uguali ($\sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_c$). Nel caso di campioni con ampiezza simile le inferenze basate sulla distribuzione *F* non sono molto influenzate da eventuali differenze tra varianze, al contrario se le ampiezze sono diverse tra loro il problema potrebbe essere serio

Analisi della varianza (ANOVA) ad una via

- **Test di Levene per l'omogeneità delle varianze**
  Questa procedura inferenziale è stata sviluppata per verificare l'ipotesi nulla H_0: $\sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_c$ contro l'ipotesi alternativa H_1: non tutte le varianze sono uguali.
  Per verificare tale ipotesi si calcola la differenza in valore assoluto tra ogni osservazione e la mediana campionaria del gruppo di appartenenza e su questi dati si conduce l'ANOVA a una via.
  Per l'esempio dei paracadute si considera

Fornitore 1	Fornitore 2	Fornitore 3	Fornitore 4
Mediana = 18.5	Mediana = 24.5	Mediana = 22.9	Mediana = 20.4
|18.5 – 18.5|= 0.0	|26.3 – 24.5|= 1.8	|20.6 – 22.9|= 2.3	|25.4 – 20.4|= 5.0
|24.0 – 18.5|= 5.5	|25.3 – 24.5|= 0.8	|25.2 – 22.9|= 2.3	|19.9 – 20.4|= 0.5
|17.2 – 18.5|= 1.3	|24.0 – 24.5|= 0.5	|20.8 – 22.9|= 2.1	|22.6 – 20.4|= 2.2
|19.9 – 18.5|= 1.4	|21.2 – 24.5|= 3.3	|24.7 – 22.9|= 1.8	|17.5 – 20.4|= 2.9
|18.0 – 18.5|= 0.5	|24.5 – 24.5|= 0.0	|22.9 – 22.9|= 0.0	|20.4 – 20.4|= 0.0

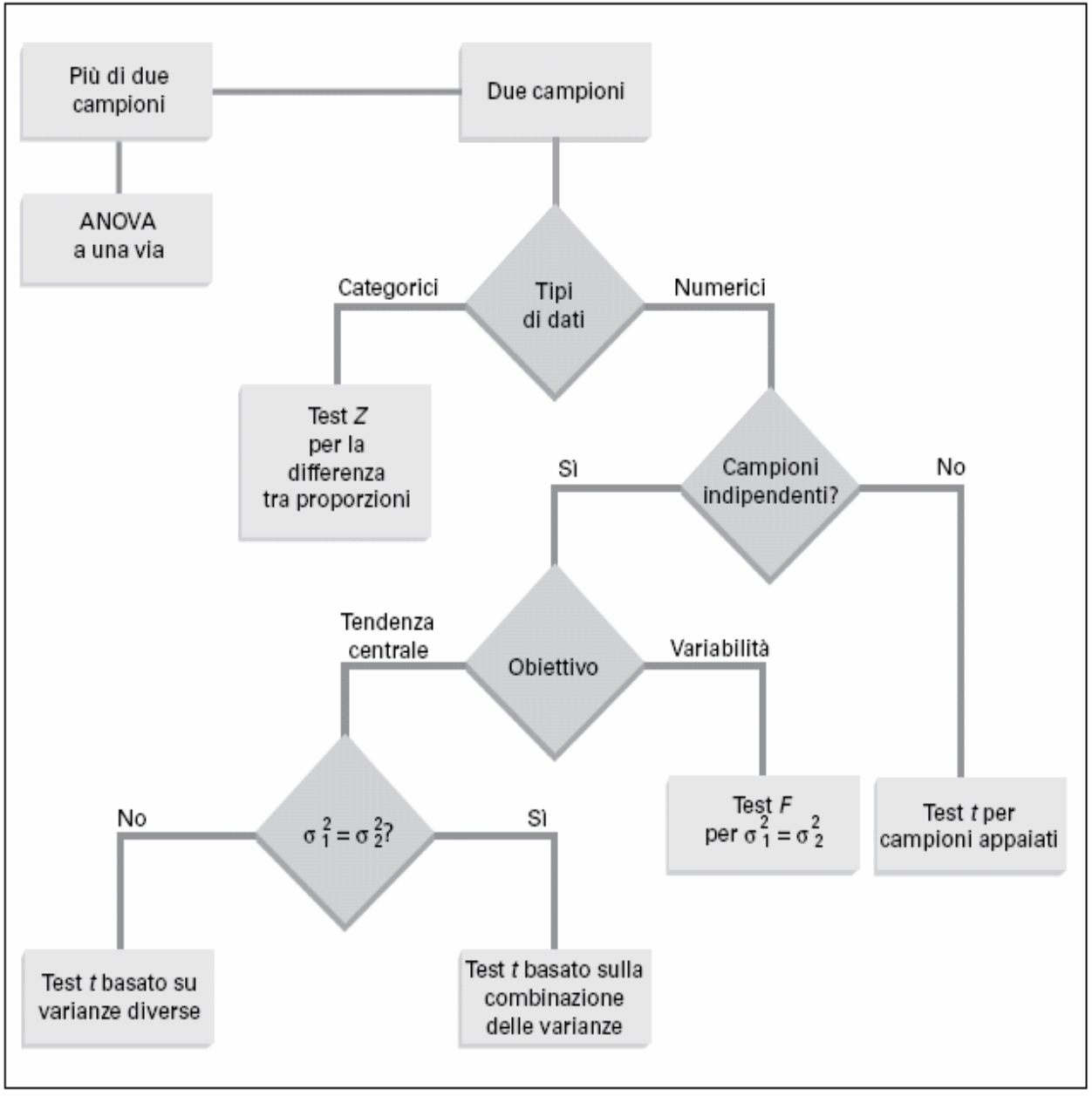
Analisi della varianza (ANOVA) ad una via

- Test di Levene per l'esempio dei paracadute

	A	B	C	D	E	F	G
1	**Analisi varianza: ad un fattore**						
2							
3	**RIEPILOGO**						
4	*Gruppi*	*Conteggio*	*Somma*	*Media*	*Varianza*		
5	**Fornitore 1**	**5**	**8,7**	**1,74**	**4,753**		
6	**Fornitore 2**	**5**	**6,4**	**1,28**	**1,707**		
7	**Fornitore 3**	**5**	**8,5**	**1,7**	**0,945**		
8	**Fornitore 4**	**5**	**10,6**	**2,12**	**4,007**		
9							
10							
11	**ANALISI VARIANZA**						
12	*Origine della variazione*	*SQ*	*gdl*	*MQ*	*F*	*Valore di significatività*	*F crit*
13	**Tra gruppi**	**1,77**	**3**	**0,59**	**0,2068**	**0,8902**	**3,2389**
14	**In gruppi**	**45,648**	**16**	**2,853**			
15							
16	**Totale**	**47,418**	**19**				

Fonte	Gradi di libertà	Somme dei quadrati	Medie dei quadrati (varianze)	*F*	*p*-value
Tra le catene	4	6536	1634.0	12.51	0.0000
All'interno delle catene	95	12407	130.6		

Riepilogo


Più di due campioni
Due campioni
ANOVA a una via
Categorici
Tipi di dati
Numerici
Test Z per la differenza tra proporzioni
Sì
Campioni indipendenti?
No
Tendenza centrale
Obiettivo
Variabilità
No
$\sigma_1^2 = \sigma_2^2$?
Sì
Test F per $\sigma_1^2 = \sigma_2^2$
Test t per campioni appaiati
Test t basato su varianze diverse
Test t basato sulla combinazione delle varianze