Introduzione alla chemiometria e disegno sperimentale

Modulo 7: Disegno sperimentale e ANOVA in R

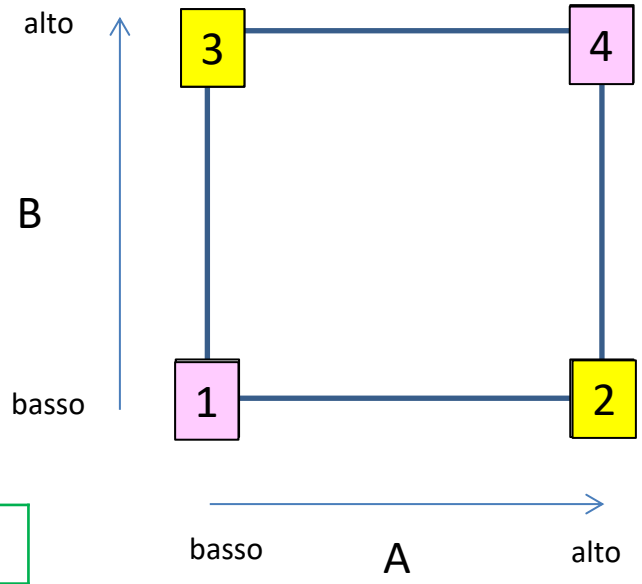
Docente: Dr. Sabina Licen (slicen@units.it)

Fractional factorial: esempio a 2 fattori

Fattori: A, B
Livelli: alto, basso codificati come +1 e -1 (es. alto e basso)
N° esperimenti: $2^{2-1} = 2$

Come si sceglie? Il sottoinsiemi di esperimenti deve prevedere che entrambi i livelli di ogni fattore siano rappresentati almeno una volta (il controllo viene effettuato osservando la tabella per colonna).

Exp	A	B
1 (full)	- 1	- 1
2 (full)	+ 1	- 1
3 (full)	- 1	+ 1
4 (full)	+ 1	+ 1



Exp	A	B
1	+ 1	- 1
2	- 1	+ 1

oppure

Exp	A	B
1	- 1	- 1
2	+ 1	+ 1

Fractional factorial: esempio a 3 fattori

Fattori: A, B, C
Livelli: alto, basso codificati come +1 e -1 (es. alto e basso)
N° esperimenti per ½ *factorial* $2^{3-1} = 4$

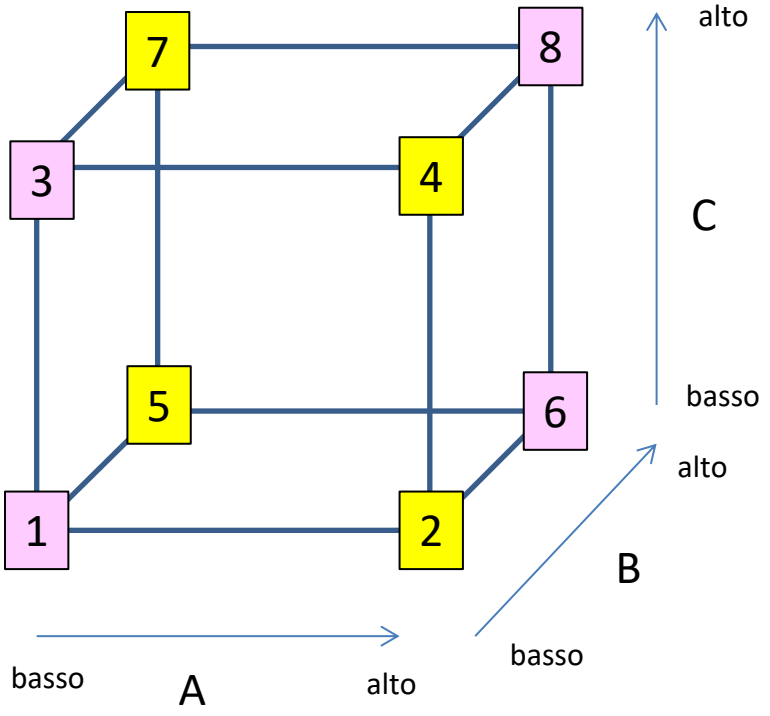
Come si sceglie? Il sottoinsiemi di esperimenti deve prevedere che entrambi i livelli di ogni fattore siano rappresentati due volte.

Exp	A	B	C
1(full)	- 1	- 1	- 1
2(full)	+ 1	- 1	- 1
3(full)	- 1	- 1	+ 1
4(full)	+ 1	- 1	+ 1
5(full)	- 1	+ 1	- 1
6(full)	+ 1	+ 1	- 1
7(full)	- 1	+ 1	+ 1
8(full)	+ 1	+ 1	+ 1

Exp	A	B	C
1	- 1	- 1	- 1
2	- 1	- 1	+ 1
3	+ 1	+ 1	- 1
4	+ 1	+ 1	+ 1

oppure

Exp	A	B	C
1	+ 1	- 1	- 1
2	+ 1	- 1	+ 1
3	- 1	+ 1	- 1
4	- 1	+ 1	+ 1



Qualità del disegno sperimentale

Ortogonalità (*orthogonality*) e disegno bilanciato (*balanced design*) sono due caratteristiche desiderabili che il disegno sperimentale dovrebbe avere.

Exp	A	B	C	*
1	+ 1	- 1	- 1	1
2	+ 1	- 1	+ 1	-1
3	- 1	+ 1	- 1	1
4	- 1	+ 1	+ 1	-1
	0	**0**	**0**	**0**

Balanced Design: la somma dei valori di colonna per ogni fattore è uguale a zero. Cioè i livelli del fattore sono rappresentati in eguale numero negli esperimenti

Orthogonality: due o più vettori sono ortogonali se la somma dei prodotti dei loro elementi è uguale a zero.

Fonte: http://users.unimi.it/gazzalab/wordpress/wp-content/uploads/2011/12/15-Design-of-Experiment-DOE.pdf

Disegno fattoriale in R

FrF2 package

Runs<- 2^2 → 2 fattori

oppure

Runs<- (2^3)/2 → 3 fattori; fractional 1/2

Design<-FrF2(nruns=Runs, nfactors=2, randomize = TRUE, seed=7)

summary(Design) → mostra i risultati

attributes(Design) → per vedere tutto il contenuto dell'oggetto creato

attributes(Design)$desnum → per estrarre solo la tabellina con il disegno sperimentale

Altri tipi di disegni sperimentali

La scelta del disegno sperimentale è fatta sulla base dei risultati desiderati. Altri esempi sono:

Completely randomized designs

Randomized block designs
- Latin squares
- Graeco-Latin squares
- Hyper-Graeco-Latin squares

Plackett-Burman designs

Response surface (second-order) designs
- Central composite designs
- Box-Behnken designs
- Response surface designs

Three-level full factorial designs

Three-level, mixed level and fractional factorial designs

In R:

agricolae package

Per vedere l'elenco dei tipi di design sperimentale ottenibili:

ls("package:agricolae", pattern = "design")

Fonte: http://users.unimi.it/gazzalab/wordpress/wp-content/uploads/2011/12/15-Design-of-Experiment-DOE.pdf

Assunzioni su ANOVA

L'utilizzo di ANOVA prevede che alcune assunzioni siano soddisfatte:

- ✓ La variabile dipendente (risposta) deve essere continua;
- ✓ La variabile dipendente (risposta) deve avere una distribuzione normale;
- ✓ Omogeneità delle varianze (omoschedasticità).

Fonte: http://www.sthda.com/english/wiki/compare-multiple-sample-variances-in-r; http://www.sthda.com/english/wiki/normality-test-in-r

Three-way ... multi-way ANOVA

Aggiungendo fattori la complessità aumenta, ma comunque si deve focalizzare l'attenzione sugli F-value.

I software statistici, come R, effettuano anche questi calcoli più complessi.

$$SST = SSG_{(F1)} + SSG_{(F2)} + ... + SSG_{(Fn)} + SSE$$

	SS	d.f.	MS	F-value
F_1				
F_2				
F_3				
$F_1 * F_2$				
$F_1 * F_3$				
$F_2 * F_3$				
$F_1 * F_2 * F_3$				

ANOVA in R

Modo 1:

```
Model<-lm(Risposta~F1+F2,data=Dati)
coefficients(Model)
anova(Model)
```

Modo 2:

```
AOV<-aov (Risposta~F1+F2,data=Dati)
summary(AOV)
```

Con interazioni tra fattori:

```
Model<-lm(Risposta~F1*F2,data=Dati)
```

```
AOV<-aov (Risposta~F1*F2,data=Dati)
```

Test preliminari per stabilire applicabilità di ANOVA

Normalità della Risposta:

```
shapiro.test(Dati$Risposta)
```

→ se p-value > 0.1 la distribuzione è normale

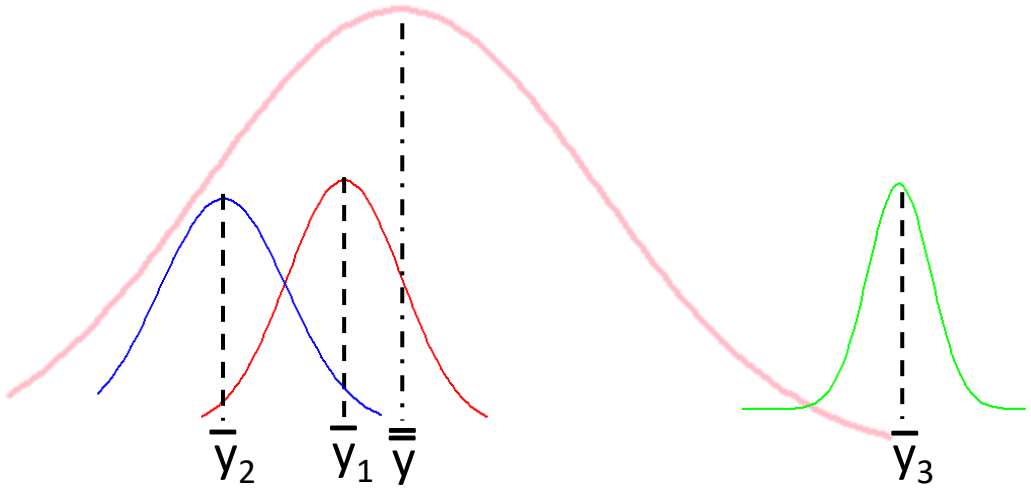
Omoschedasticità (per ogni fattore!):

```
bartlett.test(Risposta~F1,data=Dati)
```

→ se p-value > 0.1 è ok

Post-hoc tests

L'informazione che dà ANOVA, se $F_{(Fi)} > F_{critical}$, è che all'interno di quel fattore uno o più livelli appartengono a una popolazione diversa (ovviamente in caso di 3 o più livelli), ma quale? Ci sono alcuni test che consentono di capire quali sono i livelli "distanti".



Tukey's HSD (Honest Significant Difference) test

Student Newman Keuls (SNK)

Dunnett's test
(solo per trattamenti vs. controllo)

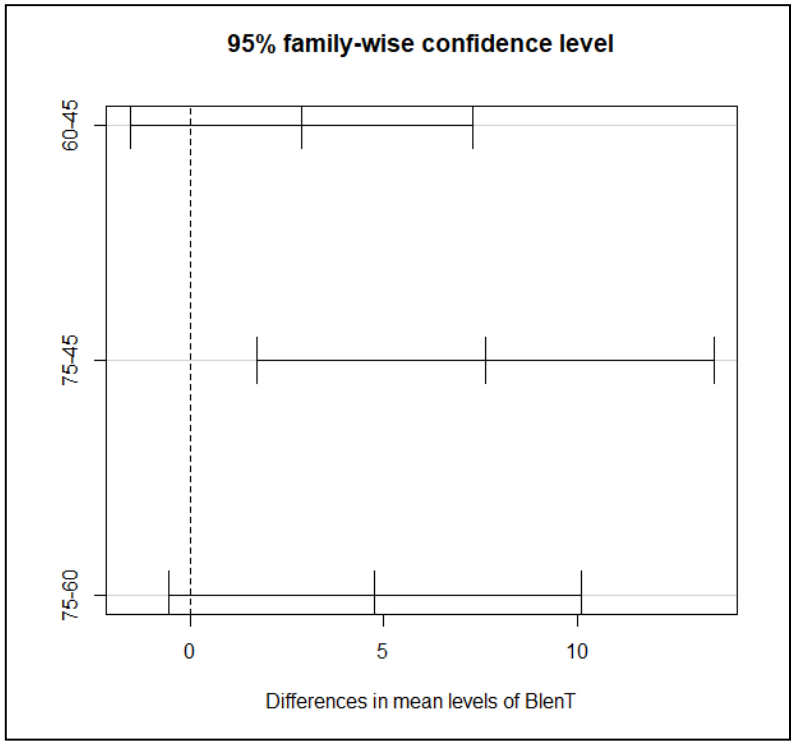
Tukey's HSD test in R

```
AOV<-aov (Risposta~F1+F2+F3,data=Dati)
```

```
TukeyHSD(AOV, "F1", conf.level = 0.95)
```

```
plot(TukeyHSD(AOV, "F1", conf.level = 0.95))
```

Nel grafico si guardano le coppie di livelli che si allontanano di più dallo zero



Analisi dati DoE - Statistica non parametrica

La statistica non parametrica si utilizza quando le variabili di interesse non seguono una distribuzione gaussiana e/o non c'è omoschedasticità.

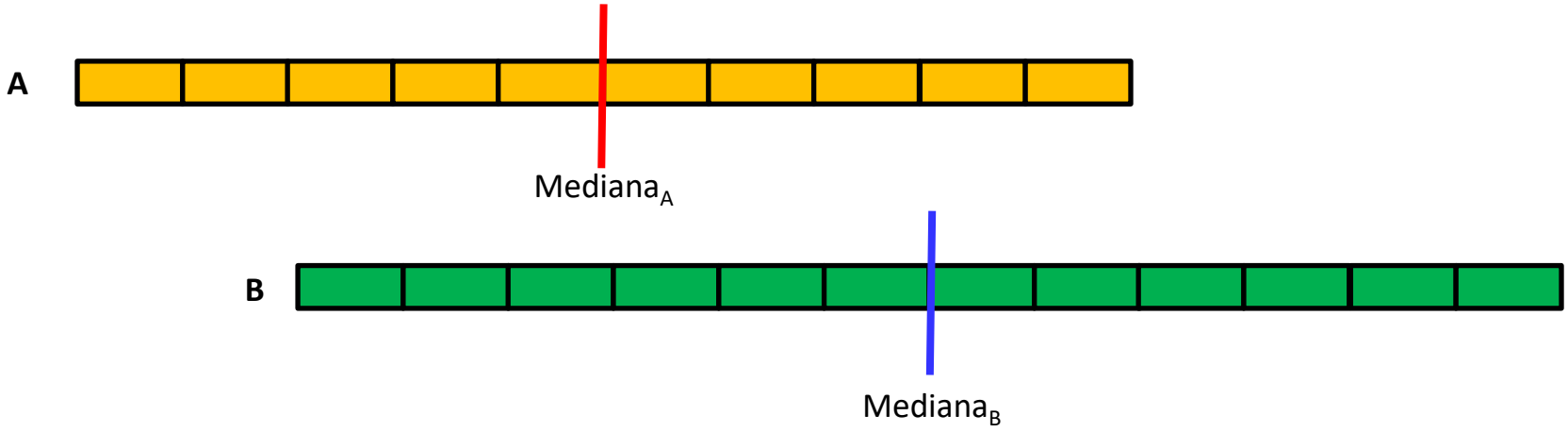
L'utilizzo di test non parametrici prevede di **ordinare** una (o più) serie di dati in ordine crescente e per confrontare la serie di dati utilizza l'indice di posizione dei dati (es. su una serie di 20 dati: posizione 4 di 20 (0.20), posizione 15 di 20 (0.75), etc...)

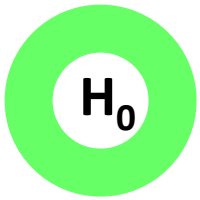
Per la comparazione di due serie di dati si utilizza il test di **Wilcoxon-Mann-Whitney** (analogo al t-test).

Per la comparazione di tre o più serie di dati si utilizza **Kruskal -Wallis** test (analogo di ANOVA)

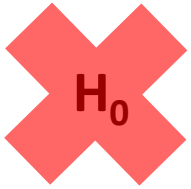
Wilcoxon-Mann-Whitney

Prevede di confrontare l'indice di posizione delle mediane di due serie di dati.





Le due serie appartengono ad una popolazione con stessa mediana

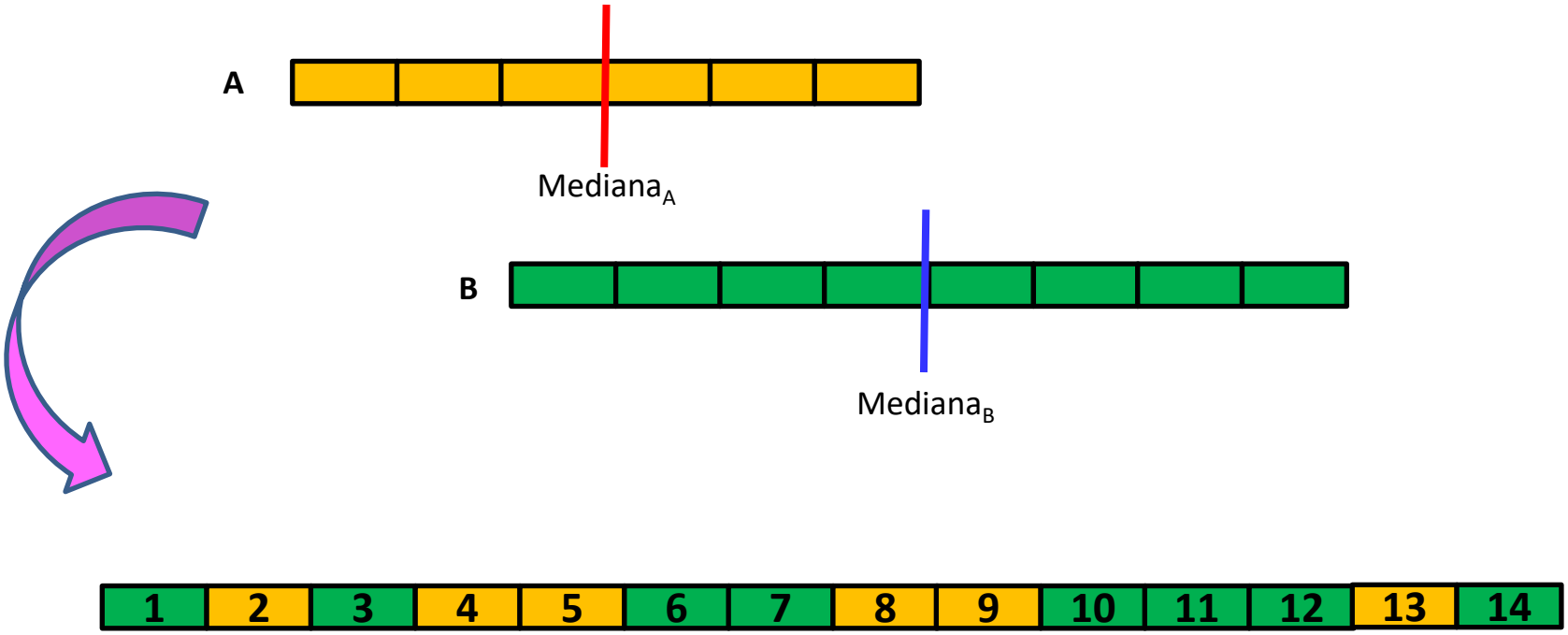


Le due serie NON appartengono ad una popolazione con stessa mediana

Fonte: https://www.bcfoltz.com/blog/stats-101/

Wilcoxon-Mann-Whitney (2)

Il metodo si basa sull'ordinamento delle due serie messe assieme e sulla conta dei rispettivi ranghi (posizioni).



$\sum A = 2+4+5+8+9+13 = 41$

$\sum B = 1+3+6+7+10+11+12+14 = 64$

$(\sum A + \sum B)\ /2 = 52.5$

H_0

$\sum A \approx \sum B \approx (\sum A + \sum B)\ /2$

H_0

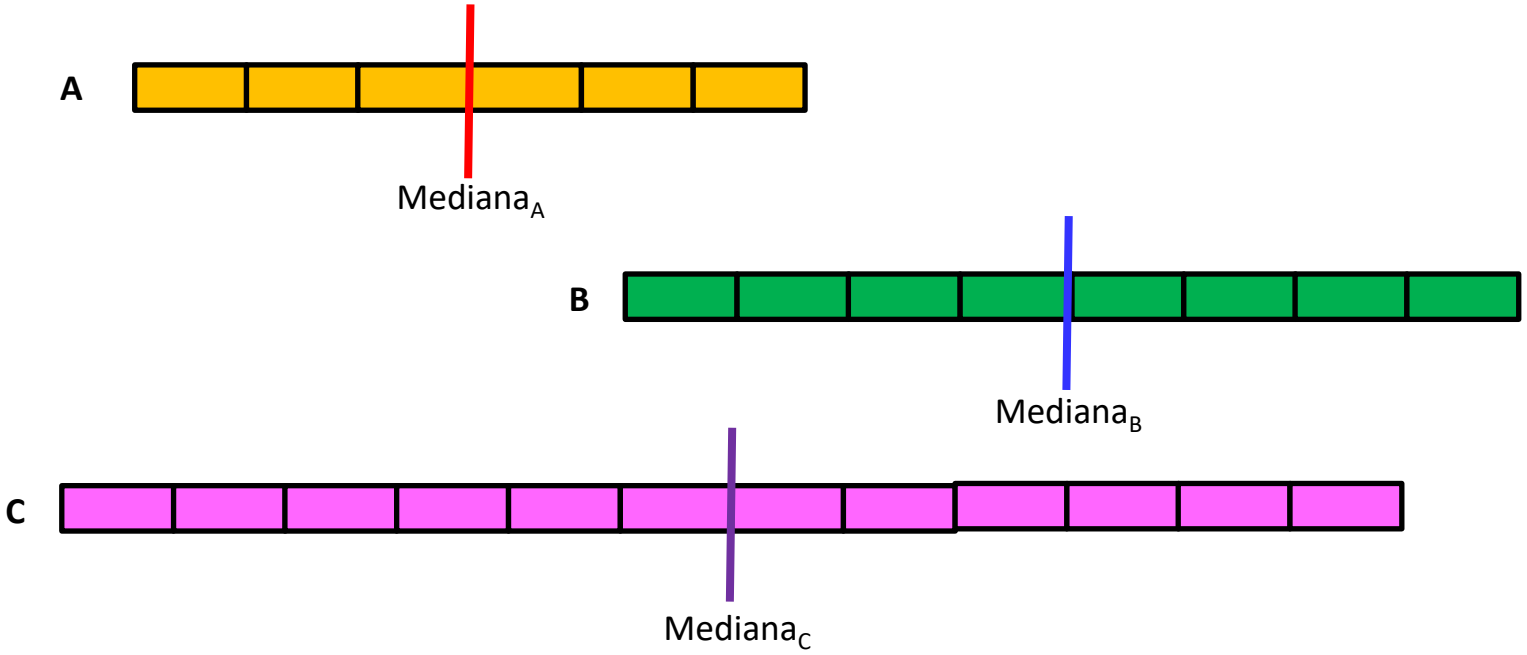
$\sum A \neq \sum B \neq (\sum A + \sum B)\ /2$

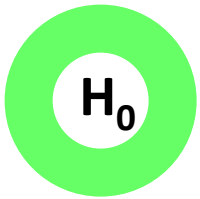
La soluzione finale comprende anche un test di significatività da cui si ottiene un p-value

Fonte: https://www.bcfoltz.com/blog/stats-101/

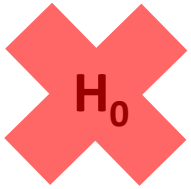
Kruskal - Wallis

Prevede di confrontare l'indice di posizione delle mediane di tre o più serie di dati.





Le serie appartengono ad una popolazione con stessa mediana

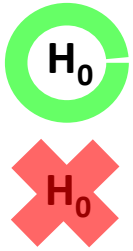


Le serie NON appartengono ad una popolazione con stessa mediana

Fonte: https://www.bcfoltz.com/blog/stats-101/

Kruskal - Wallis (2)

Anche in questo caso si utilizza la somma dei ranghi.



$$\sum A \approx \sum B \approx \sum C \approx (\sum A + \sum B + \sum C)\,/3$$

$$\sum A \neq \sum B \neq \sum C \neq (\sum A + \sum B + \sum C)\,/3$$

La soluzione finale comprende anche un test di significatività da cui si ottiene un p-value

Come per ANOVA, una volta rigettata l'ipotesi nulla, è necessario stabilire quale serie è la più "lontana" dalle altre.

Si utilizza il Wilcoxon-Mann-Whitney per ogni coppia di serie

Fonte: https://www.bcfoltz.com/blog/stats-101/