UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI SCIENZE AGRARIE E AMBIENTALI

PRODUZIONE, TERRITORIO, AGROENERGIA

Metodologia Sperimentale Agronomica / Metodi Statistici per la Ricerca Ambientale

Marco Acutis

marco.acutis@unimi.it

www.acutis.it

a.a. 2018 - 2019

CdS Scienze della Produzione e Protezione delle Piante (g59)

CdS Biotecnologie Vegetali, Alimentari e Agro-Ambientali (g61)

CdS Scienze Agro-Ambientali (g57)

Lezione 04 - Sommario

- Analisi della Varianza
  - Condizioni di applicabilità
  - Violazione dei requisiti di applicabilità

Analisi della Varianza – Condizioni di applicabilità

Il problema

Per la corretta interpretazione dei risultati dell' ANOVA sono richieste **tre** assunzioni sulle popolazioni di partenza:

1. **indipendenza delle osservazioni**
2. **normalità della distribuzione delle popolazioni di partenza**
3. **omogeneità delle varianze (o omoscedasticità o omoschedasticità)**

Formalmente: $X_j \sim N(\mu_j, \sigma^2),\ j = 1, \ldots, k$ e X_j indipendenti

Analisi della Varianza – Condizioni di applicabilità

N.B.

La verifica delle assunzioni richieste dovrebbe essere **sempre** compiuta preliminarmente all'esecuzione dell'ANOVA vera e propria. Frequentemente conclusioni errate sono state tratte per la mancanza del requisito di omogeneità, a cui l'ANOVA è particolarmente sensibile; risulta invece maggiormente robusta rispetto all'assunto di normalità.

Analisi della Varianza – Condizioni di applicabilità

indipendenza delle osservazioni

*Le osservazioni sperimentali si dicono **indipendenti** quando l'esito di ciascuna misura non è influenzato dalla precedente. Di conseguenza tale condizione viene garantita dalla natura stessa dello schema sperimentale.*

In altre parole, la randomizzazione deve essere fondata su elementi obiettivi e non lasciata all'arbitrio o all'intuito dello sperimentatore: ogni dato deve avere la stessa possibilità di essere influenzato dai fattori noti (effetto trattamento) e da quelli ignoti (effetto ambiente statistico).

Analisi della Varianza – Condizioni di applicabilità

normalità della distribuzione delle popolazioni di partenza (1/3)

Il requisito di normalità impone che la distribuzione dei dati all'interno di ciascun fattore (o trattamento) sia normale $\left(X_j \sim N\left(\mu_j, \sigma^2\right)\right)$ *e questo corrisponde a richiedere che la distribuzione dei residui sia normale* $\left(\varepsilon \sim N\left(0, \sigma^2\right)\right)$*, dal momento che questi ultimi, per definizione, sono calcolati tenendo conto delle medie di ciascun fattore.*

Test di **normalità** di **Shapiro-Wilk**

Analisi della Varianza – Condizioni di applicabilità

normalità della distribuzione delle popolazioni di partenza (2/3)

Nel caso dell'ANOVA ci sono quindi due opzioni per verificare la normalità:

- ✓ se i gruppi sono pochi (<4) e le repliche sono tante (>20), allora è possibile valutare il requisito sui dati «tal quali» di ciascun fattore e condurre così il test tante volte quanti sono i fattori;
- ✓ se i gruppi sono numerosi ed il numero di repliche limitato (questo è il caso generalmente più frequente nel nostro campo), spesso è più conveniente applicare il test di normalità una volta sola su tutti i residui assieme.

Analisi della Varianza – Condizioni di applicabilità

normalità della distribuzione delle popolazioni di partenza (3/3)

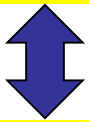
L'ipotesi nulla del test di Shapiro-Wilk è che le popolazioni di partenza siano distribuite secondo una normale.

La verifica della normalità quindi consiste nel controllare che il *p*-value (o livello di significatività osservato) del (o dei) test sia **superiore** al valore prefissato $\alpha = 0.05$.

omogeneità delle varianze (1/2)

Al di là dell'aspetto formale, ricordiamo che la varianza è una stima della credibilità di una media. In altri termini, dati molto variabili (cioè caratterizzati da una varianza ampia) hanno, a parità di numero di osservazioni, medie più variabili (come i dati di partenza) e perciò meno credibili. L'analisi della varianza confronta le medie, è quindi necessario che la loro credibilità sia simile, soprattutto quando i campioni hanno dimensioni molto differenti.

Test di **omogeneità delle varianze** di **Levene**



ANOVA sul **valore assoluto dei residui**

omogeneità delle varianze (2/2)

L'ipotesi nulla del test di Levene o dell'ANOVA applicata al valore assoluto dei residui è che le varianze siano omogenee.

La verifica dell'omogeneità quindi consiste nel controllare che il *p*-value (o livello di significatività osservato) del (o dei) test sia **superiore** al valore prefissato $\alpha = 0.05$.

Analisi della Varianza – Violazione dei requisiti

Non normalità

Premessa: spesso l'allontanamento dalla normalità non ha effetti gravi, dal momento che solo una notevole asimmetria può invalidare il test F, il quale è ritenuto robusto rispetto a questo requisito.

→ In caso di moderata non normalità (significatività del test di Shapiro-Wilk >0,01), è lecito soprassedere, soprattutto con un numero limitato di dati.

→ In caso di elevata non normalità (significatività del test di Shapiro-Wilk < 0,01), è possibile ricorrere alla trasformazione dei dati o all'uso di test non parametrici.

Trasformazione dei dati (1/2)

Alcuni suggerimenti:

«problema»	trasformazione
Asimmetria positiva moderata	$Y = \sqrt{X}$
Asimmetria positiva forte	$Y = \log_{10}(X + c)$, dove c è una costante da inserire per evitare che l'argomento del logaritmo sia minore di 1
Asimmetria negativa moderata	$Y = \sqrt{k - X}$, dove k è una costante da inserire per evitare di avere sotto radice un valore negativo
Asimmetria negativa forte	$Y = \log_{10}(c - X)$, dove c è una costante da inserire per evitare che l'argomento del logaritmo sia minore di 1

Howell, D. C. (2007). Statistical methods for psychology (6th ed.). Belmont, CA: Thomson Wadsworth.

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Boston: Allyn and Bacon.

Trasformazione dei dati (2/2)

Altri suggerimenti:

1. Se la deviazione standard è proporzionale alla media, allora la distribuzione è positivamente asimmetrica la trasformazione ideale è quella logaritmica.
2. Se la varianza è proporzionale alla media, allora è da preferire la trasformazione in radice quadrata. Questo accade sovente i dati sono frutto di conteggio.
3. Se la deviazione standard è proporzionale al quadrato della media, si può valutare l'opzione di una trasformazione reciproca ($Y = 1/X$).

N.B. I risultati di un'ANOVA condotta su dati trasformati sono relativi alle medie dei dati trasformati!

Test non parametrici (1/2)

Nel caso in cui la trasformazione dei dati non basti per avere un dataset «normale», occorre fare ricorso ai test non parametrici, i quali non richiedono ipotesi a priori circa la distribuzione delle popolazioni.

I test non parametrici in caso di non normalità funzionano piuttosto bene, tuttavia è bene ricordare che:

1. sono generalmente **meno potenti** dei test parametrici, anche se di poco;
2. talvolta forniscono risultati la cui **interpretazione** può risultare **difficoltosa** (molti ricorrono alla trasformazione dei dati originali in ranghi, quindi sapere che la differenza tra i ranghi medi di due gruppi è, ad esempio, pari a 5, non aiuta molto la comprensione dei dati).

Test non parametrici (2/2)

Modello	Test non parametrico
ANOVA a 1 via	Test di Kruskal-Wallis
Anova a 2 o più vie senza interazioni	Test di Friedman
Anova a 2 o più vie con interazioni	Test di Scheirer-Ray-Hare (estensione del Test di Kruskal–Wallis)

Analisi della Varianza – Violazione dei requisiti

Non omogeneità delle varianze

Premessa: la violazione del requisito di omogeneità ha conseguenze decisamente più rilevanti sull'attendibilità del risultato dell'ANOVA, soprattutto nel caso in cui l'esperimento sia sbilanciato.

Possibili soluzioni:

- ✓ utilizzare i test non parametrici elencati in precedenza come rimedio al problema della non normalità, soprattutto quando la disomogeneità delle varianze è da attribuirsi alla presenza di dati aberranti;
- ✓ nel caso di ANOVA a 1 via, fare ricorso al test di Welch o al test di Brown-Forsythe;
- ✓ seguire l'approccio di Conover & Iman («Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics" – 1981 - The American Statistician Vol. 35, No.3, pp 124-129), che consiste sostanzialmente nel trasformare i dati in ranghi e poi applicare l'ANOVA direttamente sui ranghi senza ulteriori passaggi;
- ✓ sfruttare le potenzialità delle tecniche di bootstrap.