**Esercitazioni Metodologia Sperimentale Agronomica 2012/2013 – Domande degli studenti**

1. Cos’è un’ANOVA *multivariata*? Che differenze ci sono rispetto ad un’ANOVA *univariata*?

La procedura dell’ANOVA multivariata consiste nel condurre l’analisi della varianza contemporaneamente su più di una variabile dipendente. In altre parole abbiamo a che fare non con un modello singolo, ma con un sistema di modelli. I risultati di ogni singola ANOVA sono indipendenti tra loro, quindi è assolutamente equivalente al ripetere più volte la procedura dell’ANOVA univariata.

1. Concettualmente a cosa corrispondono i gradi di libertà del fattore e dell’errore? Ci è stato sicuramente già spiegato, ma sentircelo dire ancora una volta ci aiuta.

I gradi di libertà di una variabile aleatoria o di una statistica in genere esprimono il numero minimo di dati sufficienti a valutare la quantità d'informazione contenuta nella statistica. La media relativa ad una popolazione non è nota, mentre quella relativa ad un campione lo è sempre; in questo caso quindi il numero minimo di dati indipendenti per valutare la quantità d’informazione dev’essere diminuito di 1, poiché ciascun valore può esser ricavato dagli altri assieme alla media. Ne consegue che per valutare la quantità d’informazione relativa alla variabilità TRA I TRATTAMENTI occorrono (#trattamenti-1) dati, mentre per valutare la quantità d’informazione relativa alla variabilità ENTRO I TRATTAMENTI, che equivale alla variabilità dell’errore, occorrono #trattamenti \*(#repliche-1). Questo si verifica perché in quest’ultimo caso è nota la media delle repliche all’interno di ciascun trattamento.

1. Relativamente alla *Tabella ANOVA* siamo in grado di calcolare tutte le sue voci anche manualmente (devianza, gdl, varianza, F) ad eccezione della significatività. Come si calcola manualmente il valore della significatività? Quale operazione matematica sta dietro la funzione di Excel “=distrib.F(X,gdl1,gdl2)”?

Non si può calcolare manualmente, si usano le tabelle oppure la funzione di Excel distrib.F(X,gdl1,gdl2), dove X è f, cioè il rapporto varianza Trattamento/varianza Errore.

Dietro distrib.F(X,gdl1,gdl2), ci sta il calcolo della funzione di densità di probabilità, che rappresenta la probabilità dell’evento: “la v.a. X è compresa entro un determinato intervallo”, della distribuzione di Fisher., la cui forma è complicata assai.

1. Relativamente al Test di Levene sappiamo che esso consiste nel calcolo dell’ANOVA sopra i valori assoluti dei residui. Senza entrare troppo nel dettaglio dei passaggi matematici, quale concetto sta dietro il test di Shapiro-Wilk per la verifica della normalità di una variabile?

Il test si basa sul calcolo di un rapporto tra una statistica d’ordine (cioè qualcosa che ha a che fare coi ranghi) e la devianza campionaria.

1. Una questione puramente terminologica: *pairwise comparison* corrisponde all’italiano *confronto a coppie,* realizzabile con alcuni test post-hoc (come Bonferroni o Tuckey)? Esiste un corrispettivo termine inglese anche per indicare i *confronti multipli* realizzabili sempre con alcuni test post-hoc (come Tuckey o REGW-F)?

Sì

Multiple comparisons

1. Ogniqualvolta l’interazione di due fattori risulta significativa, i test post-hoc sono eseguiti secondo la procedura *Analizza - Modello lineare generalizzato - Univariata - Opzioni - Visualizza medie per: … - Confronta effetti principali (Sidak)* e la successiva modifica della sintassi. È corretto, nel caso in cui vi sia significatività per l’interazione, ma io sia interessato anche ai post-hoc per un singolo fattore fisso, seguire questi passaggi: *Analizza - Modello lineare generalizzato - Univariata - Post hoc - Test post hoc per: … (ad esempio: Tuckey e REGW-F)?* Cosi facendo visualizzo dell’informazione ridondante rispetto a quanto fatto sopra o ottengo informazioni aggiuntive di possibile interesse?

Nessuna delle due perché nel caso in cui sia evidenziata un’interazione significativa non è corretto analizzare gli effetti principali, i quali a loro volta soffrono delle reciproche interazioni. In altri termini, l’effetto del fattore 1 varia al variare del livello del fattore 2 e viceversa; di conseguenza analizzare l’effetto semplice di uno dei due fattori fornisce informazioni parziali e potenzialmente errate.

1. Che informazioni mi dà la tabella “Test Univariati” visualizzata a seguito del calcolo dei post hoc quando l’interazione è risultata significativa?

Per ciascun livello del fattore 1, mostra l’eventuale presenza di differenze significative entro tutti i livelli del fattore 2. Se nella sintassi è presente il doppio comando

/EMMEANS=TABLES(F1\* F2) COMPARE (F2) ADJ(SIDAK)

/EMMEANS=TABLES(F2\* F1) COMPARE (F1) ADJ(SIDAK)

allora ottengo nell’output un’altra tabella “Test Univariati” che mostra la significatività dei singoli livelli del fattore 2 entro tutti i livelli del fattore 1.

Esempio: l’effetto del fattore Inerbimento entro i singoli livelli di Diradamento\_grappoli.

Si osservi che l’ANOVA aveva messo in luce una significatività dell’interazione e questa tabella ne dà un riscontro, mostrando un valore di significatività non sempre < 0.05 per i livelli di diradamento.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test univariati** | | | | | | |
| Variabile dipendente: GradiBrix | | | | | | |
| Diradamento\_grappoli | | Somma dei quadrati | df | Media dei quadrati | F | Sig. |
| Assente | Contrasto | 209,380 | 2 | 104,690 | 12,190 | ,000 |
| Errore | 412,237 | 48 | 8,588 |  |  |
| Precoce | Contrasto | 106,425 | 2 | 53,212 | 6,196 | ,004 |
| Errore | 412,237 | 48 | 8,588 |  |  |
| Standard | Contrasto | 371,323 | 2 | 185,661 | 21,618 | ,000 |
| Errore | 412,237 | 48 | 8,588 |  |  |
| Tardivo | Contrasto | 49,909 | 2 | 24,955 | 2,906 | ,064 |
| Errore | 412,237 | 48 | 8,588 |  |  |
| F verifica l'effetto di Inerbimento. Questo test è basato sui confronti a coppie indipendenti e lineari tra le medie marginali stimate. Questi test sono basati sui confronti a coppie linearmente indipendenti tra le medie marginali stimate. | | | | | | |

1. Tra le assunzioni della correlazione c’è anche il fatto che le due variabili varino secondo una relazione lineare. Il professore ci ha detto che tale aspetto è verificabile anche solo visivamente attraverso un grafico a dispersione. Quindi prima di ogni correlazione è necessario realizzare e commentare questo grafico?

No. non è strettamente necessario. Tuttavia può essere utile per farsi un’idea della relazione tra le due variabili nel caso in cui non vi sia linearità.

1. Differenze tra un test a una coda e un test a due code.

Al momento della definizione del comando di correlazione è necessario specificare se si conosce già in partenza il segno della relazione tra le variabili in esame (se sono quindi direttamente o inversamente proporzionali) segnalando in questo caso il test ad una coda.

1. Definizione di *potente* e *robusto* e loro differenze*.*

La potenza è un concetto riferito ai test statistici, si definisce come il complemento a 1 della probabilità di commettere errore di II specie (accettare l’ipotesi nulla, quando invece questa è falsa) ed in pratica rappresenta la capacità di evidenziare differenze.

Ricordiamo che un test è protettivo quando è bassa la probabilità di commettere l’errore di I specie (rifiutare l’ipotesi nulla, quando invece questa è vera). Ad esempio, il test di Duncan è più potente, ma meno protettivo rispetto al test di Bonferroni.

Come regola generale, i test statistici da preferire sono quelli che rappresentano un buon compromesso tra potenza e livello di protezione, poiché, a parità di numerosità campionaria, al diminuire della probabilità dell’errore di prima specie (α), aumenta la probabilità dell’errore di seconda specie (β) e cala la potenza del test (1-β).

La robustezza invece è una caratteristica non solo dei test, ma anche delle statistiche e rappresenta la sensibilità rispetto a valori/condizioni anomale. Quindi ad esempio la mediana è più robusta della media rispetto agli outlier e l’anova è più robusta in condizioni di non normalità rispetto alla condizione di non omogeneità delle varianze.

1. Qual è il comando Excel che consente di calcolare automaticamente la tabella Anova?

Dati->Analisi Dati->Analisi varianza:…

Per attivare il componente Analisi Dati:

Andare su File->Opzioni-> Componenti aggiuntivi e selezionare il componente Strumenti di analisi

1. Quali differenze ci sono fra i tipi di variabile *numerico* e *stringa* visibili in “visualizzazioni variabile”?

Una variabile numerica rappresenta un numero, quindi può avere l’attributo di scala. Questo vuol dire che il software riconosce l’ordine dei valori ed ammette che possa assumere valori intermedi, il che equivale a dire che su questa si possono fare i conti.

Una variabile stringa invece è solo un nome, quindi ha di default l’attributo nominale ed al massimo può avere quello ordinale che implica una gerarchia tra i valori (es. alto – medio - basso)

1. Una percentuale è una variabile continua?

Il fatto che una variabile sia espressa in percentuale non ha nulla a che vedere col fatto che sia continua. La continuità dipende dalla possibilità che la variabile possa assumere come valori tutti i numeri reali compresi all’interno del suo campo di variazione. Quindi se ad esempio ho a che fare con una variabile che conta il numero di individui e non ha senso parlare di 3,47 individui, allora la variabile non è continua.

1. Rivedere le ipotesi che stanno dietro la regressione: deve essere verificata l’omogeneità delle varianze? Se sì, su quale variabile?

Sì, sui valori assoluti dei residui, i quali rappresentano la differenza tra il valore osservato ed il valore stimato dal modello regressivo Y=aX+b.

1. Nell’editor di un grafico a dispersione, selezionando *Elementi - Curva di Adattamento a Totale* appare la finestra *Intervalli di confidenza* con le opzioni: *nessuna - media - singolo ( … percentuale).* È possibile riascoltare la distinzione tra queste tre opzioni?

Caso media: vengono calcolati gl’intervalli di confidenza di intercetta (a) e coefficiente angolare (b) ed in pratica vengono disegnati i limiti entro i quali con il 95% di probabilità si trova la retta “reale”. Si utilizza quando siamo interessati all’affidabilità del modello poiché se gl’intervalli sono molto ampi allora non possiamo fidarci troppo della retta stimata.

Caso singolo: vengono calcolati, oltre agl’intervalli di confidenza di intercetta (a) e coefficiente angolare, anche i residui ed in pratica vengono disegnati i limiti entro i quali si trova il 95% dei casi. Si utilizza quando siamo interessati maggiormente ai punti, ad esempio quando a valutare l’affidabilità di uno strumento di misura.

1. File dati *Poligoni* dell’esercitazione sui cluster. Nel clusterizzare i dati mediante la *two step analysis* abbiamo prima lasciato determinare ad SPSS il numero “idoneo” di gruppi (in questo caso 3); successivamente è stata fatta la stessa analisi imponendo però un numero di gruppi pari a quattro. Nel secondo caso abbiamo ottenuto una qualità del cluster maggiore rispetto al primo caso. Se SPSS nello scegliere il numero ideale di cluster non valuta la qualità del cluster allora che cosa valuta?

1° caso



2° caso



Dal manuale di SPSS: “La visualizzazione Riepilogo del modello mostra un'istantanea (o riepilogo) del modello di cluster, compresa una misura della silhouette di coesione e separazione dei cluster, che è ombreggiata per indicare risultati scarsi, discreti o buoni. Questa istantanea consente di verificare rapidamente se la qualità è scarsa, nel qual caso è possibile decidere di tornare al nodo per la creazione dei modelli per correggere le impostazioni del modello di cluster e ottenere un risultato migliore.

La qualità del risultato (scarso, discreto, buono) è basata sul lavoro di Kaufman e Rousseeuw (1990) relativo all'interpretazione delle strutture dei cluster. Nella visualizzazione Riepilogo del modello, un risultato buono equivale a quei dati che rispecchiano la classificazione di Kaufman e Rousseeuw di ragionevole o forte indizio di una struttura di cluster, un risultato discreto rispecchia la classificazione di indizio debole, un risultato scarso corrisponde alla classificazione di assenza di indizio significativo.

La misura della silhouette viene calcolata su tutti i record, (B−A) / max(A,B), dove A è la distanza del record dal centro del relativo cluster e B è la distanza del record dal centro del più vicino cluster a cui non appartiene. Un coefficiente di silhouette pari a 1 indica che tutti i casi si trovano direttamente in corrispondenza dei centri dei relativi cluster. Un valore pari a −1 indica che tutti i casi si trovano in corrispondenza dei centri di altri cluster. Il valore 0 indica, in media, che i casi sono equidistanti tra il centro del rispettivo cluster e il cluster più vicino”.

Nell’esercizio preso in esame, si voleva semplicemente mostrare l’opzione di scelta di valore impostato a priori dall’utente e valutarne la qualità, confrontando il risultato con il valore di silhouette relativo al numero di cluster indentificato come migliore dal software.