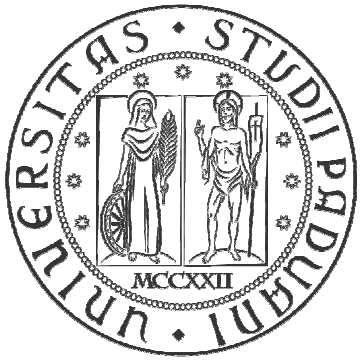
UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE



TESI DI LAUREA

Piani fattoriali multivariati: l'approccio PERMANOVA e applicazione a dati ecologici

Multivariate factorial design: the PERMANOVA approach and its application to an ecological data set

Relatore: Dott. Livio Finos

Correlatore: Dott.ssa Monica Bressan

Laureando: Francesco Gatti

Matricola: 581296

ANNO ACCADEMICO: 2010/2011

INDICE

INTRODUZIONE

In ecologia l'utilizzo dell'analisi di dati multivariati ha una rilevanza sempre maggiore per verificare gli effetti di più specie rispetto a tutto l'ambiente circostante.

Questo succede nelle ricerche e nello studio della biodiversità o degli impatti ambientali in molti habitat, sia marini che terrestri. Uno degli strumenti più importanti per testare la variabilità dei fattori e delle loro interazioni è l'ANOVA che presenta un corrispondente per il caso multivariato, la MANOVA (Mardia et alt, 1991).

Questi strumenti non si sono sempre dimostrati adeguati allo studio dei dati ecologici nei quali è frequente avere numerosità molto basse che comportano zeri nei data set. Anche per questo motivo quindi, le distribuzioni delle abbondanze sono spesso asimmetriche o aggregate e non permettono di ipotizzare una distribuzione normale per i dati. Infine non è raro lavorare con dataset che presentano più variabili che osservazioni.

Di conseguenza sono state sviluppate una serie di tecniche per partizionare la varianza a partire dalla matrice delle distanze calcolata su una qualunque misura metrica o anche semimetrica (la distanza euclidea in questi casi, infatti, non rappresenta in modo adeguato i dati) (Anderson, 2006).

La disponibilità di potenza di calcolo a basso costo ha permesso lo sviluppo di tecniche non parametriche basate sulle permutazioni che hanno portato allo sviluppo della PERMANOVA (Anderson, 2001b).

Nei capitoli successivi viene prima descritto l'approccio che ha portato al metodo di analisi della varianza basato sulle permutazioni, partendo dal modello multivariato e dalla MANOVA.

La descrizione è limitata agli strumenti utilizzati nel software statistico PRIMER e in particolare al pacchetto aggiuntivo PERMANOVA+, che contiene la routine utilizzata per testare simultaneamente la relazione tra una o più variabili rispetto a uno o più fattori in un'analisi della varianza basata su una qualunque misura di similarità, usando il metodo delle permutazioni (la PERMANOVA, appunto).

Nell'ultimo capitolo, infine, viene presentata un'applicazione ad un data set i cui dati sono stati raccolti dal dipartimento di biologia marina dell'Università di Padova per indagare sulla possibilità di allevamento di molluschi nel nord Adriatico.

CAPITOLO 1

Modello multivariato– MANOVA

Scopo del modello di regressione multivariato è quello di modellare le variabili risposta a partire dallo stesso insieme di variabili esplicative. Il modello considerato sarà definito come:

$$Y = XB + U \tag{1}$$

Dove Y è una matrice $n \times p$ contenente un numero p di variabili risposta e per ciascuna di esse un numero n di osservazioni. X, avente dimensioni $n \times q$ è la matrice contenente le variabili esplicative. La matrice B, di dimensione $q \times p$, è la matrice contenente i coefficienti di regressione e U, infine, è la matrice $n \times p$ contenente i termini d'errore del modello.

Si suppone che le righe della matrice U siano indipendenti e incorrelate rispetto alle corrispondenti righe della matrice X, cioè ci si aspetta che i termini d'errore siano indipendenti dalle corrispondenti osservazioni, e che ciascuna riga della matrice U abbia media nulla e una matrice di covarianza Σ.

Quando X è una matrice di q variabili esplicative osservate su ciascuna delle n osservazioni il modello considerato è un modello di regressione multivariata.

Le colonne della matrice Y, che rappresentano le variabili dipendenti del modello, sono le variabili che devono essere spiegate dalle colonne dalla matrice X. Da questo segue la relazione:

$$E(y_{ij}) = x_i^T \beta_{(j)}$$

quindi il valore atteso di y_{ij} dipende dall'i-esima riga della matrice X e dalla j-esima colonna della matrice B dei coefficienti di regressione.

Nella maggior parte dei casi è possibile supporre che la matrice U si distribuisca normalmente, sia cioè una matrice con distribuzione $N_p(0,\ \Sigma)$ e che sia indipendente da X.

Sotto l'ipotesi di normalità degli errori la log-verosimiglianza per la matrice Y rispetto ai parametri B e Σ è:

$$l(B,\Sigma) = -\frac{1}{2}n\log|2\pi\Sigma| - \frac{1}{2}\mathrm{tr}(\mathrm{Y}-\mathrm{XB})\Sigma^{-1}(\mathrm{Y}-\mathrm{XB})'$$

Come vedremo in seguito, quando analizzeremo dati ambientali o ecologici, le assunzioni di indipendenza delle colonne della matrice X e di normalità degli errori contenuti nella matrice degli errori U, osserveremo che queste non vengono mai o quasi mai soddisfatte.

1.1 MANOVA: analisi della varianza multivariata

Nel caso di un modello di regressione multipla definito come sopra (1) in cui Y è una matrice $n \times p$ con $p = 1$, è possibile valutare l'uguaglianza delle medie in gruppi differenti e valutare la significatività dell'effetto di una variabile esplicativa sulla variabile risposta, utilizzando l'analisi della varianza (ANOVA). E' possibile, anche nel caso si stia trattando un modello di regressione multivariato, utilizzare una procedura simile, la MANOVA (multivariate analysis of variance). Supponendo, quindi, di voler analizzare un'eventuale relazione presente tra due o più variabili dipendenti rispetto a una variabile esplicativa è possibile calcolare una statistica F, non più univariata bensì multivariata (lambda di Wilks) basata sul confronto tra la matrice di varianza e covarianza degli errori e la matrice di varianza e covarianza delle variabili esplicative. Infatti, nel caso si tratti un modello multivariato per il calcolo della statistica F, è necessario stimare anche i prodotti incrociati (crossproducts).

La procedura di analisi multipla della varianza si basa su una serie di assunzioni che, se violate, rendono non affidabili gli indicatori calcolati.

Tali assunzioni sono:

- L'indipendenza delle osservazioni, poiché nell'ipotesi in cui questa condizione non valga, allora la MANOVA non è robusta.
- Dimensioni uguali per tutti i gruppi.
- Somme dei quadrati appropriate.
- Dimensioni adeguate del campione.
- Distribuzione casuale dei residui.
- Omoschedasticità (omogeneità delle varianze e delle covarianze): la varianza di ogni variabile dipendente continua deve essere simile.
- Omogeneità della regressione: i coefficienti delle covariate sono gli stessi per ogni gruppo formato da variabili categoriali e misurato sulle variabili dipendenti.
- Distribuzione normale multivariata: si assume la normalità multivariata se ogni variabile segue la distribuzione normale. La MANOVA è robusta anche in caso di violazione di questa assunzione, se la dimensione del campione non è piccola. Per verificare tale assunzione si possono utilizzare procedure diverse: è possibile rappresentare graficamente i dati e verificare la presenza di simmetria o calcolare la curtosi oppure valutare la presenza di valori anomali; si può utilizzare il test di Shapiro-Wilks, che considera come ipotesi nulla la normalità dei dati, oppure il diagramma "quantile-quantile", che confronta i quantili della distribuzione normale con quelli dei dati. Queste procedure sono quelle utilizzate anche nel caso univariato, perché si assume che, se tutte le variabili sono normali, allora complessivamente avremo una distribuzione normale multivariata. Altri controlli si possono effettuare calcolando i quadrati delle distanze dalla media delle osservazioni, poiché se la popolazione è normale e n e $n-p$ sono maggiori di 30, allora tali distanze dovrebbero seguire la distribuzione di un χ^2_{n-p} .
- Assenza di valori anomali.
- Le covariate sono linearmente correlate o in una relazione nota con le variabili dipendenti, la forma della relazione tra covariate e variabili risposta deve essere nota. Spesso le covariate sono trasformate per stabilire una relazione lineare.

Durante l'analisi di dati eco/biologici i principali problemi che si riscontrano nell'utilizzo della MANOVA riguardano l'adeguata numerosità del campione (nella maggior parte dei casi i campioni raccolti sono poco numerosi), con il conseguente allontanamento dalla normalità e il diverso numero di repliche per gruppo. Inoltre la normale è una variabile aleatoria continua, mentre le abbondanze sono dei conteggi e quindi valori discreti.

CAPITOLO 2

PERMANOVA

Come detto in precedenza le distribuzioni delle abbondanze degli esemplari delle specie sono di solito molto asimmetriche e aggregate. Le abbondanze, sono valori discreti: le specie con una media bassa hanno spesso distribuzioni asimmetriche perchè sono necessariamente ridotte a zero, le specie rare, invece, contribuiscono a riempire i data set con un gran numero di zeri. Generalmente le numerosità sono basse ed è frequente la situazione in cui sono presenti più parametri che osservazioni ovvero il caso in cui la matrice X dei dati avrà $n < p$.

Nella maggior parte dei data set ecologici e biologici non è possibile effettuare la tradizionale MANOVA, che viene sostituita da metodi non parametrici, basati sulle permutazioni che permettono di effettuare l'analisi della varianza, utilizzando delle assunzioni meno restrittive.

Partizionare la variabilità, come nell'ANOVA multifattoriale, è particolarmente importante per testare ipotesi in un ecosistema complesso con variabilità temporale e spaziale. Il partizionamento è anche utile per testare ipotesi multivariate con disegni sperimentali che includono più fattori.

La PERMANOVA, tecnica non parametrica, fa assunzioni più deboli sulle distribuzioni dei dati, consentendo quindi analisi affidabili anche nei casi in cui la distribuzione dei dati sia ben lontana dalla normalità, non permettendo l'utilizzo della MANOVA.

Punto di partenza per il calcolo dalla PERMANOVA è la matrice di distanze che può essere calcolata utilizzando, oltre alla distanza euclidea, una qualunque distanza sia metrica che

semimetrica. In particolare, la distanza euclidea non è in grado di "adattarsi" adeguatamente ai dati ecologici, enfatizzando in modo adeguato la composizione delle specie (rif. Faith & al. 1987, Legendre & Legendre 1998, Clarke 1993, Clarke et al. 2006c).

2.1 PERMANOVA: approccio geometrico alla MANOVA

"La PERMANOVA può essere vista come un metodo che porta un approccio geometrico alla MANOVA" (Edgington 1995).

Consideriamo una matrice di dati X di dimensioni $n \times p$ con p numero di variabili e n numero di osservazioni. Possiamo pensare a ciascuna delle p variabili come dimensioni dello spazio e a ciascuna delle n osservazioni della matrice come punti disposti nello spazio p-dimensionale, in accordo con il valore assunto in ciascuna delle p variabili. Si può quindi calcolare, rispetto a tutti i punti considerati, quello centrale detto "centroide globale". Nel caso si stia utilizzando un sistema di misura euclideo, il centroide globale viene semplicemente calcolato con la media aritmetica di ciascuna variabile. Allo stesso modo è possibile calcolare un centroide rispetto a ogni variabile, il quale, sempre pensando in ottica di uno spazio p-dimensionale, sarà in mezzo alla nuvola di punti delle corrispondenti osservazioni.

In analogia con l'ANOVA calcolata per un modello univariato, è possibile considerare la distanza tra uno qualunque dei punti rispetto al centroide globale formata da due parti: la distanza tra il punto e il centroide del suo gruppo più la distanza tra il centroide del gruppo e il centroide globale.

Le somme dei quadrati possono quindi venire calcolate nel seguente modo:

SST : somma dei quadrati delle distanze tra le unità e il centroide globale

$SSres$: somma dei quadrati delle distanze tra le unità e il centroide del gruppo

SSa : somma dei quadrati delle distanze tra il centroide del gruppo e il centroide globale

La relazione $SST = SSres + SSa$ è la stessa dell'ANOVA univariata.

2.2 Teorema di Huygen

La distanza euclidea non è adeguata per rappresentare data set che misurano abbondanze delle specie, poichè non è in grado di enfatizzare in modo adeguato l'insieme delle varie specie (rif. Faith & al. 1987, Legendre & Legendre 1998, Clarke 1993, Clarke et al. 2006c).

In ecologia la misura di similarità più utilizzata è quella di Bray-Curtis (usata successivamente nell'analisi dei dati), mentre tra le altre misure comunemente adottate in ambito eco/biologico ricordiamo la misura di Kulczynski, Jaccard, Gower e la devianza binomiale.

Sfortunatamente per molte di queste misure il centroide globale non può essere calcolato semplicemente come il vettore delle medie delle singole variabili, infatti a tutt'oggi per molte misure non euclidee non è ancora possibile calcolarlo.

Il calcolo è reso possibile solo sfruttando il teorema di Huygens, secondo il quale la somma dei quadrati delle distanze dei punti dal centroide del gruppo è uguale alla somma dei quadrati delle distanze tra i punti divisa per il numero di punti. Si può calcolare la somma dei quadrati senza dover prima trovare i rispettivi centroidi, potendo così utilizzare anche misure per le quali non è ancora nota una appropriata procedura di calcolo. Sempre per il teorema di Huygens, le uniche proprietà richieste alla misura di similarità utilizzata sono:

- l'essere non negativa
- simmetrica
- la distanza di un punto da sè stesso nulla.

Una volta calcolata la matrice di distanza, possiamo definire d_{ij} come la distanza tra l'unità i-esima e l'unità j-esima. La somma dei quadrati SST può essere ridefinita come la somma delle distanze tra tutti i punti divisa per n con n numero totale di osservazioni, invece la somma dei quadrati dei residui, $SSres$, è la somma dei quadrati delle distanze delle unità

appartenenti allo stesso gruppo (supponendo che si tratti di un disegno bilanciato). La quantità SSa, infine, può essere definita allo stesso modo come la somma dei quadrati delle distanze delle diverse variabili rilevate, oppure calcolata tramite la relazione:

$$SSa = SST - SSres$$

2.3 Test pseudo-F

Per la spiegazione di tale test consideriamo il caso univariato, in quanto l'estensione al caso multivariato non implica nessuna complicazione particolare. Quindi il modello considerato sarà:

$$Y = X\beta + \varepsilon$$

Y è il vettore $n \times 1$ delle variabili risposta con n numero di osservazioni, X è la matrice delle variabili esplicative $n \times p$ con n numero di osservazioni e p numero di parametri presenti nel modello, β è il vettore dei parametri $p \times 1$ ed ε il vettore dei residui di dimensioni $n \times 1$ con distribuzione normale di media nulla e varianza comune Σ.

Dal punto di vista formale stiamo testando il seguente sistema d'ipotesi:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0 \quad \mathrm{i} = 1, .., \mathrm{p}$$

Il test pseudo-F, ripreso dal caso univariato, è il rapporto tra la varianza spiegata dal modello e la varianza residua, ciascuna divisa per un opportuno numero di gradi di libertà. Nel caso si costruisca a partire dalla somma dei quadrati delle distanze si ottiene:

$$F = \frac{SSa/(a-1)}{\mathrm{SSres}/(\mathrm{N}-\mathrm{a})}$$

Ricordiamo che SSa è la somma dei quadrati delle distanze tra il centroide del gruppo e il centroide globale, $SSres$ è la somma dei quadrati delle distanze tra le unità e il centroide del gruppo, $(a-1)$ sono i gradi di libertà associati con il fattore e $(N-a)$ sono i gradi di

libertà residui con N numero di unità. Consideriamo la matrice di distanze la calcolata con la distanza euclidea. In tal caso la distribuzione del test è una F di Snedecor con $(a-1)$ gradi di libertà al numeratore e $(N-a)$ gradi di libertà al denominatore. All'aumentare del test pseudo-F, la possibilità che l'ipotesi nulla sia vera diminuisce.

Nel caso multivariato la matrice di distanza non viene più calcolata utilizzando la distanza euclidea, quindi la distribuzione del test F non è più una F di Snedecor, ma è ignota, ed è chiaro perchè venga definita come "pseudo" F.

La pseudo-F è nota solo nel caso in cui si considera un piano sperimentale con un solo fattore, quindi se si calcola la matrice di distanze utilizzando la distanza euclidea e i residui del modello sono normali, coincide con l'usuale test F calcolato nell'ANOVA (univariata) a una via.

Qualora si voglia trovare una distribuzione appropriata per la statistica pseudo-F sotto l'ipotesi nulla si utilizzano procedure di permutazione. L'idea su cui si fondano i test via permutazione è che qualora non siano presenti effetti causati dal fattore, sia possibile scambiare casualmente le unità tra i gruppi, in tal modo scambiando casualmente i valori tra le variabili, si ottengono altri risultati per la statistica pseudo-F. Nel caso in cui l'ipotesi nulla sia vera, il valore della statistica pseudo-F, ottenuta dopo la riallocazione delle unità, sarà simile al valore ottenuto con le osservazioni reali. Formalizzando, possiamo scrivere:

$f(x_1, x_2, .., x_n) = f(x_{\pi 1}, x_{\pi 2}, \dots, x_{\pi n})$ dove $\pi_1, \pi_2, \dots, \pi_n$ sono le permutazioni di indici $1, 2, \dots, n$

La distribuzione di frequenza della statistica qui descritta è discreta, infatti i dati possono venire riallocati in modo univoco un numero finito di volte. La probabilità associata al test sotto l'ipotesi nulla è calcolata come il rapporto tra il numero di volte in cui il test pseudo-F ottenuto dopo il riordinamento dei dati è maggiore del valore della statistica test pseudo-F osservato.

$$P = \frac{(Numero\ di\ F^{\pi} \geq F)}{(Totale\ di\ F^{\pi})}$$

Nel caso univariato, considerando un solo fattore, si ottiene l'equivalente di un normale test ANOVA a una via e, se vengono effettuate tutte le possibili permutazioni invece di una

parte, il test sarà esatto. Le permutazioni possono venire eseguite utilizzando la stessa procedura sia sulla matrice dei dati che sulla matrice di similarità.

Con numerosità elevate non è possibile calcolare tutte le possibili permutazioni, che, ricordiamo, in molti casi sarebbero un numero tale da richiedere molto tempo per il calcolo. Manly B.F.J. (1997) consiglia di eseguirne almeno 1000 per un test con $\alpha = 0{,}05$, e almeno 5000 per un test con $\alpha = 0{,}01$ (Manly 1997).

Considerando l'attuale potenza di calcolo di un comune personal computer, effettuare un test anche su un disegno fattoriale complesso con un gran numero di permutazioni, se ne calcoliamo solamente una parte, il tempo richiesto per l'elaborazione è più un problema.

Nel caso in cui non vengano considerate tutte le possibili permutazioni, è conveniente effettuare il test, includendo sia nel numeratore che nel denominatore i valori osservati originariamente ("+1" nella formula). Il significato di tale correzione è che, naturalmente, una delle possibili permutazioni, anche se non viene calcolata, è l'ordinamento reale dei dati che in tal modo viene automaticamente incluso nel calcolo del test.

Poichè il test è basato su estrazioni casuali dallo spazio di tutte le possibil permutazioni, se lo ripetiamo più volte, naturalmente, i p-values risulteranno tra loro leggermente differenti. Le differenze tra i p-values, come ci aspettiamo secondo il teorema del limite centrale, diminuiscono all'aumentare del numero di replicazioni effettuate per il calcolo del test. La minima significatività raggiungibile dal p-value è direttamente influenzata dal numero di replicazioni effettuate. Nel caso di un p-value che tende a zero, replicando il test e aumentando di volta in volta il numero di replicazioni, questo p-value tenderà a diventare sempre più piccolo.

Trattando invece un insieme di dati multivariati, le unità, sulle quali sono state rilevate più variabili, vengono (come nel caso univariato) riallocate casualmente tra i gruppi. Quando viene effettuata la riallocazione, tutte le variabili di un'unità vengono riallocate nello stesso momento (quindi, parlando in termini di matrici, si rialloca un'intera riga o un'intera colonna) evitando così di mescolare casualmente i valori nella matrice dei dati.

Tuttavia l'ipotesi di indipendenza delle variabili nel caso multivariato non è "realistica", poichè stiamo trattando diverse misurazioni effettuate sulle stesse unità, quindi non è

possibile assumere questo tipo di indipendenza. Questo non è un problema perchè quando vengono effettuate le permutazioni, la riallocazione avviene a livello di unità e non di singola variabile.

Inoltre, le unità devo essere indipendenti una dall'altra, altrimenti, nel caso sia presente una qualche relazione, quest'ultima verrebbe annullata dalla procedura di permutazione.

Per poter effettuare il test dobbiamo assumere che le osservazioni siano scambiabili liberamente tra loro, ipotizzando H_0 vera. Quindi supponiamo che le osservazioni siano indipendenti e che abbiano la stessa distribuzione. E' necessario precisare che stiamo ipotizzando che le osservazioni siano scambiabili, che le variabili siano indipendenti e abbiano la stessa variabilità, altrimenti non sarebbero realmente scambiabili.

Capitolo 3

Tipi di permutazioni

Prendiamo in considerazione i seguenti metodi di permutazione:

1) Permutazioni illimitate di dati grezzi
2) Permutazioni dei residui sotto il modello ridotto
3) Permutazioni dei residui sotto il modello completo

Vengono utilizzati questi metodi di permutazione perchè non ignorano una possibile relazione tra le variabili esplicative, nel caso questa sia presente. Tale relazione non viene alterata effettuando le permutazioni sulle variabili risposta o sui residui. Limitiamo inoltre la trattazione a questi tre metodi, perchè sono quelli presenti nel software PRIMER. Anche questa volta, per semplicità di notazione, prenderemo in considerazione il caso univariato, in quanto l'estensione al caso multivariato non presenta particolari difficoltà. Tutti i metodi sono uguali ed esatti nel caso si stia testando un modello con un solo predittore, inoltre sono tutti e tre asintoticamente equivalenti, e risultano simili nel caso in cui gli errori del modello soddisfino l'assunzione di normalità. Un accurato confronto tra i metodi è stato fatto da Anderson & Legendre (1999).

Consideriamo il seguente modello:

$$Y = \mu + \beta_{12}X + \beta_{21}Z + \varepsilon$$

Dove Y è la variabile risposta, X e Z sono le variabili esplicative, β_{12} e β_{21} sono i coefficienti di regressione parziale calcolati con il metodo dei minimi quadrati sul modello di regressione multipla di Y su X, Z, ed ε termine casuale di errore. In particolare β_{21} indica il coefficiente di regressione parziale per la relazione tra Y e la seconda variabile Z al netto

dell'effetto della prima variabile X. Il modello considerato è univariato, tuttavia l'applicazione al caso multivariato resta identica. L'ipotesi nulla di interesse è $\beta_{21} = 0$, ovvero non che non sia presente un effetto significativo della variabile Z all'interno del modello. Per il calcolo viene utilizzata l'usuale statistica t che ha la seguente forma:

$$t = \frac{(b_{21} - 0)}{se(b_{21})}$$

dove b_{21} è la stima calcolata con i minimi quadrati di β_{21} e $se(b_{21})$ è lo standard error del coefficiente di regressione parziale.

3.1 Permutazioni illimitate di dati grezzi

Per il calcolo delle permutazioni illimitate sui dati grezzi l'ipotesi nulla utilizzata è $H_0 : \beta_{21} = 0$.

La procedura utilizzata per il calcolo è la seguente:

1) Vengono stimati i parametri utilizzando il metodo dei minimi quadrati di Y condizionati a X e Z contemporaneamente per stimare b_{21}. Viene poi calcolata la statistica t, definita precedentemente, sui dati originali, per verificare $H_0 : \beta_{21} = 0$. La statistica t calcolata sui dati originali, poi verrà chiamata t_{obs}.
2) I valori della variabile risposta Y vengono permutati casualmente per ottenere Y^*.
3) Vengono stimati i parametri di Y^* condizionati a X e Z (non permutate) contemporaneamente, sempre utilizzando i minimi quadrati, per stimare b_{21}^* stima di β_{21}. Dopo aver effettuato le permutazioni viene calcolato t^* sui dati permutati. Gli ultimi due passaggi vengono ripetuti un numero elevato di volte per generare una distribuzione di valori di t^* sotto permutazione.
4) Viene calcolato:

$$P = \frac{(Numero\ di\ |t^*| \geq |\mathrm{t_{ref}}|)}{(Numero\ di\ permutazioni)}$$

Non essendo valido l'assunto di scambiabilità, neppure asintoticamente, si utilizza questo tipo di permutazioni per ottenere una buona approssimazione del test per disegni ANOVA complessi. L'errore del primo tipo che si ottiene è vicino ad α, tuttavia, all'aumentare della dimensione del campione, tenderà a diventare più conservativo (meno potente) rispetto al test che effettua le permutazioni dei residui. Dal punto di vista computazionale questa è la scelta più rapida. Inoltre non è necessario un dataset di grandi dimensioni per ottenere buoni risultati. Questo metodo non altera nel corso delle permutazioni, se presenti, le covariate tra X e Z e all'interno di X, se X è una matrice. Tra tutti e tre i metodi è il migliore nel caso di covariate e in presenza di collinearità tra X e Z. Bisogna prestare particolare attenzione nel caso siano presenti outlier nella matrice X, e nel caso in cui $\beta_{21} \neq 0$ poichè, in tali occasioni si assiste a un aumento dell'errore del primo tipo che rimane anche aumentando la numerosità delle osservazioni, sconsigliando quindi l'utilizzo di questo metodo di permutazione.

3.2 Permutazioni dei residui sotto un modello ridotto

In questo caso le unità utilizzate per le permutazioni sono i residui del modello lineare.
Si considerano il modello ridotto

$$Y = \beta_0 + \beta_1 X + \varepsilon'$$

e vera l'ipotesi nulla $H_0 : \beta_{21} = 0$.
Freedman e Lane (1983) hanno proposto la seguente procedura per l'esecuzione delle permutazioni:

1) Vengono stimati i parametri di Y su X e Z contemporaneamente per ottenere b_{21} , stima di β_{21} , e un valore di riferimento t_{obs} per i dati reali.
2) Viene calcolata la regressione per la sola variabile Y su X, utilizzando il modello ridotto delineato precedentemente per stimare b_0 per β_0, b_1 per β_1 e i residui ε.
3) I residui della regressione calcolata al punto precedente vengono permutati casualmente ottenendo così e^*.
4) Viene calcolato Y^* aggiungendo ai valori stimati i residui permutati precedentemente.

5) Viene calcolata la regressione di Y^* su X e Z, in accordo con

$$\mathrm{E}(Y^*) = \beta_0^* + \beta_1^* X + \beta_2^* Z + R^*$$

per ottenere una stima b_{21}^* di β_{21} e poter calcolare la statistica t* sui dai permutati. Gli ultimi due passaggi vengono ripetuti un numero elevato di volte per generare una distribuzione di valori di t^* sotto permutazione.

6) Viene calcolato P secondo il test visto nel punto 4 del metodo precedente.

Come il metodo precedente, anche la permutazione dei residui sotto il modello ridotto permette di preservare le covarianze presenti tra le variabili esplicative e tra le variabili esplicative e la variabile risposta. Freedman e Lane (1983) indicano due condizioni per l'utilizzo di questo metodo di permutazioni: la matrice X non deve contenere outlier particolarmente lontani dal resto dei dati, e tra X e Z non vi deve essere un'alta correlazione. Il campione, inoltre, deve essere abbastanza numeroso.

Il test, a differenza del precedente, non è esatto, ma solamente asintoticamente esatto.

Questo metodo di permutazioni è quello usato di default nel software PRIMER.

Fornisce il test più potente e più accurato per l'errore del primo tipo per piani sperimentali multifattoriali nel maggior numero di circostanze (M. J. Anderson & P.Legendre 1999). Nel caso le variabili della matrice X non siano indipendenti o nel caso in cui $\beta_{21} \neq 0$, il metodo delle permutazioni sotto il modello ridotto è quello che più di tutti si avvicina a un test esatto minimizzando lo scarto dall' α nominale.

3.3 Permutazioni dei residui sotto il modello completo.

Questo metodo di permutazione, sviluppato da ter Braak (1990, 1992), utilizza i residui del modello di regressione completo come unità per generare le permutazioni.

1) Viene stimato, sempre utilizzando i minimi quadrati il modello di regressione

$$Y = b_0 + b_{12} X + b_{21} Z + \varepsilon$$

Dove b_0 è la stima della media, b_{12} e b_{21} sono le stime dei coefficienti di regressione e ε i residui. Viene inoltre calcolato $\mathrm{t_{ref}}$.

2) I residui ε vengono permutati casualmente, generando ε^*.
3) Calcoliamo Y^*, stima di Y sotto permutazione utilizzando il modello procedente, sostituendo ε con ε^*.
4) A partire da Y^* stimiamo un nuovo modello per ottenere b_{21}^*e calcoliamo $t^* = \frac{(b_{21}^* - b_{21})}{se(b_{21}^*)}$ per testare l'ipotesi nulla $b_{21}^* = b_{21}$.

   Gli ultimi due passaggi vengono ripetuti un numero elevato di volte per generare una distribuzione di valori di t^* sotto permutazione.
5) Viene calcolato P secondo il test visto nel punto 4 del primo metodo.

Anche questo metodo, come i precedenti, preserva eventuali relazioni presenti tra le variabili esplicative, ed è asintoticamente esatto. Inoltre permette di ottenere risultati comparabili col secondo metodo che è probabile sia ugualmente appropriato nella maggior parte dei casi. Solamente in casi particolari, con outlier estremi ed errori non normali su insiemi di dati molto piccoli, risulta preferibile effettuare le permutazioni sui residui del modello ridotto. Dal punto di vista computazionale, quest'ultimo metodo risulta meno gravoso solamente rispetto alla permutazione dei residui sotto il modello ridotto.

Un'ultima nota per la scelta del tipo di permutazioni: in caso di modelli con poche replicazioni non è consigliabile utilizzare i metodi che effettuano le permutazioni dei residui, poichè non è detto che in tal caso il modello sia sufficientemente rappresentativo della vera distribuzione delle osservazioni, e di conseguenza che non lo siano nemmeno i residui. Viene in tal caso raccomandato il metodo che effettua le permutazioni direttamente sui dati.

3.4 Tipi di somma di quadrati

Può capitare, per fattori esterni o ancora al momento della pianificazione degli esperimenti, che i fattori considerati abbiano un differente numero di replicazioni, e che, quindi, il disegno sperimentale risulti sbilanciato. Nel caso in cui il disegno sperimentale comprenda solamente un fattore, questo non rappresenta un problema, ed è possibile effettuare

normalmente la PERMANOVA. In tal caso i fattori presenti nella stima delle componenti di variabilità e nel denominatore del test F non saranno necessariamente dei numeri interi, a differenza del caso bilanciato. Al momento di eseguire le permutazioni sarà ancora possibile allocare casualmente le unità tra i vari gruppi, ma sarà necessario tenere in considerazione le diversa numerosità di questi ultimi. Ciascuna unità, infatti, non ha più la stessa possibilità di cadere in uno dei gruppi presi in considerazione, ma avrà una probabilità proporzionale alla numerosià dei campioni. E' tuttavia possibile continuare a utilizzare l'ipotesi base delle permutazioni: tutti i possibili riarrangiamenti sono ugualmente probabili.

I problemi nascono, naturalmente, nel momento in cui consideriamo un disegno con più fattori, ciascuno con un numero di replicazioni differente.

Inoltre, nel caso multivariato, si aggiunge il problema che gli effetti principali dei fattori e i termini di interazione non sono più indipendenti l'uno dall'altro, quindi l'ordine utilizzato per inserire i fattori nel disegno fattoriale diventa importante.

Cominciamo considerando come esempio una ANOVA a due vie, con fattori A, B e l'interazione $A \times B$. Nel caso in cui il disegno sia bilanciato, l'ammontare della variabilità spiegata da ciascuno dei termini del modello è completamente indipendente da termine a termine, ovvero ciascun termine spiega una porzione di variabilità differente rispetto agli altri. Al contrario, con un disegno non bilanciato avremo una sovrapposizione delle porzioni di variabilità spiegate da ciascun termine.

In questo caso, quindi, con un disegno non bilanciato, è possibile effettuare il partizionamento della variabilità in più modi e la scelta tra i vari tipi dipenderà da come scegliamo di trattare la possibile sovrapposizione tra i termini del modello. Sono stati definiti quattro tipi di somme dei quadrati (tipo I, II, III e IV). La terminologia è stata inizialmente coniata dagli sviluppatori del software SAS, e successivamente è diventata di uso comune. Di seguito verrà utilizzata la sigla originale inglese "SS" (sums of squares) seguendo la notazione utilizzata da PRIMER, inoltre si ricorda che la trattazione è limitata solamente ai tre metodi presenti nel software. Nel caso si stia lavorando con un modello bilanciato i tre metodi portano allo stesso risultato. Vediamo brevemente, senza addentrarci nello specifico, i vari tipi.

- SS Tipo I, comunemente chiamato “sequenziale”. Ciascun termine del modello viene stimato dopo i termini che lo precedono. L’ordine in cui i termini sono elencati nel disegno fattoriale è importante, infatti, modificandolo verrà modificata anche la stima della somma dei quadrati. In questo caso, a differenza dei successivi tipi di SS, la somma dei valori delle somme dei quadrati dei singoli termini sarà uguale alla somma totale. Questo approccio è il più indicato per modelli gerarchici per i quali esiste un ordine naturale dei termini. E’ possibile cercare di stimare l’ammontare della sovrapposizione della variabilità spiegata dai termini, cambiando di volta in volta l’ordine in cui compaiono i termini nel modello e verificando i cambiamenti del risultato.

  Facciamo un esempio con due variabili A, B e con la relativa interazione $A \times B$, inserite in questo ordine nel disegno fattoriale; verrà calcolata innanzitutto SS_A (ignorando gli altri termini), SS_B viene invece calcolata condizionando rispetto ad A. $SS_{A\times B}$, infine, viene calcolata condizionando rispetto a entrambi i termini.

  Questo tipo di SS è disponibile di default nel software statistico R al momento del calcolo dell’ANOVA. Tramite un pacchetto aggiuntivo è possibile utilizzare anche gli altri due tipi.

  Generalmente, in campo statistico, viene usato questo tipo perchè permette una stima più precisa della ripartizione della variabilità.

- SS Tipo II, generalmente definito come un’analisi condizionata. Il secondo tipo di SS per un certo termine, è dato dalla riduzione nell’SS residuo dovuta all’aggiunta del termine successivo, dopo che gli altri sono stati inclusi nel modello, al di fuori di quelli che contengono l’effetto che deve essere testato.

  Potenzialmente potrebbe succedere che, sommando le varie componenti della SS, non si giunga alla SS totale, ignorando le parti di varabilità condivise tra più termini.

  L’ordine in cui tali termini vengono inseriti nel disegno non modifica la ripartizione della variabilità tra i fattori.

  Considerando il disegno fattoriale dell’esempio precedente per il calcolo di SS_A, si condiziona A rispetto a B. SS_B viene calcolato condizionando B rispetto ad A, mentre $SS_{A\times B}$ si calcola condizionando l’interazione rispetto a entrambi i termini.

  Questo tipo in particolare è disponibile di default per il software statistico SPSS.

- SS Tipo III definito come un analisi completamente parziale. Ogni termine del modello è stimato solo dopo aver considerato tutti gli altri termini del modello completo.

  Con questo tipo l'ordine in cui i termini vengono inseriti nel disegno non ha importanza, in quanto non sono presenti aree di sovrapposizione.

  Usando questa SS viene comunque assicurata la completa ortogonalità (ovvero l'indipendenza) tra i termini di tutte le ipotesi.

  Utilizzando nuovamente il disegno fattoriale degli esempi precedenti SS_A si calcola condizionando A rispetto a B e rispetto ad $A \times B$. SS_B si calcola condizionando rispetto ad A e rispetto all'interazione, mentre $SS_{A\times B}$ si calcola condizionando rispetto a entrambi i termini.

  Nel caso si trattino dati ecologici questo tipo è il più usato dalla stampa specializzata per disegni non bilanciati, perché tende a essere il più conservativo tra i tre. Tuttavia, all'aumentare dello sbilanciamento delle replicazioni presenti nel disegno fattoriale, la quantità sovrapposta potrebbe iniziare a essere considerevole e portare a potenziali discrepanze, anche non trascurabili, nelle stime del test F.

E' doveroso aggiungere che le conclusioni dell'analisi non saranno comunque stravolte a seconda del tipo di somma dei quadrati che verrà utilizzata. Calcolando più volte la PERMANOVA sullo stesso disegno fattoriale, e cambiando unicamente il tipo di SS utilizzato, viene modificato soltanto la ripartizione della variabilità tra i fattori, mentre le quantità relative alle interazioni non cambiano.

Capitolo 4

Fattori fissi e casuali – componenti di varianza

Tutti i fattori presenti in un piano fattoriale ANOVA possono venire considerati fissi o casuali. Nel caso un modello contenga sia fattori costanti che casuali, si tratta di un modello misto (mixed model). Anche nel caso di un ANOVA uni variata, la decisione se un fattore sia fisso o casuale ha conseguenze importanti sulle assunzioni del modello, in particolare sul valore atteso della media dei quadrati (denotato con la sigla EMS, dall'inglese expectation of mean squares). Inoltre la decisione sulla tipologia del fattore influenza le conclusioni traibili dal test e il tipo di inferenza che è possibile effettuare.

Nel caso di un fattore costante, siamo a conoscenza di tutti i vari stati in cui lo possiamo trovare, generalmente perché tali stati del fattore sono discriminanti per il tipo di esperimento che si vuole indagare tramite l'ANOVA. Tutti i livelli del fattore costante, o per lo meno quelli a cui siamo interessati, vengono considerati nell'esperimento. Ciascun livello ha un significato intrinseco e, nel caso dovessimo replicare lo stesso esperimento, i livelli scelti sarebbero comunque gli stessi. Gli effetti di un fattore fisso sono considerati come valori costanti in tutti gli stati. In un modello la componente di variabilità attribuita a un fattore costante è considerata come una somma di quadrati relativa all'effetto costante, divisa per un numero appropriato di gradi di libertà. Dal punto di vista inferenziale, le conclusioni che è possibile trarre possono essere applicate solo ai livelli considerati del fattore fisso, e non ad altri. Ad esempio, effettuando uno stesso esperimento in due differenti siti ed essendo interessati alle differenze rilevate nei risultati dell'esperimento stesso dovuti alle caratteristiche specifiche del sito, lo andremo a considerare come un fattore fisso.

Per un fattore casuale, invece, i livelli inclusi nell'esperimento sono confrontati con un sottoinsieme casuale rispetto a tutti i possibili stati che avremo potuto includere. Non viene fatta una distinzione particolare tra un fattore e un altro, generalmente perché vengono considerati allo stesso modo in un elenco, e non viene dato un significato particolare ai vari livelli che vengono utilizzati per trarre delle conclusioni rispetto a una possibile popolazione. Come nell'esempio precedente, consideriamo un esperimento effettuato in due siti differenti (sito uno e sito due). Qui vengono poi effettuate delle misurazioni (che saranno i fattori fissi del disegno sperimentale) per verificare se sono presenti differenze tra i due luoghi. In questo caso i due siti sono fattori casuali, perché ciò che effettivamente ci interessa non sono le differenze dovute al sito in sé, ma quelle riscontrate sulle misurazioni.

I fattori casuali, a differenza di quelli costanti, contribuiscono ad aggiungere al modello un ulteriore fonte casuale di variabilità, oltre alla varianza dell'errore. Un fattore casuale è quindi una realizzazione di una variabile casuale i.i.d. di media nulla e di varianza σ^2.

Nel caso l'esperimento venga ripetuto, probabilmente non sceglieremo di nuovo gli stessi livelli. Quando costruiamo la statistica F e calcoliamo il p-value per un fattore casuale, le conclusioni che possiamo trarre dalla variabilità della componente sono estendibili a tutti i possibili livelli della popolazione che potrebbero venire scelti, e non solamente ai livelli inclusi nell'esperimento.

Quindi è evidente l'importanza della scelta del tipo di fattore ai fini delle conclusioni che è possibile trarre dai test eseguiti.

La presenza all'interno dello stesso modello di fattori fissi e casuali, fa di questo un modello misto. I modelli misti, rispetto a quelli fissi e a quelli casuali, hanno un campo di applicazione molto più vasto, e risultano in molti casi più appropriati rispetto a un modello contenente uno solo dei due effetti, in quanto riescono a rappresentare meglio la realtà e il modo in cui sono stati raccolti i dati.

Vediamo un esempio famoso di un modello misto per chiarire definitivamente le differenze (Searle SR, Casella G & McCulloch CE, 1992):

Supponiamo di voler testare l'effetto di I dosi di interesse di un nuovo farmaco. La prova viene eseguita in J differenti cliniche. Le cliniche sono state casualmente estratte tra tutte le

cliniche di uno Stato. Ciascuna clinica testa l'effetto del farmaco su k pazienti. Un possibile modello per il k-esimo paziente sottoposto alla i-esima dose nella clinica j-esima potrebbe quindi essere:

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Dove α_i, β_j e γ_{ij} sono gli effetti dovuti della dose i, clinica j e l'interazione tra trattamento e clinica. Le dosi, naturalmente, sono ciò su cui vogliamo indagare. Le cliniche nelle quali sono state eseguite le prove sono estratte casualmente e quindi β_j è un effetto casuale. Infine, l'interazione tra un effetto fisso e un effetto casuale è anch'essa un effetto casuale.

Capitolo 5

Analisi dei dati

Il dataset che verrà utilizzato per le analisi seguenti è stato raccolto dal dipartimento di biologia marina dell'Università di Padova. Lo scopo dell'esperimento era quello di valutare e raccogliere ulteriori dati per lo sviluppo dell'allevamento di pectinidi (comunemente detti "canestrelli"), inoltre in concomitanza con la raccolta dati è stata eseguita anche una prova di allevamento.

L'esperimento è inziato alla fine di luglio 2009 con il posizionamento delle cime e dei corrispondenti collettori, ed è finito tra dicembre 2009 e gennaio 2010. La raccolta dei molluschi è stata effettuata mediante collettori (detti comunemente "sacchetti"), in due siti distinti, collocati a nord e a sud della bocca di porto di Chioggia, a circa 2,5 km dalla costa al largo delle località di Caleri e Pellestrina. Le quantità rilevate sono le abbondanze di molluschi raccolti, suddivisi per specie.

I raccoglitori sono stati posti a dieci profondità differenti, dai 5 ai 14 metri dalla superficie. Nel sito di Pellestrina sono state effettuate cinque distinte rilevazioni, tramite il posizionamento di cinque diverse corde per ciascuna profondità, mentre nel sito di Caleri per ciascuna profondità ne sono state effettuate quattro. La singola rilevazione sarà identificata mediante una lettera "P" o "C" per indicare il sito, seguita dalla profondità a cui è stata effettuata, e infine dalla corda di riferimento, con le lettere A-D per il sito di Caleri, e lettere A-E per il sito di Pellestrina. L'esperimento fin dall'inizio si presenta sbilanciato con un diverso numero di osservazioni tra i due siti. Il fatto che un esperimento pianificato come bilanciato al momento dell'analisi dei dati non sia più tale per problemi avvenuti in fase di raccolta è abbastanza frequente.

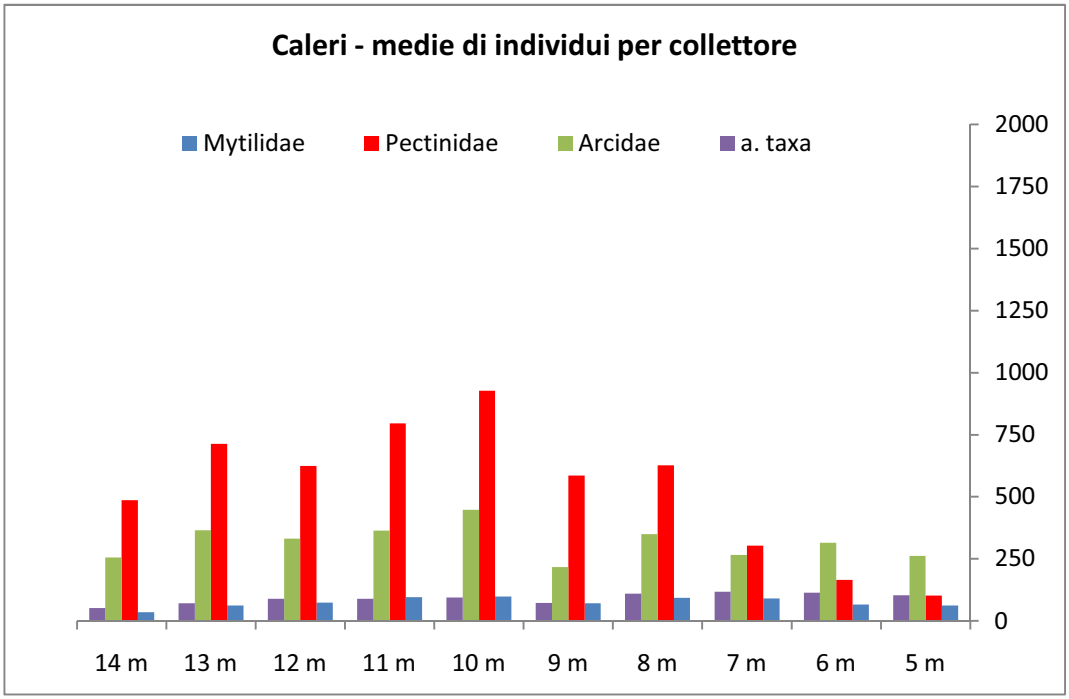
Sfortunatamente il raccoglitore posto a 5 metri di profondità sulla corda A nel sito di Caleri si è danneggiato nel corso della rilevazione, ed è stato escluso da tutte le analisi, aumentando la differenza nel numero di osservazioni disponibili tra i due siti. A tale conclusione si è giunti solamente osservando che gli esemplari raccolti con quel sacchetto erano molto inferiori rispetto agli altri, tuttavia l'ipotesi che vi sia stata effettivamente una rottura del contenitore non è tuttora verificata.

In totale la matrice dei dati presenta quindi 89 righe e 31 colonne, ciascuna riga corrisponde a un preciso raccoglitore, mentre sulle colonne troviamo le quantità di molluschi rilevate suddivise per specie.

Punto di partenza dello studio sarà una rapida analisi descrittiva, già presente nella tesi di laurea magistrale in biologia marina di Valentina Francesca Codognotto (relatore dott.ssa Monica Bressan), con la presentazione delle medie dei due siti condizionate rispetto alla profondità. Gli esemplari raccolti sono stati catalogati per famiglia, in modo da permettere una più agevole lettura delle tabelle e una rappresentazione grafica più facilmente interpretabile.

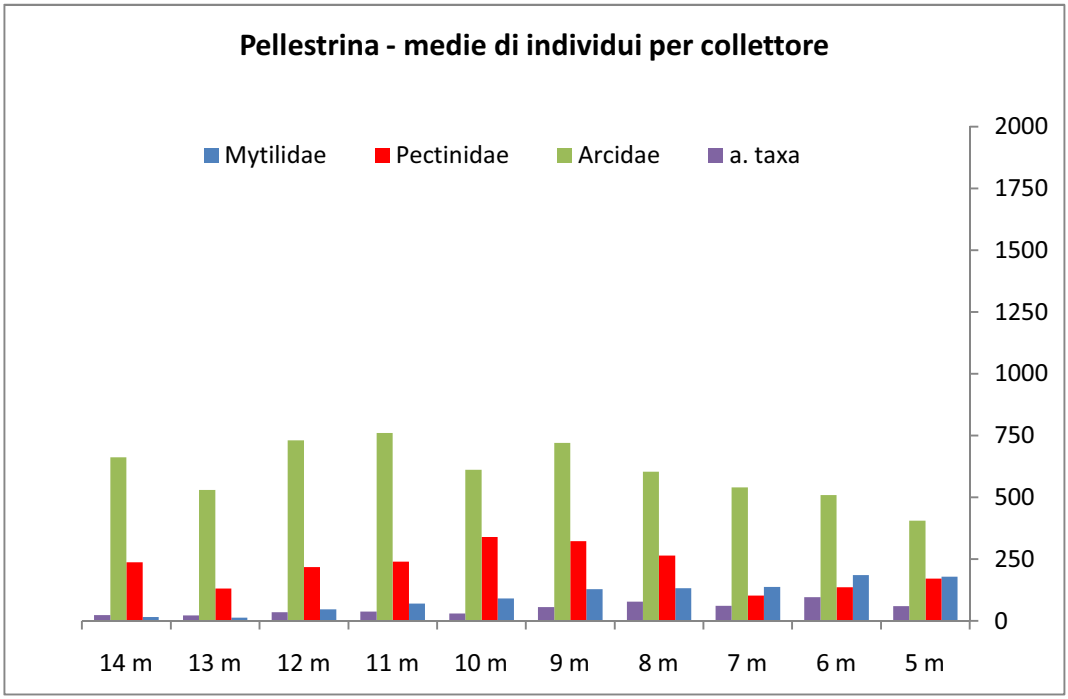
I dati saranno successivamente analizzati utilizzando il software statistico PRIMER 6, e in particolare il pacchetto aggiuntivo PERMANOVA+. Questo software è probabilmente il più utilizzato in ambito eco/biologico e permette di manipolare i dati tramite una semplice interfaccia grafica, fornendo un sufficiente numero di strumenti per l'analisi di dati multivariati.

Di seguito sono riportate le tabelle con le medie per sito condizionate per profondità, e i relativi istogrammi.



Medie di individui per collettore per il sito “Caleri”

	Mytilidae	Pectinidae	Arcidae	a. taxa	Totale
5 m	61,5	102,25	261,5	103	528,25
6 m	66,25	165,25	314,5	113,75	659,75
7 m	90,25	303	265,75	117,75	776,75
8 m	93,5	627,25	350	109,75	1180,50
9 m	71,75	586,25	216,75	72,5	947,25
10 m	97,75	927	448,25	94,5	1567,50
11 m	96,25	795,5	364,25	89,75	1345,75
12 m	74,25	624,25	332,25	88,75	1119,50
13 m	61,5	713,5	364,5	70,5	1210,00
14 m	35,25	486,5	255,25	51,5	828,50
Tot.	748,25	5330,75	3173,00	911,75	10163,75



Medie di individui per collettore per il sito “Pellestrina”

	Mytilidae	Pectinidae	Arcidae	a. taxa	Totale
5 m	179	170,6	405,2	59,2	814,00
6 m	185,2	135,8	509,2	95,8	926,00
7 m	137	102,6	539,8	60,8	840,20
8 m	132	264,2	603,6	78	1077,80
9 m	128,6	323	720	56	1227,60
10 m	91	339,2	610,8	30,2	1071,20
11 m	70,2	240,4	759,8	37,4	1107,80
12 m	47,4	218	730,8	35	1031,20
13 m	12,6	131,4	529,8	22,6	696,40
14 m	15,4	237,4	661,6	23	937,40
Tot.	998,40	2162,60	6070,60	498,00	9729,60

La quantità totale di molluschi raccolta nei due siti non è molto dissimile, tuttavia, al variare della profondità, le quantità riportate sono abbastanza differenti.

Dalle tabelle si nota che nel sito di Pellestrina le quantità riportate sono abbastanza uniformi per tutte le profondità, salvo un brusco calo a 13 metri dalla superficie. Nel sito di Caleri, invece, vi è un aumento del numero di individui fino a 10 metri di profondità, mentre tra i 10 e i 13 metri si nota una leggera diminuzione e un calo più marcato ai 14 metri. Viste le differenze sia al variare del sito che al variare della profondità si deduce la presenza di un'interazione tra le due variabili. Nei due istogrammi, in cui si è scelto di focalizzarsi principalmente sulla diversa ripartizione delle specie piuttosto che sulle abbondanze, sono riportate le frequenze rilevate solamente per le famiglie "Mytilidae", "Pectinidae", "Arcidae" e "a. taxa" (categoria in cui sono raccolte tutte le altre specie meno frequenti). Si nota chiaramente che la composizione nei due siti è molto differente. Nel sito di Caleri la famiglia più frequente, tra i 7 e i 14 metri è "Pectinidae" mentre per il sito di Pellestrina per tutte le profondità è "Arcidae".

Come già detto, sono state calcolate le medie solamente sulle famiglie più diffuse, le più rare invece sono state raccolte nella colonna "a. taxa", per rendere leggibili i grafici che altrimenti conterrebbero un numero troppo elevato di variabili, molte delle quali nulle o con valori prossimi allo zero.

Già da questa prima analisi si nota una diversa composizione delle specie presenti nei due siti.

Ricordiamo inoltre che il piano sperimentale risulta sbilanciato tra i due siti, per il sito di Pellestrina sono presenti ben undici osservazioni in più. Tuttavia, come già detto in precedenza, questo inconveniente non pregiudica l'utilizzo della PERMANOVA come tecnica per l'analisi della varianza.

Il primo accorgimento necessario, quando si trattano disegni sperimentali sbilanciati, è ricordarsi che, nel momento in cui si delinea il disegno fattoriale su cui calcolare la PERMANOVA, l'ordine in cui vengono inseriti i fattori influenza i risultati del test. Inoltre è importante verificare, come già detto in precedenza, il tipo di permutazioni effettuate e di somma di quadrati utilizzata nel calcolo.

Prima di creare la matrice di dissimilarità i dati sono stati trasformati applicando la radice quadrata. La matrice di dissimilarità è stata calcolata sulla misura di dissimilarità di Bray-Curtis che viene calcolata nel seguente modo:

$$d_{ll'} = \frac{\sum_{k=1}^{p} |x_{lk} - x_{l'k}|}{\sum_{k=1}^{p} (x_{lk} + x_{l'k})}$$

Sono state rimosse dal dataset le colonne corrispondenti alle specie meno frequenti tra tutti i raccoglitori. La matrice finale su cui calcoliamo la PERMANOVA ha solamente venti colonne, ben undici in meno rispetto alla matrice iniziale.

Nel caso si stia analizzando un disegno fattoriale con più di un fattore, PERMANOVA+ per effettuare il partizionamento della variabilità, stima un modello lineare addittivo partendo dalla matrice di dissimilarità utilizzata. Viene quindi ipotizzata una relazione addittiva lineare tra i fattori e che gli errori sono indipendenti e identicamente distribuiti attraverso tutta la matrice di dissimilarità.

Scopo del test è quello di verificare se sono presenti differenze nelle distribuzioni degli assemblaggi di molluschi nei due siti, tenendo inoltre conto della profondità alla quale viene effettuata la rilevazione. Il modello che andiamo a considerare è:

$$Y = si + co + pr + si{:}pr + si{:}co + pr{:}co + \varepsilon$$

Dove "si" rappresenta il fattore sito, "co" il fattore corda, "pr" il fattore profondità, "$si{:}pr$" l'interazione tra sito e profondità, "$si{:}co$" l'interazione tra sito e corda e infine "$pr{:}co$" è l'interazione tra profondità e corda. Il disegno che prendiamo in considerazione includerà sicuramente come primo termine la variabile "sito" come fattore fisso. Il secondo fattore da inserire invece è "corda", e infine "profondità". Il fattore "profondità" viene anch'esso considerato come fisso, infatti, insieme al sito è dell'analisi. Il fattore "corda" viene considerato come casuale, perché non ci è dato sapere quali fenomeni possono rendere differente una corda da un'altra, (come ad esempio una corrente calda che potrebbe lambire solamente alcune corde). Considerare tale fattore come casuale equivale a considerarlo come un ulteriore termine di variabilità.

Per poter permettere l'esecuzione regolare del test, non disponendo di replicazioni, escludiamo dal modello il termine relativo all'interazione di terzo livello, per disporre in tal modo di un numero sufficiente di gradi di libertà al fine di calcolare il test pseudo-F sulle rimanenti variabili.

Poichè il disegno sperimentale utilizzato non è bilanciato, e seguendo anche i consigli della stampa specializzata, utilizziamo il terzo tipo di SS (parziale) ed effettuiamo 9999 permutazioni dei residui sotto il modello ridotto, garantendo in tal modo una elevata stabilità ai p-value trovati.

L'ipotesi nulla che stiamo testando è l'assenza di differenze tra i diversi fattori.

Di seguito viene riportato l'output di Primer per il disegno fattoriale testato:

PERMANOVA table of results

Source	df	SS	MS	Pseudo-F	P(perm)	Unique perms
si	1	7625,3	7625,3	15,469	0,0226	8091
co	4	1808,1	452,02	2,6503	0,001	9915
pr	9	8747,7	971,97	5,1592	0,0001	9900
sixpr	9	4470,6	496,73	2,9124	0,0001	9905
sixco**	3	1481,2	493,75	2,8949	0,0011	9911
coxpr	36	6851,1	190,31	1,1158	0,2544	9829
Res	26	4434,4	170,55			
Total	88	38694				

** Term has one or more empty cells

L'interazione "*co:pr*", interazione tra i fattori "corda" e "profondità", presenta delle celle vuote. Il p-value è sufficientemente alto da non portarci a dubitare della veridicità dell'ipotesi nulla. Tuttavia, analizzando il significato di questa interazione, ovvero che le differenze dipendano contemporaneamente dalla corda e dalla profondità, fa riflettere sulla sua utilità, poichè "corda" è già un fattore casuale per il modello. Ricordiamo inoltre che l'interazione tra un fattore fisso e un fattore casuale è anch'essa un termine casuale. Analizzando "*si:co*", contenente l'interazione tra sito e corda, è presente una nota che

ricorda che la variabile ha una o più celle vuote. In questo caso il p-value molto alto porta a non accettare l'ipotesi nulla, tuttavia, vale la pena soffermarsi come per l'interazione "*co:pr*" sulla reale utilità dell'ipotesi che stiamo verificando.

I test sulle altre tre variabili "*si* ", "*pr*", "*si:pr*" presentano tutti i p-values sufficientemente bassi da poter rifiutare H_0.

Analizzando la colonna relativa alle somme dei quadrati SS e calcolando le somme delle righe, notiamo che non corrisponde al totale, e che, anzi, lo scarto non è indifferente. Questo è dovuto al tipo di somma dei quadrati utilizzata, cioè quella parziale, che esclude dal calcolo le quantità di variabilità sovrapposta tra i fattori presenti nel piano sperimentale. Se si cambia nuovamente la PERMANOVA, modificando solamente questo parametro, la ripartizione della variabilità tra i fattori sarà differente, modificando di conseguenza il valore corrispondente del test pseudo-F, ma non a tal punto da modificare i risultati del test e le conclusioni che è possibile trarre.

La colonna MS riporta le medie dei quadrati "mean squares" ottenute dividendo la somma dei quadrati per i gradi di libertà (df).

L'ultima colonna sulla destra "Unique perms" riporta il numero di permutazioni uniche trovate durante l'esecuzione. Tale valore non và trascurato, infatti nel caso non sia sufficientemente elevato, il software non è stato in grado di eseguire un numero sufficiente di permutazioni uniche, e quindi il p-value relativo a quel termine non è da considerarsi affidabile. In tal caso è possibile calcolare il p-value utilizzando tecniche di simulazione basate sul bootstrap, quindi eseguendo ricampionamenti con reinserimento a partire sempre dalle stesse osservazioni.

Output:

Details of the expected mean squares (EMS) for the model

Source	EMS
si	1*V(Res) + 9,6429*V(sixco) + 38,571*S(si)
co	1*V(Res) + 17*V(co)
pr	1*V(Res) + 1,5556*V(coxpr) + 7,7778*S(pr)
sixpr	1*V(Res) + 3,8889*S(sixpr)
sixco	1*V(Res) + 9,6667*V(sixco)
coxpr	1*V(Res) + 1,7222*V(coxpr)
Res	1*V(Res)

Construction of Pseudo-F ratio(s) from mean squares

Source	Numerator	Denominator	Num.df	Den.df
si	1*si	0,99754*sixco + 2,4631E-3*Res	1	3,01
co	1*co	1*Res	4	26
pr	1*pr	0,90323*coxpr + 9,6774E-2*Res	9	42,7
sixpr	1*sixpr	1*Res	9	26
sixco	1*sixco	1*Res	3	26
coxpr	1*coxpr	1*Res	36	26

Questa parte dell'output riporta i dettagli per il calcolo della media attesa dei quadrati (expected mean square) e il denominatore del test pseudo-F.

Viste le conclusioni fatte precedentemente riguardo le interazioni, possiamo costruire un modello ridotto in cui non consideriamo le interazioni con il fattore corda, guadagnando così gradi di libertà per stabilizzare maggiormente il test pseudo-F. Lasciamo inalterate, rispetto al modello precedente, le opzioni relative al tipo di somma dei quadrati (il terzo tipo), al tipo di permutazioni (permutazioni dei residui sotto il modello ridotto) e al numero di permutazioni eseguite (9999).

Il modello costruito sarà:

$$Y = si + co + pr + si{:}pr + \varepsilon$$

Di seguito viene riportato l'output:

PERMANOVA table of results

Source	df	SS	MS	Pseudo-F	P(perm)	Unique perms
si	1	7581,6	7581,6	38,743	0,0001	9948
co	4	1898	474,51	2,4248	0,001	9883
pr	9	9940,8	1104,5	5,6442	0,0001	9860
sixpr	9	4948,7	549,86	2,8098	0,0001	9874
Res	65	12720	195,69			
Total	88	38694				

Details of the expected mean squares (EMS) for the model

Source	EMS
si	1*V(Res) + 39,31*S(si)
co	1*V(Res) + 17,25*V(co)
pr	1*V(Res) + 8,7415*S(pr)
sixpr	1*V(Res) + 4,3707*S(sixpr)
Res	1*V(Res)

Construction of Pseudo-F ratio(s) from mean squares

Source	Numerator	Denominator	Num.df	Den.df
si	1*si	1*Res	1	65
co	1*co	1*Res	4	65
pr	1*pr	1*Res	9	65
sixpr	1*sixpr	1*Res	9	65

Rispetto al modello precedente, i p-value corrispondenti al sito, alla profondità e all'interazione tra sito e profondità sono diminuiti ulteriormente. Eseguendo nuovamente il test, aumentando l'ordine di grandezza del numero di permutazioni, i p-values tenderanno sempre di più a zero. Sempre rispetto al modello precedente, sono disponibili molti più gradi di libertà al denominatore per il calcolo del test pseudo-F.

I due siti, stando ai risultati del test presentano una evidente differenza per quel che riguarda la distribuzione delle specie. Il risultato viene confermato anche dagli istogrammi presentati.

Bibliografia

Anderson MJ, 2001a. A new method for non-parametric multivariate analysis of variance. Austral ecology 26: 32-46

Anderson MJ, 2001b. Permutation tests for univariate or multivariate analysis of variance and regression. Canadian Journal of Fischeries and Aquatic Sciences 58: 626-639.

Anderson MJ, 2006. Distance-based tests for homogeneity of multivariate dispersions. Biometrics 62: 245-253

Anderson MJ & Legendre P, 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation 62: 271-303

Anderson MJ & Robinson J, 2001. Permutation tests for linear models. Australian & New Zealand Journal of Statistics 43: 75-88

Anderson MJ, Gorley RN & Clarke KR, 2008. PERMANOVA+ for PRIMER: Guide to software and Statistical Method. PRIMER-E Ltd, Plymouth.

Anderson MJ & ter Braak CJF, 2003. Permutation tests for multi-factorial analysis of variance. Journal of Statistical Computation and Simulation 73: 85-113.

Clarke KR, 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology 18: 117-143.

Clarke KR, Somerfield PJ & Chapman MG, 2006c. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. Journal of Experimental Marine Biology and Ecology 330: 55-80.

Codognotto VF, 2010. Dati preliminari sulla biologia di *Flexopecten glaber* nel Nord Adriatico. Tesi di laurea magistrale in biologia marina, Università di Padova.

Faith DP, Minchin PR, & Belbin L. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69: 57-68.

Langsrud Ø, 2003. ANOVA for Unbalanced Data: Use Type II Instead of Type III Sums of Squares, *Statistics and Computing*, 13, 163-167.

Legendre P & Anderson MJ, 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs 69: 1-24.

Legendre P & Legendre L 1998. Numerical ecology, 2nd English edition. Elsevier, Amsterdam.

Manly BFJ, 1997. Randomization, bootstrap and Monte Carlo methods in biology, 2nd edition. Chapman & Hall, London.

Mardia KV & Kent JT & Bibby JM, 1992. Multivariate Analysis. Academic Press, London.

Mc Ardle BH & Anderson MJ, 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82: 290-297.

Searle SR, Casella G & McCulloch CE, 1992. Variance components. John Wiley & Sons, New York.