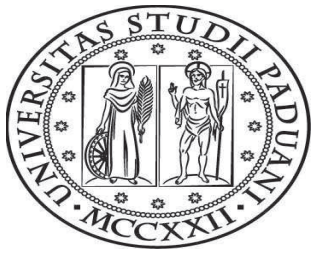
Università degli Studi di Padova

Corso di Laurea in
Statistica e informatica



ALCUNI PROBLEMI DI VERIFICA D'IPOTESI PER VARIABILI CATEGORIALI ORDINATE

Relatore: Prof. Fortunato Pesarin
Dipartimento di Scienze Statistiche

Correlatore: Dott. Livio Corain
Dipartimento di Tecnica e Gestione
dei sistemi industriali

Laureando: Yacouba Yabre
Matricola: 623340

Anno Accademico 2011/2012

Alla mia famiglia.

Indice

1 Introduzione generale. 7

2 Alcuni concetti di base. 9

2.1 Introduzione. 9

2.2 variabili categoriali ordinate. 9

2.3 Variabile aleatorie di Bernoulli e Binomiali 10

3 Metodi non parametrici di permutazione 13

3.1 Introduzione 13

3.1.1 le famiglie di distribuzioni non parametriche 14

3.2 Analisi condizionata in ambito non parametrico 14

3.3 Il principio di scambiabilità dei dati 15

3.3.1 Il principio di permutazione nella verifica d'ipotesi. 16

3.4 I diversi approcci dei test di permutazione 17

3.5 Aspetti computazionali. 18

4 Il test di permutazione 19

4.1 aspetti generali 19

4.2 Approcio univariato. 19

4.2.1 Il metodo Monte Carlo condizionale (C.M.C) 20

4.3 caso multivariato 22

4.3.1 Funzione di combinazione non parametrica 24

4.3.2 Le due fasi dell'algoritmo di combinazione non parametrica . . 26

5 Test di simmetria per variabili categoriali ordinali. 29

5.1 Caso univariato. 30

5.2 Estensione al caso multivariato. 32

5.3 Esempi 33

5.3.1 Caso univariato. 33

5.3.1.1 Analisi esplorativa dei dati

. 34

5.3.2 Caso multivariato 38

6 Test di simmetria per variabili categoriali ordinali con dati appaiati. 41

6.1 caso univariato 42

6.2 caso multivariato 44

6.3 Esempi 46

6.3.1 caso univariato 46

6.3.2 caso multivariato. 49

7 simulazioni 53

8 Conclusione 57

1 Introduzione generale.

Con l'avvento delle nuove tecnologie, la raccolta dei dati è diventata molto più semplice rispetto al recente passato. In particolare, oggi essa non avviene più solo mediante i vecchi mezzi, cioè attraverso interviste, sondaggi, questionari..., che oltre ad essere costosi, sono anche laboriosi da gestire.

Oggi infatti, la produzione dei dati è diventata una cosa scontata. In effetti, è sufficiente una carta fedeltà di un qualche esercizio per avere dei dati abbastanza completi sul consumo di una persona, oppure è addirittura possibile raccogliere tante informazioni su una persona senza che essa se ne renda conto. Questo può essere fatto ad esempio attraverso la navigazione di quest'ultima sul web.

Si capisce quindi che oggi, il problema non è più su come raccogliere i dati, ma bensì su come trarre informazioni da questa mole di dati a disposizione. Attualmente in ambito parametrico, esistono vari tipi di strumenti per analizzare questi dati. Questi strumenti si differenziano in funzione della natura dei dati che abbiamo. Per quanto riguarda l'analisi delle variabili quantitative, quando abbiamo una numerosità campionaria tale da permettere l'uso di questi strumenti, riusciamo a analizzare senza problemi sia dei dati univariati che multivariati. Al contrario, quando si tratta di dati qualitativi, allora le cose si complicano all'aumentare della dimensione del dataset e del numero di variabili coinvolte. Una di queste difficoltà è dovuta al fatto che i metodi parametrici, per l'analisi di queste variabili si basa sulla costruzione di tabelle di contingenza. Queste tabelle si possono tranquillamente analizzare quando abbiamo una sola variabile, usando ad esempio il test chi-quadro. Superata questa soglia la costruzione, e quindi l'analisi, delle tabelle diventa molto difficile.

l'approccio usato in questo lavoro per ovviare a questo inconveniente è usare la cosi detta rappresentazione unit-by-unit, quindi l'analisi non si baserà più sulle tabelle di contingenza, ma direttamente sulle risposte individuali. Questo ci ha permesso di sviluppare diverse tecniche basate sui metodi di permutazione, in particolare l'approccio usato in questo lavoro si basa sulla probabilità di osservazione di ogni unità nelle diversi classi del campione.

La nostra analisi sarà articolata in diversi capitoli, nel secondo capitolo spiegheremo alcuni concetti base della statistica che useremo del prosieguo. Nel terzo capitolo,

vedremmo alcune proprietà importanti riguardanti i metodi non parametrici, e in particolare quelli di permutazione. Nel capitolo successivo vedremmo poi gli strumenti necessari alla costruzione della distribuzione nulla dei test di permutazione. Spiegheremo poi come costruire i diversi test nel capitolo 6. Infine l'ultimo capitolo sarà consacrato all'analisi della potenza di questi test.

2 Alcuni concetti di base.

2.1 Introduzione.

In questo primo capitolo, concentreremo la nostra attenzione sulle definizioni di alcuni concetti di base che useremo in seguito. Il nostro obbiettivo non è la trattazione esaurente di questi temi, ma vogliamo semplicemente fa capire perché si è scelto di usare una approccio invece di un'altro. Le definizioni che daremo in seguito saranno quindi solo sintetiche, esse ci permetteranno di capire sia il concetto che le motivazioni che hanno portato alla loro scelta.

Nel prosieguo, useremo la variabile X per rappresentare sia la variabile casuale univariata(multivariata) di riferimento, sia l'insieme dei dati campionari. Il contesto è in generale sufficiente ad evitare fraintendimenti.

2.2 variabili categoriali ordinate.

Una variabile categoriale è semplicemente una variabile composta da un insieme di categorie. Il rincorso a queste variabili si rende necessario quando non è possibile associare nessuna struttura sensata di quantificazione numerica al fenomeno osservato. L'uso di queste variabili è ormai diffuso in quasi tutti i settori scientifici, in particolare in quelli clinici. Basti pensare, ad esempio, ad un esperimento in cui siamo interessati a valutare l'effetto di un nuovo trattamento farmacologico. Per fare ciò, sottoponiamo ad un gruppo di n individui un questionario in cui essi devono rispondere ad una serie di domande. In particolare a questi individui viene chiesto il sesso, e il rispettivo gradimento rispetto ad una caratteristica del farmaco. Supponiamo che per esprimere il proprio gradimento, ogni persona debba scegliere tra le quattro possibili scelte disponibili: scarso, discreto, buono, ottimo.

Da questo esempio, emerge che le variabili categoriali permettono di partizionare il carattere in esame, ovvero è possibile fare una classificazione delle risposte del fenomeno osservato. Vi sono tuttavia due grandi tipologie di tali variabili.

Un primo gruppo è composto dalle variabili categoriali nominali (o non ordinali). La caratteristica di queste variabili è che non ha alcun senso l'ordinamento del

carattere osservato. Il fatto di sapere che abbiamo un gruppo di maschi e un altro di femmine non ci permette di dire che abbiamo "ordinato" o "misurato" il sesso della popolazione, ci permette solo di dire che siamo riusciti a suddividere in due gruppi i nostri individui, e quindi di aver classificato in due gruppi gli intervistati.

L'altro gruppo è composto dalle variabili categoriali ordinali. La rilevazione delle diverse categorie non costituisce una rilevazione numerica, nel senso che, visto che il carattere rilevato è solitamente qualitativo, non è possibile l'uso degli indici abituali che permettono di fare inferenza, come le medie e in genere i momenti.

Tornando all'esempio precedente, le classi del carattere studiato erano: scarso, discreto, buono e ottimo. Non possiamo, in questo caso, calcolare ad esempio la media aritmetica, ma possiamo dire che $scarso \prec discreto \prec buono \prec ottimo$, ovvero ordinare le diverse categorie. Questa conoscenza ci permette ad esempio di trasformare queste variabili qualitative in variabili rango con cui poi fare diverse analisi.

2.3 Variabile aleatorie di Bernoulli e Binomiali

Supponiamo di eseguire una prova i cui esiti possono essere classificati come successo o insuccesso. Se poniamo:

$$X = \begin{cases} 1 & se\, l'esito\, è\, successo \\ 0 & altrimenti \end{cases}$$

allora la densità discreta di X è data da:

$$\begin{cases} P(0) = P\{X = 0\} = 1 - p \\ P(1) = P\{X = 1\} = p \end{cases}$$

dove $0 \leq p \leq 1$, rappresenta la probabilità che la prova abbia successo.

Una variabile aleatoria X è detta variabile aleatoria di Bernoulli(dal matematico svizzero Bernoulli) se la sua densità discreta è data dalla formula precedente per un qualche valore di p.
Supponiamo ora di eseguire n prove in maniera indipendente, ognuna delle quale abbia come possibili risultati un successo con probabilità p e un insuccesso con probabilità $(1-p)$. se X rappresenta il numero di successi che otteniamo sull'insieme delle n prove, allora X è detta variabile casuale binomiale di parrametri(n, p). La densità discreta di una variabile aleatoria binomiale(n, p) è data da:

$$p(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

Come la maggior parte delle variabili aleatorie, i momenti della binomiale sono noti. In particolare Il valore atteso e la varianza di X sono :

$$E(X) = np$$
$$V(X) = np(1-p)$$

3 Metodi non parametrici di permutazione

3.1 Introduzione

Sotto certe condizioni, la maggior parte dei problemi univariati può essere risolta in modo efficiente usando sia i metodi parametrici che non parametrici . I metodi di permutazione sono essenzialmente metodi non parametrici che sotto certe condizioni permettono inferenze esatte. Il difetto dei dei test non parametrici è che essi a volte lamentano una carenza di efficienza comparativamente a quelli parametrici nelle condizioni di quest'ultimi . Tuttavia, è da sottolineare che in ambito parametrico, ci sono diversi test le cui distribuzioni non sono note, quindi esse vengono approssimate. La relativa mancanza di efficienza dei test non parametrici per certi campioni, può in alcuni casi, essere compensata dalla scarsa approssimazione dei test parametrici. Inoltre le assunzioni su cui si basano i modelli parametrici, assunzioni come l'omoschedasticità o la normalità sono raramente soddisfatti nei casi reali, quindi le inferenze risultanti sono approssimate, e la bontà di queste approssimazioni è difficile da valutare.

In pratica, i modelli parametrici richiedono l'uso di modelli probabilistici che a loro volta richiedono l'introduzione di un insieme di assunzioni che possono essere irrealistiche , ingiustificate, o a volte costruite ad hoc per specifiche inferenze statistiche. Questo fa si che questi metodi siano validi a livello metodologico, ma poco realistici rispetto ai modelli non parametrici che cercano di ridurre al minimo tali assunzioni, il che allarga il loro range di impiego.

In più, ci sono vari problemi multivariati che sono difficili da risolvere in ambito parametrico. Questi modelli risolti in parte in questo lavoro, come ad esempio la trattazione delle variabili ordinali multivariate, le cui soluzioni sono estremamente difficili se non impossibili da trovare se si vuole usare un approccio parametrico.

3.1.1 le famiglie di distribuzioni non parametriche

> **Definizione**: una famiglia di distribuzione $\mathscr{P}$ è detta famiglia non parametrica quando non è possibile trovare uno spazio dimensionale finito Θ, tale per cui esista una relazione uno-a-uno tra Θ e P, nel senso che ogni elemento P di $\mathscr{P}$ non può essere identificato da un unico elemento θ di Θ; e viceversa.

Se questa relazione uno-ad-uno esiste, θ è chiamato parametro e quindi Θ sarà lo spazio parametrico e$\mathscr{P}$ la corrispondente famiglia parametrica.

Le famiglie di distribuzioni le cui caratteristiche non sono ancora specificate o che sono specificate ad eccezione di un numero infinito o ignoto di parametri, non soddisfano la nostra definizione, di conseguenza esse appartengono alle famiglie non parametriche. La definizione include inoltre tutti i casi in cui la dimensione n del campione è inferiore alla dimensionalità di Θ, ossia al numero di parametri.

Notiamo inoltre che il fatto di voler classificare una famiglia $\mathscr{P}$ nella classe di famiglie non parametriche dipende essenzialmente dalle assunzioni che si fanno su di essa. Se assumiamo che essa contenga tutte le distribuzione continue, allora il dataset X è statistica sufficiente e minimale.

I test di permutazione sono test condizionati al dataset osservato, dataset che sappiamo essere sotto H_0 una statistica sufficiente.

Il fatto di fare inferenza condizionatamente ad una statistica sufficiente sotto l'ipotesi nulla, e l'assunzione di scambiabilità (che verrà esplicitata più avanti) dei dati del campione fanno si che i test di permutazione sono test indipendenti dal modello assunto da P, di conseguenza la distribuzione P può essere ignota o non specificata.

3.2 Analisi condizionata in ambito non parametrico

Prima di prosegue cerchiamo di capire cosa si intende per analisi condizionata, concetto molto importante e utile in ambito non parametrico e che verrà più volte citato nel prosieguo.

Rispetto ai metodi parametrici, quelli non parametrici generalmente non fanno rincorso ai sottostanti modelli di distributivi statistici. A questi modelli , come ben sappiamo, sono associati diversi stimatori tra cui quello di massima verosimiglianza $\hat{\theta}$, che è funzione dei dati osservati, ovvero $\hat{\theta} = T(x_1, \cdots, x_n)$. Questo stimatore, qualora la statistica sufficiente sia unidimensionale contiene tutta l'informazione necessaria per fare l'inferenza, dipende esclusivamente dalla particolare scelta di T che

a sua volta non dipende dai dati osservati, ma solo dal particolare modello. Il fatto di aver osservato un dataset invece di un altro non modifica la scelta di T, ad esempio del modello normale. Questo tipo di approccio è detto non condizionato perché la scelta di T dipende essenzialmente dal modello dato. Cioè per non condizionata si deve intendere non condizionata all'esito dell'esperimento, cioè ai dati osservati.

Quindi nelle analisi non condizionate, possiamo dire che i dati giocano un ruolo meramente passivo, rispetto a le analisi condizionate in cui l'informazione principale viene estratta dai dati osservati e non dal modello solitamente assunto ignoto.

3.3 Il principio di scambiabilità dei dati

Sotto H_0, i dati osservati X costituiscono sempre una statistica sufficiente, qualsiasi sia la distribuzione sottostante. Tutte le famiglie non parametriche $\mathscr{P}$ che sono interessate in una analisi di permutazione sono assunte sufficientemente "ricche", nel senso che se X e X' sono punti qualsiasi di χ, allora $X \neq X'$ implica $f_P(X) \neq f_P(X')$ per almeno una $P \epsilon \mathscr{P}$, esclusi i punti di densità nulla. Quando assumiamo che la famiglia sottostante $\mathscr{P}$ contiene tutte le distribuzioni continue, allora l'insieme dei dati X è statistica sufficiente e minimale. Dato un insieme di punti x se $X^* \epsilon \chi^n$ è tale che il rapporto di verosimiglianza $f_P^{(n)}(X)/f_P^{(n)}(X^*) = \varrho(X, X^*)$ non dipende da f_P per qualsiasi $P \epsilon \mathscr{P}$, allora X e X' contengono essenzialmente la stessa informazione rispetto a P, e sono quindi equivalenti. L'insieme dei punti che sono equivalenti ad X, rispetto all'informazione contenuta, costituisce l'orbita associata ad X, e viene indicata con $\chi^n/_X$. Si noti che quando i dati sono raccolti con campionamento casuale e le osservazioni sono i.i.d., così che $f_P^{(n)}(X) = \prod_{1 \leq i \leq n} f_P(X_i)$, l'orbita associata $\chi^n/_X$ contiene tutte le permutazioni di X ed il rapporto di verosimiglianza soddisfa l'equazione $\varrho(X, X^*) = 1 \, \forall \, X^* \epsilon \chi^n/_X$. La stessa conclusione si ottiene se per $f_P^{(n)}(X)$ l'assunzione di indipendenza per i dati osservati è sostituita con l'assunzione di scambiabilità: $f_P^{(n)}(X_1, ..., X_n) = f_P^{(n)}(X_{u_1^*}, ..., X_{u_n^*})$, dove $(u_1^*, ..., u_n^*)$ è una qualsiasi permutazione di $(1, ..., n)$. Nel contesto dei test di permutazione, il concetto di scambiabilità si riferisce alla scambiabilità dei dati rispetto ai gruppi. Le orbite $\chi^n/_X$ sono anche chiamate spazi campionari di permutazione. E importante notare che l'orbita $\chi^n/_X$ associata al campione dei dati X contiene sempre un numero finito di punti, se n è finito. Quindi i test di permutazione sono procedure statistiche condizionate, dove il condizionamento è rispetto all'orbita $\chi^n/_X$ associata all'insieme dei dati osservati, e $\chi^n/_X$ gioca il ruolo di insieme di riferimento per l'inferenza condizionata. In questo modo sotto l'ipotesi nulla ed assumendo la scambiabilità, la distribuzione di probabilità condizionata di un generico punto $X' \epsilon \chi^n/_X$ per qua-

lunque distribuzione sottostante $P \epsilon \mathscr{P}$, è:

$$Pr\{X^* = X^{'} | \chi^n / X\} == \frac{\#[\, X^* = X^{'}, X^* \epsilon \chi^n / X]}{\#[X^* \epsilon \chi^n / X]}$$

e risulta indipendente dalla distribuzione P. Se esiste un solo punto in χ^n/X le cui coordinate coincidono con quelle di $X^{'}$ e non esiste nessun altro vincolo nell'insieme dei dati, e le permutazioni corrispondono alle permutazioni dell'argomento, allora la probabilità condizionata risulta $(1/n)!$. La probabilità $Pr\{X^* = X^{'} | \chi^n / X\}$ risulta quindi uniforme su χ^n/X per ogni $P \epsilon \mathscr{P}$. Per questo l'inferenza di permutazione sotto H_0 risulta invariante rispetto a P, ed i test di permutazione vengono classificati come distribuzione-free e non parametrici. Sotto H_1, invece, la probabilità condizionata mostra delle differenze sostanziali ed in particolare potrebbe dipendere da P. Consideriamo per esempio un problema a due campioni con $X_1 \, e \, X_2$, due insiemi di dati separati ed indipendenti con rispettive numerosità n_1 e n_2 provenienti da distribuzioni P_1 e P_2. La verosimiglianza associata all'intero insieme dei dati è $f_P^{(n)}(X) = f_{P1}^{(n_1)}(X_1) \cdot f_{P2}^{(n_2)}(X_2)$, e per il principio di sufficienza l'insieme dei dati può essere partizionato in due gruppi (X_1, X_2), che forma un gruppo di statistiche sufficienti. L'orbita associata a x risulta $(\chi_{/X_1}^{n_1}, X_{/X_2}^{n_2})$ dove $\chi_{/X_1}^{n_1} \, e \, X_{/X_2}^{n_2}$ sono le orbite parziali associate a X_1 e X_2. Questo implica che i dati X_1 non possono essere scambiati con quelli di X_2 perché sotto H_1 la scambiabilità è permessa solo all'interno dei gruppi e non tra gruppi.

3.3.1 Il principio di permutazione nella verifica d'ipotesi.

> **Definizione**: Se due esperimenti assumono valori nello stesso spazio campionario χ rispettivamente con distribuzione sottostante P_1 e P_2, entrambi appartenenti a $\mathscr{P}$, danno lo stesso insieme di dati X ed a condizione che la scambiabilità tra i gruppi risulti soddisfatta sotto l'ipotesi nulla, allora le due inferenze condizionate a X ottenute usando la stessa statistica test devono essere uguali. Se due esperimenti, con distribuzione sottostante P_1 e P_2, forniscono rispettivamente X_1e X_2, e $X_1 \neq X_2$ allora le due inferenze condizionate potrebbero essere diverse.

Dall'analisi della definizione precedente emerge che, una condizione sufficiente per poter applicare il test di permutazione è che la scambiabilità dei dati sotto H_0 sia soddisfatta. Quindi per quanto segue, quando si farà riferimento ad un test di permutazione, si assumerà implicitamente che la condizione di scambiabilità sia soddisfatta sotto l'ipotesi nulla.

Una delle caratteristiche più importanti del principio di permutazione è che in teoria e sotto un insieme di certe condizioni, le inferenze condizionate possono essere estese in maniera incondizionata a tutte le distribuzioni P di $\mathscr{P}$ per le quali la densità rispetto ad un adeguata misura ξ dominante sia positiva ad esempio ${}^{dP(x)}/_{d\xi^n} > 0$. Va sottolineato, tuttavia, che questa caratteristica deriva dai principi di sufficienza e condizionalità dell'inferenza, principio che ci permette di estendere l'inferenza a tutte le popolazioni che condividono lo stesso valore della statistica condizionata. Per esempio l'inferenza basata sulla t di student è estesa a tutte le distribuzioni normali con varianza stimata $\hat{\sigma}^2$ positiva, e quindi l'inferenza è per tutta la famiglia di distribuzioni, anche se tali estensioni incondizionate devono essere effettuate con cautela.

Un'altro aspetto importante riguarda uno dei problemi affrontati in questo lavoro, ovvero i problemi multivariati la cui risoluzione richiede i metodi di combinazione non parametrica (NPC). Per questi tipi di problemi sopratutto quando sono complessi, non è necessaria la specificazione o la modellazione della struttura di dipendenza delle variabili nella popolazione, come si farebbe ad esempio in ambito parametrico. Infatti, se si considera il caso delle variabili categoriali multivariate, è estremamente difficile trovare una soluzione valida per la risoluzione dei relativi test in ambito parametrico mentre usando l'approccio appena citato le analisi sono fattibili e i risultati risultano agevoli da interpretare.

Tuttavia il fatto di condizionare i test di permutazione alla statistica sufficiente fanno si che questi test test godano di buone proprietà. In più quando la condizione di scambiabilità dei dati sotto H_0 è rispettata queste procedure danno luogo ad inferenze esatte.

3.4 I diversi approcci dei test di permutazione

Esistono tre grandi approcci per la costruzione dei test di permutazione. Il primo approccio è un approccio essenzialmente euristico, mentre i due altri sono più formali.

L'approccio euristico è quello generalmente usato quando abbiamo dei problemi abbastanza semplici. Esso è basato su un ragionamento intuitivo, e quindi è sufficiente l'uso del buonsenso per risolvere questi problemi, in particolare, fornisce soluzioni molto chiare e semplici da interpretare.

I due approcci formali sono sostanzialmente equivalenti, essi sono più eleganti, effettivi e precisi. Uno dei due è basato sul concetto di invarianza della distribuzione nulla sotto l'effetto di un gruppo G di trasformazioni, mentre l'altro approccio si basa

sul fatto di condizionarsi ad un set di statistiche sufficienti sotto H_0 quando vogliamo fare inferenza su una distribuzione sconosciuta. Questo ultimo risulta interessante perché, oltre a permettere di stabilire se una soluzione è approssimata o esatta, è anche costruttivo.

3.5 Aspetti computazionali.

Uno dei problemi maggiori associati ai test di permutazione è il fatto che è difficile esprimere in forma chiusa le loro distribuzione nulle, specialmente nei casi multivariati. Questo è dovuto al fatto che queste distribuzioni dipendono dallo specifico dataset e quindi variano in funzione di esso nell'insieme dello spazio campionario χ^n. In più, quando la numerosità campionaria non è piccola, il calcolo diretto che consiste nell'enumerazione di tutte le possibili permutazioni è quasi impossibile perché la cardinalità dello spazio campionario di permutazione diventa proibitivamente grande e quindi difficile da gestire anche per i calcolatori di ultima generazione.

Tuttavia degli algoritmi esatti, non basati sull'enumerazione completa di tutte le possibili permutazioni sono stati implementati per diversi casi univariati permettendoci di calcolare usando appropriate tecniche le relative distribuzioni multivariate. In particolare il metodo Monte Carlo condizionato è un metodo basato sul ricampionamento senza reinserimento del dataset osservato. Questo metodo ci permette di avere una stima statistica della distribuzione desiderata di permutazione. La precisione di questa distribuzione dipende esclusivamente dal numero di repliche che vengono fatte. Si capisce quindi che per avere una buona stima di questa distribuzione dobbiamo quindi avere dei computer in grado non solo di immagazzinare tutta la quantità di dati ma anche di fare le permutazioni in un tempo ragionevole.

La disponibilità oggi di computer relativamente rapidi, poco costosi ed efficienti fa si che i metodi di permutazione siano diventati molto accessibili e competitivi rispetto sia ai metodi non parametrici basati sui ranghi sia a quelli parametrici. Infatti la loro distribuzione nulla può essere facilmente approssimata senza tuttavia compromettere le proprietà statistiche desiderate. Questo ha fatto si che oggi esistano numerosi test non parametrici e l'uso di essi è diventato più comune.

Vogliamo infine fa notare che lo strumento statistico usato in questa tesi per fare le analisi è il software statistico R.

4 Il test di permutazione

4.1 aspetti generali

Come si è detto in precedenza, i metodi di permutazione sono applicati all'intero insieme di dati osservati. Questo insieme di dati, sotto H_0, costituisce una statistica sufficiente per la rispettiva distribuzione non degenere P.

Consideriamo quindi che la nostra distribuzione P sia sconosciuta e che la relativa famiglia non parametrica $\mathscr{P}$ contenga solo distribuzioni non degeneri comprese le distribuzioni discrete, continue e miste.

La descrizione delle diverse tecniche di permutazione avverrà in due passi. In un primo tempo descriveremo come sono generalmente costruiti i sistemi di verifica di ipotesi, e come procedere per il calcolo del livello di significatività, e in un secondo momento, ci occuperemo dell'estensione del test al caso multivariato.

4.2 Approcio univariato.

L'analisi dei problemi unidimensionale è molto diffusa negli esperimenti reali, basta ad esempio pensare agli studi sperimentali casi/controllo in cui siamo solitamente interessati a sapere se statisticamente, la popolazione dei casi, ovvero la popolazione a cui è stata somministrato il trattamento è stocasticamente dominata da quelli dei controlli, cioè quelli che hanno ricevuto il placebo.

Il nostro campione sarà quindi formato da due campioni rilevati su un insieme di n soggetti. Questi soggetti vengono puoi classificati secondo i due livelli del trattamento. Consideriamo che queste variabili siano variabili ordinali, pertanto, dati due campioni indipendenti di osservazioni $X_j = \{X_{jr},\, r = 1, \cdots, n_j\}$, $j = 1, 2$, si voglia saggiare l'ipotesi nulla:

$$\begin{cases} H_0 : \{X_1 \overset{d}{=} X_2\} \equiv \{F_1 = F_2\} \equiv \{\cap_c [F_1(A_c) = F_2(A_c)],\, c = 1, \cdots, C-1\} \\ H_1 : \{X_1 \overset{d}{>} X_2\} \equiv \{F_1 \leq F_2\} \equiv \{\bigcup_c [F_1(A_c) < F_2(A_c)]\} \end{cases}$$

in cui la funzione $F_j(A_c) = Pr\{X_j \leq A_c\}$, $j = 1, 2$, gioca il ruolo di funzione di ripartizione (cumulativa) della variabile categorialeX_j. Si noti che H_1 definisce la dominanza stocastica di X_1 rispetto a X_2, in quantoX_1 manifesta "valori" tendenzialmente più grandi di quelli manifestati da X_2. Inoltre, è importante osservare che H_1 nella forma $\{\bigcup_c [F_1(Ac) < F_2(Ac)]\}$ definisce un insieme di alternative cosiddette ristrette, ossia composto da C-1 sotto-alternative unilaterali o direzionali. In questa forma il problema si presenta come notoriamente difficile si vedano ad esempio i lavori di Hirotsu (1982, 1986, 2002), Sampson and Whitaker (1989), Dykstra et al. (1995), El Barmi and Dykstra (1995),Wang (1996), Dardanoni and Forcina (1998), Cohen et al. (2000, 2003), Perlman and Wu (2002), Silvapulle and Sen (2005)).
Si noti anche che H_0 afferma che i dati dei due gruppi sono tra loro scambiabili, e che quindi per la relativa analisi risulta applicabile il principio di permutazione. L'insieme costituito dalla riunione di tutti i dati (insieme ottenuto dal concatenamento delle due liste) $X = \{X_{jr},\, r = 1, ..., n_j,\, j = 1, 2\}$ costituisce una statistica sufficiente sotto H_0.

Il primo passo da compiere per poter risolvere il problema della verifica d'ipotesi consiste ovviamente nella costruzione di un test adeguato. Una volta costruita la statistica, dobbiamo decidere se accettare o rifiutare l'ipotesi nulla utilizzando la distribuzione nulla di permutazione della statistica test che ci permetterà di calcolare il livello di significatività del test.

Nel caso in cui abbiamo una piccola numerosità campionaria possiamo calcolare il p.value esatto enumerando tutte le possibili permutazioni, altrimenti bisogna ricorrere ad altri metodi che ci permettono allo stesso tempo di non esplorare tutto lo spazio di permutazioni e di avere risultati accettabili.

Uno dei metodi più utilizzato a questo proposito è il metodo Monte Carlo Condizionale (CMC). Questa tecnica consiste nel selezionare casualmente un campione di dimensione B (B abbastanza grande), tra tutte le possibili permutazioni, e applicare il test sul campione invece di considerare l'intero spazio χ^n/χ. Molti autori hanno notato (Ernst 2004, Good 1994, Manly 1997) che una B pari a qualche migliaia è sufficiente per avere un stima accettabilmente precisa della probabilità critica (p-value) esatta.

4.2.1 Il metodo Monte Carlo condizionale (C.M.C)

Innanzitutto, notiamo che sotto l'ipotesi nulla, e con l'assunto di scambiabilità, tutti i punti dello spazio χ/χ hanno lo stesso grado di interesse. Come abbiamo fatto notare precedentemente, la trattazione matematica in forma chiusa della distribu-

zione di permutazione k-variata è notoriamente impossibile, sia generalmente per la elevata cardinalità dell'insieme delle possibili permutazioni ma soprattutto per la difficoltà nella caratterizzazione analitica della regione critica del test che si voglia adottare. È invece possibile e computazionalmente appropriata l'applicazione di un metodo di combinazione non parametrica dei test parziali, cui resta associata una regione critica più facilmente caratterizzabile ed il cui valore-p globale (nel caso di confronto di più test) viene di fatto stimato tramite un procedimento di simulazione Monte Carlo condizionale consistente nel considerare un campione casuale semplice di B elementi dall'insieme di tutte le possibili permutazioni dei dati.

Senza perdita di generalità, si assume che la statistica test T univariata sia significativa per valori grandi. La distribuzione di permutazione di T in corrispondenza del dataset X è indicata con $F_T(z|X)\,, \forall\, z\ \epsilon R^1$.

La procedura C.M.C per la stima della funzione di distribuzione cumulativa e del p-value associato λ indotto dalla statistica T è decritto nei seguenti passagi:

1. Calcolare la statistica test T sui dati osservati X: $T_0 = T(X)$, T_0 rappresenta il valore osservato del test.

2. considerare una permutazione $(u_1^*, \cdots, u_n^*)$ di $1, \cdots, n$ determinando la corrispondente permutazione $X^* = \{X(u^*),\, i = 1, \cdots, n\}$ di X e calcolare il corrispondente test di permutazione $T^* = T(X^*)$.

3. Ripetere indipendentemente B volte il punto precedente. L'insieme dei risultati C.M.C T^* ossia $\{T^*,\, r = 1, \ldots, B\}$ è il campione casuale della distribuzione nulla di permutazione univariata di T.

4. La stima non distorta e consistente di $F_T(z|X)$ di T dato$X$, è data dalla funzione di ripartizione empirica:

$$\widehat{F}_B^*(z) = \sum_{r=1}^{B} \frac{I(T_r^* \leq z)}{B}$$

dove$I(.) = 1$ se la relazione è soddisfatta e 0 altrimenti. Per il teorema di Glivenko-Cantelli, $\widehat{F}_B^*(z)$ fornisce una stima fortemente consistente della distribuzione $F_T(z|X)$ di T. Inoltre, la corrispondente funzione del livello di significatività, ovvero la percentuale dei valori peggiori o al più uguali a t è : $L(t|X) = \#(T^* \geq t)/B$ in cui $\#$ rappresenta il numero di punti(eventi) che soddisfano (.).

poiché il valore osservato T_0, se fosse vera l'ipotesi nulla, sarebbe uniformamene distribuito sul insieme dei valori di permutazione (il supporto condizionato), il p.value del test T è dato da:

$$\widehat{\lambda} = L(T_0|X) = \sum_{r=1}^{B} \frac{I(T_r^* \geq T_0)}{B}$$

La tabella seguente riassume la procedura C.M.C

X	X_1^*	$\cdots$	X_r^*	$\cdots$	X_B^*
T_0	T_1^*	$\cdots$	T_r^*	$\cdots$	T_B^*

Tabella 4.1: Metodo C.M.C

4.3 caso multivariato

Uno dei punti di forza del test di permutazione risiede della sua semplicità e facilità a trattare i casi multivariati. Infatti, una volta implementata la versione univariata del test che ci interessa, l'estensione al caso multivariato è "quasi" immediata. Quasi perché a volte è sufficiente, qualche piccola modifica per adattare la procedura che verrà esplicitata in questa parte al proprio problema. Quindi la procedura che descriveremo in questa parte costituisce il punto di partenza per la costruzione di diversi test adatti a risolvere i problemi legati alle variabili multivariate. In particolare, e come vedremmo più avanti, questa tecnica è stata leggermente modificata per permetterci di estendere i nostri test univariato ai casi multivariati.
Le assunzioni principali riguardanti la struttura dei dati, l'insieme dei test parziali e le ipotesi d'interesse per i test, nel contesto della combinazione non parametrica, possono essere così schematizzate:

1. L'insieme q-dimensionale dei dati, o il vettore q-variato delle risposte è indicato con

$$\begin{aligned} X &= \{X_j,\, j = 1, ..., C\} = \{X_{ji},\, i = 1, ..., nj,\, j = 1, ..., C\} \\ &= \{X_{hji},\, i = 1, ..., nj,\, j = 1, ..., C,\, h = 1, ..., q\} \end{aligned}$$

   Il vettore q-variato delle risposte X è definito dal modello statistico(X, χ, β, P), dove χ è lo spazio campionario, β è una $\sigma-$algebra e P è una distribuzione di

probabilità di solito non specificata proveniente da una famiglia $\mathscr{P}$ di distribuzioni non degeneri.
L'insieme dei dati X è costituito da $C \geq 2$ campioni o gruppi di ampiezza $n_j \geq 2$, con $n = \sum_j n_j$; i gruppi sono rappresentativi di C livelli di un trattamento e i dati X_j sono supposti i.i.d. con distribuzione $P_j \epsilon \mathscr{P}$, $j = 1, ..., C$. Uno dei passi che ha permesso la trattazione delle variabile discrete è la rappresentazione unità per unità in luogo di quella tabellare. L'insieme X si può quindi riscrivere con $X = \{X(i),\, i = 1, ..., n;\, n_1, ..., n_C\}$, in cui si assume che i primi n_1 vettori di dati provengano dal primo campione, i successivi n_2 dal secondo, e così via.

2. Sotto l'ipotesi nulla le distribuzioni multivariate delle risposte sono uguali nei C gruppi e i dati sono scambiabili tra i C campioni. Il sistema d'ipotesi è:

$$\begin{cases} H_0 : \{P1 = \cdots = Pc\} = \{X_1 \stackrel{d}{=} \cdots \stackrel{d}{=} X_c\} = \cap_c H_{0i} \\ H_1 : \{almeno\, una\, H_{0i}\ è\ falsa\} \end{cases}$$

Supponiamo che si vogliano analizzare e verificare più aspetti di un problema, così che l'ipotesi nulla H_0 può essere scomposta in un insieme finito di sotto-ipotesi H_{0i}, $i = 1, \cdots, k$, ciascuna adatta all'aspetto parziale di interesse. E' da notare che l'ipotesi nulla è vera se tutte le sotto-ipotesi nulle sono vere, ossia se tutte le H_{0i} sono congiuntamente vere, quindi è sufficiente che una sola H_{0i} non sia vera per dire che l'ipotesi alternativa è vera.

3. con $T = T(X)$ si indica il vettore k-dimensionale di statistiche test, le cui componenti rappresentano i test univariati non degeneri di primo ordine ciascuno idoneo alla verifica delle sub-ipotesi $H_{0i}\, vs\, H_{1i}$.

   Le assunzioni riguardanti l'insieme di test parziali $T = \{T_i, i = 1, ..., k\}$ sufficienti per la combinazione non parametrica sono:

- Tutti i test parziali T_i sono permutazionalmente esatti, ovvero le variazioni delle statistiche sotto H_0 dipendono solo dalla casualità delle permutazioni, e sono marginalmente corretti, sono stocasticamente significativi per valori grandi, vale a dire che sotto H_1 le loro distribuzioni sono stocasticamente più grandi rispetto ad H_0.

- Almeno uno dei test di permutazione T_i è consistente, cioè:

$$Pr\{T_i > T_{i\alpha}|H_{1i}\} \geq \alpha,\ \forall\, \alpha > 0,\ i = 1, ..., k$$

e al tendere di n all'infinito si ha che:

$$\lim_{n\to\infty} Pr\{T_i > T_{i\alpha}|H_{1i}\} \Rrightarrow 1$$

per almeno una $i \epsilon \{1, \cdots, k\}$, dove n è la numerosità campionaria e $T_{i\alpha}$ è il valore critico, assunto finito, di T_i al livello α.

4.3.1 Funzione di combinazione non parametrica

La combinazione non parametrica è il mezzo che ci permette di avere il livello di significatività globale del nostro test. Esso prende congiuntamente in considerazione i p-value λ_i di permutazione associati alle statistiche test T_i, $i = 1, ..., k$.
Il test combinato di secondo ordine $T'' = \psi(\lambda_1, \cdots, \lambda_k)$ si ottiene tramite una funzione $\psi \div (0,1)^k \rightarrow R^1$ continua, non crescente, univariata, non degenere, reale (misurabile) che soddisfi alle seguenti proprietà:

i) ψ deve essere non crescente in ogni suo argomento, ovvero $\psi(.., \lambda_i, ..) \geq \psi(.., \lambda_i^{'}, ..)$ se $\lambda_i < \lambda_i^{'}$, $i\, \epsilon (1, ..., k)$

ii) ψ assume il suo valore supremo $\widehat{\psi}$ che potrebbe non essere finito, quando almeno uno degli argomenti raggiunge lo 0, cioè $\psi(.., \lambda_i, ..) \rightarrow \psi se\, \lambda_i \rightarrow 0$, per almeno un $i \epsilon (1, \cdots, k)$

iii) il suo valore critico $T_\alpha^{''}$ è finito e tale che $T_\alpha^{''} \leq \hat{\psi}$, $\forall\ \alpha > 0$.

Queste proprietà definiscono una classe $\mathcal{C}$ di funzioni di combinazione. Le funzioni di combinazione ψ più utilizzate sono:

1. La funzione di combinazione di Fisher

$$T_F^{''} = -2 \cdot \Sigma_i log(\lambda_i)$$

Se tutti i k test parziali sono indipendenti e continui, sotto l'ipotesi nulla $T_F^{''}$ si distribuisce come un χ^2 centrale con $2k$ gradi di libertà

2. La funzione di combinazione di Liptak

$$T_L^{''} = \Sigma_i \Phi^{-1}(1 - \lambda_i)$$

dove Φ è la funzione di ripartizione di una normale standard. Se tutti i k test parziali sono indipendenti e continui, sotto l'ipotesi nulla $T_L^{''}$ ha distribuzione normale, con media 0 e varianza k. Una versione della funzione di combinazione di Liptak considera la trasformazione logistica dei p-values:

3. La funzione di combinazione di Tippett

$$T_T^{''} = \underset{1\leq i\leq k}{max}(1-\lambda_i)$$

la cui distribuzione sotto l'ipotesi nulla, se tutti i k test parziali sono continui e indipendenti, si comporta come il più grande di k valori casuali estratti da una variabile uniforme nell'intervallo aperto (0, 1)

4. La funzione di combinazione di Lankaster

$$T_G^{''} = \Sigma_i \Gamma_{r,a}^{-1}(1-\lambda_i)$$

dove $\Gamma_{r,a}^{-1}$ è l'inversa della funzione di ripartizione di una variabile casuale gamma centrale con parametro di scala noto a e r gradi di libertà. Se tutti i k test parziali sono indipendenti, sotto l'ipotesi nulla, $T_G^{''}$ ha distribuzione gamma centrale con parametro di scala a e rk gradi di libertà.

5. Un' interessante sottoclasse di $\mathcal{C}$ è costituita dall'insieme di funzioni CD di combinazione non parametrica diretta, espressa nella forma

$$T_D^{''} = \Sigma_i T_i$$

che si può utilizzare quando tutte le statistiche test parziali sono omogenee, per cui condividono la stessa distribuzione di permutazione asintotica (per esempio, si distribuiscono come normali standard o sono del tipo χ^2 con gli stessi gradi di libertà) e il loro comune supporto asintotico è almeno non limitato a destra. Le distribuzioni di permutazione di tutti i test parziali possono essere le stesse solo per numerosità campionarie abbastanza elevate, quindi per numerosità campionarie finite, questa condizione può essere soddisfatta solo approssimativamente poiché le distribuzioni di permutazione sono essenzialmente dipendenti dai dati osservati. La funzione di combinazione diretta consente di evitare i calcoli abbastanza intensivi dell'algoritmo ed inoltre, anche se la funzione diretta opera come nel caso univariato, è sostanzialmente una combinazione non parametrica perché il problema di verifica viene spezzato in k sotto problemi, una statistica test globale $T \div R^k \rightarrow R^1$ non è disponibile

e le relazioni di dipendenza tra i test parziali sono implicitamente catturate dalla procedura di combinazione.

4.3.2 Le due fasi dell'algoritmo di combinazione non parametrica

La procedura che consente di ottenere una stima del livello di significatività globale, tramite CMC, della distribuzione di permutazione dei test combinati avviene in due fasi: la prima relativa alla stima della distribuzione k-variata di T, la seconda riguarda la stima della distribuzione di permutazione del test combinato $T^{"}_{\psi}$ e utilizza i risultati ottenuti via C.M.C nella prima fase. In ambito multivariato dunque, l'algoritmo di simulazione per la stima della distribuzione k-variata di T può essere descritto nelle seguenti fasi:

1. si calcola il valore osservato di T : $T_0 = T(X)$

$$\left[\begin{array}{ccc||ccc} X_1(1) & \cdots & X_1(n_1) & X_1(1+n_1) & \cdots & X_1(n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_q(1) & \cdots & X_q(n_1) & X_q(1+n_1) & \cdots & X_q(n) \end{array}\right] \longrightarrow \left[\begin{array}{c} T_{10} \\ \vdots \\ T_{k0} \end{array}\right]$$

2. si considera un componente g^*, casualmente rilevato da un appropriato gruppo di trasformazioni G, e i valori del vettore$T^* = T(X^*)$ dove $X^* = g^*(X)$. Una permutazione X^* del dataset. Si può anche ottenere considerando una permutazione $(u_1^*, u_2^*, , ..., u_n^*,)$ delle etichette $(1, 2, ..., n)$, ciascuna indicante una unità del campione di partenza, e assegnando poi il vettore di dati con le etichette permutate al gruppo appropriato, per esempio nel caso di due campioni di numerosità n_1 e n_2 tali che $n_1 + n_2 = n$ si avrà che $X(u_i^*)$ viene assegnato al primo gruppo se l'etichetta i soddisfa la condizione $1 \leq i \leq n_1$, altrimenti al secondo gruppo

$$\left[\begin{array}{ccc||ccc} X_1(u_1^*) & \cdots & X_1(u_{n1}^*) & X_1(u_{1+n_1}^*) & \cdots & X_1(u_n^*) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_q(u_1^*) & \cdots & X_q(u_{n1}^*) & X_q(u_{1+n_1}^*) & \cdots & X_q(u_n^*) \end{array}\right] \longrightarrow \left[\begin{array}{c} T_1^* \\ \vdots \\ T_k^* \end{array}\right]$$

3. si ripete B volte la fase descritta in 2 l'insieme dei risultati del CMC $\{T_r^*, r = 1, ..., B\}$ è un campione casuale dalla distribuzione nulla k-variata di T

X	X_1^*	$\cdots$	X_r^*	$\cdots$	X_B^*
T_{10}	T_{11}^*	$\cdots$	T_{1r}^*	$\cdots$	T_{1B}^*
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
T_{k0}	T_{k1}^*	$\cdots$	T_{kr}^*	$\cdots$	T_{kB}^*

4. la funzione di distribuzione empirica simulata k-variata (EDF) è cosi definita

$$\widehat{F_B}(z|X) = [\frac{1}{2} + \sum_r I(T_r^* \leq z)]/(B+1),\ \forall z \epsilon R^k,$$

con I(·) che vale 1 se la relazione è soddisfatta 0 altrimenti, fornisce una stima della distribuzione di permutazione k-dimensionale$F(z|X)$ di T. Si ha poi che

$$\widehat{L}_i(z|X) = \left[\frac{1}{2} + \sum_r I(T_r^* \geq z)\right]/(B+1),\ i = 1, ..., k$$

dà una stima $\forall\, z\, \epsilon\, R^1$ della funzione di permutazione marginale del livello di significatività $L_i(z|X) = Pr(T_i^* \geq z|X)$; così che $\widehat{L}_i(T_{i0}|X) = \lambda_i$ fornisce una stima non distorta e consistente del vero p-value marginale $\lambda_i = Pr(T_i^* \geq T_{i0}|X)$ relativa al test $T_i, \forall i = 1, ..., k$. E' da notare che rispetto agli stimatori EDF standard, 1/2 e 1 sono rispettivamente aggiunti al numeratore e al denominatore allo scopo di ottenere un valore stimato nell'intervallo aperto (0, 1) in modo che le trasformazioni inverse di funzioni continue, come $-log(\lambda)$ o $\Phi^{-1}(1-\lambda)$, siano continue e sempre definite. Dato che B è molto grande queste quantità aggiunte sono alterazioni praticamente irrilevanti che non influenzano il comportamento degli stimatori sia per campioni di numerosità finita che asintoticamente. Al posto di 1/2 e 1 si può dunque mettere una qualsiasi quantità positiva ε e 2ε, assegnando ad ε un valore arbitrario in (0,1).

I prossimi passi costituiscono la seconda fase dell'algoritmo, essi ci permetteranno di fare la combinazione vera e propria dei test calcolati calcolati nella prima parte. Questi passi sono:

1. i k p-value $\widehat{\lambda}_i = \widehat{L}_i(T_{i0}|X)$ osservati sono stimati sui dati X dove $T_{i0} = T_i(X)$, $i = 1, ..., k$ rappresenta i valori osservati dei test parziali e $\widehat{L}_i$ è la i-esima funzione marginale del livello di significatività ottenuta tramite ricampionamento C.M.C sull'insieme dei dati nell'ultimo passo dell'algoritmo della prima fase.

2. il valore osservato combinato dei test di secondo ordine utilizza i risultati ottenuti nella prima fase con CMC ed è dato da $T_0'' = \psi(\widehat{\lambda}_1, ..., \widehat{\lambda}_k)$

3. l'r-esimo vettore statistico simulato combinato è calcolato partendo da $T_r^{''} = \psi(\widehat{\lambda^*}_{1r}, ..., \widehat{\lambda^*}_{kr})$, dove $\lambda_{ir}^* = \widehat{L}_i(T_{ir}^*|X)$, $i = 1, ..., k$, $r = 1, ..., B$

4. p-value globale è dato da $\widehat{\lambda_\psi^{''}} = \sum_r I(T_r^{''} \geq T_0^{''})/B$

5. si rigetta l'ipotesi nulla globale a livello di significatività fissato e pari ad α se $\widehat{\lambda}_\psi^{''} \leq \alpha$.

Le fasi appena descritte possono essere riassunte nella tabella seguente:

T_{10}	T_{11}^*	$\cdots$	T_{1r}^*	$\cdots$	T_{1B}^*
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
T_{k0}	T_{k1}^*	$\cdots$	T_{kr}^*	$\cdots$	T_{kB}^*
	$\Downarrow$				
$\widehat{\lambda_1}$	λ_{11}^*	$\cdots$	λ_{1r}^*	$\cdots$	λ_{1B}^*
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$
$\widehat{\lambda_k}$	λ_{k1}^*	$\cdots$	λ_{kr}^*	$\cdots$	λ_{kB}^*
	$\Downarrow$				
$T_0^{''}$	$T_1^{''*}$	$\cdots$	$T_r^{''*}$	$\cdots$	$T_B^{''*}$

Tabella 4.2: procedura C.M.C

Il prossimo capitolo sarà dedicato allo studio di diversi problemi legati alle variabili categoriali ordinate.

5 Test di simmetria per variabili categoriali ordinali.

Come con buona parte dei test non parametrici, ci occuperemo prima dell'applicazione del test al caso univariato, poi estenderemo la nostra analisi al caso multivariato attraverso una combinazione specifica che avremmo cura di spiegare nel dettaglio più avanti.

Questo test può essere usato per la verifica d'ipotesi in diversi casi. Si pensi, ad esempio, al caso clinico menzionato precedentemente in cui eravamo interessati a confrontare i casi con i controlli; oppure, per non limitarsi ai soli casi clinici, è possibile pensare ad altri studi come ad esempio la valutazione dei diversi corsi che si fanno nelle diverse università italiane. Possiamo quindi essere interessati a vedere se il corso è migliorato rispetto agli anni precedenti, oppure se gli studenti hanno gradito il fatto di aver cambiato il docente che erogava un dato corso sottoponendogli un questionario.

Il test può anche essere usato per confrontare i servizi offerti da due strutture diverse, a questo proposito, consideriamo uno studio in cui l'obbiettivo è la valutazione dei servizi offerti dal proprio comune rispetto ad altri comuni. Per fare ciò, è sufficiente raccogliere il parere degli utenti che usufruiscono di questi servizi in entrambe le strutture e applicare la nostra statistica.

Insomma si capisce che il range di applicazione di questa statistica è molto ampio, e il fatto di poterlo applicare sia ai casi univariati che multivariati la rende particolarmente importante.
Il modo più semplice per spiegarne la costruzione e come vengono applicati i diversi test è partire da degli esempi. Nei prossimi paragrafi spiegheremo come sono stati costruiti i diversi test partendo da semplici esempi. Inoltre, abbiamo riportato alcuni passaggi delle procedure già menzionate prima come quelle del C.M.C, perché esse differiscono leggermente dal quelle standard.

5.1 Caso univariato.

Supponiamo di voler verificare l'effetto di un farmaco su un gruppo di n persone. per fare ciò, somministriamo il farmaco a degli individui in due momenti diversi che chiameremo prima e dopo.
Una possibile rappresentazione tabellare dei nostri dati è la seguente:

Prima	Dopo
Y_1	Y_2
y_{11}	y_{21}
$\vdots$	$\vdots$
y_{1n}	y_{2n}

Tabella 5.1: rappresentazione tabellare dei dati nel caso univariato

L'obbiettivo dello studio è di cercare di capire se il farmaco è migliore del placebo. Quindi nella prima occasione, solitamente si somministra il placebo ai pazienti e poi in un secondo momento il farmaco vero e proprio. Siamo quindi interessati a sapere se c'è stato un miglioramento nei nostri pazienti, ovvero l'effetto del trattamento nella seconda occasione è migliore di quella della prima. Vogliamo quindi risolvere il sistema di verifica d'ipotesi seguente:

$$I : \begin{cases} H_0 : Y_1 \stackrel{d}{=} Y_2 \\ H_1 : Y_1 \stackrel{d}{<} Y_2. \end{cases}$$

In cui viene enfatizzata la condizione di scambiabilità sotto H_0. Visto che le nostre due variabili sono variabili categoriali ordinali definite sul comune supporto $(A_1 \prec A_2 \prec \cdots \prec A_k)$, dove con $A_i \prec A_j$ si intende che la j-esima classe è superiore all'i-esima classe per ogni indice $i < j$; possiamo quindi considerare le differenze per singola unità (individuo o soggetto) $u = 1, \cdots, n$. Sia quindi X_u la differenza tra Y_{1u} e Y_{2u}, ovvero: $X_u = \varphi(Y_{1u}; Y_{2u})$.
Sotto H_0 La variabile casuale X; per effetto della scambiabilità, sarà quindi una variabile simmetrica. Infatti, sotto l'ipotesi nulla, le variabili Y_{1u} e Y_{2u} sono scambiabili entro le unità e questo implica che:

$$F_1(t) = F_2(t), \forall\, t \epsilon R^1$$

e

$$F_{1|t}(y|Y_{1u} = t) = F_{2|t}(y|Y_{2u} = t),\ \forall\, t\, \epsilon R^2$$

in cui F_1, F_2, $F_{1|t}\, e\, F_{2|t}$ sono rispettivamente le funzioni di ripartizione delle variabili Y_{1u} ,Y_{2u}, $(Y_{1u}|Y_{2u} = t)$ e $(Y_{2u}|Y_{1u} = t)$. Queste funzioni sono note e legate a P, in particolare:

$$Pr\{(Y_{1u} - Y_{2u} \leq y) = \int_{-\infty}^{+\infty} F_{1|t}(y + t|Y_{2u} = t).dF_{2u}(t)\}$$

e

$$Pr\{(Y_{2u} - Y_{1u} \leq y) = \int_{-\infty}^{+\infty} F_{2|t}(y + t|Y_{2u} = t).dF_{2u}(t)\}$$

quindi si ha: $Pr\{X > y\} = Pr\{X < -y\}, \forall\, y\epsilon R^1$, che è la condizione di simmetria di X intorno allo 0.

Una delle conseguenze di questa proprietà è che sotto H_0, $Pr\{X > 0\} = Pr\{X < 0\}$, quindi assumendo che $E(Y)$ esista e sia finito, la variabile casuale X ha media nulla $(E(X) = 0)$, inoltre la mediana di X è anch'essa nulla se la mediana di Y è unica. Invece, sotto H_1, $Pr\{X < 0\} > (<) Pr\{X > 0\}$, a seconda che: $Y_{1u} \overset{d}{>} (\overset{d}{<} Y_2)$. Un'ulteriore conseguenza è: Il vettore dei segni $\{X/|X_i|,\ i = 1, ..., n\}$ è stocasticamente indipendente da quello delle differenze $\{X_i = Y_{1u} - Y_{2u},\ u = 1, ..., n\}$ (da sottolineare che se le $X_i = 0$ allora le X_i e i relativi segni $X_i/|X_i|$vengono esclusi dall'analisi).

Verificare il sistema d'ipotesi I, equivale pertanto a verificare, dato un qualsiasi indice di classe c, che la variabile casule X sia simmetrica intorno ad c. cioè:

$$Pr(A_c) = Pr(A_{h-c+1}).$$

Consideriamo adesso la variabile di permutazione S^*, che è una trasformata della binomiale con parametro 1 e 1/2 $(bin(1, 1/2))$. Questa variabile è cosi definita: $S^* = 1 - 2bin(1, 1/2)$. Essa assume quindi valori in $\{-1, 1\}$ con probabilità 1/2. Allora il relativo spazio campionario di permutazione χ/X sarà dato da:

$\chi/X = \{\bigcup_{s^*\epsilon[-1,+1]^n} [X_u S_u^*,\ u = 1, ..., n\}$, quindi data l'indipendenza tra i vettori individuali:

$f^*(c) = \sum_i I(X_i S_i^* \epsilon\, A_c) \sim bin(v_c, 1/2)$, con $v_c = f(c) + f(k - c + 1)$, $c = 1, \cdots, k$.

Una statistica test ragionevole utilizzabile in questo caso, assumendo che tutti i punteggi abbiano lo stesso peso, è :

$$T^* = \sum_{c=1}^{\lfloor k/2 \rfloor} \frac{[f^*(c) - \frac{v_c}{2}]}{\sqrt{v_c/4}}$$

con $\lfloor k/2 \rfloor$ si intende la parte intera di $k/2$, la cui distribuzione nulla asintotica è $N(0, \lfloor k/2 \rfloor)$

5.2 Estensione al caso multivariato.

Consideriamo di essere sempre nelle stesse condizioni di prima, solo che adesso, invece di osservare una sola variabile prima e dopo, ne osserviamo q variabili come raffigurato nella tabella sottostante.

prima			dopo		
Y_{11}	$\cdots$	Y_{1q}	Y_{21}	$\cdots$	Y_{2q}
y_{111}	$\cdots$	y_{1q1}	y_{211}	$\cdots$	y_{2q1}
$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
y_{11n}	$\cdots$	y_{1qn}	y_{21n}	$\cdots$	y_{2qn}

Tabella 5.2: rappresentazione tabellare dei dati nel caso multivariato

La nostra verifica d'ipotesi sarà quindi:

$$II : \begin{cases} H_0 : \{(Y_{11} \stackrel{d}{=} Y_{21}) \cap .... \cap (Y_{1q} \stackrel{d}{=} Y_{2q})\} \\ H_1 : \{(Y_{11} \stackrel{d}{<} Y_{21}) \cup .... \cup (Y_{1q} \stackrel{d}{<} Y_{2q})\} \end{cases}$$

espressione in cui H_0 è vera se tutte le sotto-ipotesi nulle $H_{0i} : Y_{1i} \stackrel{d}{=} Y_{2i}$ sono congiuntamente vere, mentre l'ipotesi alternativa è vera se anche una sola sotto-alternativa $H_{1i} : Y_{1i} \stackrel{d}{<} Y_{2i}$ è vera.
Le differenze X in questo caso diventano: $X = \varphi(Y_{1i}, Y_{2i},\, i = 1, ...q)$. Queste differenze costituiscono una matrice (nxq). Possiamo quindi applicare il test precedente sulle sotto-tabelle nx2, e combinarle usando una adeguata funzione di combinazione, ad esempio quella diretta.

Il test sarà quindi:

$$T^* = \sum_{i=1}^{q} \sum_{c=1}^{\lfloor k/2 \rfloor} \frac{f_i^*(c) - v_{ic}/2}{\sqrt{v_{ic}/4}}$$

Visto che il modo usato per effettuare la combinazione dei test parziali è leggermente diverso da quello usato in (Pesarin and Salmaso, 2010), riporto di seguito i diversi passi dell'algoritmo che ci permette di calcolare la matrice dei test permutati: gli T_i^*.

1. calcolare la matrice delle differenze, è una matrice nxk, con n il numero di unità e k il numero di confronti da fare.

2. calcolare allora i rispettivi valori osservati T_{0i}, $i = 1, \cdots, k$ delle statistiche T_i.

3. sia $P = 1 - 2 * rbinom(1, 1/2)$, ovvero P è una trasformata della binomiale che produce valori appartenente ad $\{-1, 1\}$. Generare n valori da P chiamiamoli "Perm".

4. moltiplicare ogni colonna della matrice per Perm, e calcolare i relativi test T_{1i}^*

5. ripetere indipendentemente B volte i passi 3 e 4, otterremo cosi una matrice Bxk di test su cui calcoleremo poi i p-value

La seconda parte dell'algorimo di CMC è la stessa di quella standard.

5.3 Esempi

5.3.1 Caso univariato.

Il primo esempio che prendiamo in considerazione riguardo la valutazione di un corso. A questo proposito è stato sottoposto ad un gruppo di $n = 50$ studenti un questionario in cui ogni studenti deve dare una valutazione che va da 1 a 10 dove 10 rappresenta il punteggio massimo, sul grado di soddisfazione di come si è svolto il corso.

Supponiamo che la probabilità di osservare ogni punteggio sia nota da studi precedenti. Infatti visto che il corso esiste da diversi anni, e quindi lo stesso questionario è stato sottoposto diverse volte a diversi studenti, siamo riusciti a ricavare la probabilità di osservare ogni punteggio.

Queste probabilità sono riassunte nella tabella seguente:

Punteggio	*Probabilità*
1	0
2	0
3	0
4	0
5	0
6	0
7	0.05
8	0.10
9	0.45
10	0.40

Tabella 5.3: distribuzione di probabilità dei punteggi

Vediamo quindi che in questo caso, il dataset è leggermente diverso da quello standard visto precedentemente. Infatti, invece di confrontare le distribuzioni dei dati rilevati in due momenti diversi, siamo interessati a confrontare la distribuzione osservata con una di riferimento.

Quindi in questo caso, consideriamo come variabile osservata nella prima occasione quella teorica e variabile osservata nella seconda occasione quella realmente rilevata. Siamo quindi interessati a vedere se la distribuzione della nostra variabile (quella realmente osservata) è stocasticamente più grande di quella di riferimento.

5.3.1.1 Analisi esplorativa dei dati

Visto che siamo interessati al confronto dei nostri punteggi con quelli di riferimento la prima cosa da fare consiste nella determinazione della variabile relativa ai punteggi di riferimento. Questa variabile può essere simulata dalla distribuzione multinomiale. Infatti, la probabilità di osservare ogni punteggio è nota, è quindi sufficiente simulare n dati da una multinomiale che abbia le probabilità elencate nella tabella precedente.

Una prima analisi dei nostri dati consiste nell'analisi delle differenze tra i ranghi attributi alle diverse classi (ranghi di riferimento - ranghi osservati). La tabella sottostante riporta rispettivamente sulle colonne i ranghi delle diverse classi della variabile realmente osservata, quelli della variabile di riferimento e le loro differenze. (per comodità abbiamo riportato solo i primi 39 dati).

$ranghi\,osservati(Y_2)$	$ranghi\,simulati(Y_1)$	$differenze(Y_2 - Y_1)$
3	9	−6
9	9	0
3	10	−7
10	10	0
9	9	0
10	9	1
3	10	−7
3	9	−6
10	10	0
5	10	−5
10	8	2
8	10	−2
9	10	−1
3	8	−5
3	10	−7
9	10	−1
3	10	−7
10	9	1
8	9	−1
3	9	−6
9	10	−1
9	7	2
2	9	−7
3	10	−7
10	8	2
10	10	0
3	9	−6
3	9	−6
9	9	0
3	10	−7
3	9	−6
10	9	2
3	9	−6
3	10	−7
3	8	−5
9	9	0
7	10	−3
3	9	−6
10	8	2
3	9	−6

Tabella 5.4: dati dell'esempio univariato

Le nostre differenze sono state calcolate facendo la differenza tra i ranghi realmente osservati e quelli simulati di riferimento, e quindi una differenza negativa significherebbe che il rango di riferimento è maggiore di quello osservato e ovviamente una differenza positiva significherebbe il contrario, cioè, quello di riferimento è inferiore a quello osservo. Allora sotto H_0, ci si aspetta che le due distribuzioni siano uguali, ovvero ci si aspetta che la distribuzione delle differenze sia simmetrica rispetto allo 0, e quindi che la percentuale di differenza negativa sia uguale a quella positiva.
La tabella che andiamo adesso ad analizzare si riferisce alle differenze dei ranghi delle due variabili (ultima colonna della tabella 5.4), questa tabella riporta sia la differenze osservate ma anche la rispettive frequenze

-7	-6	-5	-3	-2	-1	0	1	2
9	15	3	1	1	5	7	4	5

Tabella 5.5: Tabella delle frequenze delle differenze tra i ranghi attributi alle due variabili.

Si evince dall'analisi della tabella che la percentuale di differenza nulla è pari all'14% ovvero , abbiamo osservato 7 differenze nulle su 50. Questa probabilità sale all'68%, per quanto riguarda le differenze negative (34 differenze negative osservate su 50), mentre le differenze positive osservate sono pari a 9, cioè il 18% delle differenze è risultato positivo. Sembra quindi che la nostra distribuzione non sia simmetrica intorno allo zero, il che ci potrebbe portare al rifiuto dell'ipotesi nulla.

E' inoltre possibile fare un'analisi grafica dei nostri dati, quest'analisi può essere fatta attraverso le funzioni di ripartizione delle nostre variabili. Visto che la nostra ipotesi nulla è che le due funzioni di ripartizione siano stocasticamente uguali, ci si aspetta quindi che sotto H_0 queste funzioni siano il più vicino possibile.

Il grafico seguente riporta le funzioni di ripartizione delle nostre due variabili in cui abbiamo rappresentato con il colore rosso la funzione di ripartizione della nostra variabile osservata e con il colore nero, quella della variabile simulata.

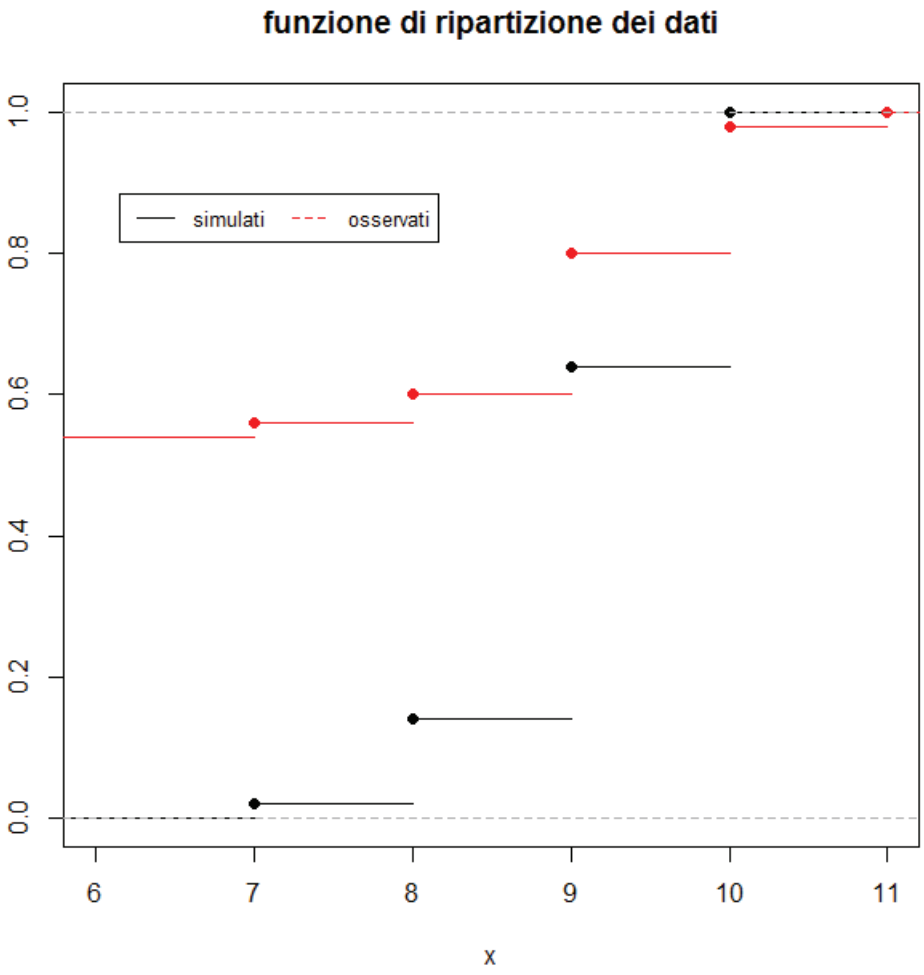


Figura 5.1: funzione di ripartizione dei dati

Infatti, vediamo dal grafico che le due distribuzioni sembrano diverse, la curva rossa è sostanzialmente sempre sopra quella nera, il che confermerebbe la nostra tesi precedente, ovvero che le differenze differenze non sono simmetriche.

A questo punto applichiamo il test per vedere se essa conferma le nostre ipotesi. Per fare ciò, abbiamo applicato il metodo Monte Carlo condizionato precedentemente visto facendo 10000 simulazioni. il livello di significatività ottenuto è : $p.\mathbf{value} = \mathbf{0.0004}$, e questo conferma le ipotesi precedenti.
Come visto prima, il livello di significatività può essere calcolato sfruttando l'approssimazione alla normale della binomiale. Infatti, visto che il nostro test non è nient'altro che la somma di $[k/2]$ test binomiali standardizzati, esso si distribuisce approssimativamente come $N(0, \lfloor {}^{k}/_{2} \rfloor)$. In questo caso k vale 19, infatti visto che abbiamo 10 classi e quindi 10 ranghi (numerate da 1 a 10), allora le differenze dei

ranghi assegnati a queste classi vanno da -9 a 9 compreso lo 0. Quindi Il livello di significatività calcolato sfruttando i quantili di una $\mathcal{N}(0, 9)$ è: 0.0028. Le differenza tra i due livelli di significatività è da attribuirsi principalmente alla scarsa numerosità in alcune classi, ossia alla non buona approssimazione della distribuzione asintotica rispetto a quella effettiva.

5.3.2 Caso multivariato

L'esempio che andiamo adesso ad analizzare è l'estensione di quello appena trattato. Ovvero, adesso siamo sempre interessati a valutare il grado di soddisfazione del corso. Ma questa volta invece di fare questa valutazione sulla base di un solo carattere, siamo interessati a valutare diversi aspetti che riteniamo importanti. Nel senso che, questi caratteri ci consentiranno poi di cercare di migliorare il corso per le prossime edizioni.
I caratteri presi in esame sono in totale 6. Le caratteristiche di queste variabili sono le medesime di prima, ovvero i punteggi di queste variabili variano sempre da 1 a 10, e le probabilità di osservare ogni punteggio sono esattamente quelle elencate nella tabella delle probabilità di prima. Siamo quindi interessati a confrontare le nostre variabili con quella di riferimento.
Visto che il p.value globale si basa sulla combinazione dei p.value parziali, è quindi possibile sapere se il livello di significatività globale sarà superiore o inferiore alla soglia(α) fissata analizzando i diversi test singolarmente.
Il prossimo grafico è relativo ai diversi test parziali effettuati. Riportiamo quindi per ogni test, le funzioni di ripartizione, in cui il colore rosso indica la funzione di ripartizione della variabile osservata e quello nero quella di riferimento. Inoltre, abbiamo aggiunto sopra il grafico il p.value del test univariato dei rispettivi confronti.

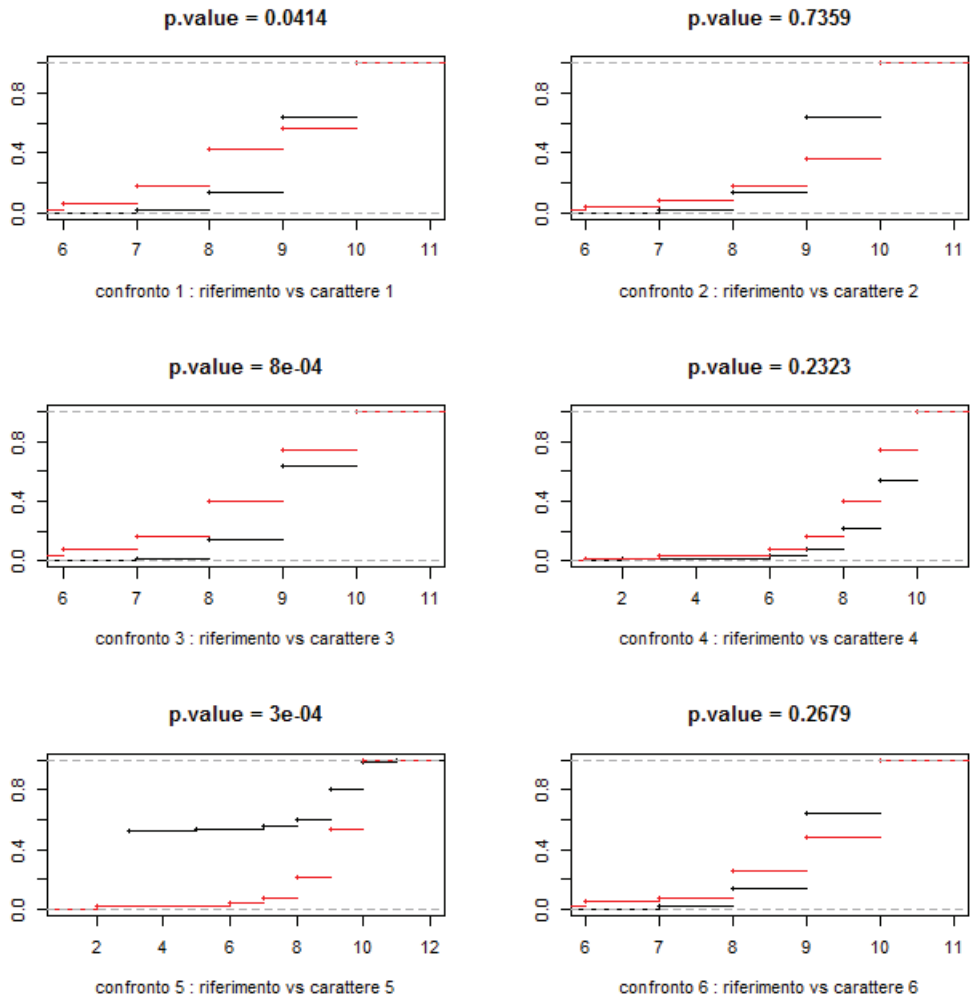


Figura 5.2: grafico delle rispettive funzioni di ripartizioni empiriche cummulate con i relativi p.value parziali

Vediamo dai grafici che non tutti i caratteri studiati (test non aggiustati per molteplicità) sono significativi al 5%, infatti il secondo, quarto e sesto caratteri hanno un $p.value > 0.05$, il che vuole dire che in questo caso accetteremo l'ipotesi nulla. Quindi in questi casi, la risposta data dagli studenti è quella prevista sotto H_0. Mentre in tutti gli altri casi, il test è risultato significativo; e come ben sappiamo, questo non ci basta a dire che il livello di significatività del test globale sarà significativo. Il livello di significatività globale, calcolato usando la combinazione di Fisher è pari ad: $p.value = 0.0051$, rifiutiamo quindi l'ipotesi nulla globale.

6 Test di simmetria per variabili categoriali ordinali con dati appaiati.

In questo capitolo consideriamo un altro test basato sullo stesso concetto di prima, ma che presenta qualche differenza rispetto al primo test.

Ovvero come prima, teniamo sempre conto dei punteggi osservati nelle diverse classi, ma diamo più importanza ai cambi di classe; nel senso che le "differenze" vengono trattate in modo diverso. Prima il test veniva effettuato direttamente sulle differenze nelle due occasioni del carattere studiato, e quindi il fatto di passare da un punteggio di 5 ad uno di 4 era uguale sotto l'ipotesi nulla a quello di passare da 10 a 9 perché la differenza fa sempre 1 e quindi il peso dato a queste due situazioni era uguale.

Il test che presentiamo adesso, invece cerca di tener conto del cambio di classe, e quindi non lavoriamo più direttamente sulle differenze, ma adesso teniamo conto delle "transizioni ", ovvero il passaggio da una categoria all'altra. E quindi sotto H_0, si assume che ci sia scambiabilità all'interno delle transizioni. Nel senso che adesso, ad esempio, si assume che passare dalla categoria 10 alla categoria 4 è uguale a passare dalla categoria 4 alla categoria 10.

Si capisce quindi che anche questo test può essere impiegato per la verifica di ipotesi in diversi ambiti. Si pensi ad esempio a quello dello marketing. A questo proposito supponiamo che una delle tante compagnie telefoniche sia interessata a valutare se la messa sul mercato di una nuova offerta telefonica abbia riscontrato il successo che ci si aspetta. Questa valutazione viene ovviamente fatta sulla base dei dati a disposizione dell'azienda. Sostanzialmente l'azienda ha classificato i clienti in diversi gruppi, e ad ogni gruppo ha dato un punteggio crescente in base all'offerta attualmente in possesso. Visto che la nuova offerta è praticamente uguale alla vecchia, solo che i servizi a disposizione si sono raddoppiati e ovviamente il prezzo ha subito un leggero aumento. Inoltre i clienti possono decidere se cambiare o meno piano tariffario, perché per la concorrenza delle altre compagnie l'azienda è costretta

a tenere tutte e due le offerte per non rischiare di perdere quelli che non hanno voglia di cambiare offerta. Alcuni clienti si trovano bene con la vecchia offerte e quindi preferiscono conservarla, altri invece non cambiano per motivi economici, e quindi possono passare da un momento all'altro alla concorrenza se le offerte sono migliori.

Si capisce che in questo caso, se si ragionasse come fatto con il test precedente, si perderebbe completamente di vista l'obbiettivo dell'azienda, ovvero non sarebbe facile tener in considerazione il cambio di piano tariffario.

6.1 caso univariato

La prima parte del nostro lavoro consiste nella costruzione del test univariato per la risoluzione del problema precedentemente illustrato. Supponiamo quindi di voler fare l'analisi di un dato carattere che viene osservato in due occasioni diverse. Questo carattere viene puoi valutato sulla base di due variabili ordinali. Supponiamo per semplicità, che le nostre variabili abbiano il medesimo supporto ($A_1 \prec A_2 \prec \ldots \prec A_k$), una possibile rappresentazione tabellare dei nostri dati è:

occasione 1	occasione 2
Y_{1i}	Y_{2i}
A_{11}	A_{21}
$\vdots$	$\vdots$
A_{1n}	A_{2n}

Tabella 6.1: tabella dei dati appaiati: caso univariato

Siamo quindi interessati a valutare sulla base delle transizioni osservate, se il carattere che stiamo studiando ha subito dei cambiamenti significativi, ovvero se il valore osservato nella seconda occasione è stocasticamente superiore a quello osservato nella prima occasione. Vogliamo quindi risolvere il sistema di verifica d'ipotesi seguente:

$$I : \begin{cases} H_0 : Y_1 \stackrel{d}{=} Y_2 \\ H_1 : Y_1 \stackrel{d}{<} Y_2. \end{cases}$$

Come abbiamo fatto notare prima, in questo caso si assume che sotto l'ipotesi nulla, i dati all'interno di ogni unità siano scambiabili. Cioè, la probabilità condizionata sotto H_0 di passare dalla $j-esima$ classe alla $h-esima$ classe nella prima occasione è la stessa di passare dalla$h-esima$ alla $j-esima$ nella seconda occasione. Quindi si ha:

$$Pr\{(A_j, A_h)/(Y_{1i}, Y_{2i}) = Pr\{(A_h, A_j)/Y_{1i}, Y_{2i}) = \frac{1}{2}$$

Sia X_i le nostre transizioni, esso ha supporto:$\{C_{11}, \cdots, C_{kk}\}$, dove $X_i = C_{hj}$, significa che lo $i-esimo$ soggetto si è mosso dalla categoria h alla categoria j.

Notiamo che, sempre per la scambiabilità delle classe all'interno di ogni individuo sotto H_0, si ha :$Pr(C_{hj}) = Pr(C_{jh})$, quindi la distribuzione delle differenze è simmetrica rispetto alla diagonale principale.
Le frequenze di permutazione f^*,si distribuiscono come una binomiale in particolare, $f^*(C_{hj}) = \sum_i I(X_i.S_i^* \epsilon\, A_{hj}) \sim bin(\nu_{hj}, 1/2)$ con $\nu_{hj} = f(C_{hj}) + f(C_{jh})$ e le $S_i^* \sim 1 - 2bin(1, 1/2)$.
La statistica che risolve il nostro test è quindi il seguente:

$$T^* = \sum_{h>j}^{k} \frac{[f^*(C_{hj}) - \nu_{hj}/2]}{\sqrt{\nu_{hj}/4}}$$

questo test non è nient'altro che la somma di $\frac{k(k-1)}{2}$ test binomiali standardizzati. E quindi, anche in questo caso, purché le frequenze in tutte le classi siano sufficientemente elevate, è possibile usare l'approssimazione alla normale per il calcolo del livello di significatività del test.

Visto che anche in questo caso, la procedura per il calcolo della distribuzione di permutazione di questo test è leggermente diverso da quelli visti precedentemente, riportiamo di seguito i passi necessari per poterla calcolare.

1. calcolare la statistica test T sui dati osservati X : $T_0 = T(X)$

2. sia $S^* = 1 - 2 * rbinom(1, 1/2)$, anche se in questa parte si poteva generare direttamente dalla binomiale standard stessa, generare n valori da S^* chiamiamoli"$Perm$".

determinare il dataset permutato : per fare ciò, scambiare le righe del dataset nx2 in base al segno $Perm_i$, e calcolare il relativo test $T^* = T(X^*)$. come mostrato nell'esempio grafico seguente ($n = 5$):

dataset iniziale		*Perm*		*dataset permutato*		
1	**3**	-1		**3**	**1**	
1	2	1		1	2	
3	2	1	$\Rrightarrow$	3	2	$\longrightarrow \quad T^* = T(X^*)$
2	**3**	-1		**3**	**2**	
3	3	1		3	3	

1. ripetere B volte i passi 2-3

2. calcolare il livello di significatività come fatto prima

$$\widehat{\lambda} = L(T_0|X) = \sum_{r=1}^{B} \frac{I(T_r^* \geq T_0)}{B}$$

6.2 caso multivariato

Supponiamo adesso che, invece di voler studiare un solo carattere siamo interessati ad analizzare q caratteri. La metodologia per poter applicare questo test ai casi multivariati è la stessa di quella vista per il caso multivariato del primo test. Le condizioni sono praticamente le stesse. La sola differenza riguarda la prima parte dell'algoritmo per la combinazione dei p.value parziali, ovvero la costruzione della tabella dei test permutati. Infatti, visto che adesso siamo interessati alle diverse transizioni, dobbiamo fare le permutazioni tenendo conto di quest'ultime. Di seguito, riportiamo i passi dell'algoritmo necessario per poter calcolare il test multivariato:

1. Sia k, il numero di confronto da fare. Calcolare allora i rispettivi valori osservati T_i, ossia $T_{0i} = T_i(X)$, $i = 1, \cdots, k$ compendiati nel vettore T_0

$$\left[\begin{array}{ccc||ccc} X_1(1) & \cdots & X_1(n_1) & X_1(1+n_1) & \cdots & X_1(n) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_q(1) & \cdots & X_q(n_1) & X_q(1+n_1) & \cdots & X_q(n) \end{array}\right] \longrightarrow \left[\begin{array}{c} T_{10} \\ \vdots \\ T_{k0} \end{array}\right]$$

2. Sia $Perm$ il vettore contenente n valori generati da S^*, considerare la matrice permutata dei dati, ovvero:

$$X^* = \{X^*(Perm_i),\ i = 1, \cdots, n;\ n_1, n_2\}$$

per spiegare meglio come permutare la matrice consideriamo l'esempio seguente in cui $n = 5$, la matrice osservata è la seguente:

occasione 1	*occasione* 2
$1 \cdots 2$	$5 \cdots 6$
$\mathbf{4} \cdots \mathbf{7}$	$\mathbf{3} \cdots \mathbf{8}$
$2 \cdots 5$	$1 \cdots 2$
$1 \cdots 4$	$1 \cdots 2$
$\mathbf{8} \cdots \mathbf{3}$	$\mathbf{7} \cdots \mathbf{4}$

supponiamo che il vettore di permutazione "$Perm$"sia il seguente:

$Perm$
1
-1
1
1
-1

visto che il confronto è tale per cui, confrontiamo l'i-esima variabile nella prima occasione con l'i-esima variabile nella seconda occasione con $(i = 1, ..., k)$, allora la matrice permutata è la seguente (abbiamo evidenziato le righe che cambiano):

occasione 1	*occasione* 2
$1 \cdots 2$	$5 \cdots 6$
$\mathbf{3} \cdots \mathbf{8}$	$\mathbf{4} \cdots \mathbf{7}$
$2 \cdots 5$	$1 \cdots 2$
$1 \cdots 4$	$1 \cdots 2$
$\mathbf{7} \cdots \mathbf{4}$	$\mathbf{8} \cdots \mathbf{3}$

calcolare quindi i corrispondenti valori delle statistiche T_i, ossia $T_i^* = T_i(X^*), i = 1, \cdots, k$ scritti nel vettore T^*

3. Ripetere B volte il punto 2, quindi l'insieme dei relativi risultati, $\{T_r^*, r = 1, ..., B\}$ costituisce una campione casuale semplice che simula la distribuzione di permutazione k-variata di T.

   i passaggi seguenti sono esattamente uguali a quello standard:

4. calcolare le associate funzioni che stimano i livelli di significatività marginali, gli $\widehat{L}_i(t)$, $t \epsilon R^1$, $i = 1, ..., k$

5. il valore osservato combinato dei test di secondo ordine utilizza i risultati ottenuti nella prima fase con CMC ed è dato da $T_0^{''} = \psi(\widehat{\lambda}_1, ..., \widehat{\lambda}_k)$

6. l' r-esimo vettore statistico simulato combinato è calcolato partendo da $T_r^{''} = \psi(\widehat{\lambda^*}_{1r}, ..., \widehat{\lambda^*}_{kr})$, dove $\lambda_{ir}^* = \widehat{L}_i(T_{ir}^*|X)$, $i = 1, ..., k$, $r = 1, ..., B$

7. Il p-value globale è dato da $\widehat{\lambda_\psi^{''}} = \sum_r I(T_r^{''} \geq T_0^{''})/B$

8. si rigetta l'ipotesi nulla globale a livello di significatività fissato e pari ad α se $\widehat{\lambda_\psi^{''}} \leq \alpha$.

6.3 Esempi

I dati che analizzeremo di seguito sono dei dati che riguardano un'analisi di mercato fatta da una azienda che fabbrica dei profumi. Essa ha deciso di mettere sul mercato una variante dei prodotti già esistenti. Prima però di commercializzare i nuovi prodotti, ha deciso di testarli su un gruppo di $n = 7$ persone. A quest'ultimi vengono quindi date in due momenti diversi, i diversi prodotti e un questionario in cui essi devono dare un punteggio che va da 1 a 10, dove 10 rappresenta il punteggio massimo. Questo punteggio rappresenta il grado di soddisfazione di ognuno di loro. L'azienda è quindi interessata a sapere se il nuovo prodotto può piacere di più rispetto al vecchio. Siamo quindi interessati a sapere se le transizioni positive sono statisticamente superiori a quelle negative. In altre parole vogliamo sapere se il numero di clienti che nella seconda occasione hanno dato un punteggio maggiore a quello della prima occasione è statisticamente superiore a quelli che hanno dato nella seconda occasione un punteggio inferiore a quello della prima occasione.

L'Azienda suppone che un cliente che ha dato un voto maggiore nella seconda occasione rispetto alla prima è un potenziale cliente, e quindi la nostra analisi consisterà nello studio congiunto delle diverse transizioni. In un primo momento, considereremo un prodotto e quindi faremmo un analisi univariata per vedere se esistono differenze significative tra i due prodotti e poi estenderemo la nostra analisi a tutti i prodotti.

6.3.1 caso univariato

La nostra analisi univariata consisterà nell'analisi di uno dei diversi prodotti. In particolare, abbiamo deciso di analizzare i dati relativi al settimo prodotto.

La tabella seguente riporta i dati relativi a questo prodotto, in cui sulla prima colonna abbiamo raccolto i dati relativi al vecchio profumo indicato con vecchio, e sulla seconda quelli relativi al nuovo, indicato con nuovo.

vecchio	*nuovo*
2	3
6	7
3	3
4	5
3	6
3	5
2	5

Tabella 6.2: Dati del settimo prodotto

La prima cosa che notiamo è ovviamente la piccola numerosità del campione, Abbiamo quindi $n = 7$ soggetti che devono "scegliere" un punteggio che va da 1 a 10, e quindi il numero di punteggi (le classi) è superiore alla numerosità campionaria. Ci si aspetta quindi di osservare 7 transizioni diverse per ogni prodotto, e quindi sarà difficile rifiutare l'ipotesi nulla. In più, avremmo una scarsa approssimazione alla normale della nostra statistica.

La tabella sottostante riporta le transizioni osservate per il settimo prodotto:

i/j	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	0	0	1	0	1	0	0
3	0	0	1	0	1	1	0
4	0	0	0	0	1	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0

Tabella 6.3: matrice delle transizioni del carattere del caso univariato

La lettura delle transizioni avviene dall'i-esima riga alla j-esima colonna. Quindi l'uno che che si trova nella seconda riga e terza colonna significa che il punteggio è passato da 2 a 3.

Come ci si aspettava, le transizioni osservate sono tutte diverse. Tuttavia, si nota che tutti gli 1 si trovano sopra la diagonale principale, questo vuole dire che tutte le transizioni osservate sono positive, ovvero, i clienti hanno tutti dato un punteggio superiore al nuovo prodotto rispetto al vecchio.

Si poteva arrivare a questa conclusione analizzando il grafico seguente. Esso riporta le votazioni di ognuna delle 7 persone. Abbiamo indicato con il rosso la votazione del primo prodotto e in verde quello del secondo.

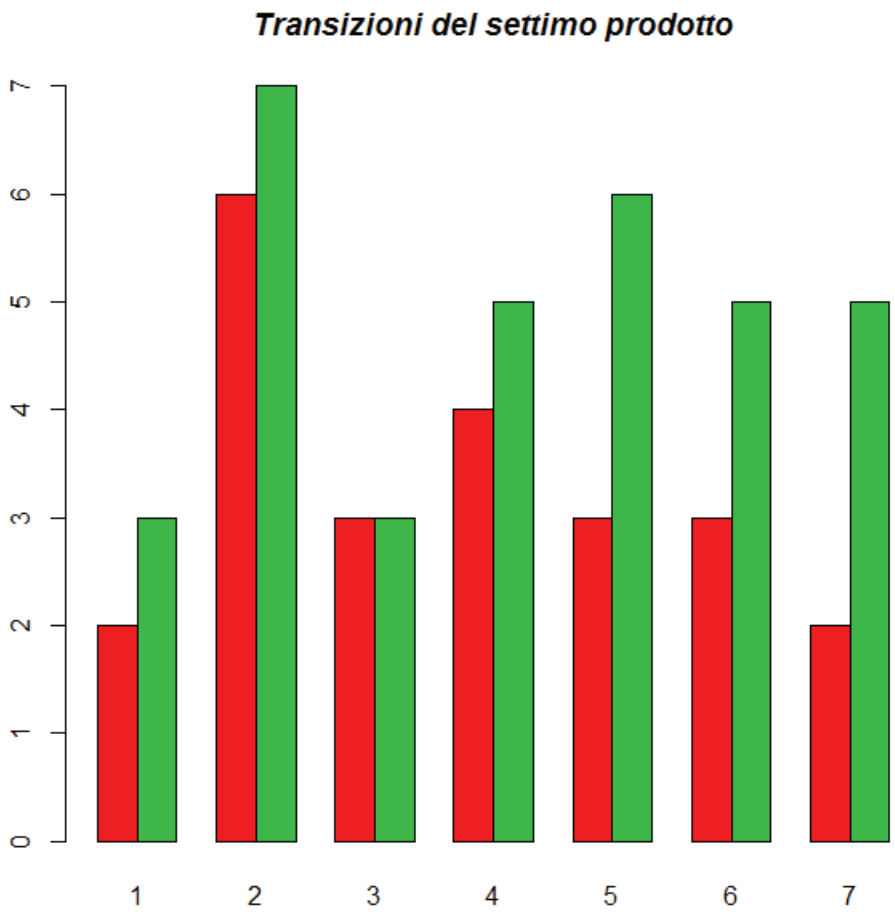


Figura 6.1: grafico delle transizioni del settimo profumo

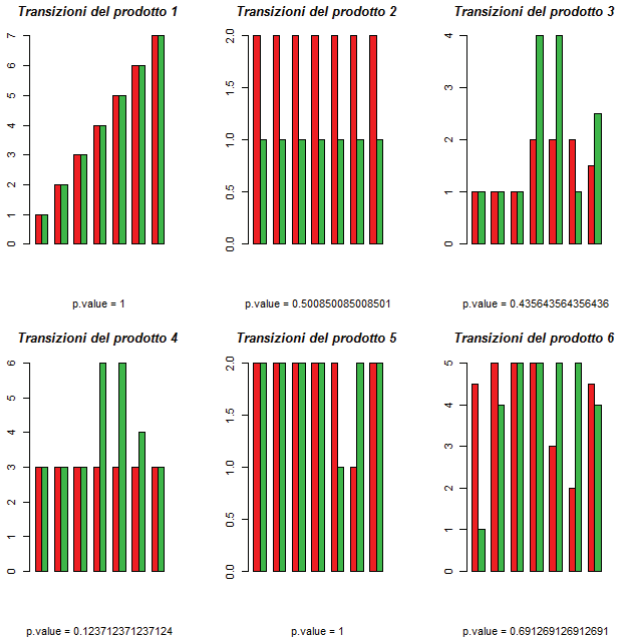
Vediamo che arriviamo alla stessa conclusione di prima. Infatti, la maggior parte degli individui ha preferito il secondo profumo, ad eccezione del terzo individuo che è rimasto imparziale; tutti gli altri individui hanno dato un punteggio migliore al nuovo prodotto. Sembra quindi che l'introduzione del settimo prodotto sul mercato può

portare all'incremento della clientela. Andiamo a vedere se è possibile confermare questa asserzione anche a livello inferenziale.

Per poter applicare la nostra statistica abbiamo anche in questo caso usato 10000 permutazioni, ottenendo cosi un livello di significatività pari ad: $p.value = 0.01610$. Quindi rifiuteremo l'ipotesi nulla, ovvero le differenze osservate sono significative al 5%. Quindi, per il settimo prodotto, il livello di gradimento per il nuovo prodotto è superiore rispetto al vecchio.

6.3.2 caso multivariato.

Adesso, invece di confrontare una sola coppia di profumi siamo interessati a confrontare $q = 15$ coppie. Nei grafici seguenti, abbiamo riportato le diverse transizioni dei diversi confronti. Abbiamo inoltre aggiunto i relativi livelli di significatività.



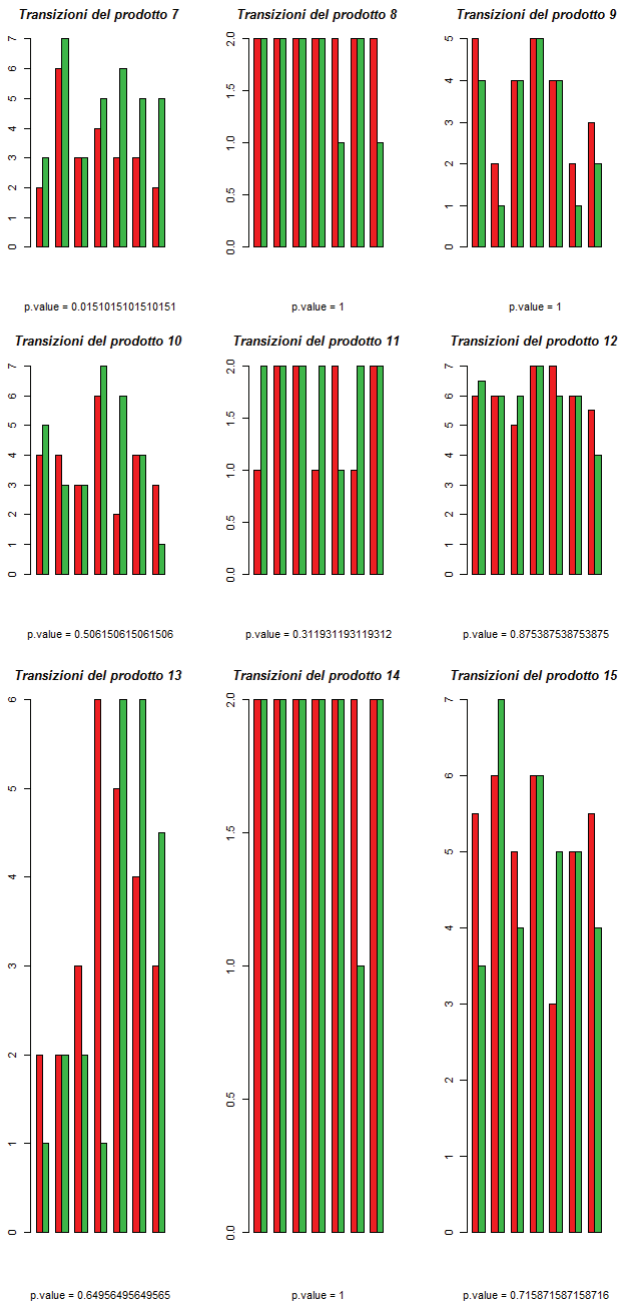


Figura 6.2: transizioni dei diversi confronti

Da questa analisi grafica, emerge che non tutti i prodotti sono stati graditi allo stesso modo dalla popolazione del campione. In particolare, mentre per il settimo prodotto, il livello di gradimento è superiore per il nuovo prodotto rispetto al vecchio. Nel caso del secondo profumo, osserviamo l'effetto contrario. Tuttavia, e come si vede dai rispettivi livelli di significatività, queste differenze risultano non significativi.

Il p.value globale è pari ad 0.0466, questo livello di significatività è stato calcolato usando la combinazione di Fisher per combinare i diversi test.

Come si vede rifiuteremo quindi l'ipotesi nulla al 5%, anche se abbiamo visto che la maggioranza dei test parziali erano non significativi(ad eccezione del settimo). Ed è proprio quello che ci porta al rifiuto di H_0. Infatti, la combinazione di test parziali è consistente se almeno un test parziale lo è. Ma la particolarità di questa tecnica è il fatto di poter poi studiare i singoli test in modo da sapere quali sono i test che hanno portato al rifiuto dell'ipotesi nulla.

7 simulazioni

Quest'ultimo capitolo riguarda lo studio della potenza dei due test implementati in questo lavoro. Lo studio sarà quindi fatto usando diverse numerosità campionarie (n) per vedere come si comporta la potenza al variare di n e della non centralità. Per la simulazione delle due variabili per ogni campione, sarà considerato un coefficiente δ (effetto del trattamento) che mi permetterà di variare le due funzioni di ripartizione. Queste variabili sono costituite da tre classi ordinate.

Di seguito riportiamo il modo in cui abbiamo costruito le diverse classi.

- per la prima variabile che chiamiamo X_2, generiamo un numero casuale X da una uniforme ($\mathcal{U}(0,1)$); a questo punto, assegniamo i ranghi alle diversi classi nel seguente modo:

$$\begin{cases} 1 & se\, X \epsilon (0, 0.28] \\ 2 & se\, X \epsilon (0.28, 0.7] \\ 3 & se\, X \epsilon (0.7, 1]. \end{cases}$$

- Il procedimento per la determinazione delle diversi classi della seconda variabile (X_1) è praticamente identico a quello precedente, solo che in questo caso, per la generazione del numero casuale X entra in gioco il coefficiente $\delta \epsilon (1,\ 0.95,\ 0.9,\ 0.75,\ 0.5)$. Invece quindi di generare X direttamente da una $\mathcal{U}(0,1)$, lo generiamo da $\mathcal{U}(0,1)^{\delta}$. In questo modo, per δ vicino a 1, ci si aspetta che le funzioni di ripartizioni delle due variabili siano identiche mentre per δ che tende a 0.5, esse devono essere tendenzialmente diverse.

Supponiamo quindi di voler verificare il sistema d'ipotesi seguente:

$$\begin{cases} H_0 :\ X_1 \overset{d}{=} X_2 \\ H_1 :\ X_1 \overset{d}{>} X_2. \end{cases}$$

Come detto in precedenza, ci aspettiamo quindi di accettare l'ipotesi nulla per valori di δ vicino ad uno, mentre la rifiuteremo mano a mano che diminuisce tale coefficiente.

In termini di potenza, sotto l'ipotesi nulla ci si aspetta una potenza vicino all'α scelto che nel nostro caso vale 0.05. mentre esso deve avvicinarsi a 1 via via che ci si allontana da H_0.
Le tabelle seguenti riguardano lo studio delle potenze per il primo test studiato, nelle condizioni appena descritte. Questo studio è stato fatto considerando diverse numerosità campionarie (n=20,50,100). In particolare lo studio della potenza è stato fatto considerando i diversi valori di δ per la generazione della seconda variabile. Inoltre la potenza è stata calcolata considerando sia l'approccio permutazionale che l'approssimazione asintotica normale.

La prima tabella che andiamo ad analizzare riguardo lo studio della potenza del primo test. In questo caso, visto che il numero delle differenze tra i ranghi(punteggi) attributi alle diverse classi vale 5 (queste differenze sono {-2,-1,0,1,2}), la potenza in questo caso è stata calcolata sfruttando i quantili di una normale $N(0, 2)$ (2 è la parte intera di 2/5). Infine, per il calcolo dei livelli di significatività dei test e il calcolo delle potenze abbiamo deciso di fare 5000 permutazioni.

		permutazione	normale
$\delta = 1$	$n = 20$	0.0362	0.0462
	$n = 50$	0.0482	0.0510
	$n = 100$	0.0450	0.0468
$\delta = 0.95$	$n = 20$	0.0492	0.0654
	$n = 50$	0.0676	0.073
	$n = 100$	0.0836	0.0868
$\delta = 0.90$	$n = 20$	0.0608	0.0788
	$n = 50$	0.0982	0.107
	$n = 100$	0.1424	0.148
$\delta = 0.75$	$n = 20$	0.1346	0.1658
	$n = 50$	0.3056	0.3212
	$n = 100$	0.4894	0.4986
$\delta = 0.5$	$n = 20$	0.4718	0.5216
	$n = 50$	0.8678	0.8796
	$n = 100$	0.9916	0.9924

Tabella 7.1: calcolo delle potenze del primo test

La prima cosa che si può notare e come ci si aspettava, è che il test di permutazione è abbastanza ben approssimato dalla distribuzione normale. Infatti le potenze

calcolate usando i due approcci sono sostanzialmente uguali. Inoltre possiamo dire che il nostro test è applicabile anche ai piccoli campioni. In effetti vediamo che, con soli 20 osservazioni divisi in tre classi e quindi mediamente 6 osservazioni per classe si arriva ad una potenza accettabile quando esiste una reale differenza tra le situazioni sperimentali, questa potenza si aggira intorno a 0.47. Essa ovviamente cresce all'aumentare della numerosità campionaria e delle differenze che esistono tra le due variabili. In particolare, si osserva che con 100 osservazioni il test ha una potenza di quasi uno. Quindi in grado di produrre statistiche accurate.
Le conclusioni che riguardano lo studio della tabella delle potenze del secondo test sono essenzialmente uguali a prima, solo che qui, l'approssimazione è stata fatta usando una $N(0, 3)$, 3 perché abbiamo considerato le tre transizioni che ci portano ad avere quindi tre somme di binomiali standardizzati ($1 \to 2$,$1 \to 3$, $2 \to 3$). Inoltre in questo caso invece di usare 5000 permutazioni per il calcolo della potenza e del livello di significatività , abbiamo usato 2000 perché questo algoritmo è leggermente più lento dell'altro.

		permutazione	normale
$\delta = 1$	$n = 20$	0.043	0.05
	$n = 50$	0.052	0.054
	$n = 100$	0.042	0.041
$\delta = 0.95$	$n = 20$	0.045	0.059
	$n = 50$	0.068	0.068
	$n = 100$	0.089	0.09
$\delta = 0.90$	$n = 20$	0.073	0.091
	$n = 50$	0.102	0.103
	$n = 100$	0.158	0.153
$\delta = 0.75$	$n = 20$	0.13	0.164
	$n = 50$	312	0.323
	$n = 100$	0.477	0.471
$\delta = 0.5$	$n = 20$	0.51	0.557
	$n = 50$	0.856	0.862
	$n = 100$	0.99	0.992

Tabella 7.2: calcolo delle potenze del secondo test

8 Conclusione

Come si è visto, il passo che ha permesso la trattazione delle variabili categoriali ordinali è stato il fatto di non lavorare direttamente sulle tabelle di contingenza, ma di usare la rappresentazione unit-by-unit. Questo metodo ci ha permesso di contare i diversi elementi di ogni classe riuscendo cosi a costruire dei test basati sulla distribuzione binomiale.

Questi test sono dei test che hanno una potenza accettabile anche quando la numerosità campionaria è bassa. Inoltre, Quando abbiamo abbastanza osservazioni per ogni classe, è possibile usare l'approssimazione alla normale per il calcolo dei livelli di significatività. Perché in questi casi, il nostro test è ben approssimato dalla distribuzione normale.

Visto che l'analisi multivariata è basata sulla combinazione dei test parziali, allora oltre ad avere un livello di significatività globale, è possibile risalire, mediante i metodi di analisi della molteplicità, alle inferenze parziali per individuare se e quali contribuiscono maggiormente alla significatività globale. Questo è quindi uno dei pregi di questo metodo.

Infine, la tecnica usata in questo lavoro è il punto di partenza per la costruzione di altri test. Infatti, in questo lavoro, abbiamo pensato solo a due modi di considerare le differenze. Ma questo non vuole dire che questi sono gli unici modi di costruire le differenze.

Elenco delle figure

5.1 funzione di ripartizione dei dati 37
5.2 grafico delle rispettive funzioni di ripartizioni empiriche cummulate con i relativi p.value parziali 39

6.1 grafico delle transizioni del settimo profumo 48
6.2 transizioni dei diversi confronti 50

Elenco delle tabelle

4.1 Metodo C.M.C 22
4.2 procedura C.M.C 28

5.1 rappresentazione tabellare dei dati nel caso univariato 30
5.2 rappresentazione tabellare dei dati nel caso multivariato 32
5.3 distribuzione di probabilità dei punteggi 34
5.4 dati dell'esempio univariato 35
5.5 Tabella delle frequenze delle differenze tra i ranghi attributi alle due variabili. 36

6.1 tabella dei dati appaiati: caso univariato 42
6.2 Dati del settimo prodotto 47
6.3 matrice delle transizioni del carattere del caso univariato 47

7.1 calcolo delle potenze del primo test 54
7.2 calcolo delle potenze del secondo test 55

Bibliografia

[1] Pesarin F., Salmaso L. Permutation Tests for Complex Data: theory, Applications and Software(2010). John Wiley & Sons

[2] Basso D. , Pesarin F., Salmaso L., Solari A.: Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications with R (2009) Springer.

[3] Pesarin, F.: Multivariate Permutation tests: with application in Biostatistics (2001). John Wiley & Sons, Chichester-New York.

[4] Miguel A.: some Bootstrap tests of symmetry for univariate continuous distributions, the annals of statistics 1991,vol. 19, No. 3, 1496-1511

[5] Ayman Baklizi (2003): A conditional distribution free runs test for symmetry , Journal of Nonparametric Statistics, 15:6, 713-718

[6] Rudolf Beran: Testing for ellipsoidal symmetry of a multivariate density , the annals of statistics 1979, vol7, No 1 150-162

[7] BHASKAR B.:Testing Conditional Symmetry Against One-Sided Alternatives in Square Contingency Tables, Metrika (1998) 47:71-84

[8] Louis N. Christofides, Thanasis S. :A non-parametric test of the symmetry of PSID wage–change distributions,Economics Letters 71 (2001) 363–368

[9] SÅNDOR CSÖRGŐ and C. R. HEATHCOTE :Testing for symmetry, 1987 Biometrika Trust

[10] Rob J. Hyndman & Qiwei Yao (2002): Nonparametric Estimation and Symmetry Tests for Conditional Density Functions, Journal of Nonparametric Statistics, 14:3, 259-278

[11] V.A. KOSTELECK´Y: TESTING CPT SYMMETRY, Physics Department, Indiana University, Bloomington, IN 47405, U.S.A.

[12] Lewbel A. :Consistent nonparametric hypothesis tests with an application to Slutsky symmetry,Journal of Econometrics 67 (1995) 379-4011

[13] Natalie Neumeyer, Holger Dette & Eva-renate Nagel (2005): A note on testing symmetry of the error distribution in linear regression models, Journal of Nonparametric Statistics, 17:6, 697-715

[14] E. D. Rothman and M. Woodroofe: A Cramer Von-Mises Type Statistic for Testing Symmetry, Ann. Math. Statist. Volume 43, Number 6 (1972), 2035-2038.

[15] Joseph Ngatchou-Wandji: Testing for symmetry in multivariate distributions, Statistical Methodology 6 (2009) 230250

[16] Li-Xing Zhua, and G. Neuhaus: Conditional tests for elliptical symmetry, Journal of Multivariate Analysis 84 (2003) 284–298

[17] Sheldom M. R.: calcolo delle probabilità, 2008 Apogeo