Università degli Studi di Padova

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E INFORMATICA

TESI DI LAUREA

TEST DI PERMUTAZIONE NEL CONFRONTO IN ETEROGENEITÀ PER DUE CAMPIONI

Relatore: Ch.mo Prof. FORTUNATO PESARIN

Laureanda: MARTA DI MARIA

ANNO ACCADEMICO 2006 - 2007

Ringrazio il Professor Fortunato Pesarin,
per la massima disponibilità dimostratami,
per essere sempre stato presente durante la stesura della tesi
e per i suggerimenti datimi.

Ringrazio la mia famiglia,
per aver contribuito al raggiungimento di questo importante traguardo,
per esser stata presente in ogni momento e avermi sopportato nei periodi di lavoro e tensione,
per aver sempre creduto in me e per il sostegno morale datomi.

Ringrazio Nicola,
per aver tenuto sempre alto il mio umore anche nei momenti di sconforto
e per essermi stato vicino.

Indice

INTRODUZIONE

In alcune discipline scientifiche assume particolare interesse stabilire se la distribuzione di una determinata variabile è più concentrata in una popolazione piuttosto che in un'altra, cioè se una certa popolazione è meno eterogenea rispetto all'altra.

Una scienza in cui l'eterogeneità risulta di rilevante interesse è la genetica, specialmente per la valutazione della biodiversità. In questo settore, gli studi si occupano di confrontare due popolazioni per verificare quale delle due presenta una differenziazione genetica maggiore, vale a dire un'eterogeneità più grande dal punto di vista delle combinazioni fenotipiche di certi fattori genetici.

Per valutare l'eterogeneità tra due popolazioni, sulla base di dati campionari, è stato considerato uno studio antropologico, svolto da C. Corrain nel 1975, su alcune popolazioni pastorali del Kenya. Per rendere possibile il confronto tra alcuni caratteri emotipologici osservati su queste popolazioni, vennero raccolti sei campioni sierologici, estratti da altrettanti tribù, stanziate per la maggior parte nelle terre semiaride a nord del Kenya. Infatti, nell'evoluzione di un determinato gruppo razziale e della sua struttura genetica, entrano in gioco molti fattori, tra i quali la selezione naturale in rapporto all'ambiente, l'isolamento e l'adattamento ad uno specifico territorio e lo stile di vita semi-nomade, che permette loro di avere contatti con altre popolazioni, con conseguenti scambi genetici e trasmissioni di caratteri ereditari.

Le sei etnie sono state confrontate, da un punto di vista genetico, analizzando le combinazioni fenotipiche di 4 fattori Gm e osservando le frequenze con cui questi diversi fenotipi si trovano presenti nel patrimonio genetico delle tribù considerate.

Dapprima è stato condotto uno studio per valutare la dissomiglianza in distribuzione tra coppie di tribù.

In seguito si è verificato se esiste dominanza in eterogeneità da parte di popolazioni nomadi nei confronti di altre più stanziali.

Infine è stato condotto uno studio per stabilire quali etnie sono maggiormente somiglianti fra loro e significativamente diverse da altre, per poterle suddividere in gruppi omogenei al loro interno, dal punto di vista genetico, ed eterogenei tra loro.

CAPITOLO 1

CENNI DI GENETICA E STORIA DEL KENYA E SUE TRIBÙ

La biometria è la disciplina preposta allo studio matematico-statistico dei fenomeni biologici ed è quindi applicabile anche alle ricerche sull'ereditarietà biologica.
Con la biometria si è dimostrato che molti fatti "qualitativi" possono essere riportati a termini numerici e studiati statisticamente. Inoltre è possibile stabilire delle relazioni che esprimano legami tra attributi presenti in gruppi di individui.
Gli strumenti logico-matematici più utili furono quelli elaborati da Fisher, che risolsero il problema della formulazione e della verifica delle ipotesi, dell'elaborazione dei programmi di esperimenti e della valutazione dei risultati.
Sulla traccia dell'opera di Fisher, la biometria attuale persegue vari indirizzi, tra i quali quello riguardante la genetica delle popolazioni.

1.1 GENOTIPI E FENOTIPI

La costituzione genetica di un individuo, ovvero il suo patrimonio ereditario, è contenuto nel genotipo. È ciò che è racchiuso nel nucleo di tutte le cellule del DNA ed è quindi immutabile.
L'insieme delle caratteristiche visibili dell'organismo, o in qualche modo evidenziabili, sono indicate invece dal fenotipo. Quest'ultimo, quindi, è l'insieme dei caratteri che l'individuo manifesta, e dipende dal suo genotipo, dalle interazioni fra i geni e anche da influenze ambientali esterne; dunque i geni ed i fattori ambientali associati ad un particolare fenotipo possono variare tra i diversi gruppi etnici.

1.2 IL KENYA: L'AMBIENTE NATURALE

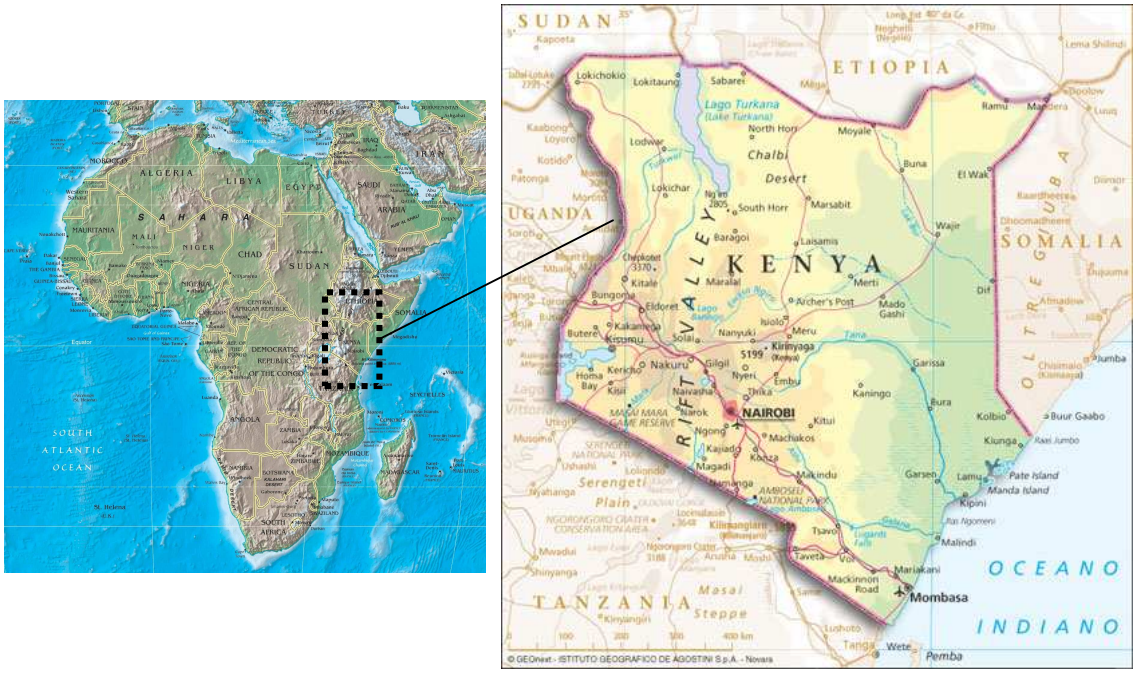
L'adattamento all'ambiente è certamente un fattore essenziale nella modificazione dei tratti somatici dell'uomo.

Secondo l'antropologia moderna, l'umanità è una composizione di diversi popoli con caratteristiche somatiche molto differenti, dovute ai processi di selezione, di isolamento e di adattamento ai diversi ambienti geografici.

In questo contesto viene preso in considerazione l'ambiente keniota.

Il Kenya, situato nell'Africa equatoriale, confina a nord con l'Etiopia e il Sudan, a ovest con l'Uganda, a sud con la Tanzania, a sud-est è bagnato dall'oceano Indiano mentre a est confina con la Somalia.

Figura 1 Cartina geografica dell'Africa e ingrandimento della cartina del Kenya con relativi confini.



La latitudine equatoriale, la conformazione del rilievo e l'apertura sull'oceano Indiano rappresentano alcuni dei numerosi fattori in grado di influenzare le condizioni climatiche del Kenya che presenta, perciò, caratteri abbastanza diversificati. Il Kenya, quindi, attraversato dall'equatore, comprende principalmente tre regioni morfologiche e climatiche ben differenziate ed è perciò caratterizzato da condizioni ambientali estremamente diversificate.

Nella sua parte centro-settentrionale, e in particolare nella porzione che si estende fra la costa occidentale del lago Turkana e il confine somalo, il bassopiano keniota assume i caratteri di una distesa arida e semidesertica. Tali regioni settentrionali e centrali sono prevalentemente caratterizzate da un clima semidesertico e il manto vegetale è rappresentato da diverse associazioni di specie in grado di resistere all'aridità dell'ambiente. Il 70% quasi del territorio keniota, quindi, è occupato da una vasta distesa di steppe e deserti pressoché disabitata. L'estrema aridità esclude ogni forma di sfruttamento agricolo, rendendo, come si vedrà più avanti, quelle terre occupate solo da pastori nomadi costretti a spostarsi per portare al pascolo il proprio bestiame.

Procedendo verso l'interno, il territorio keniota è caratterizzato da una profonda faglia, la grande fossa tettonica della Rift Valley, che attraversa il territorio da nord a sud. A nord, al confine con l'Etiopia, tale fossa è occupata dal lago Turkana, principale bacino lacustre. In corrispondenza della sezione centrale, si possono incontrare alcuni importanti complessi di origine vulcanica recente, dalle altitudini piuttosto elevate. Verso sud-ovest, il sistema degli altipiani digrada dolcemente sino a raggiungere l'ampio bacino del lago Vittoria, dove il Paese gode di un clima sub-equatoriale, proprio grazie all'influsso benefico sulla temperatura e sulle precipitazioni apportato dalla grande massa d'acqua del lago. Il bordo occidentale, quindi, appare contrassegnato da precipitazioni molto più frequenti e, in corrispondenza dei rilievi, la vegetazione diventa rigogliosa:

con l'aumentare dell'altitudine, si può notare in queste zone il cambiamento della vegetazione da foresta, a savana ed a prateria alpina.
È proprio su queste alte terre degli altipiani centro-occidentali che è avvenuta la colonizzazione agricola da parte degli Europei, di cui si parlerà nel prossimo paragrafo.
Infine, la linea di costa che si apre sull'oceano Indiano risulta notevolmente articolata: è costituita da fertili suoli alluvionali ed è interessata da una discreta piovosità, soprattutto a sud, dove è possibile praticare l'agricoltura. La vegetazione, infatti, è piuttosto fitta e la costa è caratterizzata da frequenti precipitazioni e da un clima caldo umido di tipo equatoriale.

1.3 IL KENYA: PROFILO STORICO-CULTURALE

La selezione naturale in rapporto all'ambiente è il primo elemento dell'evoluzione di un tipo razziale. Nel tempo, l'umanità ha potuto differenziarsi in razze per effetto della selezione naturale, per l'isolamento o per l'essersi adattata ad uno specifico ambiente.
Ma le differenziazioni razziali sono state originate anche per certe costrizioni bioculturali, che hanno indotto gli individui a seguire costumi e regole proprie del gruppo a cui appartengono.
Il Kenya, abitato inizialmente da genti di stirpe camitica, è stato oggetto, in passato, di colonizzazione e di contese da parte di Arabi ed Europei; l'isolamento di alcune tribù in certe sue zone e le differenziazioni tra tali etnie, dal punto di vista culturale e sociale, trovano spiegazione, quindi, dalla storia passata del Kenya.
Prima dell'arrivo degli Europei, la popolazione, di ceppo prevalentemente bantu, era insediata in modo permanente solo nelle regioni umide sud-occidentali e lungo le fertili piane costiere, favorevoli alla pratica dell'agricoltura e dell'allevamento del bestiame.

Gli Arabi, intorno al IX secolo d.C., fondarono diverse città sulla costa del Kenya ed avviarono rapporti commerciali con la tribù locale dei bantu. Per quanto riguarda questi ultimi, nello specifico i Kamba, essenzialmente agricoltori, si manifestarono presto come uno tra i gruppi più numerosi e potenti, anche se la loro supremazia fu messa in discussione dai bellicosi Masai.

Ancora quasi totalmente spopolate, le terre interne del Kenya, come le regioni degli altipiani, venivano percorse da tribù nomadi di cacciatori e pastori dalle origini nilo-camitiche, come i Masai e i Turkana.

Con la progressiva espansione delle attività agricole, praticate prevalentemente dalle genti bantu e in particolare dai Kamba, si accese un contrasto sempre più acuto fra gli agricoltori stanziali e gli allevatori nomadi, che segnò una netta linea di separazione tra i due gruppi. Tale divisione venne rinvigorita ulteriormente dall'avvento degli Europei i quali si insediarono nella regione degli altipiani, sconvolgendo la struttura socio-culturale di queste popolazioni. La colonizzazione avvenne con la collaborazione dei docili Kamba, mentre i Turkana, ma soprattutto i Masai più fieri e indipendenti, venivano sospinti nelle regioni più aride dell'interno. I Masai, ancorati alle loro tradizioni di allevatori, si trovarono quindi sempre più isolati.

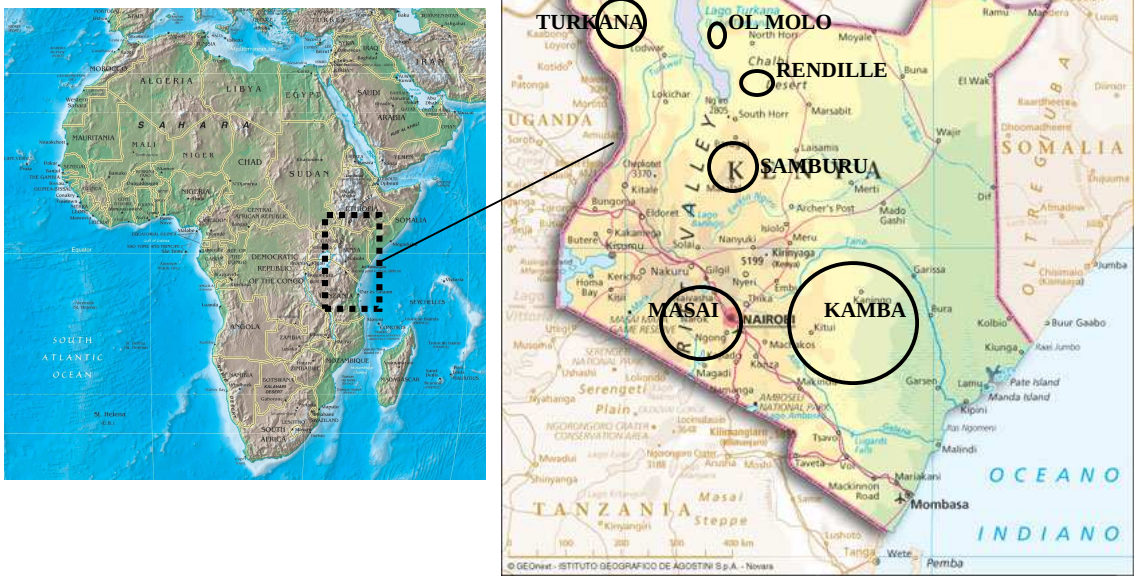
Ancora oggi, le popolazioni autoctone stanziali, che si occupano prevalentemente di agricoltura nelle terre fertili del sud, e di allevamento e pastorizia nelle terre più interne, vivono sparpagliate in villaggi. Data l'ineguale distribuzione delle piovosità e, di conseguenza, del potenziale agricolo delle terre coltivate, la densità degli insediamenti è altamente disomogenea: il 20% del territorio keniota è coltivabile e su tali terre si registra l'80% della popolazione. Quindi, mentre le zone sud-occidentali sono caratterizzate da densità elevatissime, le distese dell'interno risultano oggi scarsamente popolate.

1.4 IL KENYA: LE DIVERSE ETNIE

La maggior parte della popolazione del Kenya è formata da neri e sono divisi in due gruppi: i bantu e i nilo-camitici. Questi due gruppi sono, a loro volta, divisi in varie tribù. Tra le tribù di ceppo bantu si ricorda l'etnia Kamba, che rappresenta circa l'11% della popolazione keniota. Tra le tribù di ceppo nilo-camitica, in prevalenza nomadi stanziati nelle regioni occidentali, i Masai (1,6% della popolazione keniota) e i Turkana (1,3%) costituiscono le etnie maggiormente importanti e popolose.

Nella sottostante cartina, vengono evidenziate le sei tribù oggetto di analisi in questa tesi, e di seguito verrà tracciato un loro breve profilo storico.

Figura 2 Cartina geografica dell'Africa e ingrandimento della cartina del Kenya con indicazioni della distribuzione delle sei relative tribù esaminate.



1.4.1 I KAMBA

I Kamba, gruppo etnico di lingua bantu, costituiscono l'etnia più antica del Kenya, tra le sei che verranno di seguito trattate. Formata prevalentemente da agricoltori, si insediò fin dall'inizio nelle zone costiere affacciate sull'oceano Indiano, avviando piantagioni e coltivazioni grazie alla fertilità di tali terre.

Nel tempo, tale etnia non sembra aver subito forti selezioni naturali in rapporto all'ambiente, poiché tuttora risulta stanziata nella zona costiera: ciò porta a pensare che i Kamba si siano adattati abbastanza velocemente all'ambiente in cui si trovavano e non abbiano subito grandi cambiamenti, anche genetici, a causa di spostamenti o isolamenti.

1.4.2 I MASAI

Sicuramente una delle tribù più conosciute dell'Africa orientale, i Masai sono un gruppo nilo-camitico che vive negli spazi aperti della Rift Valley. Famosi per la loro reputazione di temibili guerrieri, fin da tempi antichi, non hanno mai abbandonato lo stile di vita semi-nomade e l'allevamento come principale fonte di sostentamento. L'abilità dei Masai sta nel sopravvivere nell'ambiente aspro e nel paesaggio accidentato della Rift Valley.

Poiché sono tutti prevalentemente pastori semi-nomadi, la loro vita è molto condizionata dalla presenza di acqua e pascoli per gli animali. Pertanto, sono costretti molto spesso a migrare in altre zone più adatte per le loro mandrie.

Secondo la storia, la maggioranza delle terre fertili del Kenya un tempo appartenevano a questi celebri pastori della savana, i quali avevano conquistato i migliori pascoli sottomettendo le altre tribù. Il declino della loro egemonia cominciò con la diffusione della peste bovina a cui si aggiunse la colonizzazione da parte degli Europei. Nonostante la

crescente occidentalizzazione, i Masai sono riusciti a conservare gran parte delle loro abitudini tradizionali.

Oggi i Masai portano liberamente le loro mandrie al pascolo, incuranti dei confini e dei regolamenti imposti dalle autorità governative per proteggere le aree dei parchi nazionali. Per questo, ancora adesso, vengono allontanati dalle loro terre e costretti sempre più in territori ridotti.

Di conseguenza, per effetto dell'isolamento e del bisogno imposto di adattarsi all'ambiente in cui vivono confinati, i Masai sono stati soggetti ad una forte selezione naturale, quindi ad un forte cambiamento sociale, anche genetico, all'interno della propria etnia.

Come si vedrà in seguito, Masai e Samburu hanno una storia che li accomuna.

1.4.3 I SAMBURU

I Samburu sono un popolo di pastori nomadi guerrieri, che vivono nella parte centro settentrionale del Kenya. Il loro territorio è molto vasto ed i loro spostamenti avvengono in un ambiente semi arido, con vegetazione molto scarsa. Ed è proprio la natura non rigogliosa che ha costretto i Samburu allo sfruttamento del terreno col bestiame.

Di origine nilo-camitica, sono per le somiglianti sembianze somatiche, oltre che per le usanze e le tradizioni antiche, parenti dei Masai, con i quali condividono anche la lingua. Solo in epoca coloniale i Samburu divennero un'etnia distinta. Alcuni Samburu, infatti, discendono dai Masai, altri sono di origine Rendille e Turkana. Proprio per questa loro discendenza è possibile ipotizzare che i Samburu presentino diversi caratteri ereditari e geni trasmessi dalle diverse etnie da cui hanno avuto origine.

1.4.4 I RENDILLE

Sono pastori nomadi che vivono in una regione semidesertica del Kenya centro-settentrionale. Pur condividendo tale territorio con altri gruppi etnici, tra cui i Turkana ma soprattutto i Samburu, essi si concentrano principalmente in un arido altopiano vulcanico, a est del lago Turkana. Proprio per l'aridità del terreno e per il clima di tipo desertico, la loro vita è organizzata attorno all'allevamento del cammello.

È un popolo poco numeroso e che, pian piano, sta scomparendo: alla minaccia di furti di bestiame, infatti, si aggiunge negli ultimi decenni un altro fattore in grado di mettere in pericolo la stessa sopravvivenza dei Rendille, cioè un sempre maggiore inaridimento del clima.

1.4.5 I TURKANA

I Turkana sono un popolo del gruppo nilotico, hanno stabilito la loro dimora nell'estremo nord-ovest del Kenya, in una regione adiacente a quelle abitate dalle popolazioni Rendille e Samburu, con le quali hanno frequenti scontri.

Tribù poligama, come le altre etnie esaminate finora, è solo grazie alla loro indole aggressiva se oggi sono ancora numerosi, pur vivendo in luoghi aridi e inospitali.

In epoca coloniale, erano ricchi di bestiame; oggi, invece, a causa dei nuovi confini tribali e del progressivo inaridimento del terreno, con piogge irregolari e pascoli che scarseggiano, sono costretti ad una vita più dura e difficile. La maggior parte sono pastori nomadi costretti a spostarsi di continuo alla ricerca di nuovi pascoli e di acqua per il bestiame.

Per questa loro continua lotta per adattarsi all'ambiente ostile, sono stati oggetto anch'essi, come i Masai, di una forte selezione naturale. Inoltre, si ricorda che l'etnia

Turkana è stata tra quelle che hanno generato la tribù Samburu, di conseguenza è stata coinvolta e interessata in alcuni scambi genetici.

1.4.6 GLI OL MOLO

La tribù degli Ol Molo è forse il più piccolo gruppo etnico del Kenya. È di etnia nilo-camitica, come le altre popolazioni del Kenya settentrionale (Turkana, Rendille e Samburu) e vive sulle rive del lago Turkana.
Le principali attività degli Ol Molo sono la caccia e la pesca, seppur scarsa, data la vicinanza al bacino lacustre e l'impossibilità di coltivare piantagioni a causa del terreno arido e semidesertico che si sviluppa nelle zone circostanti.
A differenza delle altre tribù, gli Ol Molo sono rigorosamente monogamici.

In conclusione, l'unica etnia che sembra sia rimasta fedele al proprio territorio, fin dai tempi antichi, è quella dei Kamba, dedita all'agricoltura grazie al terreno altamente fertile della zona costiera del Kenya.
Le altre popolazioni sono formate prevalentemente da pastori nomadi, quindi costrette a spostarsi continuamente in cerca di pascoli e zone più adatte al bestiame. Tra queste, merita di essere citata l'etnia dei Samburu, la quale deve la sua discendenza da altre popolazioni autoctone e che solo in epoca recente si è formata come tribù distinta dalle altre.

CAPITOLO 2

CONFRONTO IN ETEROGENEITÀ

2.1 ETEROGENEITÀ

In alcune discipline scientifiche spesso è d'interesse stabilire se la distribuzione di una determinata variabile è più concentrata in una popolazione piuttosto che in un'altra, cioè se una certa popolazione è meno eterogenea rispetto all'altra.

Il concetto di eterogeneità è molto usato nella statistica descrittiva. Data una variabile nominale X, definita su un supporto finito $(A_1,\ldots..,A_K)$, avente distribuzione di probabilità $Pr\{X \in A_k\} = p_k \geq 0$, con k= 1,…..,K e $\Sigma_k\, p_k = 1$, l'eterogeneità è minima se la distribuzione della variabile osservata è degenere, cioè presenta una singola categoria con una frequenza relativa pari a 1 e tutte le altre con frequenza pari a 0.

D'altra parte, l'eterogeneità è massima se la variabile è equamente distribuita su tutte le categorie.

In generale, quindi, l'omogeneità è la predisposizione di un fenomeno statistico a manifestarsi sempre nella stessa modalità. Un insieme di unità statistiche, infatti, è perfettamente omogeneo se tutte le unità sono caratterizzate dalla stessa categoria. Se questo non accade, cioè se vengono evidenziate molteplici categorie nell'insieme delle unità statistiche, allora ci si trova in una situazione di eterogeneità, cioè di assenza di omogeneità.

Quindi, il grado di eterogeneità dipende ovviamente dal numero di categorie osservate così come dalle loro frequenze.

L'eterogeneità può essere associata non solo al concetto di concentrazione, ma anche a quello di diversità, che è l'attitudine di una variabile qualitativa ad assumere modalità

differenti. Quindi è direttamente associabile ai concetti di incertezza e di informazione. Infatti, nel caso di minima eterogeneità, anche l'incertezza di una decisione è minima e l'informazione derivabile dalla singola osservazione è massima. Al contrario, quando si ha massima eterogeneità, c'è massima incertezza sulle decisioni e c'è minima informazione derivabile dalla singola unità statistica.

Una scienza in cui l'eterogeneità risulta di rilevante interesse è, appunto, la genetica, specialmente per la valutazione della biodiversità.

In questo settore, gli studi si occupano di confrontare due popolazioni per verificare quale delle due presenta una differenziazione genetica maggiore, vale a dire un'eterogeneità più grande dal punto di vista delle combinazioni fenotipiche di certi fattori genetici.

2.2 I DATI

Nel 1975, C. Corrain svolse uno studio antropologico[1] su alcune popolazioni pastorali del Kenya.

Obiettivo di tale progetto era l'analisi del valore discriminativo di alcuni caratteri emotipologici osservati su queste popolazioni, omogenee dal punto di vista economico, ambientale e culturale. Per poter confrontare i caratteri degli abitanti del Kenya, vennero raccolti alcuni campioni sierologici di tribù, stanziate per la maggior parte nelle terre semiaride a nord del Kenya.

Nel complesso, quindi, furono analizzati 384 individui, così suddivisi per etnia: 109 Samburu, 81 Rendille, 63 Turkana, 63 Masai, 45 Ol Molo e 23 Kamba.

Dallo studio svolto da C. Corrain, è emerso che gli Ol Molo, come pure i Masai, i Rendille e i Samburu, sono una popolazione nomade che ha frequenti scambi genetici con

[1] Studio documentato in Corrain et al. (1977).

altre tribù, in quanto ciò fa parte del comportamento sociale e culturale di questi gruppi etnici. Infatti, tali popolazioni usano praticare la adozione o la compera delle donne delle tribù vicine (Rendille soprattutto, ma anche Ol Molo).

Al contrario, i Kamba riflettono una stretta endogamia verso le altre tribù, perciò tale popolazione effettua sporadici scambi con gli altri popoli, vivendo in una sorta di isolamento genetico, sconosciuto presso gli altri gruppi etnici. Per questo motivo, ci si aspetta che la popolazione Kamba sia caratterizzata da un'eterogeneità genetica più bassa rispetto a quella delle popolazioni Ol Molo, Masai, Rendille e Samburu, e quindi presenti una più alta concentrazione relativa di combinazioni fenotipiche.

Come già detto, per confrontare due popolazioni dal punto di vista genetico, bisogna analizzare le combinazioni fenotipiche di certi fattori genetici presenti nei rispettivi campioni.

In questo caso, i fattori studiati sono: il fattore Gm(1), il fattore Gm(2), il fattore Gm(4) e il fattore Gm(12). Ciascuna combinazione fenotipica è caratterizzata dalla presenza (indicata col simbolo +) o assenza (configurato col simbolo -) di uno o due o tre o di tutti i fattori considerati; quindi il numero delle possibili combinazioni fenotipiche è pari a 2^4, cioè 16 categorie nominali. I complessivi 384 individui, perciò, sono stati classificati tra le 16 combinazioni risultate. Le frequenze assolute osservate vengono evidenziate nella tabella riportata a pagina seguente (*Tabella 1*).

Tabella 1: Distribuzione delle frequenze assolute delle combinazioni fenotipiche presenti nelle sei tribù in esame.

					Frequenze osservate					
Classe	Gm (1)	Gm (2)	Gm (4)	Gm (12)	Samburu	Rendille	Turkana	Ol Molo	Kamba	Masai
1	+	+	+	+	8	14	10	12	0	18
2	-	+	+	+	5	0	1	1	0	1
3	+	-	+	+	11	32	12	8	6	12
4	+	+	-	+	7	6	8	2	1	2
5	+	+	+	-	2	0	0	0	0	1
6	-	-	+	+	7	1	0	1	0	1
7	+	+	-	-	2	0	0	1	0	1
8	-	+	-	+	9	1	0	0	0	0
9	-	+	+	-	3	0	0	0	0	0
10	+	-	+	-	6	0	0	2	0	2
11	+	-	-	+	20	21	27	8	15	15
12	+	-	-	-	5	0	1	6	0	5
13	-	+	-	-	4	0	0	0	0	0
14	-	-	+	-	1	1	0	0	0	0
15	-	-	-	+	12	1	2	3	1	0
16	-	-	-	-	7	4	2	1	0	5
nj					109	81	63	45	23	63

Come si può notare, l'assenza contemporanea dei quattro fattori Gm appare abbastanza rara in tutti i gruppi etnici, mentre la presenza di tutti i fattori discrimina abbastanza le sei tribù considerate.

L'allotipo Gm 1, 12 è sempre molto rappresentato, se non addirittura il più rappresentato in assoluto. È pure molto frequente l'allotipo Gm 1, 4, 12.

Le restanti combinazioni fenotipiche sembrano tutte alquanto poco presenti tra le sei popolazioni autoctone, ma se si osserva meglio il prospetto, oltre alle diversità riscontrate, se ne possono evidenziare altre più o meno apprezzabili. Ad esempio, con gli allotipi Gm 1, 2, 12 e Gm 12 si possono individuare gruppi di tribù più o meno omogenee tra loro: la prima combinazione fenotipica citata appare più presente negli indigeni Samburu, Rendille e Turkana; il singolo fattore Gm(12) invece risulta prevalente nella sola etnia Samburu.

In supporto a tali considerazioni, basta pensare che il campione composto dai 109 aborigeni Samburu presenta tutti i 16 fenotipi possibili; i 45 Ol Molo e i 63 Masai conside-

rati ne presentano 11; gli 81 Rendille offrono 9 fenotipi; i 63 Turkana presentano 8 combinazioni fenotipiche e, infine, i 23 autoctoni Kamba ne offrono solo 4.

In generale, quindi, si può dire che tanto più sono frequenti le varie combinazioni teoricamente possibili dei 4 fenotipi, tanto più un dato gruppo etnico può essere ritenuto eterogeneo da un punto di vista genetico.

Viceversa, se le osservazioni sono distribuite su poche combinazioni fenotipiche, il gruppo etnico può essere ritenuto geneticamente omogeneo, e ciò sarà tanto più vero quanto più le osservazioni saranno concentrate su un più ristretto numero di combinazioni fenotipiche.

Portando al limite questo ragionamento, si può dire che se un dato gruppo etnico presentasse una sola combinazione fenotipica delle 16 possibili, esso sarebbe del tutto omogeneo, dal punto di vista genetico, per quanto riguarda i fenotipi rilevati.

Per questi motivi, è ragionevole supporre che il gruppo etnico Samburu sia il più eterogeneo tra le sei etnie considerate, proprio per il fatto che le unità statistiche si ripartiscono tra tutte le 16 categorie considerate. Al contrario, i Kamba risultano essere la popolazione più omogenea dal punto di vista genetico, poiché gli individui sono caratterizzati da sole quattro combinazioni fenotipiche. Le restanti quattro etnie sembrano abbastanza somiglianti tra loro, in quanto contano un numero pressoché simile di fenotipi presenti. Tali ipotesi verranno avvalorate tramite l'analisi grafica riportata nel successivo paragrafo.

In questo ordine di idee, quindi, non è interessante tanto il tenere presente quali particolari combinazioni fenotipiche siano più frequenti, quanto piuttosto il loro numero e la loro concentrazione.

2.3 ANALISI GRAFICA PRELIMINARE

Si procede con un'analisi preliminare grafica, in modo da mettere in evidenza i gruppi etnici kenioti che maggiormente si differenziano dagli altri.

Per lo studio, si è utilizzato il diagramma di Pareto, che permette un facile e immediato confronto tra le popolazioni considerate.

Tale diagramma consiste infatti nell'affiancare all'istogramma della distribuzione relativa delle combinazioni fenotipiche osservate nelle popolazioni, risultante ordinando in modo decrescente le frequenze osservate, la spezzata della distribuzione cumulata delle frequenze degli stessi allotipi. In questo modo, la rappresentazione così ottenuta permette di valutare a colpo d'occhio le combinazioni fenotipiche maggiormente rilevanti in ogni popolazione e di quanto queste incidono. Infatti, tanto più lentamente cresce la spezzata della distribuzione cumulata, e le frequenze osservate dei fenotipi si distribuiscono su più colonne dell'istogramma, tanto più la rispettiva popolazione può essere ritenuta eterogenea dal punto di vista genetico.

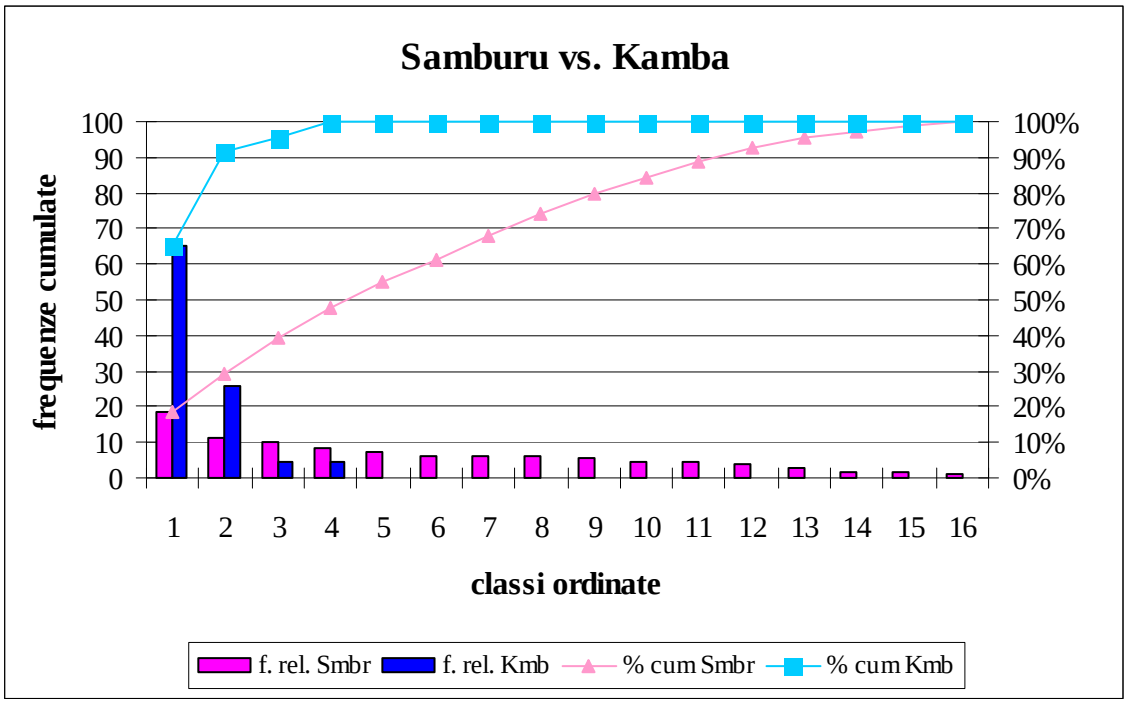
Al contrario, se le frequenze osservate si concentrano su poche classi categoriali, quindi su poche colonne dell'istogramma, tale gruppo etnico può essere considerato omogeneo, e sarà tanto più omogeneo quanto più si impenna la curva della distribuzione cumulata.

In riferimento alle considerazioni fatte nel precedente paragrafo, si è deciso di esordire nell'analisi grafica proprio con un confronto tra le popolazioni Samburu e Kamba.

Tabella 2: Distribuzione delle frequenze relative delle combinazioni fenotipiche nelle due popolazioni Samburu e Kamba.

	Classi ordinate																
Popolazione	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	
Samburu	18.3	11.0	10.1	8.3	7.3	6.4	6.4	6.4	5.5	4.6	4.6	3.7	2.8	1.8	1.8	0.9	100
Kamba	65.2	26.1	4.3	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100
Frequenze marginali	26.5	13.6	9.1	7.6	6.1	5.3	5.3	5.3	4.5	3.8	3.8	3.0	2.3	1.5	1.5	0.8	100

Figura 3: Frequenze percentuali cumulate per le classi ordinate delle popolazioni Samburu e Kamba.



Come si può vedere dal grafico in *Figura 3*, la spezzata relativa all'etnia Samburu cresce molto più gradualmente rispetto a quella della tribù Kamba, la cui funzione di distribuzione empirica raggiunge il 100% già con la quarta classe nominale. Ciò è a conferma delle ipotesi fatte precedentemente: i Samburu sono una popolazione piuttosto eterogenea, dovuto anche al loro comportamento sociale e alla pratica della compera delle donne delle altre tribù, di cui si è accennato a inizio capitolo. Al contrario, a seguito

della stretta endogamia verso le altre etnie, i Kamba risultano essere una popolazione abbastanza omogenea.

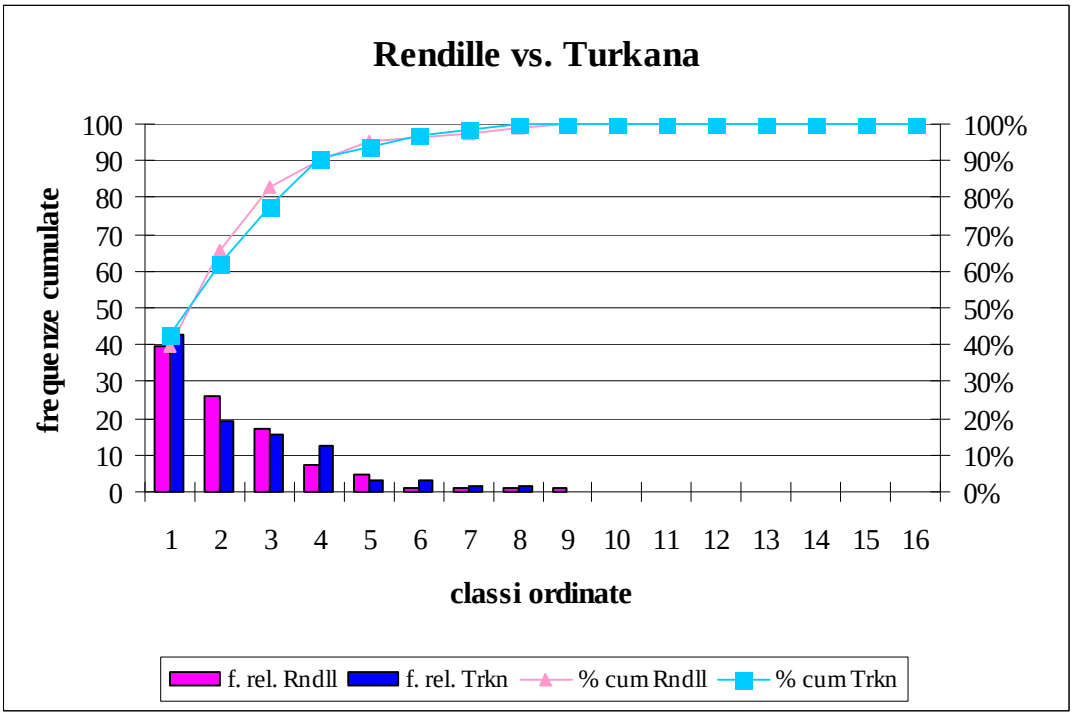
Si è ipotizzato, anche, che le rimanenti quattro etnie siano abbastanza somiglianti tra loro, in quanto presentano un analogo numero di combinazioni fenotipiche.

Si riporta dapprima il confronto grafico tra le popolazioni Rendille e Turkana; in seguito quello tra gli Ol Molo e i Masai.

Tabella 3: Distribuzione delle frequenze relative delle combinazioni fenotipiche nelle due popolazioni Rendille e Turkana.

	Classi ordinate																
Popolazione	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	
Rendille	39.5	25.9	17.3	7.4	4.9	1.2	1.2	1.2	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100
Turkana	42.9	19.0	15.9	12.7	3.2	3.2	1.6	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100
Frequenze marginali	41.0	22.9	16.7	9.7	4.2	2.1	1.4	1.4	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100

Figura 4: Frequenze percentuali cumulate per le classi ordinate delle popolazioni Rendille e Turkana.



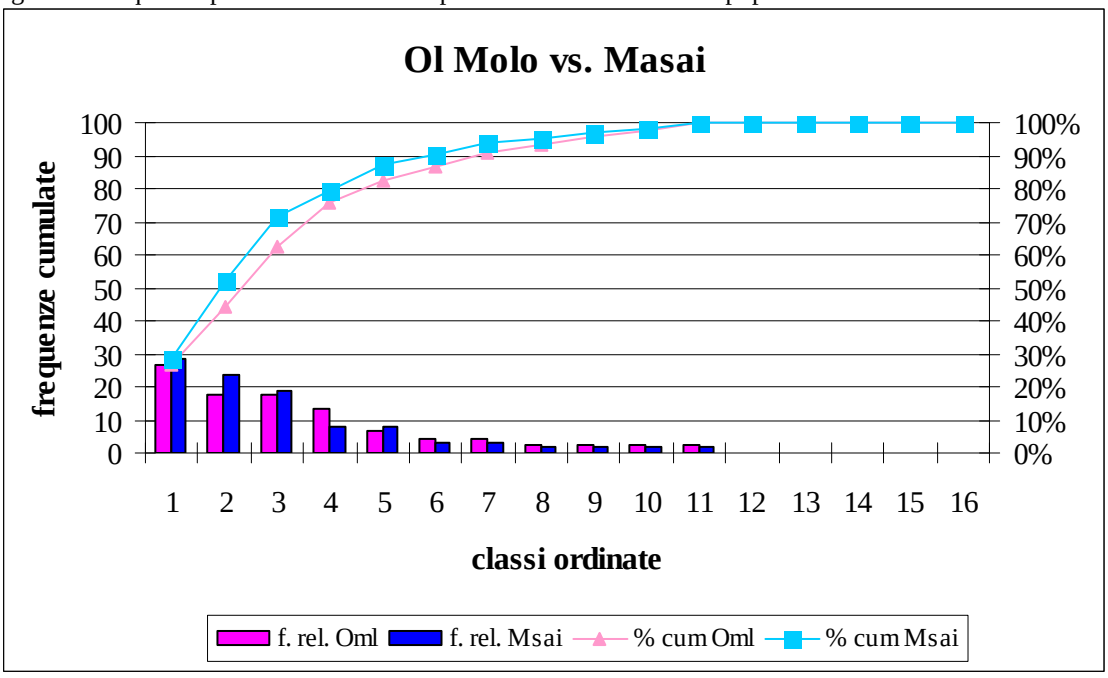
Le curve delle rispettive distribuzioni cumulate quasi combaciano ed entrambe crescono in modo graduale; ciò sta a significare che i due gruppi etnici possono essere considerati entrambi abbastanza omogenei dal punto di vista genetico, poichè non si differenziano molto tra loro.

Allo stesso modo, come si può notare dal successivo grafico, si possono ritenere geneticamente omogenei anche gli autoctoni Ol Molo e Masai. Le rispettive spezzate tendono a sovrapporsi e crescono più lentamente rispetto a quelle relative alle popolazioni Rendille e Turkana, precedentemente confrontate.

Tabella 4: Distribuzione delle frequenze relative delle combinazioni fenotipiche nelle due popolazioni Ol Molo e Masai.

	Classi ordinate																
Popolazione	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	
Ol Molo	26.7	17.8	17.8	13.3	6.7	4.4	4.4	2.2	2.2	2.2	2.2	0.0	0.0	0.0	0.0	0.0	100
Masai	28.6	23.8	19.0	7.9	7.9	3.2	3.2	1.6	1.6	1.6	0.0	0.0	0.0	0.0	0.0	0.0	100
Frequenze marginali	27.8	21.3	18.5	10.2	7.4	3.7	3.7	1.9	1.9	1.9	1.9	0.0	0.0	0.0	0.0	0.0	100

Figura 5: Frequenze percentuali cumulate per le classi ordinate delle popolazioni Ol Molo e Masai.

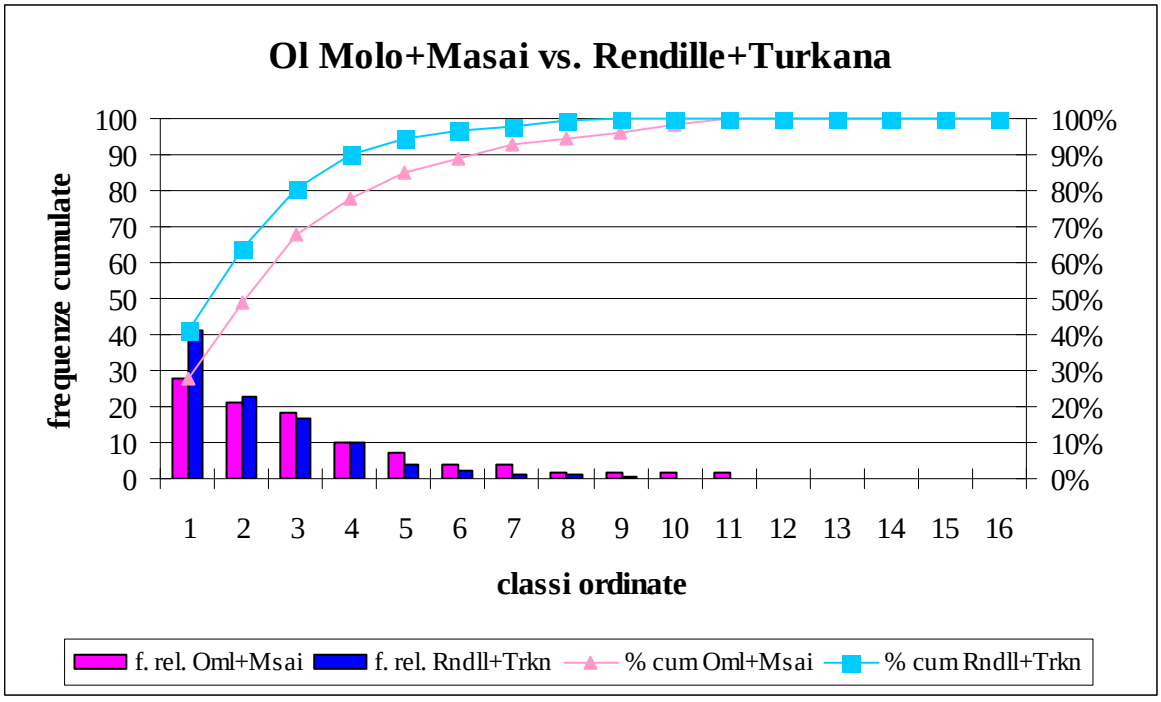


Facendo un ultimo confronto tra le due coppie di popolazioni appena analizzate (si veda il grafico *Figura 6*), si può notare che le curve delle rispettive distribuzioni cumulate hanno entrambe un progressivo andamento crescente e non si discostano molto l'una dall'altra. Le quattro etnie Ol Molo, Masai, Rendille e Turkana sembrano, quindi, somiglianti tra loro, dal punto di vista genetico.

Tabella 5: Distribuzione delle frequenze relative delle combinazioni fenotipiche nelle due popolazioni Ol Molo+Masai e Rendille+Turkana.

	Classi ordinate																
Popolazione	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	
Ol Molo+Masai	27.8	21.3	18.5	10.2	7.4	3.7	3.7	1.9	1.9	1.9	1.9	0.0	0.0	0.0	0.0	0.0	100
Rendille+Turkana	41.0	22.9	16.7	9.7	4.2	2.1	1.4	1.4	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100
Frequenze marginali	35.3	22.2	17.5	9.9	5.6	2.8	2.4	1.6	1.2	0.8	0.8	0.0	0.0	0.0	0.0	0.0	100

Figura 6: Frequenze percentuali cumulate per le classi ordinate delle popolazioni Ol Molo+Masai e Rendille+Turkana.



Resta infine un'ultima analisi da fare, quella che metta in comparazione il gruppo formato dalle quattro etnie aggregate con le rimanenti due popolazioni prese singolarmente: i Samburu, gruppo etnico più eterogeneo, e i Kamba, gruppo etnico più omogeneo.

Tabella 6: Distribuzione delle frequenze relative delle combinazioni fenotipiche nelle due popolazioni Samburu e Ol Molo+Masai+Rendille+Turkana.

	Classi																
Popolazione	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	
Samburu	18.3	11.0	10.1	8.3	7.3	6.4	6.4	6.4	5.5	4.6	4.6	3.7	2.8	1.8	1.8	0.9	100
Ol Molo+Masai+Rendille+Turkana	28.2	25.4	21.4	7.1	4.8	4.8	2.4	1.6	1.2	1.2	0.8	0.4	0.4	0.4	0.0	0.0	100
Frequenze marginali	25.2	21.1	18.0	7.5	5.5	5.3	3.6	3.0	2.5	2.2	1.9	1.4	1.1	0.8	0.6	0.3	100

Figura 7: Frequenze percentuali cumulate per le classi ordinate delle popolazioni Samburu e Ol Molo+Masai+Rendille+Turkana.

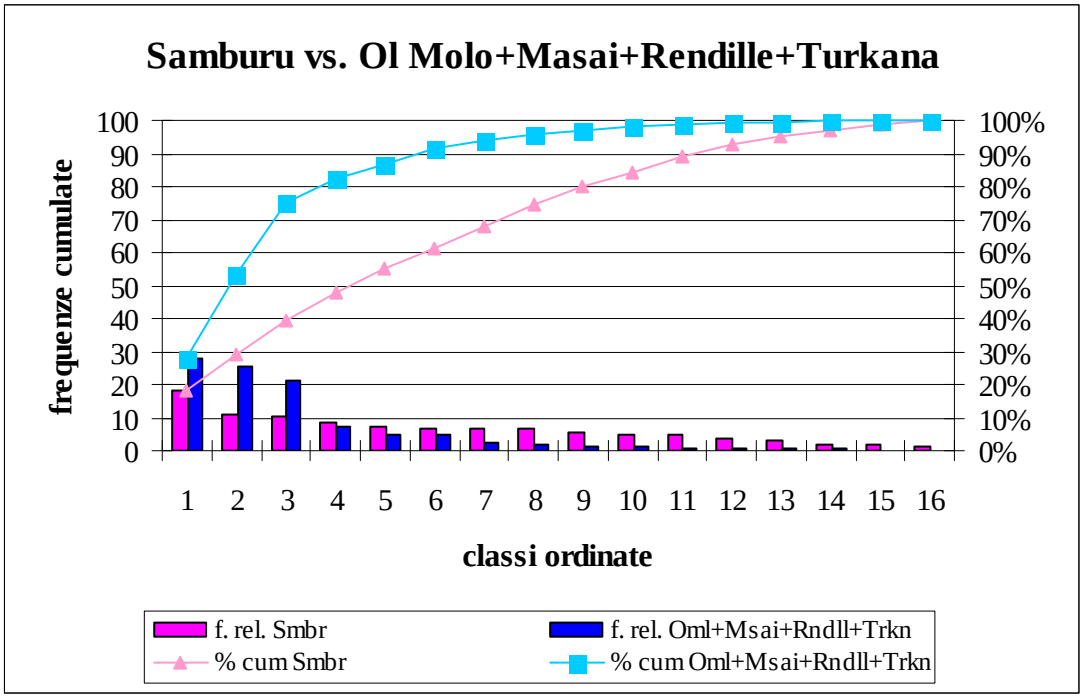
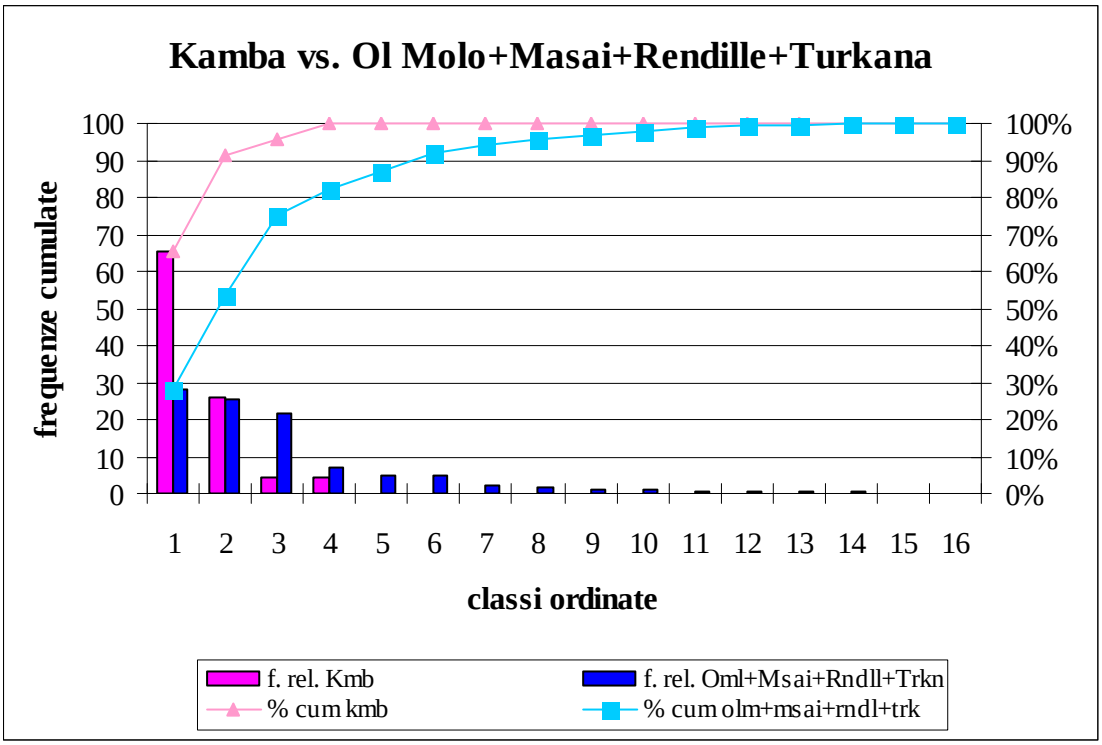


Tabella 7: Distribuzione delle frequenze relative delle combinazioni fenotipiche nelle due popolazioni Kamba e Ol Molo+Masai+Rendille+Turkana.

	Classi																
Popolazione	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	
Kamba	65.2	26.1	4.3	4.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100
Ol Molo+Masai+Rendille+Turkana	28.2	25.4	21.4	7.1	4.8	4.8	2.4	1.6	1.2	0.8	0.8	0.4	0.4	0.4	0.0	0.0	100
Frequenze marginali	31.3	25.5	20.0	6.9	4.4	4.4	2.2	1.5	1.1	1.1	0.7	0.4	0.4	0.4	0.0	0.0	100

Figura 8: Frequenze percentuali cumulate per le classi ordinate delle popolazioni Kamba e Ol Molo+Masai+Rendille+Turkana.



Dai grafici sopra riportati si può dedurre, a titolo conclusivo, che la popolazione Kamba può essere ritenuta la più omogenea tra le sei tribù considerate (si veda *Figura 8*); gli individui sono concentrati su poche combinazioni fenotipiche, a dimostrazione del numero ridotto di colonne prominenti nell'istogramma e della maggior pendenza della curva della distribuzione cumulata delle frequenze relative.

Viceversa, gli autoctoni Samburu costituiscono il gruppo etnico più eterogeneo in assoluto (si veda *Figura 7*). Sono frequenti un gran numero di combinazioni fenotipiche teoricamente possibili: la funzione di distribuzione empirica raggiunge il livello 100% con la sedicesima e ultima classe nominale, e la curva delle frequenze cumulate ha un andamento molto più lento e progressivo rispetto a quello della spezzata del cluster formato dalle quattro etnie aggregate; inoltre, le frequenze si ripartiscono su più classi categoriali, dando così alle colonne una struttura più uniforme.

Questa analisi grafica preliminare, in sostanza, ha permesso di individuare tre gruppi principali di tribù pressoché omogenee al loro interno e differenziate tra di loro: un primo, più eterogeneo, formato dalla singola etnia Samburu; un secondo, più omogeneo, rappresentato dalla sola popolazione Kamba; un terzo composto dai quattro gruppi etnici Ol Molo, Masai, Rendille e Turkana.

Si vuole far notare che tale suddivisione delle predette etnie in gruppi, omogenei al loro interno ma dissomiglianti tra loro, è stata ottenuta attraverso una semplice analisi grafica e, quindi, non è risolutiva. Infatti, nel capitolo 5, dedicato al raggruppamento delle etnie svlto tramite la procedura di Bonferroni-Holm, si vedrà che le quattro tribù qui considerate aggregate, Ol Molo, Masai, Rendille e Turkana, verranno divise tra loro e assegnate a due distinti gruppi. Da una parte gli Ol Molo e i Masai, dall’altra i Rendille i Turkana.

2.4 INDICI DI ETEROGENEITÀ

È stato detto, a inizio capitolo, che l'omogeneità è la predisposizione di un fenomeno statistico a manifestarsi sempre nella stessa modalità. Se questo non accade, allora ci si trova in una situazione di eterogeneità.

Un indice che traduca sinteticamente il grado di eterogeneità di un fenomeno osservato deve avere le seguenti caratteristiche:

1. Assumere il minimo valore in presenza di massima omogeneità, cioè quando il fenomeno sotto studio si manifesta con una singola categoria;
2. Assumere valori sempre più grandi man mano che ci si allontana dalla situazione degenere e ci si avvicina all'equidistribuzione;
3. Assumere il massimo valore in presenza di distribuzione uniforme.

Tra i vari indicatori di eterogeneità proposti in letteratura, per questo lavoro sono stati usati gli indici di Gini, di Shannon e di Renyi.

Dapprima si è testata la diversità in distribuzione tra due popolazioni:

$$H_0 : P(X) = P(Y) \qquad \text{vs.} \qquad H_1 : P(X) \neq P(Y)$$

Molto usata nelle applicazioni di confronto di eterogeneità di popolazioni è la statistica di Pearson, definita da:

$$X^2 = \sum_{k=1}^{K}\sum_{j=1}^{C} \frac{\left(f_{kj} - \frac{f_{k\cdot} f_{\cdot j}}{n} \right)}{\frac{f_{k\cdot} f_{\cdot j}}{n}}$$

dove K è il numero di classi nominali; C è il numero di campioni considerati; f_{kj} sono le frequenze assolute della k-esima classe nominale e j-esimo campione, cioè il numero di elementi del campione j, con j=1,2, che possiedono la k-esima combinazione fenotipica, con k=1,…..,K; con $f_{k\cdot}$ e $f_{\cdot j}$ vengono indicate le frequenze marginali di riga e di colonna rispettivamente[2].

[2] $f_{\cdot j}$ corrisponde alla numerosità campionaria n_j.

La statistica test X^2 ha distribuzione nulla asintotica chi-quadrato con (K-1)·(C-1) gradi di libertà.

Sempre per verificare la dissomiglianza in distribuzione tra due popolazioni, è stata utilizzata una seconda statistica test, qui denominata X^2_{AD}, definita in questo modo:

$$X^2_{AD} = \sum_{k=1}^{K} \frac{\left(\frac{f_{k1}}{n_1} - \frac{f_{k2}}{n_2} \right)^2}{\frac{f_{k.}}{n}\left(1 - \frac{f_{k.}}{n}\right)}$$

dove f_{k1} e f_{k2} sono le frequenze relative per la k-esima combinazione fenotipica nei due rispettivi campioni; n_1 e n_2 sono le numerosità campionarie delle due popolazioni prese in considerazione; n è la numerosità totale campionaria ($n= n_1+n_2$) e $f_{k\cdot}$ rappresenta sempre le frequenze marginali di riga. Essa corrisponde alla divergenza secondo Anderson-Darling tra le distribuzioni di frequenza.

Infine, si è deciso di verificare se, oltre ad una diversità in distribuzione tra i due campioni, ci fosse anche una vera e propria dominanza in eterogeneità da parte di una popolazione il cui comportamento favorisce frequenti scambi genetici, rispetto ad un'altra la cui vita sociale e culturale, invece, si sviluppa solo all'interno della propria etnia. Obiettivo di questo lavoro, infatti, è verificare l'ipotesi che una popolazione nomade presenti una eterogeneità genetica superiore a quella di una popolazione non nomade.

Di conseguenza, è stato definito il seguente sistema d'ipotesi:

$$H_0 : \text{Het}(X) = \text{Het}(Y) \qquad \text{vs.} \qquad H_1 : \text{Het}(X) > \text{Het}(Y)$$

Come accennato precedentemente, tra i diversi indicatori presenti in letteratura, per lo studio seguente sono stati utilizzati gli indici di Gini, di Shannon e di Renyi.

Per testare la dominanza in eterogeneità tra le due popolazioni, è ragionevole usare come statistica test la differenza dei due rispettivi indicatori campionari.

Ad esempio, facendo riferimento solo ad un singolo campione, l'indice di eterogeneità proposto da Gini, per una variabile categoriale X che assume valori nelle K classi nominali, con frequenze osservate f_k, k=1,….,K, è così determinato:

$$G = 1 - \sum_{k=1}^{K} f_k^2$$

e la sua versione normalizzata risulta:

$$G' = \frac{G(K-1)}{K}$$

Per il confronto in eterogeneità tra due popolazioni, si userà come statistica test la differenza dei due indici di Gini campionari, G_1 e G_2, cioè:

$$T_G = G_1 - G_2 = \sum_{k=1}^{K} \left(f_{2(k)}{}^2 - f_{1(k)}{}^2 \right)$$

L'indice di Shannon, altresì chiamato, in teoria dell'informazione, col nome di indice di entropia di una distribuzione, viene invece così calcolato:

$$S = -\sum_{k=1}^{K} f_k \log(f_k)$$

Con log(·) si intende il logaritmo naturale e si è definito che 0log(0) = 0.

La versione normalizzata dell'indice di diversità di Shannon è:

$$S' = \frac{S}{\log(K)}$$

Anche in questo caso, la statistica test, basata sull'indice di Shannon e sulla differenza degli indici campionari S_1 e S_2, è

$$T_S = S_1 - S_2 = \sum_{k=1}^{K} \left[f_{2(k)} \log\left(f_{2(k)}\right) - f_{1(k)} \log\left(f_{1(k)}\right) \right]$$

Sempre facendo riferimento ad un singolo campione, un indice generalizzato, ottenuto anch'esso dal campo della teoria dell'informazione, è l'indice di entropia generalizzato di ordine α proposto da Renyi e così definito:

$$R_\alpha = \frac{1}{1-\alpha} \log\left(\sum_{k=1}^{K} f_k^\alpha \right)$$

Per $\alpha \neq 1$, R_α è funzione decrescente di α. Al variare di α, si possono ottenere diversi indici di eterogeneità; se ne citano tre tra quelli più frequentemente usati:

$$R_1 = \lim_{\alpha \to 1} R_\alpha = -\sum_{k=1}^{K} f_k \log(f_k) = S$$

$$R_2 = \lim_{\alpha \to 2} R_\alpha = -\log\left(\sum_{k=1}^{K} f_k^2 \right) = -\log(1-G)$$

$$R_\infty = \lim_{\alpha \to \infty} R_\alpha = -log\left[sup_{k=1,\ldots,K}(f_k)\right]$$

È evidente la relazione tra l'indice R_α e l'ordine α, come si può vedere dal forte legame esistente tra l'indice generalizzato di Renyi e gli indici di Shannon (R_1 = S) e di Gini (R_2 = - log(1-G)).

Di conseguenza, per quanto riguarda l'indice generalizzato di Renyi, per risolvere il sistema d'ipotesi di dominanza in eterogeneità per due campioni, è stato preso in considerazione, tra gli ultimi tre presentati, solo l'indicatore R_∞.

Quindi, la statistica test utilizzata per il confronto in eterogeneità che calcola la differenza tra $R_{\infty 1}$ indice di Renyi sul primo campione, e $R_{\infty 2}$ indice sul secondo campione, è:

$$T_{R\infty} = R_{\infty 1} - R_{\infty 2} = log\left(max_k\, f_{2(k)}\right) - log\left(max_k\, f_{1(k)}\right)$$

Le statistiche test utilizzate per lo studio di dominanza in eterogeneità, basate sugli indici di Gini, Shannon e Renyi appena menzionati, saranno meglio descritte nel capitolo successivo.

Capitolo 2

CAPITOLO 3

TEST NON PARAMETRICI

3.1 STATISTICA NON PARAMETRICA

Uno degli scopi fondamentali della statistica è quello di predisporre metodologie che consentano di pervenire a corrette conclusioni a proposito di ipotesi o teorie formulate nei riguardi di qualche carattere manifestato dalle unità che formano una certa popolazione.

Negli ultimi tempi si sono ulteriormente sviluppate tecniche definite “non parametriche”, di importanza pari a quella delle note tecniche “parametriche”. Queste ultime richiedono, per la loro applicazione, che sia nota e che venga specificata la distribuzione della popolazione di riferimento. Le tecniche non parametriche, al contrario, non richiedono la precisazione di questa condizione e offrono il vantaggio di essere applicabili quando non si ha conoscenza della distribuzione di provenienza dei dati.

Tali procedure spesso sono abbastanza agevoli ad impiegarsi e, pur essendo talvolta meno potenti di quelle parametriche quando è richiesta la condizione di specificare la distribuzione di appartenenza dei dati, risultano insostituibili quando, invece, la stessa condizione viene a mancare.

Di conseguenza, non essendoci la cognizione della vera distribuzione dei dati relativi alle combinazioni fenotipiche presenti nelle sei tribù keniote, si fa uso delle tecniche non parametriche.

Il metodo proposto, quindi, consiste nel determinare statistiche test basate sulle probabilità attribuite alle diverse classi nominali e su test non parametrici.

È bene notare che, in un problema a due campioni, le probabilità delle due distribuzioni confrontate sono non note e verranno stimate tramite le frequenze relative; le stime di massima verosimiglianza di p_{kj}, con k=1,…..,K e j=1,2 saranno, perciò, così determinate $\hat{p}_{kj} = \frac{f_{kj}}{n_j}$, dove f_{kj} rappresenta la frequenza assoluta osservata nella classe k-esima del j-esimo campione, cioè è il numero di soggetti del campione j che possiedono la k-esima combinazione fenotipica; n_j è la numerosità del campione j.

Di conseguenza, proprio per questa mancanza di informazioni sulle probabilità, l'ordinamento di queste può essere valutato solo sulla base di dati campionari. Per questo motivo, le soluzioni proposte non possono essere esatte, ma solo approssimate, in quanto esatte asintoticamente.

Nel capitolo precedente, l'eterogeneità è stata trattata principalmente da un punto di vista descrittivo, con l'introduzione di alcuni indici che misurano il grado di eterogeneità di una distribuzione di frequenza in un certo insieme di unità statistiche.

D'ora in poi, l'argomento verrà trattato invece da un punto di vista inferenziale, confrontando l'eterogeneità campionaria di una variabile categoriale X in due popolazioni. Nel far questo verranno utilizzati gli indici descritti nel capitolo 2.

3.2 TEST DI PERMUTAZIONE

Realizzazioni della variabile categoriale X, oggetto di studio, sono le frequenze di 16 diverse combinazioni fenotipiche presenti in alcuni campioni sierologici di tribù stanziate nel nord del Kenya. Il supporto di tale variabile X, quindi, è partizionato in K=16 classi nominali (A_1,…..,A_K), appunto le 16 possibili combinazioni fenotipiche, e con $P=\{p_k, k=1,…..,K\}$ viene indicata la sottostante distribuzione della popolazione.

Si considerino due campioni indipendenti, ciascuno composto da osservazioni indipendenti e identicamente distribuite, i cui rispettivi dati campionari verranno indicati con $X_j=\{X_{ij}, i=1,\ldots..,n_j\}$, con j=1,2.

L'intero insieme di dati, quindi, è ottenuto dall'unione delle osservazioni appartenenti ai due distinti campioni, indicato con $X = X_1 \overset{+}{\cup} X_2$, dove $\overset{+}{\cup}$ simboleggia appunto l'aggregazione, concatenamento, dei due dataset X_1 e X_2.

Nel caso di variabili categoriali, i dati osservati vengono normalmente espressi tramite frequenze assolute $\left\{ f_{kj} = \sum_{i \le n_j} I(X_{ij} \in A_k), k = 1,\ldots,K; j = 1,2 \right\}$ dove con I si denota la funzione indicatrice che assume i valori: $I = \begin{cases} 1 & se\ X_{ij} \in A_k \\ 0 & altrimenti \end{cases}$.

Su tale dataset X, contenente le frequenze assolute osservate, si calcolano prima i valori osservati T° delle statistiche test necessarie per il tipo di verifica di ipotesi che si vuole testare. In seguito, occorre calcolare le possibili permutazioni di dati e ricavare, per ognuna di queste, il dataset permutato X^* che verrà utilizzato per il calcolo dei valori dei test permutati T^*.

Infatti, l'analisi di permutazione si basa, come dice il nome stesso, su test di permutazione, applicazioni che vanno dall'insieme campionario contenente tutte le permutazioni dei dati osservati, all'insieme dei numeri reali, $T : \chi^n_{/X} \to \Re^1$ e tali test sono, in generale, non distorti.

Per poter ottenere le permutazioni è stato utilizzato, tramite il pacchetto statistico R, il generatore di pura aleatorietà 'runif', col quale sono stati ottenuti n valori ragionevolmente considerabili come realizzazioni indipendenti di una variabile casuale univariata con legge U(0,1).

Sembra esserci una contraddizione fra l'aleatorietà richiesta per tali valori e il fatto che il calcolatore deve produrli attraverso un algoritmo deterministico. Ma, appunto per questo, i valori u_i^*, con i=1,…..,n, sono solo dei numeri pseudo-casuali. Tuttavia, questi ultimi sono pressoché indistinguibili da effettive osservazioni da un'uniforme U(0,1).

Di conseguenza, tali valori forniti dall'algoritmo possono essere considerati dei loro validi sostituti.

Dapprima, il generatore dei numeri casuali uniformi è stato inizializzato, tramite l'istruzione 'RNGkind', con la versione di Mersenne-Twister, un tipo di generatore avente periodo $2^{19937-1}$, il quale assicura, con una certa tranquillità, di evitare il problema della ciclicità delle successioni di realizzazioni pseudo-casuali.

Mediante il calcolo della permutazione delle etichette (1,…..,n), indicata con $\left(u_1^*,\ldots,u_n^*\right)$, si è così in grado di ottenere lo spazio campionario di permutazione $\chi^n_{/X}$, contenente tutte le possibili n! permutazioni $X^* = \left\{X\left(u_i^*\right), i = 1,\ldots,n\right\}$ corrispondenti all'insieme originario di dati osservati X. Nel caso in cui la cardinalità dello spazio campionario di permutazione sia molto elevata, si effettua un campionamento casuale dall'insieme di tutte le possibili permutazioni attraverso la procedura Monte Carlo Condizionata (CMC) che estrarrà, appunto, B insiemi casuali di permutazione dello spazio campionario. Definito, pertanto, il numero B di permutazioni dei dati originari, è possibile calcolare le frequenze assolute di permutazione: $\left\{ f^*_{kj} = \sum_{i \le n_j} I\left(X^*_{ij} \in A_k\right), k = 1,\ldots,K;\ j = 1,2 \right\}$ dove, nel caso in cui j=1, $X_{i1}{}^*$ sono i dati permutati del primo campione, $X^*_{i1} = X\left(u_i^*\right)$ per $i \le n_1$, con n_1 che indica, appunto, la numerosità del primo campione; nel caso in cui j=2, $X_{i2}{}^*$ sono quelli relativi al secondo campione, $X^*_{i2} = X\left(u_i^*\right)$ per $n_1 < i \le n$, con n che indica la numerosità totale.

Aggregando in un'unica tabella le frequenze dell'uno e dell'altro campione, si ottiene così il dataset permutato contenente, appunto, le frequenze osservate permutate per ciascuna classe categoriale A_K.

È importante notare che, in questo modo, le frequenze marginali sono permutazionalmente invarianti, in quanto $f_{.k} = f_{1k} + f_{2k} = f_{1k}{}^* + f_{2k}{}^* = f_{.k}{}^*$. Quindi le frequenze marginali di riga, calcolate sui dati originari, saranno equivalenti alle stesse calcolate, però, sui dati dei due campioni permutati.

Su ciascuna di queste tabelle verranno calcolati i valori dei test permutati T^*. Di conseguenza, calcolando questi valori per tutte le possibili permutazioni X^* dell'insieme di dati osservati X, si ottiene una distribuzione di permutazione delle statistiche test utilizzate per il sistema d'ipotesi che è oggetto di verifica.

In seguito, per ogni statistica test, è possibile calcolare il corrispondente livello di significatività osservato. Infatti, se B è il numero di permutazioni casuali considerato, il p-value, associato ad una particolare statistica, è dato da $\lambda_T = \frac{\#\left(T^* \geq T^\circ \mid X\right)}{B}$, dove con T si indica una delle statistiche test utilizzate nello studio di permutazione e con $\#(T^* \geq T^\circ \mid X)$ si indica il numero di volte che i valori di permutazione del test T sono non inferiori al valore osservato dello stesso test, condizionatamente al dataset X. Dato un generico sistema d'ipotesi, il calcolo del livello di significatività osservato di un test può essere visto come una prima fase, che non vincola ad una conclusione definitiva, in un processo di valutazione dell'evidenza empirica contro H_0. In una fase successiva, poi, si potrà utilizzare il livello di significatività osservato per accettare o rifiutare l'ipotesi nulla in un test con il livello di significatività fissato α più conveniente.

Secondo la regola generale, infatti, se il p-value λ_T risulta inferiore o uguale ad un livello di significatività fissato α, l'ipotesi nulla H_0, del sistema oggetto d'esame, verrà rifiutata a favore dell'ipotesi alternativa H_1. La scelta del livello di significatività α dipende da quanto forte deve essere l'evidenza empirica per lasciar cadere l'ipotesi nulla. Usualmente, il livello α del 5% può essere considerato l'opzione di default.

3.3 TEST SU IPOTESI DI DISSOMIGLIANZA IN DISTRIBUZIONE

Quando si desidera verificare sperimentalmente un'ipotesi statistica, tipicamente i dati non possono contrastare l'ipotesi con la forza di una contraddizione. Possono, però, essere interpretati come in disaccordo con l'ipotesi ad un livello più o meno elevato. Un livello elevato di disaccordo induce ad optare per un'altra ipotesi adatta, mentre un livello moderato di disaccordo induce a ritenere l'ipotesi statistica adeguata alla luce dei dati campionari disponibili.

Con la verifica d'ipotesi, quindi, si vuole solo saggiare la conformità dei dati ad una particolare ipotesi formulata sui dati stessi.

Un tipico test, riguardante un problema a due campioni, consiste nel verificare il seguente sistema d'ipotesi

$$H_0 : X_1 \overset{d}{=} X_2 \qquad \text{vs.} \qquad H_1 : X_1 \overset{d}{\neq} X_2$$

ossia

$$H_0 : P(X_1) = P(X_2) \qquad \text{vs.} \qquad H_1 : P(X_1) \neq P(X_2)$$

È bene osservare che in uno studio univariato a due campioni, sia l'insieme delle frequenze marginali $\{n_1,n_2,f_{\cdot 1},\ldots..,f_{\cdot k}\}$, che l'intero dataset X e ogni sua permutazione X^* sono statistiche sufficienti equivalenti, poiché contengono la stessa quantità di informazione sulla distribuzione P quando l'ipotesi nulla $H_0 : P(X_1) = P(X_2)$ è vera.

Infatti, si sa che l'insieme di tutti i dati X rappresenta sempre una statistica sufficiente per una qualsiasi distribuzione P; inoltre, ad eccezione di una permutazione irrilevante delle etichette, c'è una relazione uno-a-uno tra il dataset completo X e l'insieme delle frequenze marginali $\{n_1,n_2,f_{\cdot 1},\ldots..,f_{\cdot k}\}$ che, come già spiegato, sono permutazionalmente invarianti: $\{n_1,n_2,f_{\cdot 1},\ldots..,f_{\cdot k}\}=\{n_1,n_2,f_{\cdot 1}^*,\ldots..,f_{\cdot k}^*\}$.

Una conseguenza del fatto che i dati globali X siano una statistica sufficiente, detta proprietà di sufficienza di X in H_0, è che la distribuzione delle frequenze osservate, condizionata alla statistica sufficiente X, è indipendente dalla sottostante distribuzione P, cioè $Pr[(f_{kj};P)|X]=Pr[f_{kj}|X]$con k=1,…..,K, j=1,2.

Oltre a ciò, vale la pena osservare che le permutazioni X^* sono equiprobabili: $Pr[X=x|X]=Pr[X^*=x|X]$, per ogni x appartenente all'insieme complessivo delle permutazioni.

Considerato il sistema d'ipotesi riportato nella pagina precedente, l'ipotesi nulla indica che i due campioni hanno la stessa distribuzione e sono indipendenti. Questo implica che non c'è differenza nell'osservare dati da un campione o da un altro, quindi c'è scambiabilità tra i due gruppi di dati.

Proprio questo principio della scambiabilità dei dati, assicurato dall'ipotesi nulla, ci offre la possibilità di reperire lo spazio campionario di permutazione $\chi^n_{/X}$ con cui comparare la situazione effettivamente osservata. Una conseguenza importante della scambiabilità sotto H_0 è il conseguimento di soluzioni inferenziali esatte.

Al contrario, sotto l'ipotesi alternativa, la statistica sufficiente è data dalla coppia di insiemi di dati campionari (X_1,X_2); di conseguenza, le permutazioni non sono equiprobabili e i dati non sono più scambiabili tra campioni perché, con l'ipotesi alternativa, si sta assumendo che le due distribuzioni sono diverse.

Talvolta la scambiabilità può essere soddisfatta solo approssimativamente, perché i dati si possono permutare tra loro unicamente all'interno dei due campioni distinti. In questi casi è importante valutare il grado di approssimazione per dimensioni campionarie finite.

Come esposto nel capitolo precedente, per testare l'ipotesi di dissomiglianza in distribuzione sono state usate la statistica test X^2 di Pearson e la statistica test che è stata denominata X^2_{AD}

$$X^2_{AD} = \sum_{k=1}^{K} \frac{\left(\frac{f_{k1}}{n_1} - \frac{f_{k2}}{n_2} \right)^2}{\frac{f_{k.}}{n}\left(1 - \frac{f_{k.}}{n}\right)}$$

corrispondente alla divergenza, secondo Anderson-Darling, tra le distribuzioni di frequenza. Infatti, per variabili categoriali non ordinabili, l'unica ipotesi alternativa possibile è quella di diversità in distribuzione, in quanto le alternative unilaterali sono plausibili solo per variabili ordinabili e, in questi casi, la statistica test principalmente usata è proprio la X^2 con distribuzione nulla asintotica chi-quadrato.

A tal proposito è stato scritto un programma[3] in grado di risolvere, attraverso le tecniche di permutazione, il test di dissomiglianza in distribuzione. All'interno del programma, perciò, sono state inserite delle procedure per la lettura dei dati, per la creazione delle matrici contenenti le frequenze osservate, sia ordinate che non, necessarie per l'applicazione delle statistiche test; inoltre, sono state scritte delle funzioni per il calcolo del valore osservato assunto dai test applicati ai dati originari e per il calcolo del valore assunto dagli stessi test per ciascuna delle permutazioni dei dati ottenute. Infine, per valutare se vi è somiglianza o diversità in distribuzione tra le due popolazioni confrontate sono state inserite, all'interno del programma, delle funzioni per il computo dei p-value associati a ciascuno dei test statistici utilizzati.

Tale codice è stato scritto utilizzando R, un pacchetto statistico, ed è stato realizzato in modo tale da poter essere eseguito con qualsiasi file dati di estensione '.txt'. Infatti, è stato creato appositamente in forma generica, facendo sì che fossero le stesse funzioni interne al programma a contare il numero di classi nominali contenute in esse, in modo tale da rendere, appunto, il programma riutilizzabile per altri dataset di dimensioni anche diverse da quelle dei file dati usati per questo specifico studio. Inoltre, è un programma che richiede una certa interazione con l'ambiente esterno, in quanto viene data

[3] Il programma 'TSD2SER.txt' è allegato in Appendice.

all'utente, tramite l'apertura di un'opportuna finestra di output, la possibilità di scegliere lo specifico file dati da aprire e, sempre a lui, è richiesto l'inserimento del numero di simulazioni condizionate da realizzare per i test di permutazione.

Il programma è stato mandato in esecuzione sui file dati contenenti le frequenze di combinazioni fenotipiche osservate nei campioni kenioti, confrontati a coppie. In particolare, per questo tipo di problema a due campioni, si è voluto verificare se vi è somiglianza o diversità tra le distribuzioni delle 15 possibili coppie di popolazioni ottenibili con le sei tribù oggetto di studio in questa tesi. Per far ciò, sono state inserite, all'interno del programma, le funzioni di calcolo delle statistiche test X^2 e X^2_{AD}. Inoltre, poiché la cardinalità dello spazio campionario di permutazione è molto elevata, sono state effettuate solo B permutazioni, estratte casualmente da tutte le n! possibili. Nello specifico, per ciascuna delle coppie esaminate, sono state eseguite 2000 permutazioni dei dati.

È bene sottolineare che una statistica test permette di evidenziare se sia più ragionevole accettare l'ipotesi nulla H_0, o rifiutarla in favore di H_1, tenuto conto dei dati campionari di cui si dispone. Accettare e rifiutare, quindi, non vanno intese nel senso di stabilire o computare la verità dell'ipotesi nulla, ma solo come posizioni conoscitive provvisoriamente prese, alla luce dei dati disponibili. Infatti, data la natura campionaria dei dati, certi valori di una statistica test possono essere del tutto conformi ad H_0. Altri valori, benché non impossibili da osservare se H_0 è vera, potrebbero indicare che il campione esaminato è troppo lontano dalla forma o configurazione prevista per l'ipotesi nulla, a favore di quanto, invece, viene contemplato dall'ipotesi alternativa. Dunque, un test statistico non è un meccanismo di discriminazione infallibile tra H_0 e H_1.

Nella tabella riportata nella pagina seguente vengono evidenziati i p-value associati ai test X^2 e X^2_{AD}, calcolati per le coppie di campioni kenioti esaminati. È da tener presente che gli specifici livelli di significatività osservati, contenuti in tale tabella, possono differire da quelli che si potrebbero ottenere con esecuzioni successive del programma in

quanto, come già detto in precedenza, le permutazioni delle n etichette sono state ottenute tramite un generatore di pura aleatorietà; di conseguenza, con ulteriori permutazioni si possono ottenere p-value sempre differenti associati ad una stessa statistica test pur concordando tutti, per la specifica statistica, sull'accettazione o il rifiuto dell'ipotesi nulla di uguaglianza in distribuzione.

Tabella 8: p-value delle statistiche X^2 e X^2_{AD} per il test di dissomiglianza in distribuzione.

	p-value	
Campioni confrontati	X^2	X^2_{AD}
Samburu-Kamba	0.007	0.001
Samburu-Turkana	0.000	0.000
Samburu-Rendille	0.000	0.000
Samburu-Ol Molo	0.033	0.026
Samburu-Masai	0.000	0.000
Rendille-Turkana	0.076	0.041
Ol Molo-Masai	0.741	0.762
Rendille-Ol Molo	0.001	0.001
Rendille-Masai	0.005	0.006
Turkana-Ol Molo	0.011	0.009
Turkana-Masai	0.014	0.013
Kamba-Turkana	0.300	0.218
Kamba-Rendille	0.047	0.014
Kamba-Ol Molo	0.002	0.000
Kamba-Masai	0.005	0.002

Come si può osservare, quasi tutti i confronti tra l'etnia Samburu e le altre diverse tribù hanno associati dei p-value, sia per la statistica X^2 che per la X^2_{AD}, che portano al rifiuto dell'ipotesi nulla di uguaglianza in distribuzione, con un livello di significatività fissato α=5%. Questo significa che la distribuzione della popolazione Samburu è dissomigliante da quella delle altre popolazioni e ciò potrebbe essere dovuto principalmente al fatto che è l'unica etnia a presentare tutte le 16 combinazioni possibili di fenotipi, come spiegato nel paragrafo 1.2. Tale etnia, quindi, sembra distinguersi particolarmente dalle altre dal punto di vista genetico. Infatti, come si vedrà nel prossimo paragrafo e nel capitolo

5, in cui è illustrata la procedura di Bonferroni-Holm per la suddivisione delle etnie in gruppi, tale tribù risulterà addirittura maggiormente eterogenea rispetto alle altre.

In seguito, seguendo l'analisi grafica esposta nel capitolo 2, è stata confrontata la coppia di campioni Rendille e Turkana prima, e successivamente quella Ol Molo e Masai. Le statistiche test hanno calcolato, per entrambe le comparazioni, dei p-value che portano all'accettazione dell'ipotesi nulla; ciò significa che le tribù rapportate sono simili in distribuzione all'interno di ciascuna coppia, ma le due coppie di etnie confrontate tra loro sono dissomiglianti. Difatti, i seguenti confronti, Rendille-Ol Molo, Rendille-Masai, Turkana-Ol Molo e Turkana-Masai, hanno generato dei p-value inferiori o prossimi all'1%, inducendo quindi ad interpretare i dati come in disaccordo con l'ipotesi nulla di uguaglianza in distribuzione. Il motivo di questa somiglianza tra le due distinte coppie di popolazioni molto probabilmente è dovuto all'analogo numero di fenotipi contenuti. Si ricorda, infatti, che le popolazioni Ol Molo e Masai possiedono entrambe 11 tra le 16 possibili combinazioni fenotipiche, mentre i Rendille e i Turkana ne possiedono rispettivamente 9 e 8. Nel capitolo 5 si vedrà, infatti, che queste quattro etnie verranno riunite in due gruppi, omogenei al loro interno ed eterogenei tra loro.

Infine, la tribù Kamba è stata messa a confronto con le etnie Turkana, Rendille, Ol Molo e Masai, come già fatto in precedenza per la popolazione Samburu. Per quanto riguarda le ultime due popolazioni citate, gli Ol Molo e i Masai, i valori bassi assunti dai livelli di significatività delle due statistiche test inducono a ritenere l'ipotesi di uguaglianza in distribuzione inadeguata alla luce dei dati disponibili sulle due etnie. Al contrario, dai risultati riportati nella *Tabella 8*, sembra esserci somiglianza tra la stessa popolazione Kamba e l'etnia Turkana, in particolare. Per quanto riguarda l'etnia Rendille, i livelli di significatività osservati potrebbero indicare una certa somiglianza in distribuzione anche tra quest'ultima tribù e gli stessi Kamba, vista la prossimità dei valori assunti dai p-value al livello di significatività fissato α=5%, soprattutto per il test X^2.

3.4 TEST SU IPOTESI DI DOMINANZA IN ETEROGENEITÀ

Si prende ora in considerazione il problema relativo alla verifica di dominanza in eterogeneità da parte di una popolazione più nomade rispetto ad un'altra socialmente più stanziale.

Siano P_1 e P_2 le distribuzioni delle due popolazioni da confrontare e si indichi con $Het(P_j)$ l'indice di eterogeneità calcolato nella popolazione P_j, con j=1,2.

Il sistema d'ipotesi relativo a tale problema può essere così formulato:

$$H_0 : Het(P_1) = Het(P_2) \qquad \text{vs.} \qquad H_1 : Het(P_1) > Het(P_2)$$

Con l'ipotesi alternativa H_1, quindi, si vuole verificare la conformità dei dati della popolazione P_1 ad essere maggiormente eterogenei dal punto di vista genetico rispetto ai dati della popolazione P_2.

Si vuole ribadire il concetto che i dati campionari, di cui si dispone, non possono attribuire la verità ad un'ipotesi o all'altra, ma solo essere interpretati come in accordo o disaccordo con un'ipotesi, ad un livello più o meno elevato.

Come spiegato a inizio capitolo, le probabilità associate alle classi nominali nelle due distinte popolazioni sono parametri non noti delle due distribuzioni di probabilità sottostanti.

Se, in qualsiasi modo, tali parametri fossero noti, potrebbero essere ordinati in senso decrescente all'interno di ciascuna popolazione: $p_{j(1)} \geq \ldots\ldots \geq p_{j(k)}$. Di conseguenza, secondo le proprietà richieste dagli indici di eterogeneità, se si osserva che le due popolazioni P_1 e P_2 hanno la stessa distribuzione ordinata, cioè $\{p_{1(k)}=p_{2(k)}, k=1,\ldots..,K\}$, allora tali popolazioni sono egualmente eterogenee. Inoltre, se $\{p_{1(k)}=p_{2(k)}, k=1,\ldots..,K\}$, si deduce che i dati delle rispettive classi ordinate, nei due campioni, sono scambiabili e quindi il principio dei test di permutazione è applicabile in maniera esatta e non approssimata[4].

[4] 'A permutation approach for testing heterogeneity in two-sample problems', Fortunato Pesarin – Luigi Salmaso, sottoposto per la pubblicazione su rivista.

In conformità a tali considerazioni, l'ipotesi nulla di uguale eterogeneità tra le due popolazioni può essere espressa quindi nel seguente modo: H_0 : $\{p_{1(k)}=p_{2(k)}, k=1,\ldots..,K\}$. Viceversa, in caso di maggiore eterogeneità della popolazione P_1 rispetto alla popolazione P_2, H_1 suggerisce che le probabilità cumulate della popolazione più eterogenea saranno inferiori o uguali a quelle dell'altra popolazione per ogni classe categoriale considerata e, inoltre, saranno strettamente inferiori a quelle della popolazione meno eterogenea per almeno una delle classi nominali A_K.

In questo caso, poiché l'ordine è determinato in base al valore dei parametri di popolazione p_{jk}, con k=1,…..,K e j=1,2, che per noi sono non noti, per stabilire tale ordinamento verranno usate le loro stime campionarie $\hat{p}_{jk} = \frac{f_{jk}}{n_j}$, dove f_{jk} sono le frequenze assolute osservate nella k-esima classe nominale del j-esimo campione, e n_j sono le numerosità campionarie, ossia le frequenze marginali di colonna.

Sostituendo le probabilità ignote p_{jk} con le loro stime campionarie, si ottiene un nuovo dataset ordinato: ciò implica una trasformazione dei dati del dataset originario in una tabella di contingenza derivata, contenente le frequenze relative ordinate all'interno di ciascun campione: $\{f_{j(k)}, k=1,\ldots..,K,j=1,2\}$. Si fa notare che l'ordinamento è realizzato separatamente all'interno di ciascun campione ed è basato sulle frequenze relative e non sulle classi categoriali. Ciò significa che la i-esima colonna della tabella di contingenza può riferirsi a due classi nominali diverse per i due campioni, in quanto nella i-esima posizione si trovano le classi le cui frequenze relative osservate occupano, appunto, la posizione i nella sequenza ordinata delle stime campionarie e tali classi, quindi, possono essere differenti per i due campioni. Questo implica che, per dimensioni campionarie finite, la scambiabilità dei dati rispetto ai campioni non può essere raggiunta in maniera esatta sotto H_0, ma può essere ottenuta solo asintoticamente. Pertanto, è importante osservare che, usando le frequenze campionarie al posto delle vere probabilità, i risultati inferenziali non sono più esatti ma diventano approssimati.

Come già detto nel capitolo precedente, è ragionevole usare come statistica test la differenza degli indici di eterogeneità campionari, descritti nel capitolo 2, vale a dire la statistica

$$T_G = G_1 - G_2 = \sum_{k=1}^{K} \left(\hat{p}_{2(k)}^{\,2} - \hat{p}_{1(k)}^{\,2} \right)$$

basata sull'indice di eterogeneità di Gini, la statistica

$$T_S = S_1 - S_2 = \sum_{k=1}^{K} \left[\hat{p}_{2(k)} log\left(\hat{p}_{2(k)} \right) - \hat{p}_{1(k)} log\left(\hat{p}_{1(k)} \right) \right]$$

basata sull'entropia di Shannon e la statistica

$$T_{R\infty} = R_{\infty 1} - R_{\infty 2} = log\left(max_k \, \hat{p}_{2(k)} \right) - log\left(max_k \, \hat{p}_{1(k)} \right)$$

basata sull'indice R_∞ di Renyi di ordine $\alpha=\infty$. Il motivo della scelta solamente di quest'ultimo indice, R_∞, tra i tre discussi nel paragrafo 2.4, R_1, R_2 e R_∞, deriva dal fatto che c'è una relazione uno-a-uno tra gli indici di Shannon e Gini e quelli di Renyi di ordine $\alpha=1$ e $\alpha=2$. Infatti, come già spiegato nel precedente capitolo, T_G è permutazionalmente equivalente a R_{21}-R_{22}, e T_S è permutazionalmente equivalente a R_{11}-R_{12}. È importante far osservare che i tre indici di eterogeneità utilizzati sono invarianti rispetto all'ordine con cui gli addendi vengono processati nel calcolo, cioè il loro valore non cambia se vengono calcolati con le stime campionarie dei parametri ordinate $\hat{p}_{j(k)}$, invece che con quelle non ordinate $\hat{p}_{jk}$. Inoltre, è bene notare che, per come è stata formulata l'ipotesi alternativa H_1, tutti questi test hanno regione critica, o di rifiuto, unilaterale destra; ciò significa che i test sono significativi per valori grandi: più grandi sono i valori assunti dalle statistiche test, più forte è la segnalazione di inadeguatezza dell'ipotesi nulla H_0, la quale può venire rifiutata. Poiché il livello di significatività osservato rappresenta una misura sintetica del grado di conformità tra i dati osservati e l'ipotesi nulla, un valore del livello di significatività osservato prossimo a zero è sintomo di forte disaccordo tra gli stessi dati e l'ipotesi H_0, a favore di H_1.

Il calcolo di questi test, e dei relativi livelli di significatività osservati, è stato inserito nello stesso programma 'TSD2SER.txt' contenente anche il computo delle statistiche X^2 e X^2_{AD}.

In questo caso è di interesse mettere a confronto coppie di tribù differenti in eterogeneità poiché lo scopo di questo studio è verificare se, tra le etnie che risultano diversamente eterogenee tra loro, ce ne sono alcune che presentano un'eterogeneità maggiore rispetto ad altre. Si è visto infatti, nel precedente paragrafo, che l'etnia Samburu si differenziava maggiormente dalle altre, mentre la popolazione Kamba risultava in particolar modo dissomigliante in distribuzione con le tribù Ol Molo e Masai, e queste ultime due, a loro volta, si differenziavano in distribuzione dalle popolazioni Turkana e Rendille.

In particolare, si vuole avere conferma dell'ipotesi, già espressa durante l'analisi grafica, di dominanza in eterogeneità da parte di una popolazione più nomade che, quindi, conduce una vita sociale che prevede frequenti scambi genetici con altre popolazioni ed è costretta, per continui spostamenti, ad adattarsi ad ambienti sempre differenti, rispetto ad una popolazione più stanziale, legata al proprio ambiente geografico e ligia alla propria etnia.

Anche in questo caso sono state eseguite 2000 permutazioni dei dati, e le coppie di campioni kenioti confrontati sono evidenziate nella tabella seguente, dove sono riportati anche i p-value associati ai test di Shannon, Gini e Renyi di ordine $\alpha=\infty$.

Tabella 9: p-value delle statistiche Shannon Gini e Renyi di ordine α=∞ per il test di dominanza in eterogeneità.

	p- value		
Campioni confrontati	Shannon	Gini	Renyi ordine α=∞
Samburu-Kamba	0.000	0.000	0.000
Samburu-Turkana	0.000	0.000	0.000
Samburu-Rendille	0.000	0.000	0.000
Samburu-Ol Molo	0.001	0.009	0.016
Samburu-Masai	0.000	0.000	0.039
Turkana-Kamba	0.007	0.013	0.021
Rendille-Kamba	0.008	0.009	0.029
Ol Molo-Kamba	0.000	0.000	0.001
Masai-Kamba	0.000	0.000	0.002
Rendille-Turkana	0.560	0.490	0.400
Turkana-Ol Molo	0.996	0.994	0.981
Turkana-Masai	0.967	0.966	0.977
Rendille-Ol Molo	0.997	0.995	0.974
Rendille-Masai	0.982	0.976	0.960
Ol Molo-Masai	0.173	0.176	0.446

Come si può osservare, tutti i confronti che vedono l'etnia Samburu come principale protagonista portano al rifiuto dell'ipotesi nulla di uguaglianza in eterogeneità. Infatti, fissato il livello di significatività α pari al 5%, poiché tale è il livello usualmente considerato nelle applicazioni, la quasi totalità dei livelli di significatività osservati relativi alle comparazioni tra i Samburu e le altre etnie sono prossimi allo zero. Di conseguenza, secondo la regola generale, poiché tali livelli osservati risultano essere inferiori al 5%, i dati campionari delle popolazioni confrontate sembrano essere più conformi all'ipotesi alternativa di dominanza in eterogeneità da parte dell'etnia Samburu.

Ciò significa che l'etnia Samburu non solo risulta avere una struttura genetica diversa da quella delle altre tribù con cui è stata confrontata, ma tale struttura risulta essere addirittura più complessa e completa dato l'elevato numero di combinazioni fenotipiche contenute. Forse tale dominanza in eterogeneità, da parte della suddetta etnia, deriva dal fatto che i Samburu, fin da tempi antichi, hanno condotto una vita nomade, caratterizzata da continui spostamenti in cerca di un ambiente sempre più favorevole; inoltre, come spiegato nel capitolo 1, è da tenere presente che i Samburu devono la loro discendenza da altre tribù già esistenti, quali i Masai, i Turkana e i Rendille.

Anche i confronti tra la popolazione Kamba e le altre etnie portano al rifiuto dell'ipotesi nulla; in questo caso però, accettando l'ipotesi alternativa, si accetta l'ipotesi di dominanza in eterogeneità da parte delle altre tribù nei confronti degli specifici Kamba. Dunque i Kamba risultano essere sì diversamente eterogenei rispetto ad alcune delle altre etnie comparate, come constatato nel paragrafo precedente, ma, a differenza di quanto visto per i Samburu, risultano avere una struttura genetica molto più semplice, composta da un numero di combinazioni fenotipiche nettamente inferiore a quello posseduto dai soggetti appartenenti agli altri campioni kenioti. Ciò forse è dovuto alla loro vita sociale, caratterizzata da sporadici scambi genetici con altre popolazioni, e dalla loro storia passata, secondo la quale si sono stabiliti, fin da subito, nel territorio in cui tuttora risiedono, senza esser costretti a continui spostamenti alla ricerca di un ambiente a loro più favorevole.

Una trattazione diversa, invece, è necessaria per quanto riguarda le altre etnie: Turkana, Rendille, Ol Molo e Masai. Al riguardo si osservino i p-value calcolati per tutti i possibili confronti a coppie creati a partire dalle quattro tribù, ed evidenziati nella *Tabella 9*. Come si può constatare, i livelli di significatività osservati sono, per la quasi totalità, molto elevati, superiori perfino ad un livello fissato pari al 10%. Ciò significa che le quattro tribù confrontate presentano tutte una somigliante eterogeneità genetica e non si riscontrerebbe una dominanza da parte di alcune di queste sulle altre etnie.

Uno studio più approfondito sulla suddivisione delle tribù keniote in gruppi di eterogeneità differente verrà discusso nel capitolo 5.

Capitolo 3

CAPITOLO 4

STUDIO DI SIMULAZIONE

4.1 GENERAZIONE DEI DATI

Nel precedente capitolo sono stati illustrati gli studi svolti per verificare se vi è dissomiglianza in distribuzione tra le varie tribù keniote e se esiste dominanza in eterogeneità da parte di etnie più nomadi rispetto ad altre più stanziali. Nel far questo sono state utilizzate, rispettivamente, le statistiche X^2 e X^2_{AD}, e gli indici di eterogeneità di Shannon, Gini e Renyi.

Si vuole valutare, ora, la prestazione dei test citati, ossia la loro adeguatezza ai due tipi di studi.

In particolare, verranno stimate la loro bontà di adattamento e la loro potenza. Nel far questo, sono stati calcolati, per i cinque test, i tassi di rigetto dell'ipotesi nulla, sotto le ipotesi H_0 e H_1 contenute nei sistemi d'ipotesi relativi proprio alle analisi di dissomiglianza in distribuzione e di dominanza in eterogeneità.

A tale scopo è stato condotto uno studio di simulazione secondo il quale i dati sono stati generati in base al seguente modello: $X = 1 + Int\left[KU^{\delta}\right]$, dove δ è un parametro reale, U è un numero pseudo-casuale generato dalla distribuzione uniforme nell'intervallo (0,1) e la funzione Int[·] denota la parte intera di [·].

Per ottenere i numeri pseudo-casuali è stato utilizzato lo stesso generatore 'runif', impiegato nel capitolo precedente per ottenere le permutazioni dei dati osservati. Anche in questo caso il generatore dei numeri casuali uniformi è stato inizializzato con la versione di Mersenne-Twister.

In base al modello appena descritto, i dati generati sono quindi discreti e il supporto della variabile X consiste nei primi K numeri interi positivi.

Assegnando valori più o meno elevati al parametro δ, si possono creare distribuzioni di X più o meno eterogenee per le due popolazioni P_1 e P_2 messe a confronto. La situazione di massima eterogeneità, ad esempio, può essere simulata ponendo δ pari a 1; in questo modo, infatti, la variabile X risulterebbe uniformemente distribuita sulle K classi nominali e le frequenze relative risulterebbero equamente ripartite tra le K categorie in quanto $f_k = \frac{\#(X=k)}{n} \cong \frac{1}{K}$, con k=1,…..,K e n che indica la numerosità. Aumentando il valore del parametro δ, la distribuzione di X si allontana sempre più dalla situazione di massima eterogeneità avvicinandosi, per $\delta \to \infty$, a quella di massima omogeneità, in cui le frequenze tendono a concentrarsi tutte sulla prima classe nominale.

Un'alternativa a questo modello è la generazione dei dati mediante distribuzioni di probabilità ipotizzate per le due popolazioni.

Poiché si sta studiando la potenza di alcuni test non-parametrici, la varietà di alternative proponibili per le simulazioni è talmente vasta che sarebbe impossibile considerarle tutte[5]. Col modello che si sta assumendo ora è comunque possibile generare distribuzioni discrete con differenti gradi di eterogeneità attribuendo disparati valori ad un singolo parametro reale, invece di K, come si dovrebbe fare se i dati fossero generati da una distribuzione completamente specificata, quale: $p_1, p_2, \ldots, p_K$.

Per poter valutare il grado di approssimazione e la potenza delle statistiche test utilizzate nel precedente capitolo, è stato necessario scrivere un programma[6] che calcolasse, per i test stessi, i tassi di rigetto di H_0, sia sotto l'ipotesi nulla che sotto l'ipotesi alternativa.

Infatti è bene precisare che, per valutare il grado di approssimazione di una statistica test, è necessario calcolare, per quello stesso test, i tassi di rigetto dell'ipotesi nulla, quando però H_0 risulta essere vera. Al contrario, per valutarne la potenza, si devono cal-

[5] Lehmann, 1953.
[6] Il programma 'Stet4t-d.txt' è allegato in Appendice.

colare i tassi di rigetto dell'ipotesi nulla, quando H_0 è effettivamente falsa. Pertanto, all'interno del programma è stato necessario inserire delle procedure che permettessero di calcolare la probabilità di rigetto dell'ipotesi nulla per ogni singolo test; in riferimento, poi, ai sistemi d'ipotesi formulati per lo studio di diversità in distribuzione e di dominanza in eterogeneità, verranno attribuiti valori uguali o dissimili ai due parametri δ_1 e δ_2, permettendo così la stima, rispettivamente, del grado di approssimazione o della potenza di ciascun test.

Dapprima, perciò, sono state scritte delle funzioni per la creazione dei dati categoriali entro ciascun campione, creazione avvenuta con l'ausilio dei numeri pseudo-casuali U prodotti dal generatore precedentemente menzionato e tramite l'utilizzo dei rispettivi parametri δ_1 e δ_2. In seguito, i dati generati sono stati ordinati per frequenze decrescenti all'interno delle due popolazioni. Inoltre, sono state scritte delle funzioni per il calcolo del valore osservato assunto dalle statistiche test applicate ai dati categoriali originati, e per il calcolo del valore assunto dagli stessi test applicati ai dati permutati.

Quindi, per valutare il grado di approssimazione delle statistiche test e la loro potenza, sono state create delle procedure per il computo dei relativi livelli di significatività, necessari per poter calcolare la probabilità di rigetto dell'ipotesi nulla attraverso il seguente calcolo: $\frac{\#(\lambda \leq \alpha)}{MC}$, dove $\#(\lambda \leq \alpha)$ indica il numero di volte che i livelli di significatività osservati risultano inferiori o uguali ai livelli nominali α e MC rappresenta il numero di simulazioni di Monte Carlo.

Il codice è stato scritto utilizzando il pacchetto statistico R. Inoltre, per l'esecuzione di tale codice, è necessaria la partecipazione dell'utente finale poiché è proprio quest'ultimo a dovere inserire le numerosità dei due campioni confrontati, il numero di insiemi di dati generati attraverso la simulazione di Monte Carlo e, per ciascuno di questi dataset creati, il numero di permutazioni B che devono essere eseguite sui dati in essi contenuti. Oltre a ciò, l'utente finale deve precisare il numero di classi categoriali entro

cui ripartire i dati generati e i valori dei parametri delle due distribuzioni, δ_1 e δ_2. Va ricordato che questi ultimi dati richiesti, vale a dire la numerosità delle classi e i valori dei parametri, sono indispensabili per la creazione dei dati secondo il modello $1 + Int\left[K\,U^{\delta}\right]$, presentato a inizio paragrafo; in particolare, la scelta dei due parametri è rilevante proprio perchè caratterizza la misura di eterogeneità presente nelle distribuzioni create per le specifiche popolazioni.

I dati, perciò, sono stati generati prendendo in considerazione, di volta in volta, dimensioni campionarie differenti e anche valori diversi da attribuire sia al numero di classi categoriali K da creare, sia ai parametri δ_1 e δ_2 delle rispettive popolazioni.

Inoltre, per rendere possibile la valutazione della bontà di adattamento dei test utilizzati, sono stati stabiliti alcuni valori per i livelli di significatività nominale α, necessari per il confronto con i tassi di rigetto calcolati per le cinque statistiche test.

È bene far notare che i p-value, calcolati sia per la valutazione del grado di approssimazione dei test che della loro potenza, possono differire da quelli ottenibili con esecuzioni successive del programma scritto per lo studio di simulazione. Infatti, pur mantenendo costanti i valori assegnati ai parametri di eterogeneità e la numerosità sia di classi categoriali K create che di permutazioni B e di simulazioni di Monte Carlo, i dati generati all'interno delle due popolazioni P_1 e P_2 secondo il modello descritto a inizio capitolo possono, di volta in volta, risultare differenti proprio a causa dei numeri pseudo-casuali U, ottenuti tramite il generatore di pura aleatorietà 'runif'.

Per gli studi di simulazione, condotti sia per le statistiche test X^2 e X^2_{AD} che per gli indici di eterogeneità di Shannon, Gini e Renyi, sono stati generati 150 insiemi di dati e, per ciascuno di questi dataset, sono state eseguite 150 permutazioni dei dati stessi; così facendo, infatti, è possibile approssimare la distribuzione di permutazione relativa a ciascuno degli insiemi di dati prodotti.

In realtà, la numerosità sia dei dataset generati che delle permutazioni eseguite dovrebbe essere nell'ordine delle migliaia per poter ottenere una migliore approssimazione dei risultati. Poiché R non è un vero linguaggio di programmazione, eseguire una tale quantità di permutazioni e di generazioni di dati richiederebbe tempi di esecuzione troppo lunghi. Di conseguenza, per la necessità di avere risultati dello studio di simulazione in tempi più brevi, è stato necessario optare per numerosità più contenute, sia di simulazioni che di permutazioni.

4.2 STATISTICHE TEST X^2 E X^2_{AD}

4.2.1 GRADO DI APPROSSIMAZIONE

Nel precedente capitolo le statistiche test X^2 e X^2_{AD} sono state utilizzate per valutare se vi è somiglianza o diversità tra distribuzioni di popolazioni. Il sistema d'ipotesi testato era infatti così formulato:

$$H_0 : P(X1) = P(X2) \qquad \text{vs.} \qquad H_1 : P(X1) \neq P(X2)$$

Si valuterà ora la bontà di adattamento delle due statistiche test a questo specifico tipo di problema.

A tal proposito è stato necessario calcolare i tassi di rigetto dell'ipotesi nulla di uguaglianza in distribuzione da parte dei due test, ipotizzando però che H_0 sia vera.

Poiché l'ipotesi nulla assume che ci sia uguaglianza in distribuzione tra le due popolazioni, sono stati attribuiti uguali valori ai due parametri δ_1 e δ_2, facendo variare i gradi di eterogeneità in un intervallo che va da un valore minimo di 1.5 ad uno massimo pari a 2.5.

Se i tassi di rifiuto, associati alle due statistiche, risultano molto prossimi ai livelli nominali α di significatività, allora i due test presentano un buon grado di approssimazione

e, pertanto, hanno un buon adattamento al tipo di problema; ciò significa, infatti, che i test minimizzano la probabilità di rifiutare l'ipotesi nulla quando questa però è vera. Al contrario, se i tassi risultanti assumono valori molto discosti dai livelli nominali di significatività, in tal caso i test associati non sembrano adeguati al tipo di problema per il quale sono stati utilizzati.

Nella *Tabella 10*, riportata nelle pagine seguenti, vengono evidenziati i tassi di rigetto di H_0 calcolati per le due statistiche test di Pearson, sotto l'ipotesi nulla di uguaglianza in distribuzione tra due popolazioni.

Tabella 10: Tassi di rigetto delle statistiche test X^2 e X^2_{AD} sotto ipotesi nulla (MC=150; B=150; K=(16,64,128,256,512); $1.5 \leq (\delta_1, \delta_2) \leq 2.5$).

K	δ_1	δ_2	n_1	n_2	α nominale	X^2	X^2_{AD}
16	1.5	1.5	60	30	0.01	0.020	0.020
					0.05	0.071	0.073
					0.10	0.127	0.110
					0.20	0.247	0.253
					0.30	0.337	0.330
					0.50	0.527	0.527
16	2.5	2.5	60	30	0.01	0.020	0.027
					0.05	0.073	0.080
					0.10	0.130	0.137
					0.20	0.233	0.220
					0.30	0.333	0.353
					0.50	0.520	0.527
16	3	3	60	30	0.01	0.020	0.020
					0.05	0.047	0.060
					0.10	0.100	0.093
					0.20	0.200	0.193
					0.30	0.327	0.340
					0.50	0.527	0.547
16	1.5	1.5	100	100	0.01	0.020	0.020
					0.05	0.053	0.053
					0.10	0.093	0.093
					0.20	0.193	0.187
					0.30	0.323	0.333
					0.50	0.530	0.537

Tabella 10 (continua): Tassi di rigetto delle statistiche test X^2 e X^2_{AD} sotto ipotesi nulla (MC=150; B=150; K=(16,64,128,256,512); $1.5 \leq (\delta_1, \delta_2) \leq 2.5$).

K	δ_1	δ_2	n_1	n_2	α nominale	X^2	X^2_{AD}
16	2.5	2.5	100	100	0.01	0.013	0.007
					0.05	0.060	0.073
					0.10	0.103	0.133
					0.20	0.213	0.227
					0.30	0.313	0.320
					0.50	0.513	0.500
16	1.5	1.5	1000	1000	0.01	0.007	0.013
					0.05	0.047	0.040
					0.10	0.080	0.080
					0.20	0.213	0.220
					0.30	0.313	0.313
					0.50	0.540	0.547
64	1.5	1.5	60	60	0.01	0.000	0.000
					0.05	0.007	0.013
					0.10	0.093	0.093
					0.20	0.167	0.160
					0.30	0.273	0.280
					0.50	0.473	0.473
64	2	2	60	60	0.01	0.007	0.007
					0.05	0.047	0.047
					0.10	0.093	0.073
					0.20	0.210	0.213
					0.30	0.320	0.320
					0.50	0.500	0.493
128	2	2	60	60	0.01	0.013	0.013
					0.05	0.053	0.060
					0.10	0.113	0.123
					0.20	0.213	0.230
					0.30	0.317	0.333
					0.50	0.520	0.520
256	2	2	60	60	0.01	0.013	0.013
					0.05	0.060	0.060
					0.10	0.093	0.093
					0.20	0.167	0.167
					0.30	0.273	0.280
					0.50	0.493	0.487
512	2	2	60	60	0.01	0.013	0.013
					0.05	0.020	0.027
					0.10	0.073	0.073
					0.20	0.167	0.173
					0.30	0.220	0.227
					0.50	0.433	0.453

I p-value, contenuti nella tabella sopra esposta, sembrano dimostrare che entrambe le statistiche test presentano una buona approssimazione al tipo di problema per il quale sono state utilizzate. In generale, infatti, i tassi di rigetto non si discostano molto dai li-

velli di significatività nominale. Si può osservare che, all'aumentare della cardinalità delle classi categoriali K, i p-value associati ai due test sembrano migliorare e tendono leggermente a diminuire rispetto al livello α.

Inoltre, si nota che, pur mantenendo costanti la numerosità campionaria e il numero di classi categoriali della variabile X, il grado di approssimazione di entrambe le statistiche test sembra progredire con l'incremento dei valori assunti dai parametri δ_1 e δ_2, cioè man mano che le frequenze relative delle due popolazioni confrontate tendono a concentrarsi su un minor numero di classi categoriali.

In ogni caso si può dire che i due test risultano essere sostanzialmente ben approssimati e le loro prestazioni sono molto simili, anche se si evidenzia una sottile preferenza per la statistica X^2, poiché i tassi di rigetto ad essa associati si approssimano meglio ai livelli nominali α.

4.2.2 POTENZA

Oltre al grado di approssimazione delle statistiche X^2 e X^2_{AD}, è importante valutare anche la loro funzione di potenza, ossia la probabilità di rifiutare l'ipotesi nulla H_0 quando effettivamente essa è falsa.

La funzione di potenza è altresì rilevante perché consente di definire alcune proprietà desiderabili per un test statistico, quali la correttezza e la consistenza. Si ricorda, infatti, che un test non distorto, ossia corretto, conduce più facilmente al rifiuto dell'ipotesi nulla quando essa è falsa piuttosto che quando essa è vera. Invece un test, basato su un campione casuale semplice con numerosità n, è detto consistente se conduce assai facilmente al rifiuto dell'ipotesi nulla, quando essa è falsa, se n è sufficientemente grande.

Per poter valutare la potenza dei test X^2 e X^2_{AD}, quindi, sono stati calcolati i tassi di rigetto dell'ipotesi nulla da parte delle due statistiche test, ipotizzando H_0 falsa, ossia equivalentemente che è vera l'ipotesi alternativa.

Poiché H_1 esprime la diversità in distribuzione tra due popolazioni, sono stati assegnati valori differenti ai due parametri δ_1 e δ_2, in modo tale da creare due distribuzioni di popolazione abbastanza dissomiglianti tra loro, con le rispettive frequenze relative distribuite diversamente tra le varie classi nominali.

Al contrario di quanto detto per la valutazione del grado di approssimazione, se in questo caso si osservano p-value ben superiori ai livelli nominali α di significatività, vuol dire che il relativo test massimizza la probabilità di rifiutare l'ipotesi nulla quando effettivamente questa è falsa e, quindi, la statistica test risulta essere potente.

Nella sottostante *Tabella 11* vengono evidenziati i risultati ottenuti con lo studio di simulazione sulla potenza delle statistiche test X^2 e X^2_{AD}.

Tabella 11: Potenza delle statistiche test X^2 e X^2_{AD} (MC=150; B=150; K=(16,64,128,256,512); $1.5 \leq (\delta_1,\delta_2) \leq 4$).

K	δ_1	δ_2	n_1	n_2	α nominale	X^2	X^2_{AD}
16	2	3	60	30	0.01	0.053	0.067
					0.05	0.133	0.147
					0.10	0.207	0.213
					0.20	0.313	0.373
					0.30	0.427	0.427
					0.50	0.593	0.653
16	2	3.5	60	30	0.01	0.060	0.073
					0.02	0.147	0.200
					0.10	0.233	0.273
					0.20	0.380	0.447
					0.30	0.533	0.600
					0.50	0.700	0.747
16	2	4	60	30	0.01	0.113	0.160
					0.05	0.240	0.313
					0.10	0.320	0.393
					0.20	0.473	0.547
					0.30	0.607	0.660
					0.50	0.780	0.827

Tabella 11 (continua): Potenza delle statistiche test X^2 e X^2_{AD} (MC=150; B=150; K=(16,64,128,256,512); $1.5 \leq (\delta_1, \delta_2) \leq 4$).

K	δ_1	δ_2	n_1	n_2	α nominale	X^2	X^2_{AD}
64	2	3	60	60	0.01	0.047	0.047
					0.05	0.167	0.180
					0.10	0.247	0.267
					0.20	0.373	0.413
					0.30	0.487	0.553
					0.50	0.693	0.707
64	2	3.5	60	60	0.01	0.187	0.200
					0.05	0.280	0.327
					0.10	0.367	0.413
					0.20	0.520	0.573
					0.30	0.627	0.667
					0.50	0.813	0.847
128	2	3	60	60	0.01	0.040	0.033
					0.05	0.140	0.160
					0.10	0.220	0.260
					0.20	0.367	0.380
					0.30	0.493	0.520
					0.50	0.707	0.713
128	2	4	60	60	0.01	0.247	0.293
					0.05	0.407	0.467
					0.10	0.533	0.580
					0.20	0.680	0.760
					0.30	0.793	0.827
					0.50	0.907	0.933
256	1.5	2	60	60	0.01	0.093	0.093
					0.05	0.133	0.133
					0.10	0.160	0.167
					0.20	0.287	0.300
					0.30	0.380	0.380
					0.50	0.507	0.520
256	2	4	60	60	0.01	0.260	0.273
					0.05	0.433	0.447
					0.10	0.527	0.580
					0.20	0.680	0.713
					0.30	0.767	0.793
					0.50	0.873	0.880
512	1.5	2	60	60	0.01	0.047	0.047
					0.05	0.120	0.133
					0.10	0.187	0.187
					0.20	0.293	0.307
					0.30	0.387	0.407
					0.50	0.533	0.533
512	2	4	60	60	0.01	0.300	0.333
					0.05	0.487	0.500
					0.10	0.547	0.613
					0.20	0.733	0.760
					0.30	0.800	0.820
					0.50	0.880	0.900

Confrontando i p-value, associati alle due statistiche test ed evidenziati nella tabella riportata nelle pagine precedenti, si può constatare che sia il test X^2 che il test X^2_{AD} sembrano avere prestazioni piuttosto somiglianti tra loro, in quanto i relativi tassi di rigetto a volte sono abbastanza analoghi.

Inoltre è bene notare che la potenza dei due test migliora con l'aumentare della cardinalità K delle classi categoriali create ma, soprattutto, aumenta man mano che cresce la differenza nei valori assunti dai due parametri di eterogeneità δ_1 e δ_2.

Quindi si può concludere che sia il test X^2 che il test X^2_{AD}, equivalente alla divergenza secondo Anderson-Darling tra le distribuzioni di frequenza, sembrano adatti per uno studio sulla dissomiglianza in distribuzione tra due popolazioni; entrambe presentano una buona approssimazione e anche i p-value, risultanti dallo studio di simulazione sulla potenza, confermano la loro adeguatezza a questo tipo di problema.

4.3 INDICI DI ETEROGENEITÀ DI SHANNON, GINI E RENYI

4.3.1 GRADO DI APPROSSIMAZIONE

Nel precedente capitolo, oltre ad uno studio sull'uguaglianza o diversità in distribuzione, è stata svolta anche un'analisi sulla verifica di dominanza in eterogeneità da parte di una popolazione più nomade rispetto ad una più stanziale.

Il sistema d'ipotesi, testato per questo tipo di problema, era infatti così formulato:

$$H_0 : Het(P_1) = Het(P_2) \qquad \text{vs.} \qquad H_1 : Het(P_1) > Het(P_2)$$

Al riguardo, sono stati utilizzati l'indice di eterogeneità di Gini, l'indicatore di Shannon e l'indice di entropia generalizzato di ordine $\alpha=\infty$ di Renyi.

All'interno dello stesso programma 'Stet4t-d.txt', scritto per lo studio di simulazione sui test X^2 e X^2_{AD}, sono state incluse anche istruzioni per il calcolo dei valori osservati dei

tre indici di eterogeneità appena menzionati, e il calcolo dei valori di permutazione assunti dagli stessi test. Infine, è stato aggiunto anche il computo dei livelli di significatività ad essi associati, necessari per la successiva stima del loro grado di approssimazione e della loro potenza.

Inizialmente si valuterà il grado di approssimazione dei tre test. A tal proposito sono stati calcolati i tassi di rigetto dell'ipotesi nulla di uguaglianza in eterogeneità, ipotizzando vera H_0.

Poiché, come sopra riportato, l'ipotesi nulla assume che ci sia uguale eterogeneità tra popolazioni, vengono attribuiti pari valori ai parametri reali delle due popolazioni confrontate, facendo variare i gradi di eterogeneità in un intervallo che va da un valore minimo di 1.5 ad un valore massimo pari a 4.

Come già spiegato nel paragrafo 4.2.1, se i p-value associati ai tre indici di eterogeneità risultano molto prossimi ai livelli nominali α di significatività, ciò significa che gli indicatori utilizzati sono adatti per l'analisi di dominanza in eterogeneità che è stata condotta. Al contrario, tassi di rigetto risultanti troppo lontani dai livelli nominali α sono indice di cattivo adattamento dei test al tipo di problema considerato.

Nella *Tabella 12*, riportata nelle pagine seguenti, vengono evidenziati i tassi di rigetto di H_0 calcolati per i tre indici di eterogeneità, sotto l'ipotesi nulla di uguaglianza in eterogeneità tra due popolazioni.

Tabella 12: Tassi di rigetto degli indici di eterogeneità di Shannon, Gini e Renyi di ordine $\alpha=\infty$ sotto ipotesi nulla (MC=150; B=150; K=(16,64,128,256,512); $1.5\leq(\delta_1,\delta_2)\leq 4$).

K	δ_1	δ_2	n_1	n_2	α nominale	Shannon	Gini	Renyi $\alpha=\infty$
16	1.5	1.5	60	30	0.01	0.007	0.000	0.002
					0.05	0.053	0.049	0.033
					0.10	0.073	0.093	0.073
					0.20	0.153	0.193	0.153
					0.30	0.260	0.280	0.213
					0.50	0.453	0.480	0.447
16	3	3	60	30	0.01	0.020	0.020	0.020
					0.05	0.073	0.057	0.040
					0.10	0.133	0.110	0.073
					0.20	0.233	0.227	0.153
					0.30	0.340	0.333	0.240
					0.50	0.533	0.527	0.433
16	4	4	60	30	0.01	0.047	0.023	0.013
					0.05	0.093	0.053	0.033
					0.10	0.133	0.113	0.067
					0.20	0.233	0.220	0.127
					0.30	0.350	0.327	0.253
					0.50	0.533	0.513	0.460
64	2	2	60	60	0.01	0.027	0.023	0.027
					0.05	0.073	0.060	0.053
					0.10	0.113	0.113	0.093
					0.20	0.207	0.213	0.167
					0.30	0.313	0.300	0.233
					0.50	0.500	0.447	0.440
64	3.5	3.5	60	60	0.01	0.027	0.027	0.027
					0.05	0.080	0.063	0.060
					0.10	0.120	0.117	0.073
					0.20	0.210	0.223	0.140
					0.30	0.327	0.320	0.200
					0.50	0.507	0.510	0.427
128	2	2	60	60	0.01	0.007	0.020	0.013
					0.05	0.073	0.070	0.067
					0.10	0.153	0.140	0.140
					0.20	0.207	0.233	0.193
					0.30	0.333	0.337	0.260
					0.50	0.500	0.507	0.413
128	4	4	60	60	0.01	0.020	0.027	0.013
					0.05	0.053	0.060	0.040
					0.10	0.107	0.127	0.067
					0.20	0.187	0.230	0.140
					0.30	0.273	0.330	0.160
					0.50	0.513	0.523	0.393
256	1.5	1.5	60	60	0.01	0.007	0.023	0.000
					0.05	0.040	0.053	0.007
					0.10	0.067	0.130	0.067
					0.20	0.200	0.237	0.153
					0.30	0.293	0.337	0.260
					0.50	0.433	0.420	0.287

Tabella 12 (continua): Tassi di rigetto degli indici di eterogeneità di Shannon, Gini e Renyi di ordine $\alpha=\infty$ sotto ipotesi nulla (MC=150; B=150; K=(16,64,128,256,512); $1.5\leq(\delta_1,\delta_2)\leq 4$).

K	δ_1	δ_2	n_1	n_2	α nominale	Shannon	Gini	Renyi $\alpha=\infty$
256	3	3	60	60	0.01	0.020	0.030	0.007
					0.05	0.100	0.073	0.037
					0.10	0.127	0.133	0.087
					0.20	0.227	0.240	0.153
					0.30	0.320	0.327	0.253
					0.50	0.467	0.457	0.453
512	2	2	60	60	0.01	0.013	0.033	0.013
					0.05	0.067	0.083	0.040
					0.10	0.113	0.160	0.073
					0.20	0.247	0.267	0.193
					0.30	0.307	0.333	0.267
					0.50	0.467	0.507	0.333
512	3.5	3.5	60	60	0.01	0.033	0.047	0.007
					0.05	0.087	0.093	0.043
					0.10	0.123	0.153	0.087
					0.20	0.220	0.247	0.193
					0.30	0.310	0.327	0.267
					0.50	0.480	0.493	0.427

Come si può notare dai livelli di significatività evidenziati nella presente tabella, il test basato sull'indice di Renyi di ordine infinito è, fra i tre indicatori considerati, quello maggiormente conservativo; i tassi di rigetto dell'ipotesi nulla calcolati per tale test, infatti, risultano spesso molto inferiori ai livelli nominali α di significatività.

Per quanto riguarda gli altri due indici, quello di Gini e quello di Shannon, le loro prestazioni sono abbastanza simili e anch'essi risultano sostanzialmente ben approssimati; i due test, infatti, sono meno conservati rispetto al test di Renyi e i p-value a loro associati si approssimano abbastanza bene ai livelli nominali α.

Si può notare anche che il grado di approssimazione dell'indice di Gini, in particolar modo, aumenta progressivamente con l'assegnazione di valori sempre più elevati ai due parametri δ_1 e δ_2 e man mano che aumenta anche il divario tra i valori assunti da ciascuna coppia (δ_1,δ_2).

In ogni caso, si può concludere affermando che i tre test sono tutti sostanzialmente ben approssimati.

4.3.2 POTENZA

Nel paragrafo precedente è stato stimato il grado di approssimazione degli indici di eterogeneità di Shannon, Gini e Renyi. Oltre a ciò, si reputa interessante valutare anche la loro funzione di potenza.

A tale scopo, per i tre indicatori, sono stati calcolati i tassi di rigetto dell'ipotesi nulla di dominanza in eterogeneità da parte di una popolazione più nomade rispetto a una più stanziale, ipotizzando vera H_1.

Per come è stata formulata l'ipotesi alternativa, per poter stimare la potenza dei tre test si deve considerare la situazione in cui l'eterogeneità presente nelle due popolazioni è differente. Di conseguenza, sono stati assegnati valori dissimili ai due parametri δ_1 e δ_2, scegliendo tali gradi di eterogeneità in un intervallo contenente valori compresi tra 1.5 e 4.

Come già spiegato, p-value superiori ai livelli nominali α di significatività sono indice di buona potenza da parte del relativo test, il quale massimizza la probabilità di rifiutare l'ipotesi nulla quando H_0 è effettivamente falsa e, pertanto, risulta adatto ad essere utilizzato per questo tipo di problema.

Nella *Tabella 13*, riportata nelle pagine successive, vengono evidenziati i risultati relativi alla potenza dei tre indici di eterogeneità.

Tabella 13: Potenza degli indici di eterogeneità di Shannon, Gini e Renyi di ordine $\alpha=\infty$ (MC=150; B=150; K=(16,64,128,256,512); $1.5\leq(\delta_1,\delta_2)\leq 4$).

K	δ_1	δ_2	n_1	n_2	α nominale	Shannon	Gini	Renyi $\alpha=\infty$
16	1.5	2	60	30	0.01	0.133	0.100	0.013
					0.05	0.240	0.240	0.173
					0.10	0.373	0.373	0.280
					0.20	0.560	0.577	0.467
					0.30	0.707	0.700	0.567
					0.50	0.827	0.810	0.740
16	2	3	60	30	0.01	0.207	0.167	0.147
					0.02	0.420	0.407	0.340
					0.10	0.553	0.520	0.513
					0.20	0.693	0.680	0.667
					0.30	0.760	0.773	0.747
					0.50	0.900	0.903	0.900
16	2	3.5	60	30	0.01	0.407	0.360	0.313
					0.05	0.627	0.640	0.540
					0.10	0.780	0.770	0.640
					0.20	0.860	0.867	0.780
					0.30	0.933	0.933	0.847
					0.50	0.980	0.987	0.940
16	2	4	60	30	0.01	0.587	0.530	0.420
					0.05	0.820	0.873	0.693
					0.10	0.907	0.903	0.840
					0.20	0.920	0.923	0.920
					0.30	0.920	0.940	0.953
					0.50	0.967	0.973	0.980
64	2	3.5	60	60	0.01	0.620	0.653	0.427
					0.05	0.820	0.807	0.700
					0.10	0.873	0.873	0.820
					0.20	0.947	0.973	0.920
					0.30	0.973	0.980	0.940
					0.50	0.987	0.987	1.000
128	2	4	60	60	0.01	0.773	0.827	0.065
					0.05	0.900	0.933	0.820
					0.10	0.967	0.960	0.900
					0.20	0.987	0.993	0.960
					0.30	0.993	0.993	0.993
					0.50	1.000	1.000	0.993
256	1.5	2	60	60	0.01	0.093	0.130	0.160
					0.05	0.220	0.340	0.313
					0.10	0.367	0.467	0.433
					0.20	0.580	0.620	0.520
					0.30	0.680	0.700	0.633
					0.50	0.840	0.820	0.700
256	2	4	60	60	0.01	0.793	0.847	0.607
					0.05	0.907	0.960	0.813
					0.10	0.960	0.967	0.900
					0.20	0.987	0.993	0.947
					0.30	0.993	0.993	0.967
					0.50	0.993	1.000	0.993

Tabella 13 (continua): Potenza degli indici di eterogeneità di Shannon, Gini e Renyi di ordine $\alpha=\infty$ (MC=150; B=150; K=(16,64,128,256,512); $1.5\leq(\delta_1,\delta_2)\leq 4$).

K	δ_1	δ_2	n_1	n_2	α nominale	Shannon	Gini	Renyi $\alpha=\infty$
512	2	3.5	60	60	0.01	0.620	0.710	0.433
					0.05	0.790	0.860	0.687
					0.10	0.903	0.890	0.753
					0.20	0.920	0.930	0.867
					0.30	0.967	0.963	0.913
					0.50	0.990	0.990	0.987
512	2	4	60	60	0.01	0.800	0.890	0.993
					0.05	0.897	0.943	0.587
					0.10	0.953	0.960	0.800
					0.20	0.980	0.990	0.893
					0.30	0.997	0.993	0.927
					0.50	1.000	1.000	0.960

Come è ovvio immaginare, la potenza dei tre test confrontati aumenta man mano che incrementa la differenza nei valori assunti dai parametri di eterogeneità delle due popolazioni.

Comparando i risultati ottenuti, emerge che il test basato sulla statistica di Renyi di ordine $\alpha=\infty$ è lievemente peggiore rispetto agli altri due, in quanto i p-value ad esso associati risultano inferiori a quelli calcolati per le altre due statistiche.

Al contrario, si evidenzia una sottile preferenza per il test basato sull'indice di Gini. Il motivo di tale distinzione deriva dal fatto che, come visto nel precedente paragrafo, l'indice di Gini sembra essere il meno conservativo tra i tre indicatori di eterogeneità considerati; ne consegue che anche la sua potenza risulta migliore rispetto a quella degli indici di Shannon e di Renyi.

In conclusione, tale test sembra essere il più consono per uno studio sul confronto dell'eterogeneità genetica presente in coppie di popolazioni.

Capitolo 4

CAPITOLO 5

GRUPPI DI ETNIE

5.1 DIVERSITÀ GENETICA TRA ETNIE

Dagli studi svolti nei precedenti capitoli si è potuto constatare che esiste diversità, dal punto di vista genetico, tra le sei tribù keniote, oggetto di analisi in questa tesi. Nel capitolo 3, infatti, si è appurato che, oltre alla dissomiglianza in distribuzione tra le popolazioni considerate, è presente anche una dominanza in eterogeneità genetica, da parte di etnie più nomadi nei confronti di altre più stanziali.

Si è già discusso dell'influsso, sull'evoluzione razziale, prodotto dalla selezione naturale in rapporto all'ambiente e dai profili storici-culturali propri di ogni popolazione.

Ciascuna tribù si distingue, quindi, per costrizioni bioculturali, usi e costumi, regole proprie del gruppo di appartenenza; ma la differenziazione tra le varie etnie, soprattutto da un punto di vista genetico, trova spiegazione sia nel loro isolamento e adattamento ad uno specifico ambiente, sia nelle loro origini e storia passata e, in special modo, nei loro spostamenti e conseguenti rapporti con altre tribù.

In particolare, finora si è potuto riscontrare che la popolazione Kamba è, tra le sei considerate, quella che ha la struttura genetica più semplice, composta da sole 4 combinazioni fenotipiche tra le 16 possibili. Inoltre, come già accennato nel capitolo 1 a riguardo della storia del Kenya e delle sue tribù, sembra che i Kamba siano stati gli unici a non aver subito forti selezioni naturali in rapporto all'ambiente.

In contrapposizione a tale etnia c'è quella dei Samburu, la cui struttura genetica, come si è potuto realizzare dai precedenti capitoli, risulta alquanto complessa e completa. I

Samburu, infatti, sono i soli autoctoni a possedere tutti i possibili fenotipi nel proprio patrimonio genetico.

Le rimanenti quattro etnie, Turkana, Rendille, Ol Molo e Masai, sono risultate, invece, abbastanza somiglianti fra loro sia dal punto di vista genetico, per numerosità di fenotipi offerti, sia per la loro storia passata che per il territorio geografico sul quale sono stanziate, condiviso in parte.

5.2 STRATEGIA DI BONFERRONI-HOLM

Poiché le sei tribù, sopra citate, sembrano differenziarsi tra loro in rapporto ad origini, struttura genetica e ambiente occupato, è stato condotto uno studio per stabilire quali etnie risultassero maggiormente somiglianti, soprattutto per quanto riguarda le combinazioni fenotipiche possedute, e significativamente diverse da altre.

Quindi, scopo di tale analisi è quello di isolare le sei tribù keniote in gruppi risultanti geneticamente omogenei al loro interno, ma eterogenei tra loro.

A tale proposito sono stati presi in considerazione i livelli di significatività, associati alla statistica di Gini, che già erano stati calcolati, per tutti i possibili confronti a coppie creati a partire dalle sei tribù oggetto di analisi, relativamente allo studio di dominanza in eterogeneità, descritto nel paragrafo 3.4.

Si è fatto riferimento solo ai p-value della statistica test di Gini poiché, come visto nel precedente capitolo, tale test è risultato essere il più adatto per lo studio sulla dominanza in eterogeneità tra popolazioni. Infatti, si è dimostrato che l'indicatore di Gini ha un buon grado di approssimazione al tipo di problema posto e una buona potenza, ossia massimizza la probabilità di rifiuto dell'ipotesi nulla quando essa è effettivamente falsa.

Per stabilire, però, quali tribù sono significativamente differenti tra loro, per struttura genetica e fenotipi offerti, non è corretto selezionare solo le coppie di etnie per le quali,

individualmente, è stato calcolato un livello di significatività osservato inferiore al 5%, valore fissato per il livello nominale α. Così facendo, infatti, la probabilità di commettere almeno un errore di primo tipo, ossia di accettare, per almeno una tra le coppie di popolazioni confrontate, l'ipotesi nulla di uguaglianza in eterogeneità anche quando essa è falsa, potrebbe essere troppo elevato.

Al fine di controllare la massima probabilità con cui una o più di queste sottoipotesi nulle venga rifiutata scorrettamente, detta 'FamilyWise Error rate' (FWE), è stata utilizzata una procedura di Closed Testing. In questo modo, quindi, si garantisce che la probabilità di commettere almeno un errore di primo tipo sia inferiore o uguale al livello nominale fissato α.

In particolare, è stata utilizzata la procedura step-down di Bonferroni-Holm. Tale strategia comporta, prima di tutto, l'ordinamento, in sequenza crescente, dei p-value calcolati per le ipotesi minimali, ossia proprio dei livelli di significatività, associati alla statistica di Gini, calcolati per lo studio di dominanza in eterogeneità, su tutti i possibili confronti a coppie creati a partire dalle sei tribù keniote.

Si ricorda, infatti, che il sistema d'ipotesi relativo a questo tipo di studio a due campioni era così formulato:

$$H_0 : Het(P_1) = Het(P_2) \qquad \text{vs.} \qquad H_1 : Het(P_1) > Het(P_2)$$

Nella *Tabella 14*, riportata nelle pagine seguenti, vengono evidenziati i p-value calcolati, ordinati in maniera crescente.

Tabella 14: Livelli di significatività osservati per l'indice di Gini, calcolati per tutti i possibili confronti a coppia creati a partire dalle sei tribù keniote.

$H_{(1)}$	$H_{(2)}$	$H_{(3)}$	$H_{(4)}$	$H_{(5)}$	$H_{(6)}$	$H_{(7)}$	$H_{(8)}$
Samburu vs. Kamba	Samburu vs. Turkana	Samburu vs. Rendille	Samburu vs. Masai	Masai vs. Kamba	Ol Molo vs. Kamba	Samburu vs. Ol Molo	Rendille vs. Kamba
$p_{(1)}$	$p_{(2)}$	$p_{(3)}$	$p_{(4)}$	$p_{(5)}$	$p_{(6)}$	$p_{(7)}$	$p_{(8)}$
0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.009

Tabella 14 (continua): Livelli di significatività osservati per l'indice di Gini, calcolati per tutti i possibili confronti a coppia creati a partire dalle sei tribù keniote.

$H_{(9)}$	$H_{(10)}$	$H_{(11)}$	$H_{(12)}$	$H_{(13)}$	$H_{(14)}$	$H_{(15)}$
Turkana vs. Kamba	Ol Molo vs. Masai	Rendille vs. Turkana	Turkana vs. Masai	Rendille vs. Masai	Turkana vs. Ol Molo	Rendille vs. Ol Molo
$p_{(9)}$	$p_{(10)}$	$p_{(11)}$	$p_{(12)}$	$p_{(13)}$	$p_{(14)}$	$p_{(15)}$
0.013	0.176	0.490	0.966	0.976	0.994	0.995

Una volta ordinati in maniera crescente i p-value ottenuti col test di Gini, si procede con il computo dei p-value aggiustati, come dispone la strategia di Bonferroni-Holm. Questi ultimi sono stati calcolati seguendo un preciso algoritmo, qui di seguito esposto:

1. adj-$p_{(1)}=K\cdot p_{(1)}$;

   se adj-$p_{(1)}\leq\alpha$ si rifiuta $H_{0(1)}$ e si procede con l'algoritmo, altrimenti si accettano $H_{0(1)},\ldots..,H_{0(K)}$ e l'algoritmo termina.

2. adj-$p_{(i)}=\max[(K-i+1)\ p_{(i)},\text{adj-}p_{(i-1)}]$;

   se adj-$p_{(i)}\leq\alpha$ si rifiuta $H_{0(i)}$ e si procede con l'algoritmo, altrimenti si accettano $H_{0(i)},\ldots..,H_{0(K)}$ e l'algoritmo termina.

È bene precisare che K rappresenta il numero di possibili confronti a coppie creati a partire dalle sei tribù oggetto di studio, in questo caso pari a 15; $p_{(1)}$ indica il p-value relativo all'ipotesi minimale $H_{(1)}$ ed è il primo p-value nella sequenza, ordinata in maniera decrescente, dei livelli di significatività osservati. Il relativo p-value aggiustato, invece, è indicato con adj-$p_{(1)}$. Infine, il livello nominale α è stato posto pari al 5%, poiché tale è il livello usualmente considerato nelle applicazioni.

Nella tabella riportata alla pagina successiva vengono evidenziati solo i p-value aggiustati per i quali sono stati ottenuti valori inferiori o uguali ad α=5%, calcolati in base all'algoritmo appena descritto.

Tabella 15: P-value aggiustati, calcolati in base all'algoritmo step-down di Bonferroni-Holm, con valore $\leq\alpha=5\%$.

$H_{(1)}$	$H_{(2)}$	$H_{(3)}$	$H_{(4)}$	$H_{(5)}$	$H_{(6)}$
Samburu vs. Kamba	Samburu vs. Turkana	Samburu vs. Rendille	Samburu vs. Masai	Masai vs. Kamba	Ol Molo vs. Kamba
adj-$p_{(1)}$	adj-$p_{(2)}$	adj-$p_{(3)}$	adj-$p_{(4)}$	adj-$p_{(5)}$	adj-$p_{(6)}$
0.000	0.000	0.000	0.000	0.000	0.000

Come si può notare, i p-value aggiustati relativi ai primi sei confronti a coppie tra tribù suggeriscono di rifiutare l'ipotesi nulla di uguaglianza in eterogeneità.

Ciò significa che le popolazioni comparate all'interno di ciascuna coppia risultano significativamente dissomiglianti tra loro dal punto di vista genetico e, pertanto, non possono essere aggregate in uno stesso gruppo, ma devono essere assegnate a due gruppi distinti.

In particolare, si evidenzia che l'etnia Samburu è quella che si differenzia più significativamente dalle altre, risultando maggiormente eterogenea rispetto alle popolazioni Kamba, Turkana, Rendille e Masai.

È da tener presente, infatti, che l'etnia Samburu è l'unica a offrire tutti i 16 fenotipi possibili e, quindi, a possedere una struttura genetica piuttosto complessa. Di conseguenza, tale tribù può essere attribuita ad un gruppo separato, del quale non potranno far parte le sopra citate etnie Kamba, Turkana, Rendille e Masai.

Si può osservare, inoltre, che anche la popolazione Kamba si distingue dalle etnie Masai e Ol Molo, oltre che Samburu, risultando meno eterogenea, dal punto di vista genetico, rispetto a queste.

La popolazione Kamba, infatti, è quella che offre il minor numero di combinazioni fenotipiche. Pertanto, anche i Kamba, come precedentemente detto per i Samburu, verranno assegnati ad un gruppo differente in modo da isolarli dalle etnie da cui risultano dissomiglianti in eterogeneità.

Le rimanenti quattro tribù, invece, sono state aggregate tra loro in rapporto al numero di fenotipi presenti nelle loro strutture genetiche.

A tal proposito si ricorda che sia gli Ol Molo che i Masai presentano 11 tra le 16 possibili combinazioni fenotipiche, mentre i Rendille e i Turkana ne offrono rispettivamente 9 e 8.

Inoltre, le tribù Ol Molo e Masai sono risultate maggiormente eterogenee rispetto ai Kamba, in riferimento ai p-value aggiustati precedentemente calcolati, come pure diversamente eterogenee rispetto all'etnia Samburu. Per questi motivi, le due popolazioni, Ol Molo e Masai, sono state riunite in un gruppo a se stante.

Allo stesso modo, le popolazioni Rendille e Turkana sono state aggregate in un altro gruppo, distinto. Quest'ultime, infatti, oltre a presentare un analogo numero di fenotipi, sono risultate somiglianti anche per altri aspetti, tra i quali lo stile di vita condotto, il territorio geografico condiviso in parte, e la capacità di adattamento ad un ambiente ostile, come discusso nel già citato capitolo 1.

Nel successivo paragrafo è stato effettuato uno studio per verificare l'ipotesi di omogeneità o dissomiglianza tra i vari gruppi, all'interno dei quali le etnie sono state distribuite.

5.3 ANALISI DELLA VARIANZA

Per verificare l'adeguatezza della suddivisione delle etnie in gruppi, descritta nel precedente paragrafo, è stata condotta un'analisi della varianza ad un fattore (ANOVA). Con tale studio, infatti, si vuole accertare se ci sono differenze in media tra i quattro gruppi creati, ossia se questi sono omogenei tra loro o, come ci si dovrebbe aspettare, dissomiglianti.

Il sistema d'ipotesi da testare, quindi, è così espresso:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$ vs. H_1 : almeno una delle uguaglianze è falsa

dove μ_1, μ_2, μ_3 e μ_4 rappresentano, relativamente ai quattro gruppi creati, le medie campionarie delle popolazioni, prese in maniera aggregata all'interno di ciascun gruppo, mentre μ è la media campionaria calcolata sulle sei popolazioni, prese globalmente.

Quindi, se H_0 dovesse risultare ammissibile, la distinzione effettuata delle etnie in gruppi sarebbe inesatta. Infatti, l'ipotesi nulla esprime l'uguaglianza in media tra i quattro gruppi, ossia che le varie etnie, pur suddivise fra loro, non si differenzino dal punto di vista genetico. Ciò significa che, sotto H_0, le osservazioni dei quattro raggruppamenti di popolazioni sono scambiabili tra loro.

È bene notare che la variabile rilevata è di tipo categoriale, in quanto misura le frequenze delle 16 diverse combinazioni fenotipiche presenti nei campioni sierologici.

Quindi, per poter condurre lo studio sull'analisi della varianza, si è reso necessario creare un file dati, nel quale i quattro gruppi di etnie creati sono stati inseriti in maniera sequenziale. All'interno di ciascun gruppo, poi, sono state aggregate le classi delle popolazioni appartenenti al gruppo stesso, ordinate per frequenze decrescenti. Infine, sono stati assegnati i ranghi alle classi categoriali così ordinate.

In questo modo, infatti, a partire dalla variabile categoriale iniziale, si è potuto ottenere una variabile quantitativa, contenente i ranghi assegnati alle classi categoriali ordinate, rendendo così possibile l'applicazione dell'ANOVA.

La statistica test utilizzata per saggiare il sistema d'ipotesi formulato è il test F, con distribuzione nulla asintotica F di Fisher con (K-1) e (n-K) gradi di libertà.

La statistica F calcolata è definita da

$$F = \frac{\left(\tilde{\delta}^2 - \hat{\delta}^2\right) \Big/ (K-1)}{\tilde{\delta}^2 \Big/ (n-K)}$$

dove K rappresenta il numero di gruppi formati e n è la numerosità totale.

Inoltre si fa notare che

$$\tilde{\delta}^2 - \hat{\delta}^2 = \frac{1}{n}\sum_{k=1}^{K} n_k \left(\bar{y}_k - \bar{y} \right)^2$$

rappresenta la 'varianza tra i gruppi', mentre

$$\tilde{\delta}^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} \left(y_{ik} - \bar{y}_k \right)^2$$

è la 'varianza entro i gruppi', con n_k che indica la numerosità di ciascun gruppo, y_{ik} indica la i-esima osservazione del k-esimo gruppo e, infine, $\bar{y}_k$ e $\bar{y}$ rappresentano rispettivamente la media campionaria del k-esimo gruppo e la media campionaria calcolata su tutte le popolazioni, dei k gruppi, prese congiuntamente.

La procedura prende il nome di analisi della varianza poiché la statistica test si basa proprio sul confronto fra la 'varianza tra i gruppi ' e la 'varianza entro i gruppi', che sono le due componenti della varianza totale:

$$\hat{\delta}^2 = \frac{1}{n}\sum_{k=1i=1}^{K}\sum^{n_k} \left(y_{ik} - \bar{y} \right)^2 = \left(\tilde{\delta}^2 - \hat{\delta}^2 \right) + \tilde{\delta}^2$$

Una volta creato il file dati, come descritto precedentemente, è stata calcolata la statistica F per la quale si è ottenuto un valore pari a 36.23. Per poter stabilire se l'ipotesi nulla di uguaglianza in media tra i quattro gruppi può essere accettata o meno, tale valore calcolato deve essere confrontato con il novantacinquesimo percentile di una F di Fisher con (K-1) e (n-K) gradi di libertà. In questo caso, i gradi di libertà sono 3 e 360 e il novantacinquesimo percentile di una F di Fisher così definita è pari a 2.63.

Poiché il valore calcolato per la statistica F è superiore a 2.63, l'ipotesi nulla viene rifiutata. Ciò significa che la suddivisione effettuata delle etnie coglie al meglio le differenze genetiche tra le sei popolazioni oggetto di studio. I quattro raggruppamenti, così creati, risultano dissomiglianti tra loro e le etnie sono state adeguatamente ripartite tra i gruppi in modo tale da mettere in evidenza la loro diversità in eterogeneità.

5.4 I QUATTRO GRUPPI DI ETNIE

Nonostante i pochi dati disponibili, è stato possibile individuare quattro principali raggruppamenti di etnie, tenuto conto della diversa struttura genetica presente nelle sei popolazioni considerate. La selezione naturale, in rapporto all'ambiente, e la storia passata di questi popoli, infatti, hanno influito molto sia sull'evoluzione e la comparsa di alcuni fenotipi e caratteri ereditari, che sulla involuzione o regressione di altri.

Da una parte la popolazione Samburu, isolata all'interno di un gruppo a se stante, domina in eterogeneità le altre tribù e si distingue per la presenza, nel proprio patrimonio genetico, di tutte le 16 possibili combinazioni fenotipiche, create a partire dalla presenza o assenza dei 4 fattori Gm considerati nello studio antropologico svolto da C. Corrain. Tale etnia, pur essendo formata prevalentemente da pastori nomadi, come le tribù Rendille, Turkana, Ol Molo e Masai di cui si parlerà più avanti, si differenzia dagli altri autoctoni per essere l'unica ad avere una struttura genetica completa. Tale eterogeneità di fenotipi presenti è sicuramente dovuta ai geni e caratteri ereditati dalle tribù preesistenti, da cui i Samburu discendono. Un forte contributo a questa complessità genetica è dato anche dal nomadismo e dai continui spostamenti, che permettono loro di avere numerosi contatti con altre etnie.

Un altro gruppo è stato formato con l'aggregazione delle due etnie Ol Molo e Masai, le quali sono risultate molto somiglianti fra loro dal punto di vista genetico. Entrambe, infatti, presentano un equivalente numero, 11, e tipologia di combinazioni fenotipiche offerte. Gli Ol Molo e i Masai conducono entrambi uno stile di vita semi-nomade. Le loro esistenze sono sempre state messe a dura prova a causa delle estreme condizioni ambientali e dei territori aspri e accidentati in cui vivono. Per questi motivi sono state entrambe oggetto di una forte selezione naturale. I Masai, famosi per la loro reputazione di temibili guerrieri, vivono relegati in territori sempre più ridotti e confinati. Gli Ol Molo,

costretti a continui spostamenti in cerca di terreni più fertili per il proprio bestiame, sono il popolo meno numeroso del Kenya e, col passare del tempo, stanno lentamente scomparendo.

Anche le tribù Rendille e Turkana sono state riunite in un unico gruppo. Si è potuto riscontrare, negli studi precedenti, che tali popolazioni hanno una somigliante struttura genetica caratterizzata dalla presenza, rispettivamente, di 9 e 8 combinazioni fenotipiche tra le 16 possibili. Inoltre queste due etnie sono accomunate anche dall'inserimento su territori adiacenti, caratterizzati da terreni aridi e da un clima ostile che rende impossibile qualsiasi attività agricola ed ostacola anche l'allevamento, loro principale attività di sostentamento, costringendoli ad assidui spostamenti.

Infine la popolazione Kamba, assegnata ad un gruppo distinto dagli altri, è caratterizzata, contrariamente ai Samburu, da una bassa eterogeneità genetica, inferiore a quella presente in tutte le altre tribù, dovuta alle poche combinazioni fenotipiche possedute, solo 4 tra le 16 possibili. Si ricorda, inoltre, che i Kamba non hanno subito forti selezioni naturali in rapporto all'ambiente. Insediatisi fin da subito nei territori in cui tuttora risiedono, non sono stati costretti a continui spostamenti proprio grazie alla fertilità dell'ambiente abitato. La produttività dei terreni, infatti, ha permesso ai Kamba di differenziarsi ancor di più dalle altre tribù per essere gli unici a praticare l'attività agricola, resa possibile anche dalla loro stanzialità. Per questi motivi e per la forte endogamia verso le altre tribù, i Kamba hanno pochi e sporadici rapporti con altre popolazioni. Da ciò forse deriva il basso numero di fenotipi presenti nel loro patrimonio genetico.

CONCLUSIONI

Obiettivo di questa tesi è il confronto in eterogeneità tra due popolazioni, valutato sulla base di dati campionari. La genetica, difatti, è una tra le discipline scientifiche in cui l'eterogeneità risulta di rilevante interesse, specialmente per la valutazione della biodiversità.

A tal proposito è stato considerato uno studio antropologico, svolto da C. Corrain nel 1975, su sei popolazioni pastorali del Kenya, da cui vennero raccolti altrettanti campioni sierologici.

Le sei etnie sono state confrontate, da un punto di vista genetico, analizzando le combinazioni fenotipiche di 4 fattori Gm e osservando le frequenze con cui questi fenotipi si trovano presenti all'interno delle sei tribù considerate: Samburu, Rendille, Turkana, Ol Molo, Masai e Kamba.

Dapprima l'eterogeneità è stata trattata da un punto di vista descrittivo attraverso un'analisi grafica, in modo da individuare eventuali dissomiglianze tra i vari gruppi etnici kenioti. Da tale analisi è emerso che le popolazioni Samburu e Kamba si distinguono significativamente risultando, rispettivamente, la più eterogenea e la più omogenea in rapporto alle altre esaminate. Le rimanenti etnie Rendille, Turkana, Ol Molo e Masai, sono risultate, invece, abbastanza somiglianti fra loro dal punto di vista genetico, per numerosità di combinazioni fenotipiche offerte.

In seguito, l'argomento è stato trattato da un punto di vista inferenziale, confrontando l'eterogeneità campionaria tra coppie di tribù, utilizzando alcuni indici di eterogeneità.

Non essendoci la cognizione della vera distribuzione dei dati relativi alle combinazioni fenotipiche presenti nelle sei popolazioni keniote, sono state utilizzate le tecniche non parametriche.

In particolare, si è voluto testare, prima di tutto, l'ipotesi di dissomiglianza in distribuzione tra coppie di tribù. A tal proposito sono state usate la statistica X^2 di Pearson

$$X^2 = \sum_{k=1}^{K} \sum_{j=1}^{C} \frac{\left(f_{kj} - \frac{f_{k\cdot} f_{\cdot j}}{n} \right)}{\frac{f_{k\cdot} f_{\cdot j}}{n}}$$

e la statistica, che in questa tesi è stata denominata X^2_{AD}, corrispondente alla divergenza, secondo Anderson-Darling, tra le distribuzioni di frequenza

$$X^2_{AD} = \sum_{k=1}^{K} \frac{\left(\frac{f_{k1}}{n_1} - \frac{f_{k2}}{n_2} \right)^2}{\frac{f_{k.}}{n}\left(1 - \frac{f_{k.}}{n} \right)}$$

Successivamente è stato verificato se, oltre ad una diversità in distribuzione tra le etnie, ci fosse anche dominanza in eterogeneità da parte di popolazioni nomadi nei confronti di popolazioni più stanziali. Lo studio proposto, quindi, consiste nel determinare appropriate statistiche test e sistemi d'ipotesi da testare, basati sull'ordinamento delle probabilità attribuite alle diverse classi categoriali corrispondenti alle combinazioni fenotipiche. Il fatto che le probabilità delle due distribuzioni confrontate sono parametri non noti e, quindi, l'ordinamento delle probabilità può essere stimato solo sulla base di dati campionari, implica che le soluzioni proposte e i risultati inferenziali ottenuti sono approssimati. Per stimare le probabilità sono state usate le frequenze relative osservate.

La statistica test, utilizzata per la verifica di dominanza in eterogeneità tra coppie di tribù, corrisponde alla differenza degli indici di eterogeneità campionari calcolati per le due popolazioni messe a confronto, e tali test possono variare in base all'indicatore di eterogeneità considerato. In particolare, sono state prese in considerazione le statistiche test basate sull'indice di eterogeneità di Gini,

$$T_G = G_1 - G_2 = \sum_{k=1}^{K} \left(\hat{p}_{2(k)}^{\,2} - \hat{p}_{1(k)}^{\,2} \right)$$

sull'indice di entropia di Shannon,

$$T_S = S_1 - S_2 = \sum_{k=1}^{K} \left[\hat{p}_{2(k)} log\left(\hat{p}_{2(k)} \right) - \hat{p}_{1(k)} log\left(\hat{p}_{1(k)} \right) \right]$$

e sull'indice di entropia generalizzato di ordine infinito di Renyi,

$$T_{R\infty} = R_{\infty 1} - R_{\infty 2} = log\left(max_k \, \hat{p}_{2(k)} \right) - log\left(max_k \, \hat{p}_{1(k)} \right)$$

Dallo studio effettuato sulla dominanza in eterogeneità, è risultato che l'etnia Samburu non solo presenta una struttura genetica diversa da quella delle altre tribù con le quali è stata confrontata, ma tale struttura è addirittura più complessa e completa in quanto contiene tutte le 16 possibili combinazioni fenotipiche. Anche i Kamba sono risultati diversamente eterogenei rispetto ad alcune delle altre etnie comparate ma, contrariamente ai Samburu, possiedono una struttura genetica molto semplice, composta da soli 4 fenotipi. Le etnie Turkana, Rendille, Ol Molo e Masai, invece, presentano tutte una somigliante eterogeneità genetica e, pertanto, non è stata riscontrata alcuna dominanza da parte di alcune di queste sulle altre etnie.

Per entrambi gli studi condotti, è stato scritto un programma, utilizzando il linguaggio statistico R, che permettesse di risolvere i test X^2 e X^2_{AD} per l'analisi di dissomiglianza in distribuzione, e i test di Gini, Shannon e Renyi per lo studio di dominanza in eterogeneità.

Per valutare l'adeguatezza dei test statistici utilizzati nelle due analisi precedentemente descritte, è stato condotto uno studio di simulazione. In particolare, sono stati stimati il grado di approssimazione dei cinque test e la loro potenza.

La generazione dei dati, per lo studio di simulazione, è avvenuta in base al seguente modello, $X = 1 + Int\left[K U^{\delta}\right]$. In questo modo, infatti, è possibile generare distribuzioni discrete con differenti gradi di eterogeneità attribuendo disparati valori al singolo parametro reale di eterogeneità δ.

Dapprima sono stati valutati il grado di approssimazione e la potenza delle statistiche X^2 e X^2_{AD}, risultate entrambe ben approssimate e con una buona potenza. Le loro prestazioni sono molto simili, anche se si evidenzia una sottile preferenza per la statistica X^2, i cui tassi di rigetto si approssimano meglio ai livelli nominali α.

In seguito sono stati stimati il grado di approssimazione e la potenza degli indici di eterogeneità di Gini, Shannon e Renyi di ordine infinito. Fra le tre statistiche considerate, quella basata sull'indice di Renyi si è dimostrata maggiormente conservativa, ma con una potenza inferiore. Al contrario, il test basato sull'indicatore di Gini è risultato sostanzialmente ben approssimato, con prestazioni simili a quelle ottenute per il test basato sull'indice di Shannon, ma con una potenza maggiore rispetto a quest'ultimo. Di conseguenza, l'indice di eterogeneità di Gini sembra essere il più consono per uno studio sul confronto, in eterogeneità genetica, tra coppie di popolazioni.

Anche per lo studio di simulazione è stato scritto un programma, utilizzando il linguaggio statistico R, che calcolasse, per i cinque test, i tassi di rigetto di H_0, sia sotto l'ipotesi nulla che sotto l'ipotesi alternativa.

Infine, è stato condotto uno studio per stabilire quali etnie risultassero maggiormente somiglianti fra loro e significativamente diverse da altre, per poterle suddividere in gruppi omogenei al loro interno, dal punto di vista genetico, ed eterogenei tra loro. Nel

far ciò è stata utilizzata la procedura di Bonferroni-Holm, applicata ai p-value calcolati, per la statistica di Gini, su tutti i possibili confronti a coppie di popolazioni, realtivi allo studio di dominanza in eterogeneità. Con tale strategia è stato possibile individuare quattro principali gruppi di tribù.

Isolata all'interno di un gruppo a se stante, la popolazione Samburu domina in eterogeneità le altre tribù. Questa etnia si distingue per essere l'unica ad offrire tutte le 16 possibili combinazioni fenotipiche. Ciò è dovuto sicuramente alla loro discendenza da altre tribù preesistenti, oltre che al loro nomadismo e ai continui spostamenti, che permettono loro di avere numerosi contatti con altre etnie, con conseguenti scambi di geni e trasmissioni di caratteri ereditari.

In contrapposizione alla sopra citata etnia c'è quella dei Kamba, anch'essa assegnata ad un gruppo distinto, caratterizzata da una bassa eterogeneità genetica dovuta alla presenza di sole 4 combinazioni fenotipiche. Essa è l'unica popolazione stanziale, tra le sei considerate, e questa sua caratteristica è determinata dalla fertilità del terreno in cui tale etnia vive. Ciò ha fatto sì che i Kamba non subissero forti selezioni naturali e che si differenziassero ancor più dagli altri autoctoni per essere gli unici a praticare l'attività agricola.

Un altro gruppo è stato formato con l'aggregazione delle etnie Ol Molo e Masai, risultate molto somiglianti tra loro per l'equivalente numero e tipologia di combinazioni fenotipiche offerte.

Infine, anche le tribù Rendille e Turkana sono state riunite in un unico gruppo. Somiglianti per struttura genetica, dato l'analogo numero di fenotipi offerti, queste sono ac-

comunate anche per l'inserimento in territori adiacenti, caratterizzati da ostili condizioni ambientali e climatiche. La selezione naturale in rapporto all'ambiente, difatti, è il primo elemento a determinare l'evoluzione di un gruppo razziale.

L'adeguatezza di tale suddivisione delle etnie in gruppi è stata, successivamente, accertata attraverso un'analisi della varianza ad un fattore, per verificare la presenza di differenze in media tra i quattro raggruppamenti.

APPENDICE

PROGRAMMA ‘TSD2SER.TXT’

```
leggi<-function(classik,nm,d,ntot) {
nome<-choose.files(filters = Filters[c("txt", "All"),])
while(length(nome)==0)
{
  cat("Non è stato trovato nessun file dati.\n")
  risposta <- readline("Riprovare? Si digiti s o n ")
  capture.output(risposta)
  while(substr(risposta, 1, 1)!="n" && substr(risposta, 1, 1)!="s")
  {
  if (substr(risposta, 1, 1) != "n" && substr(risposta, 1, 1) != "s")
    cat("comando non valido.\n")
  risposta <- readline("Riprovare? Si digiti s o n ")
  capture.output(risposta)
  }
  if (substr(risposta, 1, 1) == "n")
  {
    classik<-0
    nm<-0
    d<-"non ci sono dati"
    ntot<-0
    cat("Programma terminato.\n")
    options("show.error.messages"=FALSE)
    return(list(d,classik,nm,ntot))
  }
  if (substr(risposta, 1, 1) == "s")
  nome<-choose.files(filters = Filters[c("txt", "All"),])
}
d<-read.table(nome,header=TRUE)
attach(d)
classik<-nrow(d)
ncolonne<-ncol(d)
j=1
nm<-vector(length=ncolonne)
while (j<=ncolonne)
{
 nm[j]<-sum(d[,j])
 j<-j+1
}
ntot<-sum(nm)
return(list(d,classik,nm,ntot))
}

frequenze<-function(fno,fo,fmarg,dati) {
fno<-matrix(nrow=nrow(dati),ncol=ncol(dati))
fo<-matrix(nrow=nrow(dati),ncol=ncol(dati))
```

```
fmarg<-array(vector(length=nrow(dati)),c(nrow(dati),1))
j=1
while (j<=ncol(dati))
{
  for (i in 1:nrow(dati))
  {
  fno[i,j]<-dati[i,j]
  fo[i,j]<-dati[i,j]
  fmarg[i]<-fmarg[i]+fno[i,j]
  }
j<-j+1
}
return(list(fno,fo,fmarg))
}

etichette<-function(e,numtot) {
e<-array(vector(length=numtot),c(numtot,1))
i=1
while (i<=numtot)
{
  e[i]<-i
 i<-i+1
}
return(e)
}

ripetino<-function(freqno,t,numtot,dati) {
t<-vector(length=numtot)
l<-1
for (j in 1:ncol(dati))
{
  for (i in 1:nrow(dati))
  {
   m<-freqno[i,j]
   while(m>0)
   {
    t[l]<-i
    l<-l+1
    m<-m-1
   }
  }
}
return(t)
}

ordina<-function(freqo,k,dati)
{
temp<-vector(length=k)
for ( j in 1:ncol(dati))
{
  temp<-freqo[,j]
  temp<-sort(temp,decreasing=TRUE)
  freqo[,j]<-temp
}
return(freqo)
}
```

```
tester<-
function(freqno,freqo,freqmarg,numtot,numcamp,dati,k,XP,XR,XS,XG,XR2)
{
XP<-0
XR<-0
XS<-0
XG<-0
XQ<-0
XR2<-0
n<-numcamp[1]
m<-numcamp[2]
for ( i in 1:k)
{
  f1<-freqno[i,1]
  f2<-freqno[i,2]
  n1<-freqmarg[i]
  p2<-n1*m/numtot
  p1<-n1*n/numtot
  if (p1 > 0)
  {
    XQ<-(f1-p1)^2/p1+(f2-p2)^2/p2
  }
  else
  {
    XQ<-0
  }
  XP<-XP+XQ
  if (n1 > 0  && (numtot-n1)>0)
  {
    XQ<-(numtot*(f1/n-f2/m))^2/(n1*(numtot-n1))
  }
  else
  {
    XQ<-0
  }
  XR<-XR+XQ
  p1<-freqo[i,1]/(n+5/n)
  p2<-freqo[i,2]/(m+5/m)
  XG<-XG+p2^2-p1^2
  if (p1 > 0)
  {
    XS1<-p1*log(p1)
  }
  else
  {
    XS1<-0
  }
  if (p2 > 0)
  {
    XS2<-p2*log(p2)
  }
  else
  {
    XS2<-0
  }
  XS<-XS+XS2-XS1
```

```
}
sup1<-max(freqo[,1])/n
sup2<-max(freqo[,2])/m
XR2<-log(sup2)-log(sup1)
return(list(XP,XR,XS,XG,XR2))
}

simper<-
func-
tion(NSnum,PTP,PTR,PTS,PTG,PTR2,numtot,numcamp,k,eti,temp,ttemp,freqno
,freqo,dati,freqmarg,XGo,XSo,XPo,XRo,XR2o)
{
PTG<-0
PTR<-0
PTS<-0
PTP<-0
PTR2<-0
for (hs in 1:NSnum)
{
  for(hb in 1:1000)
  {
    f9<-permet(eti,numtot,u,numcamp)
    eti<-f9[[1]]
    n<-f9[[2]]
    f10<-freqpermut(temp,eti,freqno,freqmarg,dati,n,numtot,k,numcamp)
    freqno<-f10[[1]]
    freqmarg<-f10[[2]]
    f11<-freqpermut(ttemp,eti,freqo,freqmarg,dati,n,numtot,k,numcamp)
    freqo<-f11[[1]]
    f12<-
tester(freqno,freqo,freqmarg,numtot,numcamp,dati,k,XP,XR,XS,XG,XR2)
    XG<-f12[[4]]
    XS<-f12[[3]]
    XR<-f12[[2]]
    XP<-f12[[1]]
    XR2<-f12[[5]]
    if (XG>=XGo)
    {
      PTG<-PTG+1
    }
    if (XS>=XSo)
    {
      PTS<-PTS+1
    }
    if (XP>=XPo)
    {
      PTP<-PTP+1
    }
    if (XR>=XRo)
    {
      PTR<-PTR+1
    }
    if (XR2>=XR2o)
    {
      PTR2<-PTR2+1
    }
```

```
  }
}
PTG<-(PTG/1000)/NSnum
PTR<-(PTR/1000)/NSnum
PTS<-(PTS/1000)/NSnum
PTP<-(PTP/1000)/NSnum
PTR2<-(PTR2/1000)/NSnum
return(list(PTG,PTR,PTS,PTP,PTR2))
}

permet<-function(e2,numtot,u,numcamp)
{
n<-numcamp[1]
for (i in 1:numtot)
{
RNGkind("Mersenne-Twister")
u<-runif(1,min=0,max=1)
l<-e2[i]
j<-trunc(u*i)+1
e2[i]<-e2[j]
e2[j]<-l
}
return(list(e2,n))
}

freqpermut<-function(temp,eti,freqno,freqmarg,dati,n,numtot,k,numcamp)
{
 for (j in 1:ncol(dati))
 {
   for (i in 1:k)
   {
    freqno[i,j]<-0
   }
 }
 for (i in 1:k)
 {
   freqmarg[i]<-0
 }
 xx<-1
 yy<-NULL
 for (j in 1:ncol(dati))
 {
  if ((j-1)<=0)
  {
    yy<-numcamp[j]
  }
  else
  {
    yy<-numcamp[j-1]+numcamp[j]
  }
  for (i in xx:yy)
  {
   z<-eti[i]
   h<-temp[z]
   freqno[h,j]<-freqno[h,j]+1
  }
```

```
  xx<-numcamp[j]+1

 }
 j<-1
 while (j<=ncol(dati))
 {
  for (i in 1:k)
  {
   freqmarg[i]<-freqmarg[i]+freqno[i,j]
  }
 j<-j+1
 }
 return(list(freqno,freqmarg))
}

inizio <- function() {
cat("Test di dissomiglianza (Pearson 1 e 2) e di dominanza in eteroge-
neità \n")
cat("(Shannon Gini e Renyi di ordine infinito) per due campioni.\n")
cat("Test di permutazione per: P(X) =/= P(Y)  e  Het(X) >= Het(Y)")
cat("\n")
cat("Pearson 1 = chi.quadro \n")
cat("Pearson 2= Sum((p1i-p21)^2/(ni*(nt-ni)) \n")
cat("Shannon = Sum(p1i*log(p1i)-p2i*log(p2i)) \n")
cat("Gini = Sum(p1i^2-p2i^2) \n")
cat("Renyi = log(max(p1i))-log(max(p2i)) \n")
cat("\n")
cat("Questo programma risolve,via simulazione condizionata (permuta-
zione) \n")
cat("il test di dissomiglianza in distribuzione e di dominanza in \n")
cat("eterogeneità su k <= 256 classi nominali (non ordinate) per due
\n")
cat("campioni per rispettivamente:\n")
cat("          H0 = P(X)=P(Y)              H1 = P(X)=/=P(Y)\n")
cat("          H0 = Het(X)=Het(Y)          H1 = Het(X)>Het(Y)\n")
cat("\n")
cat("Le frequenze vanno inserite da file. La dimensione del file (som-
ma \n")
cat("delle numerosità dei campioni) deve essere inferiore a 32000. I
dati \n")
cat("sono i k records con le frequenze di classe.\n")
cat("Il programma chiede:\n")
cat("     il numero B di replicazioni delle permutazioni\n")
cat("Il programma visualizza:\n")
cat("     il file dati\n")
cat("     il numero k di classi\n")
cat("     la numerosità dei due campioni\n")
cat("     i valori-p dei tre test\n")
cat("\n")
cat("\n")
cat("\n")
  ptm1 <- proc.time()
  cat("Tempo all'inizio:  ",ptm1,"\n")
cat("\n")
  cat("Ora verrà chiesto all'utente se desidera proseguire col pro-
gramma e di \n selezionare il nome del file dati da aprire.\n")
```

```
  risposta <- readline("Si vuole proseguire? Digita s per continuare
oppure n per terminare ")
capture.output(risposta)
while(substr(risposta, 1, 1)!="n" && substr(risposta, 1, 1)!="s")
{
  if (substr(risposta, 1, 1) != "n" && substr(risposta, 1, 1) != "s")
    cat("comando non valido.\n")
  risposta <- readline("Si vuole proseguire? Digita s per continuare
oppure n per terminare ")
  capture.output(risposta)
}
if (substr(risposta, 1, 1) == "n")
    cat("programma terminato \n")
if (substr(risposta, 1, 1) == "s")
{
  cat("\n")
  cat("Test di dissomiglianza (Pearson 1 e 2) e di dominanza in etero-
geneità \n")
  cat("(Shannon e Gini) per due campioni.\n")
  cat("Test di permutazione per: P(X) =/= P(Y)  e  Het(X) >= Het(Y)")
  cat("\n")
  cat("Pearson 1 = chi.quadro \n")
  cat("Pearson 2= Sum((p1i-p21)^2/(ni*(nt-ni)) \n")
  cat("Shannon = Sum(p1i*log(p1i)-p2i*log(p2i)) \n")
  cat("Gini = Sum(p1i^2-p2i^2) \n")
  cat("Renyi = log(max(p1i))-log(max(p2i)) \n")
  cat("\n")
  cat("\n")

  dati<-as.matrix
  dati<-NULL
  k<-NULL
  numcamp<-as.vector
  numcamp<-NULL
  numtot<-NULL

  f1<-leggi(classik,nm,d,ntot)
  dati<-f1[[1]]
  k<-f1[[2]]
  numcamp<-f1[[3]]
  numtot<-f1[[4]]
  cat("\n")
  cat("Dati Kenya Corrain, ")
  cat(names(dati)[1])
  cat(" VS ")
  cat(names(dati)[2])
  cat("\n")
  cat("Classi nominali contenute:")
  print(k)
  cat("\n")
  cat("Numerosità campionarie:")
  print(numcamp)
  cat("\n")
  cat("Numerosità totale: ")
  print(numtot)
  cat("\n")
```

```
  freqno<-as.matrix
  freqo<-as.matrix
  freqmarg<-as.matrix
  freqno<-NULL
  freqo<-NULL
  freqmarg<-NULL

  f2<-frequenze(fno,fo,fmarg,dati)
  freqno<-f2[[1]]
  freqo<-f2[[2]]
  freqmarg<-f2[[3]]

  eti<-as.matrix
  eti<-NULL

  f3<-etichette(e,numtot)
  eti<-f3

  temp<-as.matrix
  temp<-NULL

  f4<-ripetino(freqno,t,numtot,dati)
  temp<-f4

  f5<-ordina(freqo,k,dati)
  freqo<-f5

  f6<-ripetino(freqo,t,numtot,dati)
  ttemp<-f6

  XPo<-NULL
  XRo<-NULL
  XSo<-NULL
  XGo<-NULL
  XR2o<-NULL

  f7<-
tester(freqno,freqo,freqmarg,numtot,numcamp,dati,k,XP,XR,XS,XG,XR2)
  XPo<-f7[[1]]
  XRo<-f7[[2]]
  XSo<-f7[[3]]
  XGo<-f7[[4]]
  XR2o<-f7[[5]]
  cat("I test di dissomiglianza di Pearson 1 e 2 calcolati sui dati
\n")
  cat("contenuti nella matrice delle frequenze in entrata non ordinate
\n")
  cat("risultano pari a:\n")
  print(XPo)
  print(XRo)
  cat("\n")
  cat("I test di dominanza in eterogeneità di Shannon, Gini e Renyi di
ordine infinito \n")
  cat("calcolati sui dati contenuti nella matrice delle frequenze or-
dinate \n")
  cat("risultano pari a:\n")
```

```
  print(XSo)
  print(XGo)
  print(XR2o)
  cat("\n")

  risposta2 <- readline("Inserire il numero di simulazioni condiziona-
te in migliaia:\n")
  capture.output(risposta2)
  NS<-substr(risposta2, 1, 3)
  NSnum<-as(NS,"numeric")

  PTP<-NULL
  PTR<-NULL
  PTS<-NULL
  PTG<-NULL
  PTR2<-NULL
     f8<-
sim-
per(NSnum,PTP,PTR,PTS,PTG,PTR2,numtot,numcamp,k,eti,temp,ttemp,freqno,
freqo,dati,freqmarg,XGo,XSo,XPo,XRo,XR2o)
  PTP<-f8[[4]]
  PTR<-f8[[2]]
  PTS<-f8[[3]]
  PTG<-f8[[1]]
  PTR2<-f8[[5]]
  cat("Livelli -p dei quattro test:\n")
  print(PTP)
  cat("\n")
  print(PTR)
  cat("\n")
  print(PTS)
  cat("\n")
  print(PTG)
  cat("\n")
  print(PTR2)
  cat("\n")
}
ptm2 <- proc.time()
cat("Tempo finale:  ",ptm2,"\n")
tempo<-ptm2-ptm1
cat("Tempo trascorso:   ",tempo,"\n")
}
inizio()
```

PROGRAMMA ‘STET4T-D.TXT’

```
etichette<-function(e,numtot) {
e<-array(vector(length=numtot),c(numtot,1))
for (i in 1:numtot)
{
  e[i]<-i
}
return(e)
}

anf<-function(dtemp,fno,fmarg,n,numtot,k,eti)
{
 for (j in 1:ncol(fno))
 {
   for (i in 1:nrow(fno))
   {
    fno[i,j]<-0
   }
 }
  for (i in 1:k)
 {
   fmarg[i]<-0
 }
 for (i in 1:n)
 {
   j<-eti[i]
   h<-dtemp[j]
   fno[h,1]<-fno[h,1]+1
 }
 for (i in (n+1):numtot)
 {
   j<-eti[i]
   h<-dtemp[j]
   fno[h,2]<-fno[h,2]+1
 }
 for (i in 1:k)
 {
   fmarg[i]<-fno[i,1]+fno[i,2]
 }
 return(list(fno,fmarg))
}

ordina<-function(fo,k)
{
temp<-vector(length=k)
for ( j in 1:ncol(fo))
{
  temp<-fo[,j]
  temp<-sort(temp,decreasing=TRUE)
  fo[,j]<-temp
}
return(fo)
}
```

```
ripetino<-function(fo,tt,numtot) {
l<-1
for (j in 1:ncol(fo))
{
  for (i in 1:nrow(fo))
  {
   m<-fo[i,j]
   while(m>0)
   {
    tt[l]<-i
    l<-l+1
    m<-m-1
   }
  }
}
return(tt)
}

gndatct<-function(dtemp,numtot,n,k,d1,d2,fno,fo,fmarg,tt,eti)
{
dtemp<-array(vector(length=numtot),c(numtot,1))
fno<-matrix(ncol=2,nrow=k)
fo<-matrix(ncol=2,nrow=k)
fmarg<-array(vector(length=k),c(k,1))
tt<-array(vector(length=numtot),c(numtot,1))
for (i in 1:n)
{
  RNGkind("Mersenne-Twister")
  x<-runif(1,min=0,max=1)
  dtemp[i]<-1+trunc(k*(x^d1))
}
for (i in (n+1):numtot)
{
  RNGkind("Mersenne-Twister")
  x<-runif(1,min=0,max=1)
  dtemp[i]<-1+trunc(k*(x^d2))
}
f3<-anf(dtemp,fno,fmarg,n,numtot,k,eti)
fno<-f3[[1]]
fmarg<-f3[[2]]
for ( j in 1:ncol(fno))
{
  for (i in 1:nrow(fno))
  {
    fo[i,j]<-fno[i,j]
  }
}
f4<-ordina(fo,k)
fo<-f4
f5<-ripetino(fo,tt,numtot)
tt<-f5
return(list(dtemp,fno,fmarg,fo,tt))
}

tester<-function(freqno,freqo,freqmarg,numtot,n,k,XP,XR,XS,XG,XR2)
{
```

```
XP<-0
XR<-0
XS<-0
XG<-0
XR2<-0
m<-numtot-n
for ( i in 1:k)
{
  f1<-freqno[i,1]
  f2<-freqno[i,2]
  n1<-freqmarg[i]
  p2<-n1*m/numtot
  p1<-n1*n/numtot
  if (p1 > 0)
  {
    XQ<-(f1-p1)^2/p1+(f2-p2)^2/p2
  }
  else
  {
    XQ<-0
  }
  XP<-XP+XQ
  if (n1 > 0  && (numtot-n1)>0)
  {
    XQ<-(numtot*(f1/n-f2/m))^2/(n1*(numtot-n1))
  }
  else
  {
    XQ<-0
  }
  XR<-XR+XQ
  p1<-freqo[i,1]/(n+5/n)
  p2<-freqo[i,2]/(m+5/m)
  XG<-XG+p2^2-p1^2
  if (p1 > 0)
  {
    XS1<-p1*log(p1)
  }
  else
  {
    XS1<-0
  }
  if (p2 > 0)
  {
    XS2<-p2*log(p2)
  }
  else
  {
    XS2<-0
  }
  XS<-XS+XS2-XS1
}
sup1<-max(freqo[,1])/n
sup2<-max(freqo[,2])/m
XR2<-log(sup2)-log(sup1)
return(list(XP,XR,XS,XG,XR2))
```

```
}

permet<-function(eti,numtot,n)
{
for (i in 1:numtot)
{
RNGkind("Mersenne-Twister")
x<-runif(1,min=0,max=1)
l<-eti[i]
j<-trunc(x*i)+1
eti[i]<-eti[j]
eti[j]<-l
}
return(eti)
}

simper<-
fun-
ction(B,numtot,n,k,dattemp,eti,freqno,freqo,freqmarg,ttemp,XP,XR,XS,XG
,XR2,SMW,SAD,STS,STG,STR2)
{
 for (i in 1:B)
 {
  f8<-permet(eti,numtot,n)
  eti<-f8
  f9<-anf(dattemp,freqno,freqmarg,n,numtot,k,eti)
  freqno<-f9[[1]]
  freqmarg<-f9[[2]]
  f10<-anf(ttemp,freqo,freqmarg,n,numtot,k,eti)
  freqo<-f10[[1]]
  f11<-tester(freqno,freqo,freqmarg,numtot,n,k,XP,XR,XS,XG,XR2)
  XP<-f11[[1]]
  XR<-f11[[2]]
  XG<-f11[[3]]
  XS<-f11[[4]]
  XR2<-f11[[5]]
  SMW[i]<-XP
  SAD[i]<-XR
  STS[i]<-XS
  STG[i]<-XG
  STR2[i]<-XR2
 }
return(list(SMW,SAD,STS,STG,STR2))
}

pliv<-
func-
tion(B,PXP,PXQ,PXS,PTG,PTR2,SMW,SAD,STS,STG,STR2,XPo,XRo,XSo,XGo,XR2o)
{
f13<-livsgn(SMW,B,PXP,XPo)
PXP<-f13
f14<-livsgn(SAD,B,PXQ,XRo)
PXQ<-f14
f15<-livsgn(STS,B,PXS,XSo)
PXS<-f15
f16<-livsgn(STG,B,PTG,XGo)
```

```
PTG<-f16
f17<-livsgn(STR2,B,PTR2,XR2o)
PTR2<-f17
return(list(PXP,PXQ,PXS,PTG,PTR2))
}

livsgn<-function(SMW,B,PXP,XPo)
{
PXP<-XPo
i<-0
for ( j in 1:B)
{
  if (PXP <= SMW[j])
  {
    i<-i+1
  }
}
PXP<-i/B
return(PXP)
}

inizio <- function() {
cat("Simulazione su quattro test di dissomiglianza ed eterogeneità \n
per due campioni \n")
cat("Test di Pearson 1 e 2, Shannon, Gini e Renyi inf: \n")
cat("                      Het(X) >= Het(Y)                      \n")
cat("\n")
cat("\n")
cat("\n")
cat("Il programma chiede:\n")
cat("     n,m (N. Campionarie); k (classi); B (N. Permutazioni) \n")
cat("     d1<=d2 e d1>1.5 (parametri); MC (SMC)\n")
cat("\n")
cat("\n")
cat("     U=RND, x=1+INT(k*U^d1), y=1+INT(k*U^d2)\n")
cat("\n")
cat("\n")
cat("\n")
ptm1 <- proc.time()
inserimento1 <- readline("Numerosità del primo campione:\n  n ")
capture.output(inserimento1)
campn<-substr(inserimento1, 1, 3)
n<-as(campn,"numeric")
while(campn=="" || is.na(n)==TRUE)
{
  cat("La numerosità del primo campione deve essere di tipo numeri-
co!\n")
  inserimento1 <- readline("Numerosità del primo campione:\n  n ")
  capture.output(inserimento1)
  campn<-substr(inserimento1, 1, 3)
  n<-as(campn,"numeric")
}
inserimento2 <- readline("Numerosità del secondo campione:\n  m ")
capture.output(inserimento2)
campm<-substr(inserimento2, 1, 3)
m<-as(campm,"numeric")
```

```
while(campm=="" || is.na(m)==TRUE)
{
  cat("La numerosità del secondo campione deve essere di tipo numeri-
co!\n")
  inserimento2 <- readline("Numerosità del secondoo campione:\n  m ")
  capture.output(inserimento2)
  campm<-substr(inserimento2, 1, 3)
  m<-as(campm,"numeric")
}
numtot<-n+m
inserimento3 <- readline("Numero di classi:\n  k ")
capture.output(inserimento3)
classik<-substr(inserimento3,1,3)
k<-as(classik,"numeric")
while(classik=="" || is.na(k)==TRUE)
{
  cat("La numerosità delle classi deve essere di tipo numerico!\n")
  inserimento3 <- readline("Numero di classi:\n  k ")
  capture.output(inserimento3)
  classik<-substr(inserimento3, 1, 3)
  k<-as(classik,"numeric")
}
k2<-2*k
inserimento4 <- readline("Numero di permutazioni:\n  B ")
capture.output(inserimento4)
permB<-substr(inserimento4,1,6)
B<-as(permB,"numeric")
while(permB=="" || is.na(B)==TRUE)
{
  cat("Il numero di permutazioni deve essere di tipo numerico!\n")
  inserimento4 <- readline("Numero di permutazioni:\n  B ")
  capture.output(inserimento4)
  permB<-substr(inserimento4, 1, 6)
  B<-as(permB,"numeric")
}
inserimento5 <- readline("Parametro della prima distribuzione:\n  d1
")
capture.output(inserimento5)
dd1<-substr(inserimento5, 1, 4)
d1<-as(dd1,"numeric")
while(dd1=="" || is.na(d1)==TRUE)
{
  cat("Il parametro deve essere di tipo numerico!\n")
  inserimento5 <- readline("Parametro della prima distribuzione:\n  d1
")
  capture.output(inserimento5)
  dd1<-substr(inserimento5, 1, 4)
  d1<-as(dd1,"numeric")
}
inserimento6 <- readline("Parametro della seconda distribuzione:\n  d2
")
capture.output(inserimento6)
dd2<-substr(inserimento6, 1, 4)
d2<-as(dd2,"numeric")
while(dd2=="" || is.na(d2)==TRUE)
{
```

```
  cat("Il parametro deve essere di tipo numerico!\nn")
  inserimento6 <- readline("Parametro della seconda distribuzione:\n
d2 ")
  capture.output(inserimento6)
  dd2<-substr(inserimento6, 1, 4)
  d2<-as(dd2,"numeric")
}
inserimento7 <- readline("Numero di simulazioni:\n  MC ")
capture.output(inserimento7)
simMC<-substr(inserimento7,1,6)
MC<-as(simMC,"numeric")
while(simMC=="" || is.na(MC)==TRUE)
{
  cat("Il numero di simulazioni deve essere di tipo numerico!\n")
  inserimento7 <- readline("Numero di simulazioni:\n  MC ")
  capture.output(inserimento7)
  simMC<-substr(inserimento7, 1, 6)
  MC<-as(simMC,"numeric")
}
H<-as.matrix
H<-matrix(nrow=11,ncol=6)
H[1,1]<-0.01
H[2,1]<-0.05
H[3,1]<-0.1
H[4,1]<-0.2
H[5,1]<-0.3
H[6,1]<-0.5
H[7,1]<-0.7
H[8,1]<-0.8
H[9,1]<-0.9
H[10,1]<-0.95
H[11,1]<-0.99
for (cc in 2:ncol(H))
{
  for (rr in 1:nrow(H))
  {
     H[rr,cc]<-0
  }
}

for (hs in 1:MC)
{
  RNGkind("Mersenne-Twister")
  u<-runif(1,min=0,max=1)
  eti<-as.matrix
  eti<-NULL

  f1<-etichette(e,numtot)
  eti<-f1

  dattemp<-as.vector
  freqno<-as.matrix
  freqo<-as.matrix
  freqmarg<-as.vector
  ttemp<-as.matrix
  dattemp<-NULL
```

```
  freqno<-NULL
  freqo<-NULL
  freqmarg<-NULL
  ttemp<-NULL

  f2<-gndatct(dtemp,numtot,n,k,d1,d2,fno,fo,fmarg,tt,eti)
  dattemp<-f2[[1]]
  freqno<-f2[[2]]
  freqmarg<-f2[[3]]
  freqo<-f2[[4]]
  ttemp<-f2[[5]]

  XP<-NULL
  XR<-NULL
  XS<-NULL
  XG<-NULL
  XR2<-NULL
  XPo<-NULL
  XRo<-NULL
  XSo<-NULL
  XGo<-NULL
  XR2o<-NULL

  f6<-tester(freqno,freqo,freqmarg,numtot,n,k,XP,XR,XS,XG,XR2)
  XP<-f6[[1]]
  XR<-f6[[2]]
  XG<-f6[[3]]
  XS<-f6[[4]]
  XR2<-f6[[5]]

  SMW<-vector(length=B)
  SAD<-vector(length=B)
  STS<-vector(length=B)
  STG<-vector(length=B)
  STR2<-vector(length=B)
  SMW<-NULL
  SAD<-NULL
  STS<-NULL
  STG<-NULL
  STR2<-NULL

  XPo<-XP
  XRo<-XR
  XGo<-XG
  XSo<-XS
  XR2o<-XR2

  f7<-
sim-
per(B,numtot,n,k,dattemp,eti,freqno,freqo,freqmarg,ttemp,XP,XR,XS,XG,X
R2,SMW,SAD,STS,STG,STR2)
  SMW<-f7[[1]]
  SAD<-f7[[2]]
  STS<-f7[[3]]
  STG<-f7[[4]]
  STR2<-f7[[5]]
```

```
  PXP<-vector(length=B)
  PXQ<-vector(length=B)
  PXS<-vector(length=B)
  PTG<-vector(length=B)
  PTR2<-vector(length=B)
  PXP<-NULL
  PXQ<-NULL
  PXS<-NULL
  PTG<-NULL
  PTR2<-NULL

  f12<-
pliv(B,PXP,PXQ,PXS,PTG,PTR2,SMW,SAD,STS,STG,STR2,XPo,XRo,XSo,XGo,XR2o)
  PXP<-f12[[1]]
  PXQ<-f12[[2]]
  PXS<-f12[[3]]
  PTG<-f12[[4]]
  PTR2<-f12[[5]]

  for (h in 1:11)
  {
    if (PXP <= H[h,1])
    {
      H[h,2]<-H[h,2]+1
    }
    if (PXQ <= H[h,1])
    {
      H[h,3]<-H[h,3]+1
    }
    if (PXS <= H[h,1])
    {
      H[h,4]<-H[h,4]+1
    }
    if (PTG <= H[h,1])
    {
      H[h,5]<-H[h,5]+1
    }
    if (PTR2 <= H[h,1])
    {
      H[h,6]<-H[h,6]+1
    }
  }
}
for(cc in 2:ncol(H))
{
  for (rr in 1:nrow(H))
  {
    H[rr,cc]<-H[rr,cc]/MC
  }
}
cat("\n")
cat("\n")
cat("    L.s. T: Chi-1  Chi-2 Shan   Gini  Renyi(inf)\n")
print(H)

ptm2 <- proc.time()
```

```
cat("Tempo finale:  ",ptm2,"\n")
tempo<-ptm2-ptm1
cat("Tempo trascorso:   ",tempo,"\n")
}


inizio()
```

RIFERIMENTI BIBLIOGRAFICI

Brunner E. e Munzel U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal*; **42**, 17-25.

Chieffi G., Dolfini S., Malcovati M., Pierantoni R. e Tenchini M. L. (2005). *Biologia e Genetica*. EdiSES, Napoli.

Corrain C., Mezzavilla F., Pesarin F. e Scardellato U. (1977). Il valore discriminativo di alcuni fattori Gm, tra le popolazioni pastorali del Kenya. In *Atti e Memorie dell'Accademia Patavina di Scienze, Lettere ed Arti*, LXXXIX, Parte II: Classe di Scienze Matematiche e Naturali, 55-63.

Frosini B. V. (1976). Sulle distribuzioni campionarie di due indici di eterogeneità. *Statistica*; **38**, 43-57.

Landenna G. e Marasini D. (1990). *Metodi statistici non parametrici*. Il Mulino, Bologna.

Novelli G. e Giardina E. (2003). *Genetica medica pratica*. Aracne, Roma.

Pace L. e Salvan A. (2001). *Introduzione alla statistica* - II Inferenza, verosimiglianza, modelli. CEDAM, Padova.

Pesarin F. (2001). *Multivariate permutation tests with applications in biostatistics*. Wiley, Chichester.

Pesarin F. (2005). Permutation tests: Multivariate, In Balakrishna, Johnson and Kotz eds. *Encyclopedia of Statistical Sciences*, 2nd Ed., Wiley, New York.

Pesarin F. e Salmaso L. A permutation approach for testing heterogeneity in two-sample problems. (submitted).

Pesarin F. e Salmaso L. Permutation test for the comparison of heterogeneity. (submitted).

Pesarin F. e Salmaso L. (2006). Permutation tests for univariate and multivariate ordered categorical data. *Austrian Journal of Statistics*; **35**, 315-324.

Randles R. H. e Wolfe D. A. (1979). *Introduction to the theory of nonparametric statistics*. Wiley, New York.

Silvapulle M. J. e Sen P. K. (2005). *Constrained Statistical Inference, Inequality, Order, and Shape Restrictions.* Wiley, New York.

Siti web consultati:

http://www.moldrek.com/africa_info_02.htm

http://fc.retecivica.milano.it