Questioni di questionari*

Roma, 11 maggio 2010 - Università degli Studi di Roma Tre

prof. Stefano M. Pagnotta

Università degli studi del Sannio/Benevento

pagnotta@unisannio.it

Questionario e variabili

- Le variabili sono lo strumento attraverso cui si misura un fenomeno reale, . . . ad esempio sociologico

- In generale, quanto più è complesso il fenomeno tante più sono le variabili che bisogna coinvolgere

- Le variabili che misurano il fenomeno vengono definite a-priori e diventano numeri/parole quando sono poi osservate sulle unità statistiche

- Un questionario riguardante un tema/fenomeno generalmente complesso) è una collezioni di variabili che vengono osservate quando si somministra il questionario.

Preparazione dei dati

La preparazione dei dati per le analisi è una fase

1) imprescindibile

2) onerosa

Quel che generalmente accade è che i dati siano trascritti male (soprattutto quando questa operazione è fatta da più persone).

Succede che

- i generi si moltiplichino: Maschio, femmina, f, FEMMINA, maschio, M, machio, ...
- si rinvengono persone di 121 anni, oppure alte 75cm e che pesano 80Kg ...

Preparazione dei dati

Prima di ogni analisi statistica utile a estrarre informazioni, tutte le variabili devono essere controllate, ovvero i sistemi di modalità devono essere normalizzati

- Maschio, maschio, M, machio, ... $\rightarrow$ m/1
- femmina, f, FEMMINA, ... $\rightarrow$ f/0

Lo strumento statistico per controllare i dati, variabile per variabile, è la costruzione della distribuzione di frequenza

Mancate risposte

Accade che su alcuni questionari manchino delle risposte. Esse si traducono nei dati mancanti.

I dati mancanti possono distinguersi in due tipi principali: strutturali e volontari

- una mancata risposta è strutturale quando il rispondente "non deve" rispondere a una domanda

esempio

domanda [FILTRO] 5) se guadagni più di 100'000E al mese vai alla domanda 51.

domanda 6) quanto riesci a risparmiare ogni mese?

In questo caso, se la persona guadagna più di 100'000E annue NON DEVE rispondere alla domanda 6 (e alcune successive), in forze della struttura del questionario.

Mancate risposte

Accade che su alcuni questionari manchino delle risposte. Esse si traducono nei dati mancanti.

I dati mancanti possono distinguersi in due tipi principali: strutturali e volontari

- una mancata risposta è volontaria quando il rispondente "non vuole" rispondere a una domanda

esempio

domanda 10) quanto guadagni al mese?

In questo caso, la persona si rifiuta di (non vuole) rispondere alla domanda per cui si genera un valore mancante.

Mancate risposte

È il caso di codificare in modo distinto i due tipi di mancate risposte

- una mancata risposta strutturale $\rightarrow$ ''null'' /-1
- una mancata risposta volontaria $\rightarrow$ ''na'' /999

Le domande filtro (come la 5) che generano valori mancanti strutturali individuano sottopopolazioni fra i rispondenti, ne consegue che le analisi sulle relative risposte devono essere analizzate separatamente.

"Gli esami non finiscono mai",

diceva Eduardo De Filippo in una sua famosa commedia. Ed è proprio così. Infatti l'uomo, per sua natura, non è mai soddisfatto delle posizioni raggiunte, anzi cerca sempre di superare se stesso per migliorare. Ma l'ansia che sopraggiunge ad ogni nuovo "esame" da superare è sempre in agguato? Se ad esempio la prossima sfida che vogliamo intraprendere con noi stessi fosse quella di partecipare ad un quiz televisivo e diventare "noi" miliardari, dovremo sicuramente fare i conti con battiti cardiaci, sudorazione eccessiva, nodo allo stomaco e gola asciutta? Provate a cimentarvi in questo piccolo test per capire se siete molto ansiosi o se invece potreste andare sotto i riflettori di uno studio televisivo con una buona tranquillità d'animo e avere dunque maggiori chance.

1 Stai passeggiando nel parco quando a distanza ti sembra ti vedere un'ombra che si muove.
a) Qualcuno ti sta spiando
b) Forse c'è un cane nascosto tra i cespugli
c) Sarà stata la tua immaginazione

2 Una dura giornata di lavoro sta per cominciare. Come ti senti?
a) Teso, come sempre d'altronde
b) Un po' intimidito ma pronto ad iniziare
c) Sicuro di te

3 Cosa ti succede quando ti senti preda dell'ansia?
a) Sei disturbato sia mentalmente che psicologicamente
b) Provi un vago senso di paura, anche se non sapresti dire di cosa
c) Ti senti frustrato perché non sei riuscito a portare a termine i tuoi impegni

4 È arrivato il momento del tuo primo esame all'università.
a) Non hai dormito tutta la notte per ripetere nuovamente il programma
b) Il giorno prima ti sei preso qualche ora di relax per riposare
c) Sei convinto che riuscirai a superarlo senza grandi difficoltà

5 Ogni volta che esci con quel tuo amico così tranquillo e sicuro di sé pensi che...
a) Tanta tranquillità non farà mai parte della tua vita
b) Stai bene con lui proprio perché ti infonde calma e sicurezza
c) Sia un po' troppo borioso e ti riprometti di smettere di frequentarlo

6 Sei a casa a dormire da solo quando senti dei rumori.
a) Ti alzi per controllare che sia tutto a posto
b) Resti in attesa per cercare di capire da dove provengono
c) Saranno sicuramente quei rompiscatole dei vicini

7 La giornata in ufficio è stata particolarmente difficile. Alla sera ti senti...
a) Irrequieto. Non riesci a prendere sonno in alcun modo
b) Stanco ma soddisfatto per l'ottimo lavoro compiuto
c) Pronto ad uscire con gli amici per una pizza

8 Il tuo partner è uscito con i colleghi d'ufficio. Tu resti in casa a leggere un buon libro e...
a) Il tempo sembra fermo e tu non riesci a concentrati su ciò che leggi
b) Ti ritrovi in men che non si dica ad aver letto cinquanta pagine
c) Anche se il libro è interessante, ti sembra comunque di star sprecando il tuo tempo

9 Capita spesso che al mattino, dopo una buona dormita, tu ti senta...
a) Stanco come e più di prima
b) Sufficientemente riposato
c) Pronto ad affrontare una nuova giornata

10 Nei momenti in cui l'ansia ti assale tu...
a) Rimugini su tristi pensieri
b) Cerchi di capire come porvi rimedio
c) Ti riempi di impegni per ignorare il disagio

Leggi il tuo profilo a pag. 49

La matrice dei dati

L'organizzazione naturale dei dati raccolti attraverso il questionario è la matrice, ovvero l'organizzazione per colonna (variabili) e per riga (osservazioni) di quanto ottenuto come risposta nella somministrazione.

Uno	Due	Tre	Qua	Cin	Sei	Set	Ott	Nov	Die	gen	età	CDL
1B	2B	3C	4B	5A	6B	7C	8B	9B	aB	M	25	BIO
1B	2B	3A	4B	5C	6A	7B	8B	9B	aA	F	19	BIO
1C	2C	3A	4A	5A	6B	7B	8A	9B	aA	F	28	BIO
1B	2C	3C	4B	5B	6B	7B	8A	9B	aC	M	25	BIO
1C	2A	3B	4A	5A	6B	7C	8B	9B	aC	F	20	BIO
1B	2B	3A	4B	5C	6B	7B	8B	9B	aC	F	25	BIO
1C	2C	3A	4B	5B	6B	7C	8A	9A	aB	M	20	BIO
1B	2C	3B	4A	5C	6B	7B	8C	9B	aC	F	23	BIO
⋮									⋮	⋮		⋮

Analisi

I dati ottenuti dal questionario ammettono analisi a diverso livello di complessità.

- analisi di una singola variabile
- analisi di gruppi di variabili contemporaneamente
- analisi di tutte le variabili contemporaneamente

Analisi

Analizzare una variabile significa

- controllare che sia osservata con coerenza
- studiare la distribuzione delle frequenze (diagramma a barre/istogramma/boxplot)
  - per le variabili qualitative decidere se si devono riclassificare le risposte (modalità) e quindi, nel caso ristudiare il sistema delle frequenze
  - per le variabili quantitative valutare se sia il caso di trasformarle in qualitative
- individuare l'elemento tipico della distribuzione
- calcolare la variabilità
- *raccontare* la variabile tendo conto della forma della distribuzione, dell'elemento tipico e della variabilità

I profili

Si consideri la distribuzione doppia di frequenze relative

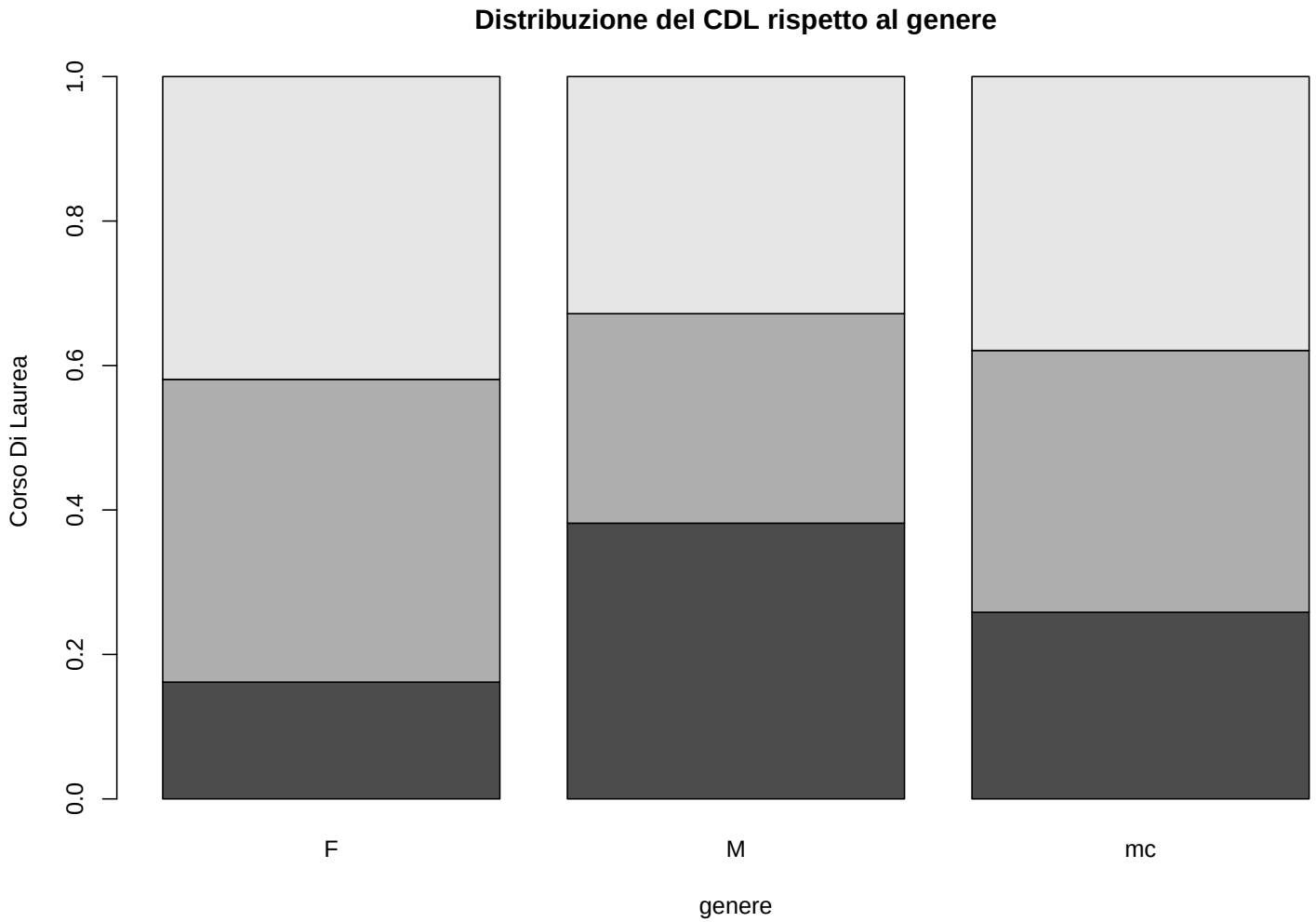
	etc	GIUR	STAT	mr
F	27	70	70	167
M	50	38	43	131
mc	77	108	113	298

Le distribuzioni condizionate di una variabile (in colonna) rispetto alle modalità dell'altra (in riga) si chiamano anche profili.

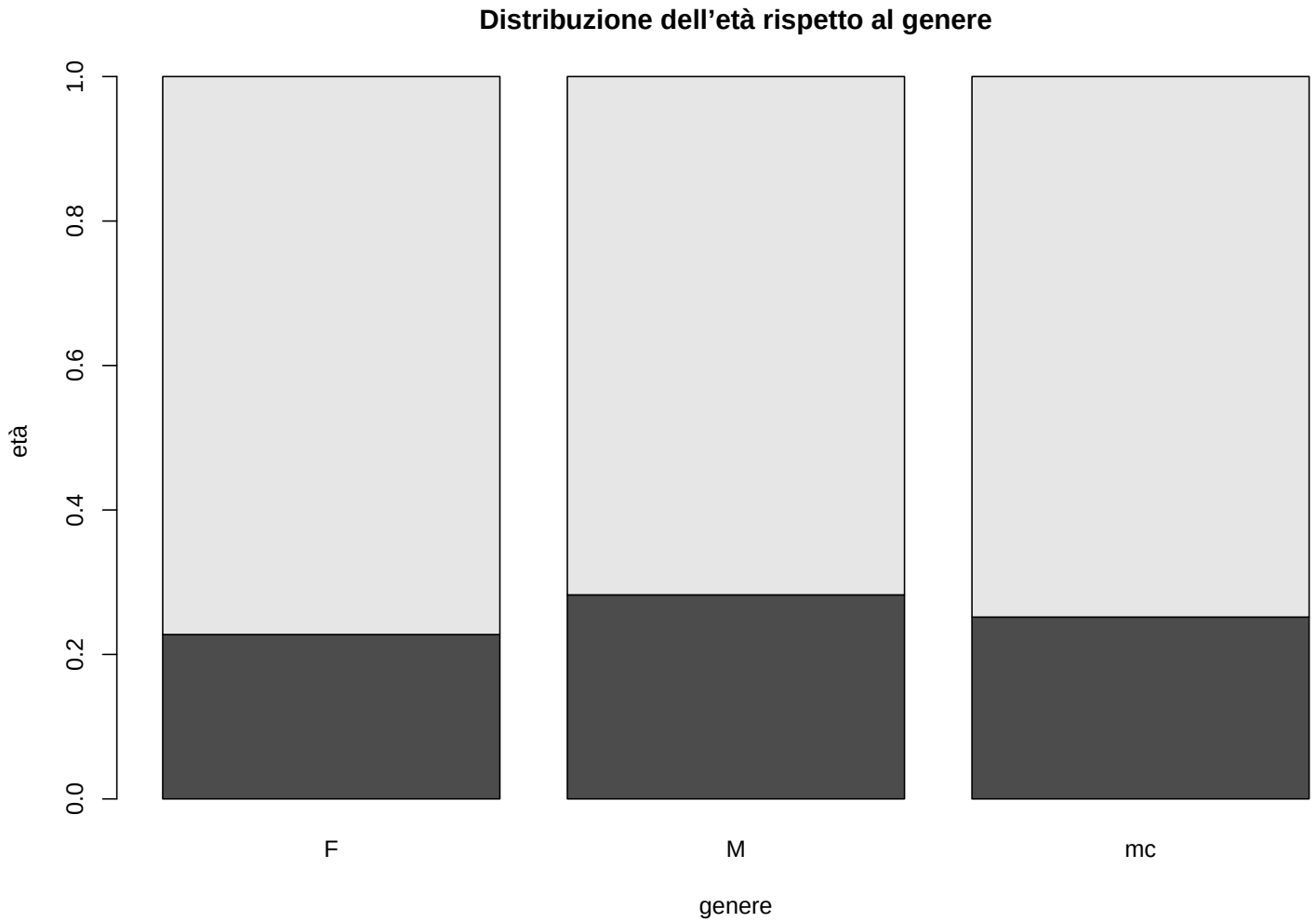
	etc	GIUR	STAT	mr
F	27/167	70/167	70/167	167/167
M	50/131	38/131	43/131	131/131
mc	77/298	108/298	113/298	298/298

	etc	GIUR	STAT	mr
F	16.17	41.92	41.92	100
M	38.17	29.01	32.82	100
mc	25.84	36.24	37.92	100

I profili


Distribuzione del CDL rispetto al genere
Corso Di Laurea
0.0
0.2
0.4
0.6
0.8
1.0
F
M
mc
genere

I profili


Distribuzione dell'età rispetto al genere
età
0.0
0.2
0.4
0.6
0.8
1.0
F
M
mc
genere

I profili

- Quando il profilo (sistemi di frequenze relative) della variabile in colonna cambia a seconda della modalità (condizionate) della variabile in riga, allora fra le variabili c'è dipendenza.

- Quando, al contrario, i profili sono pressochè simili fra loro al variare delle modalità condizionanti, allora c'è indipendenza.

Un profilo di riferimento è il sistema delle frequenze relative delle variabile in colonna (marginale colonna).

La variabilità

La nozione di variabilità per le variabili qualitative si chiama eterogeneità e si basa sulla diversità con cui si osserva la variabile.

L'eterogeneità si misura attraverso le frequenze relative associate alle modalità.

- la minima diversità (eterogeneità) si ha quando tutta la massa di frequenza è concentrata su una sola modalità

- la massima diversità si ha quando la massa di frequenza è ripartita in modo uguale su tutte le modalità

entropia di Shannon

$$H = \frac{-1}{\log K} \sum_{k=1}^{K} f_k \log f_k$$

- $H = 0$ quando tutta la massa di frequenza è concentrata su una sola modalità (assenza di variabilità/diversità)

- $H = 1$ quando tutte le frequenze sono uguali (massima variabilità)

entropia di Shannon

$$H = \frac{-1}{\log K} \sum_{k=1}^{K} f_k \log f_k$$

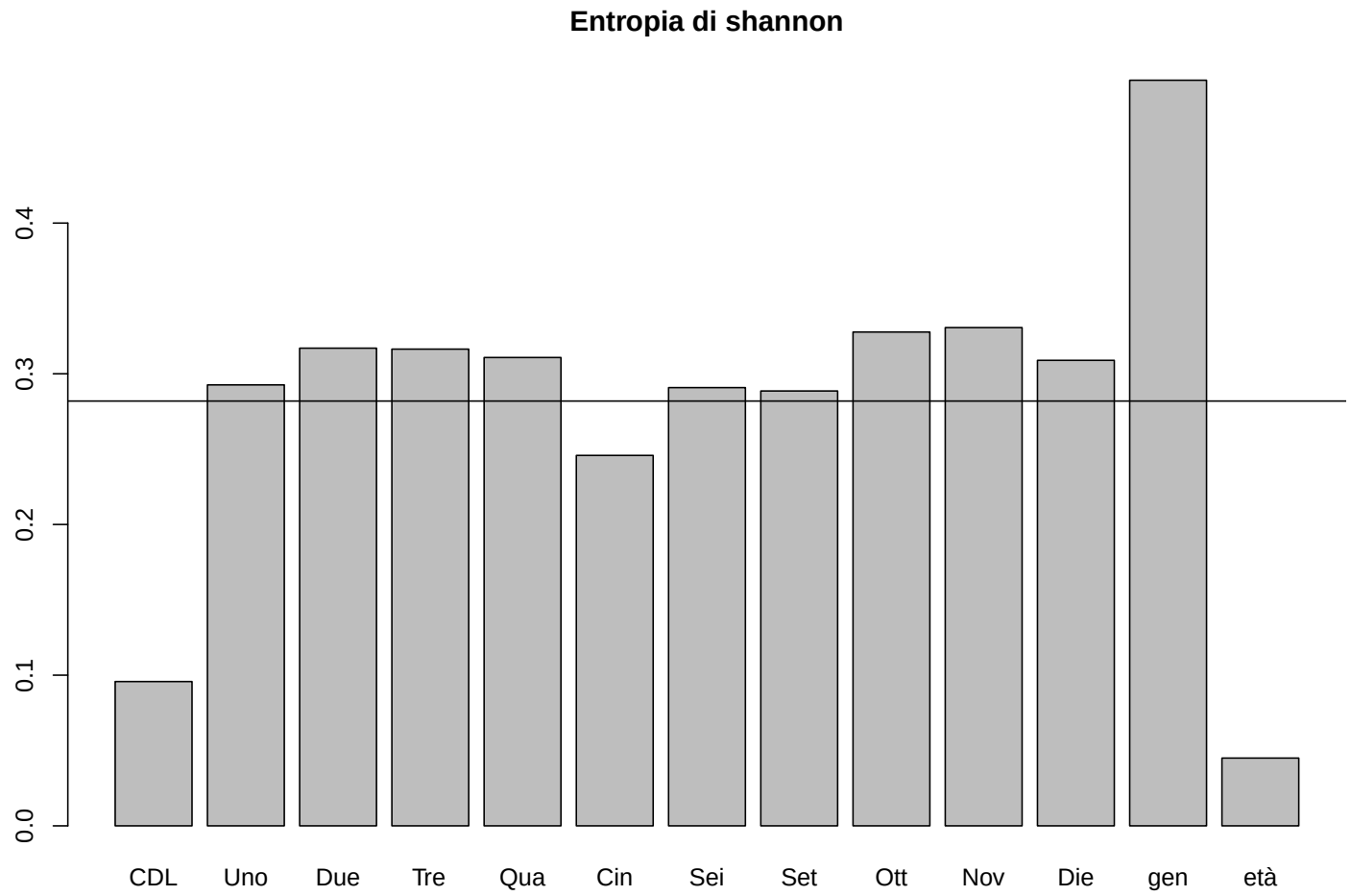
Se $H \approx 0$ bisogna prendere in considerazione l'idea di scartare la variabile!

Se la variabilità è **bassa** si può pensare di **aggregare** delle modalità per farla diventare più alta

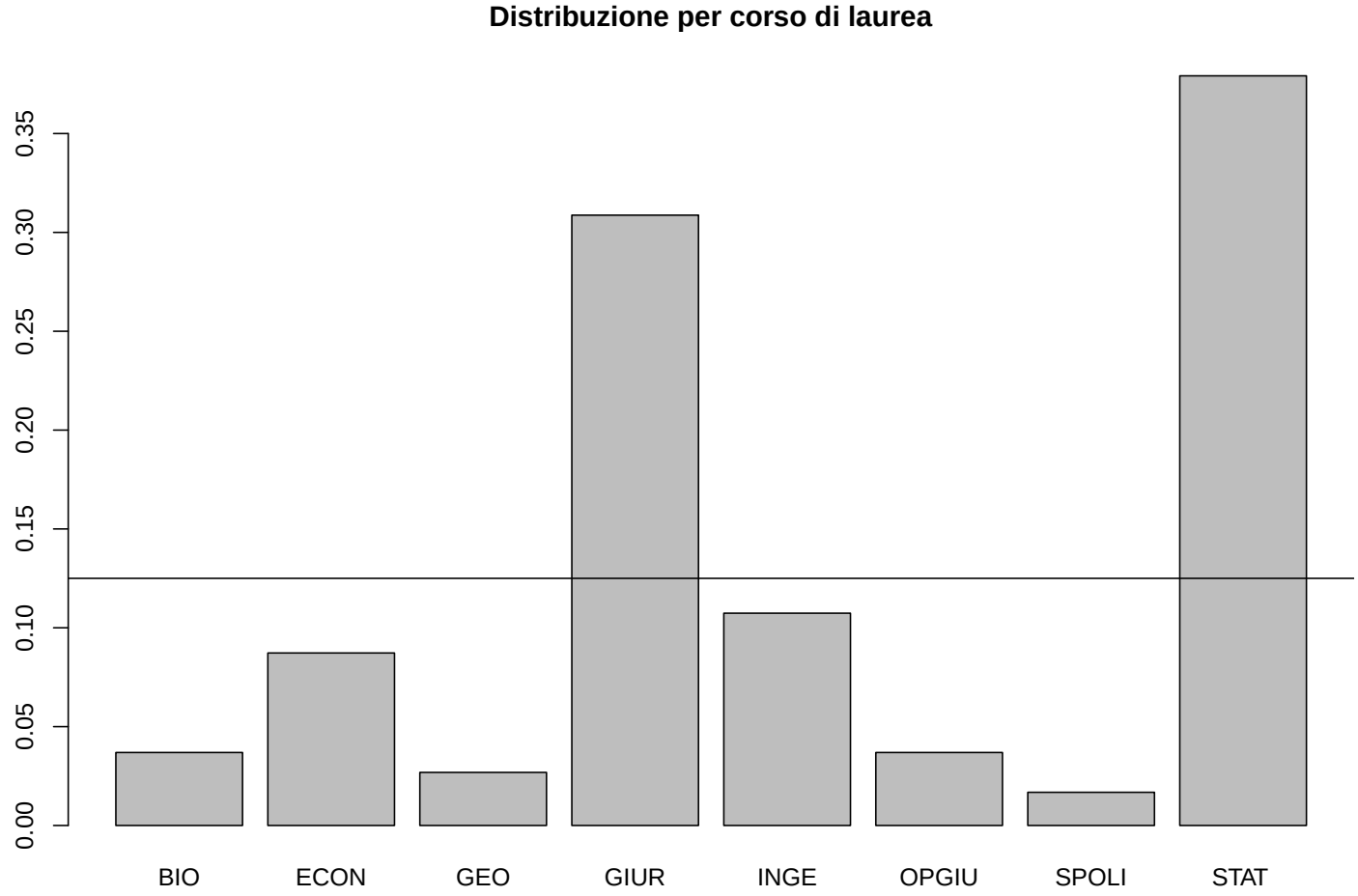
- ... si aggregano le modalità con frequenza piccola rispettando la coerenza dell'insieme delle modalità

- ... quanto bassa? Quando la H di una variabile è molto inferiore della media delle H calcolata su tutte le variabili.

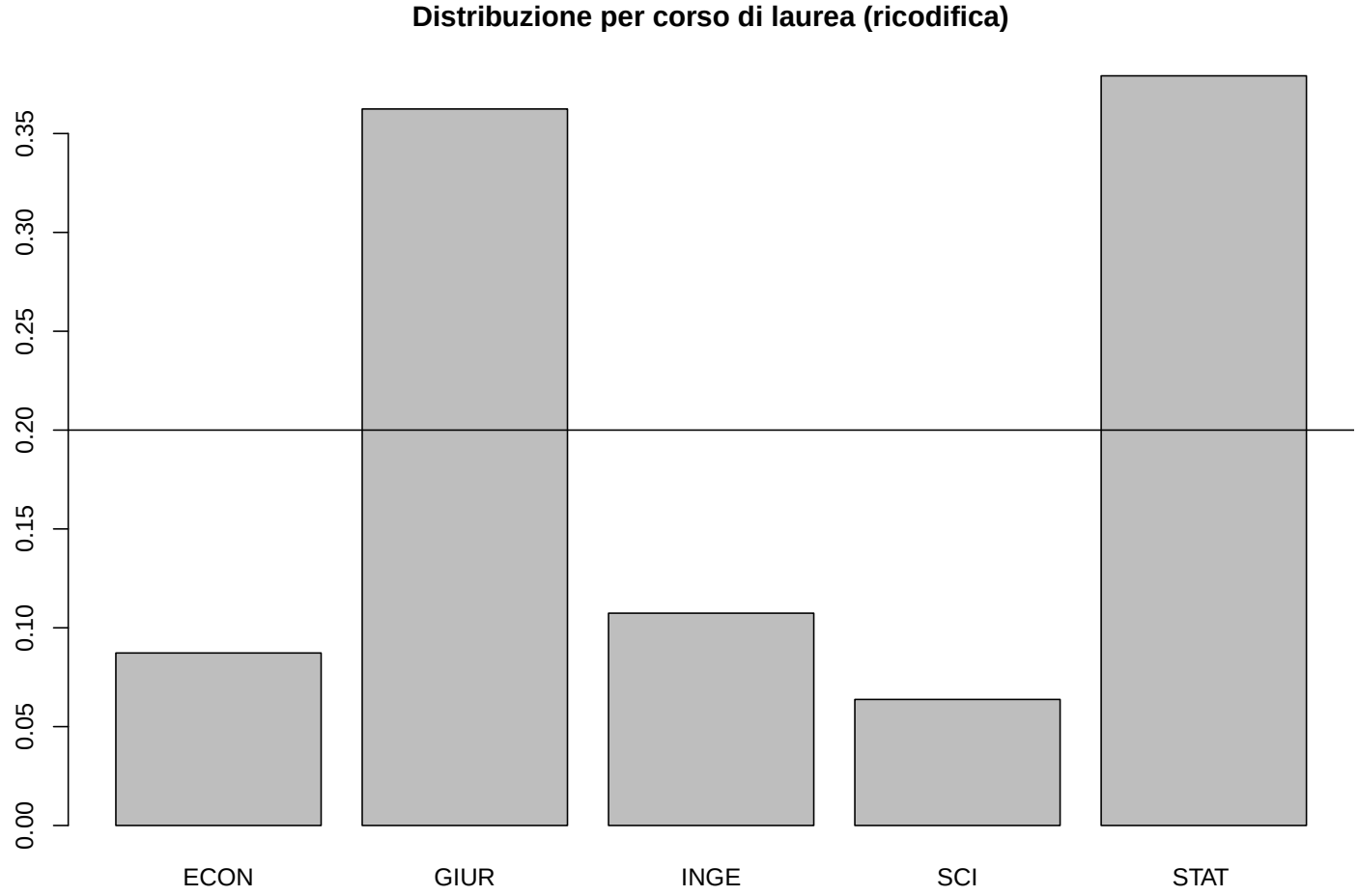
Entropia su tutte le variabili ($\bar{H} = 0.28$)



Corso di laurea ($H = 0.096$)

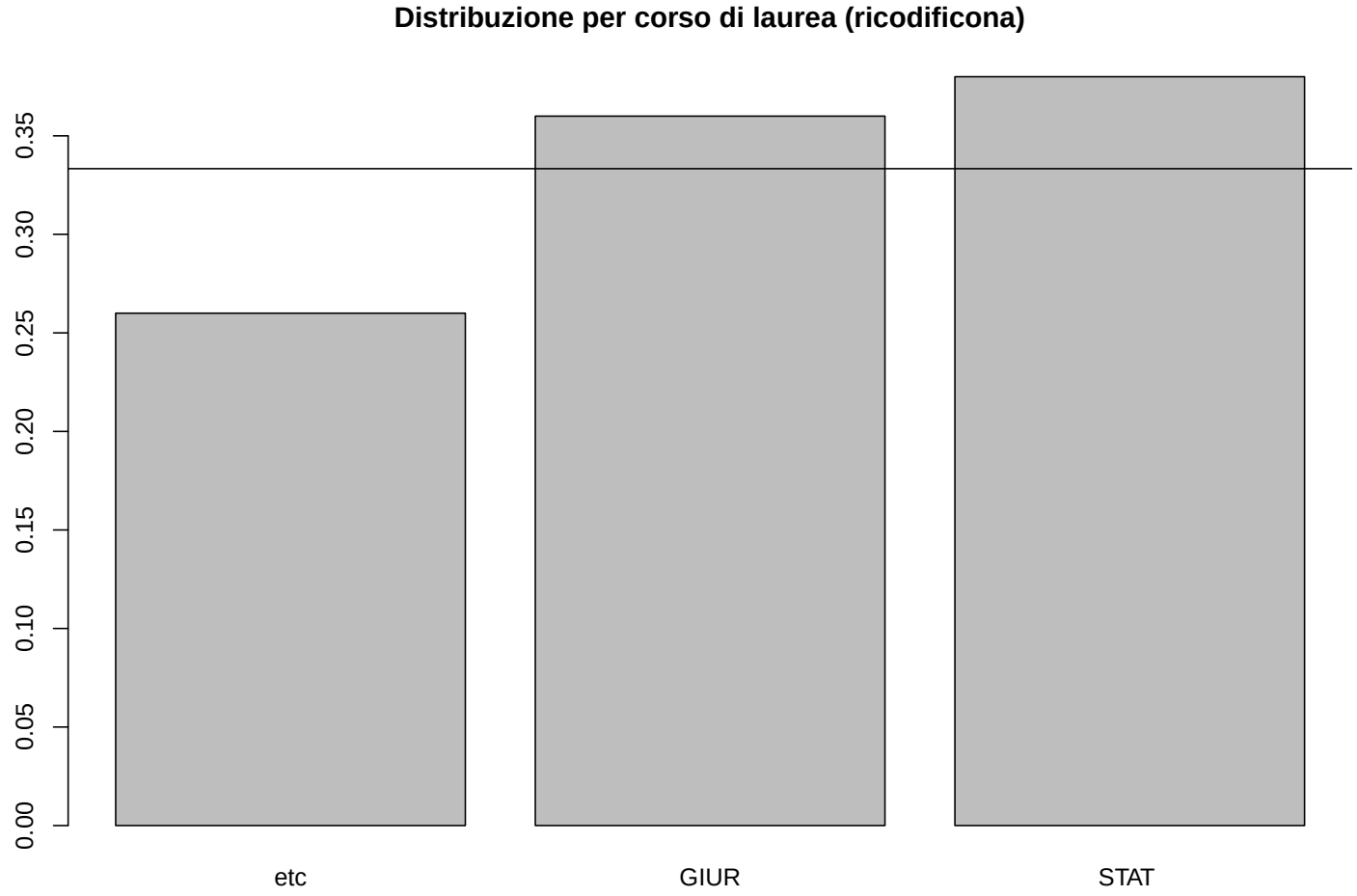


Corso di laurea ($H = 0.169$)



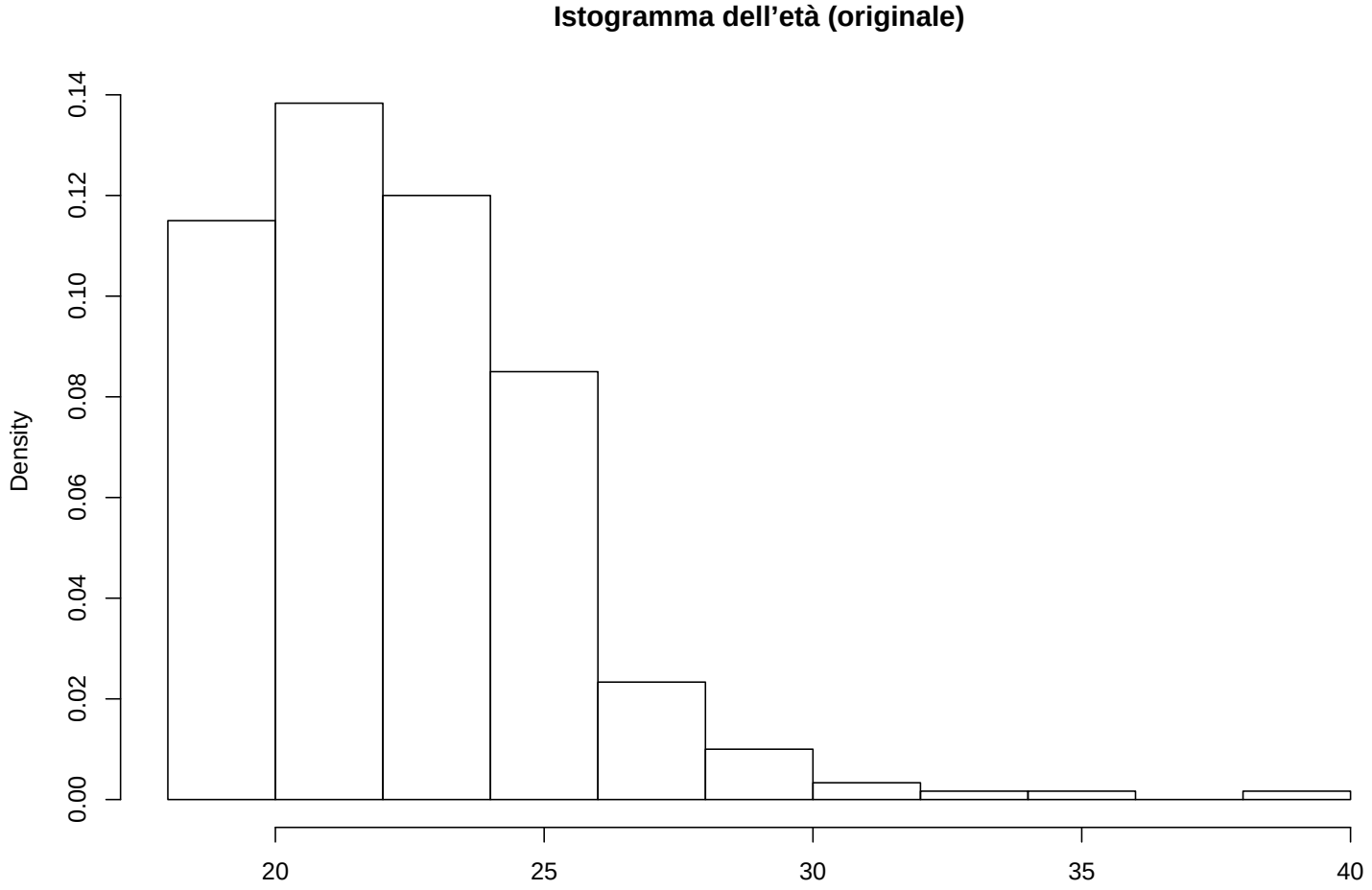
SCI = BIO+GEO, GIUR = GIUR + OPGIU+SPOLI

Corso di laurea ($H = 0.329$)



etc = BIO+GEO + SCI + ECON + INGE, GIUR = GIUR + OPGIU+SPOLI

Età ($H = 0.062$)


Istogramma dell'età (originale)
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
20
25
30
35
40

Età ($H = 0.04$)

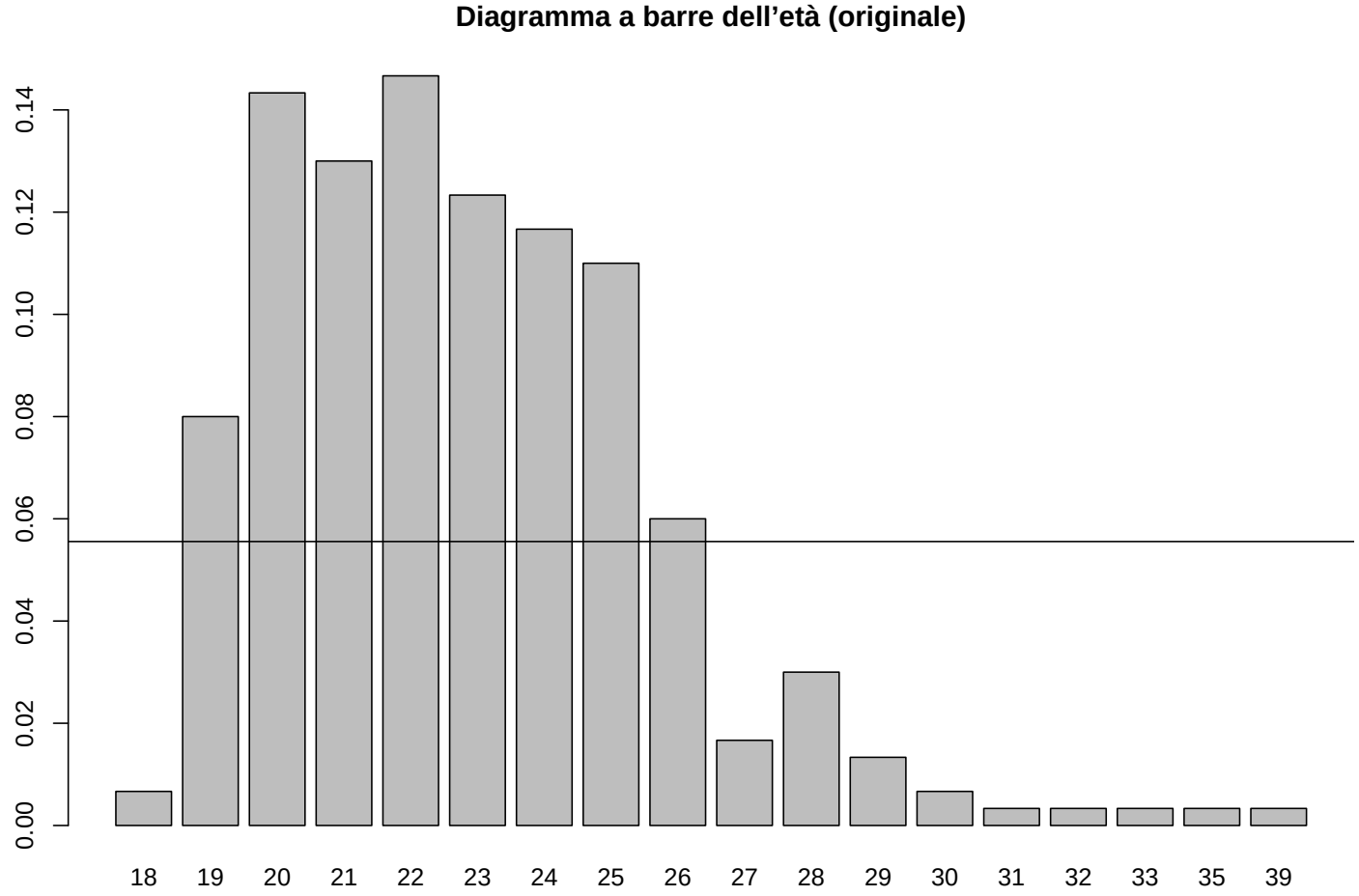
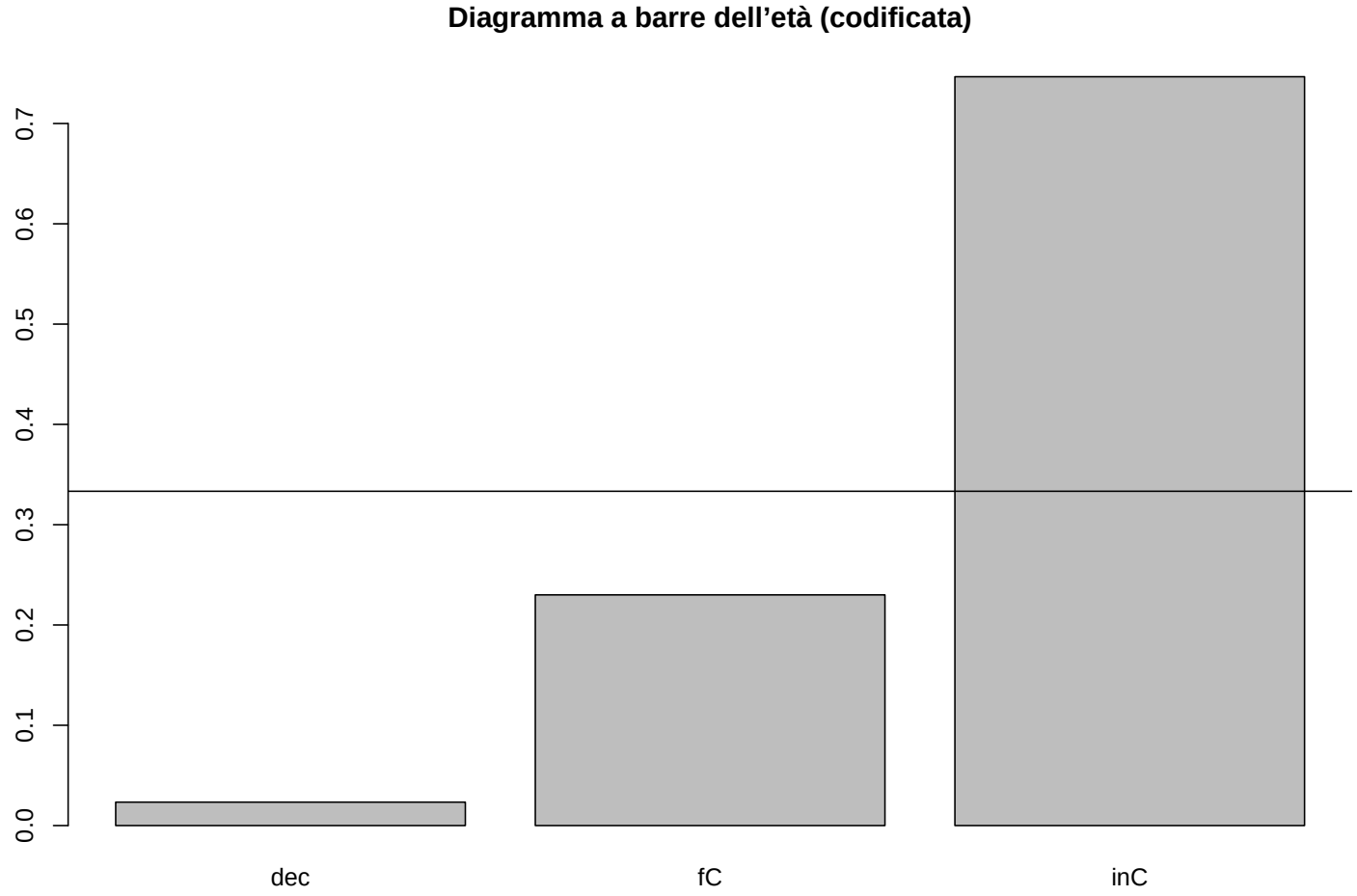


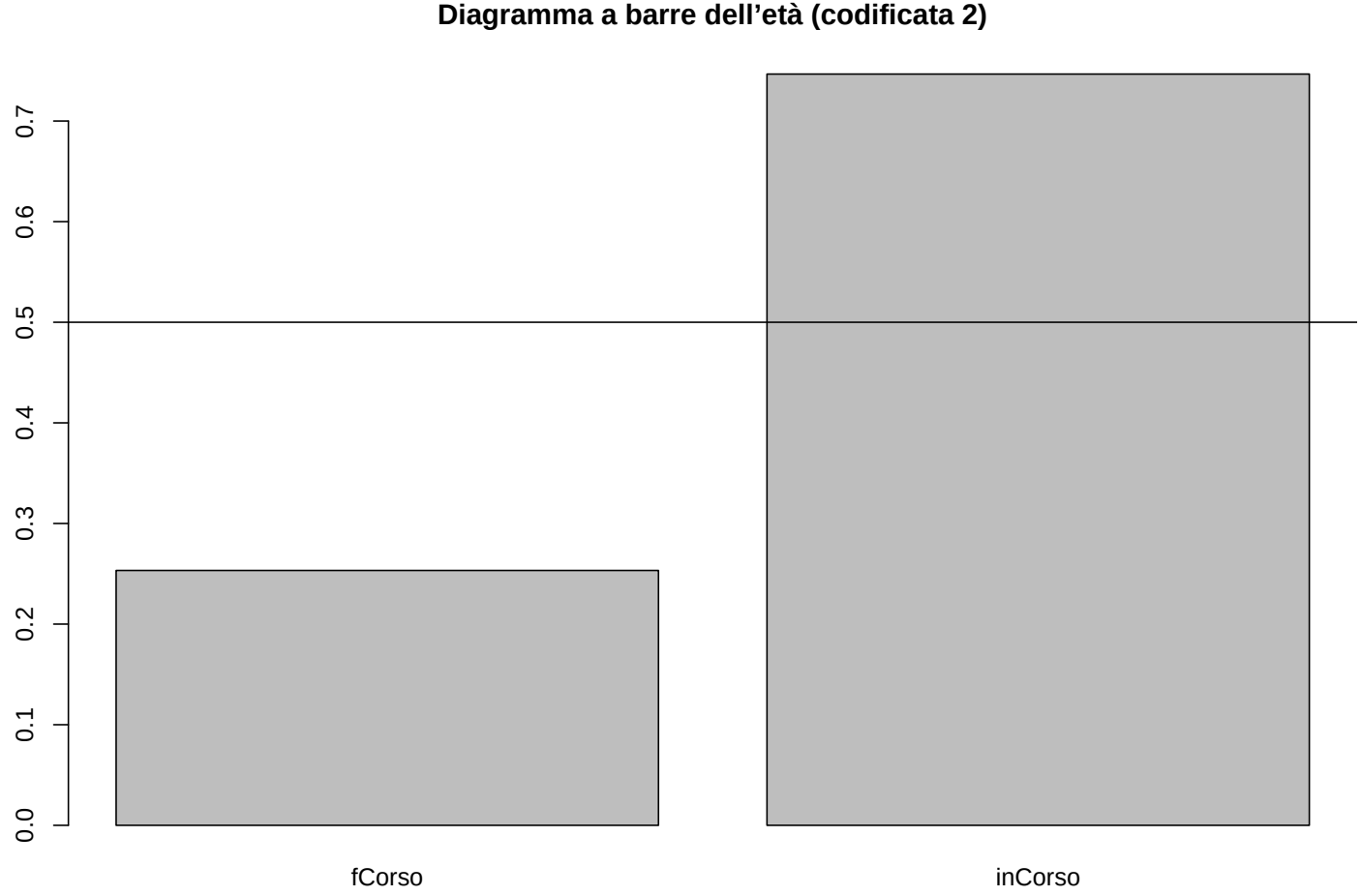
Diagramma a barre dell'età (originale)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
35
39

Età ($H = 0.195$)



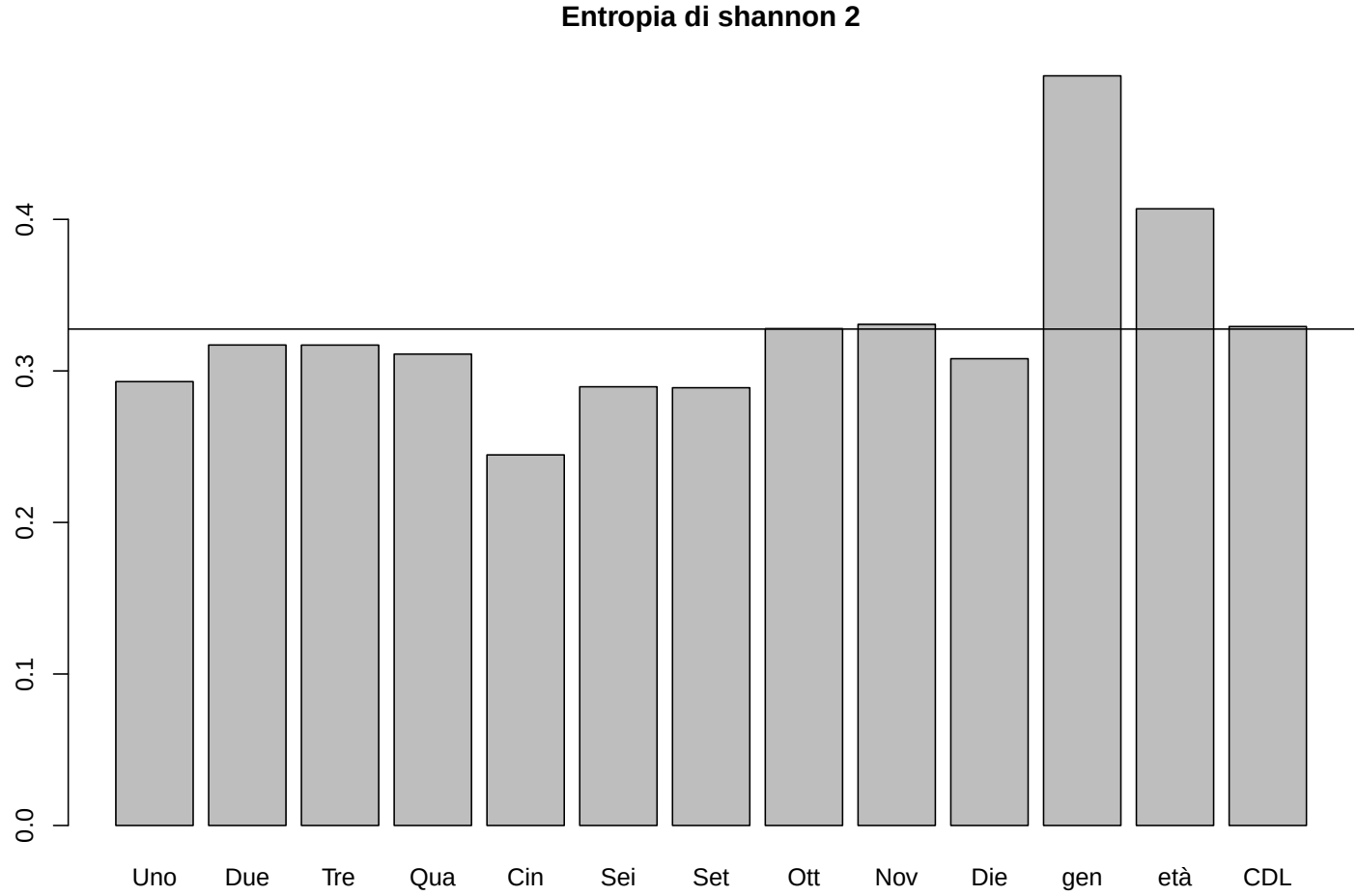
in corso (inC) = "$età \leq 25$", fuori corso (fC) = "$25 < età < 30$",
decani (dec) = "$età \geq 30$"

Età ($H = 0.407$)



fCorso = fC+dec

Entropia su tutte le variabili ($\bar{H} = 0.327$)



CDL ed età ricodificate

La misura della dipendenza nelle distr. doppie

L'indice χ^2 di Pearson misura la dipendenza nelle distribuzioni doppie di frequenza

$$\hat{\chi}^2 = \sum_{ij} \frac{(n_{ij} - n_{i\cdot} n_{\cdot j}/n)^2}{n_{i\cdot} n_{\cdot j}/n} = n \sum_{ij} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}}.$$

Esso può essere utilizzato anche per verificare (con α fissato) l'ipotesi nulla che fra due variabili qualitative X_1 e X_2 vi sia indipendenza quando i dati sono di origine campionaria:

$$\mathcal{H}_0 : X_1 \text{ e } X_2 \text{ sono indipendenti}$$

Se indichiamo con

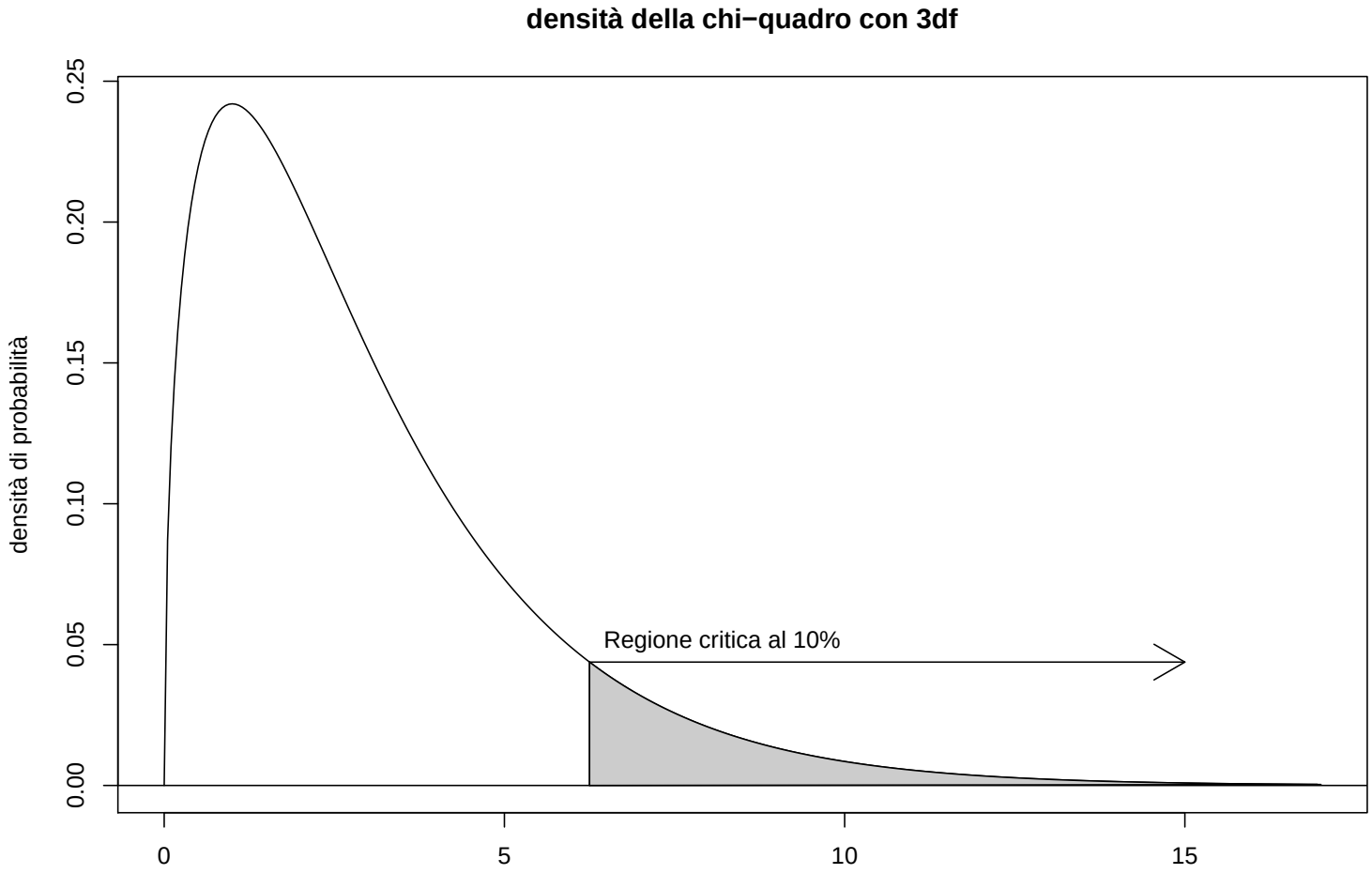
$$pv = \mathsf{P}\left[\chi^2_{(q-1)(p-1)} > \hat{\chi}^2\right]$$

il valore P del test, allora

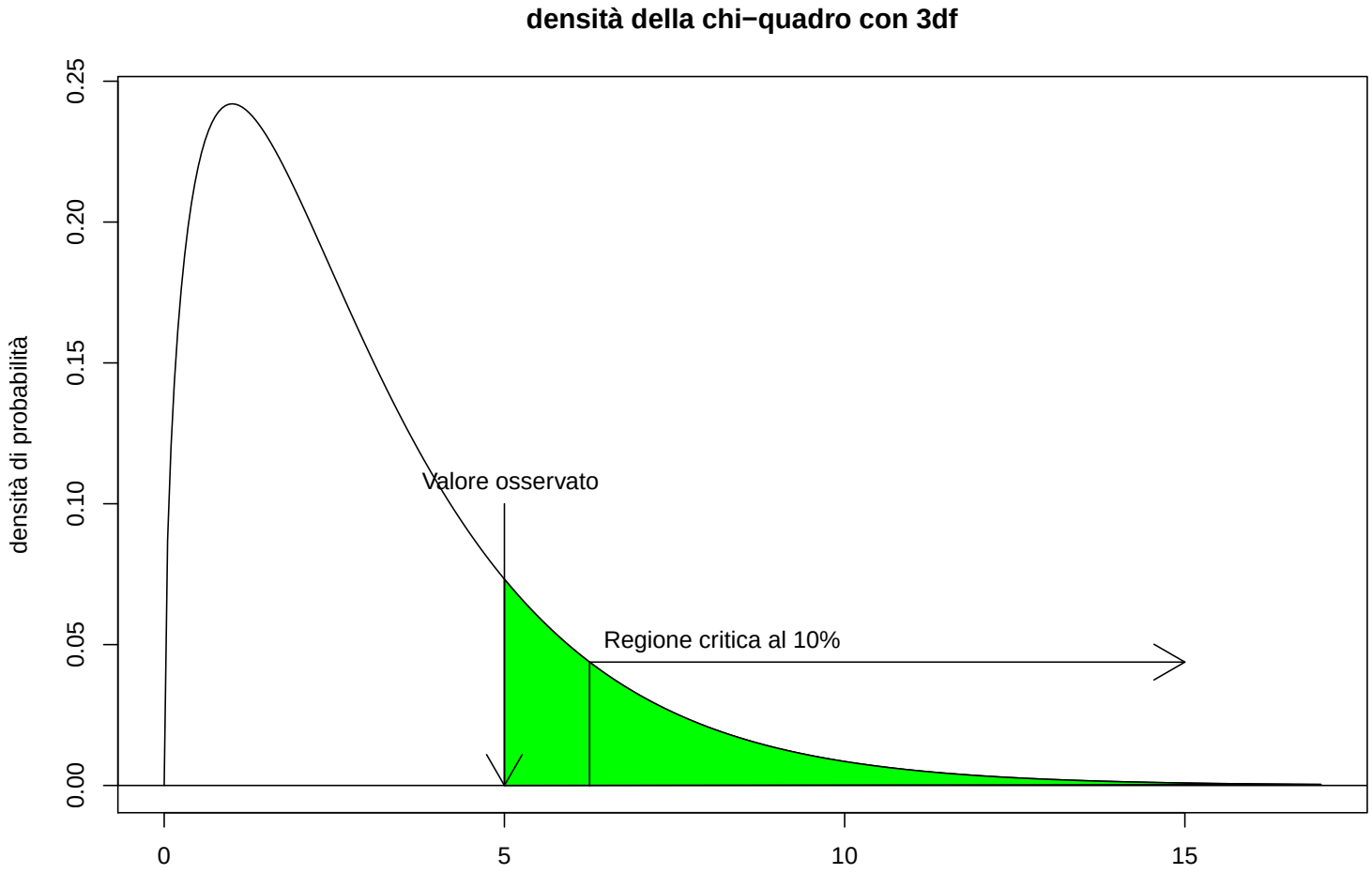
- se $pv > \alpha$ (valore osservato $<$ valore critico) non si rifiuta l'ipotesi nulla, ovvero i dati campionari osservati sono in linea con l'indipendenza delle variabili e tale caratteristica può essere considerata strutturale per la popolazione su cui i dati sono osservati;

- se $pv < \alpha$ (valore osservato $>$ valore critico) si rifiuta l'ipotesi nulla, ovvero i dati campionari osservati sono in linea con la dipendenza delle variabili, e tale ...

La misura della dipendenza nelle distr. doppie

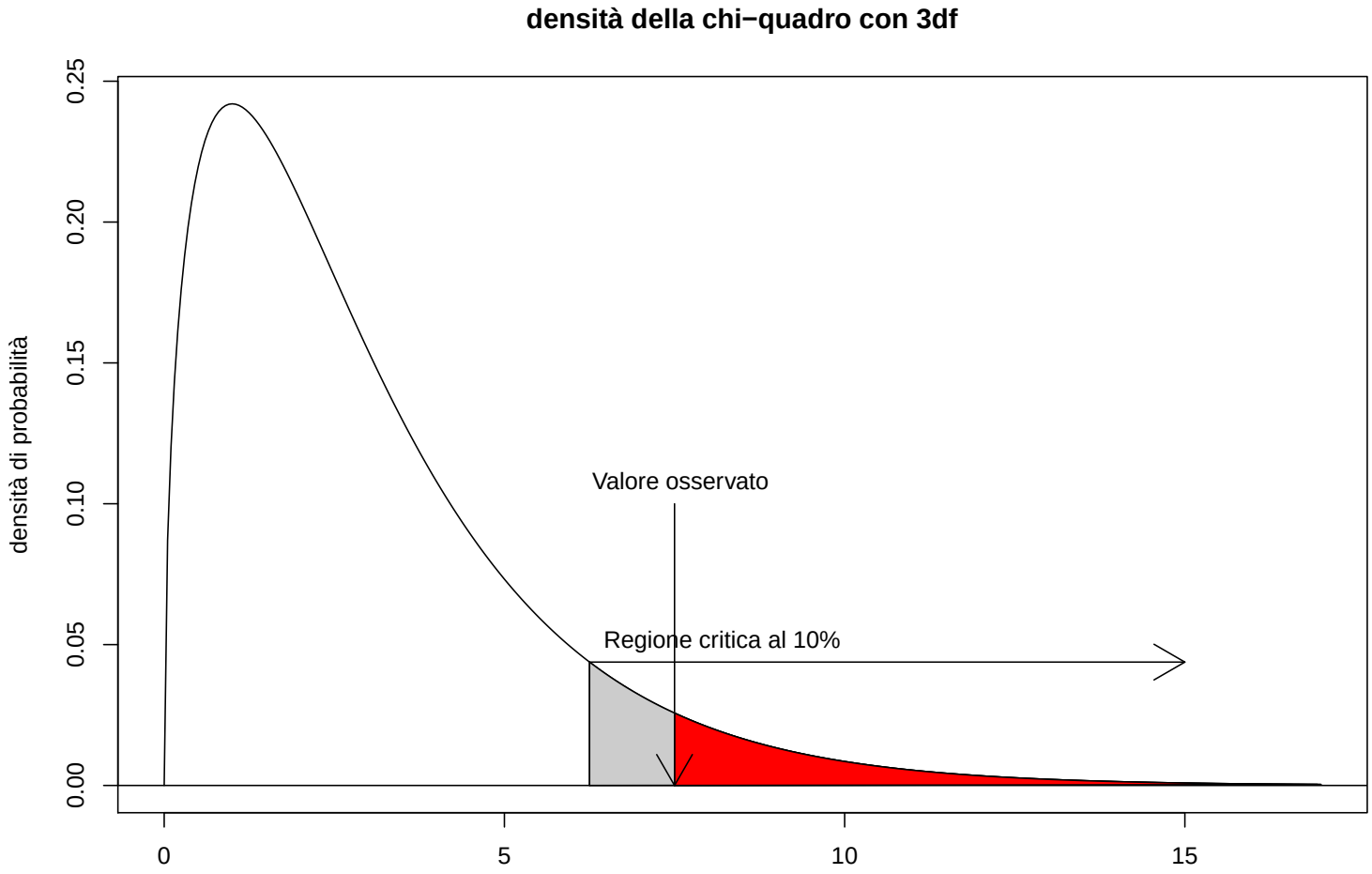


La misura della dipendenza nelle distr. doppie



se $pv > \alpha$ non si rifiuta l'ipotesi nulla

La misura della dipendenza nelle distr. doppie



se $pv < \alpha$ si rifiuta l'ipotesi nulla

tavola dei pValue

	Uno	Due	Tre	Qua	Cin	Sei	Set	Ott	Nov	Die	gen	età
Uno	0.00											
Due	0.02	0.00										
Tre	0.10	0.47	0.00									
Qua	0.02	0.00	0.12	0.00								
Cin	0.83	0.32	0.06	0.05	0.00							
Sei	0.54	0.00	1.00	0.11	0.09	0.00						
Set	0.67	0.00	0.64	0.04	0.86	0.46	0.00					
Ott	0.21	0.21	0.08	0.00	0.32	0.45	0.10	0.00				
Nov	0.16	0.00	0.43	0.00	0.09	0.10	0.06	0.26	0.00			
Die	0.05	0.01	0.29	0.00	0.00	0.01	0.00	0.30	0.12	0.00		
gen	0.19	0.00	0.10	0.00	0.22	0.02	0.08	0.03	0.07	0.00	0.00	
età	0.23	0.04	0.01	0.79	0.30	0.59	0.01	0.03	0.34	0.76	0.34	0.00
CDL	0.03	0.40	0.23	0.00	0.01	0.22	0.02	0.13	0.31	0.25	0.00	0.00

tavola dei pValue

	Uno	Due	Tre	Qua	Cin	Sei	Set	Ott	Nov	Die	gen	età
Uno	0.00											
Due	0.02	0.00										
Tre	0.10	0.47	0.00									
Qua	0.02	0.00	0.12	0.00								
Cin	0.83	0.32	0.06	0.05	0.00							
Sei	0.54	0.00	1.00	0.11	0.09	0.00						
Set	0.67	0.00	0.64	0.04	0.86	0.46	0.00					
Ott	0.21	0.21	0.08	0.00	0.32	0.45	0.10	0.00				
Nov	0.16	0.00	0.43	0.00	0.09	0.10	0.06	0.26	0.00			
Die	0.05	0.01	0.29	0.00	0.00	0.01	0.00	0.30	0.12	0.00		
gen	0.19	0.00	0.10	0.00	0.22	0.02	0.08	0.03	0.07	0.00	0.00	
età	0.23	0.04	0.01	0.79	0.30	0.59	0.01	0.03	0.34	0.76	0.34	0.00
CDL	0.03	0.40	0.23	0.00	0.01	0.22	0.02	0.13	0.31	0.25	0.00	0.00

In rosso sono evidenziati i valori $P < 0.05$,

. . . dove vi dipendenza strutturale fra le variabili (domande).

Indipendenza e associazione

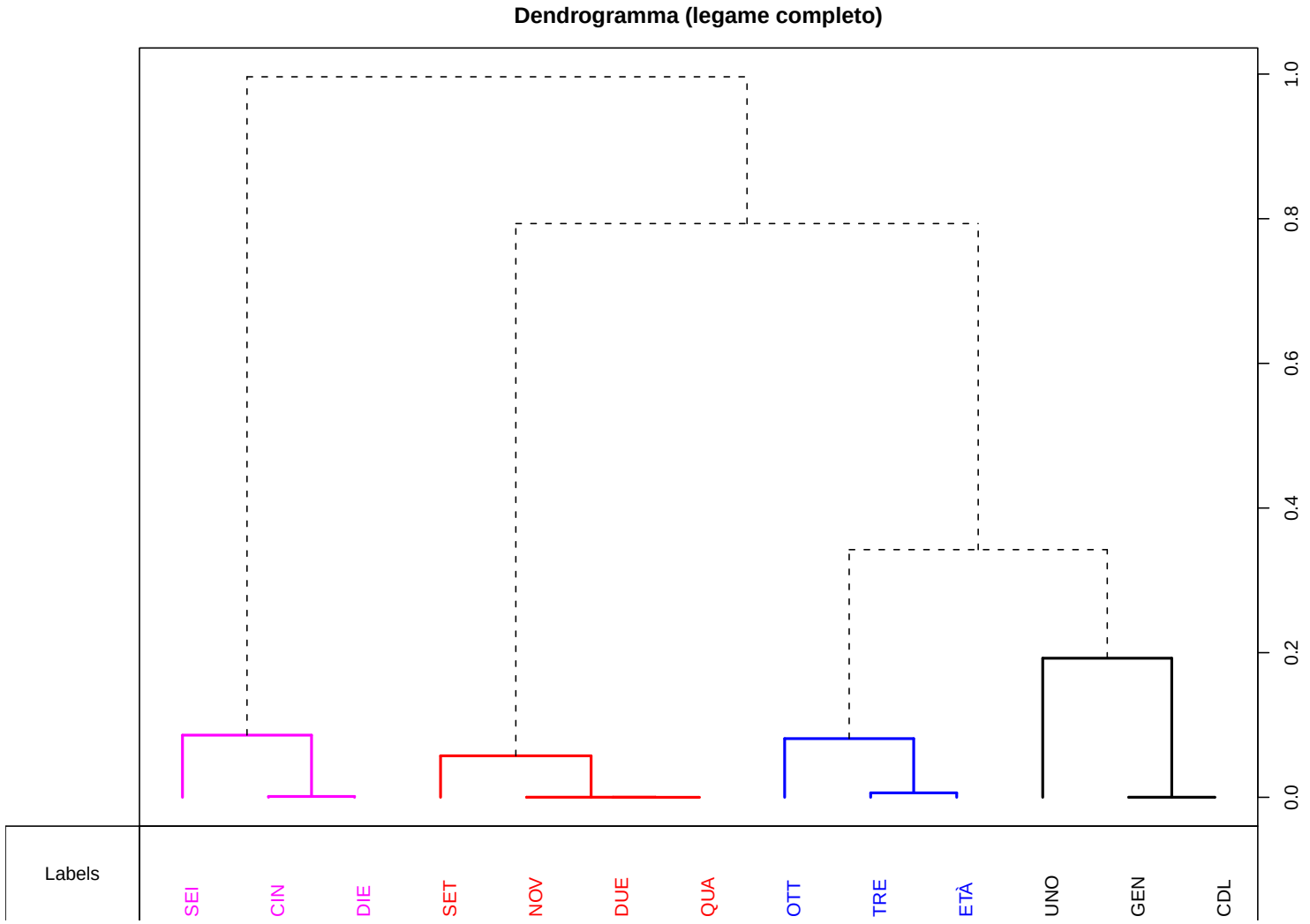
$$pv = \mathsf{P}\left[\chi^2_{(q-1)(p-1)} > \tilde{\chi}^2\right]$$

- se $pv > \alpha$ non si rifiuta l'ipotesi nulla, ovvero le due variabili portano informazioni reciprocamente indipendenti, quindi sono **distanti**
- se $pv < \alpha$ si rifiuta l'ipotesi nulla, ovvero le due variabili portano in una certa misura le stesse informazioni, quindi sono **vicine**

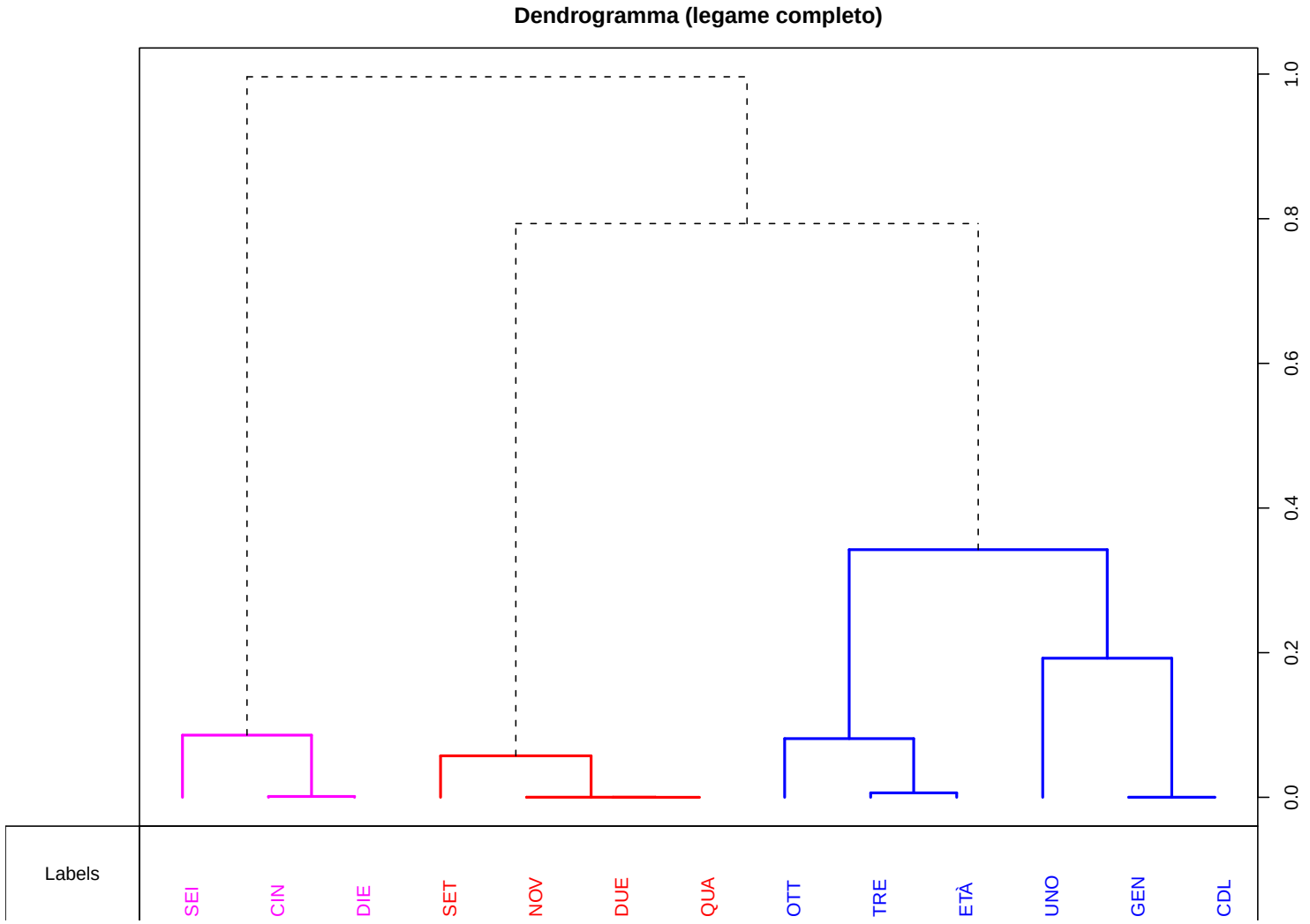
Il valore P può essere letto come una misura di dissimilarità fra le variabili!

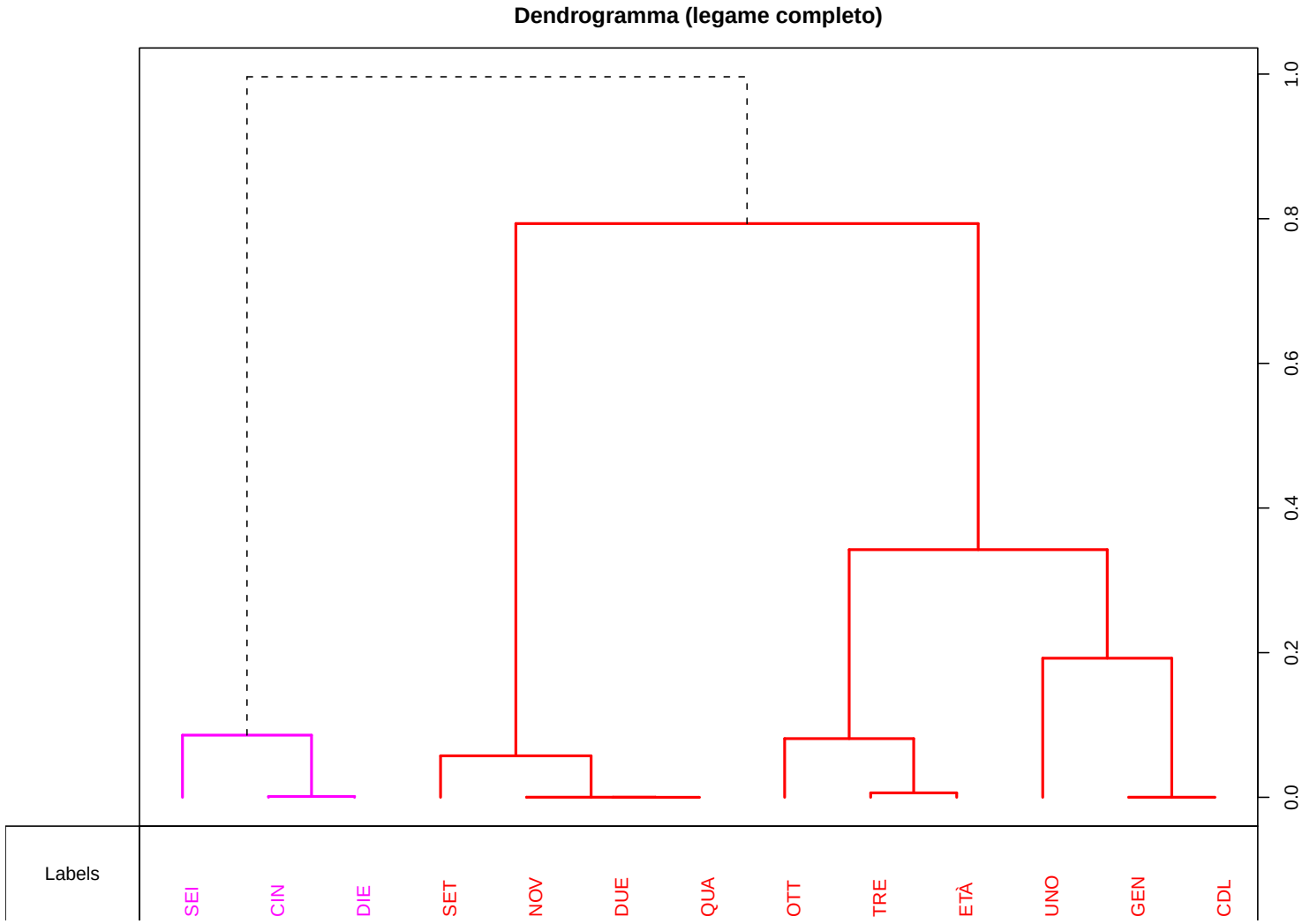
La ricerca dei gruppi di variabili mutuamente associate nel questionario può essere condotta studiando la totalità dei valori P calcolati su tutte le coppie delle variabili.

Indipendenza e associazione



Indipendenza e associazione




Dendrogramma (legame completo)
1.0
0.8
0.6
0.4
0.2
0.0
Labels
SEI
CIN
DIE
SET
NOV
DUE
QUA
OTT
TRE
ETÀ
UNO
GEN
CDL

Le variabili 2 e 9

$$\hat{\chi}^2 = n \sum_{ij} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = \sum_{ij} \frac{n\,(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} \approx 24 \gg 4$$

$100 * f_{ij} =$

	9A	9B	9C	tot
2A	8.725	5.034	5.034	18.792
2B	14.765	14.430	13.423	42.617
2C	6.376	11.074	21.141	38.591
tot	29.866	30.537	39.597	100.000

$100 * (f_{ij} - f_{i\cdot} f_{\cdot j}) =$

	9A	9B	9C
2A	3.112	-0.705	-2.408
2B	2.037	1.415	-3.453
2C	-5.150	-0.711	5.860

$$\frac{1}{9} \sum_{ij} \frac{n\,(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = 2.771$$

$\frac{n(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} =$

	9A	9B	9C
2A	**(+)5.144**	0.258	2.321
2B	0.972	0.459	2.105
2C	**(-)6.856**	0.128	**(+)6.697**

Le variabili 4 e 6

$$\tilde{\chi}^2 = n \sum_{ij} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = \sum_{ij} \frac{n\,(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} \approx 7.6 > 4; pValue = 0.105$$

$100 * f_{ij} =$

6A	6B	6C	tot	
4A	10.403	23.826	3.356	37.584
4B	11.074	26.510	8.054	45.638
4C	5.369	7.718	3.691	16.779
tot	26.846	58.054	15.101	100.000

$100 * (f_{ij} - f_{i\cdot} f_{\cdot j}) =$

	6A	6B	6C
4A	0.313	2.007	-2.320
4B	-1.178	0.016	1.162
4C	0.865	-2.022	1.158

$$\frac{1}{9} \sum_{ij} \frac{n\,(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = 0.85$$

$\frac{n(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} =$

	6A	6B	6C
4A	0.029	0.550	*2.825*
4B	0.337	0.000	0.584
4C	0.495	*1.251*	*1.576*

... ancora sui profili!

	9A	9B	9C	tot
2A	8.725	5.034	5.034	18.792
2B	14.765	14.430	13.423	42.617
2C	6.376	11.074	21.141	38.591
tot	29.866	30.537	39.597	100.000

$\cdots \rightarrow$

$\rightarrow \cdots$

	9A	9B	9C
2A	8.725/18.792	5.034/18.792	5.034/18.792
2B	14.765/42.617	14.430/42.617	13.423/42.617
2C	6.376/38.591	11.074/38.591	21.141/38.591

$= \cdots$

$\cdots =$

	9A	9B	9C	tot
2A	0.464	0.268	0.268	1.0
2B	0.346	0.339	0.315	1.0
2C	0.165	0.287	0.548	1.0

... ancora sui profili!

	9A	9B	9C	tot
2A	8.725	5.034	5.034	18.792
2B	14.765	14.430	13.423	42.617
2C	6.376	11.074	21.141	38.591
tot	29.866	30.537	39.597	100.000

$\cdots \rightarrow$

$\rightarrow \cdots$

	9A	9B	9C
2A	8.725/18.792	5.034/18.792	5.034/18.792
2B	14.765/42.617	14.430/42.617	13.423/42.617
2C	6.376/38.591	11.074/38.591	21.141/38.591

$= \cdots$

$\cdots =$

	9A	9B	9C	tot
2A	0.464	0.268	0.268	1.0
2B	0.346	0.339	0.315	1.0
2C	0.165	0.287	0.548	1.0
2A	1.0	0.0	0.0	1.0
2B	0.0	1.0	0.0	1.0
2C	0.0	0.0	1.0	1.0
centro	0.299	0.305	0.396	1.0

... ancora sui profili!

	9A	9B	9C	tot
2A	8.725	5.034	5.034	18.792
2B	14.765	14.430	13.423	42.617
2C	6.376	11.074	21.141	38.591
tot	29.866	30.537	39.597	100.000

$\cdots \rightarrow$

$\rightarrow \cdots$

	9A	9B	9C
2A	8.725/18.792	5.034/18.792	5.034/18.792
2B	14.765/42.617	14.430/42.617	13.423/42.617
2C	6.376/38.591	11.074/38.591	21.141/38.591

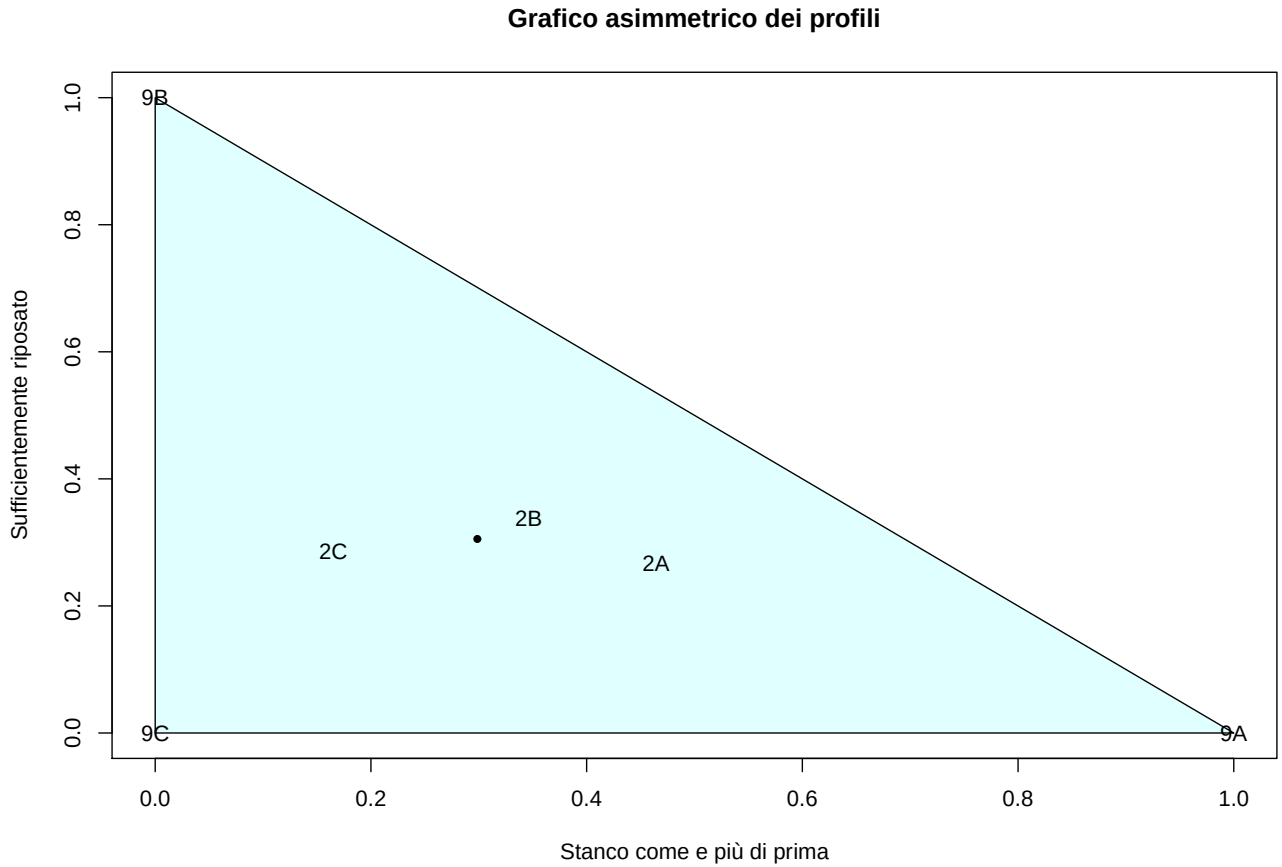
$= \cdots$

$\cdots =$

	9A	9B	9C	tot
2A	0.464	0.268	0.268	1.0
2B	0.346	0.339	0.315	1.0
2C	0.165	0.287	0.548	1.0
2A	1.0	0.0	0.0	1.0
2B	0.0	1.0	0.0	1.0
2C	0.0	0.0	1.0	1.0
centro	0.299	0.305	0.396	1.0

... ancora sui profili!

9. Capita spesso che al mattino tu ti senta ...



C. Pronto ad affrontare una nuova giornata

... ancora sui profili!

2 Una dura giornata di lavoro sta per cominciare. Tu ti senti ...

- A Teso, come sempre d'altronde
- B Un po' intimidito ma pronto ad iniziare
- C Sicuro di te

9 Capita spesso che al mattino tu ti senta ...

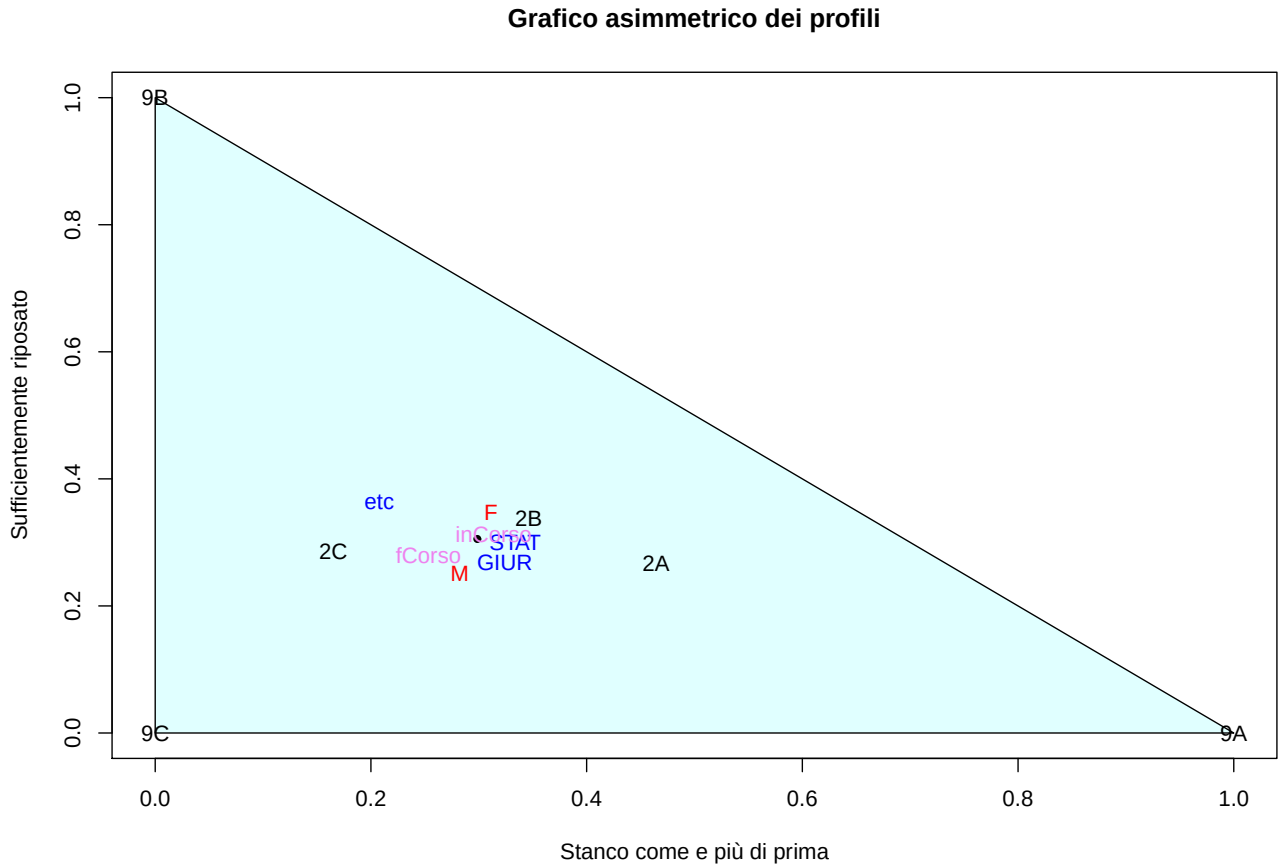
- A Sufficientemente riposato
- B Stanco come e più di prima
- C Pronto ad affrontare una nuova giornata

... ancora sui profili!

	9A	9B	9C
2A	0.464	0.268	*0.268*
2B	0.346	0.339	*0.315*
2C	0.165	0.287	*0.548*
9A	1.000	0.000	*0.000*
9B	0.000	1.000	*0.000*
9C	0.000	0.000	*1.000*
centro	0.299	0.305	*0.396*
F	0.311	0.347	*0.341*
M	0.282	0.252	*0.466*
etc	0.208	0.364	*0.429*
GIUR	0.324	0.269	*0.407*
STAT	0.336	0.301	*0.363*
fCorso	0.253	0.280	*0.467*
inCorso	0.314	0.314	*0.372*

... ancora sui profili!

9. Capita spesso che al mattino tu ti senta ...



C. Pronto ad affrontare una nuova giornata

... ancora sui profili!

1 Stai passeggiando nel parco quando a distanza ti sembra di vedere un'ombra che si muove.

   A Qualcuno ti sta spiando

   B Forse c'è un cane nascosto fra i cespugli

   C Sarà stata la tua immaginazione

2 Una dura giornata di lavoro sta per cominciare. Tu ti senti ...

   A Teso, come sempre d'altronde

   B Un po' intimidito ma pronto ad iniziare

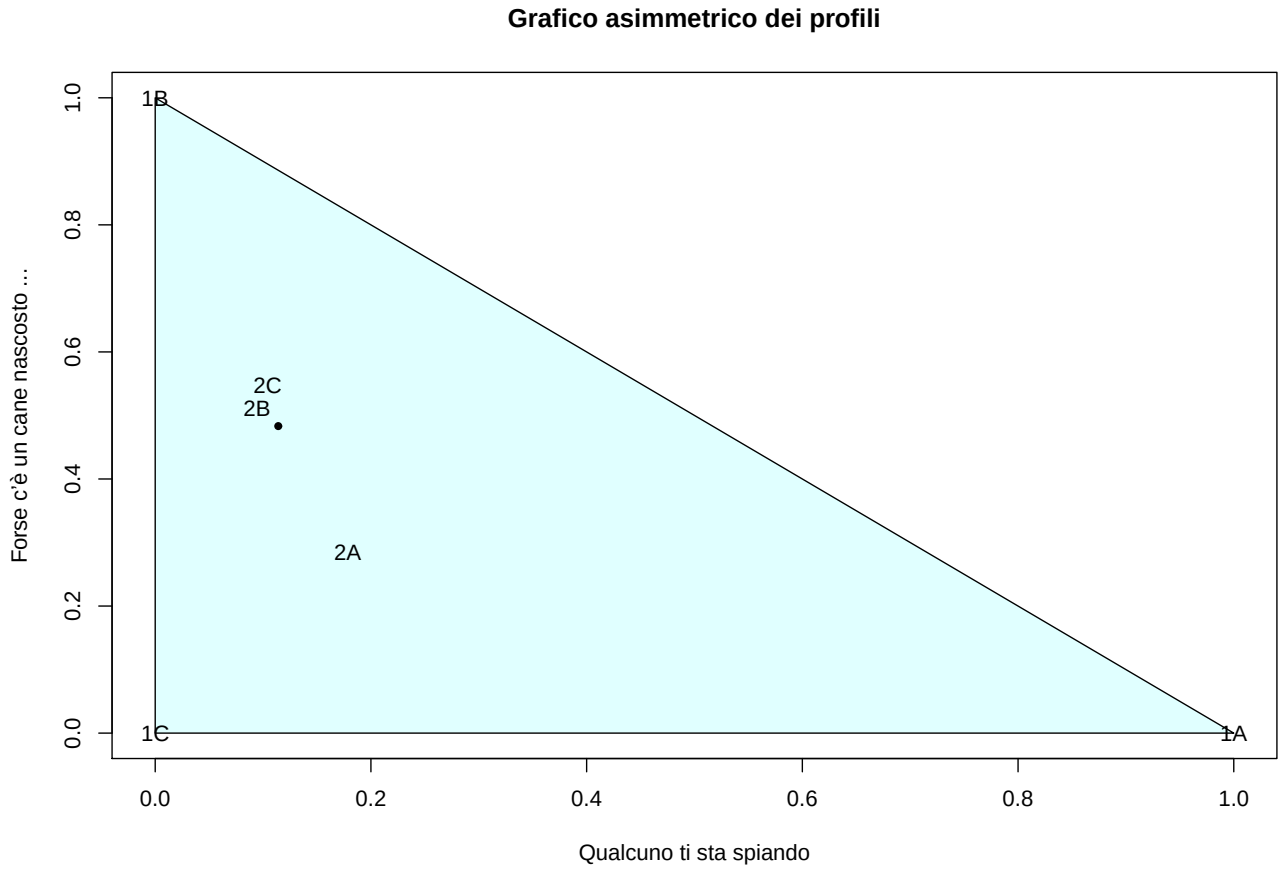
   C Sicuro di te

... ancora sui profili!

	1A	1B	1C
2A	0.179	0.286	*0.536*
2B	0.094	0.512	*0.394*
2C	0.104	0.548	*0.348*
1A	1.000	0.000	*0.000*
1B	0.000	1.000	*0.000*
1C	0.000	0.000	*1.000*
centro	0.114	0.483	*0.403*
F	0.144	0.467	*0.389*
M	0.076	0.504	*0.420*
etc	0.156	0.584	*0.260*
GIUR	0.120	0.407	*0.472*
STAT	0.080	0.487	*0.434*
fCorso	0.120	0.560	*0.320*
inCorso	0.112	0.457	*0.430*

... ancora sui profili!

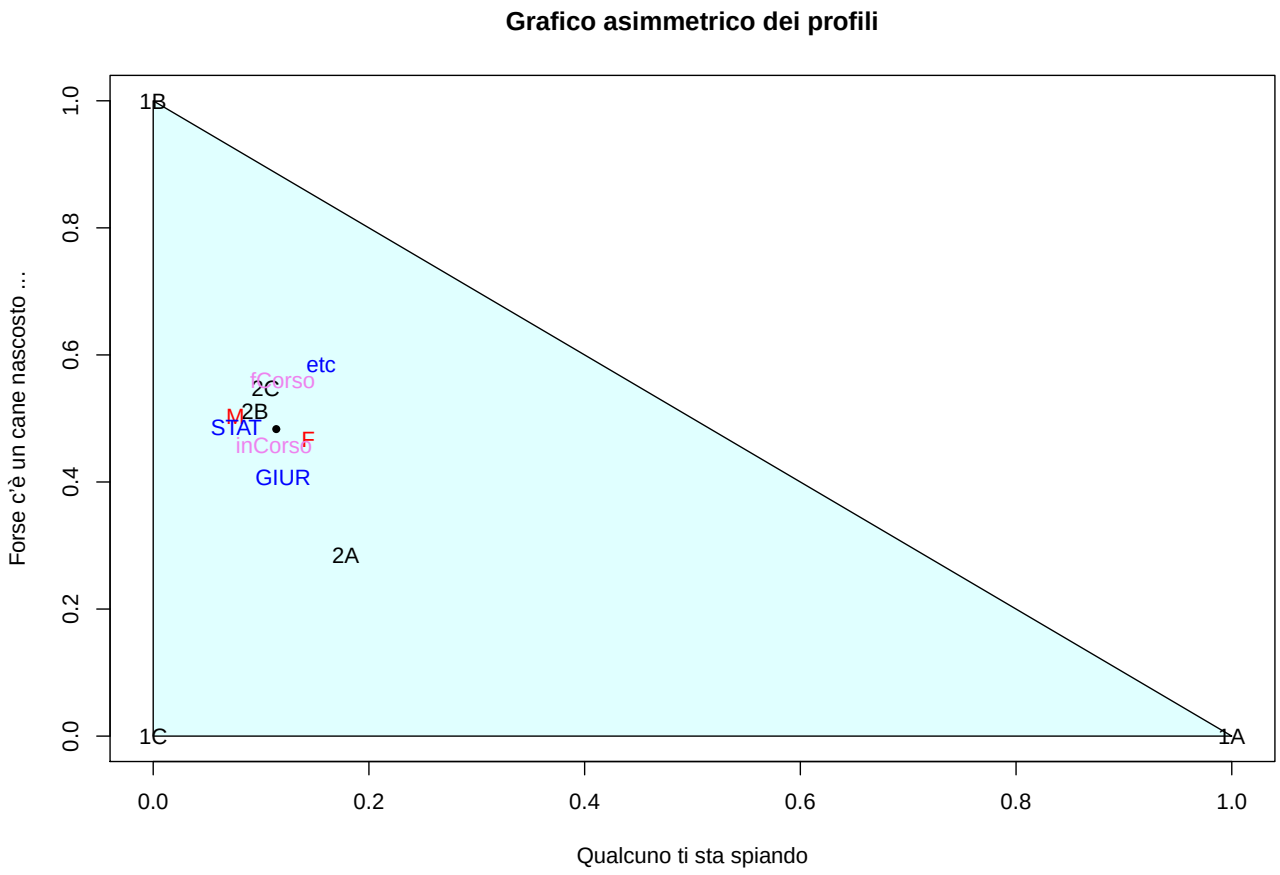
1. Stai passeggiando nel parco quando a distanza ti sembra di vedere un'ombra che si muove.



C. Sarà stata la tua immaginazione

... ancora sui profili!

1. Stai passeggiando nel parco quando a distanza ti sembra di vedere un'ombra che si muove.



C. Sarà stata la tua immaginazione

L'inerzia

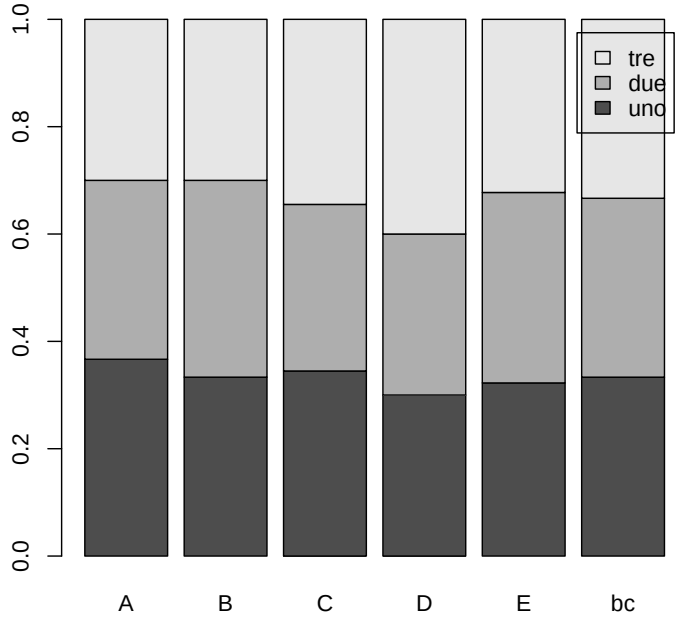
L'inerzia è il rapporto fra l'indice del $\hat{\chi}^2$ e la numerosità n dei dati.

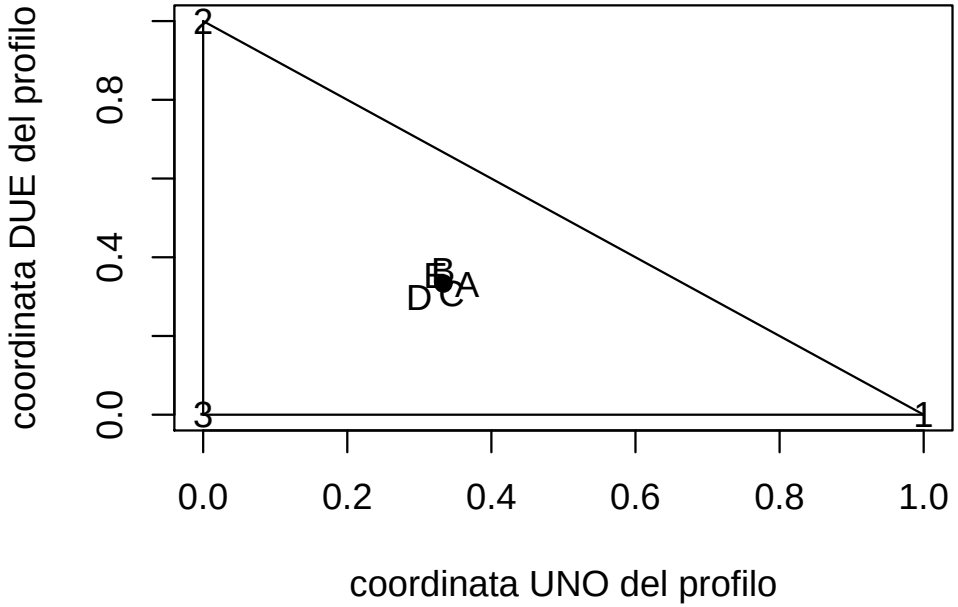
$$\Phi^2 = \hat{\chi}^2/n = \sum_{ij} \frac{\left(f_{ij} - f_{i\cdot} f_{\cdot j}\right)^2}{f_{i\cdot} f_{\cdot j}}.$$

... è anche noto come indice di contingenza quadratica media.

Inerzia $= 0.0076$ ($\chi^2 = 1.14 < 2 \cdot 4$)

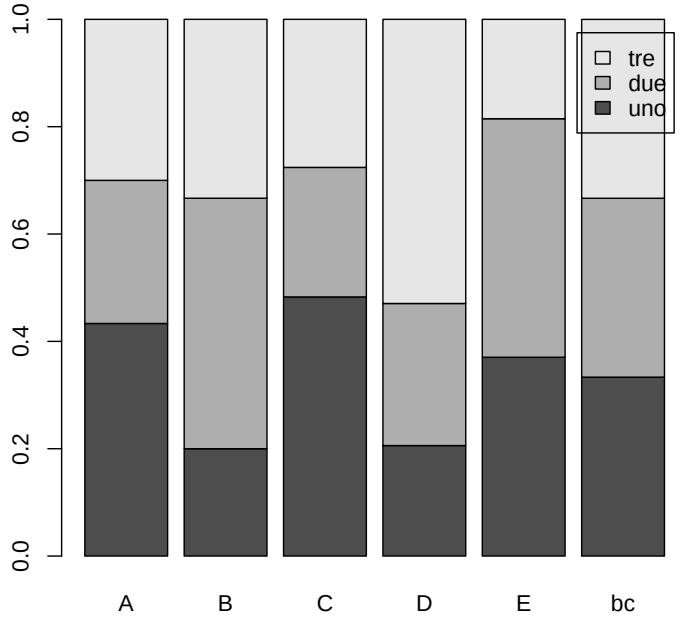
	Uno	Due	Tre	Uno	Due	Tre
A	11	10	9	*0.37*	*0.33*	0.30
B	10	11	9	*0.33*	*0.37*	0.30
C	10	9	10	*0.34*	*0.31*	0.34
D	9	9	12	*0.30*	*0.30*	0.40
E	10	11	10	*0.32*	*0.35*	0.32
	50	50	50	*0.33*	*0.33*	0.33

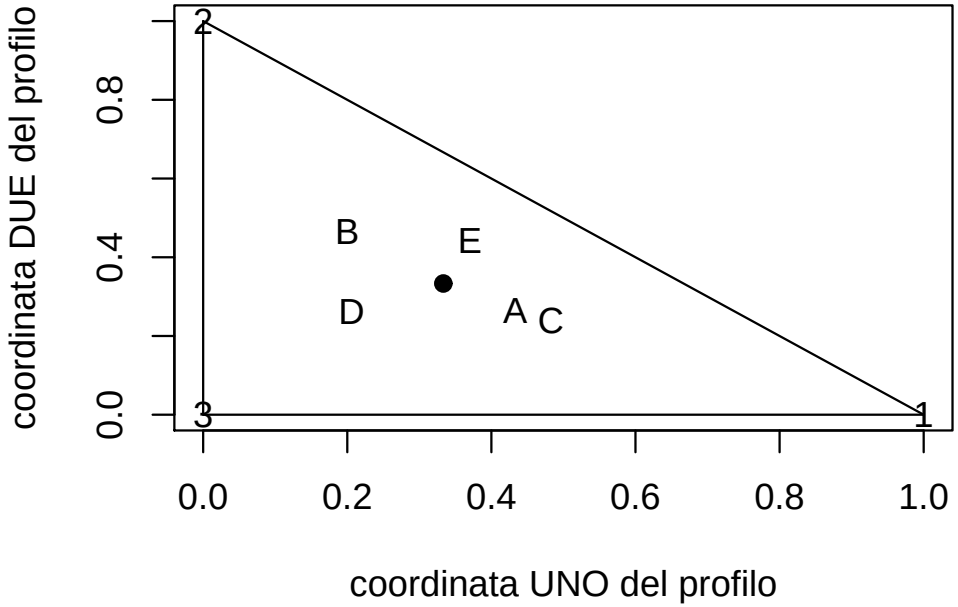




Inerzia $= 0.1101$ $(\chi^2 = 16.52 > 8)$

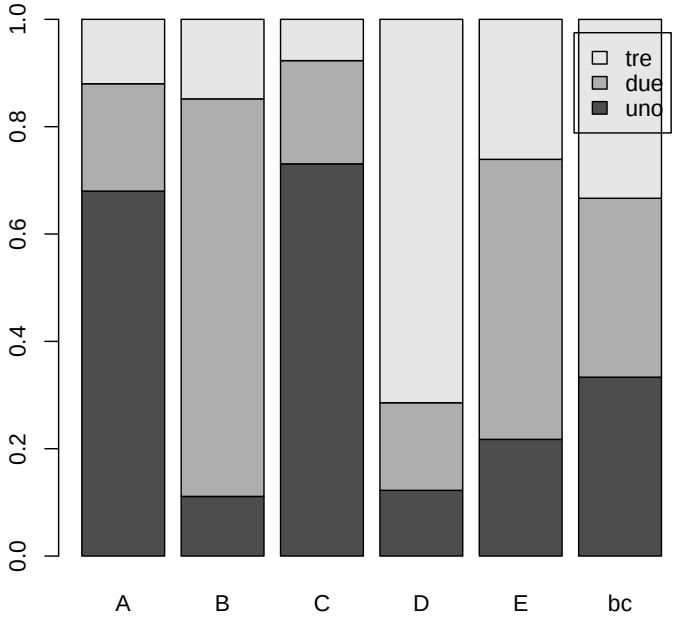
	Uno	Due	Tre	Uno	Due	Tre
A	13	8	9	*0.43*	*0.27*	0.30
B	6	14	10	*0.20*	*0.47*	0.33
C	14	7	8	*0.48*	*0.24*	0.28
D	7	9	18	*0.20*	*0.26*	0.52
E	10	12	5	*0.37*	*0.44*	0.19
	50	50	50	*0.33*	*0.33*	0.33

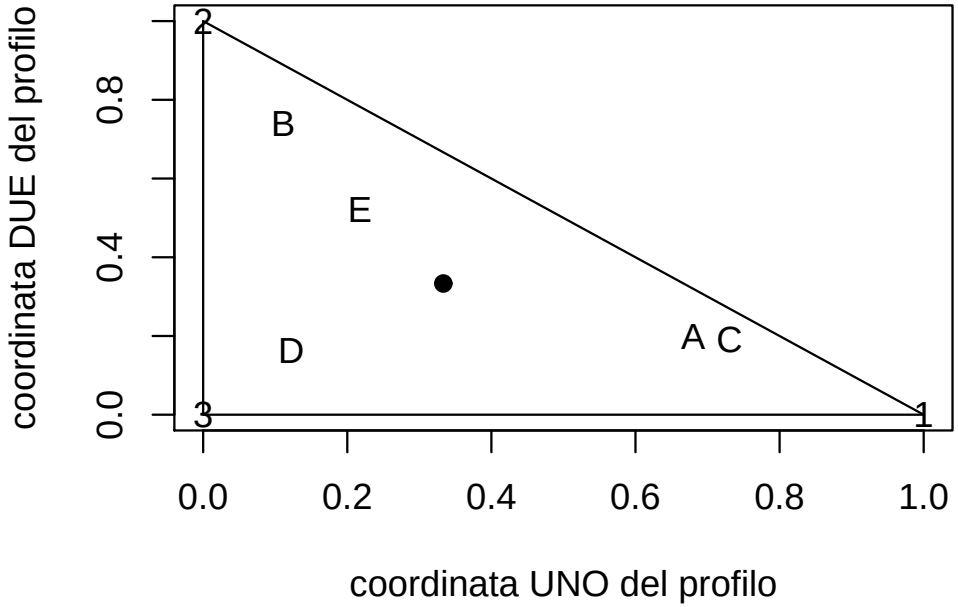




Inerzia $= 0.5923$ ($\chi^2 = 88.85 \gg 8$)

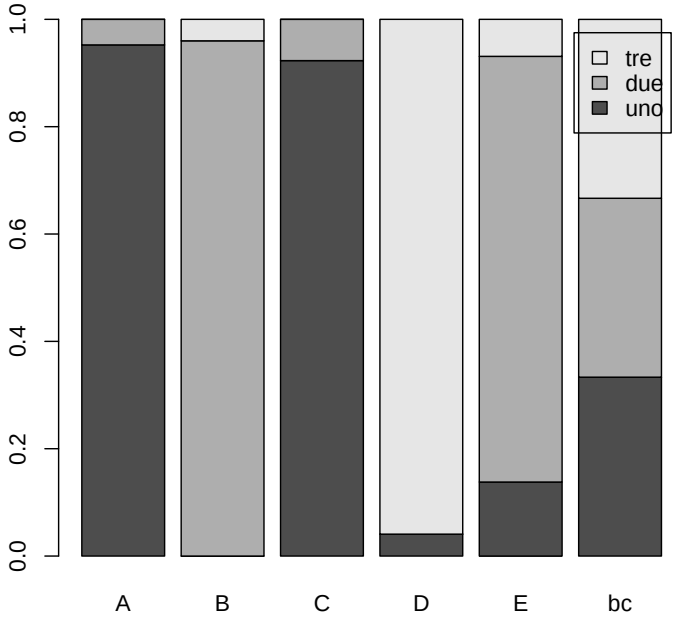
	Uno	Due	Tre	Uno	Due	Tre
A	17	5	3	*0.68*	*0.20*	0.12
B	3	20	4	*0.11*	*0.74*	0.15
C	19	5	2	*0.73*	*0.19*	0.08
D	6	8	35	*0.12*	*0.16*	0.71
E	5	12	6	*0.22*	*0.52*	0.26
	50	50	50	*0.33*	*0.33*	0.33

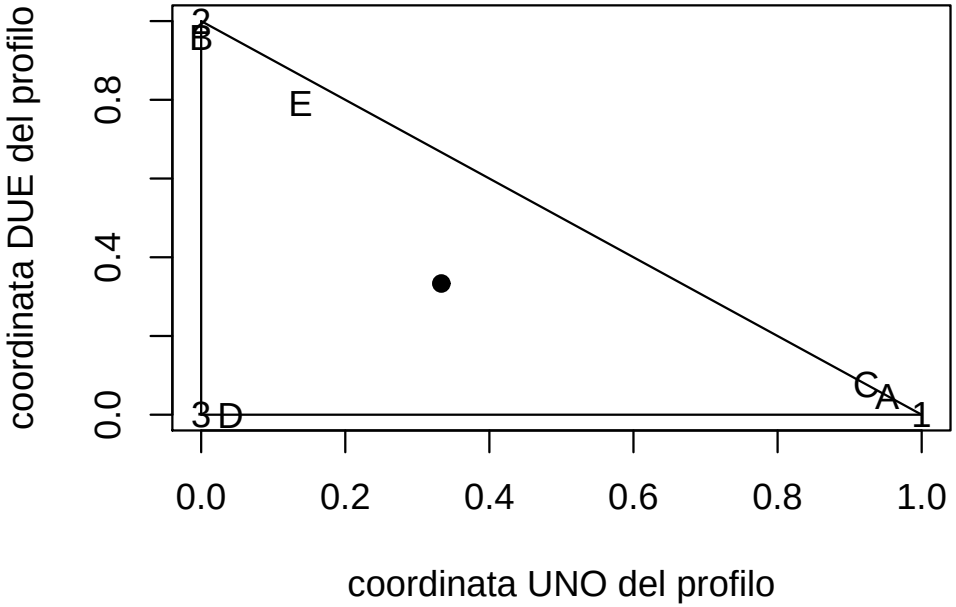




Inerzia $= 1.5715$ ($\chi^2 = 253.73 \gg 8$)

	Uno	Due	Tre	Uno	Due	Tre
A	20	1	0	*0.95*	*0.05*	0.00
B	0	24	1	*0.00*	*0.96*	0.04
C	24	2	0	*0.92*	*0.08*	0.00
D	2	0	47	*0.04*	*0.00*	0.96
E	4	23	2	*0.14*	*0.79*	0.07
	50	50	50	*0.33*	*0.33*	0.33





Analisi delle corrispondenze semplici

Quando nessuna delle variabili ha un numero di modalità pari a tre (entrambe ne hanno di più) allora per poter vedere i profili in un piano e comprendere dove si nascondono le associazioni, è necessario utilizzare un'analisi più sofisticata che si chiama

analisi delle corrispondenze semplici.

In questa analisi l'inerzia Φ^2 ha un ruolo centrale e viene scomposta in una somma di numeri $\lambda_1, \lambda_2, \ldots, \lambda_{K-1}$ che hanno la proprietà di essere ordinati, ovvero $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{K-1}$.

$$\Phi^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_{K-1} = \hat{\chi}2/n$$

I numeri λ_k si chiamano autovalori e K è il numero più piccolo fra il numero delle modalità di riga e quello di colonna.

Analisi delle corrispondenze semplici

A ciascun autovalore è associato un sistema di coordinate sia delle modalità di riga sia di quelle di colonna.

Se consideriamo i primi due sistemi di coordinate, ovvero quelli associati agli autovalori λ_1 e λ_2, spesso essi sono sufficienti a rappresentare *bene* in un piano cartesiano (piano fattoriale) tutte le modalità di riga e tutte le modalità di colonna.

Se due punti (rappresentativi di modalità/profili) che si riferiscono a variabili differenti sono vicini, allora fra le modalità c'è associazione; quando i punti sono distanti allora c'è discordanza, ovvero assenza di associazione.

Analisi delle corrispondenze semplici

La bontà della rappresentazione delle modalità nel piano fattoriale dipende dalla quota parte dell'inerzia che si concentra negli autovalori utilizzati per costruire il piano fattoriale.

Si consideri l'intensità percentualizzata dell'autovalore

$$\lambda_k \to 100\,\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_{K-1}},$$

se il piano è formato dalle coordinate corrispondenti al primo e secondo autovalore, allora si considera

$$\lambda_k \to 100\,\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \cdots + \lambda_{K-1}};$$

più questo numero è vicino a 100, tanto migliore è la rappresentazione grafica.

I bambini

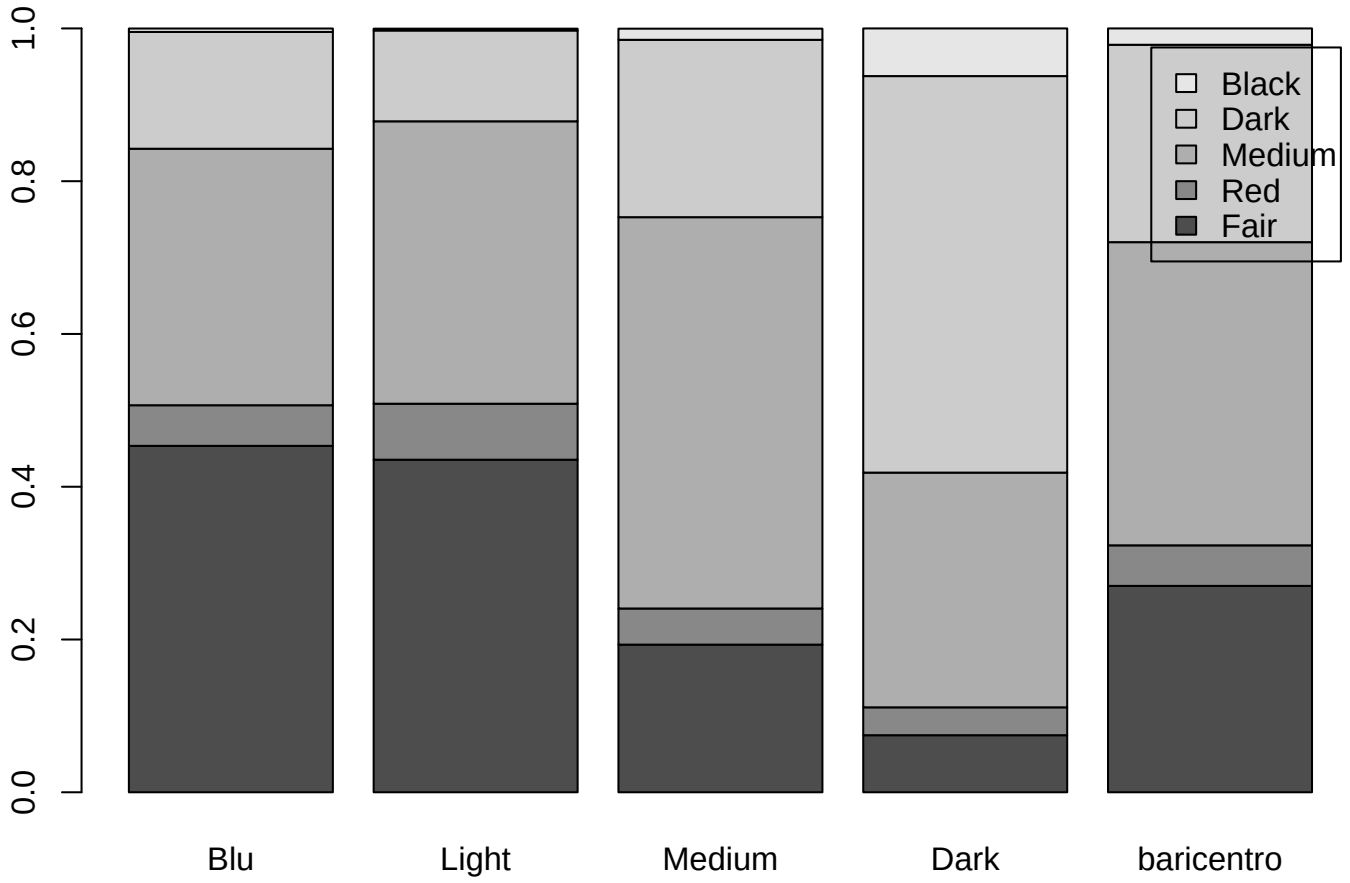
La distribuzione doppia di frequenze* raccoglie su un collettivo di 5384 bambini la frequenza contemporanea di alcuni colori osservati per gli occhi e per i capelli.

	Fair	Red	Medium	Dark	Black	
Blu	326	38	241	110	3	718
Light	688	116	584	188	4	1580
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	82	1312
	1455	286	2137	1391	115	5384

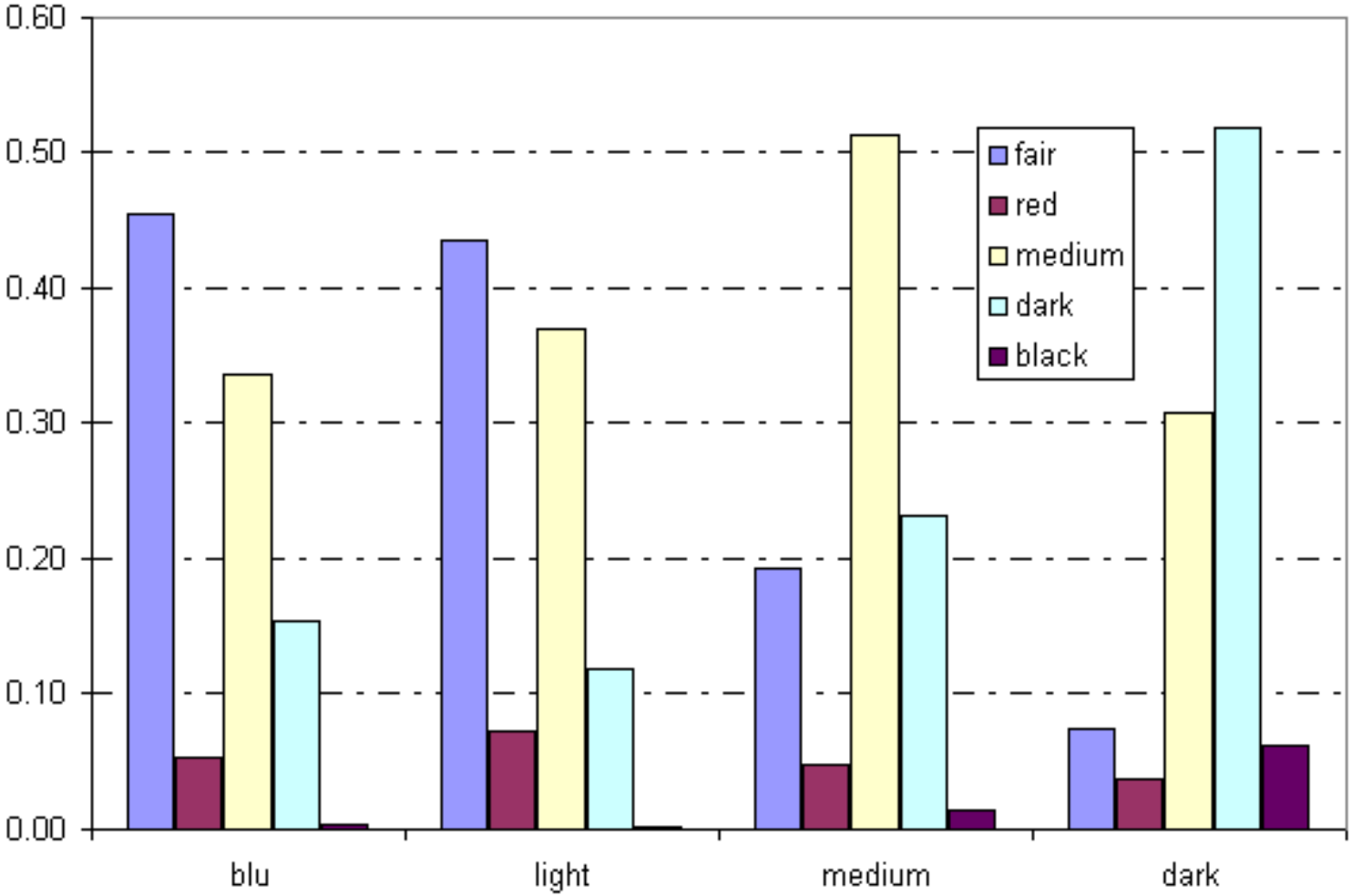
Il problema è di capire se vi sia dipendenza fra il colore degli occhi e dei capelli nel collettivo osservato.

*Small data sets, n.188/p.146

Profili riga rispetto al baricentro



Profili riga rispetto al baricentro

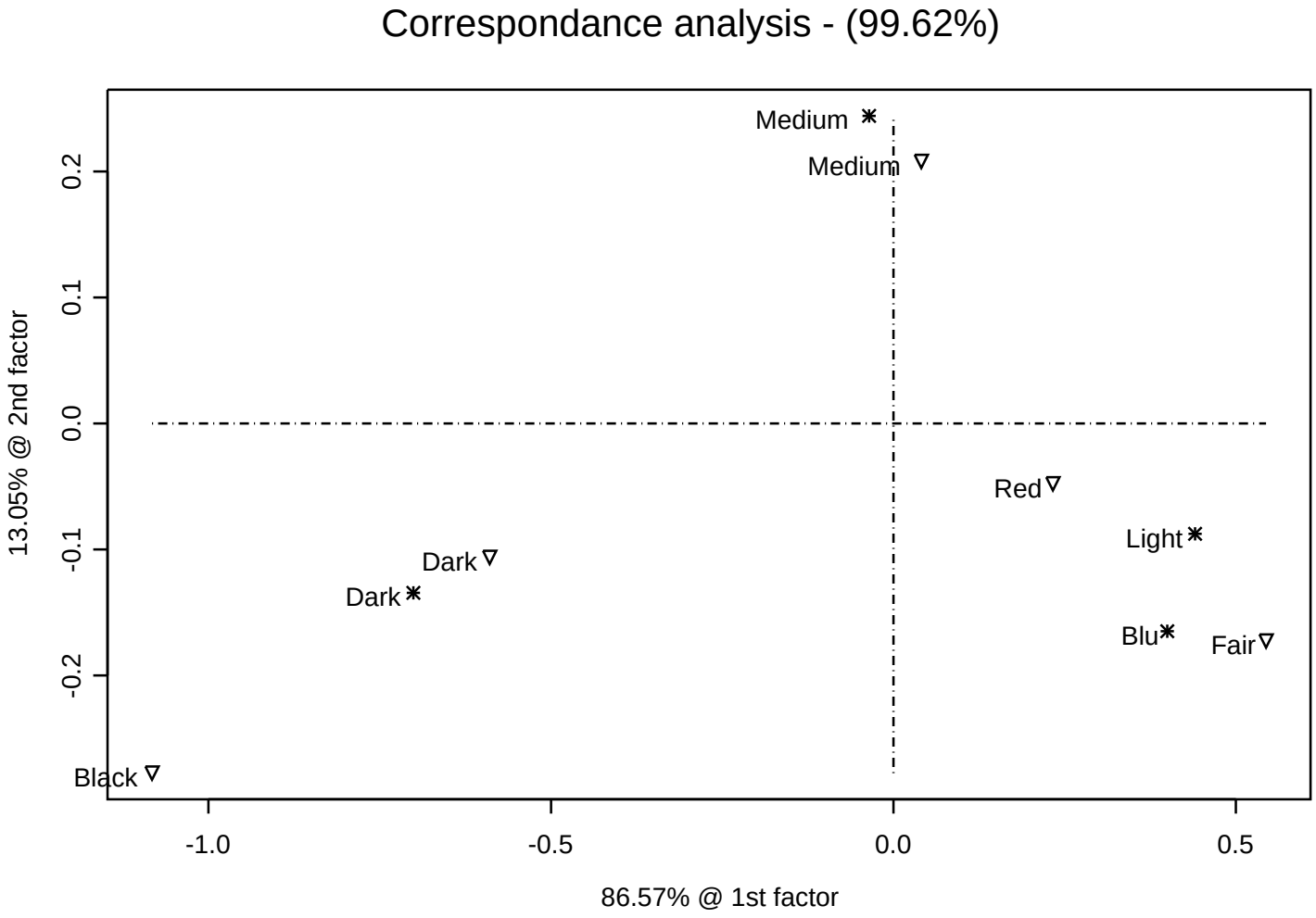


Decomposizione dell'inerzia

Distribuzione fattoriale dell'inerzia rispetto al baricentro

	F1	F2	F3	*(somma)*
Autovalore	0.198	0.030	0.001	*(0.229)*
% inerzia	86.573	13.055	0.373	*(100.0)*
% inerzia cumulata	86.573	99.627	100.000	

Decomposizione dell'inerzia



Raccontare il fenomeno

Quando si raccontano i sistemi di relazioni che si rivelano durante l'analisi dei dati, l'essere comprensibili è l'obiettivo primario.

... Per questo è importante esprimersi con chiarezza, esattezza e semplicità. La semplicità è il risultato visibile della complessità. I due termini non sono l'uno l'opposto dell'altro, ma si completano. Complesso non vuol dire complicato, semplice non vuol dire facile. Diciamo che qualcosa è complesso quando possiede una struttura e un progetto che non è evidente, e che è soltanto possibile intuire o ricostruire. Per contro, la parte visible di qualcosa di complesso ha forma semplice e riconoscibile. ...

[da "Saper Scrivere - Corso di scrittura, a cura di A.Baricco; i corsi di Repubblica-L'espresso.]

Raccontare il fenomeno

La narrazione dell'analisi deve raccontare il modello del fenomeno che emerge attraverso i dati osservati. Per modello si intende il sistema della relazioni che sussistono (o non sussistono) fra i singoli elementi (le variabili) che sono il sostegno del fenomeno.

Nel vostro caso cercate di indagare l'idea di comunità (il fenomeno) e questa si declina nei luoghi in cui essa si manifesta e nelle differenziazioni (etnia, religione, altre diversità) dei gruppi di persone che possono entrare o meno a farne parte.

Raccontare il fenomeno

La costruzione del *romanzo* del fenomeno nasce dall'annotare ogni "scoperta" (idea) che si conquista durante ogni fase del trattamento e analisi dei dati.

Le annotazioni devono essere corredate anche di aspetti tecnici, come ad esempio la misura della forza della dipendenza fra due variabili, oppure se una relazione debba essere considerata tendenziale o strutturale.

La lista disordinata delle idee ha bisogno di riflessione per assumere un percorso coerente ed in linea con il fenomeno indagato.

La lista ordinata è il sommario della narrazione cui è necessario aggiungere le parole per trasferire l'immagine del fenomeno.

L'immagine è un modello della realtà.

Tractatus (Wittgenstein)

1. Il mondo è tutto ciò che accade

   1.1 Il mondo è la totalità dei fatti, non delle cose

   1.13 I fatti nello spazio logico sono il modo

2. Ciò che accade, il fatto, è il sussistere di stati di cose

   2.01 Lo stato di cose è un nesso di oggetti

   2.02 L'oggetto è semplice

   2.03 Nello stato di cose gli oggetti ineriscono l'uno nell'altro come le maglie di una catena

   2.04 La totalità degli stati di cose sussistenti è il mondo

   2.05 La totalità degli stati di cose sussistenti determina anche quali stati di cose non sussistono

   2.06 Il sussistere e non sussistere di stati di cose è la realtà

   2.1 Noi ci facciamo immagini dei fatti

   2.12 L'immagine è un modello della realtà

   2.14 L'immagine consiste nell'essere i suoi elementi in un determinata relazione l'uno con l'altro

Tractatus (Wittgenstein)

Quanto può dirsi, si può dir chiaro;
e su ciò di cui non si può dire, si deve tacere

Giovedì 22 Aprile 2010 Corriere della Sera

I modelli matematici e la realtà come illusione

di GIULIO GIORELLO

Con il blocco dei voli e il faticoso rientro alla normalità c'è un nuovo imputato. Non più la natura che rivela il suo volto indifferente ai desideri umani o la stessa imprudenza di Homo sapiens ma... la matematica. Di fronte all'imprevista esplosione del vulcano islandese, gli esperti hanno fatto ricorso a un qualche modello matematico: nella fattispecie, a quelli che analizzano le dinamiche delle polveri inquinanti nell'atmosfera. Abitualmente noi tutti «pensiamo per modelli», anche se non ne siamo sempre consapevoli. Pensiamo alla cartina della città dove abitiamo, che consultiamo per raggiungere un qualche edificio: un cinema, un teatro, un museo, ecc. Individuato l'obiettivo sulla carta siamo in grado di ricavare anche le istruzioni per raggiungerlo nella realtà. Ma una carta geografica di Milano non è la Milano reale, è solo una sua versione semplificata e «maneggevole». Lo stesso si può dire delle equazioni matematiche che prima traducono in numeri situazioni reali e poi vengono trattate ricorrendo a potenti calcolatori. In tutti questi casi si operano semplificazioni e approssimazioni, un po' come quando i mercanti tengono conto delle merci inviate, ma non degli involucri in cui le spediscono. L'immagine era già di Galileo Galilei; ma oggi sappiamo, persino da qualche storiella di Paperon de' Paperoni, che trascurare come irrilevante qualche dato può portare a una cocente delusione e a un ingente danno economico!

Il punto che i critici impazienti della modellizzazione matematica dimenticano è che questa procedura non è solo scienza, ma anche arte, poiché sceglie fra modi diversi di analizzare il problema, modi che hanno virtù e vizi differenti. Così, nel caso in questione, alcuni modelli permettono di tracciare i flussi delle varie «polveri» per ampi settori dell'atmosfera, ma sono di grana grossa e non consentono previsioni precise sui dettagli. Altri, invece, sono di risoluzione molto più fine, ma questa maggiore esattezza locale rende difficile la generalizzazione globale. Basta un piccolo scarto dalla realtà nella rilevazione dei dati che vengono introdotti nel modello per ottenere una «previsione» sballata! Lo sapevano già — a proposito di meteorologia — grandi fisici e matematici all'inizio del secolo scorso, prima ancora che si disponesse del computer. L'impiego del quale oggi non dispensa da quella finezza interpretativa che è necessaria perché non si prenda il modello per la realtà, nell'illusione che «i numeri» sciolgano da sé le nostre preoccupazioni politiche ed economiche.

Creative commons licence

You are free:

- to Share - *to copy, distribute and transmit the work*
- to Remix - *to adapt the work*

Under the following conditions:

- Attribution. *You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).*
- Noncommercial. *You may not use this work for commercial purposes.*
- Share Alike. *If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.*

http://creativecommons.org/licenses/by-nc-sa/3.0/

End of slides