METODI STATISTICI PER IL MANAGEMENT

Lezione 19

Modelli CUB per l'analisi dei dati ordinali

Domenico Piccolo

Università degli Studi della Basilicata

domenico.piccolo@unibas.it

Sintesi

- *Introduzione*
- ***I. La problematica delle variabili ordinali***
- ***II. Fondamenti statistici dei modelli CUB***
- ***III. Alcune evidenze empiriche***
- ***IV. La classe dei modelli CUB***
- ***V. Una procedura operativa***
- ***VI. Un'applicazione per il dataset DIUBAS2023***
- *Considerazioni finali*

Introduzione

- ***Tutta la vita è risolvere problemi.***

(Popper, K.R., 2001)

- ***Choosing to do or not to do something is a ubiquitous state of activity in all societies.***

(Louvier *et al.*, 2000)

- ***Almost without exception every human beings undertake involves a choice (consciously or sub-consciously) including the choice not to choose.***

(Hensher *et al.*, 2005)

- ***Ogni azione è la conseguenza di una decisione che implica sempre una valutazione comparata.***

Che vuol dire **valutare**? (secondo alcuni Dizionari)

1. Determinare il valore commerciale di un bene esprimendolo in moneta, assegnare a un oggetto il valore di mercato, o quello che si ritiene giusto o conveniente *(apprezzare)*.
2. Determinare le qualità, l'importanza di qualcosa.
3. Stimare o calcolare approssimativamente *(stimare)*.
4. Tenere conto ai fini di un calcolo complessivo (conteggiare, contare), o ai fini di un giudizio di merito, di una classifica o graduatoria.
5. Considerare sotto ogni aspetto o punto di vista, esaminare rigorosamente ai fini di un giudizio complessivo *(soppesare, ponderare, vagliare)*.
6. Esaminare a fondo, giudicare il pro e il contro o le conseguenze di un'azione, attribuire una valutazione al lavoro di studenti.

- **Valutare con le mani**:

  CONTARE . . . . . variabili ***discrete***

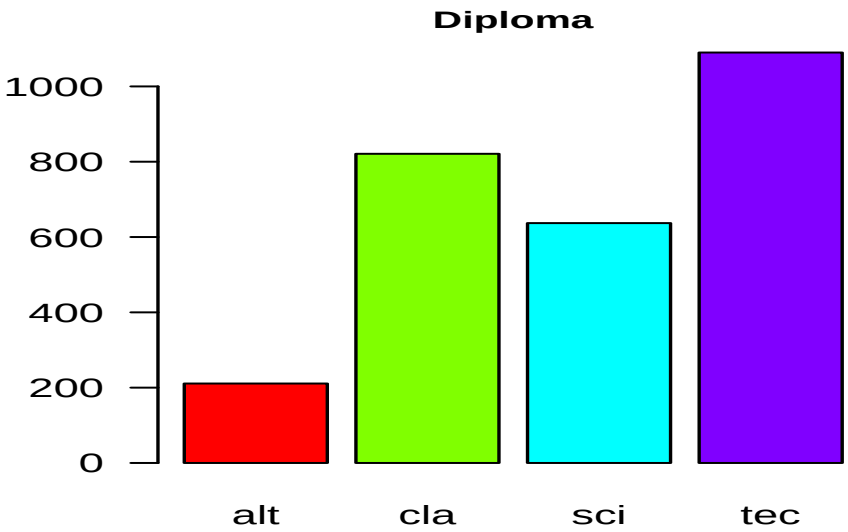
- **Valutare con oggetti di riferimento**:

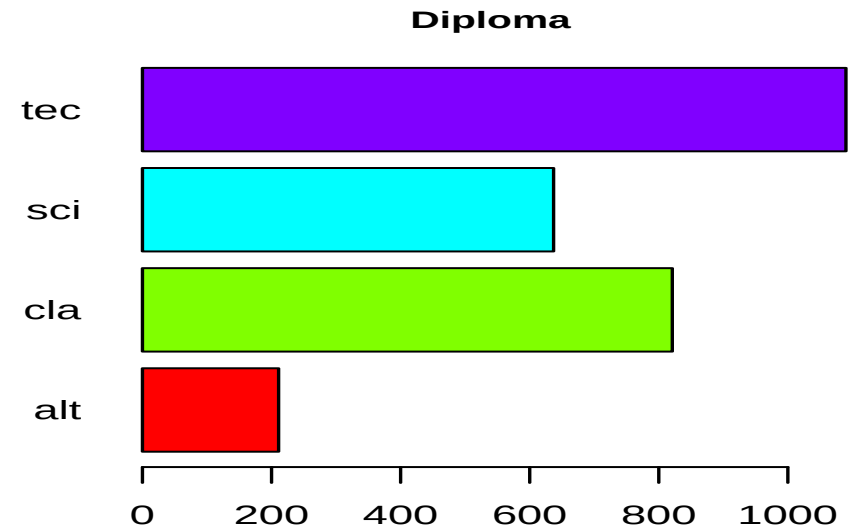
  STIMARE . . . . . variabili ***continue***

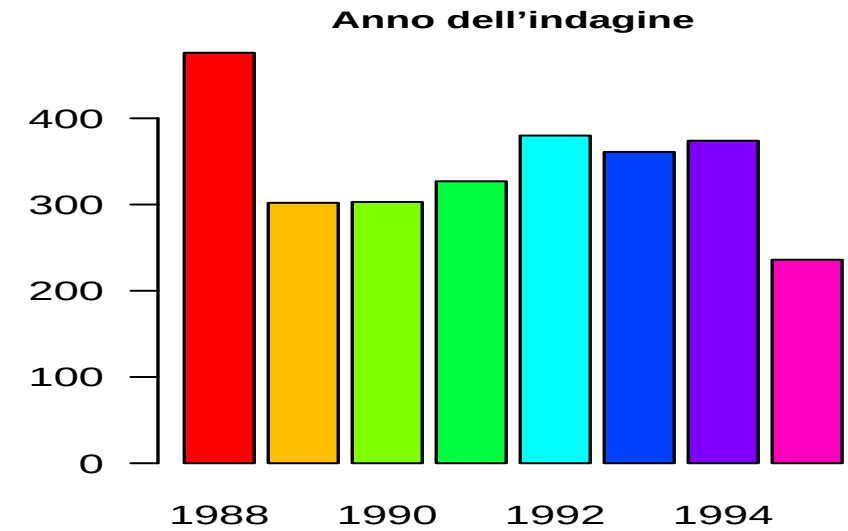
- **Valutare con la mente: opinioni, giudizi,. . .**

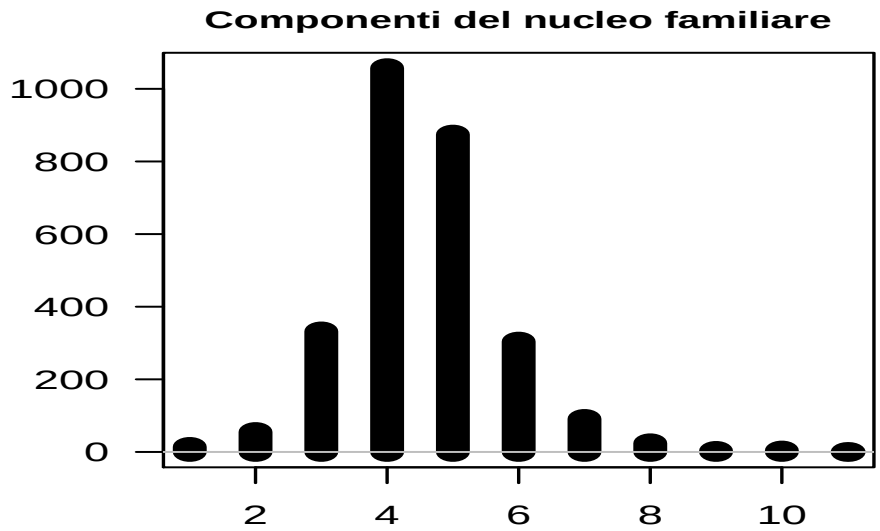
  ?!?!?! . . . . . variabili ***non osservabili***

Organizzare il "conteggio": diagramma a barre

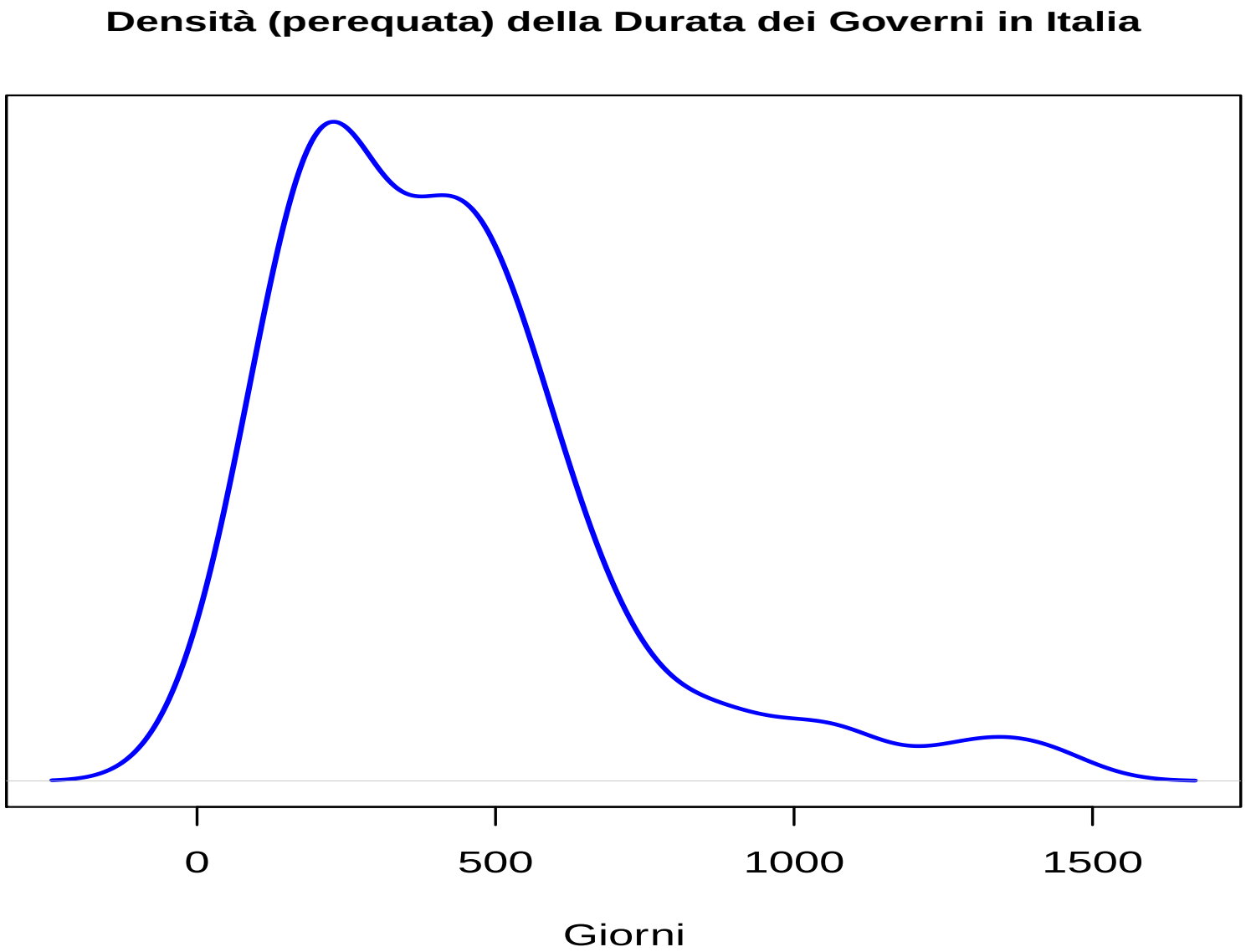








Organizzare il "continuo": la durata dei governi in Italia




Conflict among job and personal values
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Never
Seldom
Sometime
Most of time
Always

Parte I

La problematica delle variabili ordinali

- Ogni problema va affrontato/risolto mediante la sintesi di constatazioni e deduzioni, fatti e teoria, esperienza personale e storia collettiva. Tali "*informazioni*" sono la ***conoscenza umana***.
- ***La conoscenza umana sostenuta dall'evidenza empirica*** si scontra con la complessità e l'incertezza, la permanenza delle relazioni e la limitatezza delle osservazioni. L'approccio razionale per la gestione ottimale di *informazioni complesse ed incerte* si chiama ***Statistica***.
- Per prendere decisioni (lo dica oppure no), ciascuno di noi fa riferimento ad una razionalizzazione dei fenomeni il che, quando avviene in forma esplicita, implica la costruzione di un ***modello statistico***.

Cosa realmente implica un **modello statistico**

- **Il modello non esiste nella realtà fenomenica**: è solo un *costrutto mentale funzionale* ad un obiettivo decisionale .....
- .....**ma è strettamente necessario** per:
  - *Interpretazione, Previsione, Classificazione, Discriminazione, Controllo, Imputazione, Simulazione, Confronto tra scenari alternativi, Associazione, Rappresentazione,*.....
- **I "fatti" non esistono . . . esiste solo l'*interpretazione* dei fatti all'interno di un modello che interpreta quei fatti** $\rightarrow \rightarrow \rightarrow$ **Teorema di Bayes**.
- Ogni modello è la specificazione statistica del ***DATA GENERATING PROCESS*** **(DGP)** ed è strettamente necessario per inferire e, quindi, per assumere decisioni accompagnate da una probabilità di errore.
- Tale convinzione implica che non si dovrebbero studiare i dati di rating, opinioni, giudizi, valutazioni ordinali senza fare riferimento al **DGP** che li ha generati.

Percezione $\Longrightarrow$ ***Scelta*** $\Longrightarrow$ ***Decisione***

- ***La percezione di un oggetto/servizio/item è il processo psichico preliminare mediante il quale si opera una sintesi dei dati sensoriali in forme dotate di significato nella coscienza del ricevente.***
- Ogni scelta all'interno di un insieme discreto e finito di alternative possibili implica una valutazione comparata, *in modo conscio o inconscio*, che produce di fatto una graduatoria e, alla fine, una decisione giustificabile.
- In questa dinamica intervengono molteplici e complesse cause e circostanze tra le quali due ci sembrano stabilmente presenti:
  - Una ***componente primaria*** che guida/orienta la scelta.
  - Una ***componente secondaria*** che esprime l'insicurezza nella decisione.

Ordinal variables

➤ ***Ordinal variables*** associate *integers* to discrete choices in several circumstances. Each ordered category is basically **qualitative**.

Ranking

Numbers convey the location of the "object" in a given ordered list
(items, applicants, candidates, sentences, situations, teams, songs, . . .)

Rating

Numbers convey the level of a "stimulus" as perceived by the respondent
(sensation, perception, awareness, appreciation, judgment, taste, fear, worry, . . .)

➤ Hereafter, to simplify discussion, we will consider ***`ratings`*** (scores), as expressions of human decisions (choices).

Ordinal variables are ubiquitous

- Psychology and Pedagogical sciences
- Medicine
- Marketing and Economic analysis
- Evaluation of public services
- Quality control of industrial products
- Political sciences
- Sensory sciences
- Linguistics
- Sports
- .....

Origin of ordinal variables

1. *Originally ordinal*
2. *Conventionally ordered*

➤ The fundamental issue is that ordinal data are qualitative; so any metric solution is just a convention.

➤ Categories of ordinal responses may be:

- coded as **numbers**
- shown as **verbal** adjectives
- presented in a **pictorial** feature

➤ Although numerical correspondence with qualitative categories is ***ambiguous***, nevertheless it turns out to be substantially ***robust***.

➤ Some researchers prefer an even (odd) number of categories to exclude (include) an *indifferent* central option.

➤ ***As many opinions as fields of interest . . . . . .***

- In several sample surveys, human and relational variables such as *happiness, job satisfaction, well-being, quality of life, consumers' preferences, work related stress*, etc. are considered as the main topics to be investigated.

- Ordinal type scale: Respondent is asked to provide his/her agreement to a statement which provides description of ordered response levels.

Extremely unsatisfied	Very unsatisfied	Unsatisfied	Indifferent	Satisfied	Very satisfied	Extremely satisfied

- Numerical scores are often proposed in questionnaire as a method to simplify subsequent coding and analysis.

- An example in **Political Sciences**:

- A **segment** is anchored to both extremes (often denoted as "extremely low" and "extremely high", respectively). Then, interviewees are asked to put a cross on the line to indicate the level of perceived latent variable (happiness, satisfaction, stress, pleasantness, agreeableness, etc.).

extremely low ——————————————— ***extremely high***

Different modalities of collection of ordinal responses

- "Classical" questionnaires to be filled paper-pencil
- Technological tools:
  - Email interviews with a link to a website which automatically registers data
  - Responses by digits on the mobile
  - .....
- **Notice**: In surveys the scales 1-10 or 0-10 are used to relate evaluations and judgements to school marks (in Italy).
- Emoticons (or other symbols) are quite frequent in marketing surveys but may bias the result
- .....

Alternative nella scala delle risposte (ordinal hedonic scales)

➤ Several extensions, as Labeled Affective Magnitude (LAM), Just-about-right (JAR), "check-all-that-apply" (CATA), have been proved effective in specific experiments for sensory analysis of food and beverages.

Liking	*Intensity*	*Just-about-right*
Like extremely	1 Not at all	Much too strong
Like very much	2	
Like moderately	3	Somewhat too strong
Like slightly	4	
Neither like nor dislike	**5**	***Just about right***
Dislike slightly	6	
Dislike moderately	7	Somewhat too weak
Dislike very much	8	
Dislike extremely	9 Extremely	Much too weak

➤ LAM scales use 11 categories including two extremes:

"*beyond any conceivable degree* . . . "

➤ Our experience suggests that ***scales with a number of odd ordinal categories greater or equal to*** 7 have to be preferred for modelling purposes. Sometimes, a $1-10$ range may be practical.

Dimensioni della scala e incertezza delle risposte

➤ Scales should have *sufficient* levels to allow discrimination in the respondents' judgement *but not so many* to cause avoidable indecision and confusion.

➤ The argument is highly controversial:

- Some people prefer a small value of m, the number of ordinal categories, to make the selection easier
- Some people prefer a great value of m to make the choice more selective

➤ Is it correct to say: "*the larger the scale, the larger the indecision . . . ?*"

➤ The "true" problem is:

are we *creating* or *disclosing* uncertainty ?

A paradigm for statistical analysis

➤ All statistical analysis, including those pertaining to ordinal rating data, should be performed according to a scientific paradigm.

1. **Explore available data**

2. **Detect the data generating mechanism** by means of:
   - a family of probability distributions
   - statistical inference methods

3. **Build effective models** in order to:
   - understand
   - predict
   - classify
   - discriminate
   - .....

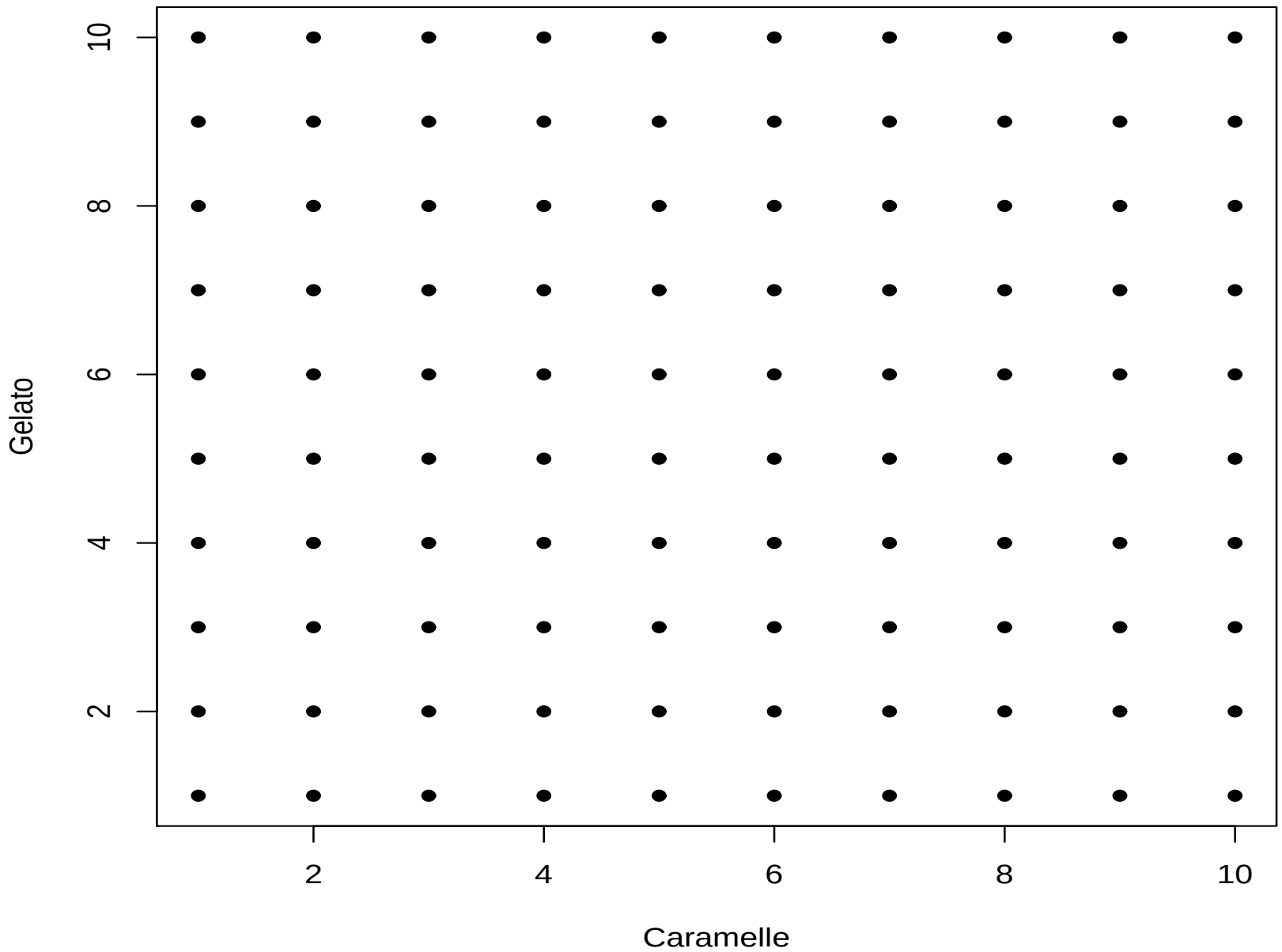
Standard approaches for the analysis of ordinal ratings

Plots
- Bar plots
- *Do not use*: kernel histograms, box-plots, etc.
- .....

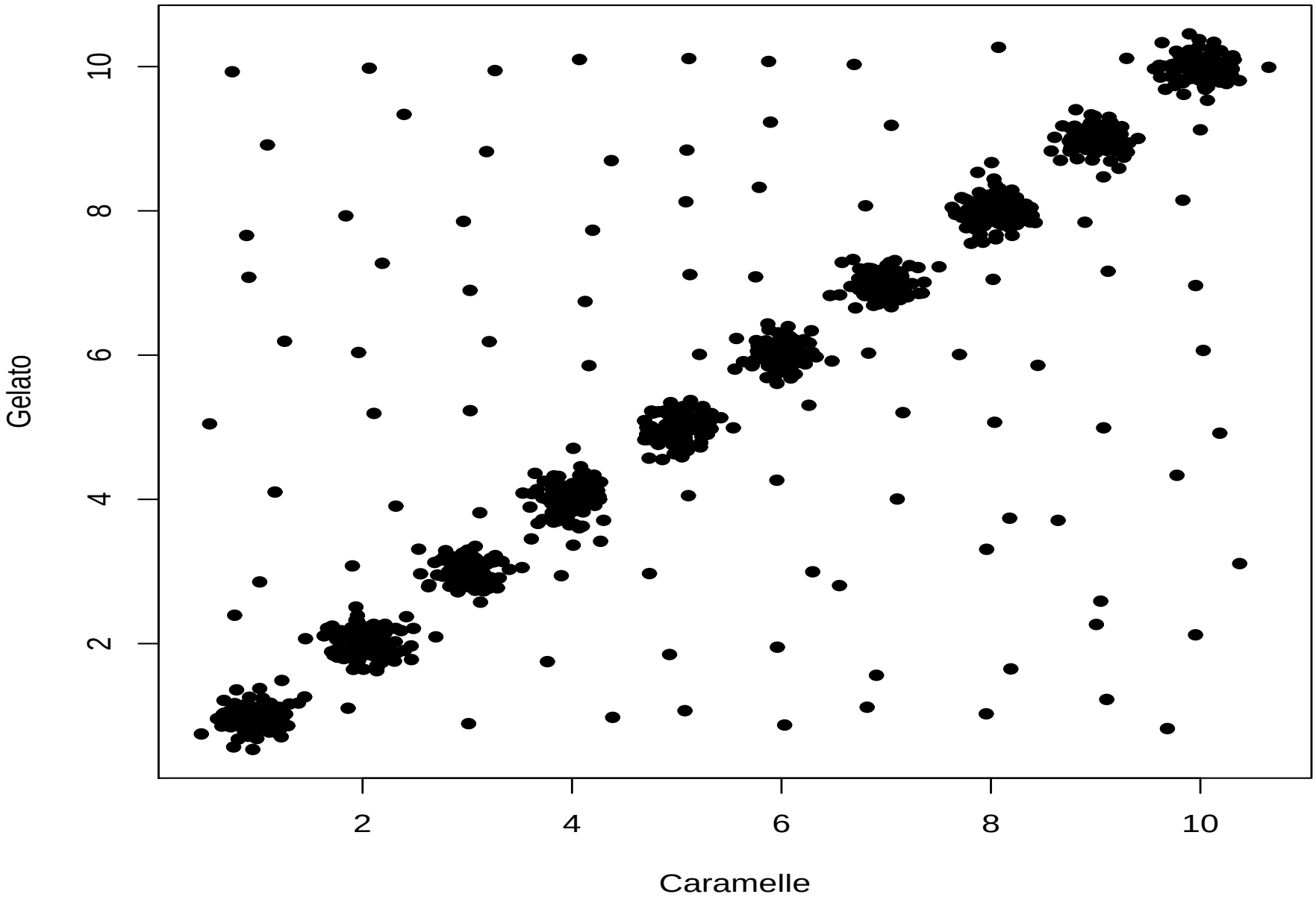
Measures
- Averages (*be careful with median!*)
- Standard deviation
- Dispersion indexes
- Mean difference
- Heterogeneity measures
- .....

Models
- Cumulative regression models
- Probit analysis
- Logit models
- .....
- **CUB models**
- .....

Grafico della relazione tra preferenze: Gelato e Caramelle



Relazione tra preferenze jittered



Il progetto RELGOODS

- **Description**: Dataset consist of the results of a survey (December 2014) aimed at measuring the evaluation of people living in the metropolitan area of Naples, Italy, with respect to of relational goods and leisure time.
- **Ordinal assignment**: Every participant was asked to assess on a 10 point ordinal scale his/her personal score for several relational goods and to leisure time.
- **Mode of collection**: questionnaire
- **Number of observations**: 2459
- **Number of subjects' covariates**: 16
- **Number of analyzed items**: 34
- **Warning**: a limited number of missing values
- **R Documentation**: **?relgoods** in the R package "CUB"

➤ Original source:

`http://www.labstat.it/home/wp-content/uploads/2015/09/relgoods.txt`

Questionario RELGOODS: informazioni sul rispondente

- Genere: *Uomo* *Donna*
- Mese di nascita:.....
- Anno di nascita:.....
- Componenti famiglia (incluso l'intervistato): .....
- Vi sono componenti di età minore di 12 anni? *Si* *No*
- Titolo di studio: *Scuola obbligo* *Diploma* *Laurea triennale* *Laurea magistrale* *Post-laurea*
- Stato civile: *Celibe/Nubile* *Coniugato/Convivente* *Separato/Divorziato* *Vedovo*
- Residenza: *Città NA* *Provincia NA* *Altre CAMPANIA* *Altre ITALIA* *ESTERO*
- Occhiali da vista (lentine): *Si* *No*
- Scrittura con la mano: *Destra* *Sinistra*
- Fumo: *Si* *No*
- Per lo più, passeggi: *Da solo* *Con altri*
- Lavori? *No* *Pensionato* *Occasionalmente* *A tempo determinato* *A tempo indeterminato*
- Pratichi qualche sport: *No* *Individuali* *Di squadra*
- Possiedi un animale in famiglia? *Si* *No*

Questionario RELGOODS: beni relazionali

➥ Nelle risposte alle domande seguenti, tieni presente che [1] significa “mai, per niente, molto raramente, pochissimo” e [10] significa “sempre, molto spesso, moltissimo”.

• Con quale frequenza fai una *passeggiata* all’aria aperta?	1	2	3	4	5	6	7	8	9	10
• Quanto spesso parli con almeno uno dei tuoi *genitori*?	1	2	3	4	5	6	7	8	9	10
• Quanto spesso incontri altri *familiari* (nonni, zii, cugini, nipoti, etc.)?	1	2	3	4	5	6	7	8	9	10
• Con quale frequenza/intensità sei *coinvolto in associazioni* culturali o religiose, gruppi di volontariato, partiti, sindacato, etc.	1	2	3	4	5	6	7	8	9	10
• Definiresti positivi i tuoi *rapporti con gli amici*?	1	2	3	4	5	6	7	8	9	10
• Definiresti positivi i tuoi *rapporti con i vicini*?	1	2	3	4	5	6	7	8	9	10
• Se hai *bisogno di aiuto*, lo chiedi facilmente agli altri?	1	2	3	4	5	6	7	8	9	10
• Definiresti positivi i tuoi *rapporti con l’ambiente* (di studio, di lavoro, di tempo libero) nel quale trascorri gran parte del tuo tempo?	1	2	3	4	5	6	7	8	9	10
• Ti senti *sicuro per le strade* del paese in cui adesso vivi?	1	2	3	4	5	6	7	8	9	10
• Tu e la tua famiglia *arrivate facilmente a fine mese* dal punto di vista economico?	1	2	3	4	5	6	7	8	9	10

Dataset RELGOODS

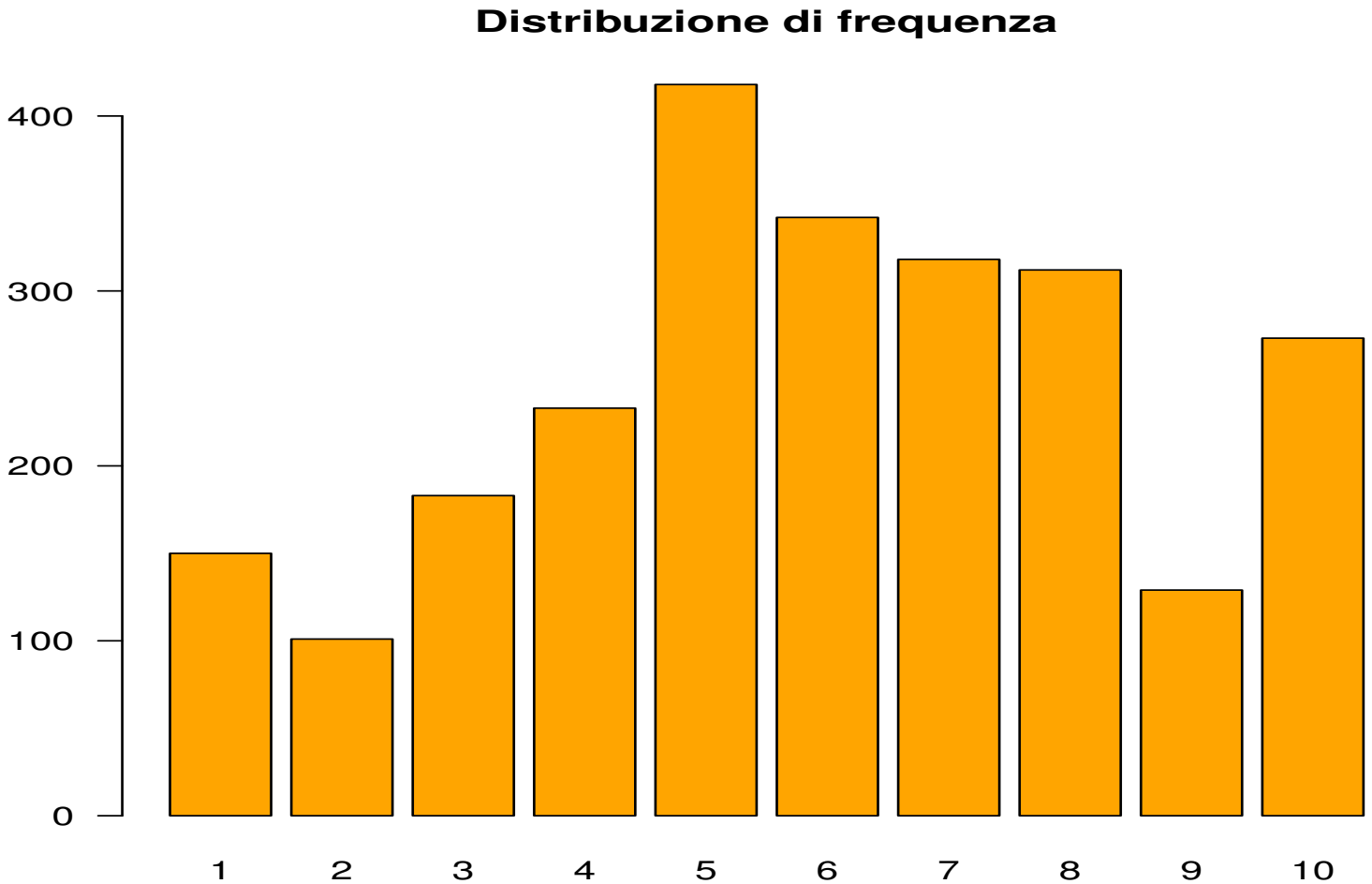
```
#############################################################################################

ID      Gender  BirthMonth      BirthYear       Family  year<12 EducationDegree MaritalStatus   Residence       Glasses RightHand       Smoking WalkAlone       Job     PlaySport       Pets    WalkOut Parents
MeetRelatives   Association     RelFriends      RelNeighbours   NeedHelp        Environment     Safety  EndofMonth      Happiness       MeetFriend      Physician       Videogames      Reading Cinema  Drawing
Shopping        Writing Bicycle Tv      StayWFriend     Groups  Walking HandWork        Internet        Sport   SocialNetwork   Gym     Quiz    MusicInstr      GoAroundCar     Dog     GoOutEat
1   1   4   1972    3   1   5   2   3   0   1   0   0   1   1   0   5   8   8   8   8   7   8   7   4   5   104 1   8   3   5   3   4   6   2   1   4   7   3   2   6   8   1   2   1   3   1   6   1   3
2   0   10  1976    3   1   4   2   3   1   0   0   1   5   1   0   6   3   3   3   5   5   4   6   4   5   95  0   7   1   8   3   1   4   5   1   5   5   1   7   1   5   1   1   1   1   1   4   1   5
3   1   12  1964    3   0   4   1   3   0   1   0   0   5   1   0   6   9   8   2   7   7   5   8   4   9   86  4   8   1   4   5   3   8   1   1   6   6   1   7   5   5   1   2   1   1   1   5   1   8
4   1   4   1958    3   0   2   2   3   1   1   0   1   2   1   0   2   1   5   1   5   5   5   5   4   7   92  2   8   1   1   1   1   1   1   1   8   4   2   3   8   1   1   1   1   1   1   2   1   1
5   0   3   1992    5   0   3   1   3   1   1   0   1   5   2   1   5   9   7   7   10  10  6   9   5   10  84  2   7   9   9   9   5   8   1   1   7   10  9   6   7   8   10  8   9   5   1   5   6   8
6   0   12  1989    4   0   3   1   4   1   1   0   0   5   1   0   7   6   2   2   7   3   7   5   7   7   52  2   7   4   3   4   3   4   1   2   3   9   5   8   5   9   8   8   8   5   1   8   1   8
7   0   7   1990    4   0   3   1   4   0   1   0   0   5   3   0   8   5   4   1   10  6   4   6   4   10  53  4   4   5   7   5   5   6   4   2   2   10  5   8   2   9   8   9   8   2   6   6   2   7
8   1   9   1991    4   1   3   1   2   0   1   1   0   3   1   1   10  10  6   1   8   1   5   8   5   4   71  6   10  1   10  7   10  7   7   5   1   10  5   10  10  7   7   7   7   5   7   7   7   5
9   0   8   1990    5   0   3   1   2   0   1   0   0   3   2   1   10  9   9   8   10  6   7   8   7   6   30  6   8   3   8   8   1   7   1   7   7   9   5   8   1   9   10  9   10  1   1   8   6   7
10  0   8   1990    6   0   3   1   2   0   1   1   0   3   1   0   5   10  7   10  10  8   7   9   9   8   55  4   10  1   5   5   5   5   7   1   5   8   10  5   6   10  5   8   1   5   10  10  1   6
11  1   10  1989    4   0   3   1   3   1   1   0   0   3   2   1   6   7   5   8   10  4   7   6   5   5   55  6   7   1   2   6   7   5   4   4   5   8   9   9   9   7   7   4   3   3   3   2   9   7
12  1   9   1987    6   0   3   1   3   1   1   0   0   5   1   0   5   9   4   2   10  7   3   8   4   3   53  3   7   1   3   5   1   5   1   1   2   6   3   2   8   7   2   3   1   1   1   1   1   5
13  1   11  1988    5   0   3   1   2   1   1   0   0   1   1   0   6   8   8   2   9   5   4   7   4   8   86  2   8   1   8   7   6   9   3   5   7   7   3   7   5   10  3   8   2   1   1   6   1   7
14  1   10  1990    4   0   3   1   2   1   1   0   0   3   2   1   8   8   10  4   8   6   6   9   9   9   61  6   9   2   8   7   8   7   4   5   2   6   4   9   3   7   8   7   8   2   1   6   1   8
15  1   5   1986    4   0   3   1   1   1   1   0   0   3   1   1   5   10  7   4   8   7   7   8   5   7   76  4   8   1   9   4   6   5   8   1   3   7   5   8   9   6   1   6   1   2   1   5   1   7
16  1   9   1991    4   0   3   1   2   0   1   0   0   3   2   0   5   3   8   2   8   9   8   9   8   9   89  1   2   1   3   4   4   6   3   7   6   6   4   6   7   6   8   5   8   2   1   4   1   7
17  1   6   1990    3   0   3   1   2   0   1   0   1   1   2   0   8   9   7   1   7   4   5   6   4   8   82  2   10  1   5   7   6   10  6   5   8   9   4   10  5   9   7   8   8   6   1   5   1   7
18  1   12  1989    5   0   3   1   2   0   1   0   0   1   1   1   6   10  10  1   3   3   4   7   5   4   93  1   9   1   2   3   1   3   1   1   3   4   1   3   1   10  1   3   1   1   1   10  1   8
19  1   10  1991    7   0   3   1   3   1   1   0   0   1   1   0   7   8   8   2   6   7   7   7   8   7   50  3   8   1   6   6   6   7   5   1   2   7   2   6   6   7   2   8   2   2   1   2   1   7
20  1   7   1989    4   0   3   1   3   0   1   0   0   4   2   1   6   8   9   3   10  9   6   8   4   9   71  3   8   1   5   5   1   6   1   1   6   8   6   6   2   9   8   1   6   1   1   5   2   6
21  1   7   1989    5   0   3   1   3   1   0   0   0   1   1   0   7   9   9   2   8   8   6   7   7   8   73  6   10  4   8   10  8   8   6   7   3   9   7   7   8   8   8   6   7   9   9   9   8   6
22  1   12  1988    4   0   3   1   4   0   1   1   0   1   1   1   1   5   4   1   8   1   5   7   2   7   65  6   10  4   6   2   2   1   1   1   9   10  1   1   1   10  1   5   1   10  1   10  10  10
23  1   7   1988    5   0   3   1   2   0   1   0   0   1   2   0   5   8   7   1   7   7   7   6   7   5   77  2   8   1   2   2   1   5   2   1   3   3   1   2   1   6   7   6   1   1   1   9   1   6
24  0   8   1991    3   0   3   1   2   1   1   0   0   1   2   0   6   4   4   1   8   1   1   9   1   7   69  6   1   1   1   2   8   6   1   1   1   8   1   6   1   6   10  10  10  1   1   10  1   10
25  0   6   1987    5   0   3   1   2   0   1   1   0   3   1   0   8   5   5   1   10  8   6   9   7   5   84  3   10  1   10  1   5   2   1   1   1   10  1   8   1   5   1   1   1   1   1   1   1   1
26  1   5   1966    4   0   1   2   2   0   1   0   0   1   1   1   8   9   10  10  10  10  6   8   6   4   78  4   8   1   7   4   10  5   4   1   7   9   10  8   3   8   4   8   4   8   1   10  10  1
27  0   8   1965    5   1   3   2   1   1   1   0   0   5   2   0   6   10  10  10  10  10  10  10  6   5   110 6   10  1   6   6   6   6   6   1   6   10  10  6   4   6   10  10  5   5   6   10  1   6
28  1   5   1961    5   0   2   2   2   0   1   0   0   1   1   1   5   1   6   6   8   8   8   8   5   6   70  4   10  1   4   5   9   3   2   1   5   10  7   6   10  1   2   1   1   4   3   3   1   1
29  1   10  1959    4   0   2   2   2   1   1   0   0   5   2   1   8   2   10  8   8   4   5   8   6   7   75  3   8   1   10  10  10  2   2   5   6   10  8   9   7   6   7   7   7   4   7   5   1   1
30  1   10  1962    4   0   2   2   1   1   1   0   1   1   3   1   8   1   6   5   7   7   1   6   1   8   105 5   8   1   7   6   1   2   4   1   5   7   6   7   9   6   9   1   8   3   1   4   4   3
31  0   5   1974    4   1   2   2   2   1   1   1   0   5   2   1   10  9   6   10  10  10  8   8   7   7   97  4   10  1   10  5   8   6   6   8   6   8   8   10  7   6   9   2   5   4   8   4   9   5
32  1   1   1960    2   0   5   3   1   1   1   0   1   5   1   0   8   10  10  8   8   6   9   8   3   10  110 6   10  1   10  5   3   5   6   1   6   10  7   6   9   6   3   2   1   5   1   7   1   2
33  1   11  1948    2   0   2   2   2   1   1   0   0   2   1   0   1   1   1   10  10  10  10  10  1   5   49  2   5   1   1   1   1   1   1   1   10  10  1   1   1   1   1   1   1   10  1   1   1   1
34  0   9   1988    4   0   2   1   2   1   1   0   0   3   1   1   0   5   3   3   10  7   6   8   6   3   3   52  6   3   1   5   3   10  5   1   1   5   9   8   5   5   10  1   10  1   3   10  10  1   7
35  1   7   1987    5   0   3   1   2   1   1   0   0   3   1   0   5   7   6   5   4   2   5   6   4   6   79  1   9   1   4   6   3   5   2   4   4   6   6   6   4   6   6   5   5   3   1   5   4   7
36  0   8   1976    3   0   1   1   2   0   1   0   0   1   1   0   5   5   1   10  10  10  5   5   10  5   96  6   10  1   6   6   6   6   1   1   1   10  10  6   6   10  8   10  6   1   1   8   5   10
37  0   4   1964    2   1   2   3   2   1   1   0   1   5   1   0   5   5   4   4   4   5   2   4   2   2   53  4   6   1   9   9   8   3   9   10  6   10  9   7   7   2   9   5   8   7   9   4   3   6
38  0   9   1954    3   0   4   2   2   1   1   0   1   5   1   1   3   1   2   3   5   5   5   8   8   7   46  1   6   1   5   2   7   1   1   2   5   6   4   8   3   6   4   3   3   6   5   3   6   2
39  1   4   1974    7   0   2   2   2   0   1   0   0   4   1   0   1   1   10  10  1   1   10  1   1   1   83  0   10  1   7   1   1   10  1   1   1   1   10  1   1   1   1   1   1   1   1   1   1   1
40  1   2   1961    4   0   2   2   2   0   1   0   0   1   1   0   7   1   6   5   7   8   3   5   1   1   110 6   6   1   7   6   6   5   1   1   10  10  5   10  10  6   6   1   7   10  1   6   1   6
41  0   10  1984    3   0   4   1   1   0   1   0   1   3   1   1   3   3   2   4   8   8   2   5   5   4   55  2   6   5   8   4   2   2   3   1   5   5   3   4   2   8   2   4   1   4   1   1   5   5
42  0   9   1987    4   0   3   1   4   1   1   0   0   1   1   0   5   8   6   1   10  5   4   10  4   7   83  6   6   3   2   10  1   2   1   1   1   10  1   5   1   7   1   3   1   7   1   1   1   8
```

Esempio: risposte alla prima domanda

➤ **WALK:** *Con quale frequenza fai una* ***passeggiata*** *all'aria aperta?*

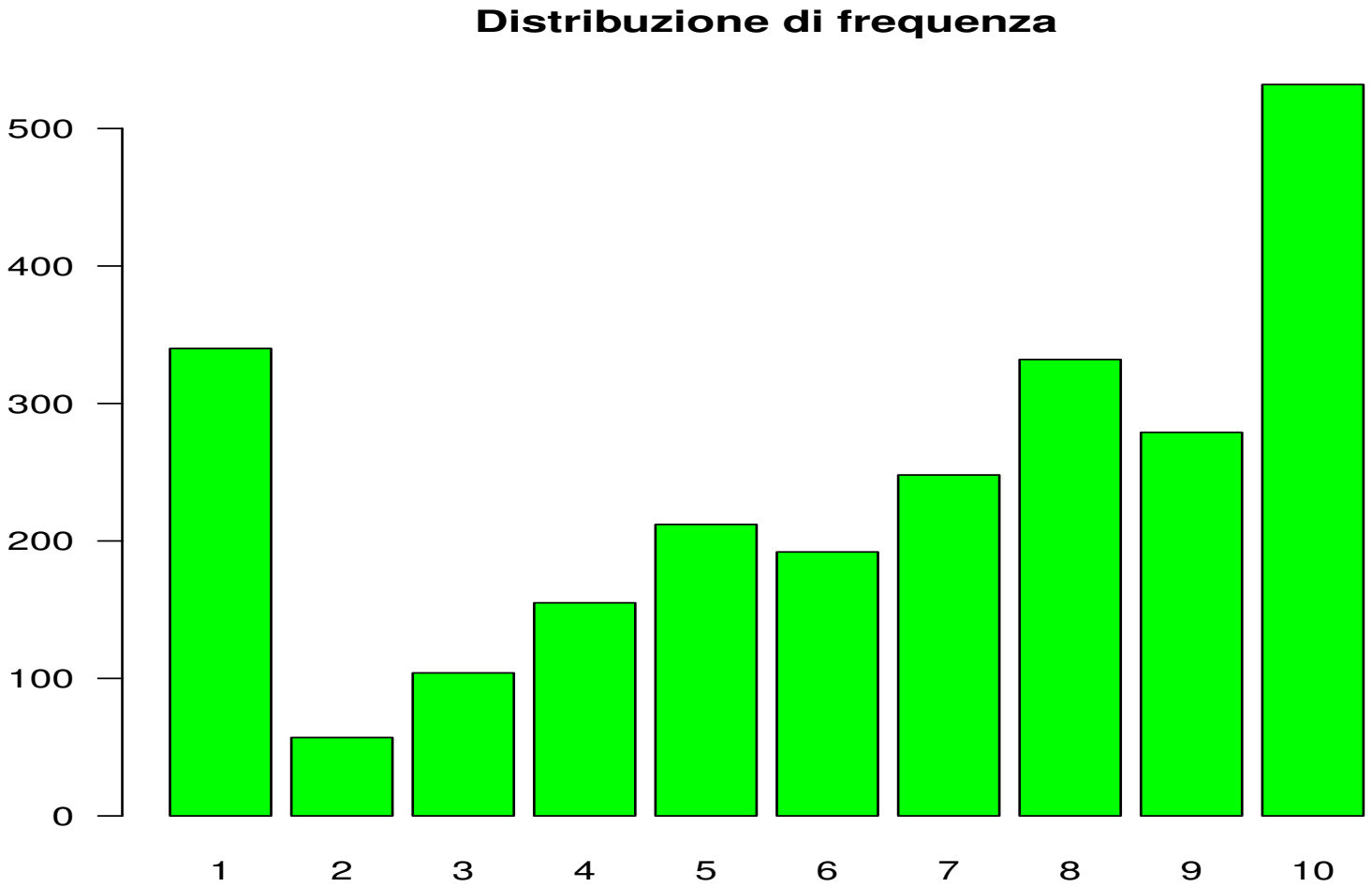
5, 6, 6, 2, 5, 7, 8, 10, 10, 5, 6, 5, 6, 8, 5, 5, 8, 6, 7, 6, 7, 1, 5, 6, 8, 8 , 6, 5 , . . .



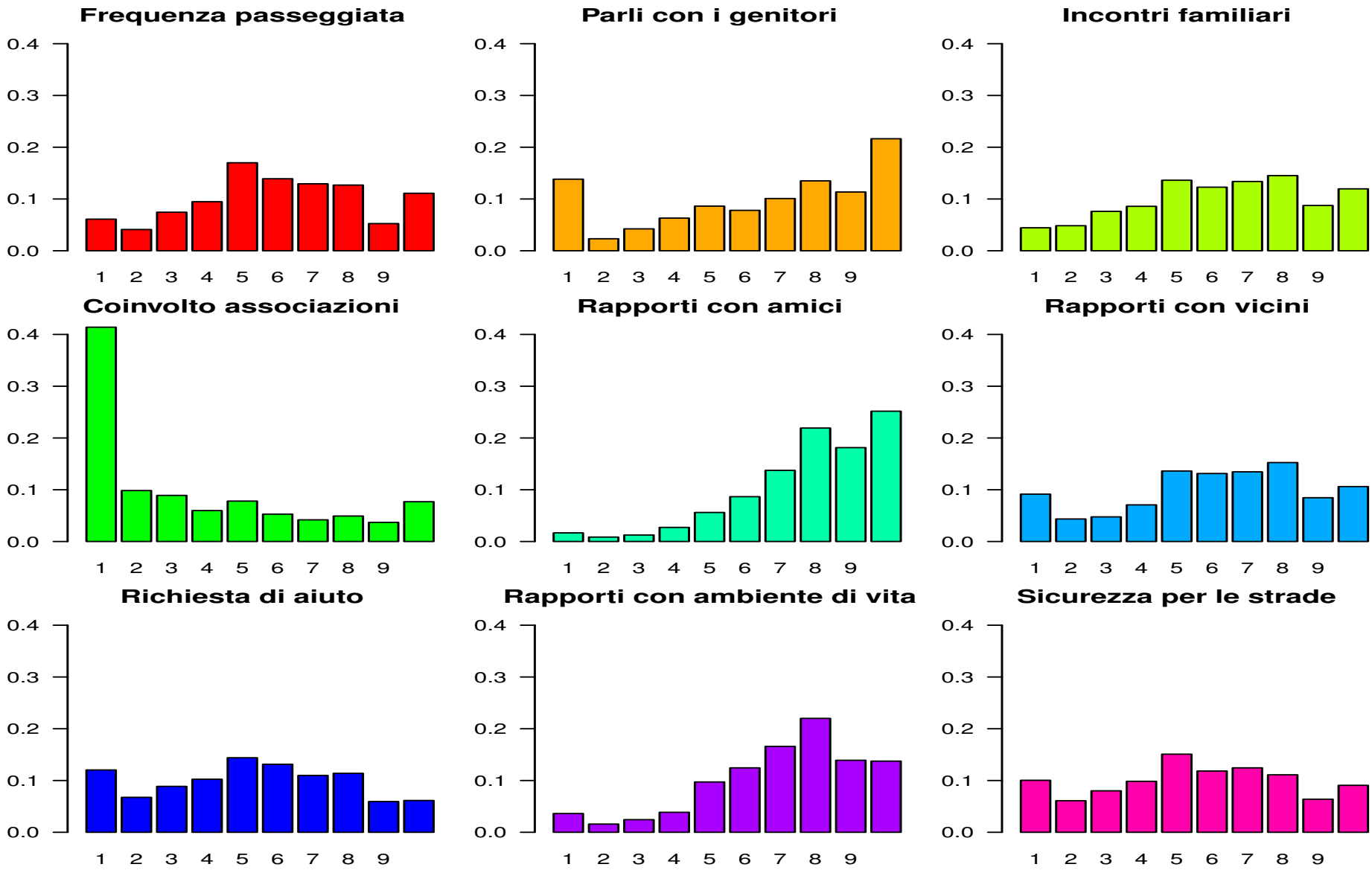
Esempio: risposte alla seconda domanda

➤ **PARENTS:** *Quanto spesso parli con almeno uno dei tuoi* ***genitori****?*

8 , 3 , 9 , 1 , 9 , 6 , 5 , 10 , 9 , 10 , 7 , 9 , 8 , 8 , 10 , 3 , 9, 10 , 8 , 8 , 9 , 5 , 8, . . .



La distribuzione di frequenza di 9 items



- Come si analizzano queste risposte?
- Come si differenziano tra di loro?
- Esistono diffenze nelle risposte in funzione della tipologia dei rispondenti, per genere, età, lavoro, cultura, sport, etc?
- Come si rappresentano (visualizzano) risposte di tipo ordinale in modo da evidenziare anche le differenze per sottogruppi?
- .....
- ***I modelli che si presenteranno cercano di offrire una risposta a queste (e a molte altre) domande sui dati ordinali mediante una nuova classe di modelli, chiamati modelli CUB , la cui introduzione deriva al processo generatore dei dati.***

Parte II

Fondamenti statistici dei modelli CUB

➤ ***Why a new statistical model for ordinal data?***

➤ A statistical model should be:

- ***identifiable*** in the probability structure
- ***parsimonious*** in the number of parameters
- ***flexible*** with respect to observed distributions
- ***interpretable*** in a simple way for the users
- ***useful*** in real applications
- .....

➤ The family of probability mass functions we will discuss is derived on the basis of the investigation of the ***psychological mechanism*** by which people choose among a list of ordered categories.

➤ As a consequence, these models try to mimic the ***data generating process*** leading to the selection of an ordinal rating.

Human mind: perceptual and decisional aspects

➤ What happens when people have to express their judgement, agreement, worry, etc. towards an item by selecting a category in a list of m ordered alternatives?

➤ Psychologists assess that human brain activates two main aspects:

Perceptual aspects: *the rater's perception of the item content*

Decisional aspects*: the rater's use of the available scale*

➤ A correct statistical model should mimic the ***Data Generating Process*** of the transition from *Perception* to a *Discrete Choice* by taking all components of the process into account.

Come (e quando) nasce l'idea dei modelli CUB ?. . .

➤ Nel 1999 bisognava pianificare la seconda edizione (2000) del mio libro "STATISTICA" ed era utile distinguerlo cambiando il *colore* alla copertina.

➤ Angela D'Elia ed io chiedemmo ad un paio di centinaia di studenti di esprimere una personale graduatoria di preferenza per una dozzina di colori.

➤ In modo netto, la moda risultò essere il colore "Blue" e, come molti sanno, questo colore tuttora è dominante anche per la III edizione (2010).

➤ Questa esperienza ci fece constatare alcuni fatti:

- Gli studenti avevano preferenze differenti.
- Quando selezionavano una casella la loro penna si orientava anzitutto in una direzione ben definita della scala ma, poi, quasi sempre, la mano "***oscillava***" prima di individuare la scelta definitiva.
- L'oscillazione della penna "cambiava" di ampiezza e di durata: da quelli *risoluti e decisi* (poca variazione, breve tempo) a quelli *più indecisi* (maggiori variazioni, più tempo)

➤ DP ha un *background probabilistico*, AD ha un *background modellistico*: così cercammo il **DGP** più idoneo per rappresentare le scelte su scala ordinale.

➤ Per un po' di tempo, leggemmo di Psicologia e parlammo con psicologici.

➤ Teoria ed esperienze ci convinsero che la scelta su una scala ordinale è il risultato di due meccanismi che –solo per semplificare la discussione– chiamai *feeling* e *uncertainty*.

➤ In effetti, come è giusto che avvenga nella ricerca scientifica, il percorso non fu né lineare né immediato:

- tra il 1999 e 2000/2001: modelli IHG e modelli SB (BIT);
- nel 2003: modelli CUB (all'epoca si chiamavano MUB);
- dopo il 2008: una sequenza ancora aperta di proposte, varianti e generalizzazioni.

Un albero che continua a crescere .....

- La reazione di tanti statistici e di *tutte le riviste*.
- Solo dopo il 2005-2008, con la pubblicazione di due papers su riviste internazionali (*CSDA* e *FQP*), la comunità scientifica comincia ad incuriosirsi ai modelli CUB.
- E, oggi ?! .....
- **Classe dei modelli CUB**
- Almeno quattro filoni di lavori/ricerche:
  - **Metodologia**: varianti e generalizzazioni
  - **Applicazioni**: Settori disciplinari molto vari
  - **Utilizzo**: Inserimento della modellistica in altre metodologie
  - **Software**: Costruzioni di packages in vari ambienti

Latent components for expressing judgments

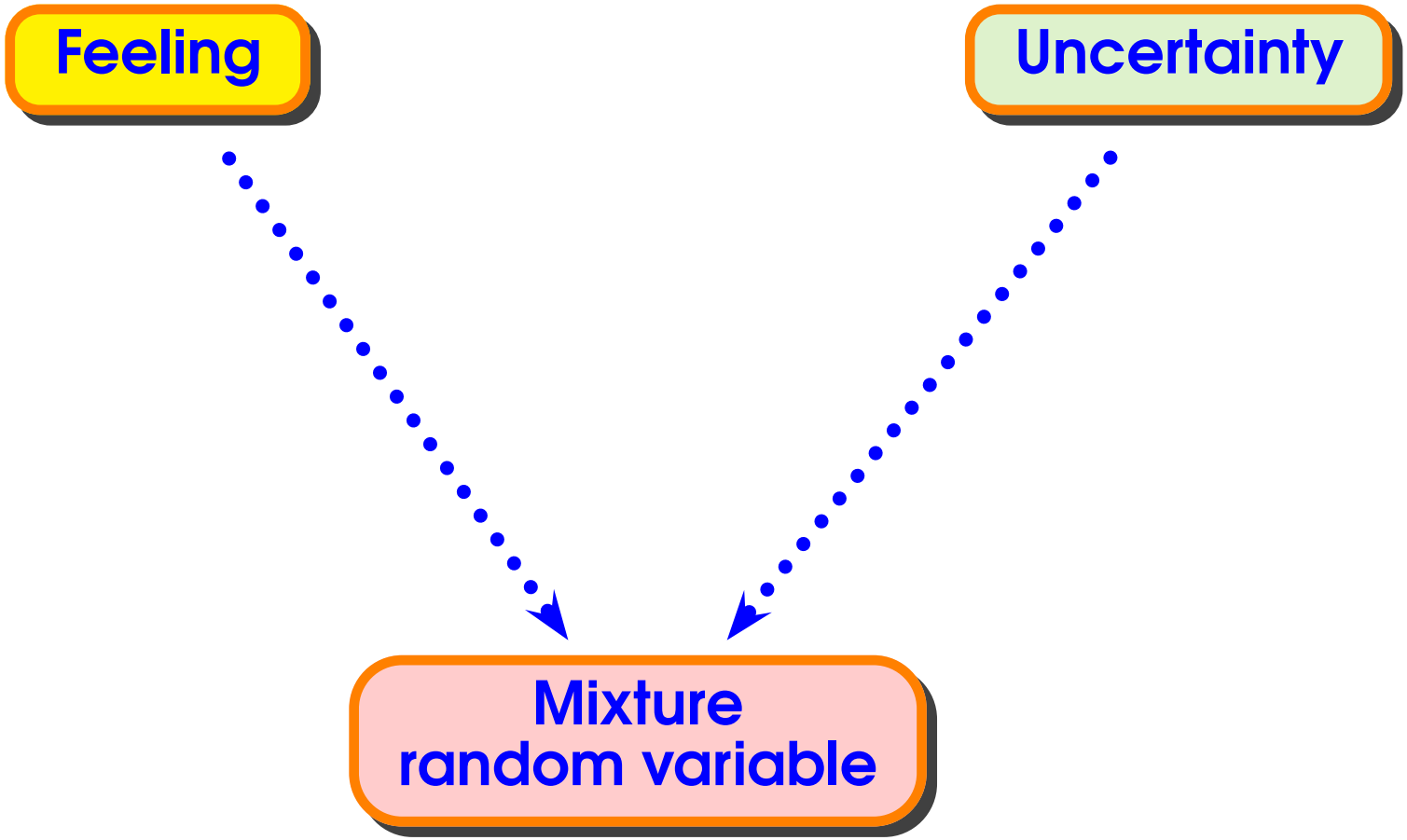
➤ Indeed, in this psychological process we recognize the response as the result of:

- a *primary* component, generated by the sound impression of the respondent, related to *awareness* and *full understanding* of the problem. We call it ***feeling*** *(agreement)*, and it is usually related to subjects' motivations;
- a *secondary* component, generated by the *intrinsic indecision* about the final choice. We call it ***uncertainty*** *(fuzziness)*, and it is mostly dependent on circumstances that surround the evaluation process.

$$\begin{cases} \textbf{FEELING} \\ \textbf{UNCERTAINTY} \end{cases} \Longrightarrow \boxed{\textit{ORDINAL CHOICE}}$$

➤ Both components will be explicitly modelled by ***discrete random variables***.

From latent variables to discrete choices



Perception

➤ ***Perception*** is a basic component in the formation of a concept: it is the ability to see, hear or understand things and it usually expresses the awareness of something via one's senses.

➤ Formally, ***perception is a cognitive act*** by which a person interprets and organizes several sensations in order to identify a specific object/situation. Thus, perception of an object/service/item is a psychological process by which a subject synthesizes sensory data in forms that are meaningful for his/her conscience.

➤ Indeed, when we ask a person to answer a specific item of a questionnaire, we are looking for his/her perception of the problem; specifically, we are asking to summarize his/her perception into a well defined category (qualitative, quantitative, mixed, verbal, and so on).

➤ Since perception is a complex function of several causes (personal, familiar, environmental, social, and so on), the expression of such a perception is affected both by the real consideration of problems and by inherent uncertainty that accompanies any human decision.

What is uncertainty?

➤ Uncertainty ***is not*** the stochastic component related to the sampling experiment (so that different people generates different ratings).

➤ Uncertainty ***is*** the result of possible convergent and related factors:

- *Limited set of information, Knowledge/Ignorance* of properties and/or characteristics of the object/item to be evaluated.
- *Personal interest/Engagement* in activities related to the specific or related field of interest.
- *Amount of time* devoted to the response.
- *Operational mode* for responding: face-to-face, questionnaire form, telephone, mobile, PC, mail, Email, etc.
- *Nature of the scale* in terms of range and wording.
- *Tiredness or fatigue* for a correct comprehension of the wording.
- *Willingness to joke and fake*.
- *Lack of self-confidence* of the respondent.
- *Laziness/Apathy/Boredom* in the selection mechanism.
- .....

➤ ***Uncertainty is not randomness*** as the first one depends on the modality of the choice while the latter is related to the experimental variability.

➤ These aspects may be connected in some way, but their concepts are quite different and require dissimilar approaches:

RANDOMNESS	$\longrightarrow$	*Variability*	$\longrightarrow$	**Random sample theory**
UNCERTAINTY	$\longrightarrow$	*Heterogeneity*	$\longrightarrow$	**Mixture modelling**

➤ It is important to distinguish between **variability** and **heterogeneity** in ordinal data distributions.

What ***really*** is uncertainty?

➤ The measure of uncertainty conveyed by $1-\pi$ includes at least three points of view:

1. ***subjective indecision***: when we examine $1-\pi_i$, it is possible to consider it as a measure of personal indecision of the i-th respondent as a function of selected covariates.

2. ***heterogeneity***: when we analyze a global CUB model for the given item, it is possible to consider $1-\pi$ as a measure of heterogeneity of the respondents.

3. ***predictability***: if we study a CUB model to predict ordinal outcomes, it is possible to consider π as a direct measure of predictability of the model with respect to two extremes:
   - a *minimum*, when responses follow a pure discrete Uniform distribution
   - a *maximum*, when responses follow a pure Binomial distribution

Rationale for modelling the uncertainty component

➤ In case the subject shows indifference (=*equipreference*) towards a given item, it seems appropriate to model the respondent's choice by means of a discrete Uniform random variable U with a probability mass function defined by:

$$Pr\,(U = r) = \frac{1}{m}, \quad r = 1, 2, \ldots, m\,.$$

➤ In this way, the choice is the result of a complete randomized mechanism since any item has the same probability of receiving any score $r \in [1, m]$.

➤ The discrete Uniform random variable maximizes the entropy, among all the discrete distributions with finite support $\{1, 2, \ldots, m\}$, for a fixed m.

➤ It adds no information to the selection mechanism preferred by respondents.

Rationale for modelling feeling

➤ In CUB model approach, the main discrete model to capture feeling of the respondent is the (shifted) **Binomial random variable**, which has been shifted to the support $\{1, 2, \ldots, m\}$ to take account of the range of modalities in common questionnaires.

➤ Here, we skip a formal and rigorous presentation of the reasons justifying this random variable to specify the feeling distribution.

➤ Briefly, arguments are based on:

- ***Discretization of Central Limit Theorem***
- ***Combinatorial aspects of the decision process***
- ***Empirical evidence***

➤ Is is possible to relate to further generalizations by using a *Beta-binomial distribution*, for instance.

Sample data and subjects' covariates

➤ From an available dataset (matrix) $\boldsymbol{T}$ we derive the $\boldsymbol{Y}$ and $\boldsymbol{W}$ matrices containing covariates useful for interpreting *uncertainty* and *feeling*, respectively.

➤ Information able to characterize both perception and uncertainty are collected in the matrices:

$$\boldsymbol{Y} = \begin{pmatrix} 1 & y_{11} & y_{12} & \dots & y_{1p} \\ 1 & y_{21} & y_{22} & \dots & y_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & y_{i1} & y_{i2} & \dots & y_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & y_{n1} & y_{n2} & \dots & y_{np} \end{pmatrix}; \qquad \boldsymbol{W} = \begin{pmatrix} 1 & w_{11} & w_{12} & \dots & w_{1q} \\ 1 & w_{21} & w_{22} & \dots & w_{2q} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & w_{i1} & w_{i2} & \dots & w_{iq} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & w_{n1} & w_{n2} & \dots & w_{nq} \end{pmatrix}.$$

➤ For compactness, we introduce $Y_0 \equiv 1$ e $W_0 \equiv 1$ and they specify the constant baselines of the model.

➤ The parameterization of CUB models requires two distinct links for feeling and uncertainty parameters. As a consequence, information set contained in $\boldsymbol{Y}$ e $\boldsymbol{W}$ may be coincident, completely different or partially completely overlapped.

Link between parameters and subjects' covariates

➤ Any function creating a one-to-one monotone correspondence between $(-\infty, \infty)$ and (0, 1) is legitimate to assess a link between subjects's covariates and parameters.

➤ Classical links as probit, logit and complementary log-log are the common solutions, being the last one more useful for asymmetric relationships.

➤ The preference is for the ***logistic function***:

- It may be given an explicit expression which is computationally ***simple***:

$$z = logit(p) = \log\left(\frac{p}{1-p}\right) \iff p = \frac{1}{1+e^{-z}},$$

  for any $p \in (0,1)\,; z \in (-\infty, \infty)$.
- It has been proved that logit is a ***robust*** link for dealing with ordinal rating.

Definition of a CUB model

➤ Let $R_i \in \{1, 2, \ldots, m\}$ the ordinal response given by the i-th subject characterized by variables $\boldsymbol{t}_i \in \boldsymbol{T}$, for $i = 1, 2, \ldots, n$.

➤ A **CUB model** (= ***C***ombination of a discrete ***U***niform and shifted ***B***inomial random variables) is defined by:

1. A ***stochastic component***:

$$Pr(R_i = j \mid \boldsymbol{t}_i, \boldsymbol{\theta}) = \pi_i \binom{m-1}{j-1} \xi_i^{m-j} (1-\xi_i)^{j-1} + (1-\pi_i) \left(\frac{1}{m}\right),$$

for $j = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, n$.

2. Two ***systematic components***:

$$\begin{cases} logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{y}_i\beta; \\ logit(\xi_i) = \log\left(\frac{\xi_i}{1-\xi_i}\right) = \boldsymbol{w}_i\gamma; \end{cases} \iff \begin{cases} \pi_i = \frac{1}{1+e^{-\boldsymbol{y}_i\beta}}; \\ \xi_i = \frac{1}{1+e^{-\boldsymbol{w}_i\gamma}}; \end{cases}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the parameters to be estimated, and $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$ are the *row vectors* containing the values of the covariates of the i-th subject, suitable to explain π_i and ξ_i, respectively.

Interpretation of a CUB model

➤ Each respondent acts with a ***propensity*** to adhere to a thoughtful and to a completely uncertain choice, measured by (π) and $(1-\pi)$, respectively.

➤ Then, in a rating question/item with positive wording:

- $(1-\xi_i)$ may be interpreted as a *measure of preference* towards the item.
- $(1-\pi_i)$ is a weight of the *uncertainty* included in the responses.

➤ When the item concerns a negative (reverse) wording the interpretation of ξ and $1-\xi$ must be reversed. This happens for worry, disagreement, stress, fear, effort, pain, etc. questions.

Explicit link between parameters and subjects' covariates

➤ A noticeable aspect of CUB models is the **direct link** between subjects' covariates and parameters.

➤ Since $1 - \xi_i$ is a direct *measure of agreement*, feeling, likeness with the item and $1 - \pi_i$ is a direct measure of the *weight of the uncertainty* distribution in the mixture, it is convenient to express those links by means of:

$$\begin{cases} logit(1-\pi_i) &= -\beta_0 - \beta_1\, y_{i1} - \beta_2\, y_{i2} - \ldots - \beta_p\, y_{ip}; \\ logit(1-\xi_i) &= -\gamma_0 - \gamma_1\, w_{i1} - \gamma_2\, w_{i2} - \ldots - \gamma_q\, w_{iq}; \end{cases}$$

➤ These expressions allow for an immediate interpretation of the effects of the selected covariates on the feeling and uncertainty components, respectively.

Notation for a *global* CUB model

- So far, CUB models with covariates have been introduced.
- However, such a specification is not compulsory:
  - Consider $\pi = \sum_i \pi_i/n$ and $\xi = \sum_i \xi_i/n$ as averages of individual parameters and use (π, ξ) to compare propensities to feeling and uncertainty of different items;
  - Consider given values $(\boldsymbol{y}_i, \boldsymbol{w}_i)$, for the i-th subject, and discuss the features of the implied model by letting $\pi_i = \pi$ and $\xi_i = \xi$. Thus, this CUB model is conditional to the values of the covariates $\boldsymbol{y}_i$ and $\boldsymbol{w}_i$.
- Then, a CUB probability distribution (Piccolo, 2003) is defined by:

$$Pr(Y = j) = \pi \underbrace{\left[\binom{m-1}{j-1} (1-\xi)^{j-1} \xi^{m-j} \right]}_{\textit{feeling}} + (1-\pi) \underbrace{\left[\frac{1}{m} \right]}_{\textit{uncertainty}}, \quad j = 1, 2, \ldots, m$$

- Notice the *different role* of $(1-\xi)$ and $(1-\pi)$.

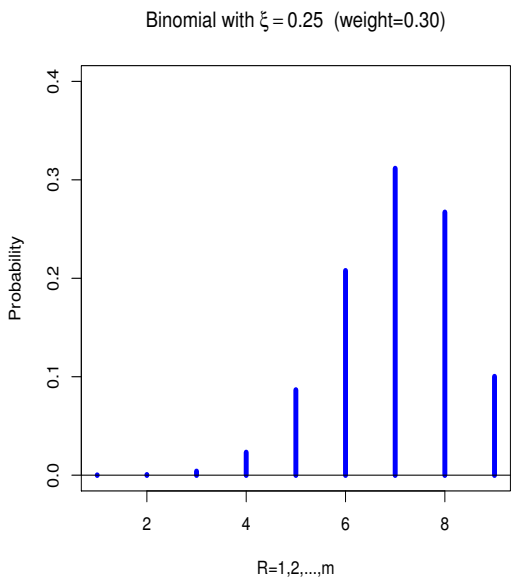
$$Pr(R_i = r|\,\boldsymbol{x}_i, \boldsymbol{w}_i) = \pi_i \underbrace{\left[\binom{m-1}{r-1}(1-\xi_i)^{r-1}\xi_i^{m-r}\right]}_{\textit{feeling distribution}} + (1-\pi_i) \underbrace{\left[\frac{1}{m}\right]}_{\textit{uncertainty distribution}}$$

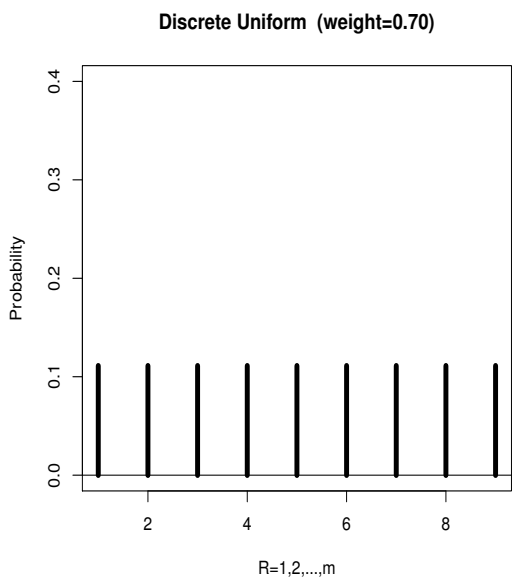
for $r = 1, 2, \ldots, m$, where $\pi_i \in (0, 1]$ and $\xi_i \in [0, 1]$, for $i = 1, 2, \ldots, n$.

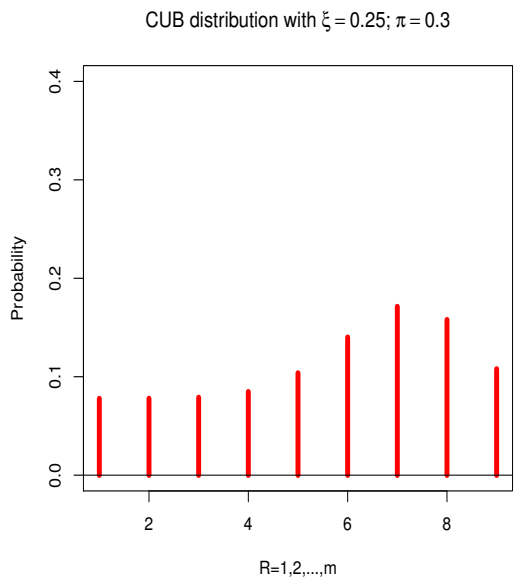
➤ CUB models estimates the *weight of uncertainty* $1 - \pi_i$, **not** the parameters of the probability distribution assumed for the uncertainty.

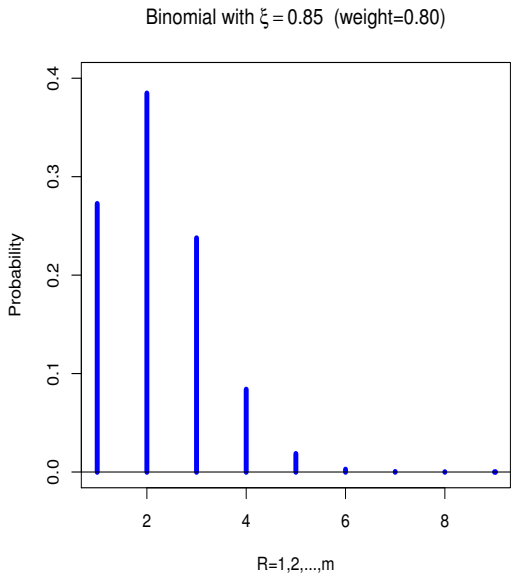
➤ As a consequence, *if covariates are significant*, any CUB model assumes a **non-constant uncertainty** $1 - \pi_i$ ***which modifies with subjects*** $(i = 1, 2, \ldots, n)$ **not with categories** $(r = 1, 2, \ldots, m)$.

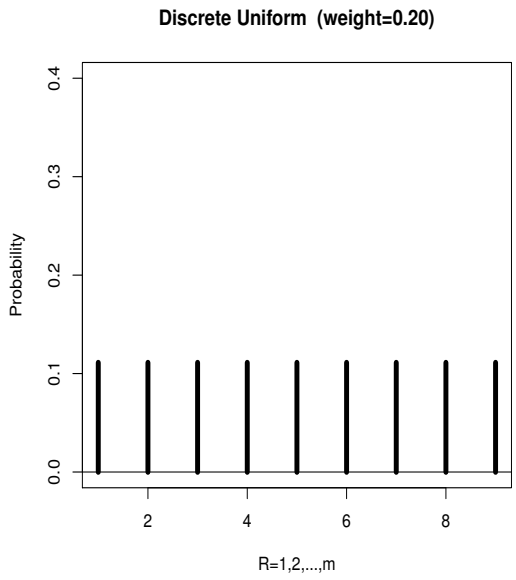
Generation of a CUB model

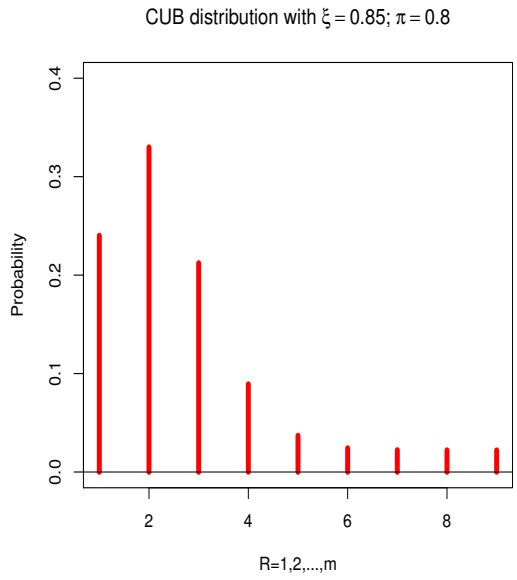












Parameter space of a CUB model

➤ A well defined CUB model requires that both parameters are constrained to the unit interval.

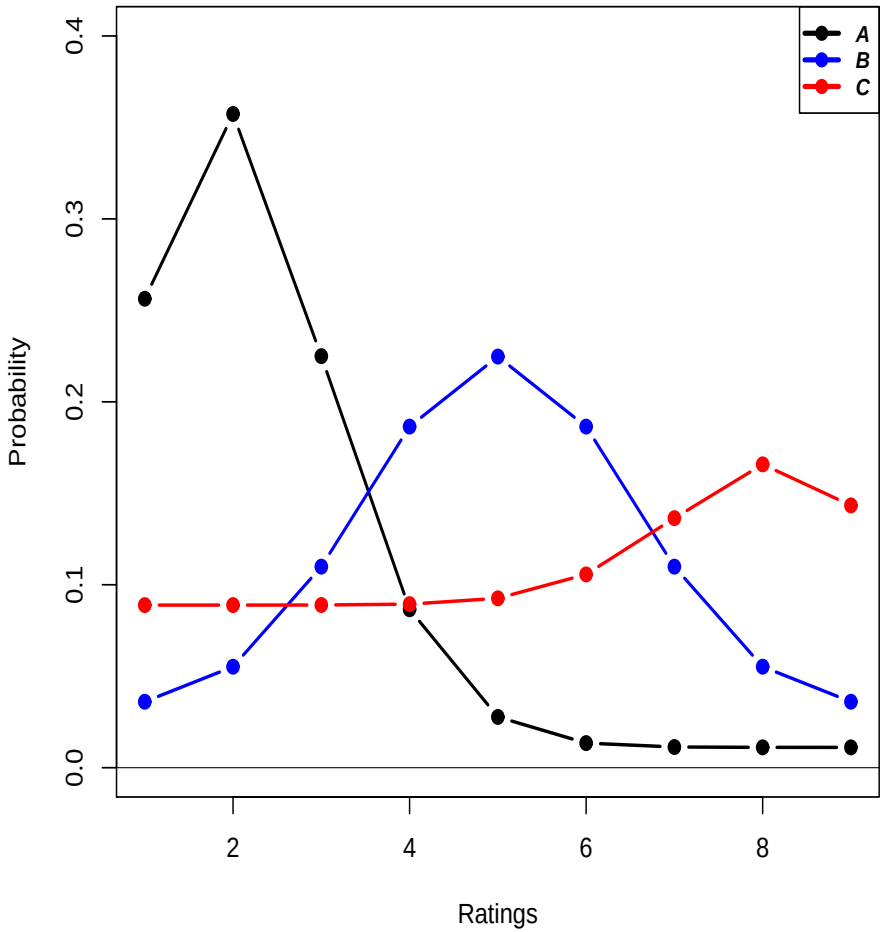
➤ As a consequence the parameter space for $\boldsymbol{\theta} = (\pi, \xi)'$ is defined by:

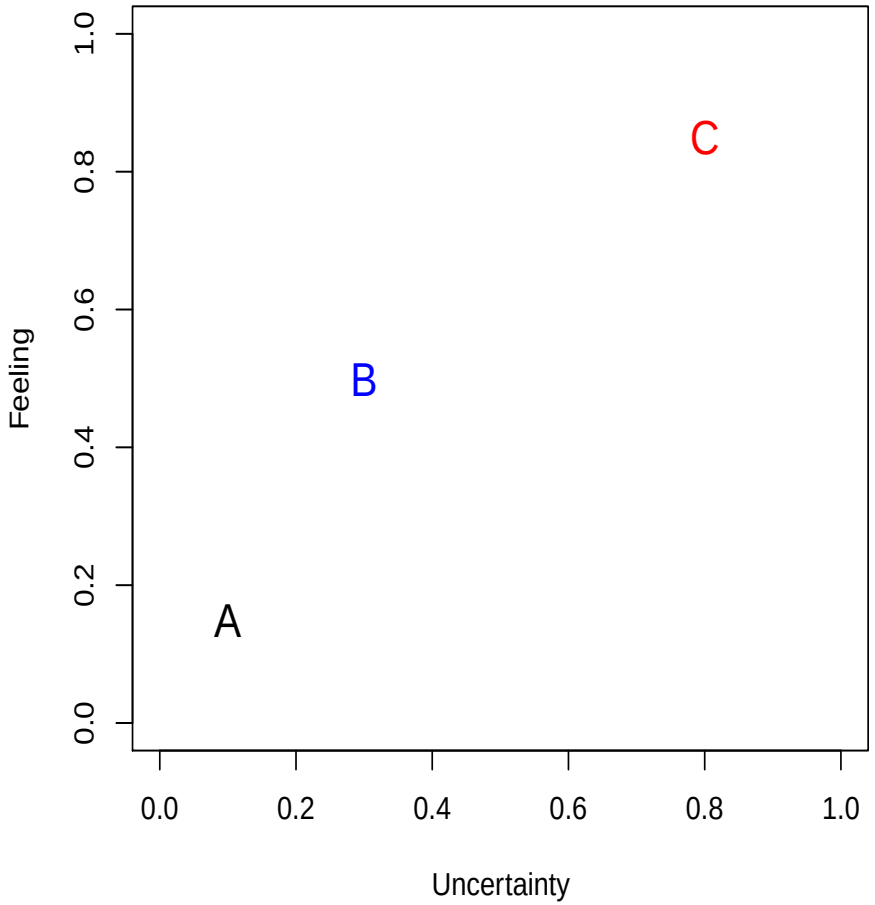
$$\Omega(\boldsymbol{\theta}) = \{(\pi, \xi) : \ 0 < \pi \leq 1, 0 \leq \xi \leq 1\} .$$

➤ Notice that this unit square misses the left side since $\pi = 0$ generates a not-identifiable model.

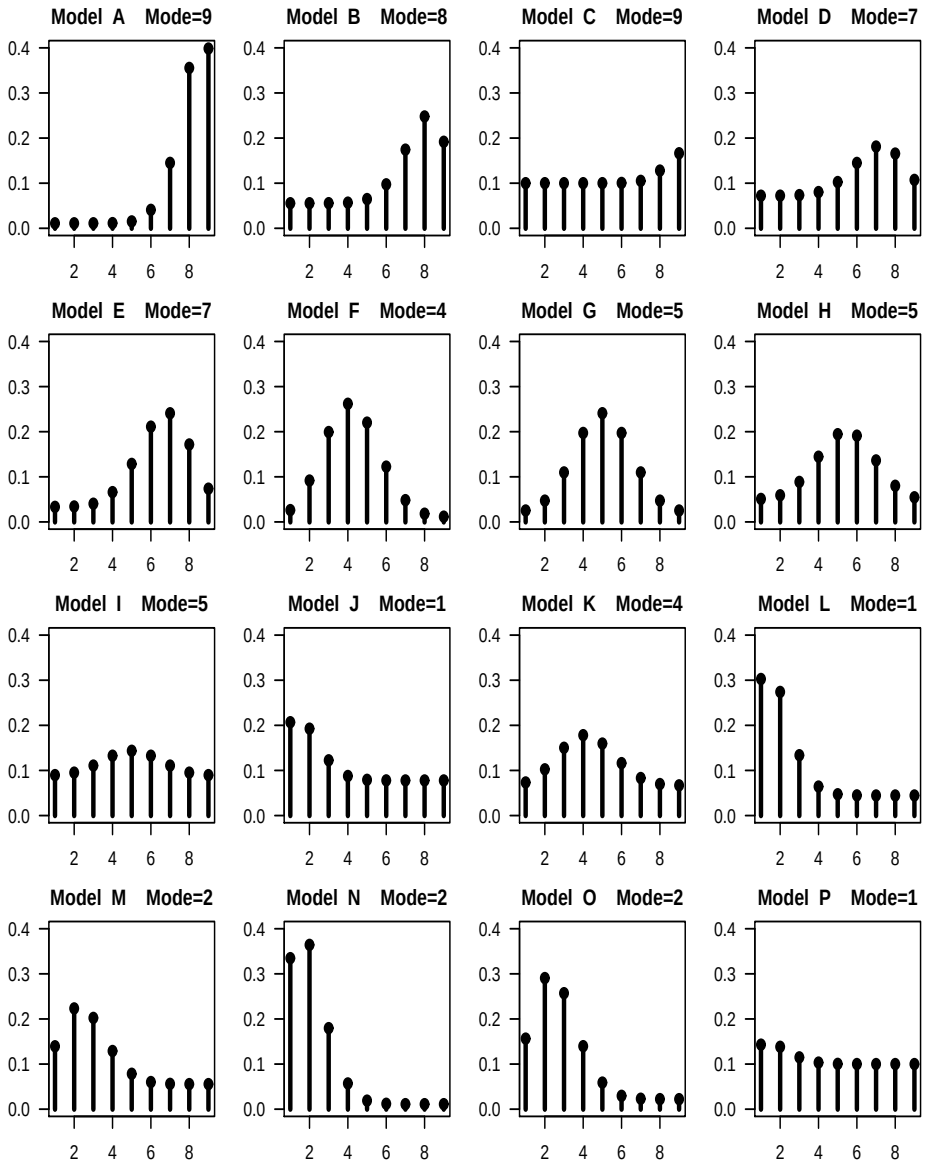
➤ Over $\Omega(\boldsymbol{\theta})$, CUB models are identifiable for any $m > 3$. The case $m = 3$ implies a saturated model.

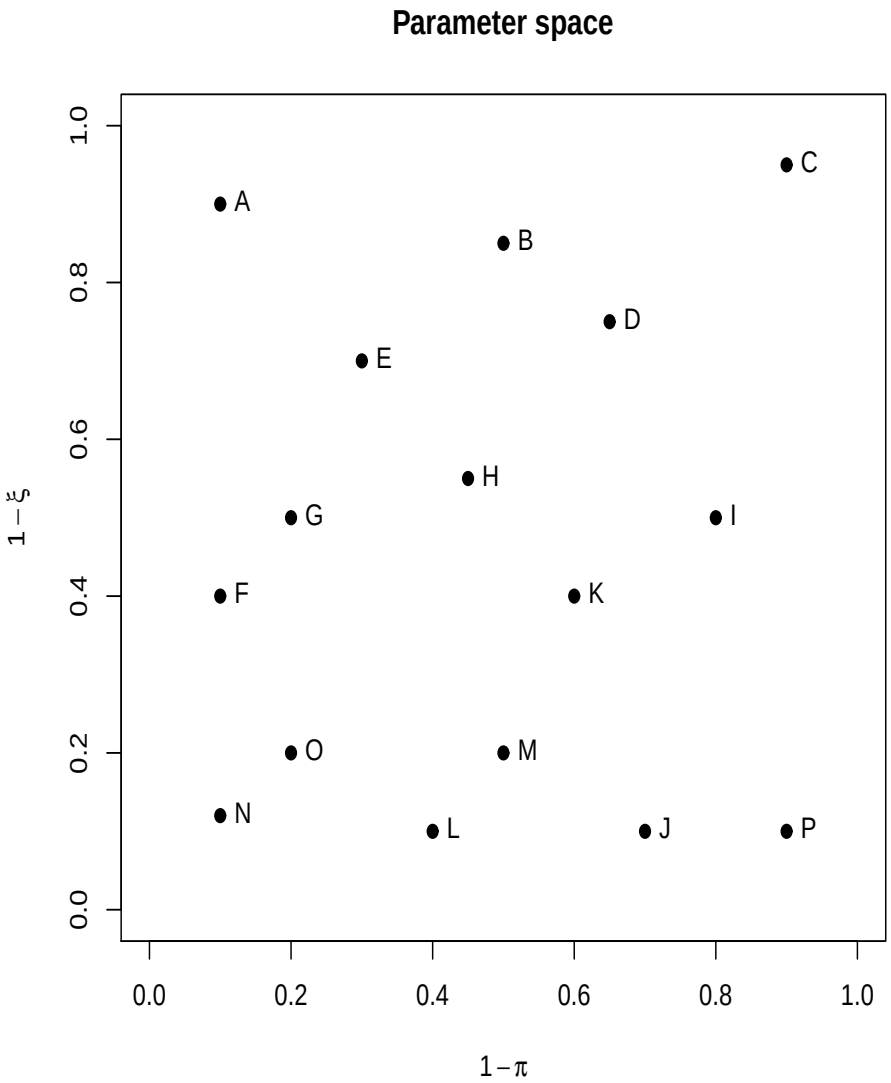
"The" appealing feature of CUB models: visualization



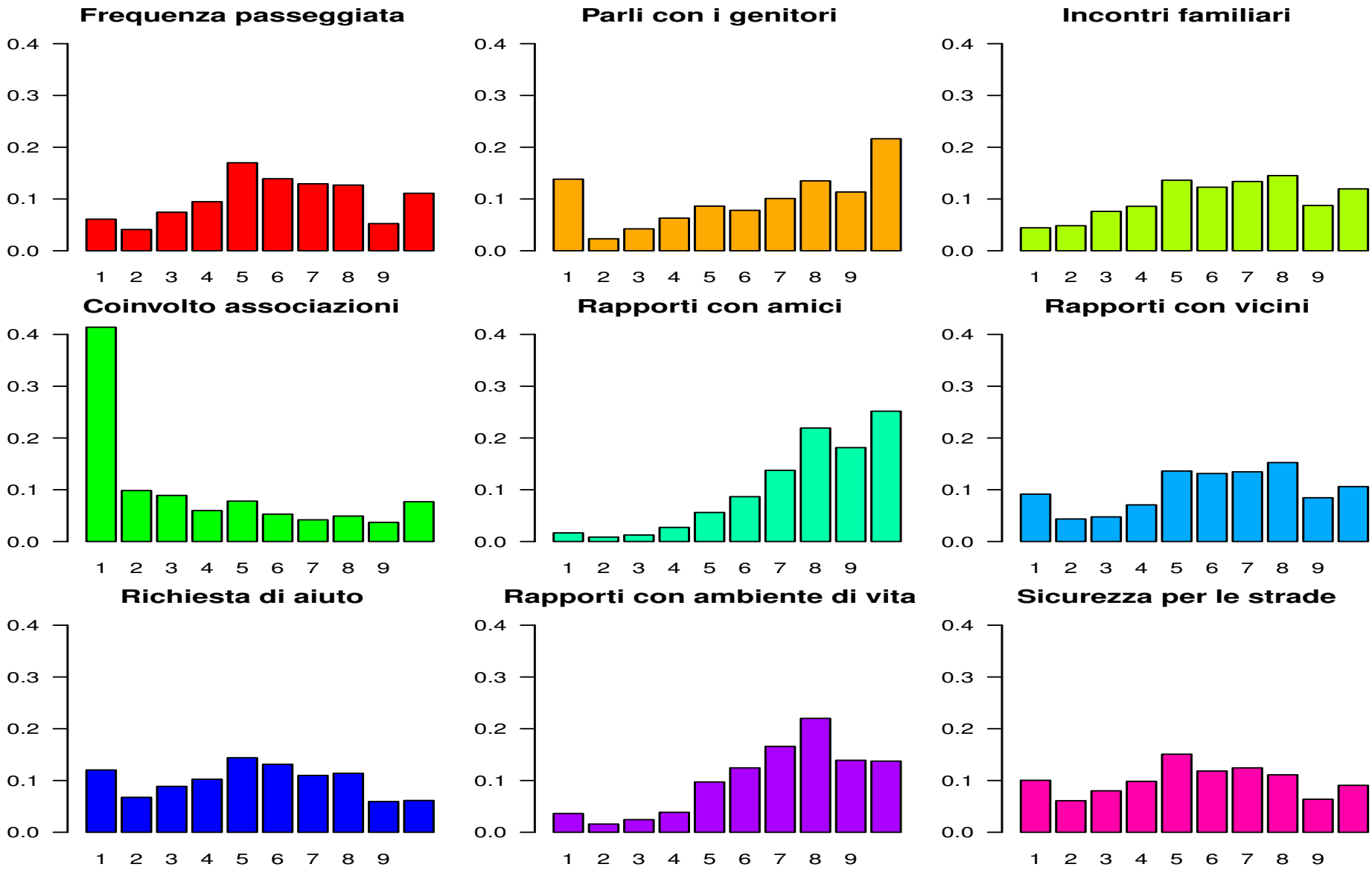


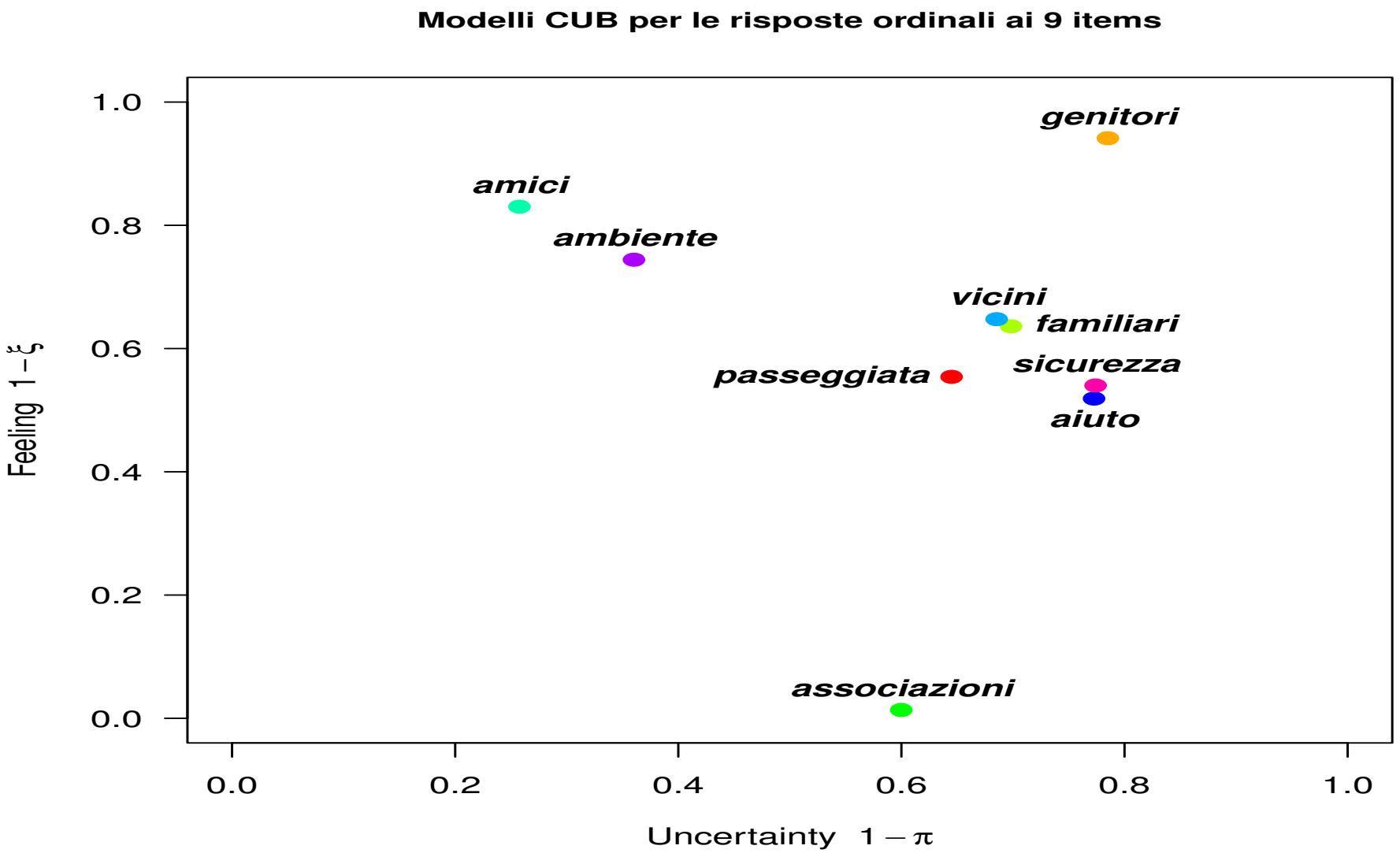
Modelli CUB e spazio parametrico



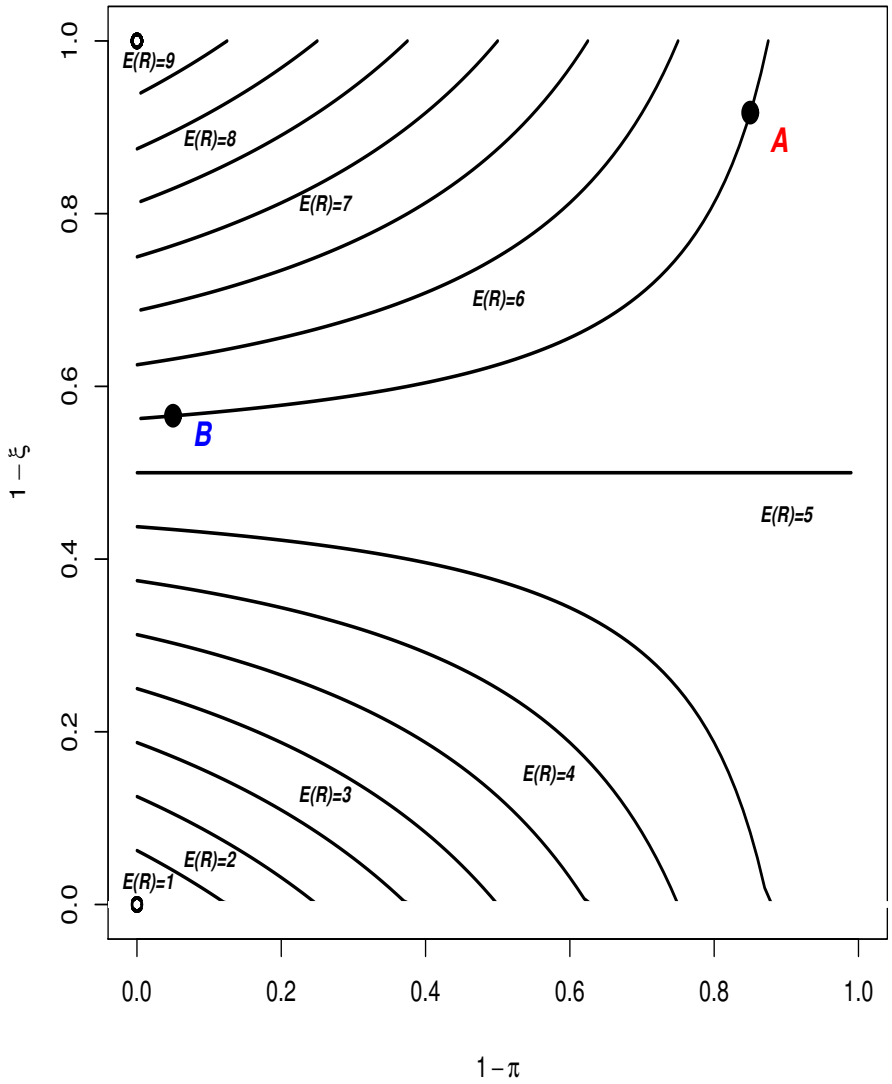


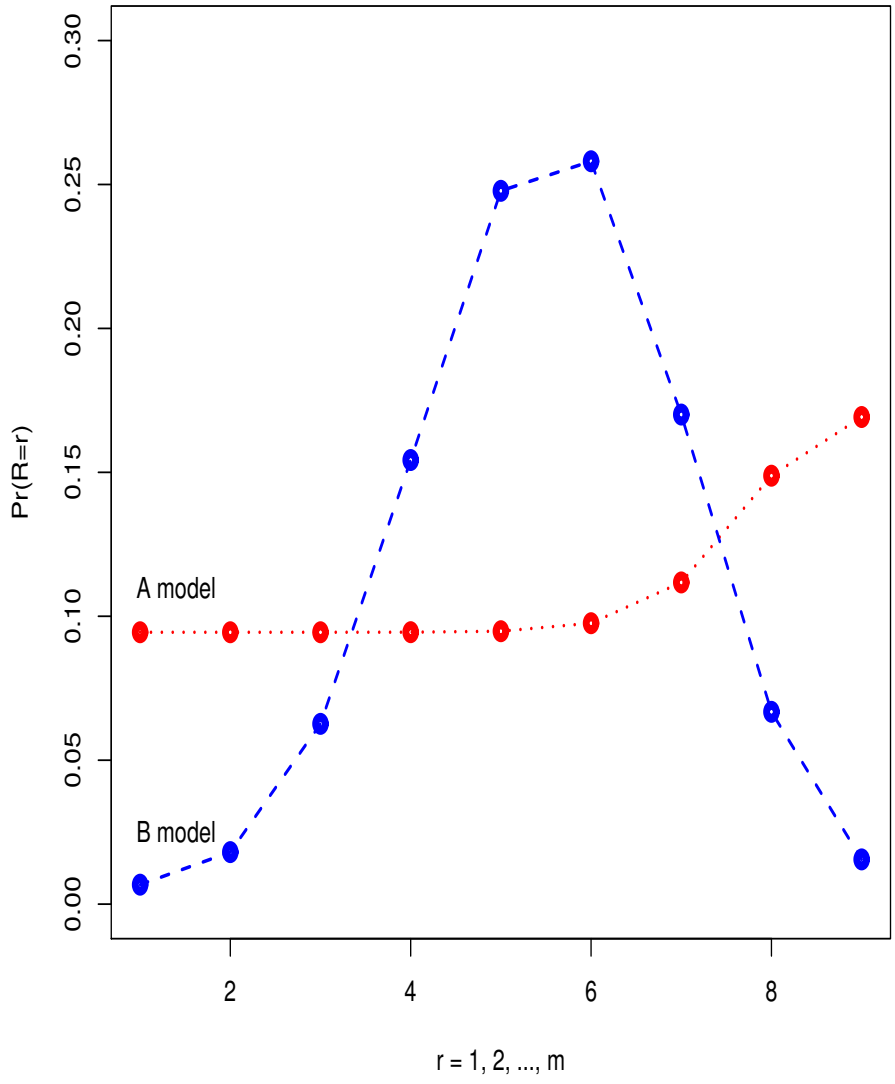
Il dataset RELGOODS: la distribuzione di frequenza di 9 items




Modelli CUB per le risposte ordinali ai 9 items
genitori
amici
ambiente
vicini
familiari
passeggiata
sicurezza
aiuto
associazioni
Feeling 1−ξ
Uncertainty 1 − π
0.0
0.2
0.4
0.6
0.8
1.0

Parametric level curves of expectations of CUB models





Parte III

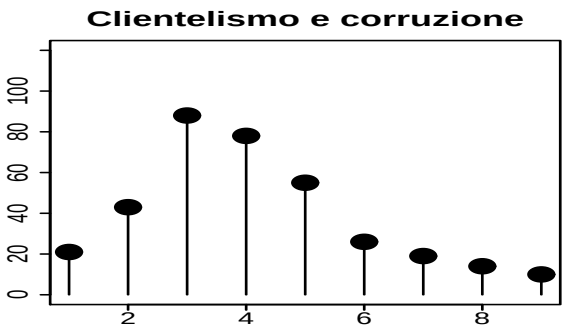
Alcune evidenze empiriche

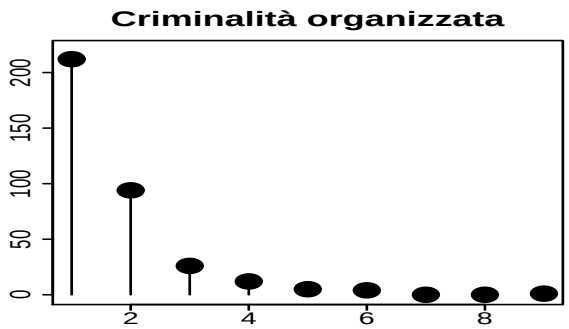
- Esprimi una tua personale graduatoria di importanza (mediante un numero da 1 a 9), mettendo in ordine le seguenti *emergenze metropolitane di Napoli* in modo da attribuire il numero 1 a quella che ritieni la più grave, il numero 2 a quella successiva in ordine di gravità, e così via, sino ad attribuire il numero 9 al problema ritenuto da te meno grave rispetto a quelli elencati.

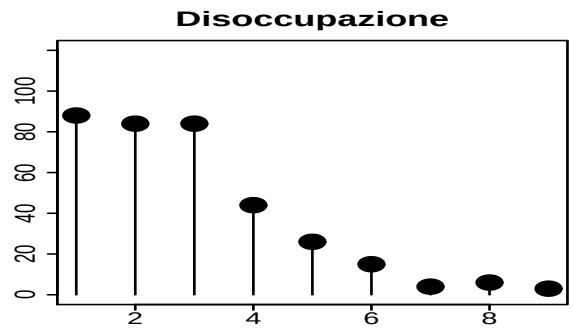
- Per favore, non dare MAI lo stesso numero a due problemi diversi.

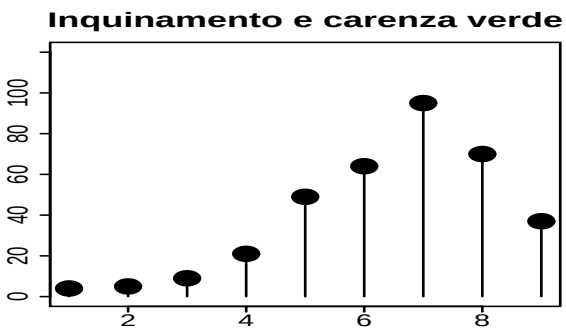
- ☐ *Clientelismo e corruzione*
- ☐ *Criminalità organizzata*
- ☐ *Disoccupazione e mancanza di lavoro*
- ☐ *Inquinamento ambientale e carenza di verde*
- ☐ *Mala-sanità*
- ☐ *Micro-criminalità*
- ☐ *Presenza di immigrati*
- ☐ *Pulizia ed igiene del territorio*
- ☐ *Trasporti-traffico-mobilità*

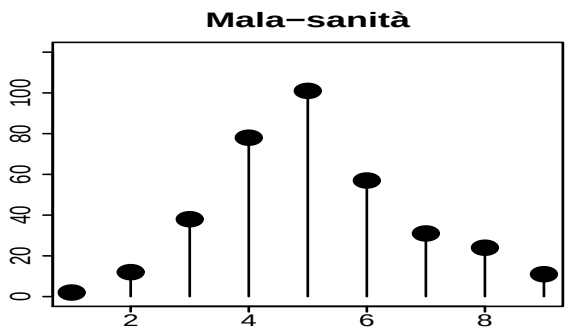
Indagine su Emergenze metropolitane (Napoli, 2004)

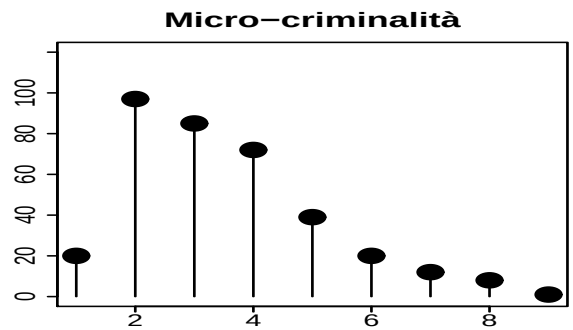


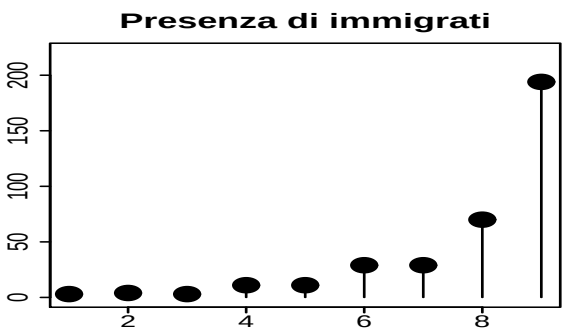


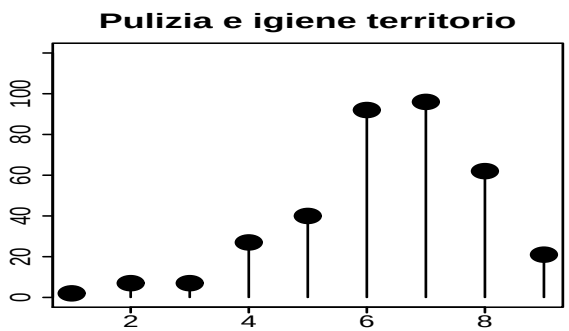


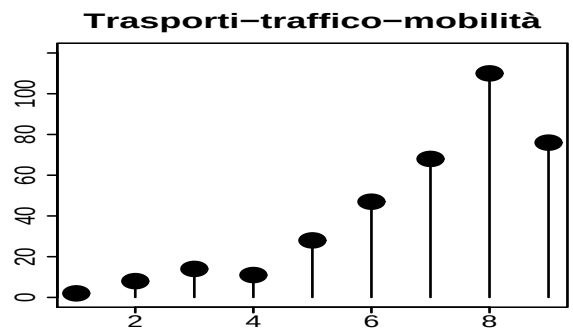


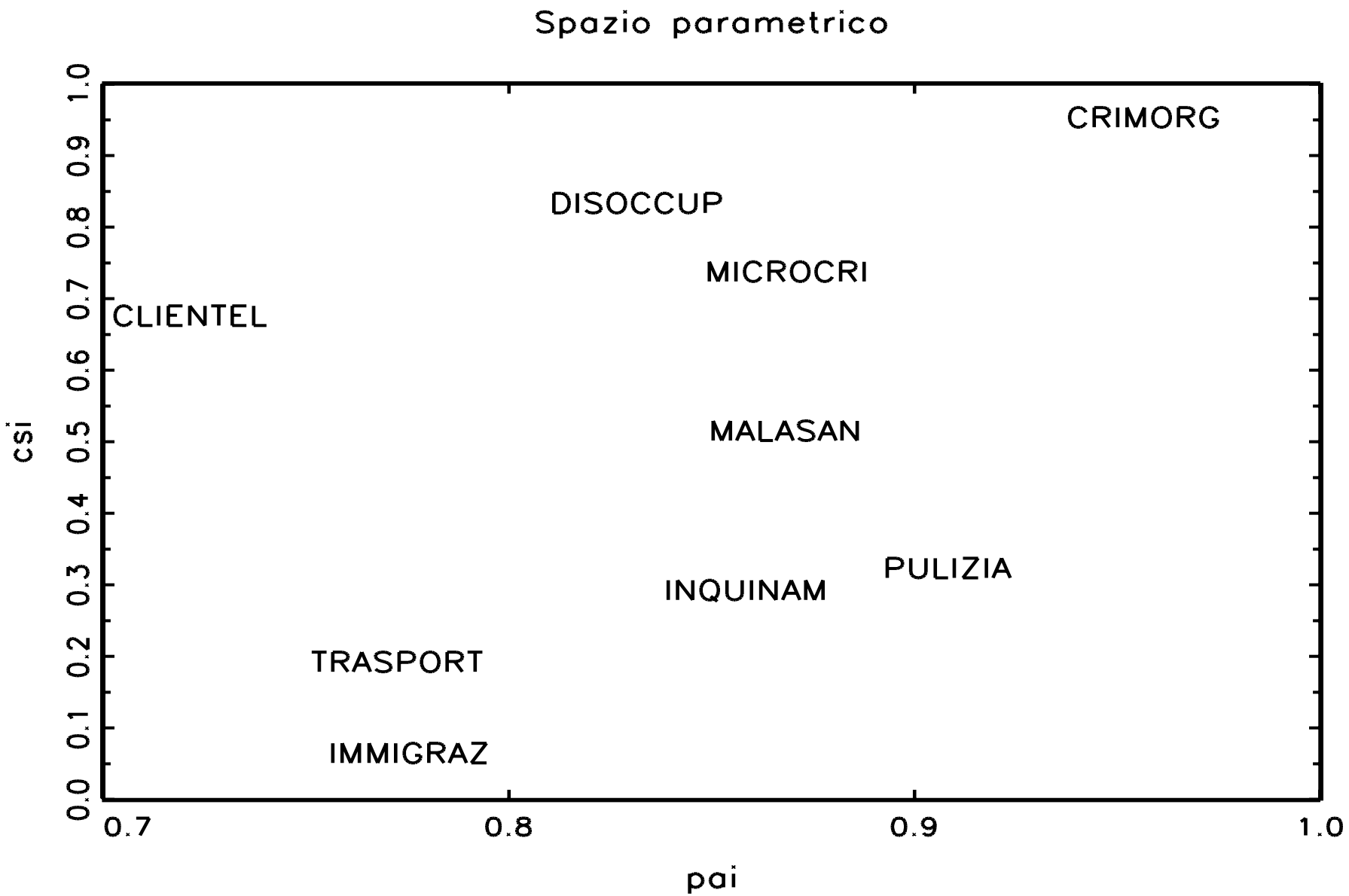




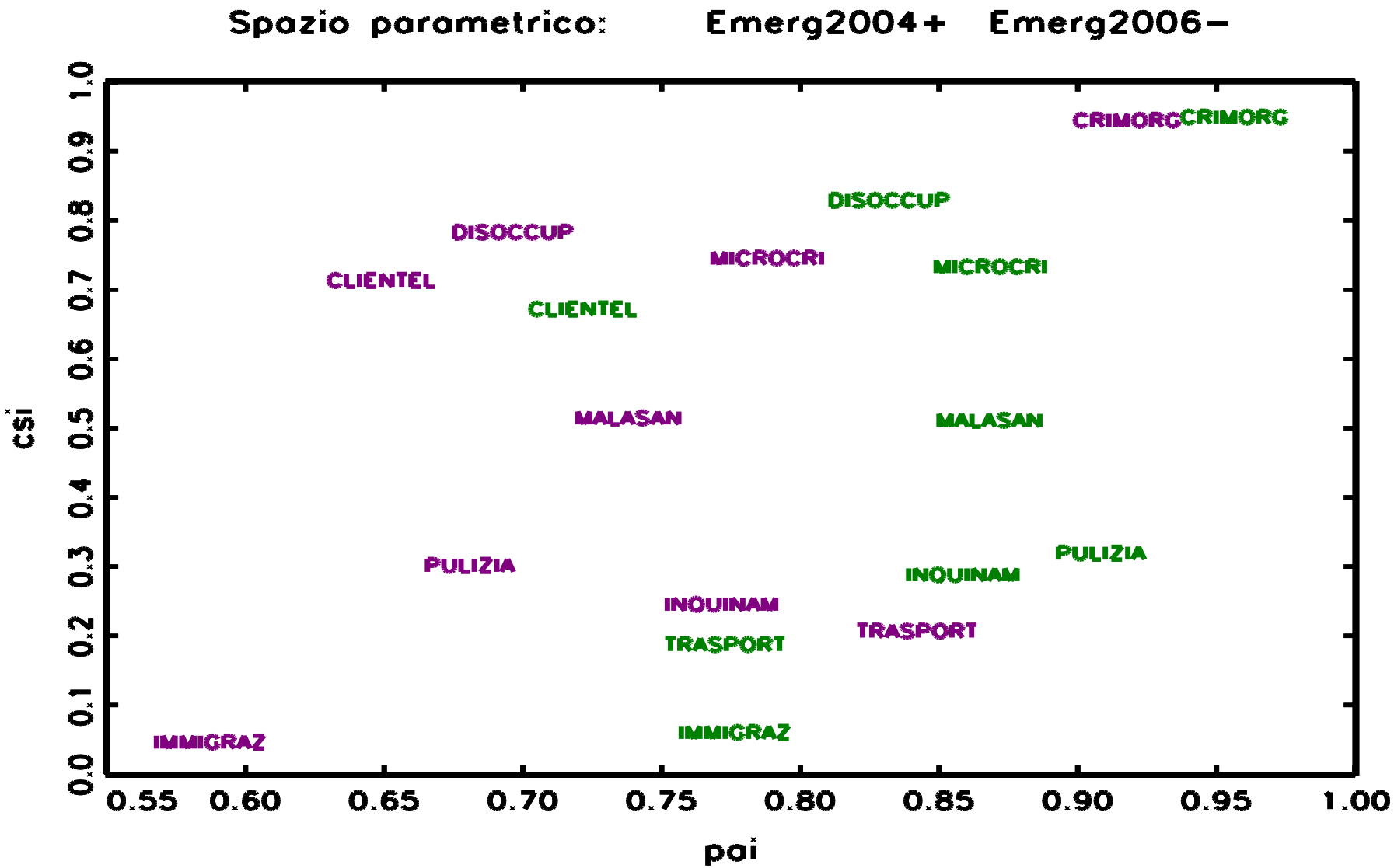




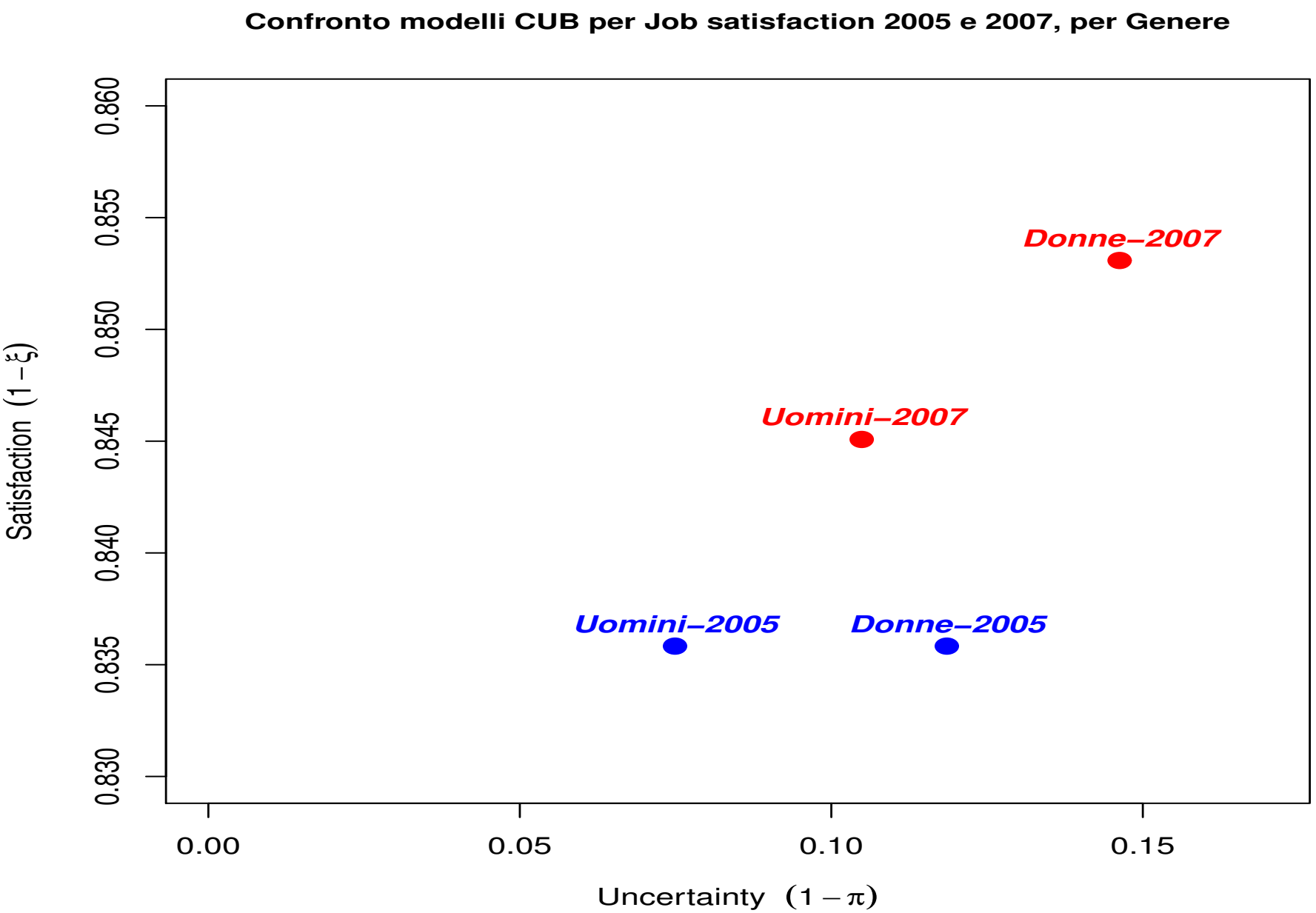



Spazio parametrico
CRIMORG
DISOCCUP
MICROCRI
CLIENTEL
MALASAN
PULIZIA
INQUINAM
TRASPORT
IMMIGRAZ
csi
pai
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.7
0.8
0.9
1.0

Confronto nel tempo fra indagini mediante modelli CUB

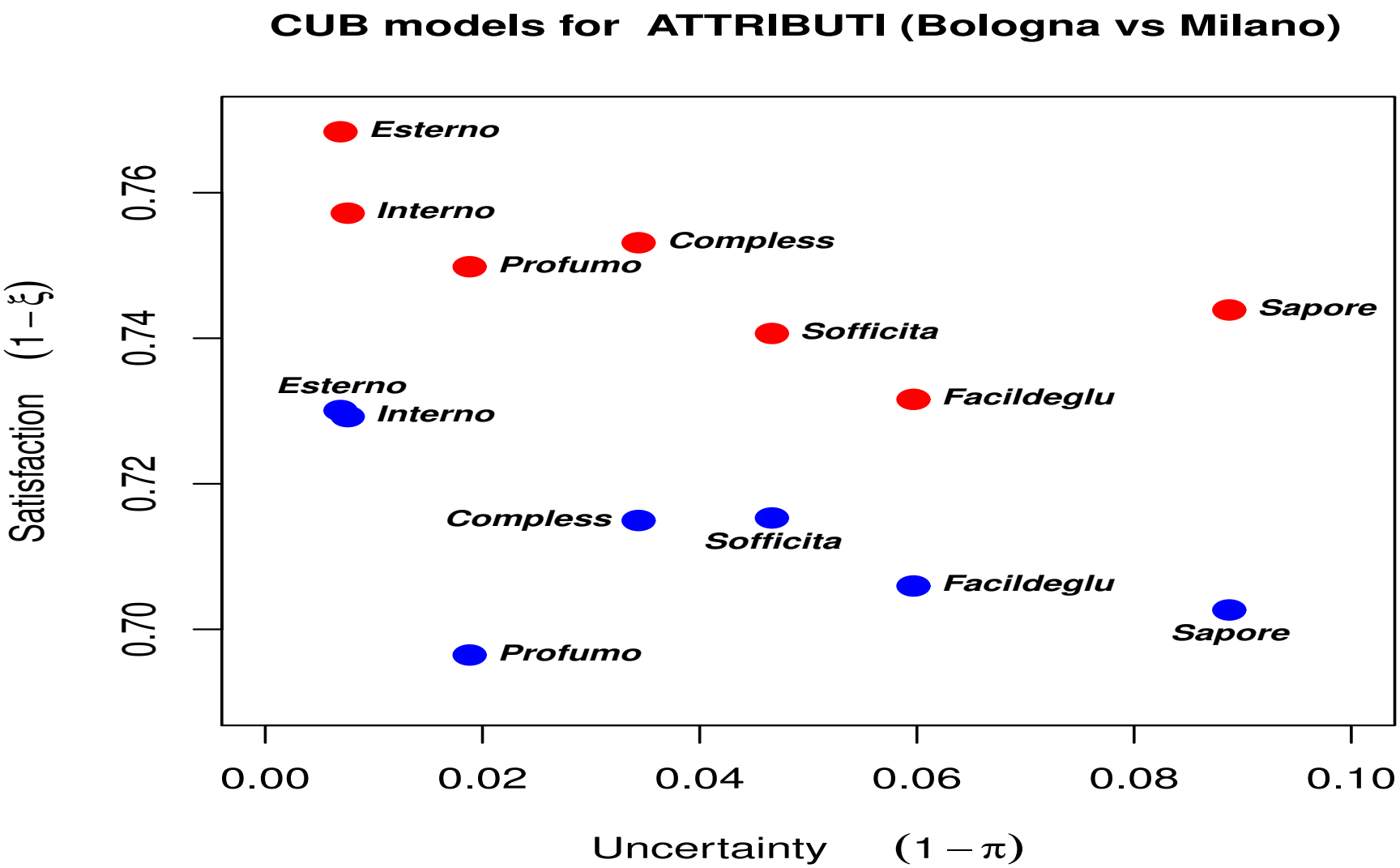


Job satisfaction: effect of gender and time

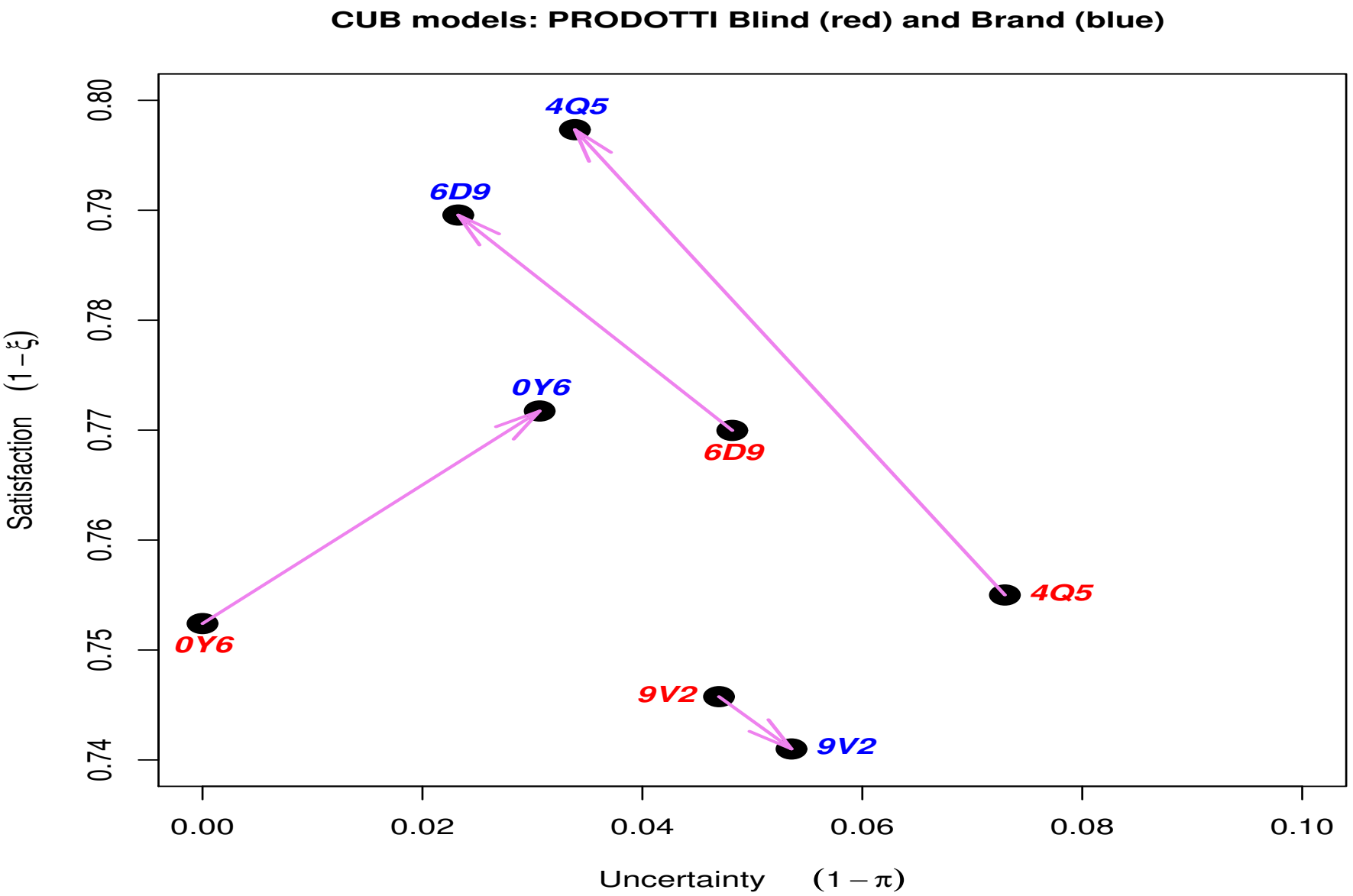


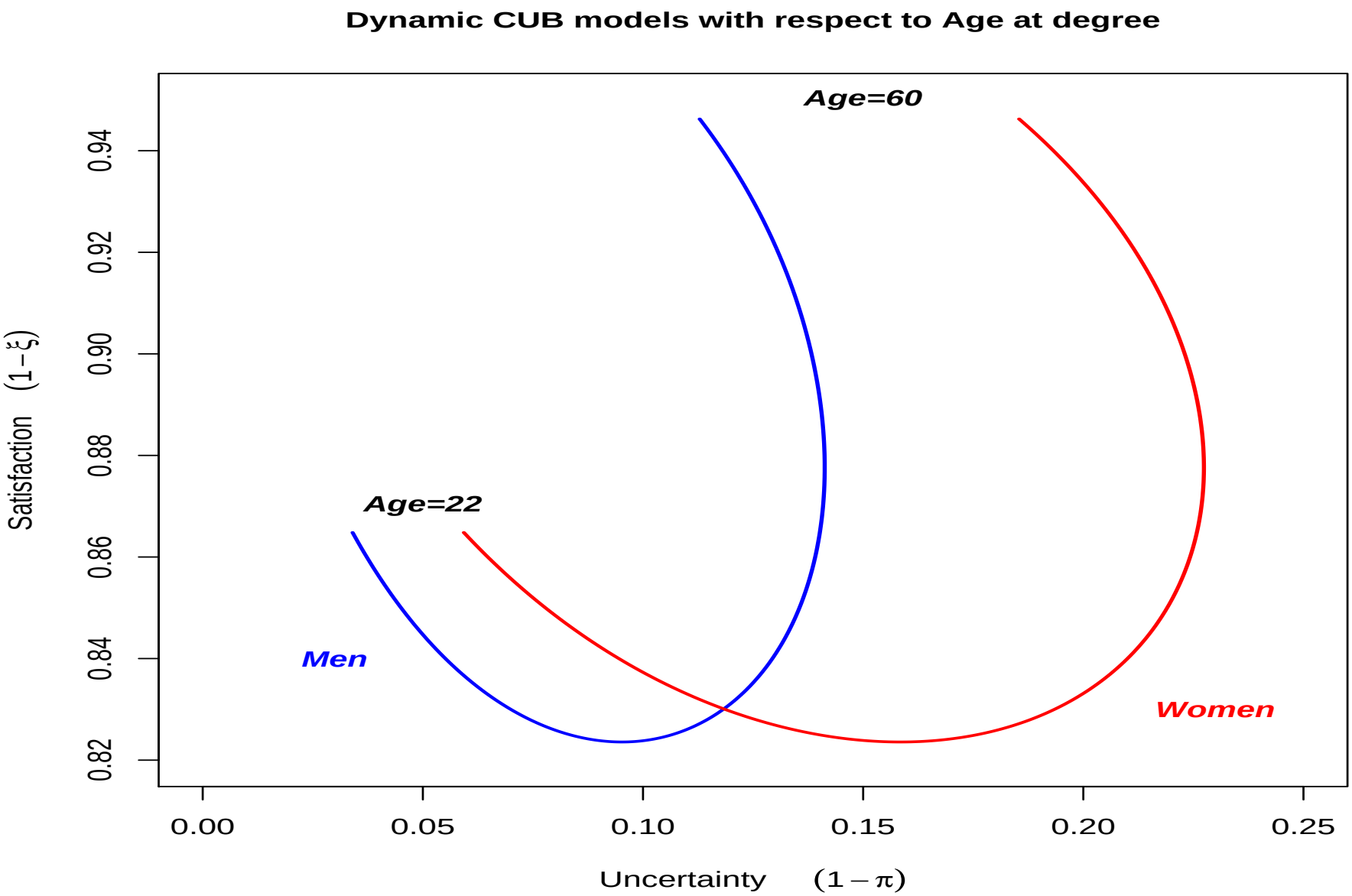
Comparison of consumers' satisfaction

➤ Satisfaction expressed by $n = 110$ consumers living in Bologna and Milan with respect to 10 aspects of a food product.



A blind/brand experiment



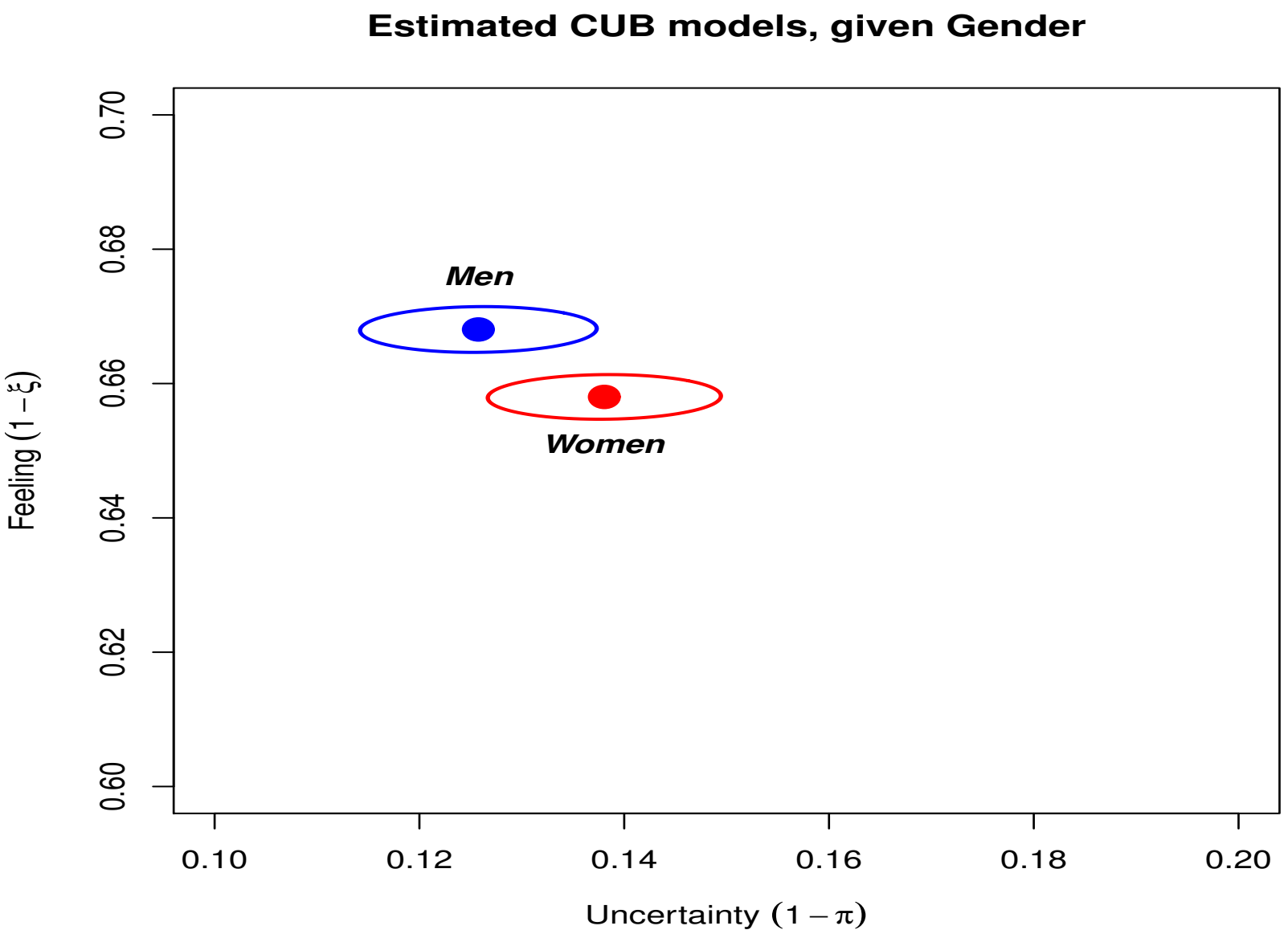

Dynamic CUB models with respect to Age at degree
Age=60
Age=22
Men
Women
Satisfaction (1−ξ)
0.82
0.84
0.86
0.88
0.90
0.92
0.94
0.00
0.05
0.10
0.15
0.20
0.25
Uncertainty (1 − π)

➤ We consider ordinal ratings on **overall life satisfaction** of a sample of Italian respondents collected on a 10-point scale: data are taken from the multiscope survey on daily life run by ISTAT (the National Institute of Statistics of Italy) in 2015.

➤ For the sake of simplicity, the case study here discussed will consider a limited set of dichotomous covariates.

- `Gender` (=1, if woman)
- `Degree` (=1, if he/she got a University degree)
- `Job` (=1, if he/she works)
- `Smoke` (=1, if smoker)
- `South` (=1, if resident in Southern Italy or Islands)
- `Married` (=1, if married/stable relationship)
- `Friends` (=1, if he/she has friends to count on)

➤ The reference data set include $38\,481$ respondents with 52% of women and 13% of graduated; 40% of respondents have a job, only 19% of them smokes whereas 40% lives in South regions and 51% is married; finally, most of them (71%) declared to trust in friends.


Estimated CUB models, given Gender
Men
Women
Feeling $(1-\xi)$
0.60
0.62
0.64
0.66
0.68
0.70
0.10
0.12
0.14
0.16
0.18
0.20
Uncertainty $(1-\pi)$

➤ After a stepwise strategy to select covariates, the best estimated CUB model is described by equations (asymptotic standard errors in parentheses):

$$
\begin{cases}
logit(1-\hat{\pi}_i) = & \underset{(0.056)}{-1.059} - \underset{(0.141)}{0.918}\, Degree_i - \underset{(0.074)}{0.824}\, Job_i + \underset{(0.073)}{0.465}\, Smoke_i \\
& - \underset{(0.061)}{0.240}\, Married_i + \underset{(0.061)}{0.864}\, Friends_i \\
logit(1-\hat{\xi}_i) = & \underset{(0.012)}{0.529} - \underset{(0.009)}{0.037}\, Gender_i + \underset{(0.012)}{0.118}\, Degree_i + \underset{(0.009)}{0.072}\, Job_i \\
& - \underset{(0.011)}{0.093}\, Smoke_i - \underset{(0.009)}{0.205}\, South_i + \underset{(0.009)}{0.116}\, Married_i \\
& + \underset{(0.010)}{0.216}\, Friends_i
\end{cases}
$$

Parte IV

La classe dei modelli CUB

The (growing) class of CUB models

- Variants of ***univariate distributions***:
  - CUB models with both subjects' and objects' covariates
  - *Hierarchical* and random effects CUB models (HCUB and RCUB)
  - *Generalized* (inflated) CUB models (GeCUB)
  - CUSH models
  - *Latent Class* CUB models (LC-CUB)
  - CUB models with "don't know" option (DK-CUB)
  - Dynamic CUB models (CUB-TS)
  - Robust links
- Variants of the ***probability distributions*** of components:
  - CUBE models
  - IHG models
  - CUB models with varying uncertainty (VCUB)
  - CAUB models
  - Non-linear CUB models
  - CUP models
  - GEM models
- ***Joint modelling*** of items:
  - CI-CUB proposal for composite indicators
  - Multi-objects modelling approach
  - Multivariate CUB models via latent variables
  - Multivariate CUB models via copula functions (CO-CUB)
  - Multivariate mixtures (SCUB and CUSCUB)
- ***Statistical usage*** of CUB models:
  - Imputations of missing values
  - CI-CUB for composite indicators
  - Classification and regression trees
  - CUB model with MIMIC structure (CUB-MIMIC)

Current applications of CUB models

- ***Nature of responses***: Ratings, marginal rankings, multivariate ratings
- ***Content of the item***: Preference, mood, agreement, likeness, agreeableness, judgements, perception, cognition, priority, assessment, similarity, changeability, attraction, qualitative distance, fear, discrimination, worry, anxiety, pain, distress, awkwardness, ...
- ***Fields of interest***: Marketing surveys, Sensory analysis, Food packaging, Tourism sustainability, Drug effectiveness, Severity of a disease, Consumer preferences, Service evaluations, Political position on a left-right ideological scale, Words synonymy, Quality of life, Job satisfaction, Stress analysis, Work discrimination, Social media reliability, Advertisement efficacy, Politicians approval, Company climate, Video recommendation, Privacy intrusion, Pharmacokinetics, Team ability, Crowd sourcing, Risk perception, Adolescent abuse substances, Cognitive dissonance, ...
- ***Widespread***: **Italy** (Naples, Brescia, Cosenza, Bergamo, Padova, Vicenza, Ferrara, Palermo, Sassari, Turin, Bari, Rome, Florence, Benevento, Milan, Pavia, Potenza, Catania, . . .), **Germany**, **Switzerland**, **France**, **Netherlands**, **Israel**, **USA**, **Argentina**, **Malaysia**, **South-Korea**, **China**, **Brasil**.

A list of references .....continuously increasing

Piccolo D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, **5**, 85–104.

D'Elia A., Piccolo D. (2005). A mixture model for preference data analysis. *Computational Statistics & Data Analysis*, **49**, 917–934.

.....*about 200 publications*.....

Piccolo D., Simone R. (2019). The class of CUB models: statistical foundations, inferential issues and empirical evidence. *Statistical Methods & Applications*, **28**, 389–435; with discussion (pp.437–475) and rejoinder (pp.477–493). (**(with hundreds of references)**

.....*recent publications* .....

Simone R. (2023). Uncertainty Diagnostics of Binomial Regression Trees for Ordered Rating Data. *Journal of Classification*, 40: 79–103, Springer.

Venson A.H., Jacinto P.A. and Sbicca, A. (2023). Cognitive Dissonance in the Self-assessed Health in Brazil: A CUB Model Analysis Using 2013 National Health Survey Data. *Integrative Psychological and Behavioral Science*, `https://doi.org/10.1007/s12124-023-09768-x`. **19 May 2023**

Sometimes. . . I cannot read a paper on CUB models

Journal of the Korean Data & Information Science Society
2015, **26**(2), 313–322

http://dx.doi.org/10.7465/jkdi.2015.26.2.313
한국데이터정보과학회지

이항-퇴화 혼합분포의 최우추정법†

황선영[1] · 손승혜[2] · 오창혁[3]

[123]영남대학교 통계학과

접수 2015년 1월 17일, 수정 2015년 2월 9일, 게재확정 2015년 3월 18일

요 약

본 연구에서는 하나의 균일분포 또는 퇴화분포와 두 개의 이항분포의 혼합분포 모형에 대하여 최우추정법을 소개하며, 제시된 모형에 대하여 시뮬레이션을 통해 최우추정량의 성질을 밝히며, 실험을 통해 얻은 강의 평가 자료에 대하여 퇴화분포를 가지는 혼합분포에 대하여 적용하여 보았다. 특히 퇴화분포는 한국의 문화 특성상 가운데 값을 선호하는 현상을 모형화하는데 유용하게 사용될 수 있음을 보였다.

주요용어: 우도함수, 이산균일분포, 최우추정법, 퇴화분포, 혼합분포.

1. 서론

혼합분포는 분포의 이질성을 나타내는 유용한 방법이며 자료가 얻어지는 모집단이 두 개 이상의 이질적 집단으로 구성되어 있는 경우에 여러 분야에서 폭넓게 사용되고 있다 (McLachlan과 Peel, 2001). 혼합분포는 몇 개의 성분분포로 이루어지며, 성분분포는 연속형 또는 이산형이 될 수 있다. 성분분포가 이산형인 경우는 이항분포, 포아송분포, 이산균일분포 등이 흔히 사용된다. 그중에서 이항분포를 성분으로 가지는 혼합분포의 이론과 적용에 대한 많은 연구가 이루어져 왔다 (Blischke, 1964; Johnson 등, 2005; Liu 등, 2006). 한편, Oh (2014)는 이동 이항분포의 혼합분포의 최우추정치를 찾는 방법을 제안하였고, Bonnini 등 (2012)은 이항분포와 이산균일분포의 혼합분포에서, Domenico (2003)는 이산균일분포와 이동 이항분포의 혼합분포에서, Lee와 Oh (2006)와 Oh (2006)는 이동 포아송분포의 혼합분포

➤ Approaching the ***Big Data era***, a simple and effective tool to capture the main features of millions of respondents in a (almost) continuous time may be a worthwhile approach to model rating data.

➤ Effective ***statistical software*** is available:

- R packages: CRAN → **`CUB`** and **`FastCUB`**; GitHub → **`cubm`**
- STATA module: ***cub***
- GRETL programme: CUB
- GAUSS software: series of functions

Parte V

Una procedura operativa

Procedura operativa "passo-passo" *(step-by-step)*

1. Organizzare i dati in un data frame che include, per ciascun soggetto, le risposte ordinali e le covariate del soggetto
2. Esplicitare con un diagramma a barre la distribuzione delle variabile ordinali
3. Costruire un modello CUB (senza covariate) per ciascuna risposta ordinale mediante i software disponibili
4. Rappresentare (e interpretare) i modelli CUB stimati sullo spazio parametrico
5. Verificare se esistono covariate dei soggetti significative
6. Interpretare l'effetto di tali covariate mediante la rappresentazione sullo spazio parametrico
7. Altre eventuali utilizzazioni . . . . . .

Questionario RELGOODS: beni relazionali

➥ Nelle risposte alle domande seguenti, tieni presente che [1] significa “mai, per niente, molto raramente, pochissimo” e [10] significa “sempre, molto spesso, moltissimo”.

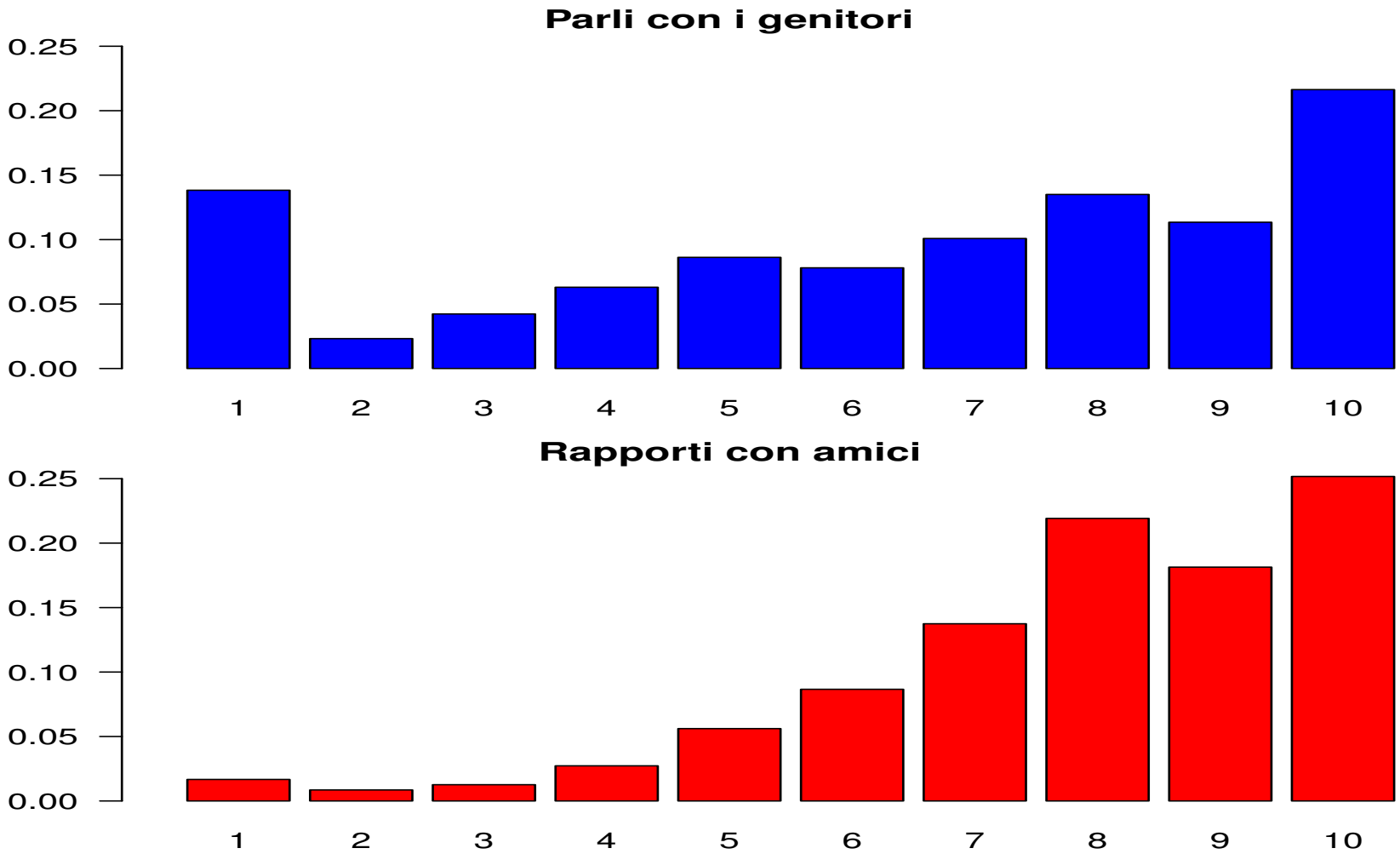
• Con quale frequenza fai una *passeggiata* all’aria aperta?	1	2	3	4	5	6	7	8	9	10
• Quanto spesso parli con almeno uno dei tuoi *genitori*?	1	2	3	4	5	6	7	8	9	10
• Quanto spesso incontri altri *familiari* (nonni, zii, cugini, nipoti, etc.)?	1	2	3	4	5	6	7	8	9	10
• Con quale frequenza/intensità sei *coinvolto in associazioni* culturali o religiose, gruppi di volontariato, partiti, sindacato, etc.	1	2	3	4	5	6	7	8	9	10
• Definiresti positivi i tuoi *rapporti con gli amici*?	1	2	3	4	5	6	7	8	9	10
• Definiresti positivi i tuoi *rapporti con i vicini*?	1	2	3	4	5	6	7	8	9	10
• Se hai *bisogno di aiuto*, lo chiedi facilmente agli altri?	1	2	3	4	5	6	7	8	9	10
• Definiresti positivi i tuoi *rapporti con l’ambiente* (di studio, di lavoro, di tempo libero) nel quale trascorri gran parte del tuo tempo?	1	2	3	4	5	6	7	8	9	10
• Ti senti *sicuro per le strade* del paese in cui adesso vivi?	1	2	3	4	5	6	7	8	9	10
• Tu e la tua famiglia *arrivate facilmente a fine mese* dal punto di vista economico?	1	2	3	4	5	6	7	8	9	10

Procedura operativa: **step 1**

➤ Per semplicità, dal dataset **RELGOODS** si scelgono due risposte ordinali (*Rapporti con 2.* **GENITORI**; *5.* **AMICI**) e due covariate (*Genere; Età*)

iden	genitori	amici	genere	età
1	8	8	1	42.667
2	3	5	0	38.167
3	9	7	1	50.000
4	1	5	1	56.667
5	9	10	0	22.750
6	6	7	0	25.000
7	5	10	0	24.417
8	10	8	1	23.250
9	9	10	0	24.333
10	10	10	0	24.333
...	...	...	...	...

➤ Si disegnano i diagramma a barre per le distribuzioni delle due variabile ordinali (*Rapporti con 2. GENITORI; 5. AMICI*)



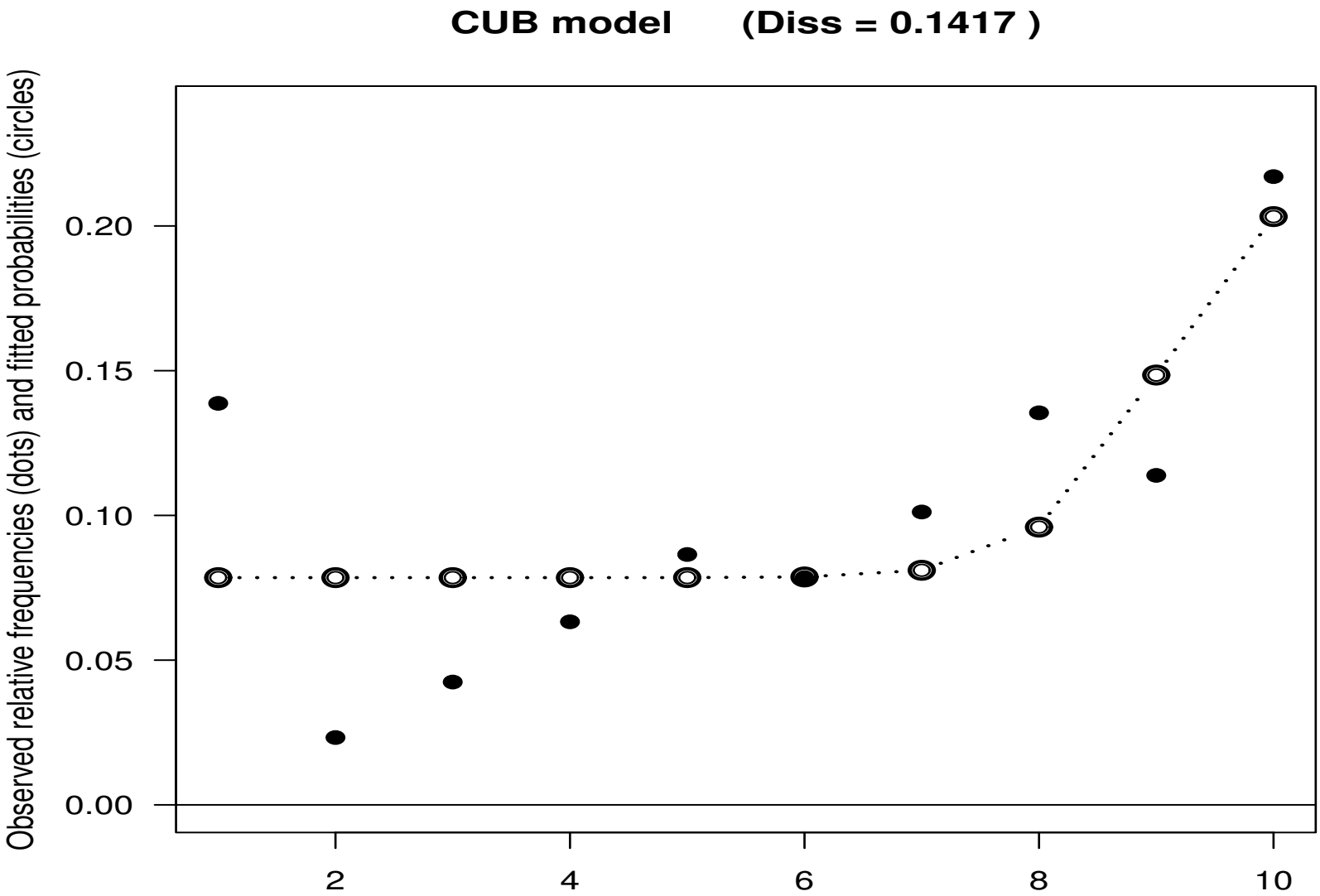
Procedura operativa: **step 3**

- Si stima un modello CUB per ciascuna risposta ordinale mediante il package **CUB** disponibile in R
- Per esempio, per la risposta ordinale GENITORI i comandi e l'output sono:

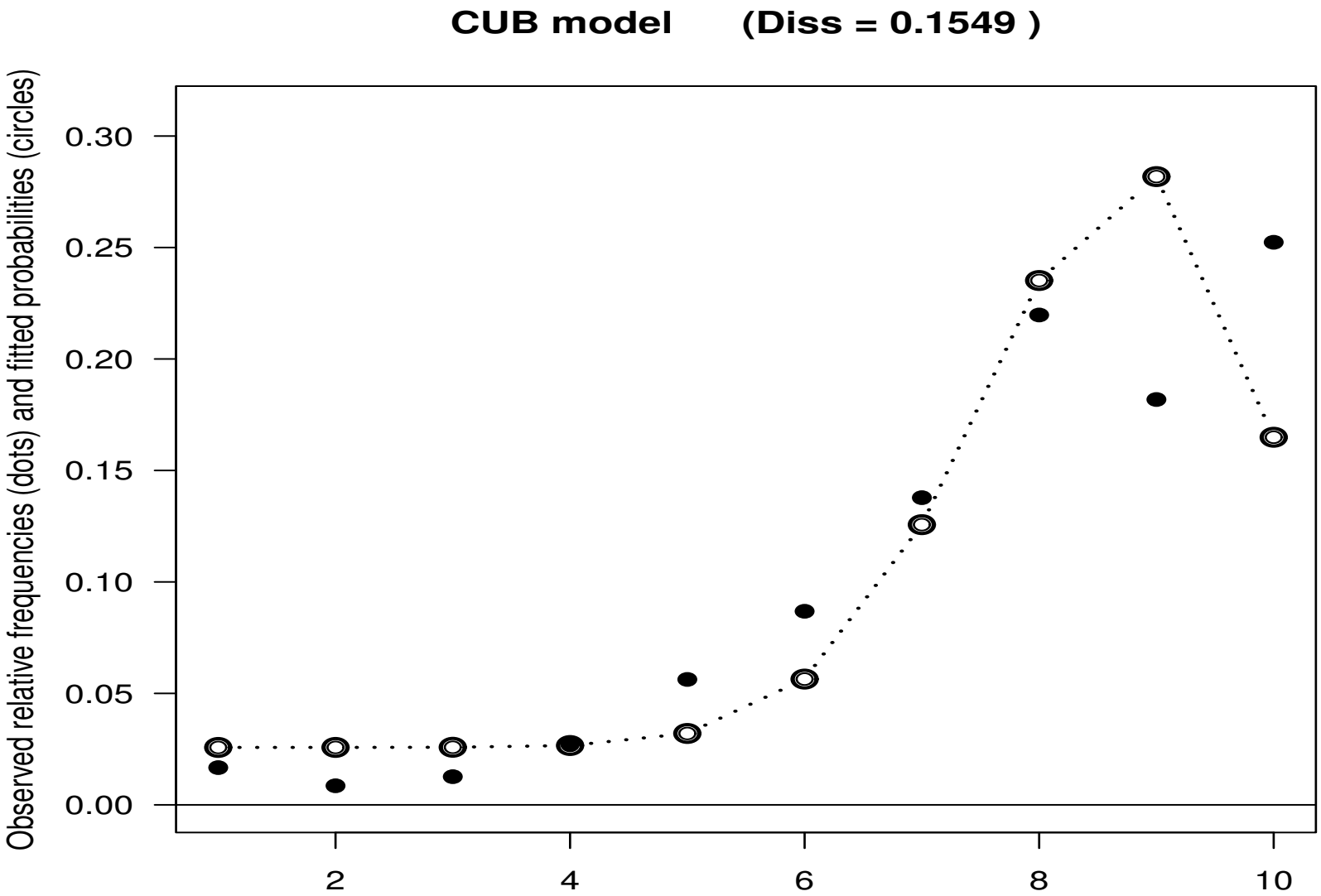
```
> modgen=GEM(Formula(genitori~0|0|0),family="cub")
> summary(modgen)

=========================================================
=====>>> CUB  model    <<<=====   ML-estimates via E-M algorithm
=========================================================
 m= 10  Sample size: n= 2451  Iterations= 21  Maxiter= 500
=========================================================
Uncertainty
   Estimates     StdErr     Wald
pai   0.21497 0.01782975 12.05681
=========================================================
Feeling
   Estimates     StdErr    Wald
csi   0.05869 0.01041633 5.63442
=========================================================
```

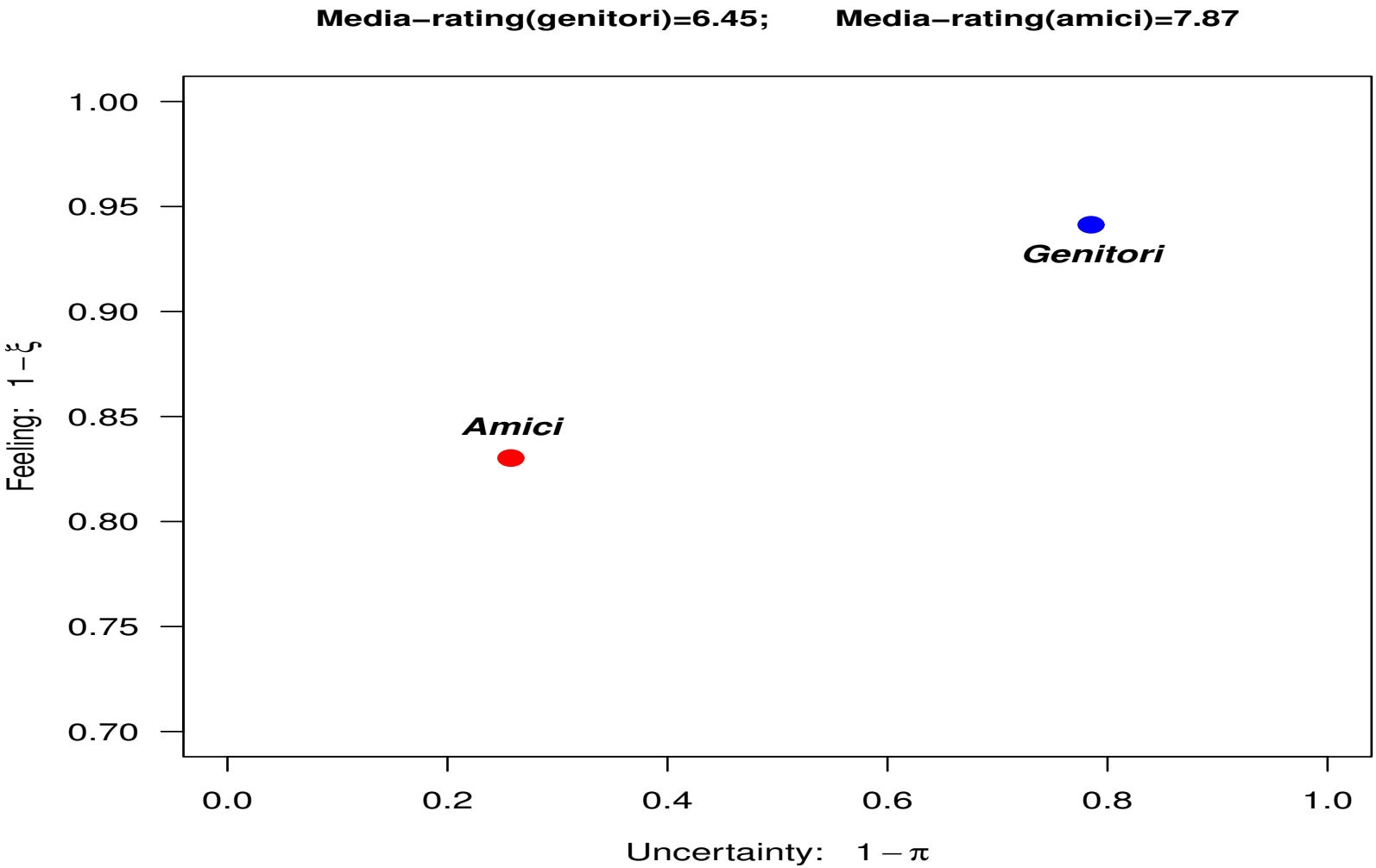
➤ Per **`GENITORI`**, si confronta il modello CUB stimato con la distribuzione osservata:



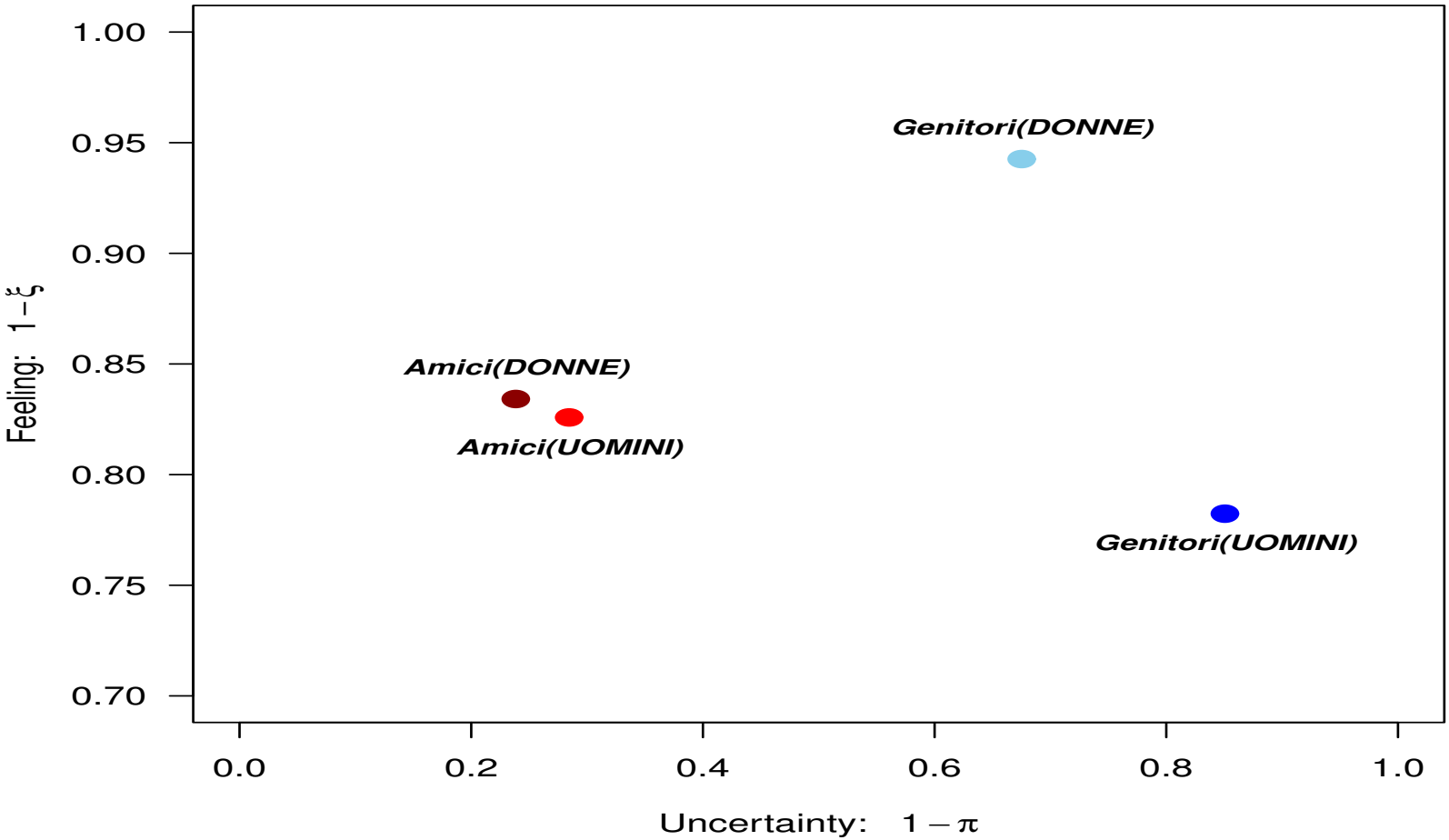
➤ Si procede analogamente anche per la variabile **`AMICI`**, ottenendo un modello CUB stimato che si confronta con la distribuzione osservata:



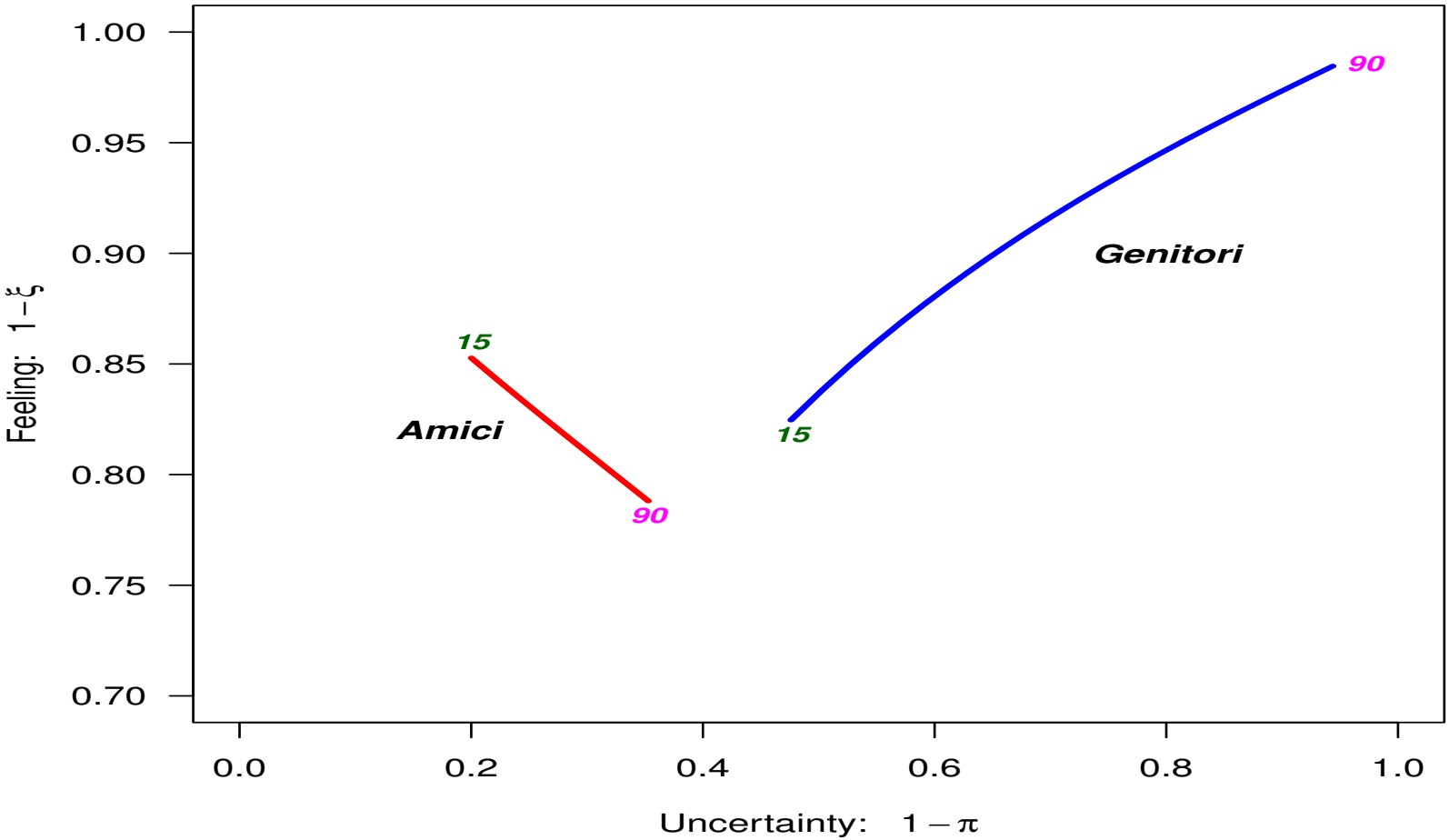
➤ Si rappresentano ed interpretano i modelli CUB stimati sullo spazio parametrico (*per comodità, si evidenzia solo una porzione*):



➤ Si controlla se la variabile **Genere** è significativa per *uncertainty* e/o *feeling* per le due risposte ordinali GENITORI e AMICI, rispettivamente.



➤ Si controlla se la variabile **Età** è significativa per *uncertainty* e/o *feeling* per le due risposte ordinali GENITORI e AMICI, rispettivamente.



Ulteriori utilizzazioni dei modelli CUB

- Confronti rispetto al **territorio** (regioni, . . .)
- Confronti rispetto al **tempo** (anni, . . .)
- Confronti rispetto al **contesto** (pre- e post-Covid, . . .)
- Costruzione di **graduatorie** a fini decisionali
- Costruzione di **indicatori composti**
- Sostituzione (*imputazione*) per **dati mancanti**
- Confronto con **questionari con differente numero di modalità**
- **Semplificazione** dei questionari
- . . . . . .

Una domanda non molto frequente. . . ma importante

- Per chi deve confrontare l'evoluzione delle preferenze di un prodotto per molto tempo può avvenire che le modalità dell'intervista ai consumatori cambi.
- Sorge, allora, la domanda:

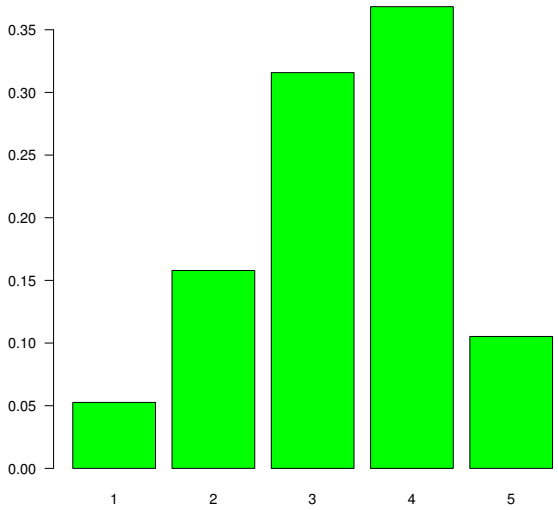
"Come faccio a confrontare le mie precedenti esperienze di valutazione e preferenze, raccolte su scale a 5 oppure a 7 modalità, con quelle più recenti, ottenute con scale a 9 modalità?"

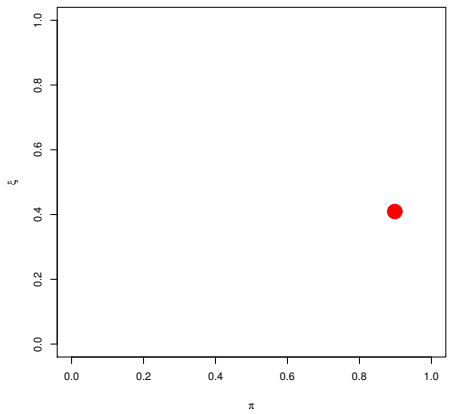
- Esistono modi semplici per rispondere fondati, per esempio, su ri-proporzionamenti.
- Per fornire risposte rigorose occorre, invece, inquadrare le risposte dei consumatori in un contesto più ampio, capace di interpretare le preferenze espresse come risultato di un ***processo generatore dei dati***.

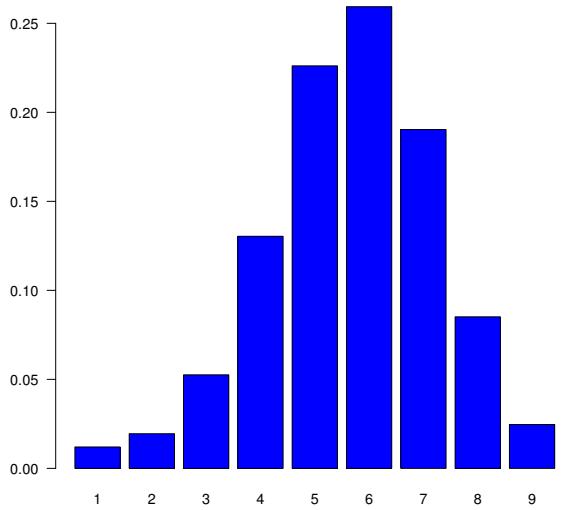
➤ Accettando la logica dei modelli CUB come *ambiente di riferimento statistico* per i dati ordinali (e quindi per le espressioni di preferenza di un prodotto), la risposta avviene secondo le seguenti tappe:

- Per ciascun prodotto valutato, per esempio su una scala con $m = 5$, si stima il modello CUB e, quindi, i parametri che lo caratterizzano.
- Ogni modello CUB è ***m-invariante***, cioé estrae dai dati le informazioni essenziali su tutta la distribuzione e le riassume nei parametri di *feeling* e *uncertainty*. ***Questo avviene a prescindere dal numero*** m ***di modalità***.
- Si calcola la distribuzione teorica delle risposte per la scala che si desidera, per esempio con $m = 9$, utilizzando come parametri quelli stimati al primo step.
- Le due distribuzioni stimate (quelle con $m = 5$ e quella con $m = 9$, per esempio) possiedono TUTTE le caratteristiche e tutti gli indicatori coincidenti.
- Il confronto tra indicatori stimati dalle distribuzioni ed indicatori calcolati sui dati è una misura per la qualità della trasformazione.

Un esempio . . . . . .







1	0.05263
2	0.15789
3	0.31579
4	0.36842
5	0.10526

$\rightarrow \quad \hat{\pi} = \mathbf{0.898}; \quad \hat{\xi} = \mathbf{0.409} \quad \leftrightarrow$

1	0.012020
2	0.019477
3	0.052555
4	0.130370
5	0.226122
6	0.259358
7	0.190327
8	0.085139
9	0.024633

$$Pr(R_i = r | \boldsymbol{x}_i, \boldsymbol{w}_i) = \pi_i \underbrace{\left[\binom{m-1}{r-1} (1-\xi_i)^{r-1} \xi_i^{m-r} \right]}_{\textit{feeling distribution}} + (1-\pi_i) \underbrace{\left[\frac{1}{m} \right]}_{\textit{uncertainty distribution}}$$

for $r = 1, 2, \ldots, m$, where $\pi_i \in (0, 1]$ and $\xi_i \in [0, 1]$, for $i = 1, 2, \ldots, n$.

➤ CUB models estimates the *weight of uncertainty* $1 - \pi_i$, **not** the parameters of the probability distribution assumed for the uncertainty.

➤ As a consequence, *if covariates are significant*, any CUB model assumes a **non-constant uncertainty** $1 - \pi_i$ ***which modifies with subjects*** $(i = 1, 2, \ldots, n)$ **not with categories** $(r = 1, 2, \ldots, m)$.

Parte VI

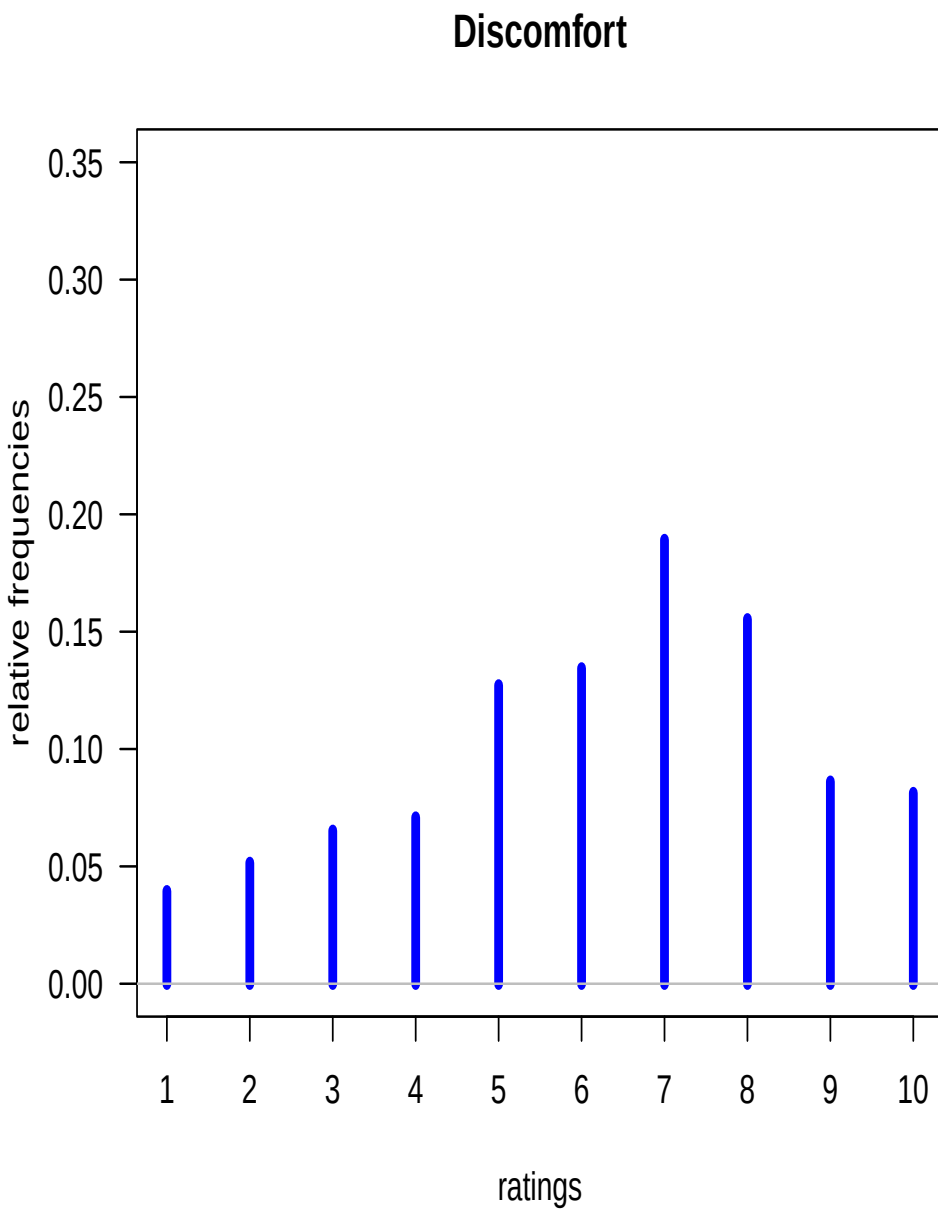
Un'applicazione per il dataset DIUBAS2023

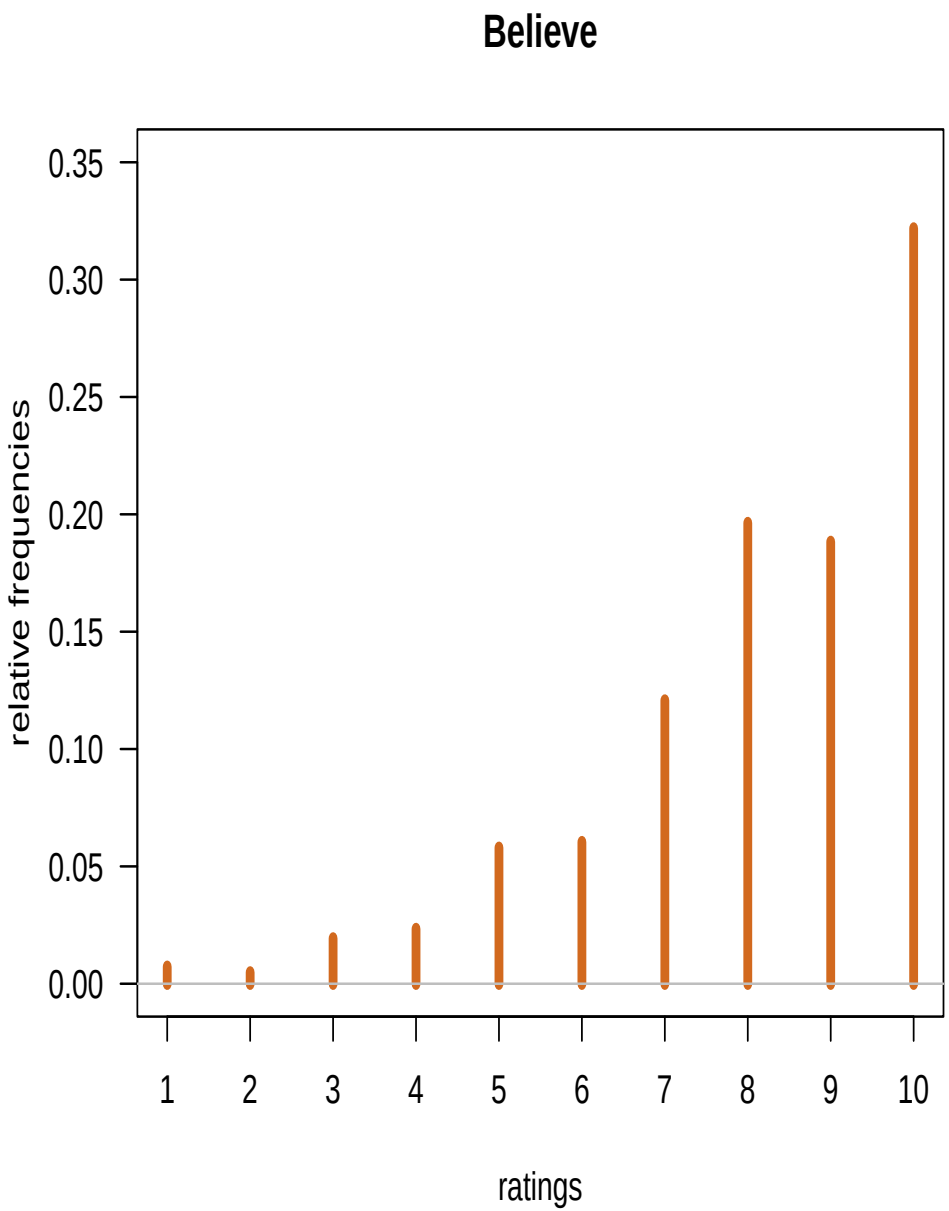
➤ A recent survey (8-22 May 2023) about the distress/discomfort (*disagio*) of University students has been planned at University of Basilicata.

➤ A large sample of students ($n = 1243$, where the population size is $N \simeq 6000$) answered on a 10 point Likert scale to several items. Two of them were:

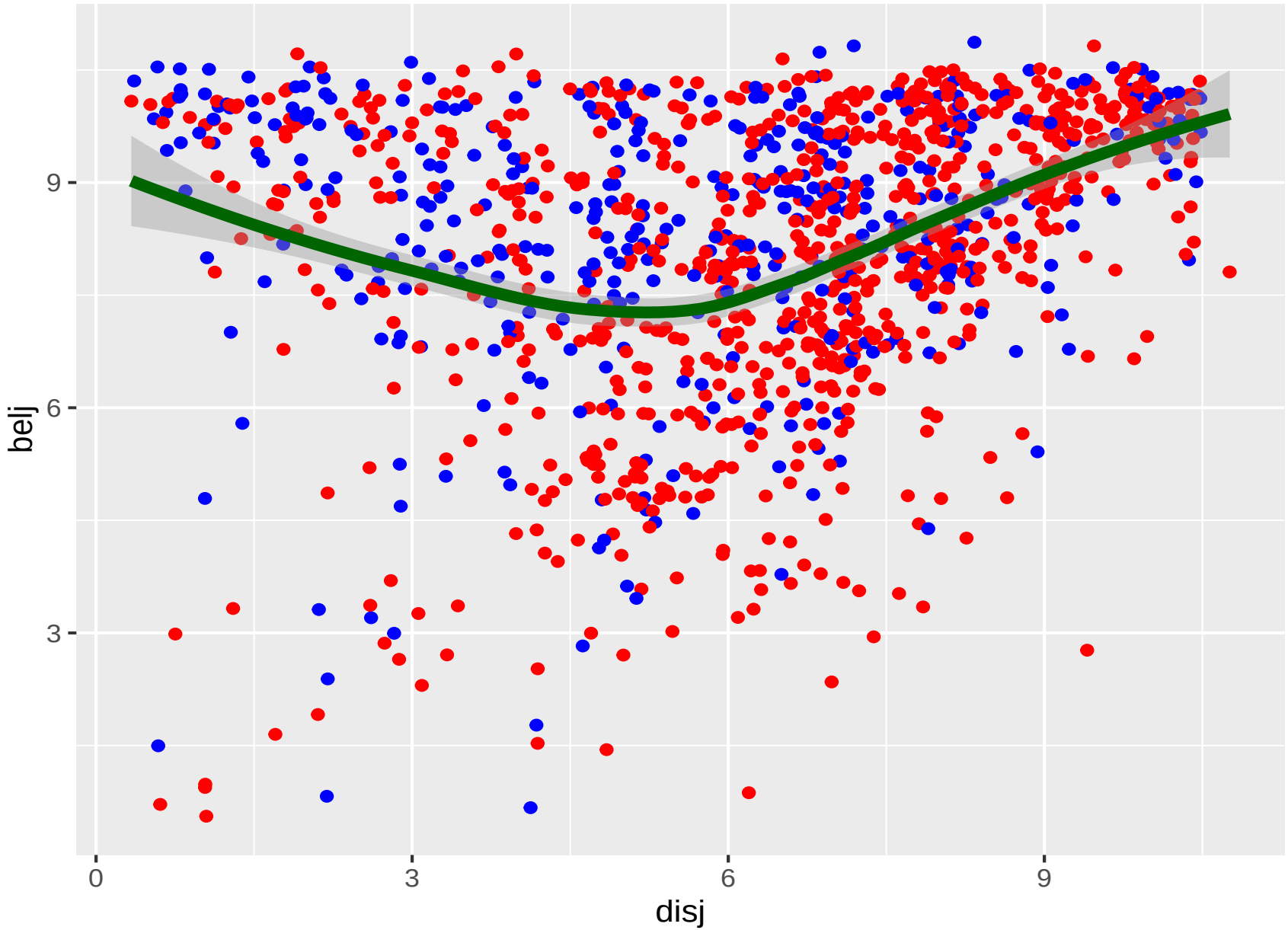
1. ***After carefully thinking, how strong is the personal discomfort/distress you are experiencing?*** $(Dist_i)$
2. ***How strongly do you believe in the answer you provided to the previous question regarding your level of discomfort/distress?*** (Bel_i)

An application of CUB models with varying uncertainty ..... (2)

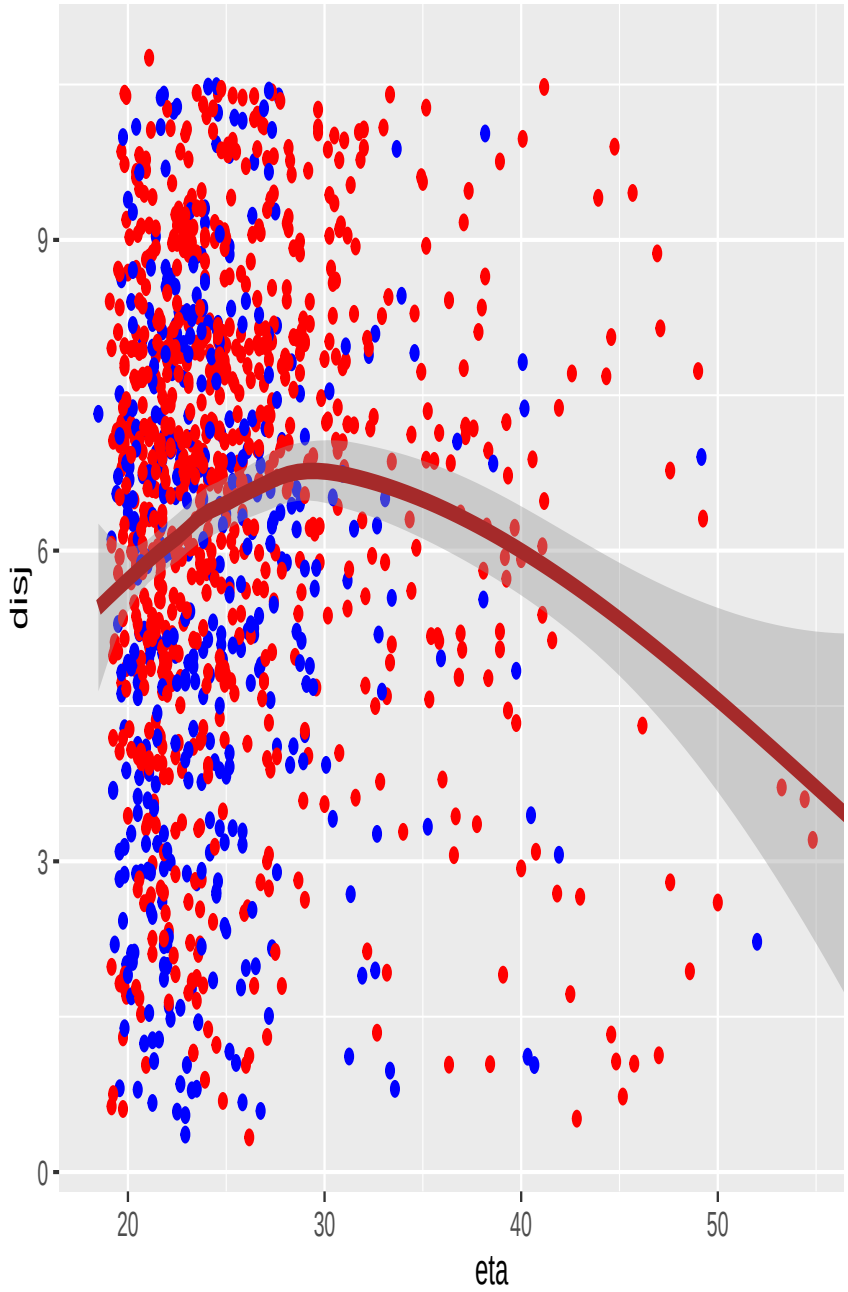


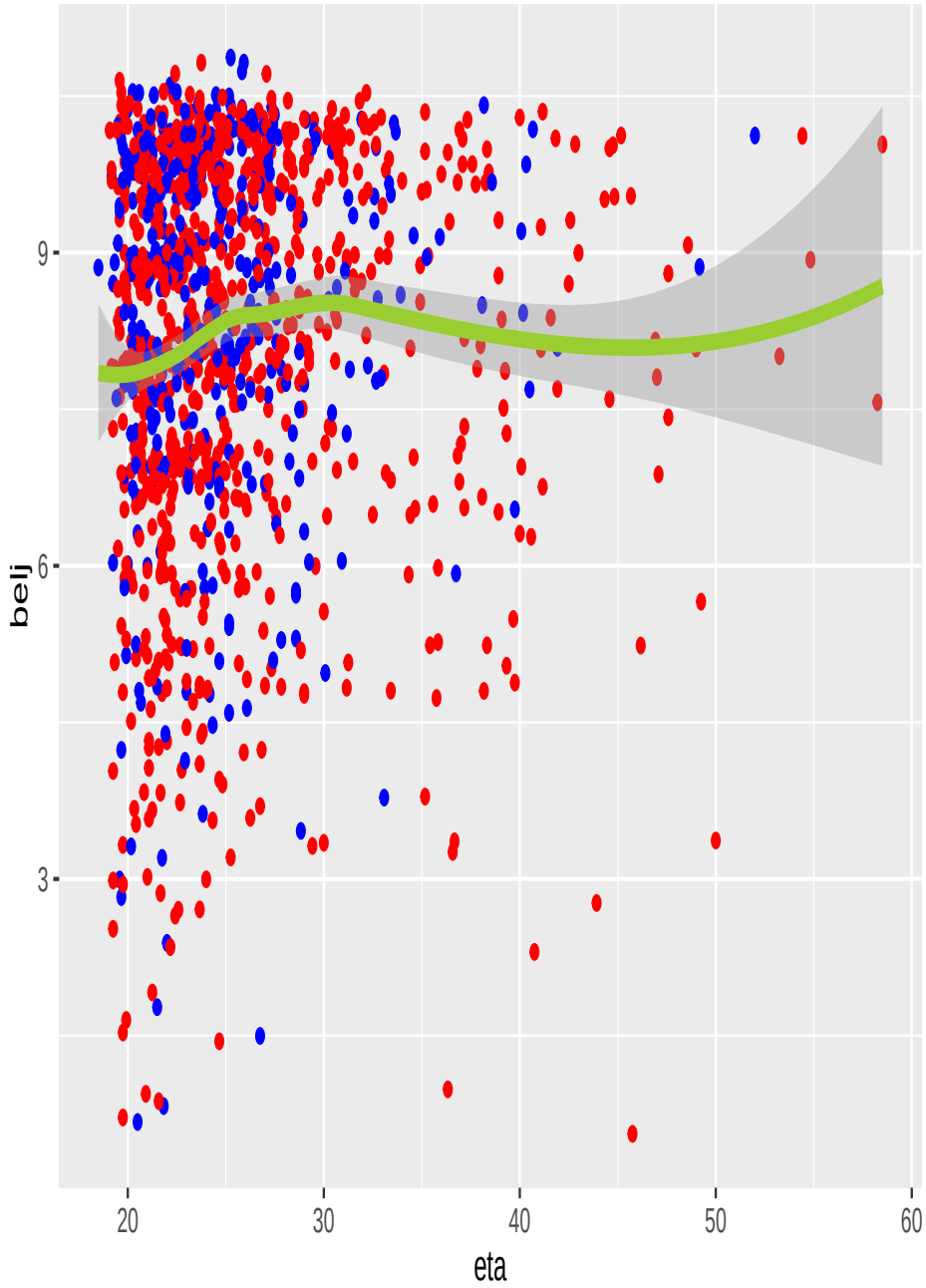


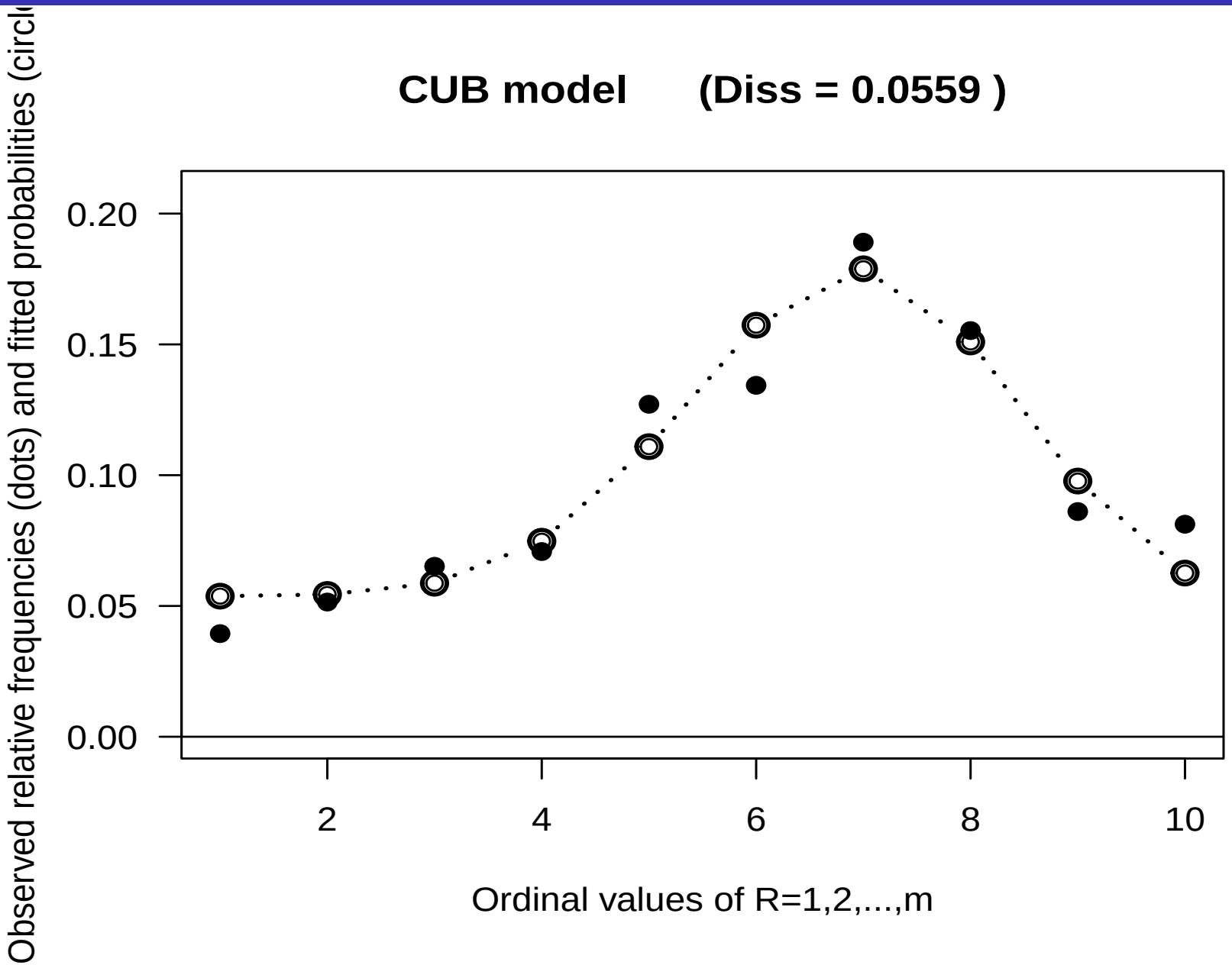
An application of CUB models with varying uncertainty ..... (3)



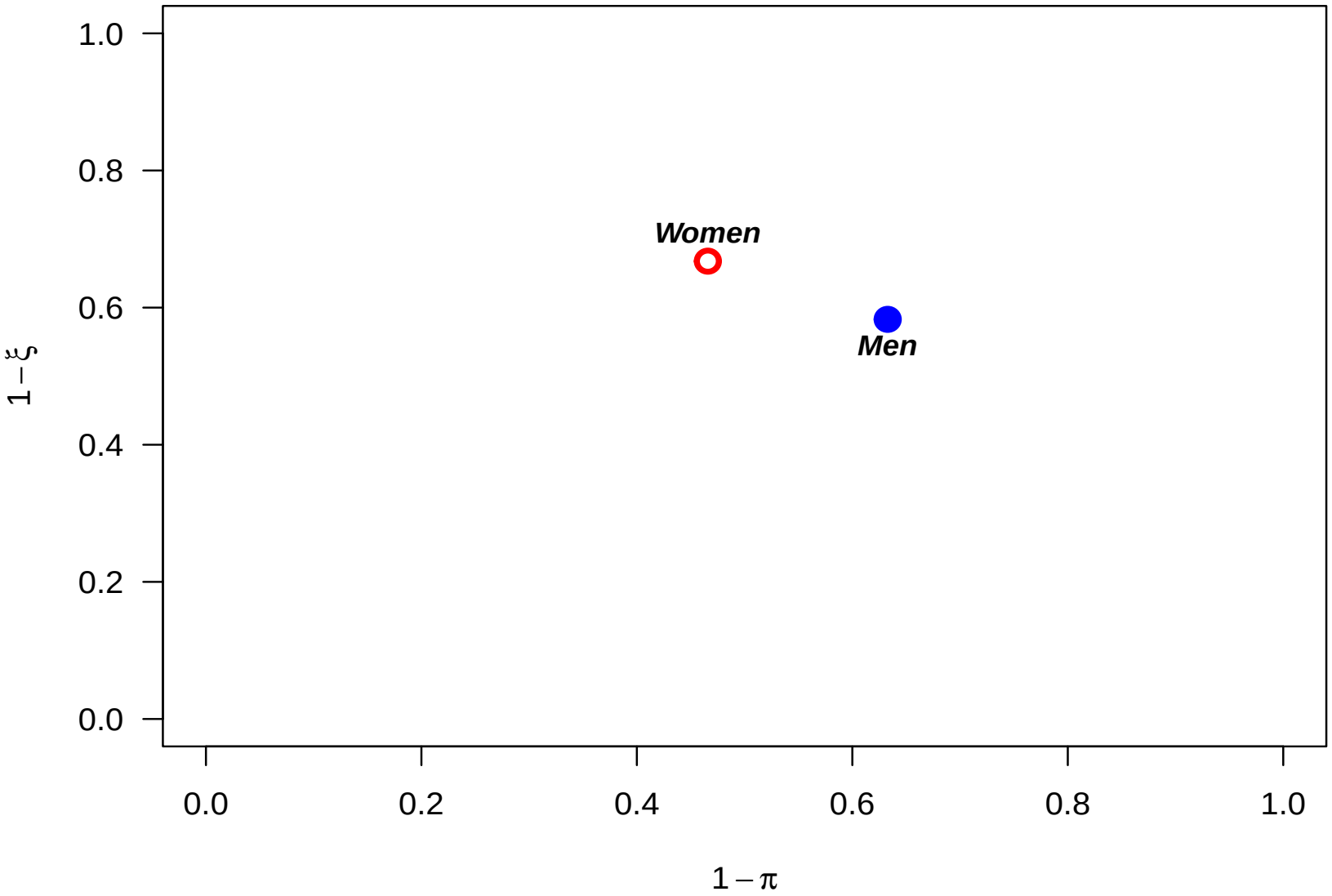
An application of CUB models with varying uncertainty ..... (4)






CUB model (Diss = 0.0559)
Observed relative frequencies (dots) and fitted probabilities (circl
0.20
0.15
0.10
0.05
0.00
2
4
6
8
10
Ordinal values of R=1,2,...,m

An application of CUB models with varying uncertainty ..... (6)



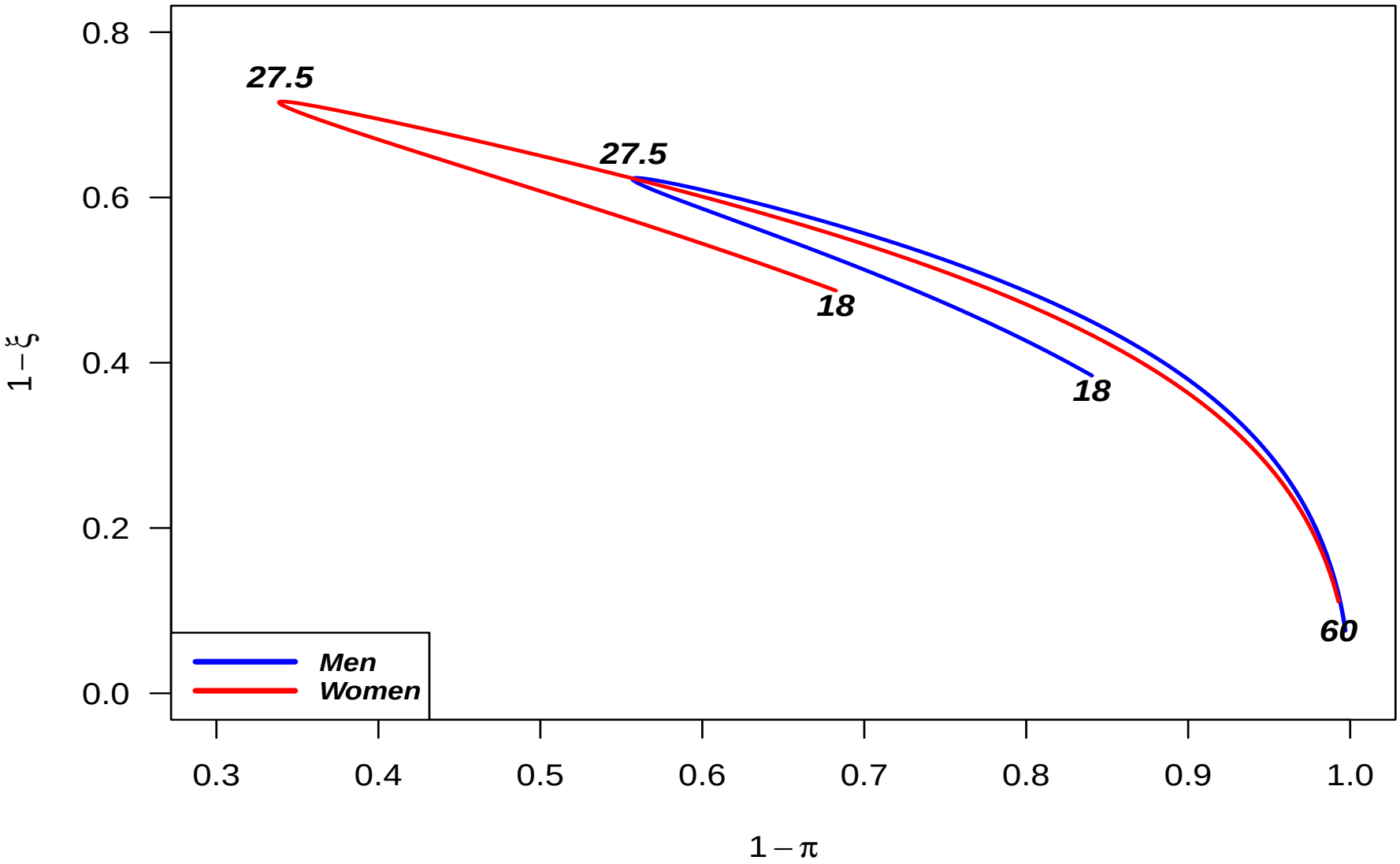
$$\widehat{Pr(Dis_i = r)} = \hat{\pi}_i \underbrace{\left[\binom{m-1}{r-1}(1-\hat{\xi}_i)^{r-1}\hat{\xi}_i^{m-r}\right]}_{\textit{feeling distr.}} + (1-\hat{\pi}_i) \underbrace{\left[\frac{1}{m}\right]}_{\textit{uncertainty distr.}} \quad , \quad r = 1, 2, \ldots, m$$

$$\begin{cases} logit(1-\hat{\pi}_i) = \underset{(33.209)}{97.271} - \underset{(20.073)}{58.856}\,[\log(Age_i)] + \underset{(3.028)}{8.925}\,[\log(Age_i)]^2 - \underset{(0.281)}{0.922}\,Gender_i \\ logit(1-\hat{\xi}_i) = -\underset{(15.726)}{57.122} + \underset{(9.617)}{34.635}\,[\log(Age_i)] - \underset{(1.469)}{5.203}\,[\log(Age_i)]^2 + \underset{(0.114)}{0.409}\,Gender_i \end{cases}$$

for $i = 1, 2, \ldots, n$.

Models	Covariates of (π_i, ξ_i)	*Log-lik*	BIC
CUB	=====	-2717.864	5497.979
CUB	*Gender*	-2730.588	5489.677
CUB	*Age*	-2728.307	5499.365
CUB	*Age, Gender*	-2712.530	5482.063
CUB	*Believe, Gender*	-2577.284	5211.571
CUB	*Believe, Age, Gender*	-2561.810	5209.123

An application of CUB models with varying uncertainty ..... (8)



Considerazioni finali

■ **paradigma**,
s. m. dal latino tardo paradigma, greco $\pi\alpha\rho\alpha\delta\iota\gamma\mu\alpha$, derivato di $\pi\alpha\rho\alpha\delta\iota\kappa\nu\nu\mu\iota$ per: "mostrare, presentare, confrontare", nome composto.

- 1. .....
- 2. .....
- 3. *Nel linguaggio filosofico, termine usato da Platone per designare le realtà ideali concepite come eterni modelli delle transeunti realtà sensibili, e da Aristotele per indicare l'argomento, basato su un caso noto, a cui si ricorre per illustrare uno meno noto o del tutto ignoto. Con altro significato, il termine è stato recentemente introdotto nella sociologia e filosofia della scienza per indicare quel* ***complesso di regole metodologiche, modelli esplicativi, criteri di soluzione di problemi che caratterizza una comunità di scienziati in una fase determinata dell'evoluzione storica della loro disciplina****: a mutamenti di paradigma sarebbero in tal senso riconducibili le cosiddette "rivoluzioni scientifiche".*

(Enciclopedia Treccani)

➤ According to Kuhn (1962), a ***paradigm*** includes "*the practices that define a scientific discipline at a certain point in time*".

➤ A change in the classical approach to ordinal data modelling has been proposed to improve the comprehension of the subjective mechanism of a discrete selection out of a list of ordinal categories.

- ***We are not working with a single model, a collection of models, a variant of existing models.***
- ***Indeed, we are proposing and implementing a whole framework (that is a "paradigm") based on the Generating Data Process of ratings.***
- ***This process includes covariates if and when their effects are significant to explain respondents' behaviour.***

- *The added value of this paradigm is a parsimonious model, a visualization feature, a better interpretation of parameters and the estimation of direct relationship with subjects' and objects' covariates.*

- *Statisticians are well aware of the role and importance of uncertainty in human decisions.*

- *Thus, CUB models may be considered as building blocks of more complex statistical specifications, that is a sort of benchmark to achieve better models . . . which in turn should ever be improved.*

- *Probably, time is not ripe yet for a paradigm shift. Nevertheless, we are supporting this prospective paradigm which is emerging as highly promising ..... with comfortable clues.*