



**UNIVERSITÀ DEGLI STUDI DI PADOVA**

***FACOLTÀ DI SCIENZE STATISTICHE***

***CORSO DI LAUREA SPECIALISTICA IN STATISTICA E INFORMATICA***

TESI DI LAUREA

**IDENTIFICAZIONE DEI PROFILI  
MULTIDIMENSIONALI DELLA SALUTE:  
UN'APPLICAZIONE DEL MODELLO GRADE OF  
MEMBERSHIP AI DATI DEL WORLD HEALTH  
SURVEY**

*Relatore: Ch.mo prof. Nadia Minicuci*

*Laureanda: Alessandra Andreotti*

ANNO ACCADEMICO 2006-2007



*Più di tutti ringrazio i miei genitori per avermi permesso di continuare gli studi e di raggiungere questo importante traguardo.*

*Ringrazio, poi, i miei nonni, Claudio e Lina, che mi hanno sempre sostenuta e che so essere molto orgogliosi di me.*

*Ringrazio Marco per avermi sopportata in questi anni di Università e per essermi sempre stato accanto durante i momenti di sconforto.*

*È doveroso, inoltre, ringraziare la prof.ssa Nadia Minicuci, la dr.ssa Paola Siviero ed il dr. Federico Campigotto, per i loro consigli e per la loro disponibilità durante tutta la durata dello stage presso il Consiglio Nazionale delle Ricerche (C.N.R.) Istituto di Neuroscienze, sezione di Padova.*

*Concludendo, ringrazio tutti gli amici ed i familiari per aver creduto in me e per aver condiviso difficoltà ed entusiasmi.*

*Alessandra*



# INDICE

<b>PREFAZIONE.....</b>	<b>1</b>
<b>Capitolo 1 INTRODUZIONE.....</b>	<b>3</b>
<b>Capitolo 2 IL MODELLO GoM .....</b>	<b>9</b>
2.1. Notazione e formulazione.....	11
2.1.1. Formulazione matriciale.....	15
2.2. Stima dei parametri.....	16
2.2.1. Altro metodo di stima.....	18
2.2.1 Scelta del numero di profili estremi $K$ .....	19
2.3. Confronto con altri metodi di classificazione.....	20
2.3.1. Analisi fattoriale.....	20
2.3.2. Analisi delle componenti principali.....	21
2.3.3. Analisi di raggruppamento, Cluster Analysis.....	22
<b>Capitolo 3 RAPPRESENTAZIONE A CLASSI LATENTI.....</b>	<b>25</b>
3.1. Modello GoM come caso particolare del modello a classi latenti.....	27
3.1.1. Modello a classi latenti.....	27
3.1.2. Modello GoM.....	29
3.2. Interpretazione.....	32
<b>Capitolo 4 APPLICAZIONE DEL MODELLO GoM A DATI     REALI.....</b>	<b>33</b>
4.1. World Health Survey.....	35
4.1.1. Disegno di campionamento.....	36
4.1.2. Il questionario.....	39

4.2. Analisi preliminare dei dati.....	41
4.3. Stima del modello GoM.....	54
4.3.1. <i>Africa</i> .....	55
4.3.2. <i>America</i> .....	62
4.3.3. <i>Mediterraneo Orientale</i> .....	66
4.3.4. <i>Europa</i> .....	71
4.3.5. <i>Sud-Est Asiatico</i> .....	76
4.3.6. <i>Pacifico Occidentale</i> .....	81
<b>Capitolo 5 DISCUSSIONE.....</b>	<b>87</b>
<b>CONCLUSIONI.....</b>	<b>93</b>
<b>APPENDICE.....</b>	<b>97</b>
<b>BIBLIOGRAFIA.....</b>	<b>103</b>

# PREFAZIONE

Attualmente, il concetto di salute è esteso ad una definizione ampia e complessa che include la capacità dei soggetti di essere in equilibrio con sé stessi e di godere di un “completo benessere fisico, mentale e sociale”<sup>1</sup>. Secondo questa prospettiva, per una valutazione globale dello stato di salute non si fa più riferimento all’assenza di malattia ma viene, invece, attribuita importanza alla percezione soggettiva delle condizioni di salute dei soggetti. Tuttavia, negli ultimi anni ha assunto sempre più evidenza il carattere multidimensionale della valutazione soggettiva dello stato di salute.

Al fine di poter valutare tale carattere multidimensionale, in questa tesi è stato impiegato uno strumento di analisi di dati multidimensionali denominato *Grade of Membership Analysis* (analisi GoM), ed applicato ai dati dell’indagine *World Health Survey* (WHS) eseguita dall’*Organizzazione Mondiale della Sanità* (OMS). Questi dati rappresentano risposte di tipo policotomico relative allo stato di salute della popolazione del mondo.

L’obiettivo di questa tesi è cercare di scorporare il concetto di stato salute in più componenti, in modo da poter comprendere le caratteristiche e le tipologie multidimensionali che meglio lo interpretano.

---

<sup>1</sup> Definizione di salute contenuta nella Costituzione dell’Organizzazione Mondiale della Sanità (OMS).

Le analisi riportate di seguito sono state svolte utilizzando principalmente due pacchetti applicativi: SAS 9.1, software per le analisi statistiche e la rappresentazione grafica dei dati, e DSIGoM 1.0, programma specifico per svolgere l'analisi GoM.



## *Capitolo 1*

# **INTRODUZIONE**

Attualmente, l'analisi di dati categoriali di grandi dimensioni è un argomento di rilevante interesse per statistici, biostatistici ed altri studiosi coinvolti in analisi quantitative. Questi tipi di dati si incontrano di frequente in studi demografici, sociali, epidemiologici, clinici, ecc. Solitamente vengono registrati in una tabella di contingenza multidimensionale, la quale diventa rapidamente vasta quando il numero di variabili aumenta. Nei problemi che trattano dati discreti, solitamente si assume che gli individui siano omogenei nelle loro risposte. Questa assunzione conduce istintivamente ad un modello di tipo log-lineare e permette lo studio dell'inter-dipendenza tra le variabili osservate.

Nella metà del ventesimo secolo, vari ricercatori hanno iniziato a sviluppare metodi che potevano essere utilizzati per capire come gli individui differiscono nelle loro risposte.

In altre più complicate situazioni, la probabilità che un individuo risponda ad una determinata modalità può dipendere da una quantità o caratteristica non osservabile, chiamata *latente*.

I principali metodi che trattano con queste entità sono: il modello a classi latenti, *Latent Class Model* (LCM), ed il modello “del grado di appartenenza”, *Grade of Membership* (GoM).

Il modello GoM è stato sviluppato da Max Woodbury verso la fine degli anni '70 [1], come tecnica statistica multivariata per classificazioni mediche.

Nel 1994, Manton, Woodbury e Tolley hanno raggruppato tutta la letteratura presente su questo argomento in un libro intitolato “*Statistical Applications Using Fuzzy Sets*” [2]. Anche se il GoM è usualmente impiegato come modello per insiemi confusi, *Fuzzy Sets Model*, può essere anche descritto come modello a strutture latenti, *Latent Structure Model* (LSM), la cui classificazione è mostrata nella tabella sottostante.

**Tabella 1.1.** Schema di classificazione dei modelli a strutture latenti.

	<b>Variabili Manifeste</b>	
<b>Variabili Latenti</b>	<i>Quantitative</i>	<i>Qualitative</i>
<i>Quantitative</i>	Analisi Fattoriale	Analisi dei Tratti Latenti
<i>Qualitative</i>	Analisi dei Profili Latenti	<b><i>Analisi delle Classi Latenti</i></b>

L'analisi fattoriale è una tecnica di riduzione dei dati che ricava fattori riassuntivi, detti anche *variabili latenti*, che concentrano le informazioni originarie in un numero minore di variabili.

Il modello a tratti latenti, invece, assume che la risposta di un soggetto ad un item può essere spiegata da una o più caratteristiche ovvero *tratti* non direttamente osservabili, a cui si accede definendo una relazione tra essi ed il punteggio riportato su quel item.

L'analisi dei profili latenti è un'estensione dell'analisi a classi latenti, nel senso che mentre quest'ultima è ristretta a variabili indicatrici di tipo dicotomico, il metodo dei profili latenti si adatta a variabili continue, ordinali e categoriali.

I modelli a classi latenti sono stati introdotti da Paul Lazarsfeld verso la fine degli anni '40 [3]. Da un primo confronto, modello GoM e modello a classi latenti possono sembrare simili in teoria ed implementazione; invece, come verrà mostrato nel seguito (Capitolo 3), questi due modelli sono concettualmente diversi tra loro.

Molti metodi statistici sono basati sull'assunzione che persone od oggetti possono essere organizzati in insiemi detti "*crisp*". Questo significa che un individuo può appartenere ad uno ed un solo insieme. Quando non è possibile adottare questa assunzione, allora si ha che i soggetti appartengono a gruppi confusi (*fuzzy*) e ad ogni individuo viene associato un grado di appartenenza all'insieme.

Durante gli ultimi 25 anni, il concetto di *fuzzy set* è stato sviluppato ed attuato nella risoluzione di molti problemi. La maggiore differenza tra insiemi *fuzzy* e *crisp* consiste nella natura dell'associazione degli individui ai vari gruppi. Più precisamente, l'appartenenza ad un insieme *crisp* richiede che gli individui siano "pieni" membri di quel insieme. Negli insiemi *fuzzy*, invece, un individuo è definito membro "parziale" di più gruppi.

Formalmente, per ogni elemento di un insieme *fuzzy* si ha un punteggio del “grado di associazione”, che rappresenta il grado con cui un individuo della popolazione appartiene ad un determinato gruppo. Questi punteggi, per assunzione, variano nell’intervallo  $[0, 1]$ , dove 0 indica che l’elemento non è membro di un determinato insieme, mentre 1 indica che l’individuo è “pieno” membro di quel dato insieme. Il punteggio del “grado di appartenenza” ad un insieme è o 1 o 0 se l’insieme è *crisp*, mentre, nel caso opposto, l’insieme è detto *fuzzy*.

Nel capitolo 2 verrà introdotto il modello GoM come strumento di analisi per questi insiemi *fuzzy*. In particolare, si tratterà la sua formulazione statistica (sezione 2.1.), la stima dei parametri ed il metodo di scelta del numero di profili estremi  $K$  (sezione 2.2.).

Inoltre, poiché tale metodo è una tecnica di analisi multivariata, nella sezione 2.3. si tenterà di impostare un confronto con altre procedure quali l’analisi fattoriale, l’analisi delle componenti principali e la Cluster Analysis (o analisi di raggruppamento).

Come affermato in precedenza, dato che il modello GoM può essere anche considerato un caso particolare del modello a classi latenti, nel capitolo 3 si introdurrà il modello a classi latenti come generalizzazione del modello GoM (sezione 3.1.): prima si presenterà il modello a classi latenti (sezione 3.1.1.), poi il modello GoM come suo caso particolare (sezione 3.1.2.), ed infine si proporrà un’interpretazione di questa comparazione (sezione 3.2.).

Per quanto riguarda la parte pratica di questo elaborato, nel capitolo 4 si cercherà di applicare il modello GoM ai dati raccolti dall'*Organizzazione Mondiale della Sanità* (OMS), mediante l'indagine denominata *World Health Survey* (WHS) sullo stato di salute della popolazione del mondo.

In particolare, si discuteranno gli obiettivi di tale indagine (sezione 4.1.), il disegno di campionamento utilizzato (sezione 4.1.1.) ed il tipo di questionario impiegato per la raccolta dei dati (sezione 4.1.2.). Successivamente verrà svolta un'analisi preliminare dei dati mediante il calcolo dei rapporti di prevalenza (sezione 4.2.) e si stimerà il modello GoM per ogni area regionale (sezione 4.3.). Nel capitolo finale, capitolo 5, verrà affrontata una discussione sui risultati ottenuti con relative conclusioni. Infine, in Appendice sarà possibile trovare l'estratto del questionario originale utilizzato per la costruzione del dataset.

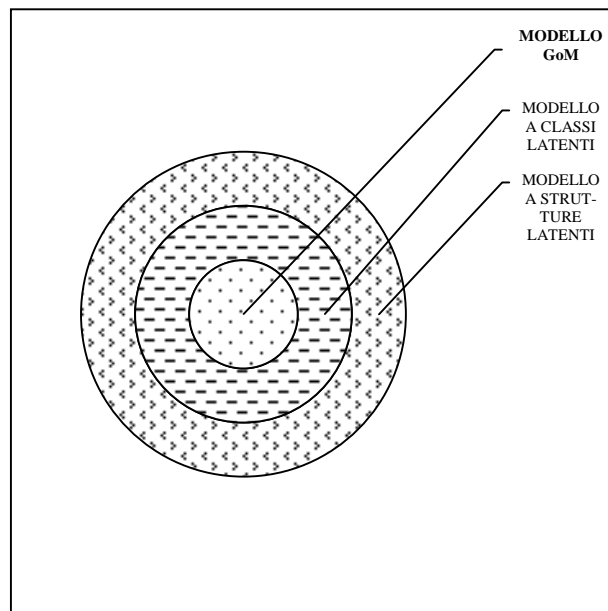


## Capitolo 2

# IL MODELLO GoM

Il modello GoM è una tecnica multivariata utilizzata per descrivere dati discreti di grandi dimensioni. La sua struttura matematica descrive risposte discrete di individui come se fossero membri “parziali” di più gruppi piuttosto che membri esclusivi di un solo gruppo.

**Figura 2.1.** Classificazione dei modelli a strutture latenti considerati.



La figura 2.1 aiuta a comprendere la classificazione dei modelli citati in questi paragrafi.

Il modello a strutture latenti, menzionato nell'introduzione, rappresenta la famiglia generale di appartenenza dei modelli GoM e a classi latenti. Infatti, come si può notare dal diagramma, il modello a classi latenti è un modello a strutture latenti, mentre il modello GoM è un caso particolare del modello a classi latenti.

Il modello GoM è stato applicato con successo ad una vasta gamma di dati ottenendo soluzioni sostanzialmente e teoricamente espressive e che manifestavano robustezza di fronte a problemi di dati mancanti ed errori di campionamento. In generale, la forma matematica del modello GoM rappresenta una struttura analitica molto flessibile, che può essere modificata in molti modi per analizzare diversi tipi di dati. In sintesi, il modello GoM è una procedura progettata specificatamente per l'analisi di dati categoriali multivariati, dove ogni variabile è misurata con errore.



## 2.1. Notazione e formulazione

La struttura generale dei dati può essere descritta come una collezione di risposte ad un numero di variabili discrete (ad esempio domande di un questionario). Si assume che gli individui siano campionati casualmente da una popolazione di interesse e che le domande del questionario siano fissate.

Si considerino le risposte discrete a  $J$  domande policotomiche<sup>2</sup> per  $I$  individui, codificate nella seguente forma binaria:

$$X_{ijl} = \begin{cases} 1 & \text{se l'individuo } i\text{-esimo assume alla domanda } j\text{-esima la modalità } l\text{-esima} \\ 0 & \text{altrimenti} \end{cases}$$

con  $i = 1, \dots, I, j = 1, \dots, J, l = 1, \dots, L_j$

dove  $x_{ijl}$  denota anche la corrispondente variabile casuale binaria. Nel seguito,  $x_{ijl}$  indicherà sia la risposta osservata dell'individuo  $i$ -esimo alla domanda  $j$ -esima, sia la corrispondente variabile casuale.

Il modello GoM assume, inoltre, che la popolazione di riferimento è suddivisa in profili estremi (*extreme profiles*, cioè sottopopolazioni) e che gli individui sono caratterizzati da parametri specifici detti punteggi di associazione (*membership scores*), i quali indicano la “proporzione” di appartenenza ad ogni profilo estremo.

---

<sup>2</sup> Policotomico è il contrario di dicotomico. Una variabile è detta policotomica se assume  $p$  modalità, con  $p > 2$ .

In questo senso, si supponga ci siano  $K$  profili estremi nella popolazione. Si assuma, poi, che ogni soggetto sia caratterizzato da un vettore di punteggi di associazione,  $g_i = (g_{i1}, \dots, g_{iK})$ , dove la  $k$ -esima componente corrisponde al punteggio di associazione attribuito al soggetto  $i$ -esimo per il  $k$ -esimo profilo estremo.

Questi punteggi sono compresi nell'intervallo  $[0, 1]$  e la loro somma è 1 per ogni soggetto:

$$\sum_{k=1}^K g_{ik} = 1, \quad i = 1, \dots, I.$$

Per capire meglio il significato di questi punteggi  $g_{ik}$ , è stata costruita la seguente tabella, la quale contiene tante righe quanti sono i soggetti del campione e tante colonne quanti sono i profili estremi.

**Tabella 2.1.1.** Punteggi di associazione.

	1	2	...	K	
1	$g_{11}$	$g_{12}$	...	$g_{1k}$	= 1
2	$g_{21}$	$g_{22}$	...	$g_{2k}$	= 1
...	...	...	...	...	...
...	...	...	...	...	...
I	$g_{I1}$	$g_{I2}$	...	$g_{IK}$	= 1

Le probabilità di risposta nel  $k$ -esimo profilo estremo, indicate da  $\lambda_{kjl}$ , sono le probabilità di risposta alla modalità  $l$ -esima nella domanda  $j$ -esima per un membro "completo" del  $k$ -esimo profilo estremo

$$\lambda_{kjl} = \Pr(x_{ijl} = 1 \mid g_{ik} = 1)^3. \quad (2.1.1)$$

<sup>3</sup> Questa è la formulazione generale delle probabilità  $\lambda_{kjl}$ .

Servono, però, delle assunzioni aggiuntive [2] per completare la formulazione di questo modello, e cioè:

- (1) La probabilità che l'individuo  $i$ -esimo risponda alla modalità  $l$ -esima nella domanda  $j$ -esima, condizionatamente al vettore punteggi GoM,  $g_i$ , è

$$p_{ijl} = \Pr(x_{ijl} = 1 \mid g_i) = \sum_{k=1}^K g_{ik} \cdot \lambda_{kjl} ; \quad (2.1.2)$$

- (2) Condizionatamente ai valori dei punteggi GoM,  $g_{ik}$ , le risposte osservate  $x_{ijl}$  sono indipendenti per diversi valori di  $j$ ;
- (3) Le risposte osservate  $x_{ijl}$  sono indipendenti per diversi valori di  $i$ ;
- (4) I punteggi GoM,  $g_{ik}$ , sono realizzazioni delle componenti di un vettore casuale con distribuzione  $D(g)$ .

Di queste quattro assunzioni le prime tre sono essenziali.

L'assunzione (1) richiede che le probabilità di risposta individuali ( $p_{ijl}$ ) siano una *combinazione convessa*<sup>4</sup> delle probabilità di risposta nei  $K$  profili estremi ( $\lambda_{kjl}$ ), pesate dai corrispondenti punteggi GoM ( $g_{ik}$ ).

L'assunzione (2) è nota anche come *assunzione di indipendenza locale*, la quale asserisce che una risposta osservata è indipendente da tutte le altre.

---

<sup>4</sup> Dati  $p_0, \dots, p_d \in \mathbb{R}^n$  e  $\alpha_0, \dots, \alpha_d \in \mathbb{R}^+ \cup \{0\}$ , con  $\alpha_0 + \dots + \alpha_d = 1$ , il punto

$\alpha_0 p_0 + \dots + \alpha_d p_d = \sum_{i=0}^d \alpha_i \cdot p_i$  è detto *combinazione convessa* dei punti  $p_0, \dots, p_d$ .

Questo significa che la risposta ad un quesito non dipende in nessun modo dalle risposte date alle altre domande. Questa è l'assunzione di base dei modelli a strutture latenti (tabella 2.1.1) e mostra che le variabili latenti spiegano le relazioni tra le risposte osservate.

L'assunzione (3), invece, corrisponde al campionamento casuale di individui dalla popolazione di riferimento.

Infine, l'assunzione (4) ha un significato anomalo in letteratura. Non è usata nella procedura di stima in Manton et al. [2] (pp. 22-24), né viene implementata nel pacchetto applicativo per la stima del modello GoM (Decision Systems, Inc. 1999) che verrà discusso nel paragrafo 2.2.1.

Versioni del modello GoM con o senza l'assunzione (4) sono definite rispettivamente modello GoM ad effetti misti (*mixed-effects*) o ad effetti fissi (*fixed-effects*). La differenza è che il primo tratta i punteggi di associazione  $g_{ik}$  come variabili casuali secondo l'assunzione (4), mentre il secondo li tratta come costanti ignote fissate.

### 2.1.1. Formulazione matriciale

Si supponga che gli individui della popolazione corrispondano alle righe e le modalità di risposta di ogni domanda corrispondano alle colonne di una matrice contenente le probabilità di risposta  $p_{ijl}$ . Allora, noti i punteggi di associazione specifici di ogni soggetto,  $g_{ik}$ , e le probabilità di risposta nei  $K$  profili estremi,  $\lambda_{ijl}$ , il modello GoM può essere riscritto nella seguente forma matriciale

$$\mathbf{p} = \mathbf{g}\boldsymbol{\lambda},$$

dove  $\mathbf{p}$  è una matrice  $I \times (L_1 + \dots + L_J)$  con  $J$  blocchi  $\mathbf{p}_1, \dots, \mathbf{p}_J$ ,  $\mathbf{g}$  è una matrice  $I \times (KJ)$  con  $J$  blocchi  $\mathbf{g}_0$ , e  $\boldsymbol{\lambda}$  è una matrice  $(KJ) \times (L_1 + \dots + L_J)$  diagonale con  $J$  blocchi  $\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_J$ .

I blocchi  $\mathbf{p}_j$ ,  $\mathbf{g}_0$  e  $\boldsymbol{\lambda}_j$  sono i seguenti:

$$\mathbf{p}_j = \begin{pmatrix} p_{1j1} & \dots & p_{1jL_j} \\ p_{2j1} & \dots & p_{2jL_j} \\ \dots & \dots & \dots \\ p_{ij1} & \dots & p_{ijL_j} \\ \dots & \dots & \dots \\ p_{1j1} & \dots & p_{1jL_j} \end{pmatrix} \quad \mathbf{g}_0 = \begin{pmatrix} g_{11} & \dots & g_{1K} \\ g_{21} & \dots & g_{2K} \\ \dots & \dots & \dots \\ g_{i1} & \dots & g_{iK} \\ \dots & \dots & \dots \\ g_{11} & \dots & g_{1K} \end{pmatrix} \quad \boldsymbol{\lambda}_j = \begin{pmatrix} \lambda_{1j1} & \dots & \lambda_{1jL_j} \\ \lambda_{2j1} & \dots & \lambda_{2jL_j} \\ \dots & \dots & \dots \\ \lambda_{kj1} & \dots & \lambda_{kjL_j} \\ \dots & \dots & \dots \\ \lambda_{Kj1} & \dots & \lambda_{KjL_j} \end{pmatrix}.$$

Si noti che quando si hanno a disposizione solo domande dicotomiche, si ha che

$\lambda_{kj2} = 1 - \lambda_{kj1}$ . Allora, in questo caso, è possibile omettere l'indice  $l = 1, 2$  ed indicare la probabilità di osservare risposte affermative alla domanda  $j$  per il profilo estremo  $k$  come:

$$\lambda_{kj} = \Pr(x_{ij} = 1 \mid g_{ik} = 1).$$

## 2.2. Stima dei parametri

Le procedure di stima dei parametri del modello GoM sono basate sul metodo di massima verosimiglianza. In particolare, ci possono essere due tipi di funzione di verosimiglianza: a) verosimiglianza GoM “condizionata”; b) verosimiglianza GoM “incondizionata”.

- a) Se si assume che i punteggi GoM,  $g_{ik}$ , sono costanti ignote fissate, quindi senza considerare l’assunzione (4), la funzione di verosimiglianza per il modello GoM è la seguente:

$$L(\boldsymbol{\lambda}, \mathbf{g} | \mathbf{x}) = \prod_i \prod_j \prod_l \left( \sum_k g_{ik} \cdot \lambda_{kjl} \right)^{x_{ijl}}, \quad (2.2.1)$$

dove  $\boldsymbol{\lambda} = \{\lambda_{kjl} : k = 1, \dots, K, j = 1, \dots, J, l = 1, \dots, L_j\}$  è la matrice dei parametri locali o strutturali che collegano i K profili estremi con le J variabili osservate,

$\mathbf{g} = \{g_{ik} : i = 1, \dots, I, k = 1, \dots, K\}$  è la matrice dei parametri specifici di ogni soggetto che descrive la distanza tra le caratteristiche osservate da un individuo I e le caratteristiche di ogni profilo K, e i dati  $\mathbf{x} = \{x_{ijl} : i = 1, \dots, I, j = 1, \dots, J,$

$l = 1, \dots, L_j\}$  sono le risposte osservate per ogni soggetto. Manton et al. [2] definiscono l’equazione (2.2.1) *verosimiglianza GoM “condizionata”*, dove il termine “condizionata” indica che il calcolo della verosimiglianza è vincolato dai valori assunti dai parametri  $g_{ik}$ .

b) Sotto l'approccio ad effetti misti, invece, si assume che i punteggi GoM sono realizzazioni casuali da una distribuzione  $D_\alpha(\cdot)$ , parametrizzata dal vettore  $\alpha$ . La funzione di verosimiglianza sotto questa assunzione diventa:

$$L(\lambda, \alpha | \mathbf{x}) = \int \prod_i \prod_j \prod_l \left( \sum_k g_{ik} \cdot \lambda_{kjl} \right)^{x_{ijl}} d D_\alpha(\mathbf{g}). \quad (2.2.2)$$

In questo caso, Manton et al. [2] definiscono questa equazione come *verosimiglianza GoM "incondizionata"*, in contrasto con la verosimiglianza "condizionata" di equazione (2.2.1).

Il metodo di stima illustrato da Manton et al. [2] (pp. 67-70) prende in considerazione la funzione di verosimiglianza "condizionata". Ad ogni iterazione dell'algoritmo, la verosimiglianza viene massimizzata considerando uno dei due parametri fissato.

Per questo motivo vengono fornite due equazioni che permettono di aggiornare le stime in modo sequenziale:

$$\begin{aligned} g_{ik} &= \frac{1}{x_{i++}} \sum_{l=1}^L \sum_{j=1}^{L_j} \left( x_{ijl} \frac{g_{ik}^* \cdot \lambda_{kjl}^*}{p_{ijl}} \right) \\ \lambda_{kjl} &= \frac{\sum_{i=1}^I \left( x_{ijl} \frac{g_{ik}^* \cdot \lambda_{kjl}^*}{p_{ijl}} \right)}{\sum_{i=1}^I \left( x_{ij+} \sum_{l=1}^{L_j} \frac{g_{ik}^* \cdot \lambda_{kjl}^*}{p_{ijl}} \right)}, \end{aligned} \quad (2.2.3)$$

dove  $g_{ik}^*$  e  $\lambda_{kjl}^*$  sono i valori calcolati nella precedente iterazione e  $x_{i++} = \sum_j \sum_l x_{ijl}$ .

Queste due equazioni sono state ottenute massimizzando la funzione di verosimiglianza  $L(\lambda, \mathbf{g} | \mathbf{x})$ , come spiegato da Erosheva [4] (pp. 23-25).

### 2.2.1. Altro metodo di stima

DSIGoM [5] è un linguaggio sviluppato dalla Decision Systems Inc. (Raleigh, North Carolina) per la stima dei parametri del modello GoM. Questo software massimizza la funzione di verosimiglianza “condizionata”, definita nel paragrafo precedente. Per ottenere le equazioni di stima, viene utilizzato il logaritmo della funzione di verosimiglianza (log-verosimiglianza) con moltiplicatori di Lagrange<sup>5</sup>:

$$l(\lambda, \mathbf{g} | \mathbf{x}) = \sum_i \sum_j \sum_l x_{ijl} \ln \left( \sum_k g_{ik} \lambda_{kjl} \right),$$

soggetta ai seguenti vincoli:

$$\sum_k g_{ik} = 1$$
$$\sum_l \lambda_{kjl} = 1 .$$

Le equazioni di stima sono ottenute eguagliando a zero le derivate parziali di primo ordine e le equazioni risultanti vengono risolte mediante applicazione dell’algoritmo di Newton-Raphson<sup>6</sup>.

Alla fine, le equazioni di stima risultanti sono le seguenti:

$$f(g_{ik} | \lambda_{kjl}) = 0$$
$$h(\lambda_{kjl} | g_{ik}) = 0 .$$

---

<sup>5</sup> Il metodo dei moltiplicatori di Lagrange serve a ricondurre una funzione il cui insieme di definizione è un aperto, ad una funzione, detta lagrangiana, definita su un insieme compatto. Data la funzione  $f(x, y)$  vincolata dalla condizione  $g(x, y)=0$ , si definisce la funzione lagrangiana  $z(x, y)=f(x, y)+\lambda \cdot g(x, y)$ , con  $\lambda$  reale. La funzione lagrangiana  $z$  è una funzione di tre variabili  $(x, y, \lambda)$  ed è pari alla somma della funzione iniziale per un multiplo del vincolo  $g(x, y)$ . Il numero  $\lambda$  è chiamato “moltiplicatore di Lagrange”.

<sup>6</sup> Il metodo di Newton-Raphson è un algoritmo iterativo che permette di risolvere numericamente un sistema di equazioni non lineari.



Il primo gruppo di equazioni viene utilizzato per ottenere le stime dei punteggi GoM,  $g_{ik}$ , assumendo fissate le probabilità  $\lambda_{kjl}$ . Analogamente, la seconda serie di equazioni viene impiegata per aggiornare la stima dei parametri  $\lambda_{kjl}$  in base ai valori ottenuti per i  $g_{ik}$  nel precedente calcolo. Quando questo procedimento iterativo arriva a convergenza, entro i gruppi l'omogeneità tra soggetti è massima, mentre tra i gruppi l'omogeneità è minima.

### ***2.2.2. Scelta del numero di profili estremi K***

Spesso il numero di profili estremi K del modello GoM è ignoto. In molti casi si è interessati ad ottenere inferenza su K oppure capire quale valore di K fornisce un migliore adattamento del modello ai dati.

L'ordine del modello K non è direttamente determinabile attraverso l'applicazione dell'algoritmo specificato nei paragrafi precedenti. Una possibilità per stimare il numero ottimale di profili estremi è quella di effettuare un test del rapporto di log-verosimiglianza tra il modello con K e quello con K+1 profili estremi [6]. Questo rapporto è distribuito secondo un  $\chi^2$  con gradi di libertà pari alla differenza del numero di parametri stimati nei due modelli. In pratica, si sceglie quel valore di K per il quale l'incremento di log-verosimiglianza non risulta statisticamente significativo.

## 2.3. Confronto con altri metodi di classificazione

Ci sono alcune analogie ed alcune diversità tra questo metodo e gli altri strumenti di analisi di dati multidimensionali. La differenza principale è che il modello GoM non assume la normalità multivariata dei dati, è quindi un metodo non parametrico. Di seguito vengono presentati i principali strumenti di analisi multivariata [7] in modo da poter eseguire un confronto con il modello GoM appena citato.

### 2.3.1. *Analisi fattoriale*

In generale, è un processo di trasformazione della realtà osservata in una nuova struttura, la cui variabilità è spiegata dai *fattori*.

Dato un insieme di variabili quantitative  $\mathbf{x}$ , con questo metodo si possono soddisfare i seguenti obiettivi:

- (i) spiegare le correlazioni esistenti tra le variabili osservate facendo ricorso ad un numero ridotto di funzioni non osservabili (*latenti*);
- (ii) evidenziare i fattori (o “dimensioni”) impliciti nei dati rilevati e che per questo vengono chiamati *latenti*.

Per raggiungere questi obiettivi, l'insieme delle osservazioni viene trasformato, attraverso tecniche dette di fattorizzazione, in un insieme che è più semplice ma altrettanto informativo di quello di partenza.

L'analisi fattoriale è un metodo fondamentale nella ricerca di pattern latenti nelle variabili osservate. Inoltre, è uno strumento statistico idoneo a ridurre un sistema complesso di correlazioni in un numero minore di dimensioni.

### ***2.3.2. Analisi delle componenti principali***

L'analisi delle componenti principali è simile all'analisi fattoriale. Gli obiettivi principali di questo strumento sono:

- (i) partendo da un insieme di  $p$  variabili originali, trovare combinazioni lineari aventi proprietà di ottimalità dal punto di vista della variabilità e della correlazione;
- (ii) riduzione della dimensionalità del problema.

L'analisi delle componenti principali si differenzia dall'analisi fattoriale nella formulazione del modello: nell'analisi fattoriale, è il risultato di un'ipotesi secondo la quale la variabilità comune alle osservazioni è in qualche modo strutturata (il problema sta nel modo di identificarne la struttura); nell'analisi delle componenti principali, invece, tutte le componenti sono comunque considerate, anche se solo alcune saranno poi utilizzate a fini interpretativi. Inoltre, nell'analisi delle componenti principali la varianza spiegata dalle componenti è il 100% della variabilità osservata. In tal caso la comunanza è 1 per tutte le variabili.

Per quanto riguarda le diversità tra i due metodi descritti in precedenza ed il modello GoM, in quest'ultimo tutti i parametri sono stimati simultaneamente, mentre i parametri dell'analisi fattoriale o del metodo delle componenti principali sono calcolati dopo l'identificazione di nuove variabili che sintetizzano l'informazione di tutti i dati raccolti (fattori o componenti principali). Inoltre, il metodo GoM è uno strumento di analisi non parametrica e la stima dei parametri non conta su nessuna assunzione distributiva. Il calcolo dei punteggi fattoriali nell'analisi fattoriale, invece, poggia su assunzioni distributive, in particolare sulla *distribuzione normale multivariata*. Infine, in opposizione a questi due metodi di analisi, il modello GoM è uno strumento di classificazione dei soggetti sulla base dei loro punteggi di associazione ( $g_{ik}$ ) [8].

### ***2.3.3. Analisi di raggruppamento, Cluster Analysis***

La Cluster Analysis è una raccolta di algoritmi e metodi utilizzati per raggruppare oggetti dello stesso tipo. È un'analisi essenzialmente esplorativa che raggruppa diverse entità in modo che il grado di associazione tra due oggetti sia massimo se appartengono allo stesso gruppo e minimo nel caso opposto. Si usa quando non si hanno ipotesi a priori sulla formulazione e sulla numerosità dei gruppi. L'obiettivo principale è di raggruppare oggetti omogenei all'interno di ogni gruppo e non omogenei al loro esterno.

L'analisi di raggruppamento si distingue dall'analisi fattoriale perché la prima è pertinente per raggruppare entità, mentre la seconda è attinente per lo studio delle relazioni tra variabili.

Inoltre, l'analisi fattoriale assume che le relazioni tra le variabili inserite nel modello sono lineari, mentre ogni vincolo di forma delle relazioni tra variabili è ignorabile nella Cluster Analysis. Questo non esclude che si possa arrivare ad identiche conclusioni adottando l'uno o l'altro metodo. Non solo, ma con opportune elaborazioni, si possono rendere evidenti, dopo aver eseguito una Cluster Analysis, le variabili più discriminanti tra le entità, e dopo un'analisi dei fattori, le unità che più sono simili o dissimili con riferimento ai fattori trovati [7].

Per quanto riguarda le diversità tra questo metodo di analisi ed il modello GoM, quest'ultimo è anch'esso un metodo di classificazione: la differenza è che non crea gruppi di simile entità come nella Cluster Analysis, ma considera l'eterogeneità tra gli individui [8].



## *Capitolo 3*

# **RAPPRESENTAZIONE A CLASSI LATENTI**

Quando si classificano gli individui come membri di classi latenti, si possono presentare due tipi di appartenenza. La prima, definita piena appartenenza (*full membership*), si ha quando ogni soggetto è membro di una ed una sola delle  $K$  classi latenti. In questo modo si forza il vettore dei punteggi,  $g_i$ , ad avere esattamente una componente diversa da zero, cioè uguale a 1. Alternativamente, il secondo tipo, definito associazione parziale (*partial membership*), si ha quando ogni soggetto è membro “parziale” di ognuna delle  $K$  classi latenti. In questo caso, le  $K$  componenti del vettore dei punteggi sono pesi o numeri reali non negativi, vincolati dal fatto che la loro somma è 1. Questo è il caso del modello GoM descritto nel capitolo precedente.

Dal momento che il vettore di associazione parziale è un caso particolare del vettore di piena appartenenza, il modello GoM può essere percepito come caso particolare del modello a classi latenti.

Ci sono, però, diversi modi di confrontare questo modello e quello a classi latenti. Per prima cosa, dal momento che è possibile trattare i punteggi GoM,  $g_{ik}$ , come o costanti ignote fissate o variabili casuali, è possibile scegliere di utilizzare un modello GoM ad effetti fissi oppure ad effetti misti, rispettivamente, per il confronto. Secondariamente, dato che i profili estremi del modello GoM e le classi di un modello a classi latenti sembrano esprimere concetti simili, è possibile effettuare una comparazione quando il numero di profili estremi è uguale al numero di classi latenti.

Nei paragrafi seguenti si cercherà di trovare una relazione tra questi due modelli.



### 3.1. Modello GoM come caso particolare del modello a classi latenti

Tutti i modelli a strutture latenti utilizzano la configurazione generale con  $J$  risposte discrete policotomiche per  $I$  individui. Per semplificare la notazione verrà omesso l'indice  $i$  degli individui. Come indicato in precedenza,  $x = (x_1, \dots, x_J)$  sono  $J$  variabili osservate (manifeste). Per  $j = 1, \dots, J$ ,  $x_j$  assume valori  $l_j \in \mathcal{L}_j = \{1, \dots, L_j\}$ . Nella formulazione di un modello a struttura latente si devono specificare le seguenti due componenti:

- (1) La distribuzione delle variabili latenti,
- (2) La distribuzione delle variabili manifeste condizionata alle variabili latenti.

Di seguito queste quantità saranno descritte sia per il modello a classi latenti che per il modello GoM.

#### 3.1.1. Modello a classi latenti

Sia  $y = (y_1, y_2, \dots, y_K)$  il vettore delle “complete” associazioni latenti definito da

$$y_k = \begin{cases} 1 & \text{se il soggetto è membro della classe } k, k = 1, 2, \dots, K \\ 0 & \text{altrimenti} \end{cases},$$

con funzione di probabilità

$$f(y) = \begin{cases} \pi_k & \text{se } y_k = 1 \text{ e } y_l = 0, l \neq k \\ 0 & \text{altrimenti} \end{cases}.$$

Quindi, se un individuo ha associato un vettore  $y$  con componenti  $(0, 0, 1, 0, \dots, 0)$  significa che questo soggetto appartiene completamente alla terza classe latente.

Si indichi, con

$$\lambda_{kjl_j} = \Pr(x_j = l_j \mid y_k = 1), \quad k = 1, \dots, K, \quad j = 1, \dots, J, \quad l_j = 1, \dots, L_j \quad (3.1.1)$$

la probabilità che la variabile manifesta  $x_j$  assuma valore  $l_j$ , condizionatamente all'appartenenza alla  $k$ -esima classe latente. È importante sottolineare che queste probabilità devono soddisfare il seguente vincolo:

$$\sum_{l_j \in L_j} \lambda_{kjl_j} = 1, \quad k = 1, \dots, K, \quad j = 1, \dots, J.$$

Questa notazione differisce leggermente da quella utilizzata nella (2.1.1), perché sono stati utilizzati dati registrati in formato multivariato ed è stato omesso l'indice  $i$  dei soggetti.

Dato che il vettore  $y$  ha esattamente una componente diversa da zero, la probabilità di osservare  $x_j$ , dato il vettore di appartenenza  $y$ , può essere riscritta come

$$\Pr(x_j = l_j \mid y) = \lambda_{kjl_j} = \sum_{k=1}^K y_k \cdot \lambda_{kjl_j}. \quad (3.1.2)$$

Considerando l'assunzione di indipendenza locale (2 a pag. 13), le variabili osservate  $x_j$  sono indipendenti, data l'appartenenza alla  $k$ -esima classe latente. Allora, la probabilità di osservare l'insieme di risposte  $l$ , condizionata al valore del vettore  $y$ , è

$$f^{LCM}(l \mid y) = \Pr(x = l \mid y) = \prod_{j=1}^J \left( \sum_{k=1}^K y_k \cdot \lambda_{kjl_j} \right).$$

Infine, integrando rispetto la variabile latente  $y$ , si ottiene la distribuzione marginale delle variabili  $x$  come:

$$f^{LCM}(l) = \Pr(x=l) = \int f^{LCM}(l|y)f(y)dy = \sum_{k=1}^K \pi_k \prod_{j=1}^J \lambda_{kjl_j}. \quad (3.1.3)$$

La probabilità di osservare l'insieme di risposte  $l$  è la somma delle probabilità di osservare  $l$  in ognuna delle  $K$  classi latenti,  $\lambda_{kjl_j}$ , pesata dalla rispettiva funzione di probabilità,  $\pi_k$ .

### 3.1.2. *Modello GoM*

Per mostrare che il modello GoM può essere pensato come un caso particolare del modello a classi latenti, di seguito viene sviluppato utilizzando il concetto di appartenenza parziale.

Come affermato in precedenza, il concetto di piena appartenenza è una generalizzazione di associazione parziale considerata per i modelli a classi latenti.

Sia  $g = (g_1, g_2, \dots, g_K)$  il vettore delle associazioni parziali di ogni soggetto. Si assume che le componenti  $g_1, g_2, \dots, g_K$ , dette punteggi di associazione, siano variabili casuali non negative che sommano a 1. Si indichi, poi, con  $D(g)$  la distribuzione di tale vettore  $g$ .

L'assunzione principale del modello GoM è la convessità delle probabilità di risposta condizionate, come visto nel paragrafo 2.1 (pag. 13).

Dati i punteggi GoM,  $g_k$ , la distribuzione condizionata della variabile manifesta  $x_j$  è data dalla combinazione convessa delle probabilità delle “piene” associazioni, cioè:

$$\Pr(x_j = l_j | g) = \sum_{k=1}^K g_k \cdot \Pr(x_j = l_j | g_k = 1), \quad (3.1.4)$$

In analogia con la terminologia utilizzata nel paragrafo precedente, si indichi la probabilità condizionata del profilo estremo come

$$\lambda_{kjl_j} = \Pr(x_j = l_j | g_k = 1), \quad k = 1, \dots, K, \quad j = 1, \dots, J, \quad l_j = 1, \dots, L_j. \quad (3.1.5)$$

Queste probabilità sono i parametri strutturali del modello, comuni a tutti gli individui.

Con questa notazione, è possibile riscrivere l’equazione (3.1.4) nel seguente modo:

$$\Pr(x_j = l_j | g) = \sum_{k=1}^K g_k \cdot \lambda_{kjl_j}, \quad j = 1, \dots, J, \quad l_j = 1, \dots, L_j. \quad (3.1.6)$$

L’assunzione di indipendenza locale (2 a pag. 13) afferma che le variabili manifeste  $x_j$  sono indipendenti condizionatamente alle variabili latenti. Allora, la probabilità condizionata di osservare un insieme di risposte  $l = (l_1, \dots, l_J)$  è la seguente:

$$f^{GoM}(l | g) = \Pr(x = l | g) = \prod_{j=1}^J \Pr(x_j = l_j | g) = \prod_{j=1}^J \left( \sum_{k=1}^K g_k \cdot \lambda_{kjl_j} \right).$$

Infine, integrando rispetto le variabili latenti, cioè i punteggi di associazione  $g$ , si ottiene la distribuzione marginale delle variabili  $x$  come:

$$f^{GoM}(l) = \Pr(x = l) = \int f^{GoM}(l | g) dD(g) = \int \prod_{j=1}^J \left( \sum_{k=1}^K g_k \cdot \lambda_{kjl_j} \right) dD(g). \quad (3.1.7)$$

Si noti che la quantità  $l_j$  appare come indice della probabilità condizionata  $\lambda_{kjl_j}$  in entrambe le equazioni (3.1.1) e (3.1.5). Inoltre, è possibile osservare che, poiché nel vettore dei punteggi  $g$  ci può essere più di una componente che assume valore diverso da zero, non è possibile semplificare l'integrale dell'equazione (3.1.7) attraverso una sommatoria, come invece è stato possibile per l'equazione (3.1.3).

## 3.2. Interpretazione

Nel modello GoM, ognuna delle  $J$  probabilità di risposta marginale è una combinazione lineare delle probabilità di risposta per i  $K$  profili estremi,  $\lambda_{kjl}$ , pesate dai rispettivi punteggi GoM,  $g_k$ . Dati i punteggi di associazione  $g_k$ , l'interpretazione usuale del modello GoM sostiene che ogni individuo della popolazione è membro parziale di ogni profilo estremo [2].

Si consideri, per esempio, un questionario sullo stato di salute contenente  $J$  domande dicotomiche. Per  $K = 2$ , la stima dei profili estremi può essere interpretata come “soggetti in salute” e “soggetti non in salute”. Un punteggio di associazione per il secondo profilo estremo,  $g_2$ , mostra quanto “non in salute” è l'individuo preso in considerazione in relazione al profilo estremo “soggetti non in salute”.

Nel modello a classi latenti, invece, una popolazione può essere suddivisa in un determinato numero di classi latenti. Ogni soggetto è considerato pieno membro di una delle classi latenti in un determinato istante nel tempo. I punteggi di associazione  $g_k$  possono essere interpretati come la “proporzione” di domande a cui quel dato individuo ha risposto come se fosse pieno membro della  $k$ -esima classe. Considerando l'esempio precedente, un punteggio di associazione per la seconda classe latente,  $g_2=1/3$ , significa che un soggetto risponde ad un terzo delle domande come un “soggetto non in salute”, e a due terzi come un “soggetto in salute”.

## *Capitolo 4*

# **APPLICAZIONE DEL MODELLO GoM A DATI REALI**

Quando si tratta il tema della salute, è opportuno fare riferimento alla Costituzione dell'Organizzazione Mondiale della Sanità (OMS), agenzia dell'ONU (Organizzazione delle Nazioni Unite) istituita il 7 aprile 1948, con l'obiettivo di operare per far raggiungere a tutte le popolazioni il livello di salute più elevato possibile. La salute, definita in tale Costituzione come *“stato di completo benessere fisico, psichico e sociale, e non semplice assenza di malattia”*, è considerata un diritto, e come tale si pone alla base di tutti gli altri diritti fondamentali che spettano alle persone. Ovviamente è molto difficile stabilire quando un individuo si trova in uno stato di “completo benessere”.

Questo principio assegna agli Stati compiti che vanno ben al di là della semplice gestione di un sistema sanitario. Essi dovrebbero farsi carico di individuare e modificare quei fattori che influiscono negativamente sulla salute collettiva, promuovendo al contempo quelli favorevoli.

In tale contesto, la salute viene considerata più un mezzo che un fine, e può essere definita come una risorsa di vita quotidiana che consente alle persone di condurre una vita produttiva a livello individuale, sociale ed economico.

Nel tempo, sul concetto di salute e sulla sua definizione, si è sviluppato un dibattito internazionale e sono state formulate alcune definizioni alternative. Fino ad ora, però, hanno avuto poco successo e quindi la definizione dell'OMS rimane ancora un punto di partenza e di riferimento.



## 4.1. World Health Survey

“*World Health Survey*” (WHS) [9] è un’indagine, intrapresa dall’Organizzazione Mondiale della Sanità nel 2002, volta a misurare lo stato di salute della popolazione del mondo. L’obiettivo principale di questo studio è fornire dati empirici sullo stato di salute nazionale, in modo da avere un migliore monitoraggio dei sistemi sanitari nazionali. In particolare, alcuni degli obiettivi generali di questa indagine sono:

- i. Esaminare il modo in cui la popolazione percepisce il proprio stato di salute;
- ii. Investigare alcune aree di interesse, quali la mortalità, la natalità, la copertura degli interventi sanitari, ecc.;
- iii. Formulare una strategia che permetta a ciascun Paese di essere in grado di condurre un’indagine di questo tipo e di sviluppare delle piattaforme di lungo termine che permettano la condivisione di questo tipo di informazioni;
- iv. Offrire una piattaforma dinamica di raccolta dati che sia trasparente ed in continua evoluzione, consentendo alla comunità internazionale la condivisione dei dati come un bene di pubblico dominio.

### ***4.1.1. Disegno di campionamento***

L'indagine WHS è stata effettuata in 74 Paesi del mondo, per ognuno dei quali è stato stilato un piano di campionamento dettagliato.

La popolazione di riferimento include tutti gli adulti, uomini e donne, d'età superiore o uguale ai 18 anni, che vivono in famiglie. Coloro che vivono in gruppi, in caserme militari o in altre sistemazioni diverse dalla famiglia non sono eleggibili per lo studio in questione. Persone che vivono in un'istituzione a causa di problemi di salute (ad esempio un ospedale, una casa di cura od un alloggio per anziani), al momento della visita alla famiglia, tali soggetti vengono intervistati nell'istituto dove si trovano oppure il questionario viene somministrato al loro ritorno in famiglia, se avviene entro un periodo di tempo di due settimane dopo la prima visita.

Il metodo di campionamento utilizzato in questa indagine è di tipo probabilistico. Significa che ogni singolo individuo della popolazione ha la stessa probabilità, diversa da zero, di entrare a far parte del campione. In particolare, per questo tipo di studio è stato impiegato il campionamento stratificato a più stadi, senza sostituzione delle unità.

Il campionamento stratificato è una delle tecniche più famose ed usate al giorno d'oggi. Consiste nel suddividere gli individui della popolazione in sottopopolazioni, *strati*, sulla base di una caratteristica comune, nell'estrarre poi un campione casuale semplice da ogni strato in modo indipendente, ed, infine, nel riunire insieme i risultati dei singoli campionamenti in modo da formare un unico campione di ampiezza predefinita.

Per il WHS, la dimensione campionaria è stata definita pari a circa 5000 soggetti per Paese. La stratificazione è stata applicata prima di effettuare il campionamento a stadi. In particolare, tutte le unità della popolazione sono state stratificate in base alle seguenti tre caratteristiche:

- i. Area geografica (Nord/Centro/Sud),
- ii. Stato socio-economico (Alto/Basso),
- iii. Presenza di centri di assistenza, *Health Care facility* (Si/No).

Il risultato è stato il seguente:

**Tabella 4.1.1.** Definizione degli strati.

<b>Strato</b>	<b>Area geografica</b>	<b>Stato socio-economico</b>	<b>Centri assistenza</b>
1	Nord	Alto	Si
2	Nord	Alto	No
3	Nord	Basso	Si
4	Nord	Basso	No
5	Centro	Alto	Si
6	Centro	Alto	No
7	Centro	Basso	Si
8	Centro	Basso	No
9	Sud	Alto	Si
10	Sud	Alto	No
11	Sud	Basso	Si
12	Sud	Basso	No

Per quanto riguarda il campionamento a stadi, in questa indagine sono stati utilizzati i seguenti 4 stadi:

1° stadio: all'interno di ogni strato sono stati estratti 100 comuni,

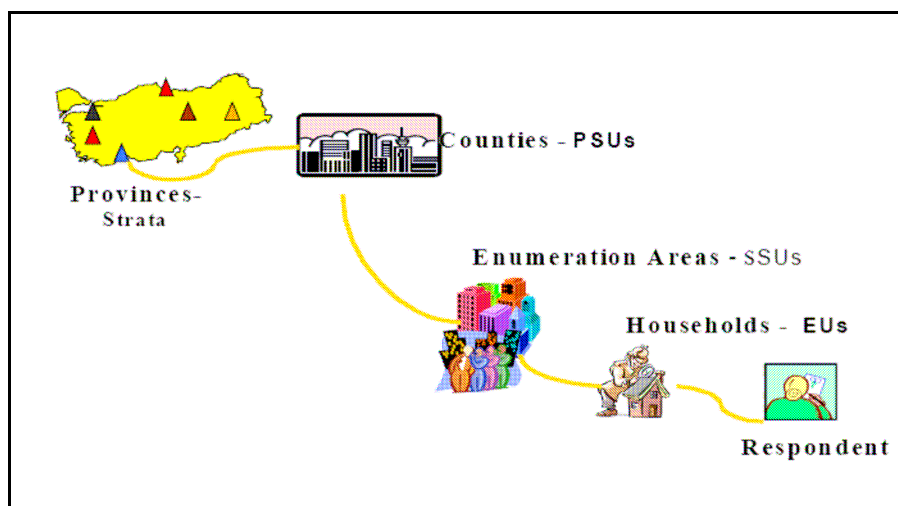
2° stadio: i 100 comuni selezionati sono stati divisi in sezioni elettorali mediante campionamento sistematico<sup>7</sup>,

3° stadio: da ogni comune sono state selezionate 50 famiglie (100 comuni \* 50 famiglie = 5000), sempre attraverso campionamento sistematico,

4° stadio: è stato somministrato il questionario ad un membro della famiglia selezionata, estratto in modo casuale tra gli eleggibili.

La figura sottostante aiuta a capire meglio lo schema di campionamento appena descritto.

**Figura 4.1.1.** Rappresentazione del campionamento utilizzato nell'indagine WHS.



<sup>7</sup> Data una popolazione di  $N$  individui e fissata la dimensione  $n$  del campione, si calcola il quoziente intero  $R = \frac{N}{n}$ . Si sceglie un numero  $k$  a caso (ad esempio da un'urna) compreso tra 1 e  $R$ . Si includono nel campione gli individui della lista che occupano i posti  $k, k+R, k+2R, \dots$

La prima vignetta di figura 4.1.1 riguarda il processo di stratificazione, e quindi alla creazione dei 12 strati, come evidenziato in tabella 4.1.1. Il termine PSU<sub>S</sub> (*Primary Sampling Units*) indica le unità di primo stadio, cioè i 100 comuni. SSU<sub>S</sub> (*Secondary Sampling Units*), invece, individua le unità di secondo stadio, ossia le sezioni elettorali in cui è stato suddiviso ogni comune. EU<sub>S</sub> (*Elementary Units*) rappresenta le unità di terzo stadio, cioè le 50 famiglie estratte da ogni comune. Infine, l'ultima vignetta (*Respondent*) indica i soggetti a cui è stato somministrato il questionario, cioè le unità di quarto stadio.

#### **4.1.2. Il questionario**

In questo studio sono state utilizzate due tipologie di questionari. Il primo è il questionario familiare (*Household Questionnaire*), impiegato per descrivere le caratteristiche della famiglia intervistata. Nel dettaglio include le seguenti sezioni:

- Informazioni sulla *struttura* della famiglia (età, educazione, occupazione, ecc.);
- Informazioni sull'*assistenza sanitaria*;
- Informazioni sulla *copertura* del sistema sanitario;
- Informazioni su un'eventuale assicurazione;
- Informazioni sulla *spesa familiare* per la salute;
- Informazioni su eventuali componenti della famiglia che *lavorano in ambito sanitario* (ad esempio: medici, infermieri, ecc.).

Il secondo questionario utilizzato, invece, è quello individuale (*Individual Questionnaire*) e comprende le seguenti sezioni:

- *Informazioni socio-demografiche* sul soggetto rispondente (età, sesso, educazione, stato civile, occupazione, ecc.);
- *Descrizione dello stato di salute* del soggetto;
- *Valutazione dello stato di salute* del soggetto;
- Informazioni su alcuni *fattori di rischio* (ad esempio: fumo, alcool, ecc.);
- Informazioni sulla *mortalità*<sup>8</sup>;
- Informazioni sulla *copertura di interventi sanitari*<sup>9</sup>;
- Informazioni sulla *sensibilità del sistema sanitario*.

Questo secondo questionario è stato realizzato in due tipologie: la versione estesa (*Long Questionnaire*) e quella ridotta (*Short Questionnaire*).

---

<sup>8</sup> Questa sezione è stata progettata per integrare le informazioni ottenute dai registri dell'anagrafe.

<sup>9</sup> Questa sezione è stata inserita per identificare i soggetti con diverse condizioni di salute che richiedono interventi specifici.

## 4.2. Analisi preliminare dei dati

In questa sezione vengono presentati alcuni risultati relativi all'analisi dei dati raccolti nell'indagine WHS.

**Tabella 4.2.1.** Elenco dei 74 paesi partecipanti all'indagine WHS.

AFRO	AMRO	EURO		EMRO	SEARO	WPRO
		Eurobarometer	Other EURO Countries			
Burkina Faso	Brazil	Austria	Bosnia	Morocco	Bangladesh	Australia
Chad	Chile	Belgium	Croatia	Pakistan	India	Japan
Comoros	Dominican Rep.	Denmark	Czech Rep.	Tunisia	Myanmar	China
Congo	Ecuador	Finland	Estonia	UAE	Nepal	Lao (PDR)
Côte D'Ivoire	Guatemala	France	Georgia		Sri Lanka	Malaysia
Ethiopia	Mexico	Germany	Hungary			Philippines
Ghana	Paraguay	Greece	Israel			Vietnam
Kenya	Uruguay	Ireland	Kazakhstan			
Malawi		Italy	Latvia			
Mali		Luxembourg	Norway			
Mauritania		Netherlands	Romania			
Mauritius		Portugal	Russian Fed.			
Namibia		Sweden	Slovakia			
Senegal		United Kingdom	Slovenia			
South Africa			Spain			
Swaziland			Turkey			
Zambia			Ukraine			
Zimbabwe			Yugoslavia			
18	8	14	18	4	5	7

Come si può notare dalla tabella 4.2.1, i Paesi partecipanti allo studio sono 74, ma, come si vedrà nel seguito, ne sono stati analizzati 70 perché i restanti 4 sono risultati o non disponibili o danneggiati. In particolare, non sono stati inclusi nelle analisi seguenti il Cile, la Romania, la Jugoslavia ed il Giappone.

Nello specifico, su 70 Paesi analizzati, 66 hanno utilizzato l'intervista faccia a faccia (52 con questionario lungo e 14 con questionario corto), 1 Paese (Ungheria) l'intervista CAPI (*Computer Assisted Personal Interviewing*, intervista personale con l'assistenza

di un Personal Computer), e 4 Paesi l'intervista CATI (*Computer Assisted Telephone Interviewing*, intervista telefonica con l'assistenza del Personal Computer).

Se si osservano queste quantità si può notare che dalla somma  $66+1+4$  si ottiene 71 e non 70 come era stato affermato in precedenza. Questo accade perché l'Australia ha somministrato il questionario ridotto sia mediante intervista faccia a faccia che intervista telefonica, quindi per questo Paese ci sono due distinti dataset.

Per costruire il dataset finale sono state prese in considerazione solo alcune informazioni ricavate dalla somministrazione del questionario individuale (vedi Appendice, pag. 95). In totale il dataset contiene 308214 dati, compresi 90740 record con osservazioni mancanti in prossimità di alcune variabili (di cui 25879 aventi i dati mancanti). Quindi, alla fine, il dataset utilizzato per le analisi contiene 217474 individui.

Le variabili prese in considerazione per l'analisi sono 18, tutte categoriali, suddivise nel seguente modo:

- Variabili socio-demografiche (*Respondent's Socio Demographic Characteristics*):
  1. sesso (*sex*);
  2. età (*age*);
  3. stato civile (*marital status*);
  4. educazione (*education*);
  5. occupazione (*current job*);



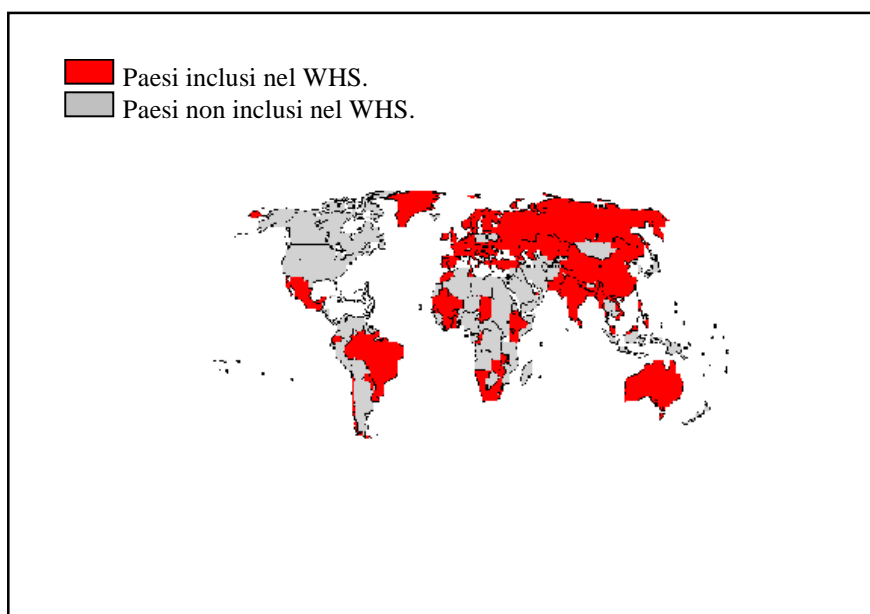
- Variabili sullo stato di salute (*Health State Descriptions*):
  6. stato di salute generale (*health status*);
  7. difficoltà di movimento (*difficulty moving around*);
  8. difficoltà nella cura di sé (*difficulty in self-care*);
  9. dolori e disagi in generale (*pain and discomfort*);
  10. difficoltà di concentrazione (*difficulty in concentration*);
  11. difficoltà nelle relazioni interpersonali (*difficulty in personal relationships*);
  12. difficoltà nel vedere e riconoscere le persone (*difficulty in seeing and recognizing persons*);
  13. problemi nel dormire (*sleeping disorders*);
  14. sentirsi tristi e depressi (*feeling sad or depressed*);
  
- Variabili sulle patologie mediche diagnosticate (*Coverage*):
  15. diagnosi di artrite (*diagnosis of Arthritis*);
  16. diagnosi di angina pectoris (*diagnosis of Angina Pectoris*);
  17. diagnosi di asma (*diagnosis of Asthma*);
  18. diagnosi di depressione (*diagnosis of Depression*).

Per quanto riguarda l'analisi preliminare dei dati a disposizione, si è pensato di creare una tabella contenente i rapporti di prevalenza di ogni variabile rispettivamente per ogni area regionale.

A tale proposito i 70 Paesi sono stati raggruppati in sei aree, secondo la classificazione di tabella 4.2.1. Queste aree sono state indicate attraverso le seguenti sigle: AFRO (*African Region Office*), che indica i Paesi dell'Africa, AMRO (*American Region Office*), che individua i Paesi dell'America, EURO (*European Region Office*), che rappresenta i Paesi dell'Europa, EMRO (*Eastern Mediterranean Region Office*), che evidenzia i Paesi del Mediterraneo Orientale, SEARO (*South-East Asian Region Office*), che indica i Paesi del Sud-Est Asiatico, e WPRO (*Western Pacific Region Office*), che identifica i Paesi del Pacifico Occidentale.

La seguente figura può aiutare a capire meglio quali e quanti Paesi sono stati inclusi nell'indagine.

**Figura 4.2.2.** Rappresentazione dei 74 Paesi inclusi nell'indagine WHS







	Extreme/Cannot do	0.95	0.71	0.50	0.49	2.31	0.23	
Difficulty in seeing and recognizing persons	None	79.29	81.56	81.95	77.88	77.27	74.32	< 0.0001
	Mild	9.29	8.12	9.96	10.89	11.21	12.35	
	Moderate	6.82	5.75	5.15	6.52	4.83	9.70	
	Severe	3.30	3.46	2.10	3.35	4.78	3.07	
Sleeping disorders	Extreme/Cannot do	1.30	1.10	0.84	1.36	1.91	0.55	< 0.0001
	None	63.98	65.87	71.18	54.50	60.95	62.31	
	Mild	18.43	14.23	16.41	21.37	18.96	19.88	
	Moderate	10.71	11.53	7.63	14.70	10.50	14.16	
	Severe	5.74	6.47	4.06	7.68	7.03	3.37	
	Extreme	1.14	1.90	0.71	1.75	2.56	0.28	

**Tabella 4.2.2.** (Continua)

variabili		AFRO	AMRO	EMRO	EURO	SEARO	WPRO	$\chi^2$
		(N= 54094)	(N= 45157)	(N= 16692)	(N= 40045)	(N= 32016)	(N= 29470)	
Feeling sad or depressed	None	58.24	56.18	66.05	53.91	51.45	70.30	< 0.0001
	Mild	22.11	20.58	19.31	23.46	24.93	16.89	
	Moderate	11.89	13.42	9.21	14.42	12.70	10.39	
	Severe	6.24	7.18	4.65	6.29	8.14	2.09	
	Extreme	1.52	2.65	0.77	1.93	2.78	0.33	
Diagnosis of Arthritis	Yes	14.70	7.48	14.21	14.38	14.35	9.85	< 0.0001
	No	85.30	92.52	85.79	85.62	85.65	90.15	
Diagnosis of Angina Pectoris	Yes	8.92	4.59	3.27	11.08	6.51	4.66	< 0.0001
	No	91.08	95.41	96.73	88.92	93.49	95.34	
Diagnosis of Asthma	Yes	3.87	8.32	4.04	6.27	4.85	4.68	< 0.0001
	No	96.13	91.68	95.96	93.73	95.15	95.32	
Diagnosis of Depression	Yes	4.33	12.95	2.71	8.57	7.90	1.96	< 0.0001
	No	95.67	87.05	97.29	91.43	92.10	98.04	



La tabella 4.2.2 riporta i rapporti di prevalenza di ogni variabile pesati con i rispettivi pesi di campionamento.

Questi pesi di campionamento sono necessari per poter estendere il calcolo delle stime campionarie alla popolazione totale. Tali pesi sono stati calcolati nel modo seguente

$$\text{peso} = \frac{1}{\text{probabilità di entrare nel campione}}$$

e dipendono dal disegno di campionamento attuato da ciascun Paese. Per ulteriori chiarimenti si veda il seguente esempio.

**Tabella 4.2.3.** Esempio di calcolo dei pesi campionari.

<b>Pesi</b>		
Soggetto 1	$1 / 0.08^{10} =$	12.5
Soggetto 2	$1 / 0.11 =$	9.09
Soggetto 3	$1 / 0.06 =$	16.67
Soggetto 4	$1 / 0.09 =$	11.11
...	...	...
...	...	...
Soggetto I	$1 / 0.03 =$	33.33

I pesi di tabella 4.2.3 descrivono il modo in cui ogni soggetto del campione “pesa” a livello di popolazione. In altre parole, il soggetto 1 rappresenta 12.5 persone della popolazione, il soggetto 2 ne rappresenta 9.09, e così via.

---

<sup>10</sup> Le probabilità di entrare a far parte del campione sono calcolate come

$$\frac{\text{dimensione campionaria}}{\text{dimensione popolazione}}$$

La prevalenza è esprimibile come la proporzione di soggetti che in un determinato istante nel tempo presentano una data patologia [10], e si calcola nel seguente modo:

$$\frac{\text{n. totale di casi di malattia in un dato periodo di tempo}}{\text{popolazione totale}} .$$

Quando il rapporto di prevalenza è calcolato su soggetti non pesati esprime la proporzione di individui del campione che presentano la patologia considerata; mentre se è calcolato su soggetti pesati, allora diventa un indice estendibile a tutta la popolazione di riferimento.

In aggiunta alle prevalenze, è stato calcolato anche il test Chi-Quadrato di Pearson ( $\chi^2$ ), in modo da verificare l'ipotesi di dipendenza tra ogni variabile considerata e le aree regionali. In pratica questo test verifica il seguente sistema d'ipotesi:

$$\begin{cases} H_0 : \text{c'è indipendenza} \\ H_1 : \text{non c'è indipendenza} \end{cases}$$

I valori riportati nella tabella 4.2.2 sono i valori  $p$  ( $p$ -value) del test  $\chi^2$ . Se il valore  $p$  è minore di un valore critico definito (solitamente 0.05), allora si rifiuta l'ipotesi nulla e quindi si conclude che c'è associazione tra le variabili considerate e le aree regionali; mentre se il valore  $p$  è maggiore del valore critico, allora l'ipotesi nulla viene accettata. Osservando i valori contenuti in tabella 4.2.2 è immediato notare che tutti i  $\chi^2$  risultano minori di 0.05; quindi è possibile concludere che le prevalenze delle variabili considerate variano da regione a regione.



Proseguendo con l'analisi di tabella 4.2.2, è possibile notare che per la regione Africa il dataset contiene 54094 soggetti, per l'America 45157, per il Mediterraneo Orientale 16692, per l'Europa 40045 individui, per il Sud-Est Asiatico 32016 ed infine per il Pacifico Occidentale 29470.

Considerando, ora, le prevalenze delle cinque *variabili socio-demografiche*, è possibile notare che, per quanto riguarda il sesso, quattro aree su sei hanno in maggioranza donne, mentre solo il Sud-Est Asiatico ed i Paesi del Mediterraneo Orientale hanno più uomini che donne. In generale, tuttavia, per ogni regione la suddivisione tra maschi e femmine è pressoché simmetrica.

Se si osserva, invece, la variabile età, si può notare che, in tutte e sei le regioni, la prevalenza più alta, compresa tra il 19.49% dell'Europa ed il 35.88% dell'Africa, si ha sempre in corrispondenza della fascia d'età 20-29. L'area regionale con più soggetti d'età compresa tra i 18 ed i 19 anni è quella africana (prevalenza del 8.28%), mentre la regione con più soggetti di età superiore a 70 anni è l'Europa (prevalenza del 13.15%).

Considerando la variabile stato civile è possibile notare che in tutte le regioni, la maggior parte dei soggetti risulta essere sposata al momento dell'intervista. In questo caso le prevalenze sono comprese tra il 72.92% del Sud-Est Asiatico ed il 47.50% dell'America.

Per la variabile educazione la situazione è diversa, nel senso che l'Africa, il Sud-Est Asiatico e il Mediterraneo Orientale hanno il maggior numero soggetti non scolarizzati.

L'America e il Pacifico Occidentale hanno più soggetti con scuola secondaria (scuola media dell'obbligo) completata. Al contrario, per l'Europa, la maggior parte del campione ha scolarità elevata perché contiene molti soggetti con scuola superiore completata. La regione avente il più basso tasso d'istruzione (*no formal schooling*) è il Mediterraneo Orientale (prevalenza del 52.91%), mentre quella con il tasso più alto (università completata) è l'Europa (prevalenza del 21.80%).

Infine, se si osserva la variabile occupazione è possibile notare che, su sei aree regionali, quattro (Africa, America, Mediterraneo Orientale, Europa) hanno un maggior numero di soggetti non lavoratori, le cui prevalenze variano tra il 31.27% ed il 51.96%; mentre in sole due regioni (Sud-Est Asiatico e Pacifico Occidentale) la maggior parte dei soggetti risulta essere lavoratore in proprio.

Considerando, ora, le successive nove *variabili sui problemi di salute* (stato di salute generale, difficoltà di movimento, difficoltà nella cura di sé, dolori e disagi in generale, difficoltà di concentrazione, difficoltà nelle relazioni interpersonali, difficoltà nel vedere e riconoscere le persone, problemi nel dormire, sentirti tristi e depressi), si può subito notare che, alla domanda sulle condizioni di salute generali (*health status*), più della metà dei soggetti ha risposto di sentirsi bene o addirittura molto bene. In questo caso le prevalenze variano tra il 53.14% dell'Europa ed il 71.61% del Mediterraneo Orientale. Il continente con il maggior numero di soggetti non in salute è il Sud-Est Asiatico con una prevalenza del 2.11%.

In tutte le restanti otto variabili le prevalenze seguono un andamento pressoché positivo, nel senso che la maggior parte dei soggetti intervistati dichiara di non soffrire dei sopraindicati disturbi.

In particolare, per quasi tutte queste variabili più del 50% dei soggetti riporta di non avere particolari problemi di salute, ad eccezione della domanda relativa alla presenza di dolori (*pain and discomfort*), nella quale le prevalenze variano dal 39.65% del Sud-Est Asiatico al 54.39% dell’Africa.

Considerando, infine, le ultime quattro domande riguardanti le *patologie mediche diagnosticate* (artrite, angina pectoris, asma e depressione), in generale, si ha che oltre l’85% del campione dichiara di non avere avuto nessuna diagnosi per queste malattie.

In particolare, osservando soltanto le prevalenze dei soggetti a cui è stata diagnosticata la patologia, è possibile notare che, per quanto riguarda l’artrite, la regione con prevalenza maggiore è l’Africa (prevalenza del 14.70%). Analogamente per le altre tre patologie si nota che, per l’angina pectoris la prevalenza maggiore si ha per l’Europa (prevalenza del 11.08%) mentre per l’asma e la depressione l’area regionale avente prevalenza maggiore è l’America (prevalenze rispettivamente del 8.32% e del 12.95%).

Concludendo, in base a queste ultime considerazioni, è possibile affermare che, la regione africana ha un maggior numero di soggetti che dichiarano di soffrire di artrite rispetto alle altre regioni, l’Europa ha più individui che affermano di avere l’angina pectoris, mentre l’America ha più soggetti che riportano di soffrire sia d’asma che di depressione.

### 4.3. Stima del modello GoM

In questa sezione viene presentata la stima dei parametri del modello GoM ottenuta mediante il programma DSIGoM, come illustrato nel paragrafo 2.2.1. Questo software è il frutto di 25 anni di teorie ed applicazioni sviluppate dal Dr. Max A. Woodbury e dai suoi colleghi, specialmente dal Dr. Kenneth G. Manton [11], docente presso il Centro di Studi Demografici dell'Università di Duke nel North Carolina (Center of Demographic Studies, Duke University, NC).

Il dataset finale è stato scomposto in sei dataset di minore dimensione, in base alla suddivisione delle aree specificata in tabella 4.2.1. In pratica, è stato creato un dataset per ogni area regionale e ad ognuno di questi dataset è stata applicata l'analisi GoM stimando le probabilità di risposta con 2, 3, 4, 5 e 6 profili estremi. Le variabili incluse in questo studio sono:

- le 9 variabili sullo stato di salute dei soggetti (*Health State Description*);
- le 4 variabili sulle patologie mediche diagnosticate (*Coverage*).

Per ogni profilo estremo, DSIGoM calcola anche il valore del rapporto di log verosimiglianza, il valore del Chi-Quadrato ( $\chi^2$ ) ed i gradi di libertà. Prima di eseguire tale programma è importante decidere quali variabili sono da definire “interne” e quali “esterne”.

La differenza principale tra questi due tipi di variabili è che le interne vengono utilizzate per identificare i K profili estremi e per stimare le probabilità  $\lambda_{kjl}$ , mentre le esterne non vengono prese in considerazione per la stima [5].

Nel seguito, vengono proposti i risultati di queste analisi per ogni area regionale considerata.

### 4.3.1. Africa

Per quanto riguarda la regione Africa, il dataset creato contiene 56703 soggetti ed è privo di osservazioni mancanti.

L'analisi con il DSIGoM ha prodotto i seguenti risultati:

**Tabella 4.3.1.** Rapporto di log-verosimiglianza e  $\chi^2$  per 2, 3, 4, 5 e 6 profili estremi.

<b>Numero profili</b>	<b>Rapporto di log verosimiglianza</b>	<b><math>\chi^2</math></b>	<b><math>\Delta \chi^2</math></b>	<b>Gradi di libertà (d.f.)</b>	<b><math>\Delta</math> Gradi di libertà (d.f.)</b>
2	135285	270570		56756	
3	188427	376850	106280	113512	56756
4	214071	428140	51290	170268	56756
5	244231	488460	60320	227024	56756
6	261462	522920	34460	283780	56756

Il passo successivo consiste nello stabilire il numero di profili estremi da considerare. Quando, in inferenza, si devono confrontare due modelli nidificati, ad esempio A e B, rispettivamente con  $p_A$  e  $p_B$  parametri ( $p_A < p_B$ ), la perdita di bontà di adattamento del modello A rispetto a B può essere valutata attraverso il test del rapporto di log-verosimiglianza.

La statistica su cui è basato questo test è  $W=2\{ l(\hat{\mu}_A; y) - l(\hat{\mu}_B; y) \}$ <sup>11</sup>, la quale ha distribuzione asintotica Chi-Quadrato ( $\chi^2$ ) con  $p_B - p_A$  gradi di libertà sotto il modello A.

Per stabilire quanti profili estremi considerare, è stato calcolato il valore del  $\chi^2$  con gradi di libertà pari alla differenza dei gradi di libertà tra il modello con K+1 e quello con K profili<sup>12</sup>. In quest'ottica, il valore critico del  $\chi^2$ , a livello di significatività del 95%, è risultato pari a 57311 ( $\chi^2_{56756, 0.95}$ ). Per verificare se accettare o meno il modello considerato, questa quantità è stata confrontata con la differenza tra il valore del test del  $\chi^2$  del modello con K+1 e quello con K profili estremi (colonna  $\Delta \chi^2$  di tabella 4.3.1). In particolare, si è verificato il seguente sistema d'ipotesi:

$$\begin{cases} H_0 : \text{accetto il modello ridotto} \\ H_1 : \text{non accetto il modello ridotto} . \end{cases}$$

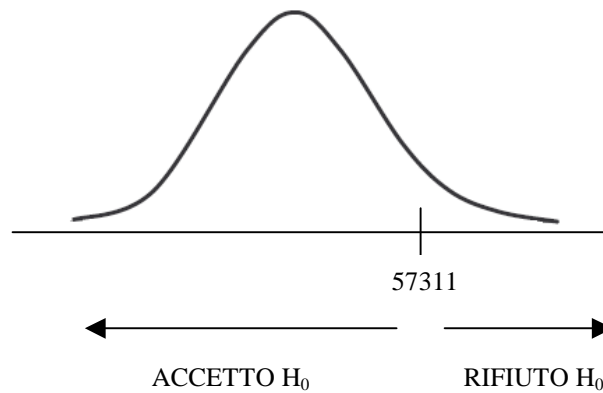
Più semplicemente, ad ogni iterazione si è confrontato il livello del test ( $\Delta \chi^2$ ) con il valore critico (in questo caso  $\chi^2_{56756, 0.95}$ ), secondo la figura 4.3.1 di pagina seguente.

---

<sup>11</sup> Dove  $\mu_A$  e  $\mu_B$  sono i parametri dei due modelli nidificati.

<sup>12</sup> **Teorema:** se  $X_1, \dots, X_n$  sono  $n$  variabili casuali  $\chi^2$  tra loro indipendenti, ciascuna con  $g_i$  gradi di libertà, allora la variabile casuale  $Y = X_1 + \dots + X_n$  è a sua volta una variabile casuale  $\chi^2$  con  $g$  gradi di libertà, ove  $g = g_1 + \dots + g_n$ .

**Figura 4.3.1.** Distribuzione  $\chi^2$  con 56756 gradi di libertà.



Osservando la figura 4.3.1 si può notare che la curva è simmetrica mentre la distribuzione  $\chi^2$  usuale possiede un'asimmetria verso sinistra molto accentuata. Questo accade perché una variabile casuale  $\chi^2$  con  $n$  molto grande (in questo caso  $n = 56703$ ) tende alla distribuzione Normale<sup>13</sup>.

Se il valore del test del  $\chi^2$  è maggiore del valore critico 57311 allora si rifiuta l'ipotesi  $H_0$ , pertanto non si accetta il modello ridotto e si prosegue con l'analisi dei successivi due modelli. Non appena il valore del test è inferiore al valore critico si accetta il modello ridotto e ci si ferma. Tale procedimento viene utilizzato da Manton et al. in [12].

---

<sup>13</sup> **Teorema:** una variabile casuale  $\chi^2$  con  $n$  molto grande ha distribuzione che tende ad essere Normale,  $N(g, 2g)$ , dove  $g$  rappresenta i gradi di libertà del  $\chi^2$ .

Nel caso dell'Africa, il ragionamento è stato il seguente:

- i. Inizialmente sono stati confrontati il modello con 2 e quello con 3 profili estremi. Si è constatato che il valore del test del  $\chi^2$  (primo elemento di colonna  $\Delta \chi^2$  in tabella 4.3.1) è maggiore del valore critico ( $106280 > 57311$ ), quindi si rifiuta l'ipotesi  $H_0$  e si procede con il successivo accostamento.
- ii. Comparando, poi, il modello con 3 e quello con 4 profili estremi, l'incremento di  $\chi^2$  è risultato minore del valore critico ( $51290 < 57311$ ), e si accetta pertanto l'ipotesi  $H_0$ .

Per tale motivo si è concluso che il modello con  $K=3$  profili estremi si adatta adeguatamente ai dati dell'Africa.

I risultati ottenuti sono riportati nella tabella sottostante.

La prima colonna contiene le frequenze di risposta per ogni modalità. Le restanti tre colonne mostrano i valori di  $\lambda_{kjl}$ , cioè le probabilità che l'individuo  $i$ -esimo risponda alla modalità  $l$ -esima nella domanda  $j$ -esima, condizionatamente al vettore punteggi GoM,  $g_i$ .

**Tabella 4.3.2.**  $\lambda_{kjl}$  per l'Africa con  $K=3$  profili estremi.

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
Health Status				
Very Good	30.00	54.92	0.00	0.00
Good	35.66	45.08	45.54	0.00
Moderate	24.85	0.00	54.46	55.12
Bad	8.11	0.00	0.00	38.32
Very Bad	1.39	0.00	0.00	6.56



**Tabella 4.3.2.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Difficulty moving around</b>				
None	68.94	100.00	0.00	0.00
Mild	15.21	0.00	100.00	0.00
Moderate	9.83	0.00	0.00	61.99
Severe	5.15	0.00	0.00	32.50
Extreme/Cannot do	0.87	0.00	0.00	5.51
<b>Difficulty in Self-Care</b>				
None	79.96	100.00	51.10	23.33
Mild	11.17	0.00	48.90	15.54
Moderate	6.06	0.00	0.00	41.76
Severe	2.29	0.00	0.00	15.74
Extreme/Cannot do	0.53	0.00	0.00	3.63
<b>Pain and discomfort</b>				
None	49.76	100.00	0.00	0.00
Mild	24.19	0.00	100.00	0.00
Moderate	15.89	0.00	0.00	60.98
Severe	8.94	0.00	0.00	34.31
Extreme	1.23	0.00	0.00	4.71
<b>Difficulty in concentration</b>				
None	63.88	100.00	0.00	0.00
Mild	19.29	0.00	100.00	0.00
Moderate	10.90	0.00	0.00	64.76
Severe	5.22	0.00	0.00	31.04
Extreme/Cannot do	0.71	0.00	0.00	4.20
<b>Difficulty in personal relationships</b>				
None	77.25	97.99	39.71	41.67
Mild	12.17	0.00	60.29	0.00
Moderate	6.35	0.00	0.00	39.82
Severe	2.21	0.00	0.00	13.89
Extreme/Cannot do	2.02	2.01	0.00	4.63
<b>Difficulty in seeing and recognizing persons</b>				
None	76.09	100.00	46.11	23.74
Mild	10.67	0.00	53.89	0.00
Moderate	7.94	0.00	0.00	45.75
Severe	4.04	0.00	0.00	23.26
Extreme/Cannot do	1.26	0.00	0.00	7.24

**Tabella 4.3.2.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Sleeping disorders</b>				
None	61.25	100.00	0.00	0.00
Mild	19.54	0.00	100.00	0.00
Moderate	11.71	0.00	0.00	60.96
Severe	6.54	0.00	0.00	34.06
Extreme	0.96	0.00	0.00	4.97
<b>Feeling sad or depressed</b>				
None	55.46	100.00	0.00	0.00
Mild	23.10	0.00	100.00	0.00
Moderate	12.97	0.00	0.00	60.48
Severe	7.08	0.00	0.00	33.03
Extreme	1.39	0.00	0.00	6.49
<b>Diagnosis of Arthritis</b>				
No	85.01	95.72	78.23	57.64
Yes	14.99	4.28	21.77	42.36
<b>Diagnosis of Angina Pectoris</b>				
No	91.19	100.00	89.21	66.74
Yes	8.81	0.00	10.79	33.26
<b>Diagnosis of Asthma</b>				
No	95.56	98.42	95.27	86.65
Yes	4.44	1.58	4.73	13.35
<b>Diagnosis of Depression</b>				
No	95.58	100.00	100.00	77.26
Yes	4.42	0.00	0.00	22.74

Ad ogni profilo estremo viene attribuita un'interpretazione in base ai valori di  $\lambda_{kjl}$  ottenuti. In particolare, si può facilmente notare che il profilo I contiene tutti i soggetti in buona salute, senza particolari disturbi e con carenza di patologie mediche. Questo profilo può essere denominato STATO DI SALUTE SODDISFACENTE.

Al contrario, il profilo II include tutti gli individui con uno stato di salute moderato, con deboli problemi di salute (in particolare difficoltà di movimento, sofferenza fisica, scarsità di concentrazione, problemi del sonno, tristezza o depressione) e con un aumento, rispetto al profilo precedente, di alcune patologie (artrite, angina pectoris e asma). Questo profilo estremo può essere indicato STATO DI SALUTE MODERATO CON LIEVI DISTURBI.

In fine, il profilo III racchiude tutti i soggetti con uno stato di salute insoddisfacente, con notevoli disturbi e con un incremento delle probabilità di soffrire di una qualche patologia (artrite 42.36%, angina pectoris 33.26% e depressione 22.74%). Questo profilo estremo può essere definito CATTIVO STATO DI SALUTE CON IMPORTANTI DISTURBI.

Da questa classificazione è possibile notare che in nessuno dei tre profili estremi appena definiti compaiono soggetti con un'elevata probabilità di soffrire di qualche patologia medica. Questo accade probabilmente perché nell'intero campione la frequenza (prima colonna di tabella 4.3.2) di soggetti con patologie diagnosticate non supera il 14.99% (le frequenze variano tra il 4.42% della depressione ed il 14.99% dell'artrite). Tuttavia è possibile osservare che, nel passaggio dal profilo I al profilo III, le probabilità di avere una qualche diagnosi pur aumentando non superano mai il 50%.

### 4.3.2. America

Per quanto concerne la regione America, il dataset prodotto contiene 64587 soggetti ed è sprovvisto di osservazioni mancanti.

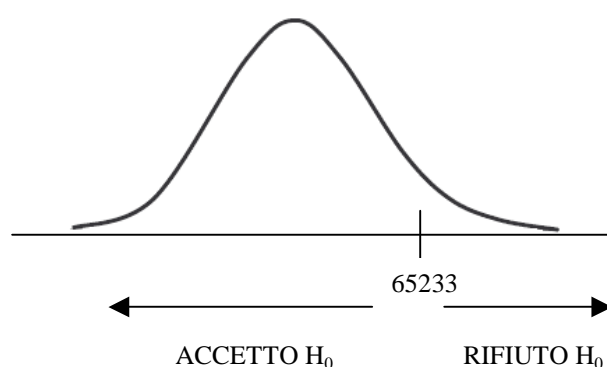
L'analisi con l'applicazione DSIGoM ha prodotto i seguenti valori:

**Tabella 4.3.3.** Rapporto di log-verosimiglianza e  $\chi^2$  per 2, 3, 4, 5 e 6 profili estremi.

Numero profili	Rapporto di log verosimiglianza	$\chi^2$	$\Delta \chi^2$	Gradi di libertà (d.f.)	$\Delta$ Gradi di libertà (d.f.)
2	119294	238590		64640	
3	170388	340780	102190	129280	64640
4	199715	399430	58650	193920	64640
5	228605	457210	57780	258560	64640
6	239513	479030	21820	323200	64640

Seguendo il medesimo procedimento approfondito nel paragrafo precedente, per stabilire quanti profili estremi considerare si è confrontato il livello del test ( $\Delta \chi^2$ ) con il valore critico ( $\chi^2_{64640, 0.95}$ ), secondo la figura 4.3.2.

**Figura 4.3.2.** Distribuzione  $\chi^2$  con 64640 gradi di libertà.



Il ragionamento seguito è stato il seguente:

- i. Inizialmente sono stati confrontati il modello con 2 e quello con 3 profili estremi. Si è constatato che il valore del test del  $\chi^2$  (primo elemento di colonna  $\Delta \chi^2$  in tabella 4.3.3) è maggiore del valore critico ( $102190 > 65233$ ), quindi si rifiuta l'ipotesi  $H_0$  e si procede con il passo successivo.
- ii. Comparando, poi, il modello con 3 e quello con 4 profili estremi, l'incremento di  $\chi^2$  è minore del valore critico ( $58650 < 65233$ ), e si accetta pertanto l'ipotesi  $H_0$ .

In tal caso si è concluso che il modello con  $K=3$  profili estremi si adatta adeguatamente ai dati dell'America.

I risultati ottenuti sono riportati nella tabella seguente.

**Tabella 4.3.4.**  $\lambda_{kjl}$  per l'America con  $K=3$  profili estremi.

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Health Status</b>				
Very Good	18.96	33.63	0.00	0.00
Good	44.46	66.37	29.79	0.00
Moderate	30.14	0.00	70.21	67.73
Bad	5.74	0.00	0.00	28.75
Very Bad	0.70	0.00	0.00	3.52
<b>Difficulty moving around</b>				
None	78.17	100.00	48.38	18.40
Mild	11.26	0.00	51.62	0.00
Moderate	6.86	0.00	0.00	52.92
Severe	3.28	0.00	0.00	25.31
Extreme/Cannot do	0.44	0.00	0.00	3.37

**Tabella 4.3.4.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Difficulty in Self-Care</b>				
None	90.84	100.00	80.49	65.31
Mild	5.45	0.00	19.51	9.65
Moderate	2.47	0.00	0.00	16.66
Severe	1.02	0.00	0.00	6.85
Extreme/Cannot do	0.23	0.00	0.00	1.53
<b>Pain and discomfort</b>				
None	52.62	100.00	0.00	0.00
Mild	22.92	0.00	100.00	0.00
Moderate	15.52	0.00	0.00	63.45
Severe	8.17	0.00	0.00	33.39
Extreme/Cannot do	0.77	0.00	0.00	3.16
<b>Difficulty in concentration</b>				
None	69.12	100.00	0.00	0.00
Mild	17.31	0.00	100.00	0.00
Moderate	9.24	0.00	0.00	68.14
Severe	3.98	0.00	0.00	29.35
Extreme/Cannot do	0.34	0.00	0.00	2.51
<b>Difficulty in personal relationships</b>				
None	88.71	100.00	71.55	63.38
Mild	6.69	0.00	28.45	6.63
Moderate	3.17	0.00	0.00	20.64
Severe	1.16	0.00	0.00	7.52
Extreme/Cannot do	0.28	0.00	0.00	1.83
<b>Difficulty in seeing and recognizing persons</b>				
None	78.68	100.00	51.08	36.24
Mild	10.52	0.00	48.92	0.00
Moderate	6.48	0.00	0.00	38.25
Severe	3.64	0.00	0.00	21.52
Extreme/Cannot do	0.68	0.00	0.00	4.00
<b>Sleeping disorders</b>				
None	69.98	100.00	25.80	0.00
Mild	15.28	0.00	74.20	0.00
Moderate	9.36	0.00	0.00	63.50
Severe	4.80	0.00	0.00	32.56
Extreme	0.58	0.00	0.00	3.94

**Tabella 4.3.4.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
Feeling sad or depressed				
None	61.39	100.00	0.00	0.00
Mild	20.79	0.00	100.00	0.00
Moderate	11.33	0.00	0.00	63.57
Severe	5.59	0.00	0.00	31.38
Extreme	0.90	0.00	0.00	5.05
Diagnosis of Arthritis				
No	93.77	100.00	100.00	64.60
Yes	6.23	0.00	0.00	35.40
Diagnosis of Angina Pectoris				
No	96.45	100.00	100.00	79.44
Yes	3.55	0.00	0.00	20.56
Diagnosis of Asthma				
No	95.47	98.29	94.75	86.03
Yes	4.53	1.71	5.25	13.97
Diagnosis of Depression				
No	93.18	100.00	100.00	61.63
Yes	6.82	0.00	0.00	38.37

Analizzando i tre profili estremi di tabella 4.3.4 si può facilmente notare che il profilo I contiene tutti i soggetti in buona salute, senza particolari disturbi e in assenza di patologie mediche diagnosticate. Questo profilo può essere denominato STATO DI SALUTE OTTIMO.

Viceversa, il profilo II include tutti gli individui con uno stato di salute moderato, con qualche disturbo (in particolare sofferenza fisica, scarsità di concentrazione, problemi del sonno, tristezza o depressione) e sostanzialmente privi di patologie mediche diagnosticate (l'unica probabilità che aumenta in modo lieve è quella riguardante

l'asma). Questo profilo estremo può essere indicato STATO DI SALUTE MODERATO CON LIEVI DISTURBI.

In fine, il profilo III racchiude tutti i soggetti con uno stato di salute insoddisfacente, con notevoli disturbi e con aumenti delle probabilità di avere una diagnosi in tutte e quattro le patologie. Questo profilo estremo può essere definito STATO DI SALUTE INSODDISFACENTE CON FORTI DISTURBI.

Anche per questa regione, come per l'Africa, accade che in nessuno dei tre profili compaiono soggetti con un'elevata probabilità di soffrire di qualche patologia medica. Questo accade presumibilmente perché nell'intero campione la frequenza di soggetti con disturbi accertati non supera il 6.82% (le frequenze variano tra il 3.55% dell'angina pectoris ed il 6.82% della depressione). Tuttavia è possibile osservare che, nel passaggio dal profilo I al profilo III, le probabilità di avere una qualche diagnosi crescono ma non superano mai il 50%.

### ***4.3.3. Mediterraneo Orientale***

Per quanto riguarda la regione Mediterraneo Orientale, il dataset prodotto contiene 16823 soggetti ed è sprovvisto di osservazioni mancanti.

L'analisi con il programma DSIGoM ha prodotto le quantità contenute in tabella 4.3.5.

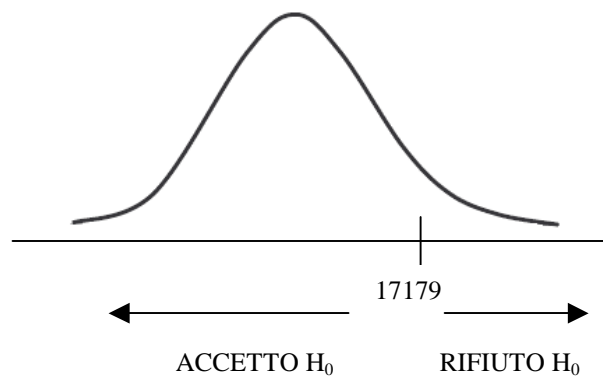


**Tabella 4.3.5.** Rapporto di log-verosimiglianza e  $\chi^2$  per 2, 3, 4, 5 e 6 profili estremi.

Numero profili	Rapporto di log verosimiglianza	$\chi^2$	$\Delta \chi^2$	Gradi di libertà (d.f.)	$\Delta$ Gradi di libertà (d.f.)
2	39285.2	78570		16876	
3	55340.9	110680	32110	33752	16876
4	63612.1	127220	16540	50628	16876
5	68993.4	137990	10770	67504	16876
6	75098.1	150200	12210	84380	16876

Seguendo il procedimento approfondito nei paragrafi precedenti, per decidere quanti profili estremi considerare si è confrontato il livello del test ( $\Delta \chi^2$ ) con il valore critico ( $\chi^2_{16876, 0.95}$ ), secondo la figura seguente.

**Figura 4.3.3.** Distribuzione  $\chi^2$  con 16876 gradi di libertà.



La procedura seguita è stata la seguente:

- i. Inizialmente sono stati confrontati il modello con 2 e quello con 3 profili estremi. Si è constatato che il valore del test del  $\chi^2$  (primo elemento di colonna  $\Delta \chi^2$  in tabella 4.3.5) è maggiore del valore critico ( $32110 > 17179$ ), quindi si rifiuta l'ipotesi  $H_0$  e si procede con il passo successivo.

- ii. Comparando, poi, il modello con 3 e quello con 4 profili estremi, l'incremento di  $\chi^2$  è minore del valore critico ( $16540 < 17179$ ), e si accetta pertanto l'ipotesi  $H_0$ .

In tal caso si è concluso che il modello con  $K=3$  profili estremi si adatta adeguatamente ai dati del Mediterraneo Orientale.

I risultati conseguiti sono riportati nella tabella seguente.

**Tabella 4.3.6.**  $\lambda_{kji}$  per il Mediterraneo Orientale con  $K=3$  profili estremi.

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Health Status</b>				
Very Good	23.60	43.39	0.00	0.00
Good	38.41	56.61	35.87	0.00
Moderate	26.70	0.00	64.13	53.72
Bad	8.62	0.00	0.00	35.33
Very Bad	2.67	0.00	0.00	10.94
<b>Difficulty moving around</b>				
None	71.35	100.00	0.00	0.00
Mild	14.22	0.00	100.00	0.00
Moderate	8.77	0.00	0.00	60.82
Severe	4.79	0.00	0.00	33.21
Extreme/Cannot do	0.86	0.00	0.00	5.97
<b>Difficulty in Self-Care</b>				
None	83.10	100.00	53.38	40.24
Mild	9.37	0.00	46.62	11.42
Moderate	4.27	0.00	0.00	27.40
Severe	2.48	0.00	0.00	15.91
Extreme /Cannot do	0.78	0.00	0.00	5.04
<b>Pain and discomfort</b>				
None	50.79	100.00	0.00	0.00
Mild	21.01	0.00	83.71	0.00
Moderate	14.99	0.00	16.29	45.19
Severe	10.43	0.00	0.00	43.25
Extreme	2.79	0.00	0.00	11.56

**Tabella 4.3.6.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Difficulty in concentration</b>				
None	62.81	100.00	0.00	0.00
Mild	17.19	0.00	100.00	0.00
Moderate	11.63	0.00	0.00	58.15
Severe	6.78	0.00	0.00	33.89
Extreme/Cannot do	1.59	0.00	0.00	7.97
<b>Difficulty in personal relationships</b>				
None	81.56	100.00	40.33	53.13
Mild	10.31	0.00	59.67	0.00
Moderate	5.03	0.00	0.00	28.99
Severe	2.26	0.00	0.00	13.06
Extreme/Cannot do	0.84	0.00	0.00	4.83
<b>Difficulty in seeing and recognizing persons</b>				
None	79.20	100.00	52.67	37.87
Mild	8.62	0.00	47.33	0.00
Moderate	6.57	0.00	0.00	33.50
Severe	4.12	0.00	0.00	21.01
Extreme/Cannot do	1.49	0.00	0.00	7.61
<b>Sleeping disorders</b>				
None	65.68	100.00	0.00	0.00
Mild	14.97	0.00	100.00	0.00
Moderate	10.86	0.00	0.00	56.13
Severe	7.17	0.00	0.00	37.05
Extreme	1.32	0.00	0.00	6.82
<b>Feeling sad or depressed</b>				
None	59.56	100.00	0.00	0.00
Mild	18.42	0.00	100.00	0.00
Moderate	11.72	0.00	0.00	53.24
Severe	8.29	0.00	0.00	37.63
Extreme	2.01	0.00	0.00	9.13
<b>Diagnosis of Arthritis</b>				
No	82.99	100.00	73.95	42.86
Yes	17.01	0.00	26.05	57.14

**Tabella 4.3.6.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
Diagnosis of Angina Pectoris				
No	96.36	100.00	100.00	82.01
Yes	3.64	0.00	0.00	17.99
Diagnosis of Asthma				
No	96.21	100.00	93.86	86.74
Yes	3.79	0.00	6.14	13.26
Diagnosis of Depression				
No	96.77	100.00	100.00	84.07
Yes	3.23	0.00	0.00	15.93

Analizzando i tre profili estremi di tabella 4.3.6, si può facilmente notare che il profilo I contiene tutti i soggetti in buona salute, senza particolari disturbi e in assenza di patologie mediche. Questo profilo può essere denominato **STATO DI SALUTE SODDISFACENTE**.

Viceversa, il profilo II include tutti gli individui con uno stato di salute moderato, con qualche disturbo (in particolare difficoltà di movimento, sofferenza fisica, scarsità di concentrazione, problemi del sonno, tristezza o depressione), e con un aumento, rispetto al profilo precedente, di alcune patologie (artrite e asma). Questo profilo estremo può essere indicato **STATO DI SALUTE MODERATO CON LIEVI DISTURBI**.

In fine, il profilo III racchiude tutti i soggetti con uno stato di salute insoddisfacente, con notevoli disturbi e con aumenti delle probabilità di avere una diagnosi in tutte e quattro le patologie, in particolare l'artrite, la cui probabilità supera il 50%. Questo profilo estremo può essere definito **STATO DI SALUTE PESSIMO CON FORTI DISTURBI E DIAGNOSI DI ARTRITE**.

### 4.3.4. Europa

Per quanto riguarda la regione Europa, il dataset prodotto contiene 42399 soggetti ed è sprovvisto di osservazioni mancanti.

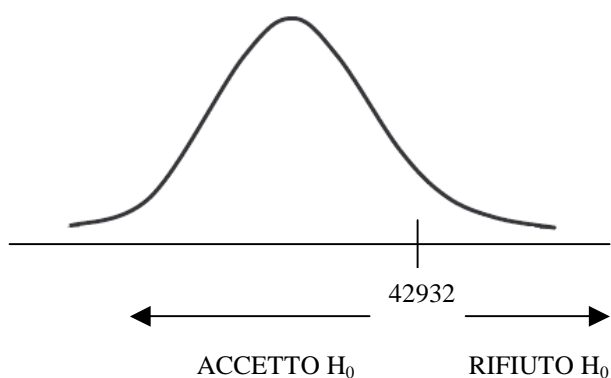
L'analisi con l'applicazione DSIGoM ha prodotto i seguenti risultati.

**Tabella 4.3.7.** Rapporto di log-verosimiglianza e  $\chi^2$  per 2, 3, 4, 5 e 6 profili estremi.

Numero profili	Rapporto di log verosimiglianza	$\chi^2$	$\Delta \chi^2$	Gradi di libertà (d.f.)	$\Delta$ Gradi di libertà (d.f.)
2	99976.7	199950		42452	
3	138952	277900	77950	84904	42452
4	160707	321410	43510	127356	42452
5	175406	350810	29400	169808	42452
6	195030	390060	39250	212260	42452

Seguendo il procedimento approfondito nei paragrafi precedenti, per decidere quanti profili estremi considerare si è confrontato il livello del test ( $\Delta \chi^2$ ) con il valore di critico ( $\chi^2_{42452, 0.95}$ ), secondo la figura seguente.

**Figura 4.3.4.** Distribuzione  $\chi^2$  con 42452 gradi di libertà.

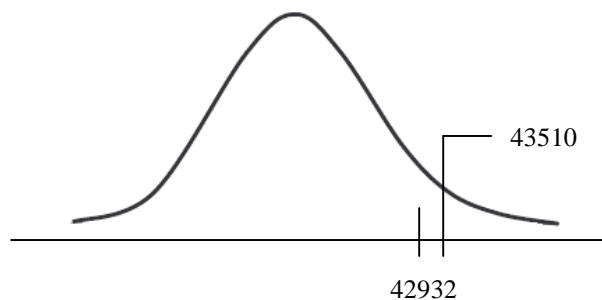


Per la scelta dei profili estremi è stata seguita la seguente procedura:

- i. Inizialmente sono stati confrontati il modello con 2 e quello con 3 profili estremi. Si è constatato che il valore del test del  $\chi^2$  (77950) è maggiore del valore critico (42932), quindi si rifiuta l'ipotesi  $H_0$  e si procede con il successivo raffronto.
- ii. Comparando, poi, il modello con 3 e quello con 4 profili estremi, si è osservato che il valore del test del  $\chi^2$  (43510) risulta di poco maggiore rispetto al valore critico (42932).

In questo caso si è di fronte alla seguente situazione:

**Figura 4.3.5.** Confronto tra valore critico e valore del test.



Osservando la figura 4.3.5 è possibile notare che, nonostante il valore del test del  $\chi^2$  si collochi nella regione di rifiuto dell'ipotesi nulla, si è deciso di accettare lo stesso l'ipotesi  $H_0$ , poiché lo scarto tra il valore del test ed il valore critico risulta essere minimo. Di conseguenza si è concluso che il modello con  $K=3$  profili estremi si adatta correttamente ai dati dell'Europa.

I risultati ottenuti sono riportati nella tabella di pagina seguente.

**Tabella 4.3.8.**  $\lambda_{kjl}$  per l'Europa con K=3 profili estremi.

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Health Status</b>				
Very Good	16.21	32.48	0.00	0.00
Good	38.50	67.52	18.32	0.00
Moderate	33.63	0.00	81.68	51.27
Bad	9.83	0.00	0.00	41.08
Very Bad	1.83	0.00	0.00	7.65
<b>Difficulty moving around</b>				
None	65.00	100.00	32.34	0.00
Mild	14.74	0.00	67.66	0.00
Moderate	12.55	0.00	0.00	61.95
Severe	6.60	0.00	0.00	32.57
Extreme/Cannot do	1.11	0.00	0.00	5.47
<b>Difficulty in Self-Care</b>				
None	86.06	100.00	100.00	17.92
Mild	7.05	0.00	0.00	41.50
Moderate	4.66	0.00	0.00	27.42
Severe	1.74	0.00	0.00	10.25
Extreme/Cannot do	0.50	0.00	0.00	2.92
<b>Pain and discomfort</b>				
None	46.86	100.00	0.00	0.00
Mild	24.37	0.00	100.00	0.00
Moderate	18.59	0.00	0.00	64.63
Severe	8.84	0.00	0.00	30.75
Extreme	1.33	0.00	0.00	4.62
<b>Difficulty in concentration</b>				
None	65.29	100.00	0.00	0.00
Mild	20.27	0.00	100.00	0.00
Moderate	10.54	0.00	0.00	72.98
Severe	3.45	0.00	0.00	23.92
Extreme/Cannot do	0.45	0.00	0.00	3.10
<b>Difficulty in personal relationships</b>				
None	77.48	100.00	49.56	39.74
Mild	14.49	0.00	50.44	16.19
Moderate	5.76	0.00	0.00	31.60
Severe	1.88	0.00	0.00	10.31
Extreme/Cannot do	0.39	0.00	0.00	2.16

**Tabella 4.3.8.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Difficulty in seeing and recognizing persons</b>				
None	78.46	100.00	64.91	31.47
Mild	11.28	0.00	35.09	15.30
Moderate	6.31	0.00	0.00	32.75
Severe	2.93	0.00	0.00	15.19
Extreme/Cannot do	1.02	0.00	0.00	5.30
<b>Sleeping disorders</b>				
None	53.62	100.00	0.00	0.00
Mild	22.06	0.00	85.44	0.00
Moderate	15.05	0.00	14.56	54.91
Severe	7.92	0.00	0.00	38.53
Extreme	1.35	0.00	0.00	6.56
<b>Feeling sad or depressed</b>				
None	56.07	100.00	0.00	0.00
Mild	23.93	0.00	100.00	0.00
Moderate	13.50	0.00	0.00	67.52
Severe	5.56	0.00	0.00	27.82
Extreme	0.93	0.00	0.00	4.66
<b>Diagnosis of Arthritis</b>				
No	83.25	100.00	100.00	31.32
Yes	16.75	0.00	0.00	68.68
<b>Diagnosis of Angina Pectoris</b>				
No	89.05	100.00	100.00	49.76
Yes	10.95	0.00	0.00	50.24
<b>Diagnosis of Asthma</b>				
No	93.11	96.04	95.22	82.87
Yes	6.89	3.96	4.78	17.13
<b>Diagnosis of Depression</b>				
No	90.94	100.00	100.00	60.06
Yes	9.06	0.00	0.00	39.94



Analizzando i tre profili estremi ottenuti in tabella 4.3.8, si può facilmente notare che il profilo I contiene tutti i soggetti in buona salute, senza particolari disturbi e con carenza di patologie mediche diagnosticate. Questo profilo può essere denominato STATO DI SALUTE OTTIMO.

Viceversa, il profilo II include tutti gli individui con uno stato di salute moderato, con qualche disturbo (in particolare sofferenza fisica, scarsità di concentrazione, problemi del sonno, tristezza o depressione), e sempre privi di patologie mediche diagnosticate (l'unica probabilità che aumenta leggermente è quella riguardante l'asma). Questo profilo estremo può essere indicato STATO DI SALUTE MODERATO CON LEGGERI DISTURBI.

In fine, il profilo III racchiude tutti i soggetti con uno stato di salute insoddisfacente, con notevoli disturbi e con aumenti delle probabilità di avere una diagnosi in tutte e quattro le patologie. In particolare risulta che il 68.68% degli Europei afferma di soffrire di artrite ed il 50,24% di angina pectoris. Questo profilo estremo può essere definito PESSIMO STATO DI SALUTE CON GRAVI DISTURBI E DIAGNOSI DI ARTRITE ED ANGINA PECTORIS.

### 4.3.5. Sud-Est Asiatico

Per quanto riguarda la regione Sud-Est Asiatico, il dataset costruito contiene 32697 soggetti ed è sprovvisto di osservazioni mancanti.

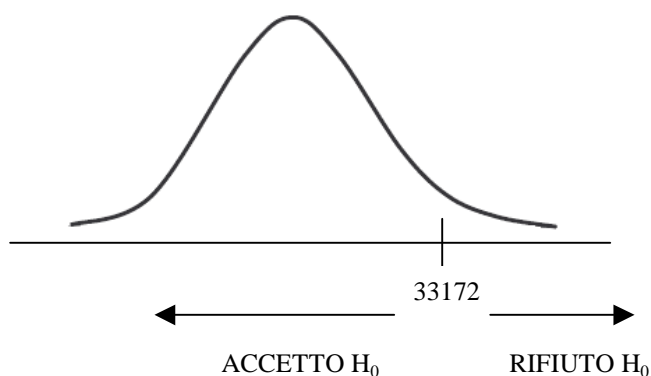
L'analisi con l'applicazione DSIGoM ha prodotto i seguenti valori:

**Tabella 4.3.9.** Rapporto di log-verosimiglianza e  $\chi^2$  per 2, 3, 4, 5 e 6 profili estremi.

Numero profili	Rapporto di log verosimiglianza	$\chi^2$	$\Delta \chi^2$	Gradi di libertà (d.f.)	$\Delta$ Gradi di libertà (d.f.)
2	80448.5	160900		32750	
3	112394	224790	63890	65500	32750
4	129199	258400	33610	98250	32750
5	145122	290240	31840	131000	32750
6	155315	310630	20390	163750	32750

Seguendo il procedimento affrontato nei paragrafi precedenti, per decidere quanti profili estremi considerare si è confrontato il livello del test ( $\Delta \chi^2$ ) con il valore critico ( $\chi^2_{32750, 0.95}$ ), secondo la seguente figura.

**Figura 4.3.6.** Distribuzione  $\chi^2$  con 32750 gradi di libertà.



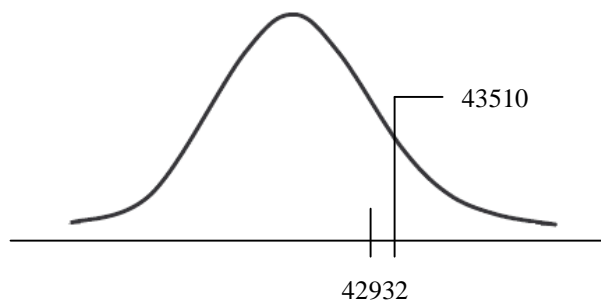
Per la scelta dei profili estremi è stata seguita la seguente procedura:

- i. Inizialmente sono stati confrontati il modello con 2 e quello con 3 profili estremi.

Si è constatato che il valore del test del  $\chi^2$  (63890) è maggiore del valore critico (33172), quindi si rifiuta l'ipotesi  $H_0$  e si procede con il successivo accostamento.

- ii. Comparando, poi, il modello con 3 e quello con 4 profili estremi, si è osservato che il valore del test del  $\chi^2$  (33610) risulta di poco maggiore rispetto al valore critico (33172). Anche in questo caso si è di fronte alla seguente situazione

**Figura 4.3.7.** Confronto tra valore critico e valore del test.



Osservando la figura 4.3.7 è possibile notare che, nonostante il valore del test del  $\chi^2$  si colloca nella regione di rifiuto dell'ipotesi nulla, si è deciso di accettarla comunque dato che lo scarto tra il valore del test (33610) ed il valore critico (33172) risulta essere minimo. Di conseguenza si è concluso che anche per questa regione il modello con  $K=3$  profili estremi è adeguato.

I risultati ottenuti sono riportati nella tabella di pagina seguente.

**Tabella 4.3.10.**  $\lambda_{kjl}$  per il Sud-Est Asiatico con K=3 profili estremi.

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Health Status</b>				
Very Good	20.11	35.80	0.00	0.00
Good	43.42	56.66	51.09	0.00
Moderate	26.40	7.54	48.91	52.38
Bad	8.71	0.00	0.00	41.21
Very Bad	1.35	0.00	0.00	6.41
<b>Difficulty moving around</b>				
None	62.43	100.00	0.00	0.00
Mild	18.25	0.00	100.00	0.00
Moderate	10.87	0.00	0.00	56.25
Severe	6.96	0.00	0.00	36.02
Extreme/Cannot do	1.49	0.00	0.00	7.73
<b>Difficulty in Self-Care</b>				
None	76.76	100.00	61.76	0.00
Mild	12.32	0.00	38.24	26.43
Moderate	6.12	0.00	0.00	41.21
Severe	3.53	0.00	0.00	23.78
Extreme/Cannot do	1.27	0.00	0.00	8.57
<b>Pain and discomfort</b>				
None	44.65	100.00	0.00	0.00
Mild	29.38	0.00	100.00	0.00
Moderate	13.40	0.00	0.00	51.57
Severe	10.53	0.00	0.00	40.55
Extreme	2.05	0.00	0.00	7.89
<b>Difficulty in concentration</b>				
None	61.18	100.00	0.00	0.00
Mild	21.38	0.00	100.00	0.00
Moderate	9.96	0.00	0.00	57.09
Severe	5.98	0.00	0.00	34.30
Extreme/Cannot do	1.50	0.00	0.00	8.61
<b>Difficulty in personal relationships</b>				
None	75.44	100.00	47.38	21.99
Mild	14.20	0.00	52.62	16.02
Moderate	5.51	0.00	0.00	32.99
Severe	3.29	0.00	0.00	19.67
Extreme/Cannot do	1.56	0.00	0.00	9.33

**Tabella 4.3.10.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Difficulty in seeing and recognizing persons</b>				
None	76.07	100.00	54.38	29.63
Mild	12.06	0.00	45.62	9.04
Moderate	5.40	0.00	0.00	27.89
Severe	4.96	0.00	0.00	25.63
Extreme/Cannot do	1.51	0.00	0.00	7.82
<b>Sleeping disorders</b>				
None	61.95	100.00	0.00	0.00
Mild	19.74	0.00	100.00	0.00
Moderate	9.48	0.00	0.00	51.75
Severe	7.08	0.00	0.00	38.66
Extreme	1.76	0.00	0.00	9.59
<b>Feeling sad or depressed</b>				
None	53.77	100.00	0.00	0.00
Mild	24.57	0.00	100.00	0.00
Moderate	10.75	0.00	0.00	49.64
Severe	8.80	0.00	0.00	40.62
Extreme	2.11	0.00	0.00	9.74
<b>Diagnosis of Arthritis</b>				
No	86.64	100.00	84.91	52.14
Yes	13.36	0.00	15.09	47.86
<b>Diagnosis of Angina Pectoris</b>				
No	94.23	100.00	100.00	72.66
Yes	5.77	0.00	0.00	27.34
<b>Diagnosis of Asthma</b>				
No	95.06	100.00	96.76	79.20
Yes	4.94	0.00	3.24	20.80
<b>Diagnosis of Depression</b>				
No	88.65	95.34	90.75	67.25
Yes	11.35	4.66	9.25	32.75

Esaminando questi tre profili estremi, si può facilmente notare che il profilo I contiene tutti i soggetti in buona salute, senza particolari disturbi e con assenza di patologie mediche diagnosticate. Questo profilo può essere denominato **STATO DI SALUTE SODDISFACENTE**.

Viceversa, il profilo II include tutti gli individui con uno stato di salute moderato, con qualche disturbo (in particolare difficoltà di movimento, sofferenza fisica, scarsità di concentrazione, problemi del sonno, tristezza o depressione), e con un aumento, rispetto al profilo precedente, di alcune patologie (artrite, asma e depressione). Questo profilo estremo può essere indicato **STATO DI SALUTE MODERATO CON LIEVI DISTURBI**.

In fine, il profilo III racchiude tutti i soggetti con uno stato di salute inadeguato, con importanti disturbi e con importanti aumenti delle probabilità di avere una diagnosi in tutte e quattro le patologie. Questo profilo estremo può essere definito come **PESSIMO STATO DI SALUTE CON FORTI DISTURBI**.

Anche per quest'area regionale, come per l'Africa e l'America, accade che in nessuno dei tre profili compaiono soggetti con un'elevata probabilità di soffrire di qualche patologia medica. Questo accade probabilmente perché nell'intero campione la frequenza di soggetti con disturbi accertati non supera il 13.36% (le frequenze variano tra il 4.94% dell'asma ed il 13.36% dell'artrite). Tuttavia è possibile osservare che, nel passaggio dal profilo I al profilo III, le probabilità di avere una qualche diagnosi crescono ma non superano mai il 50%.

### 4.3.6. Pacifico Occidentale

Per quanto riguarda la regione Pacifico Occidentale, il dataset prodotto contiene 29601 soggetti ed è privo di osservazioni mancanti.

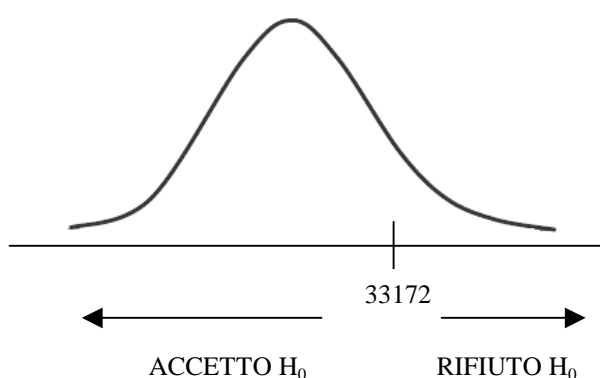
L'analisi con il programma DSIGoM ha fornito le quantità di pagina seguente.

**Tabella 4.3.11.** Rapporto di log-verosimiglianza e  $\chi^2$  per 2, 3, 4, 5 e 6 profili estremi.

Numero profili	Rapporto di log verosimiglianza	$\chi^2$	$\Delta \chi^2$	Gradi di libertà (d.f.)	$\Delta$ Gradi di libertà (d.f.)
2	67238.3	134480		29654	
3	93572.8	187150	52670	59308	29654
4	107117	214230	27080	88962	29654
5	119720	239440	25210	118616	29654
6	127883	255770	16330	148270	29654

Seguendo il procedimento utilizzato nei paragrafi precedenti, per decidere quanti profili estremi considerare si è confrontato il livello del test ( $\Delta \chi^2$ ) con il valore critico ( $\chi^2_{29654, 0.95}$ ), secondo la seguente figura.

**Figura 4.3.8.** Distribuzione  $\chi^2$  con 29654 gradi di libertà.



La procedura seguita è stata la seguente:

- i. Inizialmente sono stati confrontati il modello con 2 e quello con 3 profili estremi. Si è constatato che il valore del test del  $\chi^2$  (52670) è maggiore del valore critico (30056), quindi si rifiuta l'ipotesi  $H_0$  e si procede con il successivo paragone.
- ii. Comparando, poi, il modello con 3 e quello con 4 profili estremi, l'incremento di  $\chi^2$  è minore del valore critico ( $27080 < 30056$ ), e si accetta pertanto l'ipotesi  $H_0$ .

In tal caso si è concluso che il modello con  $K=3$  profili estremi si adatta adeguatamente ai dati del Pacifico Occidentale.

I risultati ottenuti sono riportati nella tabella sottostante.

**Tabella 4.3.12.**  $\lambda_{kjl}$  per il Pacifico Occidentale con  $K=3$  profili estremi.

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Health Status</b>				
Very Good	20.16	37.58	0.00	0.00
Good	47.29	62.42	38.66	16.04
Moderate	28.10	0.00	61.34	59.55
Bad	4.10	0.00	0.00	22.50
Very Bad	0.35	0.00	0.00	1.91
<b>Difficulty moving around</b>				
None	68.32	100.00	27.97	0.00
Mild	15.69	0.00	72.03	0.00
Moderate	13.88	0.00	0.00	86.78
Severe	1.91	0.00	0.00	11.93
Extreme/Cannot do	0.21	0.00	0.00	1.29



**Tabella 4.3.12.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
<b>Difficulty in Self-Care</b>				
None	84.61	100.00	86.20	0.00
Mild	7.60	0.00	13.80	35.16
Moderate	7.00	0.00	0.00	58.27
Severe	0.67	0.00	0.00	5.54
Extreme/Cannot do	0.12	0.00	0.00	1.04
<b>Pain and discomfort</b>				
None	49.15	100.00	0.00	0.00
Mild	29.83	0.00	100.00	0.00
Moderate	17.17	0.00	0.00	81.68
Severe	3.55	0.00	0.00	16.89
Extreme	0.30	0.00	0.00	1.43
<b>Difficulty in concentration</b>				
None	59.80	100.00	0.00	0.00
Mild	24.01	0.00	100.00	0.00
Moderate	13.58	0.00	0.00	83.87
Severe	2.39	0.00	0.00	14.73
Extreme/Cannot do	0.23	0.00	0.00	1.40
<b>Difficulty in personal relationships</b>				
None	75.94	100.00	55.53	0.00
Mild	12.18	0.00	44.47	13.66
Moderate	10.16	0.00	0.00	73.82
Severe	1.39	0.00	0.00	10.09
Extreme/Cannot do	0.33	0.00	0.00	2.43
<b>Difficulty in seeing and recognizing persons</b>				
None	77.69	100.00	58.30	20.46
Mild	11.79	0.00	41.70	9.96
Moderate	7.61	0.00	0.00	50.35
Severe	2.51	0.00	0.00	16.57
Extreme/Cannot do	0.40	0.00	0.00	2.66
<b>Sleeping disorders</b>				
None	60.89	100.00	0.00	0.00
Mild	21.60	0.00	100.00	0.00
Moderate	13.54	0.00	0.00	77.31
Severe	3.57	0.00	0.00	20.39
Extreme	0.40	0.00	0.00	2.30

**Tabella 4.3.12.** (Continua).

	<b>Freq</b>	<b>I</b>	<b>II</b>	<b>III</b>
Feeling sad or Depressed				
None	65.99	100.00	0.00	0.00
Mild	19.96	0.00	100.00	0.00
Moderate	11.07	0.00	0.00	78.82
Severe	2.53	0.00	0.00	18.03
Extreme	0.44	0.00	0.00	3.15
Diagnosis of Arthritis				
No	87.01	100.00	70.10	66.90
Yes	12.99	0.00	29.90	33.10
Diagnosis of Angina Pectoris				
No	95.34	100.00	93.63	81.16
Yes	4.66	0.00	6.37	18.84
Diagnosis of Asthma				
No	93.16	97.27	88.73	84.76
Yes	6.84	2.73	11.27	15.24
Diagnosis of Depression				
No	96.79	100.00	100.00	80.94
Yes	3.21	0.00	0.00	19.06

Analizzando i tre profili estremi di tabella 4.3.12, si può facilmente notare che il profilo I contiene tutti i soggetti in buona salute, senza particolari disturbi e in assenza di patologie mediche. Questo profilo può essere denominato STATO DI SALUTE SODDISFACENTE.

Viceversa, il profilo II include tutti gli individui con uno stato di salute moderato, con qualche disturbo (in particolare sofferenza fisica, scarsità di concentrazione, problemi del sonno, tristezza o depressione) e con un aumento, rispetto al profilo precedente, di alcune patologie (artrite, angina pectoris ed asma). Questo profilo estremo può essere indicato STATO DI SALUTE MODERATO CON LEGGERI DISTURBI.

In fine, il profilo III racchiude tutti i soggetti con uno stato di salute insoddisfacente, con notevoli disturbi e con importanti aumenti delle probabilità di avere una diagnosi in tutte e quattro le patologie diagnosticate. Questo profilo estremo può essere definito STATO DI SALUTE INSODDISFACENTE CON GRAVI DISTURBI.

Anche per questa regione, come per l'Africa, l'America ed il Sud-Est Asiatico, accade che in nessuno dei tre profili compaiono soggetti con un'elevata probabilità di soffrire di qualche patologia medica. Questo accade probabilmente perché nell'intero campione la frequenza di soggetti con disturbi accertati non supera il 12.99% (le frequenze variano tra il 3.21% della depressione ed il 12.99% dell'artrite). Tuttavia è possibile osservare che, nel passaggio dal profilo I al profilo III, le probabilità di avere una qualche diagnosi crescono ma non superano mai il 50%.



## *Capitolo 5*

# **DISCUSSIONE**

In questo capitolo viene presentata una breve discussione sui risultati ottenuti. In particolare si cercheranno di confrontare le probabilità di risposta delle sei aree regionali considerate, ottenute mediante l'applicazione del modello GoM.

Per prima cosa è importante evidenziare che in tutte le regioni sono stati ottenuti  $K=3$  profili estremi. Questo è un vantaggio perché permette di confrontare i risultati ottenuti.

Da un primo sguardo alle tabelle contenute nel paragrafo 4.3 è immediato notare che per le regioni Africa, America, Sud-Est Asiatico e Pacifico Occidentale si sono raggiunte conclusioni simili. Questo potrebbe indurre a ritenere che gli abitanti di queste quattro aree regionali possiedono uno stato di salute pressoché somigliante. È naturale che se si osservano le probabilità per ogni modalità di risposta si avranno risultati diversi da regione a regione.

Per quanto riguarda le restanti due aree regionali, Mediterraneo Orientale ed Europa, i profili ottenuti dalla stima del modello GoM sono diversi. In particolare il Mediterraneo Orientale comprende anche i soggetti che dichiarano di soffrire di artrite, mentre i profili dell'Europa includono sia i soggetti affetti da artrite che da angina pectoris.

In questo modo si sono formati principalmente tre categorie di risultati:

- la prima composta dai profili di Africa, America, Sud-Est Asiatico e Pacifico Occidentale, cioè
  - I. *Stato di salute soddisfacente*
  - II. *Stato di salute moderato con leggeri disturbi*
  - III. *Stato di salute insoddisfacente con forti disturbi*
- la seconda composta dai profili di Mediterraneo Orientale, cioè
  - I. *Stato di salute soddisfacente*
  - II. *Stato di salute moderato con lievi disturbi*
  - III. *Stato di salute pessimo con forti disturbi e diagnosi di artrite*
- la terza composta dai profili dell'Europa, cioè
  - I. *Stato di salute ottimo*
  - II. *Stato di salute moderato con leggeri disturbi*
  - III. *Pessimo stato di salute con gravi disturbi e diagnosi di artrite ed angina pectoris.*

Confrontando tra loro queste tipologie di risultati è possibile notare che i soggetti appartenenti al terzo gruppo (europei) percepiscono uno stato di salute peggiore rispetto alle altre due categorie.

Se si effettua un confronto all'interno della *prima categoria*, è possibile osservare che, in generale, non ci sono grosse differenze tra lo stato di salute percepito dalla popolazione dell'America e quello avvertito dalla popolazione africana.

Ad esempio, se si osservano le frequenze di risposta alle otto domande riguardanti le condizioni di salute, si può notare che in ognuno di questi quesiti i soggetti americani hanno probabilità lievemente maggiori di assenza di disturbi. Al contrario, per quanto riguarda le ultime quattro domande concernenti le patologie diagnosticate, gli abitanti dell'America sembrano soffrire di più d'asma e di depressione e meno di artrite e di angina pectoris.

Se si confrontano, poi, queste due aree con il Sud-Est Asiatico, è possibile osservare che, in generale, gli abitanti di quest'ultima regione sembrano godere di uno stato di salute pressoché simile da quello avvertito dalla popolazione dell'Africa.

Di fatto, se si osservano le frequenze di risposta alle otto domande riguardanti le condizioni di salute, si può facilmente notare che in ognuno di questi quesiti gli abitanti del Sud-Est Asiatico hanno probabilità di assenza di disturbi minori rispetto l'America ed abbastanza simili a quelle dell'Africa.

Inoltre, per quanto riguarda le ultime quattro domande riguardanti le patologie diagnosticate, rispetto la regione Africa, gli abitanti del Sud-Est Asiatico sembrano soffrire più d'asma (soprattutto nell'ultimo profilo) e depressione, ma meno di artrite (tranne nell'ultimo profilo) e angina pectoris. Al contrario, in relazione all'America, il Sud-Est Asiatico possiede frequenze di risposta maggiori in tutte le quattro patologie.

Concludendo, si può affermare che la popolazione dell'Sud-Est Asiatico possiedono uno stato di salute pressoché simile a quello dell'Africa ma sicuramente peggiore di quello avvertito dall'America.

Infine, se si confrontano separatamente tutte e quattro le regioni appartenenti alla prima categoria, è possibile osservare che, in generale, gli abitanti del Pacifico Occidentale sembrano percepire uno stato di salute pressoché simile a quello avvertito dalla popolazione del Sud-Est Asiatico e dell'Africa.

Ad esempio, se si osservano le frequenze di risposta alle otto domande relative alle condizioni di salute, si può facilmente notare che in ognuno di questi quesiti gli abitanti del Pacifico Occidentale hanno probabilità di assenza di disturbi minori rispetto l'America, ma abbastanza simili a quelle dell'Africa e del Sud-Est Asiatico. Inoltre, per quanto riguarda le ultime quattro domande riguardanti le patologie diagnosticate, rispetto l'Africa ed il Sud-Est Asiatico, gli abitanti del Pacifico Occidentale sembrano soffrire più d'asma e meno di artrite, angina pectoris e depressione.

Al contrario, in relazione al continente America, gli abitanti del Pacifico Occidentale sembrano soffrire meno di depressione e più di artrite, asma ed angina pectoris.



Concludendo, si può affermare che la popolazione del Pacifico Occidentale gode di uno stato di salute pressoché simile a quello dell’Africa e del Sud-Est Asiatico ma certamente peggiore di quello avvertito dall’America.

Dopo aver analizzato tutte le aree regionali della prima categoria, sono state confrontate tra loro le regioni Mediterraneo Orientale ed Europa. Infatti, è immediato notare che, in generale, gli abitanti dell’Europa sembrano percepire uno stato di salute moderatamente peggiore rispetto a quello avvertito dagli abitanti del Mediterraneo Orientale.

Di fatto, se si osservano le frequenze di risposta alle otto domande riguardanti le condizioni di salute, si può notare che, in generale, in ognuno di questi quesiti gli abitanti dell’Europa hanno probabilità di assenza di disturbi minori rispetto al Mediterraneo Orientale. Inoltre, per quanto concerne le ultime quattro domande riguardanti le patologie diagnosticate, rispetto al Mediterraneo Orientale, gli abitanti dell’Europa sembrano soffrire meno di artrite e più di angina pectoris, asma e depressione.

Concludendo, si può affermare che la popolazione dell’Europa possiede uno stato di salute sicuramente peggiore di quello degli abitanti del Mediterraneo Orientale.



# CONCLUSIONI

Lo sviluppo del modello Grade of Membership (GoM) ha avuto inizio negli anni '70. Molti studiosi hanno cercato di servirsene per l'analisi di dati di vario genere, tuttavia questo metodo non ha ottenuto molte attenzioni nella letteratura statistica.

L'analisi del modello GoM ha obiettivi simili all'analisi delle classi latenti utilizzata in psicometria<sup>14</sup>. Ci sono alcuni collegamenti teorici tra questi due tipi di modelli, che sono stati presentati in questa tesi. In particolare, è stato dimostrato che il modello GoM può essere interpretato come caso particolare del modello a classi latenti. Definendo costanti la struttura e la distribuzione delle classi latenti, si è ottenuta la rappresentazione a classi latenti del modello GoM (vedi Capitolo 3, pag. 25).

Per raggiungere l'obiettivo preposto (vedi Prefazione, pag. 1) si è mostrato che il metodo GoM è il procedimento più adatto per riassumere il concetto dinamico multidimensionale di stato di salute.

---

<sup>14</sup> La psicometria è il campo di studio della **teoria e della tecnica della misura in psicologia**. Il campo di studio è particolarmente volto verso lo studio delle differenze tra individui. Gran parte del lavoro teorico ed applicativo della psicometria è stato svolto nel tentativo di misurare l'intelligenza.

A tale scopo sono stati analizzati dati forniti dall'Organizzazione Mondiale della Sanità (OMS), ottenuti mediante l'indagine World Health Survey (WHS), con i quali è stato creato un dataset contenente solo alcune variabili di interesse (vedi Appendice, pag. 95).

In primo luogo, per avere un'idea generale delle caratteristiche generali del campione, è stata costruita una tabella contenente le prevalenze, da cui è emerso che la maggior parte dei soggetti intervistati percepisce uno stato di salute buono od assolutamente ottimo.

Successivamente, è stato applicato il metodo GoM, da cui è risultato che, per ogni regione, il concetto di stato di salute può essere ripartito attraverso le seguenti tre caratteristiche (profili estremi):

- I. Soggetti con uno stato di salute soddisfacente. Questa categoria include tutti gli abitanti delle sei aree regionali considerate che dichiarano di non avere nessun disturbo e quasi assenza di patologie diagnosticate.
- II. Individui con uno stato di salute moderato con lievi disturbi. Questa classe contiene tutta la popolazione che manifesta qualche disturbo (in particolare difficoltà di movimento, sofferenza fisica, scarsità di concentrazione, mancanza di relazioni personali, problemi del sonno, tristezza o depressione) e con un aumento, rispetto al profilo precedente, della presenza di patologie mediche diagnosticate.
- III. Soggetti che possiedono uno stato di salute pessimo con forti disturbi ed alcune patologie diagnosticate.

Questa categoria include tutti gli individui che mostrano tutti o gran parte dei disturbi evidenziati nella seconda sezione del questionario (2000. *Health State Descriptions* ) e con la presenza di patologie diagnosticate, in particolare artrite ed angina pectoris.

Concludendo, sembra che, in generale, la popolazione mondiale possiede uno stato di salute abbastanza soddisfacente, nonostante sia presente qualche disturbo od una qualche patologia.



# APPENDICE

Nelle pagine seguenti sono riportate le domande selezionate dal questionario per l'applicazione del modello *Grade of Membership* ai dati della *World Health Survey*.

## 1000. Respondent's Socio Demographic Characteristics

<b>Q1001</b>	<b>Record sex as observed</b>	<b>1. Female</b>			<b>2. Male</b>		
<b>Q1002</b>	How old are you? (Years)	_____			<b>888. DK</b>		
<b>Q1008</b>	What is your current <u>marital status</u> ?	<b>1. Never Married</b>	<b>2. Currently Married</b>	<b>3. Separated</b>	<b>4. Divorced</b>	<b>5. Widowed</b>	<b>6. Cohabiting</b>
<b>Q1009</b>	What is the <u>highest level of education</u> that you have completed?	<b>1. No formal schooling</b>					
		<b>2. Less than primary school</b>					
		<b>3. Primary school completed</b>					
		<b>4. Secondary school completed</b>					
		<b>5. High school (or equivalent) completed</b>					
		<b>6. College / pre-university / University completed</b>					
		<b>7. Post graduate degree completed</b>					
<b>Q1012</b>	What is your <u>current job</u> ?	<b>1. Government employee</b>	<b>2. Non-government employee</b>	<b>3. Self-employed</b>	<b>4. Employer</b>	<b>5. Not working for pay</b>	



## 2000. Health State Descriptions

<b>Q2000</b>	In general, how would you rate your health today?	<b>1. Very Good</b>	<b>2. Good</b>	<b>3. Moderate</b>	<b>4. Bad</b>	<b>5. Very Bad</b>
--------------	---	---------------------	----------------	--------------------	---------------	--------------------

### Mobility

<b>Q2010</b>	Overall in the last 30 days, how much difficulty did you have with <u>moving around</u> ?	<b>1. None</b>	<b>2. Mild</b>	<b>3. Moderate</b>	<b>4. Severe</b>	<b>5. Extreme/ Cannot do</b>
--------------	---	----------------	----------------	--------------------	------------------	----------------------------------

### Self Care

<b>Q2020</b>	Overall in the last 30 days, how much difficulty did you have with <u>self-care</u> , such as washing or dressing yourself?	<b>1. None</b>	<b>2. Mild</b>	<b>3. Moderate</b>	<b>4. Severe</b>	<b>5. Extreme/ Cannot do</b>
--------------	---	----------------	----------------	--------------------	------------------	----------------------------------

### Pain and Discomfort

<b>Q2030</b>	Overall in the last 30 days, how much of <u>bodily aches or pains</u> did you have?	<b>1. None</b>	<b>2. Mild</b>	<b>3. Moderate</b>	<b>4. Severe</b>	<b>5. Extreme</b>
--------------	---	----------------	----------------	--------------------	------------------	-------------------

### Cognition

<b>Q2050</b>	Overall in the last 30 days, how much difficulty did you have with <u>concentrating or remember things</u> ?	<b>1. None</b>	<b>2. Mild</b>	<b>3. Moderate</b>	<b>4. Severe</b>	<b>5. Extreme/ Cannot do</b>
--------------	--	----------------	----------------	--------------------	------------------	----------------------------------

## Interpersonal Activities

<b>Q2060</b>	Overall in the last 30 days, how much difficulty did you have with <u>personal relationship or participation in the community</u> ?	<b>1. None</b>	<b>2. Mild</b>	<b>3. Moderate</b>	<b>4. Severe</b>	<b>5. Extreme/ Cannot do</b>
--------------	---	----------------	----------------	--------------------	------------------	----------------------------------

## Vision

<b>Q2071</b>	In the last 30 days, how much difficulty did you have in seeing and recognizing <u>a person you know across the road</u> (i.e. from a distance of about 20 meters)?	<b>1. None</b>	<b>2. Mild</b>	<b>3. Moderate</b>	<b>4. Severe</b>	<b>5. Extreme/ Cannot do</b>
--------------	---	----------------	----------------	--------------------	------------------	----------------------------------

## Sleep

<b>Q2080</b>	Overall in the last 30 days, how much of a problem did you have with sleeping, such as <u>falling asleep</u> , waking up <u>frequently during the night</u> or waking up <u>too early</u> in the morning?	<b>1. None</b>	<b>2. Mild</b>	<b>3. Moderate</b>	<b>4. Severe</b>	<b>5. Extreme</b>
--------------	---	----------------	----------------	--------------------	------------------	-------------------

## Affect

<b>Q2090</b>	Overall in the last 30 days, how much of a problem did you have with <u>feeling sad, low or depressed</u> ?	<b>1. None</b>	<b>2. Mild</b>	<b>3. Moderate</b>	<b>4. Severe</b>	<b>5. Extreme</b>
--------------	---	----------------	----------------	--------------------	------------------	-------------------

## 6000. Coverage

<b>Q6000</b>	Have you ever been diagnosed with <u>arthritis</u> (a disease in the joints)?	<b>1.</b> Yes	<b>5.</b> No	<b>8.</b> DK
<b>Q6009</b>	Have you ever been diagnosed with <u>angina</u> or <u>angina pectoris</u> (a heart disease)?	<b>1.</b> Yes	<b>5.</b> No	<b>8.</b> DK
<b>Q6017</b>	Have you ever been diagnosed with <u>asthma</u> (an allergic respiratory disease)?	<b>1.</b> Yes	<b>5.</b> No	<b>8.</b> DK
<b>Q6025</b>	Have you ever been diagnosed with <u>depression</u> ?	<b>1.</b> Yes	<b>5.</b> No	<b>8.</b> DK



# BIBLIOGRAFIA

- [1] Woodbury M. A., Clive J., Garson A. (1978), *Mathematical typology: A Grade of Membership technique for obtaining disease definition*, Computers and Biomedical Research, **11**, pp. 277-298.
- [2] Manton K. G., Woodbury M. A., Tolley H. D. (1994), *Statistical Applications using Fuzzy Sets*, Wiley-Interscience.
- [3] Lazarsfeld P., Henry N. W. (1968), *Latent Structure Analysis*, Boston, MA: Houghton Mifflin.
- [4] Erosheva E. A. (2002), *Grade of Membership and Latent Structure Models With Applications to Disability Survey Data*, Thesis, Department of Statistics, Canegie Mellon University.
- [5] Decision System, Inc. (1999), *User Documentation for DSIGoM. Version 1.0*.
- [6] McNamee P. (2004), *A comparison of Grade of Membership measure with alternative health indicators in explaining costs for older people*, Health Econ., **13**: pp. 379-395.
- [7] Fabbris L. (1991), *Analisi esplorativa di dati multidimensionali*, Cleup.
- [8] Portrait F., Lindeboom M, Deeg D. (1999), *Health and Mortality of the Elderly: the Grade of Membership method, Classification and Determination*, Health Econ., **8**: pp. 441-457.
- [9] <http://www.who.int/whs>
- [10] Vineis P., Duca P., Pasquini P. (1988), *Manuale di metodologia epidemiologica*, Cosmopress.
- [11] Manton K. G., Woodbury M. A. (1991), *Grade of Membership Generalizations and Aging Research*, Experimental Aging Research, **17**(4): pp. 217-226.
- [12] Manton K. G., Gu X., Huang H., Kovtun M. (2004), *Fuzzy set analyses of genetic determinants of health and disability status*, Statistical Methods in Medical Research, **13**: pp. 395-408.