I modelli di scaling. Confronto tra ipotesi complesse per la misurazione del soggettivo

Applicazioni STatistiche nella RIcerca Sociale - 2

Filomena Maggino

I modelli di scaling

Confronto tra ipotesi complesse per la misurazione del soggettivo

Università degli Studi di Firenze

Indice

1. I MODELLI DI SCALING

Come sappiamo, le misure multiple rendono la misurazione più sensibile consentendo di rilevare l'attributo con maggiore precisione; in altre parole le misure multiple consentono di:

- discriminare tra gli oggetti in modo più chiaro,
- aumentare l'affidabilità, infatti se per stimare il valore posseduto dall'oggetto si utilizza una combinazione di più misure, gli errori causali presenti nelle diverse misure tendono a compensarsi, rendendo la misurazione più accurata. A tale proposito occorre tenere presente che maggiore è la componente d'errore in ogni singola misurazione, maggiore è il numero di misurazioni richiesto per produrre una misura affidabile.

Ciascun modello di *scaling* è definito dall'insieme dei metodi e delle tecniche finalizzate alla sintesi di un insieme di dati per renderli significativi e gestibili[1] e prevede la definizione di tre approcci:

- *teorica*, che definisce il tipo di approccio allo *scaling*,
- *sperimentale*, che definisce il procedimento per rilevare i dati,
- *analitica*, che definisce il trattamento dei dati.

Ciascun modello di *scaling* è costituito da un insieme di assunti definiti all'interno di un'ipotesi di misurazione. Gli assunti che definiscono i diversi modelli riguardano:

- la dimensionalità, ovvero il numero di componenti del concetto da misurare;
- la verifica del modello di *trace line*[2] dell'indicatore (item) adottata, sulla base del livello di approssimazione tra *trace line* e frequenze.

La definizione delle misure multiple per ciascuna variabile formalizza lo *standard di misurazione*, ovvero il vero e proprio procedimento di misurazione; possono essere considerate misure multiple:

- gli indicatori, in questo caso l'insieme delle misure multiple consente di registrare in maniera più accurata il caso; in questo caso l'obiettivo della misurazione è la classificazione dei casi rispetto all'attributo;
- i casi, in questo caso l'insieme delle misure multiple consente di registrare una misura più accurata dell'indicatore; in questo caso l'obiettivo della misurazione è la classificazione dell'indicatore rispetto all'attributo.

Vediamo un esempio in cui l'interesse è concentrato sullo studio del livello di "la distanza sociale verso un particolare gruppo sociale" in una comunità; in questo caso le misure multiple potrebbero essere rappresentate da diversi item rappresentati da affermazioni su comportamenti rispetto ad un componente del gruppo sociale (per esempio "non voglio averci a che fare", "accetterei di sedergli accanto nell'autobus", "lo accetterei come compagno di lavoro", "lo inviterei a casa mia", "lo accetterei come amico", "lo accetterei come parente acquisito"[3]).

---

[1] In questo senso esso può essere definito come un "piano internamente consistente di sviluppo di una nuova misura" (Nunnally, 1978) per costruire e applicare le regole di misurazione.

[2] Il modello di *trace line* definisce la relazione tra il continuum osservato e la probabilità di ottenere un certo valore di misurazione, esempio una risposta corretta ad un item. Dato che tale probabilità sottintende la presenza di una porzione d'errore, è possibile associare a ciascun item due distribuzioni corrispondenti a due tipi di probabilità:

- *alfa*, corrispondente alla probabilità del valore atteso, interpretabile come risposta "corretta", superamento dello stimolo, accordo con l'affermazione presentata o valore della misurazione che va nella direzione della dimensione misurata;
- *beta*, corrispondente alla probabilità del valore non atteso interpretabile come risposta "sbagliata", mancato superamento dello stimolo, disaccordo con l'affermazione presentata o valore della misurazione che non va nella direzione della dimensione misurata.

[3] Si tratta di uno strumento utilizzato dall'ideatore E.Bogardus, sociologo austro-americano, per misurare e studiare il razzismo; una forma alternativa di scala Bogardus da utilizzare nel nostro esempio può essere: lo accetterei come

Se l'obiettivo è quello di misurare

- il livello di distanza sociale di ciascun individuo, le misure multiple saranno rappresentate dall'insieme degli item;
- il livello di distanza sociale rilevata da ciascun item, le misure multiple saranno rappresentate dall'insieme degli individui.

L'esempio ci consente di mettere in evidenza come la risposta individuale possa assumere un peso e un significato diverso; in altre parole a seconda dello standard di misurazione, la risposta individuale può assume un ruolo diverso; in particolare:

	Standard di misurazione	
	"Oggetto" della misurazione	Misure multiple
ciascun individuo reagisce a ciascuna variabile sulla base della propria posizione rispetto all'attributo	caso	item
ciascun individuo valuta ciascun item rispetto all'attributo che deve essere misurato	item	caso

Naturalmente non in tutte le occasioni è possibile scegliere tra i due diversi standard. Esiste infatti un'area in cui la distinzione tra i due approcci sulla base del "ruolo" degli individui si rivela insufficiente; tale area comprende le preferenze, i giudizi estetici, i giudizi di piacevolezza, ecc.; in questi casi se le risposte riflettono sempre la posizione soggettiva rispetto all'item su un dato attributo.

1.1 LE PROCEDURE

Per ciascun modello di *scaling* è possibile identificare due diverse procedure:

- *criterio di scaling*: per il quale l'obbiettivo può essere
  - *confermativo* al fine di verificare un'ipotesi sul modello in termini di valutazione dell'adattamento del modello di misurazione ipotizzato e formulato all'insieme di dati; ciò vuol dire principalmente accertare se una particolare struttura dimensionale rappresenta in modo accurato un insieme di dati empirici; la valutazione del livello di adattamento della struttura ai dati avviene attraverso particolari misure di bontà di adattamento;
  - *esplorativo* al fine di scoprire e identificare il numero delle dimensioni latenti necessarie a descrivere un insieme di osservazioni ovvero identificare la struttura dei dati; ciò può essere fatto attraverso particolari strategie di *analisi dimensionale*, importanti nelle scienze sociali, nei casi in cui i ricercatori non conoscono a priori le fonti di variazione empirica nelle osservazioni;
- *tecnica di scaling*: stabilito che una particolare struttura dimensionale è appropriata, l'obbiettivo è quello di misurare gli oggetti; ciò equivale a trovare uno specifico numero di valori (coordinate) che consenta di collocare ciascun punto-oggetto in uno spazio definito dalla dimensionalità del modello adottato; quindi il risultato delle tecniche di *scaling* è rappresentato dai valori assegnati alle osservazioni; ciò consentirà anche di correlare il concetto misurato con altre misure di interesse attraverso l'attribuzione dei punteggi agli oggetti (obbiettivo *descrittivo*).

La maggior parte dei modelli di *scaling* presentano entrambe le procedure.
Gli errori di *scaling* vengono interpretati in maniera diversa a seconda della procedura coinvolta.

"vicino di casa", "amico", "parente", "partner", "figlio".

1.2 INTERPRETAZIONE DEGLI ERRORI DI SCALING

Le strutture geometriche generate dalle procedure di *scaling* raramente si adattano perfettamente ai dati; la discrepanza tra le posizioni dei punti e gli oggetti empirici sono dette *errori di scaling*. La natura dell'errore dipende dalla specifica procedura di *scaling* ma è possibile identificare in generale due possibili fonti di errore:

a. essi possono riflettere la presenza di un inappropriato modello di *scaling* (la dimensionalità è sbagliata oppure il modello geometrico non è coerente con le osservazioni empiriche); tale interpretazione è possibile nell'ambito del criterio di *scaling*;
b. essi possono rappresentare semplicemente delle fluttuazioni che si verificano perché le osservazioni sono affette da errori di misurazione, errori di campionamento o altri fattori stocastici; tale interpretazione sottintende una accettazione della struttura dimensionale e quindi è possibile nell'ambito di una tecnica di *scaling*; il principale obbiettivo di una tecnica di *scaling* è quello di eliminare tali tipi di errore e produrre un insieme di numeri che meglio rappresenti la variabilità tra gli oggetti[4].

Il problema del ricercatore è quello di capire davanti a quale tipo di errore ci si trova. Nella maggior parte dei casi tale decisione dipende dalla quantità di errore presente in una data soluzione di *scaling*. Una piccola porzione di errore viene di solito attribuita a fluttuazioni relativamente poco importanti, mentre una grande porzione di errore viene spesso interpretata come una prova della mancanza di validità del particolare modello di *scaling* nella sua applicazione ai dati. La quantità di errore in una soluzione di *scaling* è misurata dalla misura di bontà di adattamento definita all'interno della particolare procedura.

L'interpretazione degli errori di *scaling* fornisce importanti informazioni sapendo che

- i criteri di *scaling* danno risposte sui modi in cui gli oggetti differiscono tra loro,
- le tecniche di *scaling* forniscono misurazioni di attributi la cui presenza è stata precedentemente accertata.

1.3 LA VERIFICA DEL MODELLO: I CRITERI DI SCALING

In generale, per verificare il modello di *scaling* e gli assunti previsti si procede osservando empiricamente il livello di adattamento del modello ai dati ottenuti dall'applicazione ad oggetti del mondo reale ovvero osservando se le proprietà dello *scaling* specificate dal modello sono osservabili nei dati. In particolare con l'applicazione di un modello di misurazione si verifica l'adattamento del modello ai dati, sapendo che:

$$dati = modello + residuo$$

Lo scarto tra dati e modello è attribuito alla presenza dell'errore; come sappiamo il residuo viene ritenuto troppo elevato quando esso non può essere attribuito al caso ma rappresenta una reale

[4] Alcuni analisti prendono in considerazione anche una particolare proprietà delle procedure di *scaling* detta *vulnerabilità*; questa si riferisce alla tolleranza di una procedura agli errori di *scaling* cioè la quantità d'errore che pur presentandosi non ostacola il raggiungimento di una soluzione di *scaling*. Le procedure che sono molto vulnerabili (come, secondo Coombs, è la tecnica di *unfolding* unidimensionale) sono più utili come criteri di *scaling* in quanto indicano in modo immediato se un particolare modello dimensionale funziona o meno nell'applicazione empirica. Le procedure meno vulnerabili (come lo *scaling* Likert) sono più utili come tecniche di *scaling* in quanto consentono la costruzione di una scala anche nei casi in cui si presenta una quantità consistente di errore. A tale proposito occorre notare che le routine di *scaling* computerizzate producono e indicano la soluzione di *scaling* con il migliore adattamento indipendentemente dalla quantità di errore che si può presentare.

divergenza tra valori attesi (modello) e osservati (dati). L'obbiettivo è quello di rendere il residuo più piccolo possibile ovvero di rendere minimo il contributo della componente residua. La maggior parte dei procedimenti di verifica dei modelli di *scaling* cercano di mettere in evidenza o di valutare la presenza della componente residua, anche se non è sempre facile riuscire a valutarne la dimensione.
Il residuo può essere dovuto principalmente a due componenti:

- errore casuale o *disturbo* (*e*),
- errore sistematico o *bias* (*b*).

Quindi:

$$dati = modello + (e+b)$$

Non essendo facile valutare la dimensione dell'errore, ancor più difficile sarà riuscire a distinguere esattamente tra le due componenti residuali. Uno dei problemi è quello di capire quando tale residuo può essere considerato errore casuale e quando deve essere interpretato invece come frutto di una scorretta operazionalizzazione del costrutto da misurare (definizione errata delle varaibili) ovvero quando è frutto di una rappresentazione errata della realtà da parte del modello.
La verifica del modello, fatta sottoponendo lo strumento definito dal modello di *scaling* ad analisi di attendibilità e di validità, prevede l'organizzazione di un "esperimento"; tali esperimenti devono soddisfare alcune condizioni definite, all'interno delle procedure di verifica, *standard di validazione*; tali standard riguardano:

- la definizione dei campioni per la sperimentazione,
- la definizione delle procedure di rilevazione
- la definizione dell'approccio alla valutazione dell'adattamento del modello.

➢ CAMPIONE PER LA SPERIMENTAZIONE

In teoria la sperimentazione per la messa a punto di uno strumento dovrebbe essere condotta e riguardare l'universo (di casi e/o di variabili)[5]. Per superare gli ostacoli che impediscono la rilevazione totale di tutti gli oggetti, la fase di validazione e di messa a punto di uno strumento richiede la definizione di un modello statistico che consenta l'applicazione su un *campione* (*campione per la sperimentazione*) tratto dall'universo oggetto della misurazione; per questo si parla di

- *campioni di casi*, comprensivi di tutta la variabilità dei casi della popolazione sulla quale verrà utilizzato lo strumento validato; tali campioni sono definiti secondo i classici modelli statistici induttivi;
- *campioni di item*, comprensivi della variabilità degli item considerati teoricamente significativi.

Nella definizione del campione è importante tener conto di due caratteristiche:

- Rappresentatività: il campione dovrebbe essere rappresentativo della popolazione di oggetti cui è rivolta l'applicazione dello strumento[6].
- Dimensione: è importante che l'analisi per la taratura dello strumento avvenga su campioni composti da un numero piuttosto alto di oggetti in modo da minimizzare la fonte di errore dovuta all'errore di campionamento, sappiamo infatti che maggiore è la dimensione del

[5] Ciò è una necessità in quanto, data la rilevanza delle differenze tra gli oggetti, l'affidabilità di uno strumento non può essere stabilita osservandone uno solo; è necessario per questo identificare con criteri appropriati una pluralità di oggetti. In altre parole la verifica dell'affidabilità dello strumento di misurazione (scala) richiede un riscontro sulla consistenza delle misure multiple non con un singolo oggetto ma su gruppi di oggetti. Per fare ciò si assume che l'*affidabilità di uno strumento sia la stessa per tutti gli oggetti del gruppo considerato e possa essere espressa da una unica misura (coefficiente di affidabilità)*.

[6] Non sempre nella ricerca sociale tale requisito viene soddisfatto. Gli esperimenti di affidabilità spesso (il più delle volte per mancanza di sufficienti risorse) vengono condotti su campioni di convenienza, costituiti dai soggetti più facilmente reperibili. Tale approccio però non consente di calibrare lo strumento rispetto ad una particolare popolazione.

campione, più precisa è la stima[7].

Nel caso in cui la sperimentazione riguardi due universi (per esempio casi e item) è virtualmente e teoricamente impossibile prendere in considerazione simultaneamente entrambe le dimensioni di campionamento senza addentrarci in grosse complessità statistiche; anche la considerazione di una sola delle due dimensioni risulta particolarmente complessa. Nella pratica spesso si cerca di prendere in considerazione una dimensione di campionamento in maniera esplicita e l'altra come possibile influenza sui risultati sperimentali. Tale necessario modo di procedere non altera il lavoro empirico. E' importante che la generalizzazione dei risultati venga fatta nell'ambito dei campionamenti effettuati (sui casi o sugli item), rimandando a studi e lavori successivi l'ampliamento delle applicazioni e l'estensione delle generalizzazioni. Una possibile accortezza è quella di riservare il campionamento ad una delle due dimensioni mantenendo l'altra più estesa possibile: in questo modo l'errore di campionamento, anche se presente, influenza maggiormente solo una delle due dimensioni.

- PROCEDURE DI RILEVAZIONE E DI RACCOLTA DATI

  La procedura di raccolta dei dati dovrebbe essere uguale a quella che si pensa sarà utilizzata una volta che lo strumento di misurazione sia stato validato. In altre parole le circostanze in cui viene svolto l'esperimento di validazione devono essere identiche a quelle che verranno utilizzate nell'applicazione dello strumento validato. Ciò deve consentire il rispetto della condizione secondo la quale la misurazione di un oggetto non deve essere influenzata da quella di altri. Vedremo quali possono essere i fattori e le condizioni relative alle procedure di somministrazione che possono influenzare il livello di affidabilità.
- VALUTAZIONE DELL'ADATTAMENTO DEL MODELLO

  La valutazione del modello viene fatta principalmente attraverso la verifica degli assunti sottostanti il modello. Indipendentemente dal modello scelto, per verificare se il livello di adattamento è accettabile, è necessario seguire determinati criteri statistici che consentono di indicarne la rilevanza reale; a tal fine in molti casi è possibile scegliere tra diverse opzioni al fine di massimizzare il livello di adattamento.

  Nel tentativo di raggiungere migliori livelli di adattamento è possibile selezionare gli item che, contenendo un livello di errore troppo elevato, possono essere esclusi. Tali tentativi possono anche valutare la rilevanza del campione osservato rispetto al modello di *scaling* definito. Un esempio di osservazioni che possono influenzare in maniera negativa il livello di adattamento sono gli *outlier* (ovvero individui se seguono un modello di *scaling* anomalo). Tali sforzi assicurano che i risultati siano robusti e stabili.

  Come vedremo la verifica dei modelli può assumere caratteristiche diverse; può essere, per esempio, di tipo deterministico o probabilistico; può richiedere il calcolo di un coefficiente oppure può comportare una valutazione più complessa.

1.3.1 Verifica della dimensionalità

L'applicazione di un criterio di *scaling* può essere utile per ottenere risultati per diverse strutture dimensionali. In genere l'analisi comincia con la rappresentazione geometrica più semplice, ovvero quella unidimensionale. Se i dati risultano coerenti con il modello (la dimensione degli errori di *scaling* è relativamente piccola), il criterio è stato soddisfatto in caso contrario si procede alla verifica di una struttura bidimensionale; anche in questo caso l'analisi del criterio termina se la verifica dell'adattamento produce un livello soddisfacente altrimenti, in presenza di un'inaccettabile quantità di errore, si procede alla verifica di un modello con soluzioni dimensionali superiori.

Un tale tipo di approccio è giustificato dall'assunto secondo il quale una singola caratteristica

[7] Ricordiamo che l'errore standard di stima di un parametro statistico è inversamente correlato alla radice quadrata della dimensione del campione.

dovrebbe condurre alla variabilità osservata tra gli oggetti studiati; in questo caso i punti-oggetto possono essere posizionati lungo una singola retta. In caso contrario la variabilità dovrà essere attribuita all'esistenza di molte caratteristiche che operano simultaneamente: il modello geometrico è multidimensionale in quanto vengono utilizzate più coordinate per stabilire le posizioni relative dei punti-oggetto.
Tale strategia progressiva, che procede da una soluzione unidimensionale a soluzioni multidimensionali sempre più complesse, rappresenta uno standard in molti approcci di *scaling* (come vedremo nel caso dei modelli fattoriali e del *multidimensional scaling*).
L'approccio progressivo si presta però ad alcune critiche; occorre infatti osservare che i modelli multidimensionali richiedono assunti che non riguardano quelli unidimensionali. Per esempio, le soluzioni multidimensionali assumono che tutte le dimensioni operano simultaneamente nel contribuire alle differenze tra gli oggetti da "scalare"; parallelamente a ciascun oggetto si attribuisce una coordinata per ogni dimensione contenuta nello spazio.
Alcuni hanno puntualizzato che tali assunti sono problematici. Anche se un insieme di oggetti possiede *K* caratteristiche oggettive, non esiste alcuna particolare ragione per la quale tutte le caratteristiche siano utilizzate per differenziare gli oggetti tra loro.
Per modellare tale situazione, l'analista potrebbe verificare, all'interno di un singolo insieme di oggetti, la presenza di *rappresentazioni unidimensionali multiple*. Dopo aver individuato sottoinsiemi di oggetti si verifica la presenza di una struttura unidimensionale per ciascuno di essi. Tale approccio genera spesso molte dimensioni per un singolo insieme di dati. Comunque i risultati non possono rilevare realmente una struttura multidimensionali in quanto per ciascun sottoinsieme vi è comunque solamente una singola rilevante dimensione.
Per identificare i sottoinsiemi di oggetti è possibile utilizzare più procedure di *scaling*, consentendo la verifica di dimensioni multiple.
Gli approcci allo *scaling* unidimensionale multiplo forniscono degli interessanti strumenti per integrare l'eterogeneità nelle osservazioni empiriche con soluzioni di *scaling* parsimoniose. In questo senso sono un'utile alternativa ai modelli multidimensionali relativamente più complessi.

1.3.2 Risultati dello scaling e interpretazione delle dimensioni

Una procedura di *scaling* applicata ad un insieme di dati produce virtualmente una struttura geometrica. D'altra parte le procedure di *scaling* non possono dare alcuna indicazione sul significato dei risultati. Solamente l'analista può determinare se le dimensioni della struttura spaziale corrispondono in modo sostanziale alle caratteristiche degli oggetti. E' sempre importante ricordare che le dimensioni in qualsiasi soluzione di *scaling* rappresentano semplicemente un sistema di coordinate utilizzato per posizionare un insieme di punti; in questo senso esse possono o non possono avere un significato sostanziale. Il rischio è che il ricercatore sia tentato a forzare un significato in una dimensione semplicemente perché la dimensione esiste, producendo così risultati non validi e senza senso. Se le posizioni dei punti appartenenti ad un insieme non possono essere interpretate in termini di una o più caratteristiche degli oggetti, sarà necessario prendere in considerazione la possibilità che la variabilità nelle osservazioni non è conforme ad un singolo sistematico modello. Allo stesso modo all'aumentare del numero di dimensioni richieste per ottenere un modello ragionevole (ovvero il rapporto tra il numero di dimensioni e il numero di oggetti si avvicina a 1) è necessario in primo luogo chiedersi se vi è una reale struttura sottostante gli oggetti.
Infine, anche nei casi in cui non sia possibile ottenere uno spazio con un numero di dimensioni non superiore a tre, la rappresentazione visiva dei risultati dello *scaling* può comunque essere vantaggiosa ai fini interpretativi. Per questo, le soluzioni di *scaling* con poche dimensioni sono preferite a quelle con tante dimensioni. Ciò è comunque coerente con il principio scientifico di spiegazioni parsimoniose. L'analisi visuale consente inoltre di individuare modelli negli oggetti che

non necessariamente possono essere trovati con altri metodi. Una rappresentazione grafica sostituisce molte spiegazioni verbali. Le presentazioni grafiche sono sempre più facili da comunicare rispetto a quelle numeriche, anche quando contengono esattamente la stessa informazione. Tali considerazioni pratiche spingono gli analisti verso soluzioni con tre o meno dimensioni, anche se gli strumenti informatici consentono sempre più di superare tali limiti.
La distinzione tra modelli *scaling* unidimensionali e multidimensionali è essenzialmente artificiosa. La dimensionalità si riferisce semplicemente al numero di differenze rilevanti in un insieme di dati. I quattro tipi di dati sono insensibili ad essi: le relazioni di dominanza e di prossimità tra coppie di punti possono verificarsi sia lungo un'unica dimensione che in uno spazio definito da più dimensioni. Naturalmente i metodi e gli algoritmi specifici possono cambiare a seconda che l'applicazione sia unidimensionale o multidimensionale; tali applicazioni rappresentano però semplici strumenti verso un fine. In ogni analisi di *scaling* l'obiettivo è di modellare la variabilità nelle osservazioni nel modo più accurato possibile. Ciò potrebbe richiedere più dimensioni oppure potrebbe essere sufficiente una singola dimensione. La scelta tra le due situazioni rappresenta una questione empirica.
Fin di primi decenni del Novecento sono stati definiti diversi modelli di *scaling* unidimensionali per ciascuno dei quali sono state proposti diversi criteri ed applicate varie tecniche che hanno avuto origine da studi di diversi ricercatori (Thurstone, Bogardus, Likert, Guttman, ecc.); lo sviluppo di modelli teorico-matematici ha consentito ulteriori sviluppi soprattutto nell'ambito della misurazione multidimensionale. Tra i modelli di *scaling* multidimensionali più diffusi ricordiamo il modello fattoriale e i modelli di *Multi-Dimensional Scaling* (*MDS*), che comprendono anche il modello fattoriale non lineare, delle preferenze individuali, di *unfolding*, delle differenze individuali, *biplot*.

1.4 TEORIA DEI DATI E CRITERI DI SCALING

Il principale obiettivo di una teoria dei dati è quello di razionalizzare l'uso delle procedure di *scaling*; in altre parole una teoria dei dati consente di scegliere tra i diversi metodi di *scaling* quello che si presenta come il più appropriato ad una determinata situazione. La scelta di una procedura di *scaling* dipende sempre dall'interpretazione che il ricercatore fa delle osservazioni (natura dei dati).
Vedremo ora quali sono le procedure di *scaling* più appropriate ai diversi tipi di dati, focalizzando l'attenzione, più che sulle tecniche di *scaling*, sui modelli ovvero sulle rappresentazioni astratte della variabilità all'interno delle osservazioni.
Non esiste un solo tipo corretto di dati che devono essere estratti da un certo insieme di osservazioni empiriche. L'interpretazione dei dati è sempre basata su una combinazione di considerazioni sostanziali (quale interpretazione delle osservazioni ha più senso) e obiettivi analitici (quale procedura di *scaling* produce il tipo di informazione desiderata).
La natura dei dati non è mai predeterminata ma dipende dall'interpretazione che il ricercatore fa delle osservazioni. Interpretazioni differenti conducono ad applicazioni di diverse procedure di *scaling* che influiscono direttamente sul tipo di informazione che è estratto dall'analisi. Il ricercatore deve decidere quale interpretazione è più appropriata e coerente. Ciò rappresenta un'importante componente creativa della ricerca empirica nelle scienze sociali.

- Dati del tipo *stimolo-unico*. Essendo i dati più comuni, non meraviglia che per essi sia stata concepita la maggior parte dei modelli di *scaling*, tra i quali ricordiamo il modello additivo, i modelli cumulativi (deterministico e probabilistico) e il modello fattoriale.
- Dati del tipo *confronto di stimoli*. Questo tipo di dati è relativamente poco utilizzato nelle situazioni non-sperimentali e si rivela il più delle volte poco interessanti ai fini dello *scaling*. Assumere l'esistenza di uno standard comune comporta che sia possibile collocare lungo una dimensione una serie di oggetti sulla base del confronto reciproco rispetto allo standard. Tra i modelli di scaling che utilizzano questo tipo di dati ricordiamo quello dei confronti

accoppiati di Thurstone, i metodi psicofisici, il *Q-Sort*.

- Dati di *somiglianza*. I dati di questo tipo trovano i loro modelli di riferimento nel *multidimensional scaling.*
- Dati del tipo *scelta di preferenza*. I dati di questo tipo vengono di solito analizzati con il modello detto *unfolding*.

L'osservazione delle caratteristiche che contraddistinguono i modelli di *scaling* consente di identificare alcuni criteri per una loro classificazione. In generale è possibile riconoscere principalmente quattro criteri:

a. la causa della variabilità dei dati rilevati (Torgerson, 1958),
b. l'errore di misurazione (Torgerson, 1958),
c. la dimensionalità,
d. il tipo di dati e la tecnica di *scaling*

Modello di *scaling*			Criterio di *scaling* (verifica del modello)	Tecnica di *scaling* (come fare ad ottenere i dati)	Tipo di dati	Dimensionalità	Punteggio finale attribuito a
Additivo			Approccio classico (confronto tra due misurazioni)	Non-comparativa	Stimolo-unico	Uni	Casi
Cumulativi-Deterministici	Guttman		Analisi dello scalogramma (riproducibilità, scalabilità, predicibilità)	Non-comparativa o comparativa	Stimolo-unico	Uni	Casi e stimoli
	Modelli alternativi	*Multidimensional Scalogram Analisys* (MSA)	Verifica della regionalità e contiguità			Bi	Casi e stimoli
		Partial Ordered Scalogram Analysis (POSA)	Verifica della corretta rappresentazione			Bi	Casi e stimoli
Cumulativi-Probabilistici	Monotoni (con uno o più parametri)		• Stima dei parametri (massima verosimiglianza) • Verifica dell'adattamento del modello (analisi del *misfit* e dei residui)	Non-comparativa	Stimolo-unico		Casi e stimoli
Thurstone				Comparativa (confronti accoppiati/*rank-order*)	Confronto	Uni	Stimoli
Perceptual Mapping	*Multidimensional scaling*		Bontà di adattamento tra prossimità e distanze (stress, alienazione)	Comparativa (confronti accoppiati)	Somiglianza	Multi	Stimoli
	Unfolding			Comparativa	Preferenza	Uni e Multi	Stimoli e casi
Modello congiunto			Bontà di adattamento tra ordinamento degli stimoli osservato e stimato dal modello (*part-worth*)	Comparativa (*rank-order*)	Preferenza	Multi	Stimoli
Metodologia *Q*				Comparativa (*rank-order*, *rating* comparativo)	Confronto		Stimoli

1.5 DOPO LA VERIFICA DEL MODELLO: L'INDIVIDUAZIONE DEI VALORI-NORMA

La verifica del modello consente e autorizza l'utilizzo del punteggio che sintetizza in uno o più valori le misure multiple. Come sappiamo il punteggio totale consente di posizionare ciascun oggetto misurato sul continuum che rappresenta la caratteristica rilevata. Per rendere operativo il procedimento di *scaling* può essere necessario individuare dei criteri che consentano di interpretare tale posizione rispetto alla caratteristica da misurare.
In genere mentre l'interpretazione dei punteggi estremi appare piuttosto chiara, più difficile è l'interpretazione di tutti gli altri; si pensi a tale proposito a quanto sia problematico identificare e interpretare il punto centrale di tale continuum che non necessariamente corrisponde e può essere interpretato come il punto di equilibrio tra i due estremi. Se, per esempio, la scala misura il livello di depressione, è possibile dire che i punteggi estremi indicano da una parte la presenza al massimo livello di tale caratteristica dall'altra la sua completa assenza ma non è possibile dire (al termine del procedimento di validazione) cosa indicano i punteggi intermedi.
La difficoltà deriva anche dal fatto che il punteggio prodotto, così come la verifica del modello di *scaling*, non è indipendente dal campione utilizzato per la validazione.
Quando lo strumento valicato deve essere utilizzato a fini diagnostici e non solo descrittivi è necessario procedere ad un ulteriore procedimento di validazione sulla base di uno schema di riferimento, che consneta di individuare i cosiddetti *valori-norma*, ovvero punteggi di riferimento che consentano di interpretare i punteggi individuali.
Per l'individuazione dei *valori-norma* non esiste una procedura standard; quasi sempre però si richiede l'utilizzo di variabili esterne; per questo motivo il procedimento spesso si identifica con quello di verifica della validità dello strumento.
L'impegno necessario per la definizione dei *valori-norma* è notevole e spesso può richiedere la stessa mole di lavoro necessaria per la costruzione della scala stessa.
La principale informazione sulla quale si base tale procedimento è quella relativa alla distribuzione di frequenza[8]. Nel caso in cui sia possibile disporre deei dati relativi a campioni diversi è importante poter osservare e confrontare la forma delle diverse distribuzioni. In alcuni casi, se il tipo di costrutto lo consente, può essere utile definire punteggi-norma diversi per i diversi gruppi (per esempio si possono definire livelli diversi per maschi e femmine).
In genere i valori-norma possono essere espressi secondo diverse forme

- punteggi standard: ricordiamo che oltre alla classica procedura di standardizzazione dei dati a volte ne viene utilizzata un'altra che consente di riferire i punteggi originali ad una distribuzione con media 500 ed deviazione standard 100;
- percentili: ricordiamo che un percentile consente di indicare la percentuale di casi che, nel campione sul quale è stato messo verificato il modello di *scaling*, si è trovato al di sotto di un particolare punteggio; se, per esempio, l'80% dei casi ha un punteggio minore di 122, un caso che registra un punteggio di 120 è all'80° percentile; i percentili possono naturalmente essere utilizzati su distribuzioni continue;
- punteggi normalizzati: ricordiamo che in quest'ultimo caso i valori-norma consentono particolari interpretazioni: se il punteggio ottenuto da un caso è superiore alla media nella misura di 2 volte la deviazione standard; ciò vuole dire che solo il 2.2% dell'intero gruppo ha ottenuto lo stesso punteggio.

Infine occorre dire che, anche se i *valori-norma* forniscono utili informazioni in senso valutativo e diagnostico, essi non sono assolutamente essenziali per un utilizzo descrittivo dello strumento.

[8] Come sappiamo, la natura delle distribuzioni può essere descritta calcolando alcune statistiche descrittive come la media, la deviazione standard, l'asimmetria, la curtosi.

2. IL MODELLO ADDITIVO

2.1 OBIETTIVI E ASSUNTI

Il modello *additivo*[1] ha l'obiettivo di misurare, classificare e ordinare gli oggetti rispetto all'attributo al fine di individuare le differenze individuali rispetto alla caratteristica[2]; alla base del modello vi sono due importanti assunti riguardanti la natura degli indicatori:

a. *l'insieme degli indicatori misura solamente un attributo* ovvero, facendo riferimento ad un modello fattoriale, gli indicatori linearmente combinati dovrebbero essere correlati solamente con un singolo fattore comune[3];
b. *ciascun indicatore è monotonamente correlato al continuum dell'attributo latente*: ciò vuol dire che nel caso si misuri, per esempio, un atteggiamento, più favorevole (sfavorevole) è l'atteggiamento del soggetto, maggiore (minore) è il suo punteggio atteso.

Dati tali obiettivi, questo modello di *scaling* richiede:

- indicatori dello stesso tipo selezionati sulla base della loro capacità nel discriminare tra casi e non sulla base delle loro posizioni relative sul continuum,
- livelli di classificazione dello stesso tipo associate a tutti gli indicatori; le diverse risposte di ciascun caso devono essere combinate in modo tale che sia possibile rappresentare differenze valide e affidabili tra i casi.

Tale modello può essere visto secondo due prospettive:

- Prospettiva geometrica: l'approccio additivo assume che uno degli insiemi di punti varia sistematicamente rispetto alla dimensione, mentre l'altro fluttua in modo casuale. In altre parole:
  - le variazioni sistematiche osservate nelle reazioni dei casi agli indicatori[4] è attribuita alle differenze individuali,
  - gli indicatori vengono considerati *misure ripetute*.

  Per ciascun oggetto nel primo insieme si sommano i valori di tutti gli oggetti nel secondo insieme in modo tale che le fluttuazioni si annullano a vicenda fornendo così una stima accurata

[1] Tale modello è molto utilizzato nella rilevazione del soggettivo (atteggiamenti, disposizioni o opinioni personali, attitudini, ecc.). In genere la serie di item definiti è rappresentata da affermazioni rispetto alle quali i soggetti riferiscono il proprio grado di accordo su una scala ordinale composta, in genere, da cinque livelli: "molto d'accordo", "d'accordo", "indeciso", "in disaccordo", "molto in disaccordo" (scala Likert); a ciascun livello è associato un valore, nell'ordine da 1 a 5 o da 0 a 4. Assumendo che le risposte a ciascuna affermazione siano tra loro equivalenti è possibile calcolare un punteggio totale individuale sommando i valori associati a tutte le affermazioni.

[2] Il modello additivo è generalmente quello più utilizzati nello *scaling* di soggetti rispetto a tratti psicologici.

[3] Il modello additivo presenta alcuni tratti comuni all'approccio fattoriale; in particolare gli assunti comuni sono:

- gli indicatori sono tra loro indipendenti e legati al tratto latente (punteggio vero),
- gli errori di misurazione non sono correlati tra loro, né con la dimensione latente,
- gli errori di misurazione sono dovuti al caso.

Infine possiamo aggiungere che il modello fattoriale consente di superare la condizione di parallelismo, che assume che tutti gli indicatori siano tra loro equivalenti, vista nel caso del modello additivo; attraverso il *factor loading* è infatti possibile sapere quanto ogni indicatore contribuisce alla formazione e composizione dei fattori.
I due approcci si differenziano rispetto alla definizione di "errore": nel modello fattoriale l'errore è formato da due componenti, una casuale, come per il modello additivo, l'altra attribuita alla variazione caratteristica di ciascun indicatore, non presente nel modello additivo; occorre però ricordare che il modello fattoriale consente di stimare solo la somma di tali due componenti (*unicità*).

[4] In questo caso gli indicatori sono rappresentati da variabili o item.

della posizione del punto del primo insieme lungo la dimensione sottostante. Notare che i punti che rappresentano gli oggetti nell'insieme non scalato non sono fissati lungo la dimensione; essi variano in modo marcato da una riga all'altra della matrice dei dati.

- Prospettiva di misurazione: gli item individuali che vengono sommati per produrre la scala rappresentano funzioni di livello ordinale della dimensione latente. In altre parole i valori numerici assegnati alle righe della matrice dei dati sono correlati in modo monotono alla caratteristica sottostante. Sommare gli item comporta sommare le funzioni. Le funzioni monotone sommate dovrebbero essere lineari, in quanto le peculiarità delle specifiche funzioni monotone degli item dovrebbero compensarsi. Come già sappiamo se vi è una funzione specifica (lineare in questo caso) tra la caratteristica sottostante e un insieme empirico di assegnazioni numeriche, è stato raggiunto un livello di misurazione a intervalli.

Esistono diverse tecniche di *scaling* che, pur essendo trattate come differenti, rispondono al modello additivo ovvero rappresentano manifestazioni differenti dello stesso modello di base[5].

Dati e matrici. Le misure costruite sulla base di questo modello semplificano la rappresentazione delle osservazioni empiriche sommando i livelli in almeno uno dei modi di una matrice *multi-way multi-mode*; nella seguente figura una matrice *n*k* viene ridotta sommando le colonne all'interno delle righe della matrice; il risultato è una matrice *n**1 contenente punteggi di scala per gli oggetti rappresentati nelle righe; il modello di *scaling* riguarda i punti per gli *n* oggetti posizionati lungo la dimensione sommando i *k* oggetti all'interno di ciascuna riga:

		Matrice di input *V: two-way, two-mode* v_{ij} : presenta il grado in cui l'oggetto *i* domina l'oggetto *j*					
		Oggetti-colonna (variabili)					
		1	2	…	*j*	…	*k*
Oggetti-riga (unità)	1	v_{11}	v_{12}	…	v_{1j}	…	v_{1k}
	2	v_{21}	v_{22}	…	v_{2j}	…	v_{2k}
	…	…	…	…	…	…	…
	i	v_{i1}	v_{i2}	…	v_{ij}	…	v_{ik}
	…	…	…	…	…	…	…
	n	v_{n1}	v_{n2}	…	v_{nj}	…	v_{nk}

		Matrice di output X x_i : fornisce la stima della posizione dell'oggetto *i* lungo la dimensione
		Punteggi di scala
Oggetti-riga (unità)	1	x_1
	2	x_2
	…	…
	i	x_i
	…	…
	n	x_n

[5] Le più conosciute sono quelle che utilizzano l'approccio Likert o il metodo Thurstone degli intervalli apparentemente uguali (in quest'ultimo approccio di solito si adotta la mediana come valore sintetico da attribuire all'oggetto di riga rispetto alla serie di oggetti di colonna.

2.2 LA VERIFICA DEL MODELLO

Nel caso del modello additivo l'obbiettivo dell'analisi è principalmente quello di verificare che l'insieme degli indicatori misuri la stessa caratteristica ovvero condivida una comune varianza (*omogeneità* e *solidità*). Per fare questo è possibile distinguere principalmente tra due approcci:

- due componenti (parallele o non parallele); in questo caso lo schema sperimentale prevede la stima dell'affidabilità attraverso la verifica dell'*equivalenza tra componenti*; tra le tecniche per individuare le componenti la più utilizzata è lo *split-half*; le misure di equivalenza sono il coefficiente *Spearman-Brown*, se le componenti sono parallele, e il coefficiente *Rulon*, se le componenti risultano *tau-equivalenti* o *tau-essenzialmente-equivalenti*;
- più componenti ciascuna delle quali viene considerata come una misurazione separata (item); in questo caso lo schema sperimentale di stima dell'affidabilità è indicato come *consistenza interna*; l'affidabilità è stimata verificando l'*equivalenza tra* n *componenti*; le misure di equivalenza sono i coefficienti *KR-20*, *KR-21*, *alfa*, L_2*;* se le *n* componenti sono parallele, *tau-equivalenti* o *tau-essenzialmente-equivalenti*.

APPROCCIO		STIMA DELL'AFFIDABILITA'			PROBLEMI
		TIPO DI VERIFICA	METODI	TECNICHE E STRUMENTI	
Componenti	Parallele	equivalenza tra le due componenti (split-half)	confronto tra componenti	▪ coefficiente Spearman-Brown ▪ correlazione tra componenti ▪ coefficiente Rulon	identificazione delle componenti parallele
	Non parallele				identificazione delle componenti
Analisi della consistenza interna		confronto tra le n componenti	confronto tra item e tra ciascun item e lo strumento intero	▪ correlazione tra item ▪ correlazione item-totale ▪ coefficienti alfa, KR-20, KR-21, L_1, L_2	individuazione della omogeneità degli item

2.2.1 L'affidabilità

2.2.1.1 Le componenti parallele

Sappiamo che nel metodo *test-retest* i punteggi sono ottenuti tenendo costanti 'strumento' e 'modalità di somministrazione' mentre con il metodo delle forme parallele si tengono costanti momento e modalità di somministrazione e si varia lo strumento. Con le componenti parallele è possibile stimare l'affidabilità dell'intero strumento correlando due parti confrontabili dello strumento; tale approccio, detto anche *split-half*, consente di evitare la maggior parte dei problemi incontrati nell'applicazione dei due precedenti metodi.
Il limite del metodo è rappresentato dal fatto che la correlazione può variare anche in modo considerevole a seconda della tecnica di suddivisione dello strumento; a tale proposito è possibile identificare principalmente due tecniche:

- *suddivisione degli item secondo l'ordine di somministrazione* (prima metà e seconda metà degli item); tale tecnica non è consigliabile in quanto ciascun raggruppamento di item può

risultare influenzato, compromettendo il livello di affidabilità, in modo differenziato da particolari fattori e condizioni soggettive e/o ambientali non sempre controllabili o eliminabili (stanchezza, fatica, confidenza, noia, ecc.);

- *separazione degli item numerati con cifre dispari da quelli numerati con cifre pari* (*odd-even method*); è la tecnica più utilizzata in quanto consente di superare i problemi osservati con la precedente tecnica: non potendo escludere i fattori e le condizioni non controllabili si fa in modo che influenzino le due componenti in misura equivalente.

Ottenere la stima dell'affidabilità secondo questo metodo non è semplice, in quanto non sempre la suddivisione produce due componenti perfettamente parallele; è per tale motivo che la stima dell'affidabilità prevede due approcci diversi:

- Stima dell'affidabilità per componenti parallele: si procede correlando le due serie di punteggi ottenuti con i due sottogruppi di item (coefficiente di equivalenza); in realtà tale correlazione sottostima il valore di *rho* in quanto l'affidabilità è direttamente correlata con la numerosità del campione di item; questo vuol dire che il risultato di tale correlazione rappresenta l'affidabilità di due gruppi di item (strumenti paralleli) e non l'affidabilità dell'intero strumento; per questo il coefficiente di equivalenza viene sottoposto ad un correttivo che corrisponde ad una delle applicazioni più frequenti dell'equazione *Spearman-Brown* (di cui si parlerà più avanti) indicata come *formula Spearman-Brown per componenti parallele*:

$$rho_x = \frac{2r_{12}}{1+r_{12}}$$

dove
r_{12} correlazione tra le due componenti.
che produce una stima dell'affidabilità dell'intero strumento.

- Stima dell'affidabilità per componenti non parallele: le componenti si dicono "non parallele" quando i punteggi di ciascuna di esse sono uguali o diversi in quantità uguale per ciascun soggetto e quando gli errori standard di misurazione delle diverse parti non sono correlati tra loro. In questo caso le componenti sono dette, analogamente a quanto succedeva per gli strumenti, *tau-equivalenti* o *tau-essenzialmente-equivalenti*. Se si soddisfano i requisiti specificati, l'affidabilità può essere stimata utilizzando una formula riportata in letteratura ad opera di P.Rulon (nel 1939) ma attribuita a J.Flanagan (tutti studiosi di psicometria) che, utilizzando i punteggi osservati, produce una buona stima dell'affidabilità per l'intero strumento:

$$rho_x = 2\left(1 - \frac{\sigma^2_{y_1} + \sigma^2_{y_2}}{\sigma^2_x}\right)$$

dove
$\sigma^2_{y_1}$ e $\sigma^2_{y_1}$ varianza dei punteggi osservati delle due parti dello strumento
σ^2_x varianza dell'intero strumento.

Al contrario della stima dell'affidabilità per componenti parallele, qui non si richiede alcuna correzione che tenga conto della lunghezza dello strumento.
E' stato dimostrato che la formula di Rulon fornisce il limite inferiore dell'affidabilità dello strumento nel caso in cui le parti definite ricadono in una delle tre tipologie (componenti parallele, componenti *tau-equivalenti*, componenti *tau-essenzialmente-equivalenti*) ma anche nel caso in cui le componenti definite non ricadono in tali tipologie. Quindi più il risultato della formula di Rulon si avvicina ad 1, minore è l'incertezza riguardo all'affidabilità dello strumento anche quando le componenti dello strumento non soddisfano i requisiti di *parallelismo*.

2.2.1.2 La consistenza interna

L'analisi della *consistenza interna* (*internal consistency reliability*) mira a verificare l'entità della

componente comune a tutti gli item (omogeneità degli item); tale componente comune è attribuibile non solo alla capacità di un gruppo di item di misurare insieme qualcosa ma riflette la presenza di un sottostante costrutto comune. Gli item che riflettono una componente comune, dovrebbero condividere una comune varianza e, quindi, registrare alte correlazioni[6].
Le correlazioni tra gli item rappresentano il reciproco dell'errore; in altre parole la componente degli item non correlata con gli altri item è considerata "errore". Maggiore è la correlazione tra gli item, minore è la componente "errore".
Stimare l'affidabilità di uno strumento attraverso la verifica dell'omogeneità degli item richiede l'utilizzo di tutte le informazioni contenute nei punteggi dei singoli item; il procedimento è di tipo iterativo e richiede l'analisi delle correlazioni tra item e delle correlazioni tra ogni item e tutti gli altri; tali informazioni trovano la finale informazione nei coefficienti di affidabilità, i cui valori (compresi tra 0 e 1) sono funzione del numero di item e del livello di intercorrelazione. Vediamo tali coefficienti:

Kuder-Richardson Formula 20 e 21

Kuder e Richardson hanno puntato la loro attenzione su strumenti composti da item dicotomici-binari. Delle molte equazioni da loro studiate, la 20ª (1937) è divenuta quella più utilizzata:

$$KR-20=\left(\frac{k}{k-1}\right)\left(1-\frac{\sum p_i q_i}{\sigma_x^2}\right)$$

dove

k	numero di item
p_i	proporzione di risposte positive di ciascun item *i*
q_i	$1-p_i$
$p_i q_i$	varianza dell'item *i*
σ_x^2	varianza dell'intero strumento (punteggio totale).

Il risultato prodotto dal coefficiente *KR-20* può essere più correttamente interpretato come il limite inferiore del livello di affidabilità dello strumento.
Esiste un'altra versione del *KR-20* che impone una condizione molto restrittiva in quanto assume che per tutti gli item vi sia la stessa proporzione di risposte positive (media per le variabili dicotomiche):

$$KR-21=\left(\frac{k}{k-1}\right)\left(1-\frac{\mu_x-\mu_x^2/k}{\sigma_x^2}\right)$$

dove

k	numero di item
μ_x	media dei punteggi osservati dell'intero strumento
σ_x^2	varianza dell'intero strumento (punteggio totale).

Coefficiente *alfa*

Per stimare l'affidabilità di uno strumento nel 1951 Cronbach ha proposto l'utilizzo del coefficiente *alfa*, che confronta la varianza del punteggio totale con le varianze di tutti gli item.
In presenza di una totale mancanza di correlazione tra gli item, la varianza dello strumento corrisponde alla somma delle varianze di tutti gli item che lo compongono; un tale risultato indica che gli item osservati misurano aspetti specifici e non sono tra loro omogenei. *alfa* consente di confrontare tali varianze al fine di stimare il livello di affidabilità:

$$alfa=\left(\frac{k}{k-1}\right)\left(1-\frac{\sum \sigma_i^2}{\sigma_x^2}\right)$$

dove

k	numero di item
σ_i^2	varianza dell'item *i*

[6] E' questo un principio che è alla base anche dell'analisi fattoriale che, come vedremo, costituisce uno degli approcci alla misurazione multidimensionale.

σ_x^2 varianza dell'intero strumento (punteggio totale).

L'evidente analogia della formula di *alfa* con quella proposta da Kuder e Richardson consente di dire che il coefficiente *alfa* rappresenta la generalizzazione del *KR-20* al caso di strumenti composti da item non dicotomici.

Se si utilizza la matrice di correlazione tra gli item, il calcolo di *alfa* diviene:

$$alfa = \frac{k * r_m}{1 + r_m(k-1)}$$

dove

k numero di item

r_m media delle correlazioni tra tutti item.

Il coefficiente *alfa* può essere calcolato su dati sia in forma originale che standardizzata; in quest'ultimo caso si utilizza la matrice di varianza-covarianza[7] che, se calcolata su dati standardizzati, corrisponde alla matrice di correlazione con in diagonale valori unitari; in tal caso la formula per calcolare *alfa* diviene:

$$alfa = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum diag}{\sum tot}\right)$$

dove

k numero di item

diag elementi nella diagonale della matrice di varianza-covarianza (varianze dei punteggi dei singoli item)

Σtot somma di tutti gli elementi della matrice di varianza-covarianza.

Siccome nella matrice di correlazione in diagonale vi sono valori unitari, *Σdiag* corrisponde al numero di item (*k*).

Non è facile decidere se utilizzare dati in forma originale o in forma standardizzata; indicativamente se è possibile assumere per gli item uguaglianza tra le varianze della popolazione e quelle campionarie allora è possibile assumere che il calcolo di *alfa* non produca valori diversi utilizzando dati standardizzati o in forma originale. In caso contrario è più corretto utilizzare la prima versione della formula. Sia *alfa* che *KR-20* stimano l'affidabilità di uno strumento se le parti che lo compongono (gli item) sono parallele, *tau-paralleli* o *tau-essenzialmente-paralleli*, altrimenti fornisce un valore che può essere interpretato come limite inferiore dell'affidabilità.

Altri coefficienti

Prima del lavoro di Cronbach sul coefficiente *alfa*, L.Guttman nel 1945 aveva già provato a definire un coefficiente di affidabilità. In particolare Guttman era giunto a definire tre coefficienti L_1, L_2 e L_3 quest'ultimo formalmente corrispondente ad *alfa*. Guttman ha dimostrato che questi coefficienti stimano limiti inferiori dell'affidabilità anche quando non è possibile fare alcun assunto sulla natura della relazione tra le parti dello strumento.

Il più semplice tra i tre coefficienti, L_1, è definito nel seguente modo:

$$L_1 = 1 - \frac{\sum \sigma_i^2}{\sigma_x^2}$$

dove

σ_i^2 varianza dell'item *i*

σ_x^2 varianza dell'intero strumento (punteggio totale).

Il migliore coefficiente di stima del limite inferiore di affidabilità è considerato L_2, definito da Guttman nel seguente modo:

[7] Ricordiamo che in tale matrice i valori in diagonale rappresentano la varianza di ciascun item mentre il resto dei valori è rappresentato dalle covarianza tra tutte le possibili coppie di item.

$$L_2 = 1 - \frac{\sum \sigma_i^2}{\sigma_x^2} + \frac{\sqrt{\frac{k}{k-1} \sum_i \sum_j \sigma_{ij}^2}}{\sigma_x^2}$$

dove

k numero di item

σ_{ij}^2 covarianza tra i punteggi osservati per tutte le coppie di item ($i, j = 1,2,...,n$ con $i \neq j$)

σ_x^2 varianza dell'intero strumento (punteggio totale).

E' stato osservato che in genere L_2 produce un valore uguale o maggiore a quello ottenuto con il coefficiente *alfa*. Nella pratica il coefficiente più utilizzato è comunque *alfa* anche per la sua maggiore semplicità di calcolo. E' probabile che l'inserimento degli altri coefficienti nei più diffusi *package*, consentirà una loro maggiore diffusione. Infatti esistono *package* che consentono di calcolare alcuni o tutti tali coefficienti.

Occorre ricordare che i coefficienti L_1, L_2, *alfa* e *KR-20* non necessariamente producono un valore uguale o superiore a 0. Se la somma delle covarianze delle componenti è negativa, i valori dei coefficienti possono risultare negativi. In tal caso, naturalmente, i coefficienti forniscono praticamente un'informazione inutilizzabile sull'affidabilità dello strumento ma indicativa di una presenza di più dimensioni al di sotto delle diverse componenti (item). La seguente tabella riporta l'*output* per alcuni coefficienti relativi ad una scala.

INTERNAL CONSISTENCY DATA	
• SPLIT-HALF CORRELATION	.573
• SPEARMAN-BROWN COEFFICIENTI	.729
• GUTTMAN (RULON) COEFFICIENTI	.728
• COEFFICIENT ALPHA - ALL ITEMS	.715
• COEFFICIENT ALPHA - ODD ITEMS	.579
• COEFFICIENT ALPHA - EVEN ITEMS	.518

Come sappiamo in genere ci si attende un valore piuttosto alto di tali coefficienti per poter affermare che la scala è affidabile anche se non vi è concordanza tra gli autori sul valore minimo richiesto per ciascuno di tali coefficienti che può andare da .70 a .90.

Nel caso in cui sia stato registrato un valore non soddisfacente, si procede ad un'analisi approfondita dei singoli item per una loro selezione; tale selezione avviene sulla base di particolari caratteristiche.

2.2.2 Standard di affidabilità

Stabilire qual è il livello di affidabilità soddisfacente è un problema che dipende dal tipo di utilizzazione che deve essere fatto dello strumento sul quale è misurata l'affidabilità.

Tenendo presente che la messa a punto di uno strumento richiede la realizzazione di più esperimenti per la verifica dell'affidabilità, si può dire che per strumenti con obbiettivi prevalentemente descrittivi è possibile accontentarsi di un valore di affidabilità anche modesto; cercare in questi casi un valore di affidabilità maggiore può essere costoso in termini di tempo e di risorse non proporzionati agli obbiettivi dello strumento. Secondo alcuni autori (Nunnally, 1978) in questi casi il livello minimo accettabile di consistenza interna è 0.70. Quando però la misura dell'affidabilità riguarda strumenti le cui misurazioni prodotte dalla sua applicazione devono essere utilizzate per assumer importanti decisioni (si pensi al campo clinico o al campo sociale), il livello minimo di

affidabilità deve essere sufficientemente alto (0.90 o 0.95)[8].
Il mancato raggiungimento del livello di affidabilità auspicato può essere attributo

- alla scorretta formulazione del modello e della definizione del costrutto (per esempio errata individuazione della dimensionalità),
- alla presenza di item non corretti ovvero affetti da errore sistematico (item *biased*).

Mentre nel primo caso è necessario procedere ad una revisione del modello ed, eventualmente, ad una sua riformulazione nel secondo caso è necessario identificare gli item *biased* utilizzando strategia iterativa.

2.2.2.1 La selezione degli item

La strategia utilizzata per identificare gli item affetti da errore sistematico assume che gli item biased sono quelli che non condividono con gli altri la misurazione di una dimensione[9].
Per poter verificare tale condizione è necessario eseguire un procedimento iterativo i cui momenti sono:

a. calcolo del punteggio totale e osservazione della sua distribuzione di frequenza,
b. calcolo di indici relativi a ciascun item (indice di difficoltà, coefficiente di discriminazione),
c. calcolo e analisi delle correlazione tra tutti gli item e tra ciascun item e il punteggio totale (correlazione *item-totale*),

Al termine del procedimento si calcola il coefficiente di affidabilità; se i risultati del procedimento hanno individuato un item biased si prosegue ripetendo il procedimento stesso dopo aver eliminato tale item.[10]

A. Calcolo del punteggio totale
   In genere si calcola sommando semplicemente i valori che gli item hanno registrato per ciascun individuo. Il trattamento statistico del punteggio totale rappresenta un problema statistico risolto nell'ambito del più generale trattamento delle combinazioni lineari.
B. Calcolo degli indici individuali
   - *Proporzione di risposte corrette*:

[8] Una scala utilizzata per la stima dell'autosufficienza fisica di un gruppo a rischio (per esempio anziani) e i cui risultati devono consentire la programmazione di determinati servizi territoriali, deve necessariamente presentare un livello di affidabilità piuttosto alto.

[9] Date le sue caratteristiche, tale approccio si presenta particolarmente adatto nell'ambito dell'applicazione del modello additivo.

[10] Un altro approccio per l'identificazione degli item biased è quello basato sull'*Analisi della Varianza* (*ANalysis Of VAriance*, *ANOVA*), che, come sappiamo, rappresenta una tecnica statistica molto potente e utile nell'ambito di molti ambiti ma piuttosto delicata e fragile quando utilizzata nell'ambito dell'individuazione del *bias* (altre strategie sono più convenienti e appropriate a questo fine). D'altra parte in questo ambito l'ANOVA è ancora molto utilizzata a causa della sua popolarità e della sua familiarità e riveste un'importanza storica (i primi studi di individuazione del *bias* ponevano al centro tale tipo di analisi).
Secondo tale approccio, dopo aver sottoposto lo stesso gruppo di item a due o più campioni estratti dalla stessa popolazione, si cerca di indagare il significato delle differenze riscontrate nei punteggi individuali. In particolare l'attenzione è concentrata sull'*interazione gruppi*item* anziché sugli effetti principali.
Come sappiamo gli item che misurano tratti o attributi diversi per sotto-gruppi estratti dalla stessa popolazione violano l'assunto richiesto di unidimensionalità. Parlando, per esempio, in termini di difficoltà, la perdità dell'unidimensionalità attraverso i sotto-gruppi molto probabilmente indica variazioni nei diversi livelli di difficoltà indipendentemente dalla differenza dei livelli di difficoltà totale dei gruppi. Un item è considerato non affetto da *bias* quando il livello di difficoltà ("probabilità di successo") su un item è lo stesso per i soggetti di uguale capacità della stessa popolazione indipendentemente della loro appartenenza ad un gruppo. Quindi è possibile inferire la presenza di *bias* ognivolta che si osserva un'interazione significativa *gruppi*item*. Per una trattazione sistematica dell'analisi della varianza si può ricorrere ad un manuale di statistica metodologica.

$$P_i = \frac{y_i}{N}$$

dove

y_i numero di risposte considerate "corrette" all'item *i*

N numero soggetti

Negli item dicotomici, corrisponde alla media.

- *Coefficiente di affidabilità* dell'item: prodotto tra deviazione standard dell'item e correlazione item-totale:

$$s_i = r_i \sqrt{P_i(1 - P_i)}$$

- *Coefficiente di discriminazione*:

$$D_i = (P_a - P_b)_i$$

dove

P_a proporzione di risposte corrette all'item *i* di soggetti che sono nel 27% superiore dei punteggi totali

P_b proporzione di risposte corrette all'item *i* di soggetti che sono nel 27% inferiore dei punteggi totali.

Un'altra versione del coefficiente di discriminazione di un item è derivata da quella definita da Ferguson per un insieme di item (Guilford, 1954):

$$\delta = \frac{(n+1)\cdot\left(N^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} f_{ij}^2\right)}{nN^2}$$

dove

δ coefficiente di discriminazione

N dimensione del campione

n numero di item del gruppo

k numero di punteggi definiti per ciascun item

f_{ij} frequenza del *j*-esimo punteggio per l'*i*-esimo item

Per un item singolo diventa

$$\delta_i = \frac{N^2 - \sum_{j=1}^{k} f_j^2}{N^2}$$

dove

δ_i coefficiente di discriminazione per l'item *i*

f_j frequenza del *j*-esimo punteggio

C. <u>Calcolo e analisi delle correlazione tra tutti gli item e tra ciascun item e il punteggio totale (correlazione *item-totale*),</u>

Un primo indice della capacità dei singoli item di condividere la misurazione di una dimensione comune è rappresentato dal livello di intercorrelazione presente tra i singoli item. La media di tali correlazioni può indicare la dimensione della componente comune a tutti gli item. La dispersione di tali correlazioni rispetto alla media indica quanto gli item tendono a variare nel contribuire alla componente comune[11].

Un altro indice è rappresentato dal livello di correlazione tra ciascun item e l'insieme degli altri item. La media di tali correlazioni può rappresentare un indice del livello di omogeneità di contenuto: se gli item sono tra loro omogenei dovremmo attenderci alte correlazioni tra il punteggio totale e i singoli item. L'osservazione di bassi valori di correlazione *item-totale* rilevano la presenza di item che, dimostrando di non condividere alcuna dimensione con gli altri item, possono essere definiti biased.

[11] Occorre a tale proposito ricordare che se da una parte la registrazione di un alto livello di correlazione può indicare e rilevare anche una sovrapposizione tra gli item e rivelare una ridondanza di informazioni (item che misurano la stessa cosa) dall'altra la mancanza assoluta di sovrapposizione può far pensare ad item che non appartengono alla stessa dimensione.

E' possibile stabilire in anticipo il numero di item che devono andare a comporre lo strumento finale e quindi selezionare un numero corrispondente di item tra quelli che registrano le correlazioni più alte. E' possibile anche selezionare solo quegli item che registrano almeno un dato livello di correlazione (in genere 0.40).

Se un item registra una correlazione *item-totale* negativa vuol dire che non è classificato nella stessa direzione degli altri, ovvero che il contenuto dell'item è orientato in senso positivo quando gli altri lo sono in senso negativo, o viceversa; in questi casi l'item deve essere sottoposto a *riflessione*, invertendo il punteggio attribuito alle singole risposte dell'item in questione per uniformare la direzione delle risposte. Dopo la riflessione si procede nuovamente al calcolo del punteggio totale e delle correlazioni; in questi casi il numero di correlazioni positive aumenta come pure la dimensione media delle correlazioni.

Nella seguente tabella sono riportati i risultati di una delle iterazioni e relativi ad un gruppo di 10 item:

Item	Label	Item-total (*R*)	Item-reliability	Excl. item	
				R	*alpha*
1	VAR83	0.695	0.467	0.623	0.837
2	VAR84	0.701	0.671	0.594	0.837
3	VAR85	0.646	0.454	0.561	0.841
4	VAR86	0.711	0.502	0.637	0.835
5	VAR87	0.523	0.425	0.402	0.853
6	VAR88	0.770	0.729	0.682	0.828
7	VAR89	0.542	0.414	0.432	0.850
8	VAR90	0.591	0.602	0.448	0.853
9	VAR91	0.727	0.599	0.642	0.833
10	VAR92	0.724	0.592	0.639	0.833
alpha	0.854				

Nell'esempio il gruppo di item risulta essere piuttosto omogeneo; l'esclusione a turno degli item dal calcolo del coefficiente alfa non ne fa aumentare il valore.

Una particolare attenzione va dedicata al coefficiente che si intende adottare per il calcolo della correlazione *item-totale*; in particolare se

- l'item è dicotomico puro allora di utilizza il coefficiente *punto-biseriale*,
- l'item e dicotomizzato allora si utilizza il coefficiente *biseriale* (v. glossario).

Dopo l'identificazione e l'eliminazione dell'item (o degli item) biased si procede al calcolo del nuovo punteggio totale, delle correlazioni *item-totale*. Il procedimento iterativo termina quando i risultati osservati risultano soddisfacenti.

Come si è visto, gli item identificati come *biased*, in genere, vengono eliminati; in determinati casi è però anche possibile procedere in altro modo:

- ridefinendo e/o riconsiderando le categorie di risposta,
- riformulando la struttura verbale degli item e delle istruzioni,
- sostituendo l'item (l'item presenta problemi di validità).

In quest'ultimo caso può essere utile stimare il numero di item che dovrebbe essere aggiunto per raggiungere un livello accettabile di consistenza interna. A tale scopo è possibile utilizzare il coefficiente *Spearman-Brown (S-B)*

2.2.2.2 Il coefficiente Spearman-Brown

Tale coefficiente consente di stimare:

a. gli effetti sul livello di affidabilità di un aumento o di una diminuzione del numero di item, per esempio stimare l'affidabilità di uno strumento con un numero maggiore di item a partire dall'affidabilità determinata per lo strumento originale,

b. il numero di item necessari per raggiungere un determinato livello di affidabilità.

Dato il valore di *alfa* per uno specifico numero di item, tale formula indica gli effetti su *alfa* di aumento o di una diminuzione del numero di item. Tale procedimento si basa sull'assunto che gli

item aggiunti o eliminati presentano la qualità (affidabilità individuale) degli item iniziali; in caso contrario la formula può sovra-stimare o sotto-stimare il numero di item.

Stima dell'affidabilità di uno strumento con un numero diverso di item a partire dall'affidabilita dello strumento originale

Sapendo che l'affidabilità di uno strumento aumenta con il numero degli item, si può concludere che il principale modo per costruire strumenti più affidabili è quello di renderli più lunghi. In altre parole sappiamo che nei casi in cui è ragionevole aumentare la lunghezza dello strumento - tenendo presente che l'incremento non dovrebbe essere così grande da rendere inutilizzabile lo strumento - è possibile ottenere misurazioni maggiormente affidabili.
Il problema può essere quello di sapere quanto aumenta l'affidabilità con l'aumento del numero di item. Ciò ha condotto alla definizione di una particolare formula di Spearman e Brown detta *prophecy formula*. Proposta nello stesso momento dai due autori separatamente nel 1910, la formula consente di calcolare le variazioni del coefficiente di affidabilità di uno strumento in funzione dell'aumento (o della riduzione) del numero di item:

$$\textit{formula generalizzata di Spearman-Brown} = rho_x = \frac{k * rho_y}{1 + (k-1) rho_y}$$

dove

rho_y affidabilità nota dello strumento

rho_x affidabilità da determinare

k a / b

dove

a numero di item dello strumento su cui calcolare l'affidabilità

b numero di item dello strumento con l'affidabilità nota.

Non è necessario che il fattore della lunghezza k nell'equazione sia un numero intero.
Disponendo di uno strumento di affidabilità rho_y=0.60, il raddoppio del numero di item produce un'affidabilità di

$$rho_x = (2*0.60)/(1+0.60) = 1.20/1.60 = 0.75$$

mentre un quadruplicamento

$$rho_x = (4*0.60)/(1+3*0.60) = 2.40/2.80 = 0.857.$$

Il livello di incremento del valore del coefficiente diminuisce gradualmente mano a mano che il valore di affidabilità si avvicina a 1.00. Se il livello di affidabilità previsto con la formula non aumenta in misura interessante si può concludere che lo strumento si presenta sufficientemente solido. La formula può essere utilizzata anche per stimare gli effetti sull'affidabilità della *riduzione del numero di item*[12].
E' importante sottolineare che, in relazione al numero di item maggiore o minore, la precisione della stima ottenuta con la formula dipende principalmente dalla differenza del numero di item tra uno strumento e la sua versione allungata o ridotta. Ciò vuol dire che non ci si dovrebbe aspettare una stima molto precisa se a partire dalla conoscenza dell'affidabilità di un'area di 5 item pretendiamo di stimare l'affidabilità dell'area con 40 item o viceversa. Nel caso in cui si voglia stimare l'affidabilità di uno strumento cui s'ipotizzi di raddoppiare il numero di item, la formula diviene:

$$rho_x = \frac{2 rho_y}{1 + rho_y}$$

dove

rho_x affidabilità dello strumento con lunghezza raddoppiata

rho_y affidabilità dello strumento originario.

Tale caso particolare ha condotto alla più nota applicazione della formula *S-B* ovvero quella che consente di stimare l'affidabilità di uno strumento attraverso il modello delle componenti parallele (*split-half*):

---

[12] Supponiamo di avere uno strumento con un livello di affidabilità di 0.75; se il valore di k è uguale a 0.5, la nuova affidabilità sarà:

$$(0.5*0.75)/[1+(0.5-1)0.60] = (0.375)/(1-0.375) = (0.375) / (-0.625) = 0.60$$

$$rho_x = \frac{2r_{12}}{1 + r_{12}}$$

dove

rho_x affidabilità dell'intero strumento

r_{12} correlazione tra le due componenti parallele.

Per poter utilizzare correttamente la formula è necessario soddisfare i seguenti assunti:

- gli item aggiunti (o eliminati) hanno le stesse caratteristiche statistiche (affidabilità individuale) degli item iniziali, conseguentemente gli item hanno uguale effetto sull'affidabilità,
- la correlazione media dello strumento di partenza deve essere la stessa dello strumento ipotizzato.

Tali assunti non vengono soddisfatti quando gli item nuovi e quelli vecchi differiscono sistematicamente e in modo significativo

- nel contenuto (ovvero provengono da campi di contenuto diversi),
- nell'affidabilità (ovvero se la correlazione media in un gruppo è più alta di quella dell'altro).

Il nome, non privo di ironia, attribuito alla formula suggerisce quanto sia difficile soddisfare i requisiti richiesti e conseguentemente quanto il risultato ottenuto sia da considerare con le dovute cautele. Ciò vale soprattutto per la seguente applicazione.

Numero di item necessari per raggiungere un determinato livello di affidabilità

La formula *Spearman-Brown* è stata generalizzata anche per stimare il numero di item necessari per raggiungere un desiderato livello di affidabilità. Per determinare quanto dovrebbe aumentare il numero di item per raggiungere il livello di affidabilità desiderato la formula *S-B* viene trasformata nel modo seguente:

$$k = \frac{rho_x\left(1 - rho_y\right)}{rho_y\left(1 - rho_x\right)}$$

dove

rho_x affidabilità desiderata

rho_y affidabilità nota

k di quanto deve aumentare il numero di item per ottenere l'affidabilità desiderata

L'errore standard di k è dato da

$$\sigma_k = \frac{k\left(1 - rho_y^2\right)}{rho_y^2\sqrt{N}}$$

L'applicazione di tale formula conduce alla conclusione che per raggiungere un'affidabilità moderatamente alta si dovrebbe aumentare il numero di item spesso anche in modo considerevole (e, il più delle volte, impraticabile)[13].

2.2.2.3 Un altro approccio alla selezione degli item: Transformed Item Difficulties

L'approccio noto come *Transformed Item Difficulties* (*TID*) assume che un item è *biased* quando risulta più "difficile" per un gruppo rispetto all'altro: l'errore è indicato da una differenza

[13] Supponiamo di avere uno strumento composto da 20 item con una affidabilità di 0.50; secondo il modello proposto, volendo raggiungere un livello di affidabilità di 0.80, dovremmo aumentare il numero di item di:

k = [0.80 (1-0.50)]/[0.50 (1-0.80)] = 0.40/0.10 = 4

ovvero è necessario aumentare il numero di item di quattro volte (20*4=80 item!!).

Nel caso in cui lo strumento disponga di 40 item e presenti un'affidabilità di solamente 0.20 e si desideri un'affidabilità di almeno 0.80:

k = [0.80 (1-0.20)]/[0.20 (1-0.80)] = 0.64/0.04 = 4

ovvero sarebbe necessario arrivare a 40*4=160 item!!. Ma sappiamo anche che con uno strumento con un così basso livello di affidabilità è principalmente necessario mettere in discussione l'intera sua formulazione.

significativa tra gruppi rispetto alla difficoltà relativa dell'item. Tale approccio è concettualmente lineare e chiaro[14] anche grazie alla possibilità che offre di utilizzare tecniche grafiche per la rappresentazione delle difficoltà degli item; inoltre consente di rappresentare in forma di equazione sia i punteggi individuali (capacità) che quelli degli item (difficoltà).
L'ipotesi da verificare attraverso tale strategia non è solamente la presenza o l'assenza dell'interazione *gruppi*item*, anche se ciò non è fatto in senso formale di accettazione o di rifiuto dell'ipotesi nulla attraverso un criterio statistico, ma anche il livello relativo in cui determinati item possono variare tra gruppi.
Il procedimento è piuttosto semplice e si sviluppa nei seguenti momenti:

1. Determinazione dell'indice di difficoltà dell'item (per esempio *p*) per ciascuno dei gruppi da confrontare.
2. Conversione o trasformazione di tale indice in un punteggio standardizzato.
   Come sappiamo, l'indice di difficoltà più semplice da utilizzare è *p* che per poter essere utilizzato deve essere standardizzato. Nell'ambito dell'approccio *TID* la trasformazione di *p* in scala standardizzata è detta *delta*; la scala così ottenuta assume particolari valori; in particolare per l'item *i* e il gruppo *j*:
   $$delta_{ij} = 4z_{ij} + 13$$
   dove *z* rappresenta il valore *p* standardizzato.
   I valori *delta* così ottenuti compongono una scala ad intervalli, con punti tra loro equidistanti consentendo trasformazioni lineari, con media 13 e deviazione standard 4.
   Uno dei vantaggi dell'utilizzo dei valori *delta* sta nel fatto che consentono di evitare i valori negativi; infatti essi vanno da 0 a 26 che corrispondono, rispettivamente, ai valori *p* di .999 e .001; quindi valori alti di *delta* indicano item difficili mentre valori bassi indicano item facili[15].
   E' comunque possibile adottare altre trasformazioni in cui *z* rappresenta il *(1-p)-esimo* percentile della distribuzione normale standardizzata sarà sufficiente.
3. Creazione di un diagramma a punti in cui
   a. in ascissa vi sono i valori di difficoltà per il I gruppo,
   b. in ordinata vi sono i valori di difficoltà per il II gruppo,
   c. ciascun punto rappresenta un item.

   Il livello di dispersione dei punti nel grafico costruito è considerato una misura dell'interazione *gruppo*item*, una specie di coefficiente di correlazione inverso.
4. Interpolazione di una retta, detta asse maggiore dell'ellisse descritta dai punti (*major axis line*), interpretata come indice della relazione bivariata dei valori *delta* dei due gruppi.
   A partire da tale retta è possibile individuare la presenza di item *biased*. L'asse maggiore minimizza la distanza tra i diversi *delta* ed è individuato attraverso una procedura matematica che non coincide con quella dei minimi quadrati, in quanto, diversamente dall'analisi di regressione, qui non è possibile individuare una "variabile indipendente" e una "variabile dipendente" ovvero non esiste alcun criterio che consenta di decidere quale gruppo di punteggi deve essere regredito a partire dall'altro: per tale motivo la retta migliore è considerata quella che presenta la minima distanza perpendicolare dai punti[16]. A parte questa differenza l'approccio

[14] Questo approccio assume che gli item siano unidimensionali.

[15] Inoltre l'errore standard associato a *delta* rimane costante per tutti i livelli di difficoltà dell'item L'errore standard di $delta_{ij}$ è calcolato nel modo seguente:

$$ES_{delta_{ij}} = \frac{4}{N_j - 1}$$

L'errore standard di una singola porzione dovrebbe essere massimo .01; se risulta essere maggiore può sorgere il sospetto che la dispersione dei punti possa essere attribuita a fluttuazioni campionarie.

[16] La distanza perpendicolare di ciascun punto dall'asse maggiore è considerata una funzione dell'interazione di *gruppo*item* ed è data da:

appare molto simile a quello che consente di individuare la retta di regressione sulla base della seguente equazione:

$$y = a + bx$$

In questo caso le due costanti vengono calcolate nel modo seguente:

$$b = \frac{\left(\sigma_y^2 - \sigma_x^2\right) \pm \sqrt{\left(\sigma_y^2 - \sigma_x^2\right)^2 + 4r_{xy}^2\sigma_{xy}^2\sigma_y^2}}{2r_{xy}\sigma_x\sigma_y}$$

$$a = m_x - bm_y$$

dove

x e y simboli che si riferiscono ai due gruppi

σ_x e σ_y deviazioni standard rispettivamente del gruppo x e del gruppo y

r_{xy} coefficiente di correlazione tra i due gruppi di valori di *delta*.

5. Definizione di una funzione di distanza che consenta di verificare la distanza minima di ciascun item dalla retta.
   La presenza del *bias* è rilevata per quegli item che sono relativamente distanti dalla retta. Il problema a questo punto è quello di capire qual è il limite di tolleranza accettabile oltre il quale la distanza dall'asse maggiore rivela la presenza di un item *biased*. Uno degli approcci più comuni è quello che determina gli intervalli di confidenza dell'asse maggiore. Gli item che nel grafico appaiono al di fuori di tali intervalli possono essere giudicati "deviati". Spesso per stabilire i confini accettabili di grandezza del *bias* si usa il limite delle unità di *punti-z* a ±.75; in alcune occasioni tale limite è ritenuto troppo rigoroso e si preferisce fissare il limite delle unità di *punti-z* a ±1.5.

Nell'applicazione di questo approccio occorre prestare particolare attenzione al possibile insorgere di alcuni problemi.

- Abbiamo visto come la strategia basata sul *TID* di identificazione degli item *biased* è basata sull'assunto secondo il quale l'interazione *gruppi*item* rappresenta una valida indicazione per identificare gli item deviati; il livello di deviazione può essere considerato misura del *bias*.
  Nell'applicazione pratica di tale approccio occorre però tenere presente che in determinate circostanze tale assunto non può essere sostenuto; se per esempio gli item sono di varia difficoltà, l'interazione *gruppi*item* può esistere anche in un item perfettamente corretto (*unbiased*); quindi l'associazione tra item *bias* e interazione *gruppi*item* può risultare inappropriata.
- La proporzione di risposte corrette non rappresenta necessariamente una reale misura della difficoltà degli item, in quanto non ci si può aspettare che i componenti di due diversi gruppi reagiscano agli item con la stessa proporzione (i punti non necessariamente si posizionano esattamente sull'asse maggiore) anche nel caso di item *unbiased*. Se esistono due diversi livelli di forza discriminante per gli item (uno per ogni gruppo) il grafico dei valori di *delta* potrebbero ricadere su curve diverse, causando confronti ineguali. In questi casi è opportuno ricorrere all'approccio basato sull'*Item Response Theory*, di cui parleremo più avanti.
- La proporzione di casi che risponde correttamente ad un item così come è indicata dal valore di *delta* descrive nell'approccio *TID* non solamente l'item, in termini di difficoltà, ma anche il gruppo osservato, in termini di capacità; ciò rappresenta un grave limite all'identificazione di item affetti da errori sistematici soprattutto se basata sul valore *p*.

$$D_i = \frac{bX_i + a - Y_i}{\sqrt{\left(b^2 + 1\right)}}$$

dove

b pendenza dell'asse maggiore

a intercetta

X_i e Y_i punteggi *delta* per ciascun gruppo rispetto all'item i.

Per superare alcuni dei problemi visti in molte applicazioni pratiche la metodologia *TID* presenta alcune interessanti varianti:

a. Per evitare i difetti attribuiti teoricamente alle trasformazioni viste in precedenza, la trasformazione i valori *p* in *punti-z* può utilizzare le medie e le deviazioni standard per gruppo; in questo caso la funzione di distanza applicata è quella vista in precedenza, anche se la pendenza dell'asse maggiore viene convenzionalmente definita come unitaria (*b*=1); concettualmente tale modifica rimane molto vicino alla strategia tradizionale.

b. Si assume una retta fissa con un'inclinazione di 45°. Dopo aver calcolato i valori *delta* per ciascun gruppo nel solito modo, anziché confrontare i valori *delta* tra loro, si verifica la normalità della distribuzione $delta_{i1}$-$delta_{i2}$ considerando la media e la varianza delle differenze come parametri. Tali parametri sono considerati limiti di confidenza rispetto ai quali è possibile fare inferenze per identificare gli item *bias* e *unbiased*; tale procedura è essenzialmente un test di bontà di adattamento (del tipo Kolmogorov-Smirnov).

c. Per l'identificazione degli item *biased*, l'interazione *gruppi*item* viene definita in maniera diversa:
- *effetti non-ordinali*: conseguenza di item che hanno diversi ranghi di difficoltà tra i due gruppi; tali effetti, considerati un potente indicatore della presenza di item *biased*, possono essere osservati calcolando la cograduazione (in genere *rho* di Spearman) tra i valori dei due gruppi; il valore $1\text{-}rho^2$ sta ad indicare l'aspetto puramente non-ordinale del'interazione *gruppi*item*;
- *effetti ordinali*: conseguenza delle differenze relative nelle difficoltà degli item indipendentemente dai ranghi di entrambi i gruppi; per stimare gli effetti ordinali può essere utilizzato il coefficiente di correlazione *r*, consentendo in questo modo il confronto tra effetti ordinali e disordinali; per il calcolo di *r* si propone di non utilizzare i valori *delta* semplici ma quelli modificati calcolando la distanza tra valori di *delta* ordinati ($delta_1$-$delta_2$, $delta_2$-$delta_3$, ecc.); tali nuovi valori sono detti *decrementi di delta* e sono ottenuti all'interno di ciascun gruppo; *r* è quindi calcolato su tali valori.

Il valore $rho^2(1\text{-}r^2)$ è considerato una stima degli effetti puramente ordinali dell'interazione *gruppi*item*. La porzione residua (contributo congiunto alla varianza dell'interazione degli effetti ordinali e disordinali) dell'interazione *gruppi*item* è uguale a $rho^2\ r^2$.

d. Per superare uno dei principali limiti dell'approccio *TID*, ovvero la sua incapacità a definire la difficoltà degli item indipendentemente dai gruppi osservati, è possibile calcolare la correlazione parziale tra successo rispetto ad un determinato item e l'appartenenza di gruppo (o punteggi individuali attesi). In pratica questo approccio, controllando la capacità, supera uno dei suoi principali limiti mantenendone la sua semplicità.

2.3 FATTORI CHE INCIDONO SULLA VERIFICA DEL MODELLO

In realtà gli errori di misurazione non sono dovuti tutti al campionamento degli item. Sono molti i fattori che influenzano la verifica del modello e, conseguentemente, del livello di affidabilità; la loro quantità e la loro tipologia dipendono dalla natura della caratteristica da misurare e dalla utilizzazione e applicazione che si fa dello strumento. Tra i fattori che possono incidere sul livello di affidabilità occorre considerare anche l'approccio sperimentale adottato; infatti ciascuno degli approcci è sensibile a fonti diverse di variazione nei punteggi individuali, ognuna delle quali produce un errore di misurazione in misura non sempre valutabile. Vediamo quali sono le fonti di variazione per ciascun metodo.

L'errore di stima nel caso dell'approccio delle *componenti parallele* in questo caso è attribuibile sia all'interazione *soggetto-contenuto dello strumento* che all'interazione *soggetto-condizioni di*

somministrazione. In quest'ultimo caso si fa riferimento anche fattori che possono incidere in maniera diversa sugli item, quali la progressiva stanchezza del soggetto durante la somministrazione. Come abbiamo già visto tale problema può in parte essere gestito adottando, per la suddivisione dello strumento, il metodo *odd-even* che consente di attribuire l'influenza di tale fattore in misura uguale alle due parti.
Accanto ai problemi attribuibili all'approccio sperimentale è possibile identificare altri fattori disturbanti tra i quali ricordiamo:

a. limiti di tempo, considerato come *tempo impiegato per sottoporre lo strumento* e inteso anche come accuratezza di esecuzione: maggiore è il tempo, maggiore è l'affidabilità. Come si può intuire tale fattore è comunque molto legato al numero di item impiegati. In ogni caso, superato un certo limite di tempo ottimale, l'affidabilità tende comunque a diminuire.
b. Caratteristiche degli item ovvero
   - *omogeneità degli item*: maggiore è l'*omogeneità*, maggiore è l'affidabilità;
   - *dipendenza tra item*: item troppo interdipendenti tendono a ridurre l'affidabilità in quanto rispetto a tali item i soggetti tendono a comportarsi nello stesso modo; tale situazione presenta lo stesso effetto della diminuzione del numero di item;
   - *capacità discriminante degli item*: minore è l'estensione del continuum misurato da ciascun item, minore è l'affidabilità; questo fattore è collegato con il successivo.
c. Qualità dello *scoring* inteso come
   - *oggettività e accuratezza di attribuzione dei punteggi agli item*: i giudizi soggettivi introducono maggiore variabilità; da ciò si può dedurre che maggiore è l'oggettività, maggiore è l'affidabilità; la mancanza di accuratezza è comunque difficilmente controllabile;
   - *livello di casualità delle risposte agli item*: si tratta della classica posizione di chi *tira a indovinare* ovvero di chi nel rispondere agli item sceglie la risposta a caso. Ciò causa alcune variazioni nel risultato da item a item ed abbassa l'affidabilità di tutto lo strumento; maggiore è la probabilità di rispondere bene (o comunque nella direzione della dimensione misurata) anche casualmente, minore è l'affidabilità; ricordiamo che la casualità è maggiore negli item dicotomici;
   - *response set*, ovvero tendenza di un soggetto a rispondere nello stesso modo a tutti gli item.
d. Eterogeneità/omogeneità del campione della sperimentazione: nella stima dell'affidabilità è importante considerare le caratteristiche sia della popolazione da cui è stato estratto il campione sul quale viene effettuato l'esperimento di affidabilità che della popolazione sulla quale lo strumento verrà utilizzato in pratica, tenendo presente che maggiore è l'omogeneità dei soggetti che compongono il campione, maggiore risulta essere l'affidabilità dello strumento e minore sarà la possibilità di estendere l'utilizzo dello strumento.
e. Livello di comprensione e interpretazione degli item: la riduzione dell'affidabilità può essere dovuta anche a cattive interpretazioni dell'item soprattutto quando:
   - le domande presentano particolari valenze emozionali;
   - le istruzioni sono inadeguate o difettose;
   - le domande presentano tranelli, trucchetti o inganni linguistici.

   Ecco perché occorre prestare molta attenzione nella scelta delle parole e delle espressioni; per esempio item espressi in una forma verbale sintetica e con costruzioni sintattiche semplici risultano di solito i migliori.
f. Altri fattori, così sintetizzati:
   - velocità di somministrazione;
   - velocità individuale nel rispondere;
   - distribuzione del tempo per rispondere agli item;
   - predisposizione soggettiva alla precisione;
   - reazione ad elementi e stimoli disturbanti, per esempio alla stanchezza;

- persistenza di determinati atteggiamenti mentali o emozionali (positivi o negativi) durante tutto il tempo di somministrazione;
- malattie, preoccupazioni, eccitamenti, ecc.
- livello di apprendimento del soggetto.

Per esempio un soggetto potrebbe avvertire improvvisamente mal di testa a metà della somministrazione, potrebbe non accorgersi di non aver risposto a determinati item, potrebbe accorgersi a metà di aver compreso male le istruzioni per rispondere, ecc.

2.3.1 Numero ottimale di item

E' difficile conoscere in anticipo quanti item devono essere utilizzati per la costruzione di uno strumento. Per determinare il numero di item da impiegare esistono alcune regole empiriche come quella che ritiene 30 item dicotomici in grado di raggiungere un buon livello di consistenza interna. Non sempre però è possibile giungere a definire 30 item, a causa di problemi di applicabilità dello strumento ma anche per problemi di definizione dell'area stessa. In genere se gli item presentano una scala di risposta a più livelli per ottenere lo stesso livello di affidabilità è sufficiente un numero inferiore di item (a volte 10 item con una scala di risposta a sette punti possono raggiungere un'affidabilità di .80).
In fase di messa a punto dello strumento, se si conosce molto poco sull'omogeneità degli item, è possibile seguire una delle seguenti strategie:

- cominciare con un numero di item più grande per eventualmente eliminarli in fase di verifica del modello di scaling (*item analysis*),
- utilizzare un numero di item minore di quello ritenuto adeguato e di aggiungere via via nuovi item fino a quando si osserva un aumento dell'affidabilità della scala.

A tale proposito occorre tenere presente che l'aggiunta o l'eliminazione di un item comporta non solo la verifica dello strumento di misurazione ma anche della sua struttura concettuale; può essere interessante confrontare gli item scartati con quelli conservati: se gli item eliminati sono differenti nel contenuto è necessario riconsiderare la definizione della caratteristica misurata dai rimanenti item.
Contemporaneamente è necessario valutare anche la dimensione del campione di soggetti su cui verificare il modello; una regola empirica è quella di disporre di un numero di osservazioni uguale a cinque/dieci volte il numero di item.

2.3.2 Ulteriori verifiche, valutazioni e controlli

L'osservazione di un valore d'affidabilità soddisfacente non garantisce la reale verifica del modello di *scaling* e in primo luogo dell'unidimensionalità della scala; se per esempio si combinano due gruppi di item che misurano due diversi costrutti, tra loro correlati e con alti valori di affidabilità, è possibile ottenere ugualmente un buon livello di consistenza interna pur essendo il modello diventato, nel frattempo, bidimensionale.
Tutti i passaggi dell'analisi devono consentire di rivalutare tutti i momenti che hanno condotto ai risultati, dalla formulazione degli item al campione utilizzato per la validazione, alla definizione del costrutto.
Una volta terminato il procedimento di validazione della scala è possibile passare ad un ulteriore stadio di sviluppo della scala. E' possibile ripetere l'analisi su dati rilevati su un nuovo campione, per verificare nuovamente il livello di affidabilità e di validità della scala. La disponibilità di più stime di affidabilità, ottenute su diversi campioni, consente di generalizzare l'affidabilità della scala a settori più ampi di soggetti. Un'altra buona abitudine è quella di calcolare sempre il coefficiente di

affidabilità ad ogni applicazione, anche nel caso di scale già validate.

2.4 LIMITI DEL MODELLO ADDITIVO

Il modello additivo ha trovato e trova larga applicazione nella ricerca sociale; ciò è dovuto sia alla sua base logica chiara che alle sue semplici procedure di applicazione e di verifica. Esso però è stato oggetto di critiche per diversi motivi.
Esso dipende interamente dall'assunto di fluttuazioni casuali tra gli item che vengono sommati per creare la scala. I criteri stabiliti per selezionare gli item (*item analysis*) utilizzati nelle tecniche *Likert* e *Thurstone* possono essere visti come strategie per cercare di assicurare la casualità delle differenze interitem. Nel caso in cui sia possibile soddisfare tale assunto l'approccio additivo risulta essere una tecnica di *scaling* molto potente. D'altra parte il modello additivo presenta due limiti prodotti dallo stesso assunto:

a. il metodo assume che in un insieme di punti tutti gli errori sono attribuibili a fluttuazioni casuali. Tale fluttuazioni potrebbero però verificarsi per altre ragioni come l'influenza simultanea di più dimensioni sottostanti. Il modello additivo però non considera a priori la multidimensionalità come presenza possibile; infatti una delle critiche rivolte a questo modello riguarda proprio il metodo di valutazione della reale unidimensionalità del gruppo di item in quanto basato su assunti troppo deboli[17]; in termini strettamente pratici, ciò significa che la verifica del modello di scala potrebbe produrre un buon adattamento anche quando le vere fonti di variazioni sono più di una. Ciò significa che tale approccio è *molto utile come tecnica di scaling* ma *molto carente come criterio di scaling*.
b. con l'approccio additivo è possibile scalare solo un'unico insieme di punti tra i due che di solito costituiscono i dati del tipo *stimolo-unico*. Ciò è dovuta al fatto che, assumendo che il secondo insieme di punti varia in modo casuale, le stime precise per le posizioni dei punti appartenenti al secondo insieme non avrebbero significato. La scelta riguardante quale insieme di punti scalare dipende interamente dagli assunti fatti sui dati e dagli obiettivi analitici. Per questo è possibile scalare:
   - casi, quando gli indicatori vengono considerati ripetizioni casuali (approccio *Likert*),
   - item, quando i casi vengono considerati ripetizioni casuali (approccio degli *intervalli che appaiono uguali* di Thurstone).

Un altro appunto rivolto a tale metodo riguarda l'approccio che il modello additivo utilizza per la valutazione dell'adeguatezza (in termini di affidabilità e validità) dello strumento in quanto troppo legata alle informazioni ricavate dai dati empirici.
Un altra caratteristica criticata riguarda la definizione delle scale di risposta troppo rigidamente vincolata per tutti gli item (si richiede infatti lo stesso numero di livelli di risposta per tutti gli item).

Ma la critica più severa rivolta all'approccio prende spunto dal fatto che secondo il modello additivo, richiamandosi al modello classico di misurazione, la dimensione da misurare è espressa come punteggio vero definito come *il valore atteso di una caratteristica presente in un certo soggetto*. Tale dimensione misurata in un soggetto è definita solamente in funzione di un particolare strumento di misurazione. Se per esempio la dimensione misurata è rappresentata da una capacità, lo strumento sarà considerato:

- *difficile* quando i soggetti risultano avere bassa capacità;

[17] Vedremo infatti come una verifica indiretta dell'unidimensionalità è l'analisi della consistenza interna, basata sulle correlazioni tra ciascun item e il punteggio totale (correlazione *item-totale*); tale criterio è considerato insufficiente come prova di unidimensionalità in quanto non è in grado di registrare l'eventuale presenza di due o più sottoinsiemi di item corrispondenti a sottodimensioni della caratteristica misurata.

- *facile* quando molti soggetti risulteranno avere alta capacità.

Analogamente definendo la *difficoltà* come la *proporzione di soggetti che risponde correttamente ad un item*, ne consegue che:

- la *facilità/difficoltà* di un item dipende dalla capacità dei soggetti misurati,
- la *capacità* di un soggetto dipende dalla facilità o difficoltà degli item.

Ciò vuol dire che la verifica dell'affidabilità (e quella della validità) dello strumento è troppo legata al campione utilizzato per la sperimentazione con la conseguenza che

- le caratteristiche di uno strumento cambiano in funzione del gruppo di soggetti (*group-dependent*),
- le caratteristiche di un soggetto cambiano in funzione degli strumenti di misurazione (*test-dependent*).

Conseguentemente è molto difficile confrontare

- soggetti misurati con strumenti diversi,
- item per la cui validazione sono stati utilizzati campioni per la sperimentazione composti da soggetti con caratteristiche diverse.

Proviamo a questo punto a riassumere i limiti riscontrati che hanno condotto ad una valutazione insoddisfacente dell'approccio additivo:

- *Group-dependent*: la dipendenza delle caratteristiche degli item dai campioni di soggetti utilizzati per la sperimentazione rende gli indici di uso limitato nella pratica, soprattutto nei casi in cui le caratteristiche di tali campioni sono diverse rispetto a quelle della popolazione sulla quale verrà utilizzato lo strumento costruito e validato.
- *Test-dependent*: è difficile confrontare direttamente tra punteggi individuali ottenuti con strumenti diversi non essendo sempre possibile osservare e identificare relazioni funzionali tra due strumenti[18]. Se i soggetti presentano diversi livelli, per esempio, di capacità (lo strumento è più difficile per un gruppo che per un altro), i punteggi osservati contengono una diversa quantità di errore, ovvero sono affidabili in diversa misura[19]. Per ottenere informazioni più precise sulla reale capacità di un soggetto sarebbe quindi necessario osservare il risultato ottenuto in ciascun item. La difficoltà di confrontare punteggi individuali ottenuti con strumenti diversi è dovuta anche alla diversa precisione nel misurare la dimensione. Per superare il problema della presenza di livelli di versi di errori di misurazione sarebbe necessario stabilire diversi livelli rispetto alla dimensione, per esempio livelli diversi di capacità e di difficoltà.
- *Definizione degli strumenti paralleli*: l'affidabilità è valutata sulla base di correlazioni tra punteggi ottenuti con forme parallele che sono difficili, se non impossibili, da definire;
- *Stime di affidabilità con significatività ignota*: la valutazione dell'affidabilità è effettuate con coefficienti che forniscono stime che non possono essere sottoposte a verifica della significatività statistica;
- *Errore di misurazione uguale*: il modello classico assume che l'errore di misurazione sia funzione dell'affidabilità e della varianza della distribuzione dei punteggi e che, conseguentemente, sia uguale per tutti i soggetti; tale assunto però non è accettabile in quanto, come abbiamo visto, ciascuno strumento misura in modo impreciso e in modo disuguale soggetti di capacità differenti.

Un ultimo limite è dato dal fatto che l'approccio additivo è *test-oriented* piuttosto che *item-oriented*; infatti l'adozione del concetto di punteggio totale non consente alcuna valutazione delle risposte individuali ai singoli item; ciò vuol dire che a partire dal punteggio totale individuale non è possibile stimare le risposte ai singoli item.

[18] Tale limite è uguale a quello riscontrato per il modello sperimentale degli *strumenti paralleli*.

[19] Il punteggio zero ottenuto da un soggetto indica che il soggetto presenta un basso livello nella capacità misurata ma non fornisce alcuna informazione su quanto esattamente basso soprattutto se confrontato con lo stesso punteggio ottenuto da un altro soggetto.

3. I MODELLI CUMULATIVI. L'APPROCCIO DETERMINISTICO

Al contrario del modello additivo, i modelli cumulativi posizionano entrambi gli insiemi di punti. Infatti secondo tali modelli l'*assegnazione dei punteggi riguarda sia gli item che compongono lo strumento che i casi cui è sottoposto*; in altre parole la misurazione deve riguardare contemporaneamente sia i casi che gli stimoli. Su questa base è possibile distinguere due distinti modelli di misurazione che si differenziano rispetto al diverso trattamento del problema dell'errore di misurazione, inteso come variazione non sistematica nelle risposte, o della varianza non sistematica (varianza dell'errore) e conseguentemente alla diversa definizione dei modelli di risposta (*trace line*); tali modelli sono:

- modello deterministico, secondo il quale non è possibile prevedere la definizione esplicita di errore o di variazione non sistematica nella misurazione; tutta la variazione nelle risposte è interamente attribuita alla posizione dei soggetti e alla posizione dell'item lungo il continuum che rappresenta la dimensione studiata; all'interno di questo modello non si fa alcuna previsione di varianza dell'errore; conseguentemente la probabilità di dare una determinata risposta ad un item può essere solo 0 (*beta*) o 1 (*alfa*) in qualsiasi punto del continuum sottostante l'attributo misurato;
- modello probabilistico, che prevede la definizione di errore casuale; secondo questo modello la probabilità di dare una determinata risposta ad un item può variare da 0 a 1 e non è ristretta a tali posizioni estreme.

Di solito tali modelli sono applicati ad una matrice di dati rettangolare con *n* risposte soggettive e *k* item (di solito dicotomici). Si assume che sia gli item che i soggetti occupano posizioni lungo la dimensione sottostante. Tale assunto, combinato con la relazione di dominanza dei dati *stimolo-unico*, conduce ad un ordinamento dei punti-soggetto e dei punti-item basato sull'accumulazione delle risposte. La versione più comune dei modelli cumulativi-deterministici è quella conosciuta con il nome *Guttman*, che ha trovato applicazione nello *scaling* di atteggiamenti.

3.1 L'IPOTESI DI SCALOGRAMMA

Il modello deterministico ha l'obbiettivo di classificare sia i soggetti che gli item lungo un continuum. Esso è detto *deterministico* in quanto nella sua formulazione è prevista la definizione di un modello ideale di riferimento del quale deve essere verificato l'adattamento ai dati osservati. In realtà tale approccio non consiste tanto nel determinare se il modello si adatta o meno ai dati quanto piuttosto nell'utilizzare il modello come approssimazione adeguata ai dati; in altre parole, essendoci sempre, tranne casi eccezionali, uno scarto tra previsione e realtà, il problema non è quello di controllare se c'è differenza, ma se questa rientra in certi margini di tolleranza; quindi nei casi di mancata corrispondenza perfetta tra modello e dati, il modello può servire come approssimazione ai dati reali. Ciò fa sorgere due importanti questioni:

- come calcolare lo scarto tra dati ideali e dati osservati,
- come definire i margini di tolleranza.

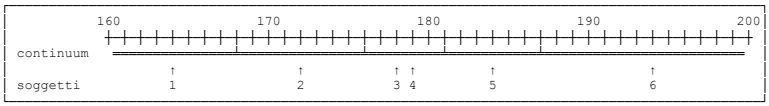
A tale proposito sono state sviluppate diverse tecniche che servono a indicare la bontà dell'approssimazione. Dato che i modelli ideali possono servire per rappresentare i dati empirici all'interno di un certo errore, espresso in termini di differenza tra il risultato atteso e il risultato

ottenuto, la definizione di questo consente di valutare, attraverso un indice, il grado di approssimazione all'ideale.
Secondo il modello deterministico per poter valutare se l'insieme di item consente di misurare sia gli item che i soggetti è necessario assumere che l'attributo rispetto al quale tutti gli item e i soggetti vengono misurati sia *scalabile*. Individuato un tale attributo (status socio-economico, pregiudizio, ecc.) è necessario fare due assunti:
- i soggetti possono essere classificati rispetto a tale attributo su un singolo continuum;
- esiste una relazione tra gli item e il continuum postulato.

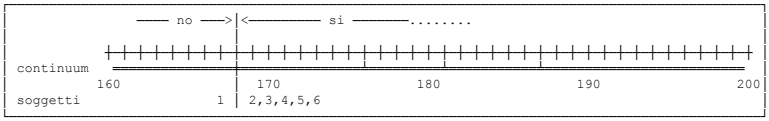
A questo punto è possibile definire e selezionare gli item che insieme devono rispondere a determinate caratteristiche. A tale proposito è possibile distinguere due diversi approcci: *Guttman* e alternativo, cui corrispondono due tipi di item.
In termini geometrici ciascun item rappresenta, come per il modello additivo, una ripetizione della dimensione sottostante anche se, come vedremo, in punti diversi del continuum.

3.1.1 Caratteristiche degli item

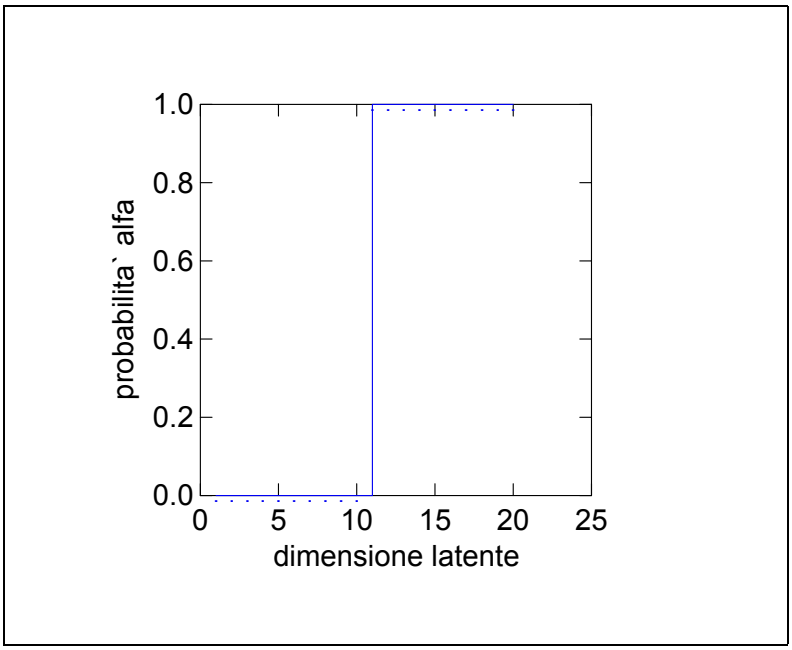
Poniamo di voler collocare lungo il continuum, relativo all'"altezza in centimetri", cinque soggetti con le seguenti altezze:
1:164 cm. 2:172 cm. 3:178 cm 4:179 cm. 5:184 cm. 6:194 cm.
Lungo il continuum reale i soggetti sono così posizionati:



Per poter posizionare i soggetti lungo il continuum è possibile definire come indicatore un item che può assumere la seguente caratteristica:
altezza maggiore di 168 cm. ☐ *si* ☐ *no*
Vediamo se e in che modo tale item riesce a collocare i soggetti lungo il continuum.



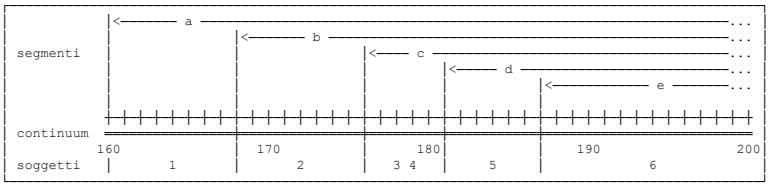
Le risposte a tale item hanno consentito di ordinare i soggetti lungo il continuum rispetto al punto di discriminazione individuato dall'item. Ciò è reso possibile dal fatto che ciascuna risposta è interpretabile come indicazione della direzione della relazione tra soggetto e item: infatti una risposta positiva indica che il soggetto è al di sopra del limite indicato dall'item mentre una risposta negativa indica che il soggetto si posiziona sul limite o al di sotto. Gli item di questo tipo sono detti *monotoni* e sono definiti da una *trace line* in cui fino ad un certo punto del continuum dell'attributo la probabilità di risposta *alfa* è 0 (mentre *beta* è 1); oltre tale punto la probabilità di risposta *alfa* è 1; tale tipo di *trace line* è basata su una funzione detta *step function* e ha la seguente forma:



L'utilizzazione di un solo item di questo tipo non consente però l'ordinamento preciso dei casi lungo il continuum relativo all'attributo scalabile; per fare ciò sono necessari più item (*scaling*). Tali item devono essere in grado di discriminare in punti diversi e ordinati del continuum ovvero devono essere non solo monotoni ma anche *cumulativi*. Il modello deterministico-cumulativo, che risponde alla definizione *Guttman*, richiede item (dicotomici nel caso più semplice) di questo tipo.
Tornando all'esempio, definiamo cinque item:

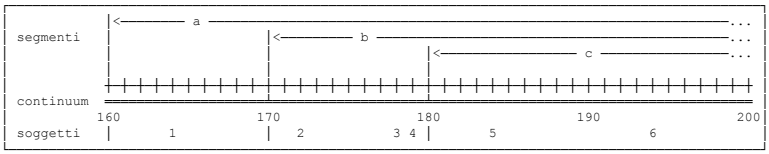
a. altezza maggiore di 160 cm. ☐ *si* ☐ *no*
b. altezza maggiore di 168 cm. ☐ *si* ☐ *no*
c. altezza maggiore di 176 cm. ☐ *si* ☐ *no*
d. altezza maggiore di 181 cm. ☐ *si* ☐ *no*
e. altezza maggiore di 187 cm. ☐ *si* ☐ *no*

Di seguito vediamo, relativamente ai cinque item, quali sono i punti di discriminazione lungo il continuum e a quali segmenti di tale continuum corrispondono le risposte positive; ciascun soggetto viene classificato nell'item, corrispondente all'ultima risposta positiva riferita in base alla propria altezza (classificazione ordinale).

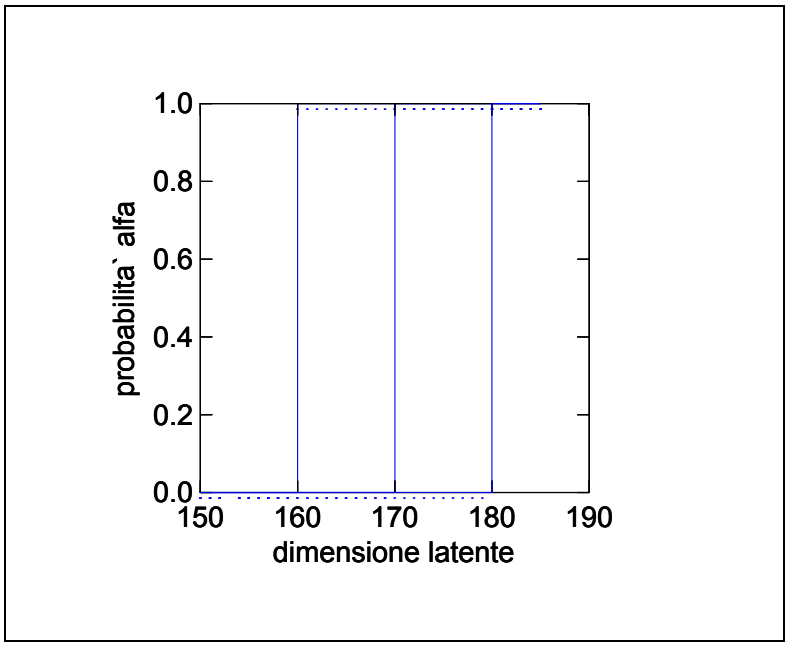


Proviamo a definire altri item che suddividano il continuum in modo diverso:

a. altezza maggiore di 160 cm. ☐ *si* ☐ *no*
b. altezza maggiore di 170 cm. ☐ *si* ☐ *no*
c. altezza maggiore di 180 cm. ☐ *si* ☐ *no*



Appare evidente come con tre item la discriminazione tra i soggetti sia meno precisa di quella ottenuta con i precedenti cinque item. Di seguito sono raffigurate le tre *trace line* corrispondenti ai tre precedenti item che soddisfano i requisiti del modello deterministico monotono:



Concettualmente gli item che soddisfano il modello di *scaling Guttman* possono essere visti in successione su una sola dimensione. In questa ottica ogni item rappresenta, alla luce dell'ipotesi, un punto intermedio fra l'item precedente e il successivo ovvero deve discriminare perfettamente in un particolare punto del continuum relativo all'attributo, diverso dagli altri. Il modello può quindi essere applicato quando è necessario misurare caratteristiche e disposizioni per le quali è possibile identificare livelli crescenti di intensità come capacità, impegno, difficoltà, ecc.; quindi il gruppo di item che soddisfà il modello cumulativo *Guttman* deve essere in grado di misurare e ricoprire la variabilità di caratteristiche, disposizioni, difficoltà, capacità crescenti; questo vuol dire che essi devono essere ordinabili ovvero devono rispondere al concetto di *scalabilità*. Nel misurare capacità, per esempio, in pratica dovremmo ottenere che gli item considerati facili devono poter essere superati da tutti mentre gli item considerati difficili devono essere superati da alcuni (ricordiamo che maggiore è la proporzione degli individui che superano un item, più facile e meno discriminante esso si presenta).

Quando

- gli item misurano livelli crescenti di intensità (gli item discriminano in punti diversi e ordinati sul continuum),
- i soggetti presentano livelli crescenti di intensità (i soggetti che rispondono positivamente ad un item rispondono positivamente anche ai precedenti e i soggetti che rispondono negativamente ad un item rispondono negativamente anche ai successivi),

è stato identificato il modello ideale che assume la forma dello *scalogramma* e che riflette la caratteristica di *perfetta scalabilità*. Lo scalogramma perfetto presenta una matrice con due triangoli, uno formato da tutti segni "+" (risposte positive) e uno formato da tutti segni "-" (risposte negative); con i dati del nostro esempio (tre item e sei soggetti) è stato raggiunto lo scalogramma perfetto:

Soggetti	Item		
	a	b	c
1	+	-	-
2	+	+	-
3	+	+	-
4	+	+	-
5	+	+	+
6	+	+	+
"+": risposta positiva			
"-": risposta negativa			

Quando viene identificato il modello ideale triangolare è possibile misurare ovvero ordinare lungo il continuum

- ciascun item, sulla base della proporzione di risposte positive registrata,
- ciascun soggetto, sulla base del numero di risposte positive rilevate (punteggio totale individuale).

L'identificazione del modello triangolare, secondo i criteri visti, consente di stabilire con esattezza a quali item ciascun soggetto ha risposto positivamente e a quali negativamente; la possibilità di prevedere per ciascun soggetto le risposte date a tutti gli item sulla base del punteggio totale rappresenta la prova che un item appartiene ad una singola dimensione sottostante.

3.1.2 Il posizionamento

Una scala deterministica, ma in generale le scale cumulative, semplifica l'informazione contenuta nei dati riducendo la singola matrice *n*k* a due vettori separati:

- *n**1 per le posizioni dei soggetti,
- *k**1 per le posizioni degli item.

		Matrice di input *V: two-way, two-mode* v_{ij} : presenta "1" se *i* domina *j*; "0" se *j* domina *i*					
		Oggetti-colonna (item)					
		1	2	…	*j*	…	*k*
Oggetti-riga (soggetti)	1	v_{11}	v_{12}	…	v_{1j}	…	v_{1k}
	2	v_{21}	v_{22}	…	v_{2j}	…	v_{2k}
	…	…	…	…	…	…	…
	i	v_{i1}	v_{i2}	…	v_{ij}	…	v_{ik}
	…	…	…	…	…	…	…
	n	v_{n1}	v_{n2}	…	v_{nj}	…	v_{nk}

Matrici di output					
X: punteggi dei soggetti			Y: punteggi degli item		
x_i : numero di punti-item posizionati a sinistra del soggetto i lungo la dimensione			y_i : numero di punti-soggetto posizionati a sinistra del punto-item lungo la dimensione		
soggetti	1	x_1	item	1	y_1
	2	x_2		2	y_2
	…	…		…	…
	i	x_i		*i*	y_i
	…	…		…	…
	n	x_n		*k*	y_k

Il modello di *scaling* riguarda

- i punti degli *n* soggetti ordinati lungo la dimensione secondo il numero di stimoli che dominano,
- i *k* punti-item ordinati lungo la dimensione secondo il numero di soggetti che dominano ciascuno di essi.

In termini geometrici il valore di ciascuna cella nella matrice dei dati fornisce un'informazione sulla coppia di punti soggetto-item. Tale valore è:

- "1" se il punto-soggetto è posizionato a destra del (domina il) punto-item lungo la dimensione sottostante,
- "0" se il punto-item è posizionato a destra del (domina il) punto-soggetto.

Ciascun punto è posizionato lungo la dimensione sommando il numero di "1" contenuti all'interno

della propria riga o colonna della matrice dei dati.
La posizione del punto-soggetto sarà tanto più a destra quanto più grande è il numero di risposte positive date dal soggetto, in quanto domina tanti item. La posizione del punto-item sarà tanto più a sinistra quanto più grande è il numero di risposte positive ottenute dall'item, in quanto è dominato da tanti più soggetti.
Una *scala Guttman* composta da dati dicotomici rappresenta l'operazionalizzazione più semplice del modello di *scaling* cumulativo-deterministico. Il modello può comunque essere generalizzato anche ad item non dicotomici. L'interpretazione dei punti-item risulta in questo caso però un pò diversa. Per quanto si è detto il punto-item può essere interpretato come un *cut-point* lungo la dimensione in quanto serve come confine tra risposte positive e risposte negative per quell'item. In altre parole se tutti i soggetti ad un determinato item danno risposte

- negative, i loro punti saranno posizionati a sinistra del *cut-point*,
- positive, i loro punti saranno posizionati a destra del *cut-point*.

Nel caso di item dicotomici il punto-item e il corrispondente *cut-point* coincidono. Nel caso di item non dicotomici si ottengono più *cut-point*; il numero di *cut-point* è uguale a q-1, dove q corrisponde al numero di categorie. Gli item non dicotomici suddividono la dimensione in q segmenti. Ciò in realtà non complica lo *scalogramma* che viene costruito nello stesso modo. La differenza sta nel fatto che la matrice dei dati rappresentano in modo esplicito i *cut-point* invece che gli item; questo vuol dire che la matrice degli item dovrà presentare più colonne.

3.2 LA VERIFICA DEL MODELLO: LO SCALOGRAM ANALYSIS

Tale modello è basato sui seguenti assunti:

- *omogeneità*: tutte le grandezze rilevate esprimono una quantità della stessa natura in modo da legittimare la riunione di più item in un unico punteggio[1]; in altre parole il criterio è soddisfatto quando tutte le grandezze rilevate (risposte agli item) sono quantità della stessa natura e possono essere riunite in un punteggio generale che rappresenta la misura di un solo fattore;
- *esuastività*: l'insieme degli item deve rappresentare un inventario completo del dominio reale di una "dimensione" ovvero gli item devono ricoprire tutta la variabilità osservabile in modo da consentire una valutazione globale;
- *unidimensionalità*: l'insieme degli item è determinato da un insieme di attitudini strettamente connesse e/o dipendenti da una sola dimensione;
- *gradualità/scalabilità*: gli item devono essere scelti in modo tale che risultino essere superabili con livelli diversi della stessa attitudine; in altre parole deve essere possibile ordinare gli item secondo un livello crescente di intensità (capacità, disposizioni, ecc.); gli item così selezionati presentano una parziale sovrapposizione di significato; ciò consente di ottenere una *gradualità* della valutazione.

Se tali assunti vengono soddisfatti ne consegue la giustificazione teorica di un punteggio globale.
Il modello teorico si realizza perfettamente quando, per superare un item, sono indispensabili tutte le attitudini utilizzate per l'item precedente più un'attitudine aggiuntiva. Per poter verificare l'adattamento del modello ai dati, ovvero per poter determinare se un campione di item e un campione di soggetti sono conformi allo specifico insieme di criteri considerati requisiti dello *scaling* Guttman, si applica un procedimento detto *scalogram analysis*. In particolare l'analisi dello

[1] Nel caso di alta consistenza interna tra item, si può non solo affermare che la scala è omogenea, ma anche che le risposte degli intervistati sono coerenti. Questa deduzione è la diretta conseguenza del fatto che gli intervistati, nella grande maggioranza dei casi, hanno compreso le domande, ossia che la scala, così come è stata applicata, è idonea per lo studio.

scalogramma consente di verificare se e in che misura la distribuzione reale delle risposte si discosta dalla distribuzione ideale o dalla combinazione ideale (scalogramma) ovvero se le risposte riproducono il modello triangolare teorizzato; quindi l'analisi dello scalogramma consente il confronto di una distribuzione concreta di soggetti rispetto a un modello teorico di perfetta scalabilità. Si registra una scala perfetta quando tutti i modelli individuali seguono questo andamento, ossia nessun individuo supererà un item se non ha superato un item più facile per l'intero gruppo.
Se gli item e i soggetti coprono la variabilità della caratteristica ordinale misurata, la distribuzione delle frequenze registrate dovrebbe presentare un numero decrescente di risposte dello stesso tipo. Se per esempio si dispone di 5 item dicotomici con difficoltà crescente e 10 soggetti con diverse attitudini si dovrebbero ottenere le seguenti distribuzioni di frequenza:

item	Numero di "si"	Numero di "no"
1	10	0
2	7	3
3	5	5
4	3	7
5	0	10

Per illustrare la procedura vediamo un semplice esempio in cui vi è un perfetto adattamento al modello deterministico; naturalmente la procedura può essere estesa ai casi più complessi. Poniamo di avere ottenuto, con 5 item dicotomici a 6 soggetti, i seguenti risultati:

Soggetti	Item				
	1	2	3	4	5
a	-	+	+	+	+
b	-	-	+	-	+
c	-	-	-	-	+
d	+	+	+	+	+
e	-	-	-	-	-
f	-	+	+	-	+
P	0.17	0.50	0.67	0.33	0.83

dove
- "+" indica *superamento dell'item/della prova* o *risposta affermativa*,
- "-" indica *non superamento dell'item/fallimento della prova* o *risposta negativa*.

A questo punto assegniamo alla risposta
- "+" (superamento dell'item/della prova o risposta affermativa) punteggio *1*,
- "-" (non superamento dell'item/fallimento della prova, risposta negativa) punteggio *0*.

Alla matrice così modificata aggiungiamo:
- il punteggio totale di ogni soggetto (PT),
- la proporzione di risposte positive per ogni item (P).

Soggetti	Item					Punteggio (PT)
	1	2	3	4	5	
a	0	1	1	1	1	4
b	0	0	1	0	1	2
c	0	0	0	0	1	1
d	1	1	1	1	1	5
e	0	0	0	0	0	0
f	0	1	1	0	1	3
P	0.17	0.50	0.67	0.33	0.83	

Avendo selezionato gli item sulla base dell'assunto di *scalabilità* si ricostruisce e si verifica tale ordine in sede d'analisi in funzione della proporzione di soggetti che hanno superato la prova. A tal fine si modifica

A. la posizione degli item (colonne della matrice) in modo tale che risultino ordinati secondo la proporzione di risposte positive (dalla più alta alla più bassa);
B. la posizione dei casi (righe della matrice) in modo tale che i soggetti siano ordinati secondo

il punteggio totale (dal più alto al più basso).

A								B						
SOGGETTI	ITEM					PT		SOGGETTI	ITEM					PT
	5	3	2	4	1				5	3	2	4	1	
a	1	1	1	1	0	4		d	1	1	1	1	1	5
b	1	1	0	0	0	2		a	1	1	1	1	0	4
c	1	0	0	0	0	1		f	1	1	1	0	0	3
d	1	1	1	1	1	5		b	1	1	0	0	0	2
e	0	0	0	0	0	0		c	1	0	0	0	0	1
f	1	1	1	0	0	3		e	0	0	0	0	0	0
P	0.83	0.67	0.50	0.33	0.17			P	0.83	0.67	0.50	0.33	0.17	

Riattribuendo alle risposte *1* il segno + e alle risposte *0* il segno - potremo osservare come tali dati riproducano in modo preciso il modello triangolare (*perfetta scalabilità*): in ogni profilo, dopo un insuccesso in una prova, si incontrano solo insuccessi nelle prove più difficili; un punteggio di "4" indica che il soggetto ha superato i primi quattro item ma non l'ultimo:

B						
SOGGETTI	ITEM					PT
	5	3	2	4	1	
d	+	+	+	+	+	5
a	+	+	+	+	-	4
f	+	+	+	-	-	3
b	+	+	-	-	-	2
c	+	-	-	-	-	1
e	-	-	-	-	-	0

In questa breve analisi il modello ideale di riferimento è stato perfettamente osservato; nella pratica è però molto difficile poter osservare una perfetta riproduzione del modello triangolare; per questo motivo, per stabilire se il modello costituisce un'adeguata rappresentazione dei dati empirici, occorre definire il livello di deviazione tollerabile (valutazione della bontà di adattamento). Per fare ciò non basta la semplice osservazione delle distribuzioni di frequenza, occorre effettuare una validazione più approfondita, basata sui concetti di *riproducibilità*, *scalabilità* e *predicibilità*.

- *Riproducibilità*: *possibilità di riprodurre per ciascun soggetto, a partire dal punteggio totale, le risposte date a ciascun item*. Per tutti i soggetti e tutti gli item, è possibile calcolare la percentuale di riproducibilità.
- *Scalabilità*: *osservazione di item a difficoltà crescente*; l'osservazione della scalabilità consente di introdurre il concetto di predicibilità.
- *Predicibilità*: *possibilità di inferire dalla risposta data da un soggetto ad un item posto ad una certa soglia di difficoltà la risposta data alla domanda posta al di sotto di tale soglia*.

La predicibilità fra le risposte di item contigui è ricavata da due diversi punti di vista:

1. dall'item più facile al più difficile: chi non ha superato un item non supera l'item successivo più difficile: P_i;
2. dal più difficile al più facile: chi ha superato un item deve aver superato l'item precedente più facile: P_{i+1}.

Nel caso di predicibilità perfetta se un soggetto supera il terzo item deve aver superato necessariamente anche il secondo e il primo. A ciascun item si attribuisce un punteggio nel modo seguente:

- all'item *predetto* dall'item precedente, ovvero quando i soggetti che hanno superato l'item ha superato anche l'item precedente più facile, si attribuisce punteggio "1";
- all'item non *predetto* dall'item precedente (*indifferenza predittiva*), ovvero quando tra i soggetti che hanno superato l'item solo il 50% supera l'item precedente più facile, si attribuisce punteggio "0".

La perfetta *scalabilità* e la perfetta *predicibilità* possono verificarsi solo nei casi in cui esistano item più "facili" essenziali per la definizione di quelli più "difficili". Per questo lo *scaling* Guttman è

particolarmente adatto a misurare attitudini e qualità crescenti.

3.2.1 La deviazione dal modello: l'errore

La procedura deterministica per stabilire se item e soggetti rispondono al modello (*scalogram analysis*) è basata sull'analisi dei modelli di risposta dei singoli soggetti, *response pattern*, all'insieme di item. Un modello di risposta indica l'insieme di risposte date da un soggetto agli item. Con n item dicotomici vi sono 2^n modelli di risposta possibili; se gli item formano davvero uno *scaling* cumulativo, si devono verificare solamente $n+1$ di tali modelli.
La presenza in un profilo individuale di una risposta positiva preceduta da una negativa costituisce un errore elementare; quindi i profili individuali che registrano errori di scalabilità sono quelli che presentano passaggi diversi da insuccesso a successo, e seguono andamenti del tipo:

−+−−− +−+−+− ++−++ −−+−+ −−++−

E' proprio l'analisi di tali errori che conduce ad una valutazione globale dell'insieme di item. La mancata occorrenza dei modelli devianti consente alla procedura dello scalogramma di ordinare gli individui, gli item e gli intervalli categorici sul continuum sottostante a partire dai dati osservati (*scalogramma perfetto*).
Le deviazioni osservate dal modello perfetto e dalla forma ideale di scalogramma richiesta sono definiti *errori*. Si assume che la quantità di deviazione, o errore, osservata sia una funzione del fallimento degli item e dei soggetti nel conformarsi alle procedure di ordinamento.
La determinazione della quantità di errore non è però univoca: è possibile infatti identificare principalmente due forme di conteggio dell'errore cui corrispondono due approcci diversi alla valutazione dello *scaling*.

3.2.2 Tecniche di valutazione dell'errore

Tecnica Cornell: *Minimizzazione dell'errore* (Guttman, 1947)

La tecnica Cornell è basata sul criterio di minimizzazione dell'errore di una serie di modelli di risposte. L'assunto base di questa tecnica è che nessun item può possedere più errore che *non-errore*. L'ordinamento degli item sul continuum sottostante è una funzione di minimizzazione dell'errore tra le risposte osservate. Tale tecnica presenta il vantaggio di consentire inversioni nell'ordinamento degli item basati sul decremento delle probabilità marginali per raggiungere la più alta riproducibilità possibile. L'ordinamento degli item e i punteggi possono essere una funzione degli errori casuali anziché rappresentare un costrutto di risposte sottostanti agli item. *Il numero di errori rappresenta il numero minimo di risposte positive che devono essere cambiate in negative o di negative che devono essere cambiate in positive per trasformare le risposte osservate nel modello ideale di risposta*. Vediamo un esempio in cui per ciascun profilo di risposta (su quattro item) è riportato il corrispondente conteggio degli errori presenti:

Profili di risposta ⇒	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
	+	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-
	+	+	-	-	+	+	-	-	+	+	-	-	+	+	-	-
	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Assegnazione dell'errore ⇒	0	0	1	0	1	1	1	0	1	1	2	1	2	1	1	0

Il criterio di minimizzazione dell'errore (suggerito da Guttman) appare debole se confrontato con la teoria alla base del modello *Guttman*, in quanto, focalizzandosi solo sugli errori di risposte positive o negative, sottovaluta gli errori dovuti alla posizione. Vediamo un esempio; secondo la tecnica della minimizzazione dell'errore descritta, il modello di risposta -++- osservato contiene un solo

errore rispetto al modello ideale più vicino che è +++-. Visto da un altro punto di vista però il profilo di risposta osservato potrebbe riflettere in realtà due errori: infatti se consideriamo che il modello ideale più vicino è ++-- (due risposte positive e quindi stesso punteggio), per trasformare il profilo osservato (-++-) in quello ideale il numero minimo di segni che devono essere cambiati è due. In questo senso con l'applicazione di questo criterio lo *scaling Guttman* perde una certa componente d'interpretabilità in quanto conduce ad un indebolimento dell'assunto cumulativo su cui è basata e ad una contraddizione della teoria alla base dello scalogramma.

Tecnica Goodenough-Edwards: *Deviazione dalla Perfetta Riproducibilità* (Edwards, 1957)
Una diversa procedura, nota con il nome di *GoodEnough-Edwards*, è basata su due principi tra loro collegati:

1. il modello ideale di risposte di un soggetto è funzione diretta del numero di item cui il soggetto ha risposto in modo positivo;
2. gli item devono essere perfettamente riproducibili a partire dalle risposte dei soggetti.

Quindi l'errore deve essere

- misurato rispetto al numero di risposte che si allontanano dai modelli previsti,
- assegnato sulla base dell'assunto della perfetta riproducibilità,

assicurando il massimo rispetto della posizione degli item e dell'ordinamento dei soggetti.
Aldilà della maggiore plausibilità teorica, questa tecnica presenta anche alcuni vantaggi pratici in quanto l'analisi è basata sulle risposte osservate anziché sulle risposte derivate dalla procedura di minimizzazione dell'errore. Vediamo come vengono valutati gli errori se riprendiamo i profili con quattro item, precedentemente presentati:

Profili di risposta ⇒	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
	+	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-
	+	+	-	-	+	+	-	-	+	+	-	-	+	+	-	-
	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-
Assegnazione dell'errore ⇒	0	0	2	0	2	2	2	0	2	2	2	2	4	2	2	0

Anche se tale procedura conduce all'identificazione di un numero maggiore di errori rispetto al metodo precedente, rappresenta comunque una descrizione più accurata dei dati basati sulla teoria dello scalogramma.

3.2.3 La valutazione dell'adattamento del modello

Dato che è difficile raggiungere una perfetta corrispondenza tra modello teorico e modello osservato, è necessario stabilire, sulla base degli errori rilevati, se il modello osservato risponde alle caratteristiche dal modello teorico di *scaling* cumulativo ovvero è necessario verificare se i modelli di risposta osservati riflettono i modelli ideali di risposte. A tale fine sono stati definiti alcuni indici e coefficienti.

❖ COEFFICIENTE DI RIPRODUCIBILITA' PER CIASCUN ITEM (CR_i)

Sappiamo che la scalabilità è funzione di quanto i modelli di risposta osservati possano essere accuratamente riprodotti sulla base dei punteggi assegnati. Per questo per valutare il livello di scalabilità dei dati empirici ovvero per verificare complessivamente la predicibilità per ciascun item, Guttman propose un coefficiente che, confrontando gli errori osservati con tutti gli errori possibili, può essere interpretato *errore di riproducibilità*:

$$CR_i = 1 - \frac{n_{ie}}{n}$$

dove

n_{ie} errori di riproducibilità tra l'item i e l'item $i+1$ (ovvero numero totale di errori per l'item i)
n numero di risposte all'item (o numero di soggetti).

I valori di CR_i compresi tra 0.85 e 1 sono indici di bontà dello scalogramma. Secondo Torgerson (1958) gli item che presentano valori di CR_i minori di 0.85 dovrebbero essere scartati in quanto, non soddisfacendo il postulato della predicibilità, rivelano una loro appartenenza a una diversa dimensione.

❖ COEFFICIENTE DI RIPRODUCIBILITA' PER TUTTI GLI ITEM (CR)

Eliminati gli item con bassi valori CR_i, si calcola il coefficiente di riproducibilità per l'insieme di item, che dovrebbe assumere un valore centrale fra i valori dei singoli coefficienti degli item grazie alla compensazione degli errori. Per l'insieme di item il coefficiente di riproducibilità (CR) è interpretabile come la *proporzione di risposte agli item che possono essere correttamente riprodotte conoscendo il punteggio totale del soggetto* e rappresenta una misura della bontà di adattamento tra il modello di risposte osservato e quello ideale o previsto:

$$CR = 1 - \frac{\sum n_{ie}}{N}$$

dove

N numero di risposte (*nitem* * *nsogg*)
dove
nitem numero di item
nsogg numero di soggetti.

Il valore minimo accettabile del coefficiente di riproducibilità è considerato 0.90; tale valore indica che l'errore osservato nella riproduzione non supera il 10% delle risposte totali e consente

- di interpretare il punteggio totale e
- di considerare scalabili gli item e rappresentabili in uno *scaling* cumulativo,
- di identificare un'unica dimensione sottostante.

Appare chiaro come il valore di CR dipenda dal metodo con cui si conteggiano gli errori. E' quindi importante, nel presentare i risultati, indicare il metodo utilizzato specificando CR_{error} per il metodo *Guttman* e CR_{ge} per il metodo *GoodEnough-Edwards*.

Esiste la possibilità di stimare l'*errore standard di riproducibilità*, ovvero la generalizzabilità (significatività statistica) della scala:

$$SE_{cr} = \sqrt{(1 - CR)\frac{CR}{N}}$$

❖ *PERCENT IMPROVEMENT* (*INP*%)

L'item cui è associato il valore di CR_i più basso rappresenta il punto più debole del gruppo di item. In tal senso è possibile calcolare un indice del miglioramento che si può ottenere eliminando tale item (*percent improvement*):

$$INP\% = CR - \min(CR_i)$$

❖ *MINIMA RIPRODUCIBILITA' MARGINALE* (*MMR*)

Siccome esiste la possibilità che il coefficiente CR venga influenzato da distribuzioni marginali estreme, è consigliabile confrontare tale valore con la *Minima Riproducibilità Marginale (Minimal Marginal Reproducibility, MMR)*, equivalente alla *sommatoria delle sole frequenze modali di ciascun item*, ovvero il valore minimo di CR tra due item contigui per item dicotomici[2]; ciò richiama l'attenzione sul fatto che la riproducibilità di un item non può essere minore della proporzione delle risposte nella sua categoria modale; analogamente la riproducibilità totale non può essere minore della somma delle proporzioni delle risposte nella categoria modale per ciascun item, diviso per il numero di item.

Il calcolo di *MMR* può avvenire nel seguente modo:

$$MMR = \frac{\sum nm_i}{N}$$

dove

[2] Edwards A.L., "Modal Categories and Minimal Marginal Reproducibility", pp. 191-198, *Techniques of Attitude Scale Construction*, Appleton, Century-Crofts, Inc., New York, 1957 e *package* BMD.

nm_i numero di risposte nella categoria modale (la più *scelta*) dell'item *i*
N numero di risposte (*nitem* * *nsogg*).

Il valore di tale coefficiente riflette la riproducibilità di una serie di item basata solo sulla conoscenza delle distribuzioni marginali degli item[3].

La riproducibilità totale non può essere minore di MMR, ovvero la differenza tra CR e MMR deve essere di tale grandezza da poter attribuire un miglioramento nella previsione dei modelli di risposta allo *scalogram analysis*. Sappiamo che CR e MMR sono nella seguente relazione:

$$CR = \frac{(N - se)}{N} \qquad MMR = \frac{N - me}{N}$$

dove
N numero di risposte (*nitem* * *nsogg*)
se errori di *scaling*
me errori marginali, somma di tutte le frequenze non-modali[4].

I due coefficienti sono nella seguente relazione:

$$CR - MMR = \frac{N - se}{N} - \frac{N - me}{N} \qquad CR - MMR = \frac{me - se}{N}$$

Quindi la differenza tra i due coefficienti è funzione del miglioramento nella previsione fornito dall'insieme di item rispetto alle frequenze marginali degli item individuali.

Il valore di tale differenza va da 0 (nessun miglioramento nella previsione) a me/N (i dati si adattano in maniera perfetta allo *scaling Guttman*). Poiché l'errore massimo marginale che può verificarsi è 50%, tale differenza ha un massimo teorico di .50. Interpretare la differenza tra CR e MMR su una scala da 0 a me/N può essere difficile; per questo motivo sono stati suggeriti altri indici alternativi.

❖ COEFFICIENTE DI SCALABILITA' (CS)

Per meglio interpretare i precedenti coefficienti è importante applicare anche il *coefficiente di scalabilità* (CS) che consente di misurare la *capacità di un insieme di item di prevedere le risposte rispetto alle previsioni basate sulle frequenze marginali*:

$$CS = 1 - \frac{\sum n_{ie}}{me}$$

dove
n_{ie} numero di errori dell'item *i*
me errori marginali, somma di tutte le frequenze non-modali.

Tale coefficiente può essere espresso anche nel modo seguente:

$$CS = \frac{INP\%}{1 - \min(CR_i)}$$

Esso può assumere valori tra 0 e 1. Se le previsioni sono perfette, ovvero non vi sono errori di

---

[3] Di seguito vediamo alcuni esempi in cui si considerano i marginali di risposte positive per quattro item. Prendiamo il caso in cui le probabilità marginali dei quattro item siano .8, .6, .4 e .2 (quinta riga della tabella riportata di seguito) il valore della minima riproducibilità marginale sarà uguale a

$$MMR = (.8 + .6 + .6 + .8) / 4 = .7$$

Osserviamo inoltre come nel caso in cui i marginali osservati siano .95, .85, .15 e .05, MMR risulta essere uguale a .90; da ciò si può concludere che il valore di MMR è funzione dei marginali estremi.

item	1	.50	.60	.70	.70	.80	.80	.90	.85	.90	.95	.95
	2	.50	.60	.50	.70	.60	.80	.70	.85	.90	.85	.95
	3	.50	.40	.50	.30	.40	.20	.30	.15	.10	.15	.04
	4	.50	.40	.30	.30	.20	.20	.10	.15	.10	.05	.05
MMR		.50	.60	.60	.70	.70	.80	.80	.85	.90	.90	.95

[4] La formula generica per calcolare l'errore marginale dell'item *i* secondo l'ipotesi casuale è

$$me_i = c(n - nm)$$

dove
c probabilità di avere una risposta giusta per caso (con item dicotomici = .50)
n numero di risposte
nm numero di risposte nella categoria modale (la più *scelta*) dell'item.

scaling, il valore di CS è 1. Se il gruppo di item non fornisce alcun miglioramento nella previsione (gli errori di *scaling* sono uguali agli errori marginali) il valore è 0. L'insieme di item presenta comunque una buona scalabilità se registra un valore di CS di almeno .60.

❖ COEFFICIENTE DI PREDICIBILITA' (CP_j)

Per ciascun soggetto è possibile calcolare un *coefficiente di predicibilità* (CP_j) basato sulle *previsioni realizzate e le previsioni possibili*:

$$CP_j = \sum \frac{pr_j}{pp_i}$$

dove
pr previsioni realizzate
pp previsioni possibili.

❖ VERIFICA DELL'UNIDIMENSIONALITA'

Come sappiamo per poter adottare un item all'interno di un certo modello occorrono delle ragioni teoriche e delle ipotesi iniziali; in altre parole l'appartenenza degli item ad un unico ambito di contenuto è condizione essenziale per poter assumere un modello triangolare. Come abbiamo visto la verifica dell'unidimensionalità è fatta essenzialmente osservando il livello di riproducibilità, il livello di scalabilità e la casualità degli errori di risposta (la frequente osservazione di un particolare modello di errore fa ipotizzare più dimensioni sottostanti il gruppo di item[5]). L'osservazione di un soddisfacente livello di scalabilità tra item non consente di verificare l'unidimensionalità: infatti un item può risultare scalabile ma non avere alcun contenuto in comune con gli altri item e con la dimensione che si vuole misurare; esistono dei casi in cui item pur molto distanti e diversi tra loro anche nel contenuto consentono comunque di soddisfare il modello triangolare, come nel seguente esempio con quattro item:

a. *risolvi rispetto a* x *la seguente equazione:* $x^2+2x+9=16$
b. *qual è il significato della parola* severo*?*
c. *quanto fa "10*38"*
d. *quando usi l'ombrello?*

E' possibile che tali item somministrati a ragazzi tra i 10 e i 16 anni possano riprodurre in modo eccellente il modello triangolare dello *scalogramma*, in quanto qualsiasi soggetto che esegue il primo item correttamente probabilmente esegue correttamente anche gli altri, pur non appartenendo e non misurando lo stesso attributo.[6]

Vediamo schematicamente i diversi momenti di verifica del modello:

---

[5] In alcuni casi l'osservazione di errori sistematici può risultare utile all'identificazione di particolari gruppi di soggetti; nel caso per esempio di scale per la misurazione dell'autosufficienza fisica negli anziani è possibile individuare soggetti portatori di disabilità legate ad aspetti specifici, e quindi sul piano epidemiologico meno frequenti o imputabili ad aspetti non connessi con i processi di invecchiamento, che richiedono interventi individuali e particolari (ad esempio, una mutilazione ma anche fattori ambientali, barriere architettoniche ecc.).

[6] Secondo alcuni autori l'unidimensionalità di una scala Guttman può essere verificata indirettamente tramite il coefficiente di dipendenza di Yule: prendiamo in considerazione due item dicotomici *i* e *j* e costruiamo la loro distribuzione incrociata:

		i		
		si	no	
j	si	a	b	*(a+b)*
	no	c	d	*(c+d)*
		(a+c)	*(b+d)*	*n=(a+b+c+d)*

se *i* esprime un intensità maggiore rispetto a *j*, la frequenza *c* tenderà a 0 (compatibilmente con la presenza di errori casuali); il coefficiente *Q* di Yule consente di verificare l'ipotesi che *c* sia significativamente uguale a 0:

$$Q = \frac{(ad - bc)}{(ad + bc)}$$

Q, che non è altro che il coefficiente *gamma* (v.) nel caso particolare di dati dicotomici, assume valore 0 quando le variabili sono indipendenti mentre raggiunge l'unità tutte le volte che una qualsiasi delle quattro celle si presenta vuota. Secondo gli autori un valore di $Q \geq 0.80$ dovrebbe essere considerato soddisfacente.

VERIFICA DEL MODELLO DETERMINISTICO

⇓

A. ANALISI DELLA RIPRODUCIBILITA'

1. SELEZIONE DEGLI ITEM

a. *Coefficiente di riproducibilità* per ciascun item CR_i

$$CR_i = 1 - \frac{n_{ie}}{n}$$

bontà di adattamento $\Rightarrow CR_i \geq 0.85$

↓

b. *Percent Improvement INP%*

$$INP\% = CR - \min(CR_i)$$

Indice del miglioramento che si può ottenere eliminando l'item con il valore CR_i più basso

↓

c. Eliminazione degli item con $CR_i > 0.85$

⇓

2. RIPRODUCIBILITA' DELLA SCALA

a. *Coefficiente di riproducibilità CR*

$$CR = 1 - \frac{\sum n_{ie}}{N}$$

proporzione di risposte agli item che possono essere correttamente riprodotte conoscendo il punteggio totale del soggetto

↓

misura della bontà di adattamento tra

- il modello di risposte osservato
- il modello di risposte ideale o previsto

$CR > 0.90 \rightarrow$

- la scala riflette il 10% di errori
- consente di interpretare la scala
- consente di considerare scalabili gli item

b. *Minima Riproducibilità Marginale (Marginal Reproducibility, MMR)*

$$MMR = \frac{\sum nm_i}{N}$$

Sommatoria delle sole frequenze modali di ciascun item.
La riproducibilità totale non può essere minore di *MMR*

B. SCALABILITA'

1. Coefficiente di scalabilità (CS)

$$CS = 1 - \frac{\sum n_{ie}}{me} \quad \text{oppure} \quad CS = \frac{INP\%}{1 - \min(CR_i)}$$

Capacità di una scala di prevedere le risposte rispetto alle previsioni basate sulle frequenze marginali.

$CS \geq 0.60 \rightarrow$ scala con buona scalabilità
$CS = 1 \rightarrow$ previsioni perfette $\rightarrow$ non ci sono errori di scala
$CS = 0 \rightarrow$ scala non fornisce alcun miglioramento nella previsione

C. ANALISI DELLE PREDICIBILITA'

1. Coefficiente di predicibilità per ciascun soggetto (CP_j)

$$CP_j = \sum \frac{pr_i}{pp_i}$$

dove

n_{ie}	numero totale di errori perl'item i = errori di riproducibilità
n	numero di risposte all'item (o numero di soggetti)
N	numero di risposte (*nitem*nsogg*)
nitem	numero di item
nsogg	numero di soggetti
nm	numero di risposte nella categoria modale dell'item
me	errori marginali, somma di tutte le frequenze non-modali
pr	previsioni realizzate
pp	previsioni possibili

L'adattamento è considerato soddisfacente quando sono soddisfatti almeno tre requisiti:

a. $CR \geq .90$;
b. valore di *MMR* non eccessivamente alto;
c. differenza tra *CR* e *MMR* che indichi che si sia realizzato un miglioramento nella scalabilità in funzione della conoscenza dei punteggi totali;
d. errori non sistematici nei modelli di risposta.

3.3 LIMITI DEL MODELLO DETERMINISTICO

Il più serio limite dello *scaling Guttman* è dato proprio dall'approccio deterministico. Tale modello non fornisce alcuna spiegazione delle deviazioni dai modelli di risposte perfettamente scalabili, ovvero non presenta alcuna teoria riguardo agli errori di *scaling*. Questi si verificano quando un soggetto dà una risposta negativa ad un item il cui punto è posizionato a sinistra del punto-soggetto o quando un soggetto dà una risposta positiva ad un item il cui punto è posizionato a destra del punto-soggetto. In entrambi i casi le risposte contraddicono il modello, data la geometria della soluzione di *scaling*. Nello standard dell'analisi dello *scaling Guttman* gli errori rimangono inspiegati: a partire dalla loro conoscenza si cerca solo, come vedremo, di sviluppare metodi che consentano di assegnare punteggi di scala alle osservazioni che presentano tali errori. Essi rimangono però un impedimento alla spiegazione della variabilità dei dati.

3.3.1 Problemi di applicazione dello scalogramma

3.3.1.1 Il calcolo del punteggio individuale

Considerando item a risposte dicotomiche, il calcolo dei punteggi individuali è effettuato sommando il numero di risposte positive; in questo senso è identico al corrispondente metodo utilizzato per il modello additivo (scale *Likert*). Relativamente al punteggio totale, tra i due modelli esistono però delle differenze:

- modello additivo (scala *Likert*):
  - *calcolo punteggi*: quando l'affidabilità dell'insieme degli item è elevata;
  - *interpretazione dei punteggi*: un punteggio di 2 indica che il soggetto ha risposto in modo positivo a 2 tra gli item.
- modello cumulativo (scala *Guttman*):
  - *calcolo punteggi*: quando il valore del *CR* è maggiore di .90;
  - *interpretazione dei punteggi*: un punteggio di 2 indica che il soggetto ha risposto in modo positivo a 2 specifici item, quelli più *facili* o più *accettabili*.

3.3.1.2 Calcolo del punteggio con item a risposte multiple

Sicuramente il modello deterministico-cumulativo è più semplice da comprendere e da applicare nel caso in cui si utilizzino item dicotomici. Per questo la maggior parte dei modelli di scalogramma sono costruiti con item a risposte dicotomiche o riconducibili a valutazioni del tipo *successo*|*fallimento* nella prova. D'altra parte esistono, come già Guttman aveva indicato, applicazioni delle tecniche di *scalogram analysis* che utilizzano valutazioni graduate delle attitudini; si tratta di item a risposte multiple (ovvero un numero di categorie maggiore di due) con valutazione decrescente (Tesi, 1993). Le risposte multiple *graduate* permettono una maggiore finezza nella rilevazione delle posizioni individuali, ma rendono più complesso il calcolo degli errori di scalabilità (per ogni soggetto e ogni item).
E' però difficile trovare item con molte categorie tra loro distinguibili in modo chiaro e non ambiguo per tutti; per tale motivo spesso si consiglia di combinare tra loro le categorie che non si presentano chiaramente distinte. Sono state per questo stabilite diverse strategie di combinazione delle categorie; tra queste ricordiamo il metodo delle approssimazioni successive; tale criterio è basato sull'osservazione delle *abitudini verbali* dei soggetti che condividono posizioni simili pur indicando categorie diverse. In tale ottica si possono combinare tra loro le categorie considerate *neutrali*, quelle considerate *positive*, quelle considerate *negative*.
Come si può immaginare la validità di tale metodo è stata messa in discussione a causa del suo carattere empirico e soggettivo. Con altre tecniche, le risposte vengono dicotomizzate con una soglia arbitraria, sotto la quale le risposte sono valutate negative, sopra positive. Quando non è possibile una chiara dicotomizzazione per tutti gli item o si ritiene sia più corretto prevedere un sistema di punteggi differenziato per ogni item, è necessario porre particolare attenzione al peso da attribuire alle singole risposte per ciascun item. Quando la scala delle risposte è di questo tipo il coefficiente *MMR* assume una particolare importanza per la validazione dello scalogramma.

3.3.1.3 Assegnazione dei punteggi nel caso di errori

Analizzata e verificata la rispondenza della scala al modello cumulativo, il passaggio successivo è

quello dell'assegnazione dei punteggi ai soggetti. Nel caso in cui la scala utilizzata è perfetta, l'assegnazione dei punteggi è molto semplice e chiara: il punteggio di ciascun soggetto sarà uguale al numero di risposte positive. Il problema sorge quando le risposte di un soggetto non sono scalate ovvero quando il modello di risposta non riflette un modello ideale.
Uno dei problemi che si pongono nell'analizzare tali tipi di scale riguarda infatti l'opportunità di considerare anche i soggetti il cui modello di risposta devia da quello ideale e quindi di calcolare il punteggio totale. Non è possibile dare una soluzione unica a tale problema; il ricercatore può sostenere che l'individuo che non rispetta l'ordine degli elementi è in errore solo se il modello teorico sostiene un perfetto schema ordinale degli item. E' necessario essere consapevoli del fatto che la decisione di accettare o respingere la scala è comunque arbitraria e che il ricercatore dovrà prima o poi risolvere il problema dell'*errore*. La violazione del modello triangolare di risposta può suggerire che

a. uno degli assunti sottostanti il modello *Guttman* non è valido rispetto ai dati;
b. lo spazio in cui gli stimoli e i soggetti sono posizionati non è unidimensionale[7].

La raccomandazione è quindi quella di adottare una scala Guttman solo se il numero di errori è abbastanza piccolo (alto valore del coefficiente di riproducibilità). In questi casi occorre comunque prestare molta attenzione a come si calcola il punteggio totale.
Assumiamo che, su uno scalogramma con 5 item, sia stato osservato il seguente modello di risposta -++--. La prima osservazione che si può fare è che non esiste alcun numero da 0 a 5 che consenta di definire in modo accurato il soggetto sul continuum sottostante. D'altra parte è necessario affrontare il problema di quale punteggio assegnare in modo che l'interpretazione della scala ne sia influenzata il meno possibile.
Come abbiamo visto in questi casi si procede all'assegnazione dell'errore rispetto alla quale sono state sviluppate due forme:

1. assegnazione basata sui modelli di risposta ideali più vicini (*tecnica Cornell*),
2. assegnazione basata sul numero di risposte positive (*tecnica GoodEnough-Edwards*).

Se applichiamo il primo criterio, il modello osservato viene riclassificato nel modello più vicino +++-- cui viene assegnato il punteggio di 3.
Se applichiamo il secondo criterio a tale *pattern* di risposta, nel quale sono presenti solamente due "+", viene assegnato il punteggio 2.
E' evidente quindi come la possibilità di gestione diversa dell'errore introduce una seria limitazione nella interpretazione dei risultati della scala per i soggetti che presentano risposte non scalari.
Poniamo di avere osservato un altro modello: -+-+-.
Se applichiamo il primo criterio tale modello di risposta può essere riferito a diversi modelli ideali: ++++-, ++--- oppure -----; conseguentemente il punteggio che può essere assegnato è 4 oppure 2 oppure 0; quindi l'applicazione del primo criterio induce una certa arbitrarietà di assegnazione sia dell'errore che del punteggio individuale.
L'applicazione del secondo criterio conduce comunque all'assegnazione del punteggio 2.
E' comunque vero però che il problema dell'assegnazione è uno pseudo-problema dal momento che per poter procedere al calcolo del punteggio individuale è necessario che la scala risponda al modello cumulativo ovvero deve registrare un elevato valore di riproducibilità ($CR \geq .90$). E' comunque consigliabile utilizzare sempre il secondo criterio ovvero quello che registra la deviazione dalla perfetta riproducibilità.

3.3.1.4 Calcolo degli errori con item a risposta multipla

L'adozione di item a risposta multipla pone un altro problema. Usando i modelli di validazione per item dicotomici, ogni singolo errore di predicibilità è valutato 1, qualunque sia l'entità della

[7] Una ulteriore validazione di queste scale è infatti quella che riguarda la verifica dell'omogeneità degli item (tutte le domande devono interessare o misurare una sola capacità); a tal fine è possibile utilizzare l'analisi fattoriale.

differenza fra i punteggi di due item contigui. Ma con una valutazione a più punti il significato della differenza può essere più o meno rilevante; è quindi necessario sviluppare indicatori in grado di pesare gli errori di predizione (un punteggio più basso in un item più difficile) sulla base della distanza tra le valutazioni contigue. In pratica, l'entità di questa differenza è ricuperabile calcolando per ogni item il numero di errori commessi e la sommatoria dell'entità degli scarti.
E' per questo che accanto all'indice classico (CR_i) è stato sviluppato un altro coefficiente di scalabilità (CR_{iw}) (Tesi G. e al., 1993) che viene ricavato

- dalla proporzione di soggetti che non hanno superato un item ma hanno superato un item successivo più difficile diviso il totale dei soggetti che hanno superato la prova *i*+1 (n_e),
- dalla sommatoria della *distanza* fra le risposte negli stessi soggetti con errori di scalabilità (n_s).

Sulla base di queste due valutazioni è possibile calcolare gli indici di riproducibilità, rapportando la sommatoria degli scarti al valore massimo della sommatoria:

$$CR_i = 1 - \frac{n_e}{n} \qquad CR_{iw} = 1 - \frac{n_s}{n * p_{mm}}$$

n_e errori di riproducibilità tra l'item *i* e l'item *i+1* (ovvero numero totale di errori per l'item *i*)
n_s scarti negativi tra l'item *i* e item *i+1*
n numero risposte (o numero di soggetti)
p_{mm} punteggio_*max* - punteggio_*min*

Analogamente il *coefficiente globale di riproducibilità* dello scalogramma, che esprime quanto l'insieme di item soddisfi la condizione *cumulativa* delle attitudini richieste per svolgere il compito preso in esame, è calcolato a partire dal totale degli errori in tutti gli item:

$$CR_w = 1 - \frac{\sum n_s}{N * p_{mm}}$$

n_s scarti negativi tra item contigui (*i* e *i+1*)
N numero di risposte (*nitem* * *nsogg*)
p_{mm} punteggio_*max* - punteggio_*min*

3.3.1.5 Matrice troppo grande

Nello *scalogram analysis* la presenza di troppi dati (item e soggetti) rappresenta un problema di complessa gestione. La complessità di analisi e di controllo consigliano in genere di adottare scale Guttman composte da pochi item anche se ciò potrebbe non consentire una corretta discriminazione tra i soggetti. D'altra parte la disponibilità di molti dati è auspicabile nei momenti di messa a punto della scala. Per ridurre la matrice analizzata ad una dimensione gestibile, sono stati messi a punto diversi metodi; tali metodi possono rappresentare dei validi sostegni nello sviluppo di una scala.

- Tecnica *H* (o degli *item inventati*)
  Se gli item da scalare fanno tutti parte dello stesso teorico universo di contenuto, ma nello stesso momento non sono importanti individualmente, è possibile creare *item inventati* combinando le risposte in due o più item. Gli item combinati sono quelli che presentano alla luce dei risultati la stessa difficoltà ovvero occupano la stessa posizione sul continuum di misurazione. Ciascuna risposta individuale ricondotta in un item è determinata sulla base delle risposte del soggetto a tutti gli item che costituiscono la nuova variabile inventata. L'utilizzazione di *item inventati* consente di eliminare risposte non scalari all'interno del gruppo di item combinati; utilizzando le risposte è possibile individuare la posizione individuale più probabile rispetto al gruppo degli item combinati. Tale tecnica si presta a diverse applicazioni, ma anche critiche, ed è stata utilizzata in origine per incrementare la riproducibilità. Può essere comunque utilizzata per consolidare i dati prima che l'analisi cominci.
- Tecnica dei *soggetti inventati*
  Si tratta di un metodo studiato per incrementare l'adattamento tra dati e modello cumulativo (riproducibilità) senza modificare, eliminare o combinare gli item. Nel caso in cui l'interesse è più orientato verso una analisi globale del fenomeno (attitudini del gruppo più che dei singoli individui) non è necessario conservare i modelli di risposta individuali. Conseguentemente è possibile creare un soggetto inventato allo stesso modo degli item inventati; ovvero è possibile combinare gli individui che presentano lo stesso punteggio totale in un unico soggetto ipotetico assumendo come punteggi per gli item le risposte predominanti dei componenti del gruppo.

- Eliminazione di item con marginali estremi
  Gli item con i marginali estremi non aumentano in modo significativo il numero di errori di scala e forniscono poche informazioni sulla loro reale appartenenza ad una data scala.
- Eliminazione di item con marginali simili agli item presenti nella scala
- Campionamento di item
- Mantenimento dei soli item significativi
- Analisi di sottoinsiemi di contenuto.

3.3.1.6 Dati missing

Un altro importante problema che l'analisi dello scalogramma deve affrontare è rappresentato dal modo di trattare i dati *missing*. Tale problema è comunque comune alla maggior parte delle ricerche empiriche ed è difficilmente evitabile (è difficile ricontattare i soggetti per completare le informazioni mancanti). Per tale motivo molti ricercatori sono spesso obbligati a sviluppare qualche procedura statistica per evitare la perdita di dati in parte già raccolti. Tali procedure riguardano i particolari criteri di analisi delle matrici che presentano dati *missing*. Alcuni criteri prendono in considerazione i modelli di risposta del soggetto a tutti gli item della scala:

a. se il soggetto non ha risposto a più del 50% degli item, il soggetto risulterà *non classificato*;
b. se il soggetto non ha risposto a meno del 50% degli item, alle risposte *missing* si attribuisce un valore corrispondente al punteggio medio della totalità dei soggetti.

Esistono altri complessi schemi per prevedere e attribuire un punteggio al valore *missing* a partire dalle risposte osservate o da altre informazioni. In alcuni casi la complessità e l'arbitrarietà di tali schemi annullano qualsiasi beneficio derivato dal loro utilizzo.

3.3.1.7 Significatività statistica

A tale modello di *scaling* vengono mosse altre critiche riguardanti in particolare la sua *significatività statistica*; secondo tali critiche le scale Guttman presentano le seguenti caratteristiche:

- *Instabilità: tali scale, soprattutto se composte da pochi item, sono soggette a variazioni ed errori casuali*. In realtà la scala Guttman, anche se composta da pochi item, conserva tutti gli elementi per il controllo della consistenza interna attraverso gli aspetti della predicibilità.
- *Poca significatività: lo scalogramma rappresenta un modello poco generalizzabile.* In effetti la verifica della scalabilità di un gruppo di item su un particolare campione non consente di affermare che applicando la stessa su un altro campione si ottengano gli stessi risultati complessivi. Si comprende quindi l'importanza della rappresentatività del campione su cui viene effettuata la validazione della scala.
  Anche se, come abbiamo visto, sono stati fissati alcuni criteri di accettabilità della scala *Guttman* $(CR > 0.9, MMR < 0.9, CS > 0.6)$ è possibile che la loro soddisfazione avvenga per caso e che quindi non fornisca una prova definitiva della aderenza degli item al modello cumulativo. A tale riguardo sono stati sviluppati diversi test di significatività come supporto aggiuntivo alla scalabilità degli item. Per verificare ciò è possibile effettuare controlli attraverso il *chi-quadro*, come confronto tra distribuzione reale e distribuzione teorica (casuale) e l'errore standard di riproducibilità (Guttman:

$$ES_{rep} \cong \sqrt{(1 - CR) * CR / N}$$

In conclusione va comunque ricordato come particolari modelli, nonostante le difficoltà applicative, spesso risultano importanti nella ricerca per lo sviluppo di teorie e dei relativi modelli matematici.

3.4 ALLE ORIGINI DEL MODELLO DETERMINISTICO: LA SCALA BOGARDUS

La scala Bogardus propone un modello che può essere considerato il precedente storico di quello deterministico. Tale scala è nata per poter misurare gli atteggiamenti verso gruppi etnici ma ha trovato applicazioni anche per misurare gli atteggiamenti verso classi sociali, gruppi religiosi, e così via. In questa scala gli item vengono definiti come gradini di una scala con *problematicità* crescente. La scala originariamente misurava la *distanza sociale* (Bogardus, Emory, "Racial Distance Changes in the United States During the Past Thirty Years", *Sociology and Social Research*, Vol. XLIII, pp. 127-130, November, 1958); tale concetto si riferisce al modo e all'intensità con cui le persone percepiscono e tendono ad avere rapporti con altre persone nelle diverse situazioni di interazione sociale; essa doveva individuare i diversi gradi in cui tale fenomeno si manifesta, in modo da fornire dei dati per una adeguata interpretazione; le affermazioni che definivano gli item indicano gerarchicamente diversi livelli di "intimità" o di contatto sociale, con un certo gruppo di individui: si passa infatti dallo stretto vincolo di parentela, attraverso i legami di amicizia, di lavoro, fino al rifiuto di un qualsiasi contatto. In particolare la scala originaria era basata su una serie di domande che richiedevano se si era disposti ad accettare i neri come residenti nella stessa città, come vicini di casa, come amici, come mariti della figlie. La scala assume risposte cumulative in quanto chi accetta un negro in casa lo accetta anche come vicino di casa. Di seguito vediamo una sequenza di insieme di item che definisce una tipica scala *Bogardus*:

1.	Lo accetterei come parente stretto	5.	Lo accetterei come concittadino
2.	Lo accetterei al mio club, come amico personale	6.	Lo accetterei nel mio paese come turista
3.	Lo accetterai come vicino di casa	7.	Lo escluderei dal mio paese
4.	Lo accetterei come compagno di scuola		

L'ipotesi dell'autore è che quanto più è stretto il contatto che il soggetto è disposto ad ammettere con una persona ad un certo gruppo, tanto più favorevole è il suo atteggiamento nei confronti delle persone di tale gruppo; al contrario, quanto meno intimo è il contatto che il soggetto è disposto ad ammettere, tanto meno favorevole è il suo atteggiamento. Dal punto di vista dell'analisi il giudizio emesso dal soggetto può essere sintetizzato in due indici:

a. *Ampiezza del contatto sociale* (*Social Contact Range*, *SCR*): indice rappresentato dal numero di item di contatto sociale che il soggetto è disposto ad ammettere.
b. *Distanza del contatto sociale* (*Social Contact Distance*, *SCD*): indice rappresentato dal numero di item di più stretto contatto sociale che non vengono scelte dal soggetto.

Tra i due indici vi è una relazione inversa: quanto minore è il numero di item di contatto ammesse, tanto maggiore è la distanza sociale che il soggetto intende stabilire tra sé e la persona giudicata. Naturalmente nel caso di rifiuto totale della persona il *SCR* assume valore 0, dato che non è ammesso nessun tipo di contatto, mentre *SCD* diventa pari al massimo ottenibile. Nel caso in cui si sottoponga la scala ad un campione di soggetti, è possibile calcolare

- *SCR-medio*: indice medio del numero di item di contatto sociale scelte dai soggetti;
- *SCD-medio*: indice medio del numero di item di contatto sociale non accettate dai soggetti.

E' importante che i due indici siano utilizzati contemporaneamente in quanto il solo indice di ampiezza del contatto sociale non fornisce un'informazione completa. Uno dei problemi riscontrati in questo tipo di scala è dato dal fatto che non esiste alcun modo per determinare la reale distanza tra i vari punti sulla scala. Tale scala risulta comunque utile nei casi in cui si vogliano studiare gli atteggiamenti degli individui verso particolari gruppi di persone ma anche nei casi in cui si vogliano misurare gli atteggiamenti verso classi sociali, occupazioni e gruppi religiosi. Infatti le categorie di accettazione o di rifiuto possono essere formulate in maniera diversa a seconda delle esigenze della ricerca condotta; è possibile per esempio affrontare ricerche riguardanti contesti culturali diversi o aree di problemi diversi (negli anni passati in Italia si sono visti esempi di applicazione nello studio del pregiudizio nei confronti dei meridionali o degli atteggiamenti nei confronti degli obiettori di coscienza mentre di recente si sono viste applicazioni che riguardavano i rapporti con gli extra-comunitari).

3.5 ALTRI MODELLI DETERMINISTICI

Il modello di scalogramma che abbiamo appena analizzato richiede che vengano soddisfatti diversi assunti riguardanti la caratteristica misurata (unidimensionale) e gli item (ordinabili lungo il continuum della caratteristica). E' comunque possibile ipotizzare scalogrammi più complessi, in cui non necessariamente si ipotizza una caratteristica unidimensionale e gli item sono ordinabili in modi diversi. A tale proposito è possibile identificare due diversi approcci:

- *Multidimensional Scalogram Analysis* (*MSA*), il cui obiettivo è quello di rappresentare i profili attraverso punti in uno spazio in modo tale che tale spazio possa essere suddiviso da ciascun item secondo le sue categorie di misurazione;
- *Partial Ordered Scalogram Analysis* (*POSA*), rappresenta i profili individuali in uno spazio geometrico, preservando le relazioni d'ordine tra essi.

In generale tali approcci risultano utili come:

a. strumenti per la definizione di ipotesi strutturali riguardanti le osservazioni;
b. procedure per valutare fenomeni la cui struttura empirica richiede più di un punteggio.

3.5.1 Modelli alternativi di scalogramma

Non sempre gli item utilizzati consentono di descrivere un perfetto scalogramma cumulativo; tra di essi è possibile osservare, all'interno del concetto di scalogramma relazioni diverse da quella osservata (cumulativa). Infatti la mancata osservazione di uno scalogramma perfetto non indica necessariamente una scorretta definizione di un modello o la sbagliata costruzione di uno strumento. In questi casi è forse possibile definire un modello di scalogramma che descriva delle nuove relazioni tra item o, meglio come vedremo, tra profili. Quando i profili individuali ottenuti con le osservazioni empiriche risultano mutuamente non confrontabili può essere comunque possibile identificare un'ipotesi di scala sottostante che ci consenta di sintetizzare in un punteggio interpretabile le informazioni degli item. Come vedremo il punteggio può non essere unico.
Secondo un approccio è necessario perseguire due obiettivi:

a. individuazione per ciascun caso di più profili e assegnazione a ciascuno di essi di un punteggio;
b. analisi delle variabili sottostanti (diversamente dallo scalogramma Guttman che ne presenta solo una).

Osserviamo i seguenti profili individuali ottenuti con quattro item (il codice '1' indica "fallimento" e il codice '2' indica "superamento"):

		item				n. di performance positive
		I_1	I_2	I_3	I_4	
casi	1	1	1	1	1	0
	2	1	1	2	1	1
	3	1	1	1	2	1
	4	2	1	1	1	1
	5	1	2	1	1	1
	6	1	1	2	2	2
	7	2	1	1	2	2
	8	2	2	1	1	2
	9	2	1	2	2	3
	10	2	2	1	2	3
	11	2	2	2	2	4
n. di performance positive		*6*	*4*	*4*	*6*	

Come si può facilmente osservare l'ipotesi cumulativa dello scalogramma Guttman qui non è sostenibile. Infatti pur avendo ordinato i punteggi individuali, questi non consentono di *riprodurre* le *performance* individuali relative ai singoli item.
Se però si ritiene che i profili ottenuti siano comunque legittimi è necessario operare in modo diverso[8]. Con il *diagramma Hasse* (detto anche *partial order diagram*) è possibile rappresentare i profili osservati; in esso ciascun profilo è scritto solamente una volta. Vediamo la rappresentazione degli undici profili presentati (in questa rappresentazione è rispettata la sequenza degli item: 1, 2, 3, 4):

Sequenza degli item 1234	Numero di performance positive
1111	0
1121 1112 2111 1211	1
1122 2112 2211	2
2122 2212	3
2222	4

Nel tracciare tale diagramma può essere utile mettere sullo stesso livello orizzontale i profili che presentano lo stesso numero di *performance* positive. Tale diagramma, ottenuto per prove ed errori, descrive le relazioni di confrontabilità tra i profili e soddisfa le condizioni del diagramma di ordine parziale: solo i profili confrontabili sono collegati da linee discendenti; notare inoltre che non presenta linee intrecciate. Un gruppo di profili che risponde a tali caratteristiche è almeno bidimensionale.

3.5.1.1 Il modello diamante

Riordinando gli item rispetto ai totali di colonna visti nella loro presentazione (la nuova sequenza è I_3, I_4, I_1 e I_2) è possibile osservare una certa interessante regolarità; il nuovo ordinamento ha tenuto conto di una nuova caratteristica osservabile in senso verticale: la *media dei ranghi delle posizioni*

[8] Un modo per esempio può essere quello di identificare e isolare all'interno del gruppo di profili più sottogruppi che descrivono uno scalogramma Guttman.

dei successi (*center of success*):

Sequenza degli item 3412							Numero di performance positive
			1111				0
2111		1211		1121		1112	1
	2211		1221		1122		2
		2221		1222			3
			2222				4
1	1.5	2	2.5	3	3.5	4	⇐ Media dei ranghi delle posizioni dei successi

Per riprodurre ciascun profilo è necessario sia il punteggio di riga che quello di colonna; per esempio i punteggi 2 (di riga) e 3.5 (di colonna) consente di identificare solo il profilo 1122.
Tale configurazione è detta "configurazione diamante" o "diagramma diamante", la cui logica può essere estesa ad un numero superiore di variabili.
La generalizzazione può riguardare anche il numero di possibili categorie (1, 2, 3, ..); in questo caso è necessario fare specificazioni aggiuntive riguardanti i pesi relativi da attribuire alle categorie.
L'ipotesi *diamante*, quando confermata dai dati, presenta potenti implicazioni teoriche e pratiche. Essa determina:

- la struttura dei concetti che sono misurati,
- gli strumenti di misurazione necessari per valutare completamente i casi rispetto a tali concetti.

Tale scalogramma presenta aspetti radicalmente diversi da quelli molto rigidi dello scalogramma Guttman; quest'ultimo può essere considerato un caso particolare della configurazione *diamante*.
Quando la stessa configurazione di profili ricorre in più applicazioni empiriche, aumenta la fiducia negli aspetti strutturali manifestati dai dati.

La logica interna della struttura e dei processi

Come abbiamo visto, tra i pochi concetti applicabili allo *scaling* Guttman vi è quello di "difficoltà"; nello scalogramma tipo *diamante* abbiamo notato come nella configurazione è possibile riscontrare un ordine tra gli item in modo tale che ogni profilo presenti una singola sequenza di *performance* positive (ovvero serie di *performance* positive non interrotte).
Un altro assunto che consente di stabilire il criterio di appartenenza di un profilo ad una configurazione *diamante* è il seguente: dato un particolare ordine tra gli item, se un caso ha superato due item deve aver superato anche tutti quelli posti tra i due in questione.
A partire da tale assunto diviene evidente un'interessante caratteristica della configurazione *diamante*: ogni profilo presenta un *primo* item superato e un *ultimo* item superato. Nel profilo "112221" il primo item superato è il terzo, mentre l'ultimo è il quinto.
E' possibile a questo punto definire una configurazione diamante sulla base di due punteggi:

a. posizione nel profilo del *primo* item superato,
b. posizione nel profilo dell'*ultimo* item superato.

Osserviamo il seguente diagramma in cui è possibile realizzare contemporaneamente due forme di *scaling*.

1° punteggio di scala: posizione del primo item superato							
	6	211111	221111	222111	222211	222221	222222
	5		121111	122111	122211	222221	122222
	4			112111	112211	112221	112222
	3				111211	111221	111222
	2					111121	111122
	1	111111					111112
		1	2	3	4	5	6
		2° punteggio di scala: posizione dell'ultimo item superato					

Notare che nel 1° punteggio di scala, il punteggio più alto viene attribuito a quei profili in cui, lungo la sequenza di item determinata dalla configurazione *diamante*, il primo "successo" avviene *presto*. Analogamente nel 2° punteggio di scala, il punteggio più alto viene attribuito a quei profili in cui l'ultimo "successo" avviene *tardi*. Si fa questo in modo da poter associare punteggi alti in entrambe le scale con la situazione in cui il caso "supera" molti item. Quindi il significato generale delle due scale rimane lo stesso ("successo" o "tendenza al successo") mentre ciascuna di esse mantiene il proprio specifico e ben definito significato ("inizio" e "fine" dei "successi").
Utilizzando la rappresentazione cartesiana (anche in termini ordinali più che quantitativi) ha consentito di posizione (o riprodurre) i profili osservati; tali scale presentano una simmetria semantica nei loro significati, cosa che non era stata riscontrata con le due precedenti scale (*numero di performance positive* e *media delle posizioni dei successi*). Tale caratteristica (condivisione di un generale significato comune) estende la nozione di *scaling* a complesse configurazioni di profili, compreso quelle di alta dimensionalità.

Scaling che preserva l'ordine

Come abbiamo visto nel caso della configurazione *diamante*, studiando la struttura dei profili è stato possibile assegnare due punteggi ad ogni profilo: il primo item superato - l'ultimo item superato; tali punteggi sono determinati con riferimento ad uno specifico ordine tra gli item. Dato un insieme di profili che risponde al criterio *diamante*, l'assegnazione dei due punteggi consentono di riprodurre un unico profilo.
Un modo per estendere tale metodo di *scoring* a altre configurazioni più complesse è quello di aderire alla nozione di ordine tra gli item e di classificare i modelli di successo che possono apparire relativi ad un determinato ordinamento fisso. Un ulteriore elemento di complessità potrebbe essere quello di considerare i profili che contengono due sequenze di "successi"; tale situazione si presenta tridimensionale in quanto per poter riprodurre ciascun profilo sono necessari tre punteggi di scala.
Non sempre però i ricercatori hanno una chiara idea della specifica struttura dei concetti utilizzati o dei meccanismi dei processi che studiano in modo da poter formulare un'ipotesi in termini di modelli attesi e di profili legittimi.
Prendiamo in considerazione i punteggi assegnati ai profili nel diagramma precedente. Risulta abbastanza chiaro come rispetto al primo punteggio, i profili che condividono lo stesso punteggio definiscono una scala Guttman. Tale *scalabilità* può essere chiamata scalabilità condizionale: il condizionamento riguarda i valori del 1° punteggio. In questo caso la scalabilità è indicata dai valori del 2° punteggio (ricordiamo che in questo caso i profili selezionati condividono lo stesso punteggio rispetto al primo). E' possibile procedere anche in senso inverso.
Quando due profili sono confrontabili (uno è maggiore dell'altro) anche i rispettivi punteggi di scala manterranno la stessa relazione, come si può osservare dalla seguente tabella:

		punteggi	
profili	111111	1	1
	111211	3	4
	112211	4	4
	112221	4	5
	122221	5	5
	222221	6	5
	222222	6	6
		⇓	⇓
		Punteggi di tipo Guttman	

A questo punto è possibile definire la proprietà del diagramma: *il diagramma converte ogni profilo (composto da* n *item) in due punteggi di scala in modo tale che ogni coppia di profili confrontabili diviene una coppia di due punteggi confrontabili che conservano tra loro la stessa relazione d'ordine. Solamente profili confrontabili sono convertibili in punteggi di scala confrontabili*. Tale proprietà è associata ad un'altra: in studi ben disegnati *le scale identificate condividono lo stesso significato di un concetto esteso del quale rappresentano aspetti diversi*.
L'obiettivo della rappresentazione di una configurazione *diamante* è non solo quello di economizzare l'informazione (obiettivo di *scaling*) ma anche quello di consentire l'identificazione di fattori sottostanti i concetti studiati. Un possibile criterio per definire molte scale può essere il seguente: dato un insieme di profili osservati, si potrebbe cercare di identificare uno spazio di coordinate cartesiane, con il numero più piccolo di assi, che consenta di rappresentare tutte le possibili relazioni tra i profili, compresa la "non-confrontabilità".
Il compito di identificare scale che preservino l'ordine può essere complesso e arduo anche nel caso bidimensionale. Le procedure automatiche *POSAC* (*Partially Ordered Scalogram Analysis with Coordinates*) consentono di affrontare tale compito in modo più agevole.
Vedremo ora come sia possibile ottenere due punteggi di scala (valori di coordinate) per profili che sono strutturati da un meccanismo interno specifico.

3.5.1.2 Il modello action system

Abbiamo visto che:

- lo *scaling* "Guttman" implica il concetto di ordine tra gli item e tra le categorie di risposta (per esempio in termini di "difficoltà");
- lo *scaling* "diamante" è sostenuto dal concetto di ordine (per esempio in termini di tempo o in termini di passaggi all'interno di un procedimento).

E' possibile però osservare un altro tipo di *scaling* ordinale a partire da altri tipi di configurazioni, i cui profili manifestano altri tipi meccanismi o di logiche interne. Tra i diversi tipi ne è stato identificato uno in grado di riflettere alcune situazioni reali piuttosto complesse. Come vedremo, in questo caso, pur aumentando la complessità strutturale, non aumenta necessariamente la dimensionalità dell'insieme dei profili, che rimane bidimensionale. Tale diversa configurazione è stata scoperta e formulata da Shye (1985) nell'ambito del lavoro sull'*action system* (noto anche come *living system* o *open system*) ed è per questo chiamata "configurazione *action system*". La caratteristica essenziale comune a tutte le definizioni di *action system* è quella seconda la quale in un tale tipo di sistema è possibile "agire" ed "essere agiti" (per esempio si può "dare" e "ricevere").
Sapendo ciò possiamo definire i seguenti item:

- *G*: indica se in un certo momento un determinato sistema "dà" (1) o meno (0);
- *R*: indica se in un certo momento un determinato sistema "riceve" (1) o meno (0).

E' possibile a questo punto identificare quattro stati sistemici:

G	R	il sistema
1	1	dà e riceve
1	0	dà e non riceve
0	1	non dà e riceve
0	0	non dà e non riceve

Seguendo la teoria dei sistemi, è necessario introdurre due nuovi item:

- *X*: indica se in un certo momento un determinato sistema vi è interazione (1) o meno (0) con l'ambiente esterno;
- *I*: indica se in un certo momento per un determinato sistema vi è equilibrio interno (1) o meno (0).

Vediamo a questo punto quali sono le relazioni esistenti tra i quattro concetti sistemici (detti anche modi sistemici).

1ª condizione: l'*interazione* può verificarsi sia con il *dare* che con il *ricevere* ed è sicuramente assente se sono assenti *dare* e *ricevere*;

2ª condizione: l'*equilibrio interno* può essere interpretato come una specie di equilibrio tra *dare* e *ricevere* e può essere valutato come "presente" se *dare* e *ricevere* sono presenti.

Se *G*=0 e *R*=0 allora *X*=0. Se *G*=1 e *R*=1 allora *I*=1.

Date queste premesse, è possibile identificare i profili assenti (ovvero non osservabili):

GRXI	
0010	esclusi dalla 1ª condizione
0011	
1100	esclusi dalla 2ª condizione
1110	

Quindi dei 16 profili teoricamente possibili, si assume che solo 12 sono empiricamente osservabili:

GRXI
0000
0001
1000
1001
1010
1011
0100
0101
0110
0111
1101
1111

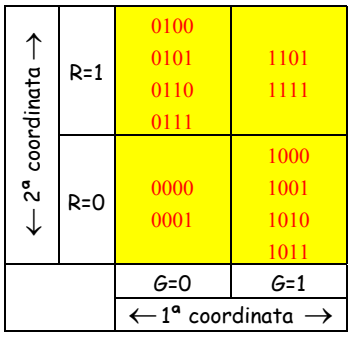
A questo punto l'obiettivo è quello di definire uno spazio bidimensionale che preservi l'ordine.

Cominciamo con il distinguere i profili con *G*=0 da quelli con *G*=1:

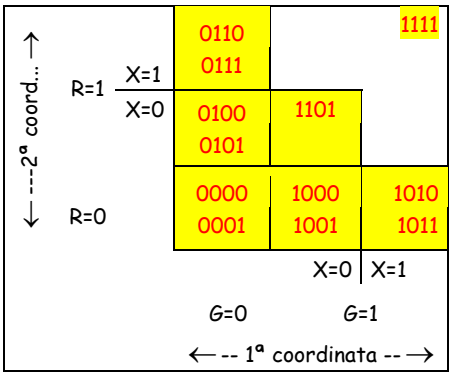
G=0	G=1
GRXI	GRXI
0000	1000
0001	1001
0100	1010
0101	1011
0110	1101
0111	1111

← 1ª coordinata →

A questo punto proviamo a distinguere i profili rispetto al valore *R* tracciando una retta orizzontale che divida ciascuna colonna in due.



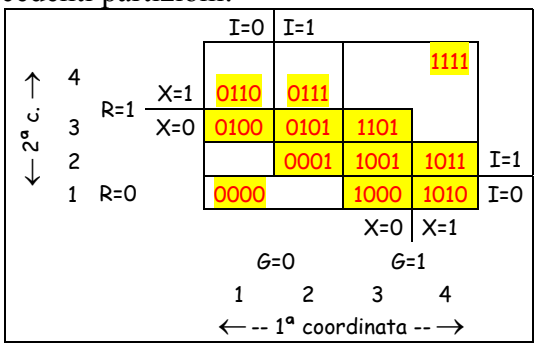
A questo punto in questo spazio bidimensionale è possibile separare i profili presenti secondo i valori di *X*. Per fare ciò è necessario individuare una linea di suddivisione appropriata che lasci tutti i profili con *X*=0 da una parte e i profili con *X*=1 dall'altra, senza alterare le classificazioni esistenti ottenute con i punteggi *G* e *R*.



Notare che la nuova linea lascia tutti i profili con *X*=0 al di sotto e a sinistra e i profili con *X*=1 al di sopra e a destra (verso i valori alti della due coordinate). La precedente suddivisione tra profili viene comunque mantenuta.

Dall'osservazione del diagramma risulta che *X*=1 è associato con valori estremi in almeno una coordinata. Per questo motivo un item che si comporta come *X* è detto *polarizzato* o *accentuato*.

A questo punto è possibile individuare una nuova linea che consenta di operare nello stesso modo ma con i valori *I*; tale linea dovrebbe lasciare tutti i profili con *I*=0 da una parte e quelli con *I*=1 dall'altra senza alterare le precedenti partizioni.



La nuova linea lascia al di sotto e a sinistra i profili con *I*=0 e al di sopra e a destra quelli con *I*=1. Notare che *I*=1 è associato con valori moderati in entrambe le coordinate. Per questa ragione un item che si comporta in questo modo è detto *moderante* o *attenuante*. Le coordinate sono ora ridefinite secondo 4 valori distinti (indicati con 1, 2, 3, 4). In questo modo a ciascun profilo vengono associati due nuovi punteggi, relativi alle due coordinate. Quindi a ciascun profilo vengono associati due punteggi. Un esame delle coppie di profili in tale diagramma consente di verificare se le relazioni d'ordine sono confermate. Per esempio il profilo 0111 (punteggio 24) è maggiore del profilo 0001 (punteggio 22) mentre il profilo 1010 (punteggio 41) e il profilo 0001 (punteggio 22) non sono tra loro confrontabili.

E' possibile a questo punto definire le regole strutturali per assegnare i valori delle coordinate (punteggi) a ciascun profilo della configurazione dell'*action system*.

1. Assegnazione del primo punteggio (prima coordinata) ad un profilo *GRXI*:
   - se G=0 e I=0 → assegnare al profilo il punteggio più basso (1)
   - se G=0 e I=1 → assegnare al profilo il punteggio successivo più basso (2)
   - se G=1 e X=0 → assegnare al profilo il punteggio successivo più alto (3)
   - se G=1 e X=1 → assegnare al profilo il punteggio più alto (4).
2. Assegnazione del secondo punteggio (seconda coordinata) ad un profilo *GRXI*:
   - se R=0 e I=0 → assegnare al profilo il punteggio più basso (1)
   - se R=0 e I=1 → assegnare al profilo il punteggio successivo più basso (2)
   - se R=1 e X=0 → assegnare al profilo il punteggio successivo più alto (3)
   - se R=1 e X=1 → assegnare al profilo il punteggio più alto (4).

Le configurazioni *action system* differiscono dalle altre per il preciso schema logico che governa le relazioni tra gli item in ogni profilo. Vi sono comunque molte caratteristiche essenziali che qualificano una configurazione come *action system*:

a. la configurazione presenta due item, detti *polari* che sono strutturalmente indipendenti tra loro ovvero: qualsiasi valore relativo ad uno, si può verificare con qualsiasi altro dell'altro item; i concetti basilari rappresentati dagli item polari sono di solito opposti e complementari nei loro significati (come *dare* e *ricevere*);
b. la configurazione presenta un item, detto *polarizzante* o *accentuante*, che è strutturalmente dipendente dagli item polari ovvero alcune combinazioni dei suoi valori con quelli degli item polari non sono osservabili nella configurazione; un tale tipo di item accentua le differenze delineate dagli item polari, nel determinare la configurazione; il concetto basilare rappresentato dall'item polarizzante varia con i contenuti del sistema studiato; il suo significato si caratterizza nella *incapacità di assemblare i poli*;
c. la configurazione presenta un item, detto *moderante* o *attenuante*, che è strutturalmente dipendente dai polari; il suo valore tende ad aumentare con l'aumento simultaneo in entrambi i polari; il suo ruolo nel determinare la struttura della configurazione è quello di attenuare o ridurre la differenziazione delineata dagli item polari; il concetto basilare rappresentato dall'item moderante è tipicamente distinto nel suo significato dai due polari; esso può "de-enfatizzare" il significato di qualsiasi polo concorrente relativo a quello dell'altro polo.

Tali caratterizzazioni della configurazione *action system* sono sufficientemente generali e consentono di definire molti modelli concreti che sono definibili in termini di relazioni logiche specifiche tra diversi item sistemici.
Nei dati empirici, le regole viste possono essere soddisfatte solo parzialmente o in modo approssimato. Inoltre la generalizzazione di configurazioni sistemiche a dimensionalità maggiori è possibile ma poco indagata.

Può essere utile a questo punto presentare uno schema riassuntivo dei tre modelli di scalogramma presentati:

Configurazione dei profili	Dimensionalità (n. di scale)	Esistenza di ordine tra gli item	Una possibile interpretazione	N. di classificazioni di item	Dimensionalità dello spazio degli item secondo LSA[1]
Guttman	1	si	difficoltà	1 (0) difficoltà[2]	0[3]
Diamante	2	si	processi finiti continui	1 posizione nel tempo	1
Action System	2	no	comportamento tipo action system	3 (2) polare, accentuante, attenuante[4]	2

1. Versione particolare della più generale procedura di analisi *Smallest Space Analysis* (*Lattice Space Analysis*).
2. L'ordine in una scala Guttman è in realtà determinato dall'ordine delle categorie rispetto alla loro relativa difficoltà; quindi, in presenza di dicotomie, gli item non sono differenziati dalle classificazioni; si può dire che il numero di classificazioni è nullo.
3. Per definizione.
4. Gli item accentuanti e attenuanti possono essere formulati come elementi della stessa classificazione, riducendo il numero di classificazioni a 2.

3.5.2 L'analisi di scalogrammi multidimensionali

Tale approccio (*Multidimensional Scalogram Analysis*, *MSA*) cerca di rappresentare profili come punti in uno spazio multidimensionale in modo tale che lo spazio possa essere suddiviso in regioni che contengano solo punti (profili) con la stessa struttura. In generale ciò è possibile solo se i dati presentano una particolare struttura.
Con *k* item osservati su *n* casi si registreranno *n* profili, ciascuno dei quali può essere considerato come una cella di una tabella di contingenza a "*k* vie" oppure come un punto in uno spazio a *k* dimensioni. I valori che compongono un profilo possono essere così considerati come coordinate in tale spazio. Se per ciascun item l'ordine delle categorie è arbitrario, la rappresentazione spaziale non è unica.
L'obbiettivo di tale approccio è quindi quello di verificare se è possibile definire una rappresentazione equivalente in uno spazio a bassa dimensionalità, soggetto alla restrizione chiamata *regionalità*; ciascun profilo apparirà come una cella o un punto in uno spazio a bassa dimensionalità in modo tale che tutti i profili che presentano la stessa struttura risulteranno essere contigui tra loro.
La rappresentazione dei profili come punti in uno spazio geometrico avviene in modo tale che

a. lo spazio possa essere suddiviso da ciascun item del profilo,
b. la soluzione sia ottenuta in uno spazio a bassa dimensionalità.

Esistono molti algoritmi per l'analisi di scalogrammi multidimensionali che si differenzianoscono tra loro per il livello di scala assegnato agli item e per la forma delle linee di partizione dello spazio:

- *MSA-I*: rappresenta l'algoritmo più generale; assume item nominali e lascia indefinita la forma delle linee di partizione;
- *MSA-II*: assume che tutti gli item hanno lo stesso *range* e che le linee di partizione siano circolari;
- *MSA-III*: assume linee di partizione, item per item, rette tra loro parallele.

Il primo algoritmo è basato sul concetto di contiguità e sulla definizione di *outer-point* e *inner-point*. Data una particolare distribuzione dei punti, osserviamo un item (X) ed un suo particolare valore individuale (x_i); è possibile determinare il limite, il confine della regione dei punti che appartengono a x_i attraverso l'individuazione degli *outer-point*. Tra tutti i punti che non appartengono a x_i ne esiste uno che risulta essere il più vicino a x_i; tale punto è detto *outer* rispetto a x_i. Tutti gli altri punti che non sono *outer* rispetto a x_i sono *inner*. L'insieme di tutti i punti appartenenti allo spazio di x_i si dice che occupano una regione contigua se ciascun *inner-point* di x_i è più vicino ad alcuni *outer-point* di x_i di quanto sia un *outer-point* di ciascun altro elemento di X.
Le deviazioni dalla perfetta contiguità sono prodotte da *inner-point* che sono più vicini ad alcuni *outer-point* di un'altra regione di un *outer-point* della propria regione. Per una data soluzione di *MSA-I* il *coefficiente di contiguità*, che può assumere valori da -1 (minima contiguità) a +1 (massima contiguità), tiene conto del numero dei punti devianti e della dimensione delle deviazioni calcolate per tutti gli item. La procedura è iterativa ed utilizza un algoritmo *steepest ascent*[9].
Nella pratica, gli spazi individuati attraverso questo tipo di analisi non sono facili da esplorare e interpretare, soprattutto nei casi in cui la dimensionalità è maggiore di 2 e quando la soluzione non risulta essere pienamente soddisfacente (coefficiente di contiguità basso). La metodologia si presenta comunque utile soprattutto nei casi in cui si ha l'obiettivo di suddividere gli item secondo diverse tipologie o mettere in rilievo particolari legami esistenti tra le diverse categorie.

[9] A tale proposito si veda il *MultiDimensional Scaling*.

3.5.3 Analisi di uno scalogramma parzialmente ordinato

Alla base dell'approccio *POSA* vi è l'assunto che secondo l'universo di contenuto indagato è rappresentato da tutti gli item prescelti, ciascuno dei quali presenta un *range* di possibili punteggi ordinati rispetto a tale universo di contenuto. L'analisi effettuata secondo tale approccio riguarda *profili*, ciascuno dei quali è definito come *insieme di soggetti con punteggi identici in tutti gli item*. Tale analisi consente una rappresentazione spaziale senza il bisogno di assunti e di statistiche aggiuntive; tale rappresentazione spaziale facilita l'interpretazione delle direzioni nello spazio; il significato generale del tratto viene comunque associato ad una direzione specificata in precedenza dello spazio geometrico derivato. Al termine dell'analisi sarà possibile assegnare a ciascuno profilo un numero di punteggi minore del numero di variabili originarie, in modo tale che una volta nota la configurazione spaziale, i profili originali possono essere riprodotti a partire dalle coordinate nello spazio geometrico.
Il *POSA* può essere considerato una estensione dello *scaling* tipo Guttman, caratterizzato come abbiamo visto dall'unidimensionalità e da un'unica soluzione di scalabilità tra i profili empirici. Occorre però dire che, dal punto di vista matematico, il passaggio dal caso speciale a quello generale non è risultato né semplice né ovvio; ciò non ha sicuramente favorito lo sviluppo di procedure analitiche d'analisi. Tra gli approcci analitici è possibile identificarne principalmente due, uno dei quali è stato definito da Coombs.

Approccio Coombs

Nel tentativo di proporre un procedimento di analisi di modelli di scalogramma parzialmente ordinali, Coombs, a metà degli anni '60, giunge a descriverne tre basati su un'analisi non-metrica che consente di produrre una configurazione multi-dimensionale. La selezione del modello più appropriato dipende dagli assunti adottati:

1. *Modello congiuntivo*, secondo il quale un individuo deve "superare" tutti gli item perché si possa affermare che possiede la caratteristica misurata. Per l'analisi si suggerisce di costruire molte scale. Sia gli item più difficili[10] che i profili con il maggior numero di errori/fallimenti individuano dimensioni distinte, ognuna delle quali definisce una scala. Con item dicotomici, ogni scala così ottenuta, definisce un ordine tra tali elementi. Tutte insieme le scale risultano in un ordine parziale tra gli item.
2. *Modello disgiuntivo*, secondo il quale un individuo deve "superare" almeno uno degli item perché si possa affermare che possiede la caratteristica misurata. Il tipo di analisi è molto simile al precedente, si invertono i ruoli dei valori "0" (fallimento) e "1" (superamento). L'analisi descritta può produrre due diverse dimensionalità per la stessa matrice dei dati a seconda del modello adottato.
3. *Modello compensatorio*, secondo il quale un individuo deve "superare" molti item perché si possa affermare che possiede la caratteristica misurata; una mancanza in determinati item può essere compensata dalla previsione in altri. La procedura può essere applicata solo a configurazioni bidimensionali. Essa prevede che si determini l'ordine tra gli item che compongono una tripletta, successivamente la combinazione di tali ordini consente di ottenere due ordini generali tra tutti gli item. Ciò limita l'applicazione ad una sottoclasse molto speciale di tali configurazioni. Tale sottoclasse è, inoltre, caratterizzata dal fatto che l'ordine degli item in una dimensione deve essere contrario di quello nell'altra dimensione.

Approccio Guttman

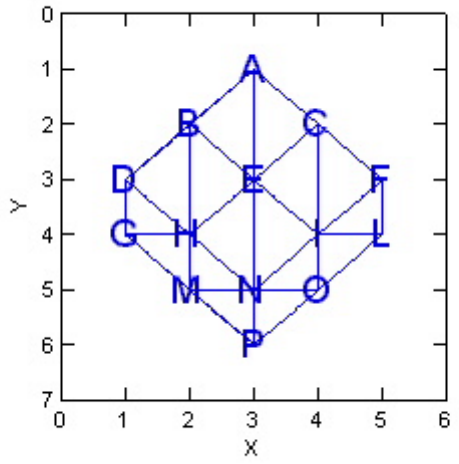
Questo approccio consente la rappresentazione dei profili e delle loro reciproche relazioni d'ordine parziale, attraverso un *diagramma Hasse*. In questo tipo di diagramma i profili vengono presentati come punti ed ogni coppia confrontabile di profili viene collegata da un segmento. Un profilo maggiore di un altro è posizionato nella parte alta del diagramma, rispetto ad una direzione pre-determinata. Con item dicotomici, l'insieme dei possibili profili è rappresentato in modo tale che le proiezioni perpendicolari sugli assi diagonali, che collegano il vertice "00..0" con il vertice "11..1",

[10] Si assume che l'item più difficile sia quello che all'interno di un profilo risulta essere l'unico a non essere superato.

riflettano le relazioni d'ordine tra i profili confrontabili. Nei casi in cui, a causa dell'interdipendenza tra item, si ottengono solamente alcuni dei possibili profili, la costruzione del *diagramma Hasse* risulta particolarmente semplificata. Tale diagramma utilizza due dimensioni, una detta *joint* (congiunta) e l'altra *lateral* (laterale). Vedremo come in tale rappresentazione i profili che presentano lo stesso punteggio totale presentano la stessa posizione rispetto alla dimensione (o direzione) congiunta ovvero la stessa latitudine; i profili confrontabili ma con punteggio totale diverso presentano posizioni diverse rispetto alla direzione congiunta ma la stessa longitudine.
Di seguito osserviamo la configurazione ottenuta dall'analisi dei profili relativi alla partecipazione alle attività culturali di associazioni sindacali (esempio tratto da Shye, 1985); attraverso tale analisi la sequenza degli item è risultata la seguente:

1°: biblioteca	2=si 1=no
2°: dibattiti	
3°: attività sportive	
4°: cine-forum	
5°: escursioni	

La sequenza determina il contenuto della direzione laterale della configurazione (x); tale direzione va da *preferenza per programmi educativi* (valori bassi di x) a *preferenza per programmi di intrattenimento* (valori alti di x):



PROFILI	SIMBOLI	X	Y
11111	A	3.000	1.000
21111	B	2.000	2.000
11121	C	4.000	2.000
22111	D	1.000	3.000
21121	E	3.000	3.000
11122	F	5.000	3.000
22211	G	1.000	4.000
22121	H	2.000	4.000
21122	I	4.000	4.000
11222	L	5.000	4.000
22221	M	2.000	5.000
22122	N	3.000	5.000
21222	O	4.000	5.000
22222	P	3.000	6.000

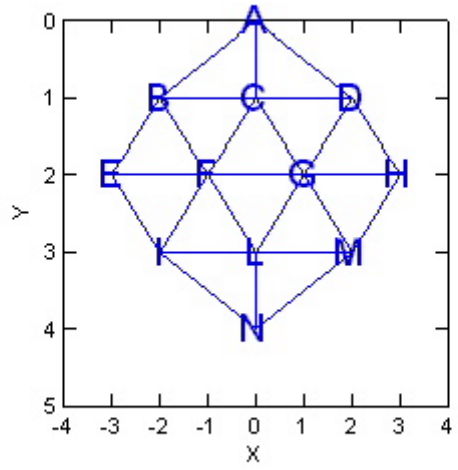
x: direzione laterale $\Rightarrow$ tipo di partecipazione

y: direzione congiunta $\Rightarrow$ estensione della partecipazione totale

Di seguito vediamo i risultati ottenuti in una indagine sulle preoccupazioni rispetto alla perdita del lavoro in seguito alla computerizzazione in un gruppo di lavoratori (esempio tratto da Shye, 1985). In seguito all'analisi la sequenza degli item è risultata essere la seguente:

1°: lavoro	1=si 0=no
2°: stabilità	
3°: prospettive	
4°: interesse	

Gli item 1 e 4 definiscono la polarità del contenuto della direzione laterale: punteggi alti nell'item 1 sono associati con la parte sinistra del diagramma mentre i punteggi alti nell'item 4 sono associati con la parte destra. Gli altri due item, 2 e 3, risultano avere la funzione rispettivamente di polarizzatore e moderatore (v.) dell'asse laterale: punteggi alti per l'item 2 sono generalmente associati ai punteggi laterali estremi (molto alti o molto bassi) mentre i punteggi alti nell'item 3 sono associati con punteggi laterali mediani.
La sequenza determina il contenuto della direzione laterale della configurazione (*x*); tale direzione va da *enfasi sulla perdita di compensi estrinseci* (valori negativi di *x*) a *enfasi sulla perdita di compensi intrinseci* (valori positivi di *x*):



PROFILI	SIMBOLI	X	Y
0000	A	0.000	0.000
1000	B	-2.000	1.000
0010	C	0.000	1.000
0001	D	2.000	1.000
1100	E	-3.000	2.000
1010	F	-1.000	2.000
0011	G	1.000	2.000
0101	H	3.000	2.000
1110	I	-2.000	3.000
1011	L	0.000	3.000
0111	M	2.000	3.000
1111	N	0.000	4.000

x: direzione laterale ⇒
y: direzione congiunta ⇒

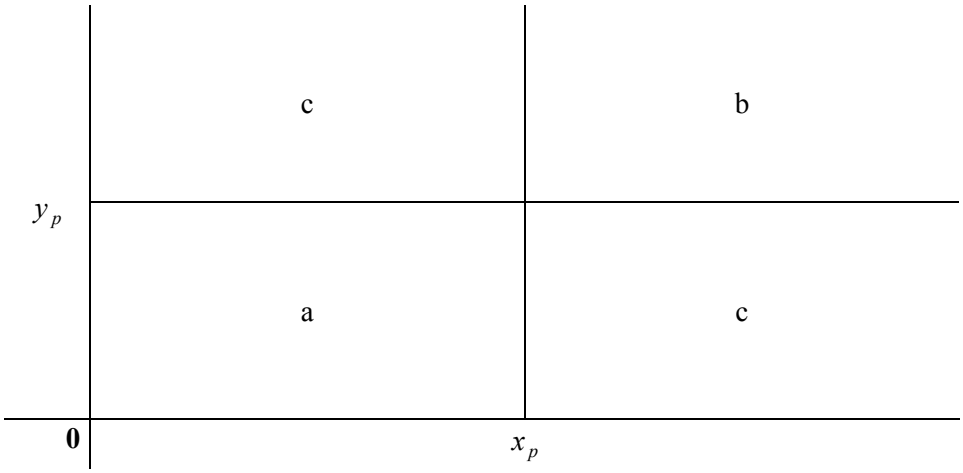
Entrambi gli esempi dimostrano gli sforzi effettuati per attribuire significati sostanziali alla seconda direzione nello spazio dei profili (*direzione laterale*). Il significato della *direzione congiunta* è

invece determinato dal significato generale del tratto misurato, comune a tutti gli item.
Come risulta chiaro a questo punto, in tale configurazione bidimensionale ciascun profilo individuale (ovvero l'insieme dei punteggi originali degli item per ciascun caso) può essere riprodotto a partire da due punteggi, quello relativo alla direzione laterale e quello relativo alla direzione congiunta. In particolare, nel secondo esempio, un caso che registra un punteggio laterale di -1 e un punteggio congiunto di 2 (ovvero ha un livello intermedio di preoccupazione con una leggera enfasi sulle preoccupazioni riguardanti la perdita di ricompense esterne) ha un profilo di 1010 ovvero egli è preoccupato della perdita del lavoro e della perdita di prospettive ma non della stabilità e dell'interesse.
Per grandi matrici di dati, tale procedura può essere ardua da applicare. Inoltre l'assegnazione del significato alla direzione laterale diviene complessa se non impossibile. E' stata però definita una procedura analitica detta *POSAC*, (*Partially Ordered Scalogram Analysis with Coordinates*) che consente di calcolare e presentare graficamente una configurazione bidimensionale dello scalogrammi e di interpretare la direzione laterale di tale scalogramma.

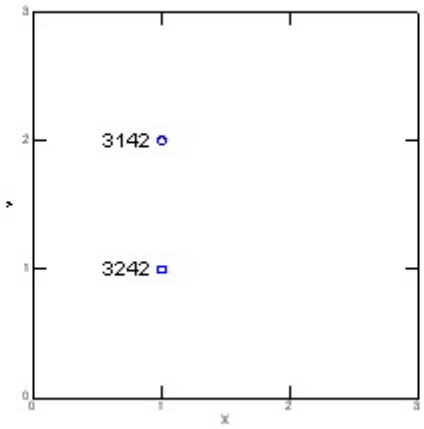
3.5.3.1 Il POSAC (Partially Ordered Scalogram Analysis with Coordinates)

Il *POSAC* rappresenta essenzialmente una procedura tecnica che consente di determinare l'adattamento dei profili osservati in uno spazio bidimensionale in modo tale da preservare la condizione di ordine ipotizzata. Il successo dell'applicazione di tale approccio ai dati empirici dipende dalla qualità del disegno sperimentale (lo schema concettuale, la scelta delle variabili, la popolazione osservate, ecc.).
L'obiettivo del *POSAC* è quello di, dato un insieme *A'* di profili osservati con *n* item, identificare il numero minimo di punteggi che consentano di collocare i profili osservati in modo tale che conservino le relazioni osservate di ordine e di confrontabilità, in altre parole, di verificare se è possibile assegnare due punteggi (corrispondenti a un punto in un piano cartesiano) a ciascun profilo di *A'*, in modo tale che per qualsiasi coppia di profili, la loro relazioni possa essere rappresentata in modo corretto per mezzo dei loro corrispondenti profili di coordinate.
Si ottiene una rappresentazione perfetta quando i punteggi individuati descrivono perfettamente l'ordine e la confrontabilità dei profili originari.
Nel caso in cui non sia possibile ottenere per un certo scalogramma una perfetta rappresentazione in uno spazio bidimensionale l'approccio *POSAC* ricerca una soluzione ottimale attraverso l'osservazione dei valori ottenuti con il *coefficiente di corretta rappresentazione* (*CORREP*), che specifica la proporzione di coppie di profili, pesati attraverso le loro frequenze osservate, la cui relazione di confrontabilità sia perfettamente rappresentata. Il valore del coefficiente di corretta rappresentazione va da 0 a 1 (soluzione perfetta).
Risulta chiaro a questo punto come l'insieme delle coppie di punteggi per ciascun profilo può essere pensato come uno spazio bidimensionale a coordinate cartesiane in cui le coordinate *X* e *Y* rappresentano rispettivamente il primo e il secondo punteggio. Le due coordinate per ciascun punto nello spazio *XY* possono essere considerate un profilo che consente ai punti del piano di formare un insieme parzialmente ordinato. Per ciascun punto nello spazio, è possibile individuare nello spazio tre diverse regioni:

a. regione dei punti che sono minori di un dato punto *P*;
b. regione dei punti che sono maggiori di un dato punto *P*;
c. regione (composta da due diverse sezioni) dei punti che non sono confrontabili con il punto *P* (al centro della configurazione).

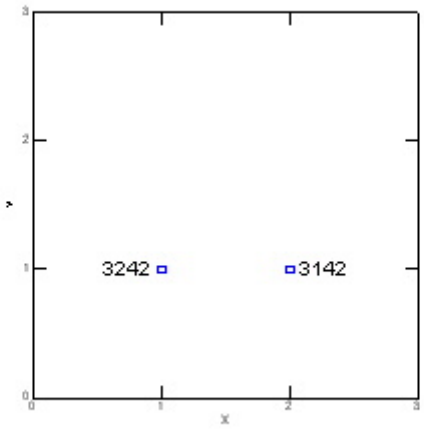


Poniamo di avere tre profili: "3142", "3242" e "1118": all'interno di un diagramma cartesiano esistono tre possibili diverse collocazioni, e tre possibili punteggi, per poter rappresentare le posizioni reciproche dei primi due profili in modo tale che conservino la loro reciproca relazione ("3242" > "3142"):

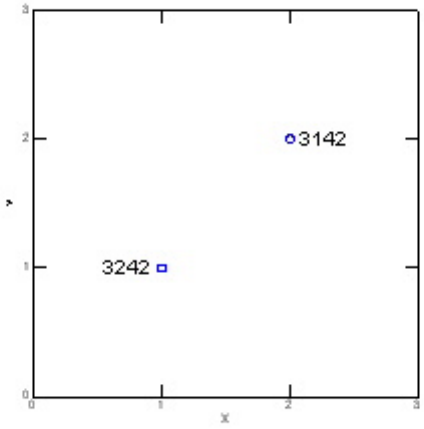


Punteggi

- 3142 → 11
- 3242 → 12



Punteggi
- 3142 → 11
- 3242 → 21



Punteggi
- 3142 → 11
- 3242 → 22

Per poter rappresentare nello stesso diagramma anche il terzo profilo occorre individuare la regione dei punti non confrontabili con le coordinate 1,1 e 2,2 (ovvero all'interno dell'intersezione delle regioni dei punti non confrontabili a 11 e dei punti non confrontabili a 22) inserendo per esempio la coordinata 3; in questo modo il profilo "1118" può assumere la posizione, e il punteggio, 03. Naturalmente non è possibile individuare una nuova posizione nel piano per qualsiasi nuovo profilo in modo tale che vengano preservate le sue relazioni con tutti gli altri; ciò vuol dire che non esiste sempre la possibilità di rappresentare correttamente in uno spazio bidimensionale tutte le relazioni d'ordine esistenti tra i profili di un certo gruppo.

***POSAC*: Bontà di adattamento**

Dato un insieme di profili, è necessario definire un criterio di bontà di adattamento che consenta di

stabilire qual è il tipo di collocazione bidimensionale che meglio descrive le relazioni d'ordine tra loro osservate. Un primo indice che consente tale valutazione è rappresentato da un coefficiente basato sulla proporzione di profili rappresentati in modo corretto; tale coefficiente tiene conto anche della frequenza registrata da ciascun profilo. L'algoritmo è basato sulla minimizzazione di una funzione proposta da Louis Guttman negli anni '80. La soluzione *POSAC* rappresenta una approssimazione allo spazio minimo definito. Tale algoritmo è molto sensibile alla *approssimazione iniziale* utilizzata; questa può essere basata su un'ipotesi riguardante, per esempio, la polarità tra item; in altri casi il procedimento per definire l'approssimazione iniziale prevede che vengano eseguiti in successione i seguenti momenti:

a. calcolo della matrice dei coefficienti di debole monotonicità (*weak monotonicity coefficients*, *wm*)[11];
b. identificazione dei due item (i_0 e j_0) che presentano la minore correlazione positiva (item estremi);
c. determinazione della posizione di ciascun profilo $a = a_{i_0} \ldots a_{j_0} \ldots a_n$:
   - calcolo del punteggio del profilo (somma dei punteggi registrati dagli item),
   - *livellamento* rispetto alla somma $x_a + y_a$ delle sue coordinate;
   - valutazione della sua prossimità relativa ad uno o all'altro degli item estremi attraverso $a_{i_0} - a_{j_0}$ che equalizzata alla differenza tra le coordinate $x_a - y_a$.

In questo modo è possibile ottenere i valori di x_a e y_a per tutti i profili a.

Riprendendo i dati di un esempio presentato da Shye (1985) e relativi a quattro item, osserviamo in pratica il procedimento:

identificazione profilo	profilo	punteggio	item$_1$ - item$_4$	coordinata x dell'approssimazione iniziale	coordinata y dell'approssimazione iniziale
1	2222	8	0	4	4
2	2221	7	1	4	3
3	2212	7	0	3.5	3.5
4	1222	7	-1	3	4
5	2121	6	1	3.5	2.5
6	2211	6	1	3.5	2.5
7	1212	6	-1	2.5	3.5
8	1122	6	-1	2.5	3.5
9	2111	5	1	2.5	2
10	1211	5	0	2.5	2.5
11	1112	5	-1	2	3
12	1111	4	0	2	2

I coefficienti *wm* per i quattro item di tale configurazione, assumendo frequenze uguali per tutti i profili, sono date nella seguente matrice.

[11] Il coefficiente di monotonicità tra due item con range ordinati, *A* e *B* è:

$$\textit{coefficiente di monotonicità} = \frac{\sum_q \sum_p (a_p - a_q)(b_p - b_q)}{\sum_q \sum_p |a_p - a_q||b_p - b_q|} \qquad q, p = 1,\ldots,N$$

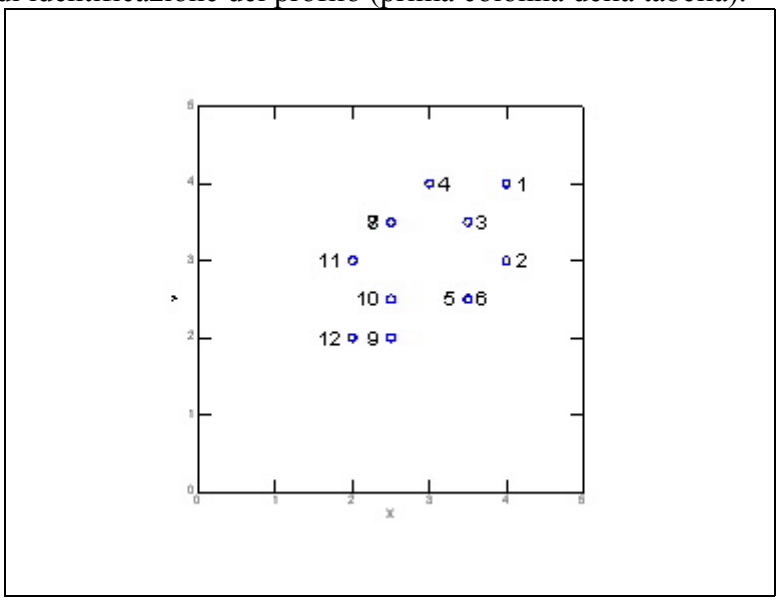
dove

a_p punteggio del caso *p* per l'item *a*
a_q punteggio del caso *q* per l'item *a*
b_p punteggio del caso *p* per l'item *b*
b_q punteggio del caso *q* per l'item *b*

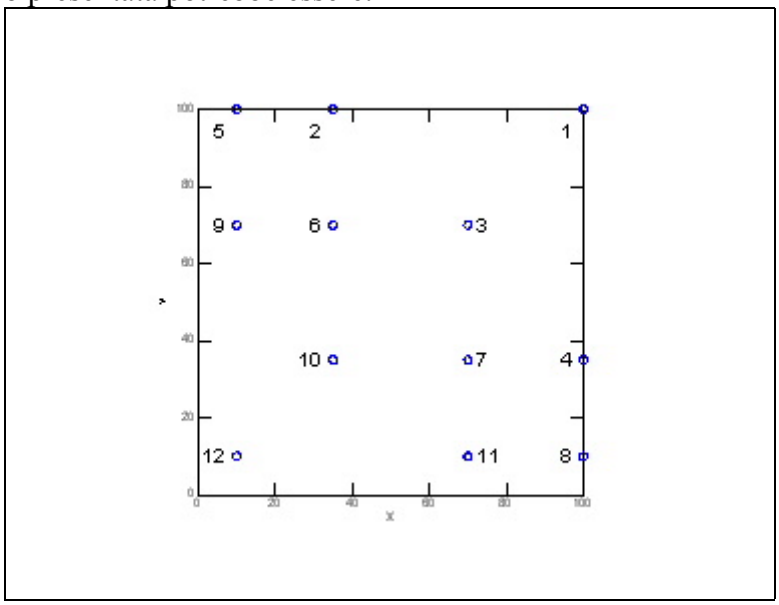
I valori di questo coefficiente vanno da +1 a -1 e indicano quanto un aumento nei punteggi di un item è accompagnato da un aumento (a da una diminuzione) nei punteggi dell'altro item.

	I_1	I_2	I_3	I_4
I_1	1			
I_2	1/3	1		
I_3	1/3	1/17	1	
I_4	-3/5	1/3	1/3	1

In tale matrice osserviamo come gli item 1 e 4 sono quelli con la più bassa correlazione positiva. Nella terza e quarta colonna della tabella vengono presentati i due parametri iniziali per la configurazione dei profili. Quando questi due parametri iniziali sono equalizzati rispettivamente a $x_a + y_a$ e $x_a - y_a$, è possibile individuare i valori di x_a e y_a per ogni profilo. Tali coordinate sono presentate nelle colonne 5 e 6 della tabella e proiettate nel seguente diagramma in cui ciascun profilo dal numero di identificazione del profilo (prima colonna della tabella).



L'approssimazione iniziale viene migliorata attraverso ripetute iterazioni, effettuate in due fasi, nel tentativo di trovare una approssimazione soddisfacente. L'output della soluzione finale presenta per ciascun profilo i valori *X*, *Y*, *J=X+Y*, *L=100+X-Y* e quindi la configurazione bidimensionale che, nel caso dell'esempio presentata potrebbe essere:



Uno dei pochi *package* che presentano al loro interno la procedura *POSAC* è il *Systat* il cui procedimento pratico, in sintesi, prevede i seguenti passaggi:

a. ordinamento degli item da sinistra a destra in modo tale che la dimensione orizzontale mostri i valori 1 che si spostano da sinistra a destra all'interno dei profili;

b. ordinamento dei profili in senso verticale rispetto al punteggio totale;
c. ordinamento dei profili da sinistra a destra;
d. individuazione di profili che non si adattano al modello e verifica che l'intera soluzione che venga influenzata da essi.

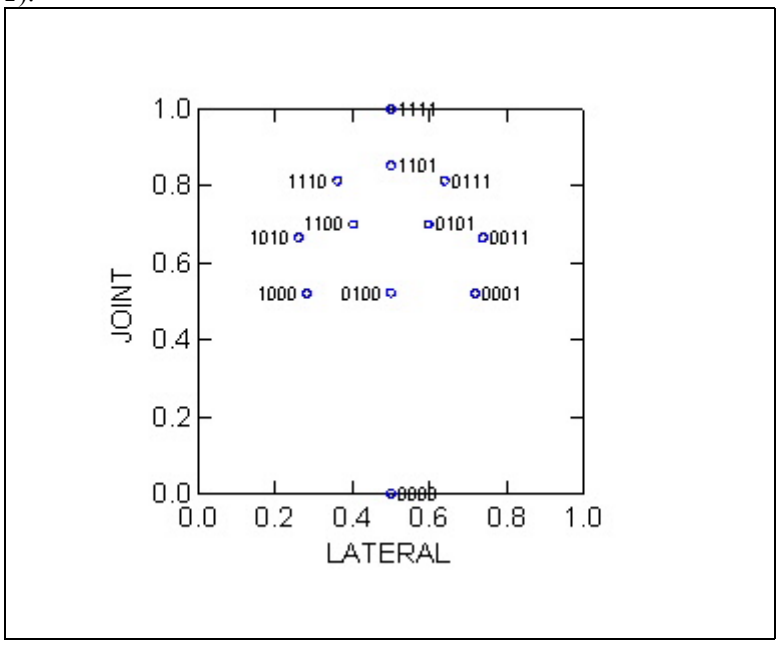
In un certo senso l'ultimo momento rappresenta un requisito ambiguo che comunque dipende dal primo momento; se per esempio avessimo i due profili "1010" e "0101", scambiando il secondo con il terzo item, avremmo i profili "0011" e "1100", corrispondenti a quelli estremi, che ci consentirebbero di ottenere una soluzione ben adattata.

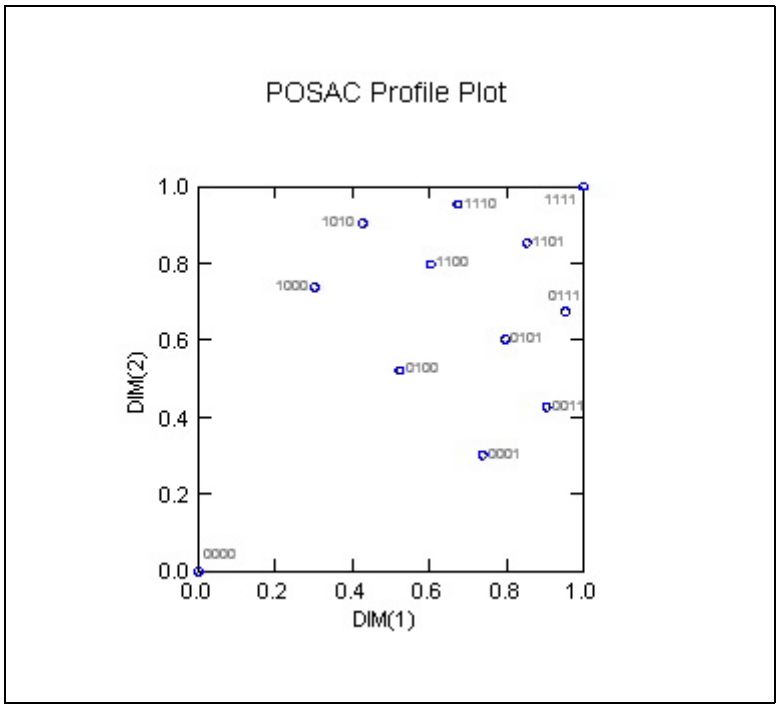
Per realizzare il primo passaggio il procedimento *POSAC* prevede la definizione della matrice di monotonicità attraverso il coefficiente di monotonicità e quindi l'ordinamento della matrice ottenuta per mezzo di un algoritmo di *multidimensional scaling*. Il procedimento è iterativo e produce al termine l'ordine degli item per la definizione del profilo, le coordinate per ciascun profilo e un valore che consente di valutare la rappresentazione ottenuta in termini di perdita (*final loss value*); minore è tale valore, migliore è la rappresentazione. Prima di predisporre la rappresentazione grafica, la procedura programmata all'interno del *package Systat* calcola la radice quadrata delle coordinate per rendere la direzione laterale lineare piuttosto che curvilinea; in questo modo il grafico risulta ruotato di 45°.

Riprendendo i dati dell'ultimo esempio si ottengono per ciascun profilo le coordinate per la direzione *joint* e la direzione *lateral* e le corrispondenti nuove coordinate (rispettivamente Dim_1 e Dim_2):

Profili	Dim 1	Dim 2	*Joint*	*Lateral*
1111	1.000	1.000	1.000	0.500
1101	0.853	0.853	0.853	0.500
1110	0.674	0.953	0.814	0.360
0111	0.953	0.674	0.814	0.640
1100	0.603	0.798	0.700	0.403
1010	0.426	0.905	0.665	0.261
0011	0.905	0.426	0.665	0.739
0101	0.798	0.603	0.700	0.597
1000	0.302	0.739	0.520	0.281
0001	0.739	0.302	0.520	0.719
0100	0.522	0.522	0.522	0.500
0000	0.000	0.000	0.000	0.500

Di seguito vediamo la rappresentazione spaziale con le coordinate originali (*lateral*joint*) e con le coordinate trasformate (*dim 1*dim 2*):





I punteggi finali ottenuti con la procedura *POSAC* possono essere interpretati osservando le relazioni tra i contenuti degli item e il ruolo che giocano nello strutturare lo spazio *POSAC*.

4. I MODELLI CUMULATIVI. L'APPROCCIO PROBABILISTICO

I modelli cumulativi-probabilistici si richiamano all'*Item Response Theory* (*IRT*), che a sua volta si rifà alla teoria del tratto latente, formalmente attribuisce la variazione nelle risposte a parametri riguardanti sia gli item che i casi; in questo senso le risposte individuali costituiscono un indicatore della relazione tra item e soggetto, in quanto ciascuna di esse è frutto della posizione rispetto all'attributo sia dell'item che del soggetto. Quindi la posizione di accettazione (superamento, accordo) o di rifiuto (fallimento, disaccordo) del soggetto rispetto ad un particolare item dipende sia dalle sue caratteristiche (capacità, atteggiamenti, attitudini, opinioni, ecc.) che da quelle riflesse dall'item (difficoltà, presentazione di un'opinione, ecc.).
In altre parole, secondo la *teoria del tratto latente*, la 'performance' di una data misura è funzione della posizione del caso lungo il continuum del tratto misurato e dell'errore casuale. L'obiettivo è quello di stimare tale funzione, assumendo l'indipendenza locale (la 'performance' di un caso rispetto ad un item è statisticamente indipendente ed è spiegabile solo in funzione delle caratteristiche individuali). La relazione è descritta dalla curva caratteristica dell'item (*Item Characteristic Curve*, *ICC*).
La definizione di modelli matematico-probabilistici ha consentito la definizione di criteri statistici per la valutazione della bontà di adattamento del modello ai dati e l'individuazione di procedure soddisfacenti di accettazione o rifiuto dell'ipotesi di misurazione.
I modelli di *Item Response Theory* si distinguono dagli altri in quanto non sono *test-dependent*, ovvero sono espressi a livello di item (*item oriented*) e non a livello di gruppo di item (*test-oriented*) e producono misurazioni precise (quantitative) sia dell'item (in genere in termini di difficoltà) che dei casi (in genere in termini di capacità) sulla stessa scala,
Inoltre secondo tali modelli le caratteristiche di ciascun item e di ciascun soggetto non cambiano in funzione dei campioni per la sperimentazione, ovvero

- gli item selezionati *non* sono *group-dependent*,
- i punteggi individuali *non* sono *test-dependent*.

Ciò consente

- la validazione sulla base di una sola applicazione, senza il bisogno di definire strumenti paralleli in senso stretto,
- la costruzione di banche di item validati su gruppi diversi (economicità),
- la scelta di item che coprano l'intera estensione del continuum (ottimizzazione della selezione degli item).

I metodi di analisi degli item inoltre consentono di ottenere precise informazioni sulle modalità interne di risposta utilizzabili con funzioni *diagnostiche*.

4.1 GLI ASSUNTI

Gli assunti che definiscono tali modelli sono l'*unidimensionalità*, l'*indipendenza locale* e la funzione che lega risposta e item.

- UNIDIMENSIONALITA'
  Si assume che la risposta di un soggetto ad un item sia determinata e possa essere spiegata da una sola componente, da un solo fattore unico dominante chiamato *tratto latente*; gli item

individuati devono misurare solo tale caratteristica. In realtà l'assunto (comune ad altri modelli) è difficile da soddisfare in modo rigoroso.

A tale proposito si pensi come la misurazione di certi attributi sia difficile

- perché non necessariamente immodificabili per l'intervento di componenti (apprendimento, esperienza, memoria, ecc.) che possono modificare nel tempo il tratto misurato,
- per l'influenza che in misura variabile e non nota possono esercitare altri fattori (cognitivi, di personalità, di motivazione, di ansietà, di capacità di lavorare velocemente, di tendenza a rispondere in maniera casuale nel caso di risposte dubbiose, ecc.).

Nell'ambito dell'*IRT*, comunque, sono stati definiti anche modelli *multidimensionali* che possono essere adottati nei casi in cui si ipotizzano risposte spiegabili da più di una dimensione.

- INDIPENDENZA LOCALE

Secondo l'assunto di <u>indipendenza locale</u>, tenuto costante il valore del tratto latente che influenza la risposta, non esiste alcuna relazione tra le risposte di ciascun soggetto ai diversi item; in altre parole le risposte di un soggetto agli item sono statisticamente indipendenti e sono spiegabili solo in funzione delle caratteristiche individuali. La dimensione misurata costituisce lo *spazio completo latente* (*complete latent space*). Quando l'assunto di unidimensionalità è soddisfatto, tale spazio riguarda una sola capacità. La proprietà d'indipendenza locale può essere matematicamente formalizzata nel modo seguente:

$$P(U_1,U_2,...,U_i,...,U_n|d) = P(U_1|d)P(U_2|d)...P(U_i|d)...P(U_n|d) = \prod_{i=1}^{n} P(U_i|d)$$

dove

n numero totale di item

d capacità che influenza la risposta del soggetto ad uno strumento

U_i risposta del soggetto all'item i (i=1, 2, ..., n)

$P(U_i|d)$ probabilità di risposta di un soggetto con capacità d

con

$P(U_i = 1|d)$ probabilità di risposta corretta

$P(U_i = 0|d)$ probabilità di risposta scorretta

ovvero la probabilità di risposta di un soggetto ad un gruppo di item è uguale al prodotto delle probabilità associate alle risposte del soggetto a ciascuno degli item[12].

Con tale definizione, l'assunto di indipendenza locale sembra complicato da soddisfare: è difficile pensare che le risposte di un soggetto a molti item non siano tra loro correlate ovvero che siano tra loro indipendenti; in realtà l'unico elemento che deve legare tali risposte è il valore del tratto latente: tenendo costante tale valore (analisi di correlazione parziale), le risposte devono risultare non correlate. Quindi se la relazione tra le diverse risposte di un soggetto ad un gruppo di item è spiegata da un unico aspetto (caratteristica misurata), rendendo costante tale aspetto, le risposte divengono indipendenti[13]. Per questa ragione l'assunto di indipendenza locale viene detto anche <u>assunto di indipendenza *condizionale*</u>.

Gli assunti di *unidimensionalità* e di *indipendenza locale* sono molto collegati tra loro, infatti se il primo risulta valido, lo sarà conseguentemente anche il secondo: identificato uno spazio latente completo, che influenza le risposte, è soddisfatto l'assunto di indipendenza locale.

Vediamo un esempio. Poniamo di avere un item che misura la capacità matematica ma che richiede anche un alto livello di padronanza linguistica; possiamo trovarci davanti a due situazioni:

- i soggetti presentano tra loro diversi livelli di competenza linguistica; in questo caso si può ipotizzare che i

[12] Per esempio, se il modello di risposta a tre item di un soggetto è *1,1,0* (ovvero $U_1=1$, $U_2=1$ e $U_3=0$) allora l'assunto di indipendenza locale implica che:

$$P(U_1 = 1, U_2 = 1, U_3 = 0|d) = P(U_1 = 1|d)P(U_2 = 1|d)P(U_3 = 0|d) = P_1P_2Q_3$$

[13] Si tratta di un principio che si ritrova anche in altri modelli come quello fattoriale.

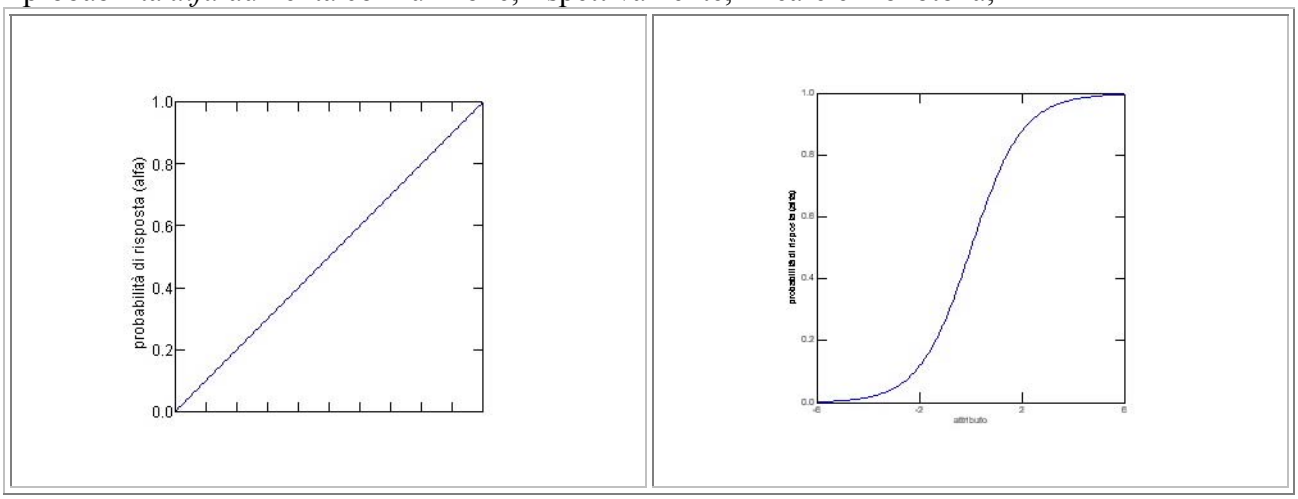
soggetti con bassa capacità linguistica non rispondano correttamente all'item indipendentemente dalla loro competenza matematica; quindi una dimensione estranea alla capacità matematica influenza la risposta all'item; conseguentemente l'assunto di indipendenza locale non può venire soddisfatto;
- i soggetti presentano tra loro gli stessi livelli di competenza linguistica; in questo caso si può ipotizzare che le risposte ottenute possano essere attribuite solo alla capacità matematica: l'assunto di indipendenza locale è soddisfatto.

Altri casi in cui non è possibile sostenere l'indipendenza locale sono quelli in cui, per esempio, l'item contiene già elementi o informazioni tali da influenzare la risposta all'item in questione o ad altri item. E' possibile che solo alcuni soggetti percepiscano tali elementi e che quindi ne risultino influenzati; tale capacità di percezione risulta essere una dimensione al di fuori della capacità realmente misurata.

- FUNZIONE CARATTERISTICA DELL'ITEM

  Tale assunto riguarda la relazione tra la variabile non direttamente osservabile (tratto latente) e la variabile realmente osservabile (risposta all'item): maggiore è il valore del tratto latente maggiore è la probabilità di rispondere affermativamente (o, secondo i casi, correttamente). La relazione tra risposta e dimensione latente crescente può essere descritta da una funzione monotona, secondo la quale all'aumentare del livello di una data caratteristica (per esempio di capacità), aumenta la probabilità di rispondere in modo affermativo o nella direzione della dimensione misurata (per esempio in modo corretto) ad un item; è quindi possibile affermare che *la probabilità di un soggetto di rispondere correttamente ad un certo item dipende dalla disposizione del soggetto e dalle caratteristiche dell'item* ovvero dipende dalla correlazione tra caratteristica misurata e caratteristiche dell'item. La probabilità di dare una risposta affermativa/corretta in funzione del valore individuale del tratto latente per ciascun item è definita da una funzione matematica (*Item Characteristic Function*, *ICF*) e rappresentata da una curva (*Item Operating Characteristics* o *Item Characteristic Curve*, *ICC*).

  E' possibile identificare e definire diverse *ICC* come si può osservare dai due esempi rappresentati di seguito che indicano come all'aumentare dell'intensità dell'attributo la probabilità *alfa* aumenta con funzione, rispettivamente, lineare e monotona;



4.1.1 L'invarianza dei parametri

La caratteristica che maggiormente distingue l'*IRT* dalla teoria classica della misurazione è la proprietà dell'invarianza della capacità del soggetto e dei parametri relativi all'item. Secondo tale proprietà i parametri che caratterizzano
- ciascun item, non dipendono dalla distribuzione della capacità dei soggetti,
- la capacità di un soggetto, non dipendono dal gruppo di item.

Se il modello di *IRT* si adatta ai dati, la *ICC* di un determinato item risulta la stessa, indipendentemente dalla distribuzione della capacità del campione di soggetti utilizzati per stimare i

parametri degli item.
In altre parole i valori di capacità più alti presentano anche probabilità più alte di rispondere correttamente all'item rispetto ai soggetti con valori più bassi, indipendentemente dalla distribuzione del gruppo di appartenenza; quindi i soggetti che presentano lo stesso livello di capacità hanno la stessa probabilità di fornire una risposta corretta all'item, indipendentemente dal gruppo di appartenenza.
Sapendo che la probabilità di successo per un soggetto con una data capacità è determinata dai parametri dell'item, anche i parametri dell'item per i due gruppi devono essere uguali.
La proprietà di invarianza è importante non solo teoricamente in quanto consente importanti applicazioni come quella di effettuare confronti, di costruire banche di item, di analizzare l'errore sistematico di particolari item e di stimare gli errori standard delle stime delle capacità individuali, diversamente dal modello classico di misurazione in cui viene definito un solo errore uguale per tutti i casi[14].

4.2 I MODELLI

Esistono molte tecniche che consentono di stimare la versione probabilistica del modello di *scaling* cumulativo. Tutte sono in grado di gestire le deviazioni empiriche dalla forma perfetta di scala cumulativa. Con tali tecniche è inoltre possibile valutare l'adattamento del modello di *scaling* rispetto all'ipotesi nulla che le risposte ai singoli item sono tra loro statisticamente indipendenti.
La distinzione tra i diversi modelli basati sull'*IRT* è fatta sulla base del numero e del tipo di parametri utilizzati nel definire la funzione adottata:

- Modello logistico con un parametro: si tratta di uno dei modelli probabilistici più noti, conosciuto spesso nella versione detta *Rasch* dal nome dello studioso che la ha sviluppata (1960); secondo questo modello le *ICC* sono rappresentate da curve logistiche parallele i cui valori variano tra 0 e 1; questo modello ha la proprietà dell'"oggettività specifica": i soggetti sono misurati su una scala a rapporti indipendente dagli item utilizzati e viceversa; rappresenta il miglior approccio alla individuazione di gruppi di item per i quali le risposte individuali ai primi item determinano la selezione, tra i successivi item, dei migliori per la misurazione del particolare soggetto.
- Modello con due parametri: Lord nel 1952 è stato il primo a sviluppare un modello con due parametri basato sull'assunto rappresentato dall'*ogiva normale* (distribuzione normale cumulata). Una *ICC* con tale forma descrive un item che è maggiormente discriminante nei punti in cui la curva è più ripida. Più ripida è tale sezione della curva, maggiore è la correlazione biseriale dell'item con l'attributo[15]. Al diminuire della correlazione tra item e attributo, la curva tende ad appiattirsi fino ad assumere la posizione di linea retta orizzontale. L'adozione del modello che fa riferimento all'ogiva normale consente di identificare per ciascun item una zona critica (individuata da un punto critico di discriminazione) corrispondente al livello di incertezza delle risposte dei soggetti; allontanandosi da tale zona in entrambe le direzioni, l'incertezza si riduce in modo evidente: la probabilità di dare una risposta positiva all'item è bassa al di sotto del valore che identifica la zona critica e alta al di sopra. Questo consente di definire, nel caso in cui si misurino capacità, la difficoltà dell'item. L'individuazione del punto critico per ogni item consente di selezionare gli item che consentono di discriminare in

[14] La caratteristica dell'invarianza non è nuova per la statistica. Essa riguarda infatti anche i modelli di regressione lineare: stabilita l'equazione e verificato il modello di regressione, la pendenza e l'intercetta della retta sarà la stessa in qualsiasi sottopopolazione della variabile *X*. Ricordiamo invece che un indice come il coefficiente di correlazione, parametro non caratterizzante direttamente la retta di regressione, cambia se calcolato su un sottogruppo. Quindi mentre il parametro di pendenza non dipende dalle caratteristiche della sottopopolazione, il coefficiente di correlazione sì.
L'applicazione dello stesso concetto anche ai modelli di *IRT* ci permette di considerare questi come modelli di regressione non lineare.

[15] Se tale sezione fosse verticale, le code scomparirebbero, l'item correlerebbe perfettamente con l'attributo; in questo caso l'item risponderebbe alle caratteristiche proprie del modello deterministico-cumulativo.

determinati punti.
Birnbaum successivamente sostituì la funzione normale con la funzione *logistica*, che presenta il vantaggio di poter essere più facilmente e convenientemente trattata dal punto di vista matematico e di godere di alcune importanti proprietà statistiche[16].

- Modello Birnbaum (tre parametri): le *ICC* sono rappresentate da curve logistiche che possono presentare diverse pendenze e asintoti che riflettono la possibilità di rispondere correttamente anche a caso e, in alcune varianti, prevedono item a scelta multipla; tale modello consente di selezionare il gruppo di item più "informativo" per un dato gruppo di soggetti.
- Modello Mokken: modello non parametrico che richiede curve (con valori che vanno da 0 e 1) che non si intersecano e che conducono ad un numero limitato di violazioni delle caratteristiche degli item di una scala Guttman (item ordinali); tale modello, oggetto di un vasto dibattito, presenta gli assunti meno restrittivi e un concetto diverso di qualità della misurazione.

Anche se i modelli probabilistici in genere fanno riferimento:

- alla caratteristica del soggetto in termini di *capacità*,
- alla caratteristica dell'item in termini soprattutto di *difficoltà*,

tali modelli possono trovare applicazione e quindi essere generalizzati anche ad altri casi.
Tra i più noti e diffusi modelli di *IRT* di seguito esamineremo i più diffusi modelli logistici.

4.2.1 I modelli logistici

4.2.1.1 Modello con un parametro

Il modello con un parametro assume che l'unica caratteristica che influenza la risposta del soggetto sia la difficoltà dell'item stesso (b_i). La *ICF* che descrive tale modello è la seguente:

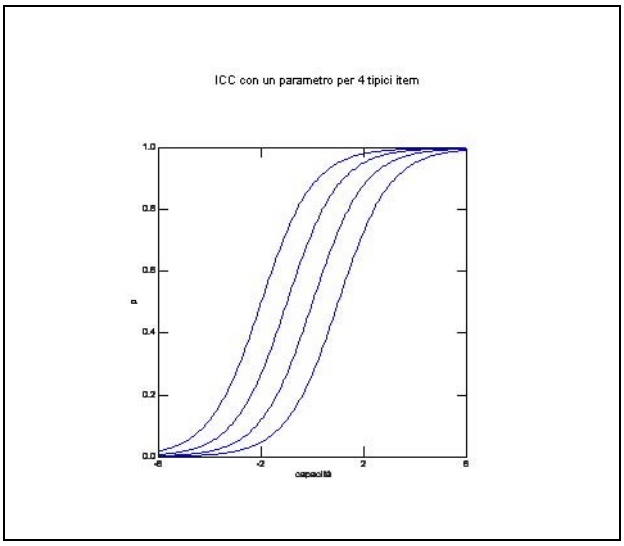
$$P_i(d) = \frac{e^{d-b_i}}{1+e^{d-b_i}} \qquad i = 1,2,...,n$$

dove

$P_i(d)$	probabilità del soggetto *j* con capacità *d* di rispondere correttamente all'item *i*
e	numero trascendente (come il *p-greco,* π) il cui valore è 2.718 (corretto ai tre decimali)
d	parametro di capacità del soggetto (*j*)
b_i	parametro di difficoltà dell'item *i*
n	numero di item.

Il parametro b_i può essere interpretato come indicatore della posizione della *ICC* in relazione alla scala di capacità. Tale parametro infatti rappresenta il punto della scala di capacità dove la probabilità di dare una risposta corretta è 0.5; maggiore è il valore del parametro b_i, maggiore è il livello di capacità richiesto ad un soggetto per aver una possibilità del 50% di rispondere in modo corretto, maggiore è la difficoltà dell'item.
In genere, quando i valori di capacità dei soggetti sono standardizzati, i valori di difficoltà degli item (b_i) variano da -2.0 a +2.0. Di seguito sono presentati alcuni esempi di *ICC* per il modello con un parametro (lateralmente sono riportati i relativi valori del parametro).

[16] Per alcuni utili richiami sulle definizioni di logaritmi e *logit* si veda l'Appendice *A*.



Le curve differiscono solo per la loro posizione rispetto alla scala di capacità; infatti da tale grafico si osserva che gli item più

- difficili sono localizzati all'estrema destra ovvero nella parte in cui i valori di capacità sono più alti (valori di b_i vicini a 2.0);
- facili sono posizionati all'estrema sinistra ovvero nella parte in cui i valori della scala di capacità sono più bassi (valori di b_i vicini a -2.0)

Secondo questo approccio i soggetti con capacità molto bassa hanno probabilità nulla di rispondere correttamente all'item. Questo modello logistico, matematicamente equivalente al modello *Rasch*, è basato su assunti restrittivi, la cui adeguatezza dipende dalla natura dei dati e dall'importanza dell'applicazione stessa.

4.2.1.2 Modello con due parametri

Il modello con due parametri può essere considerato l'estensione del precedente modello rispetto al quale presenta un nuovo parametro: discriminazione dell'item[17] (a_i); questo è proporzionale alla pendenza della *ICC* nel punto b_i della scala di capacità.

In teoria il parametro di discriminazione dell'item è definito su una scala che va da $-\infty$ a $+\infty$; nella pratica tale parametro difficilmente supera 2.0; i valori molto bassi definiscono *ICC* per le quali la capacità aumenta gradualmente mentre i valori molto alti di a_i identificano *ICC* con le maggiori pendenze corrispondenti quelli che sono gli item più utili a discriminare i soggetti in punti diversi del continuum di capacità. Accanto a tale parametro è stato introdotto anche un fattore di *scaling* che rende la funzione logistica più simile possibile alla funzione normale (D[18]).

L'equazione matematica alla base del modello con due parametri (Birnbaum, 1968) è la seguente:

$$P_i(d) = \frac{e^{Da_i(d-b_i)}}{1+e^{Da_i(d-b_i)}} \qquad i = 1,2,...,n$$

dove

n	numero di item
$P_i(d)$	probabilità del soggetto j con capacità d di rispondere correttamente all'item i
e	numero trascendente (come il π) il cui valore è 2.718 (corretto ai tre decimali).
d	parametro di capacità del soggetto (j)

[17] A differenza del modello classico, che considera tutti gli item ugualmente discriminanti, il valore di tale parametro cambia da item a item.

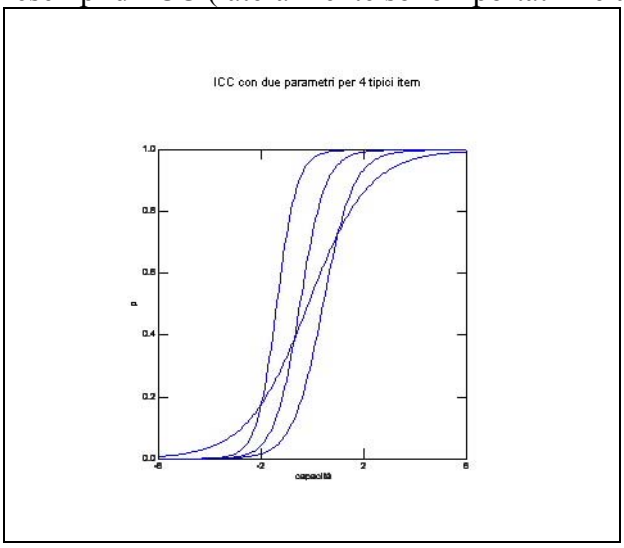
[18] E' stato dimostrato che quando D=1.7, i valori di $P_i(d)$ per il modello dell'ogiva normale con due parametri e per il modello logistico con due parametri differiscono nel valore assoluto di meno di 0.01 per tutti i valori di d.

b_i parametro di difficoltà dell'item *i*

a_i parametro di discriminazione dell'item *i*

D fattore di *scaling* che avvicina la funzione logistica a quella normale.

Di seguito vediamo alcuni esempi di *ICC* (lateralmente sono riportati i relativi valori dei parametri).



Le curve non sono parallele come nel caso del precedente modello; ciò è dovuto ai diversi valori del parametro di discriminazione cui corrispondono pendenze diverse. Inoltre per ciascuna curva le asintoti sono uguali a zero; ciò indica che oltre i parametri considerati non esiste alcuna altro elemento che spieghi le risposte.

4.2.1.3 Modello con tre parametri

Il nuovo parametro introdotto con questo modello rappresenta la probabilità dei soggetti con bassa capacità di rispondere correttamente all'item (c_i). Tale parametro consente di prendere in considerazione l'esecuzione rispetto al limite inferiore della scala di capacità, dove *tirare a indovinare* può rappresentare un fattore che influenza la risposta. Tale parametro risulta particolarmente adatto ai casi in cui si sottopongono scale di risposta in cui appaiono scelte convincenti ma scorrette.

La funzione matematica alla base del modello logistico con tre parametri è la seguente:

$$P_i(d) = c_i + (1 - c_i)\frac{e^{Da_i(d-b_i)}}{1+e^{Da_i(d-b_i)}} \qquad i = 1,2,...,n$$

dove

$P_i(d)$ probabilità del soggetto *j* con capacità *d* di rispondere correttamente all'item *i*

e costante (come il π) il cui valore è 2.718 (corretto ai tre decimali)

c_i parametro del *livello pseudo-casuale*

d parametro di capacità del soggetto (*j*)

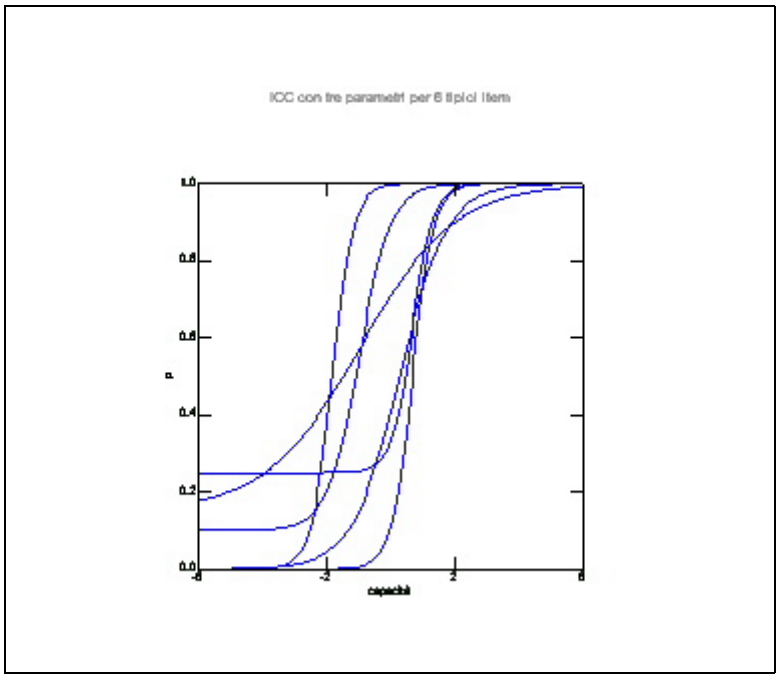
b_i parametro di difficoltà dell'item *i*

a_i parametro di discriminazione dell'item *i*

D fattore di *scaling* che avvicina la funzione logistica a quella normale

n numero di item.

Di seguito vediamo alcuni esempi tipici di *ICC* definiti da tale modello (lateralmente sono riportati i relativi valori dei parametri).



Il confronto tra le diverse *ICC* corrispondenti ai diversi item consente di evidenziare il ruolo dei diversi parametri:

- *parametro di difficoltà*: gli item più difficili (1, 2, 3) sono posizionati sull'estremo più alto della scala di capacità;
- *parametro di discriminazione*: confrontando gli item 1 e 2 (o gli item 1, 3 e 4 con gli item 2, 5 e 6) si evidenziano le diverse "ripidità";
- *parametro di casualità*: confrontando le asintoti appare evidente come per gli item 3, 5 e 6 vi sia, anche se a livelli diversi, la possibilità di rispondere correttamente tirando a indovinare.

4.2.2 Altri modelli

Riassumendo i modelli logistici visti, possiamo dire che i modelli di risposta ad un item di un soggetto con una determinata capacità sono definiti da:

- difficoltà dell'item (modello con un parametro);
- difficoltà e discriminazione dell'item (modello con due parametri);
- difficoltà, discriminazione dell'item e casualità di risposta (modello con tre parametri).

Nell'ambito dell'*Item Response Theory* sono stati sviluppati altri modelli, molti dei quali possono essere applicati ad item non dicotomici. E' il caso del modello con due parametri presentato da Bock nel 1972 (HAMBLETON & altri, 1991), sviluppato per essere applicato nei casi di item a risposta multipla. Esso viene detto *modello a risposte nominali* (*nominal response model*) in quanto considera le scale di risposta nominali ovvero non prevede alcun ordine a priori nelle modalità di risposta. Lo scopo di tale modello è quello di massimizzare la precisione delle stime di capacità utilizzando tutta l'informazione contenuta nelle risposte del soggetto, non solo nel caso di risposta corretta all'item. Secondo tale modello, la probabilità che un soggetto selezioni una particolare risposta *k* (su *m* risposte disponibili) all'item *i* è data da

$$P_{ik}(d) = \frac{e^{a_{ik}^{*}(d-b_{ik}^{*})}}{\sum_{h=1}^{m} e^{a_{ih}^{*}(d-b_{ih}^{*})}} \qquad i = 1,2,...,n;1,2,...,m$$

Per ciascuna *d*, la somma delle probabilità per le *m* opzioni è uguale a uno ovvero $\sum P_{ik} = 1$. Le

quantità b_{ik}^* e a_{ik}^* rappresentano i parametri di item relativi alla *k-esima* opzione[19].

Esiste un altro modello detto *graded response* presentato da Samejima nel 1969 che può essere applicato ad item con scale di risposte con categorie ordinate. In questo modello, come nel precedente, si cerca di ottenere più informazioni relative alle risposte scorrette. Dato il maggiore utilizzo di scale di risposta ordinali nella ricerca sociale si comprende l'interesse suscitato da tale modello. Supponiamo che le categorie di risposta di un item siano ordinate dal valore più piccolo a quello più grande e che siano indicate con $x_i = 0,1,...,m_i$ dove $m_i + 1$ rappresenta il numero di categorie per l'*i-esimo* item. La probabilità di un soggetto di rispondere ad un item in una particolare categoria o ad una più alta può essere determinata attraverso una estensione del modello logistico con due parametri:

$$P_{xi}^*(d) = \frac{e^{Da_i(d-b_{xi})}}{1+e^{Da_i(d-b_{xi})}}$$

dove

b_{x_i} livello di difficoltà della categoria m_i.

Con $m_i + 1$ categorie, è necessario stimare m_i valori di difficoltà per ciascun item, più un parametro di discriminazione di item. La reale probabilità di un soggetto di ottenere un punteggio di x_i è data dalla seguente espressione

$$P_{xi}(d) = P_{xi}^*(d) - P_{xi+1}^*(d)$$

Per esempio con 50 item e 5 punti di scala per item, dovrebbero essere stimati un totale di (50*4)+50=250 valori di parametri di item!!

4.3 LA VERIFICA DEL MODELLO

La verifica dei modelli probabilistici prevede principalmente due passaggi:

a. STIMA DEI PARAMETRI che caratterizzano il modello prescelto. Essendo tali parametri ignoti, il primo e più importante momento nell'applicazione dei modelli di *IRT* è quello della loro stima. Tale stima, come in altri ambiti della statistica, avviene necessariamente a partire dalle osservazioni sperimentali (risposte dei soggetti). Il momento della stima dei parametri è particolarmente delicato in quanto la validità dell'applicazione dell'*IRT* dipende dalla possibilità di disporre di soddisfacenti procedure di stima; tali parametri sono la capacità e quelli che riguardano gli item; ricordiamo che nei modelli logistici che abbiamo visto questi ultimi sono, a seconda del modello,
   - difficoltà dell'item *i* (b_i),
   - discriminazione dell'item *i* (a_i),
   - casualità di risposta all'item *i* (c_i), corrispondente alla possibilità anche con bassa capacità di rispondere correttamente.
b. VERIFICA DELL'ADATTAMENTO DEL MODELLO AI DATI ovvero della possibilità di questo di prevedere e/o spiegare in modo adeguato i dati.

4.3.1 La stima dei parametri

Come sappiamo la probabilità di dare una risposta corretta è definita dai parametri dell'item e della capacità; tali parametri sono però sconosciuti: l'unico elemento noto è rappresentato dalle risposte

[19] Il simbolo "*" indica "fino a"; quindi $P_{x_i}^*$ indica "*P* fino a x_i".

dati dagli individui. Per tale motivo il primo è più importante momento nell'applicazione dell'*IRT* è quello della stima dei parametri che caratterizzano il modello di risposta degli item prescelti ovvero la capacità per ciascun soggetto e i parametri assunti per ciascun item. Alla possibilità di disporre di soddisfacenti procedure per la stima dei parametri del modello è legato il successo delle successive applicazioni dell'*IRT*.
La stima dei parametri può essere eseguita in diversi modi. La strategia adottata per la stima dei parametri, nel caso di dati campionari, è quella che identifica i valori dei parametri che producono la migliore curva di adattamento. Tale problema è simile a quello che viene affrontato nell'ambito dell'analisi di regressione: i parametri che caratterizzano il modello di regressione (coefficienti di regressione) devono essere stimati a partire dai dati osservati; in tale sede il criterio che consente di definire il migliore adattamento, come sappiamo, è quello dei minimi quadrati.
I modelli di *IRT*, a differenza di quelli di regressione, non consentono l'utilizzazione di tale criterio in quanto non sono lineari e non dispongono dei valori osservabili per la variabile indipendente (*d* non è infatti osservabile direttamente). Vedremo come per tali modelli si utilizza principalmente l'approccio di stima basato sulla *massima verosimiglianza*.[20]

4.3.1.1 Stima della capacità

Riprendendo l'assunto di indipendenza locale, il prodotto delle probabilità di osservare le risposte di ciascun item è uguale a:

$$P(U_1,U_2,...,U_n|d)=P(U_1|d)*P(U_2|d)*...*P(U_i|d)*...*P(U_n|d)=\prod_{i=1}^{n}P(U_i|d)$$

dove

n	numero di item
d	capacità che influenza la risposta del soggetto
U_i	risposta del soggetto all'item i (i=1,2,...,n)
$P(U_i\|d)$	probabilità di risposta di un soggetto con capacità d

Tenendo conto del fatto che
$P(U_i=1|d)\rightarrow$ probabilità di risposta corretta
$P(U_i=0|d)\rightarrow$ probabilità di dare una risposta scorretta
la funzione di probabilità può essere riscritta:

$$P(U_1,U_2,...U_i,...,U_n|d)=\prod_{i=1}^{n}P(U_i|d)^{U_i}\left[1-P(U_i|d)\right]^{1-U_i} \text{ ovvero } \qquad \prod_{i=1}^{n}P_i^{U_i}Q_i^{1-U_i}$$

dove

P_i	$P(U_i\|d)$
Q_i	$1-P(U_i\|d)$

Tale equazione rappresenta un'espressione della probabilità congiunta di un modello di risposta.
Quando il modello di risposta non è teorico ma osservato $(U_i=u_i)$, l'interpretazione probabilistica

[20] Il metodo di *stima di massima verosimiglianza* (in inglese *Maximum Likelihood Estimation*, *MLE*) si basa su una serie di approssimazioni successive ai valori incogniti dei veri parametri. A differenza del metodo dei minimi quadrati comuni, che valuta la *bontà di adattamento* ai dati del modello calcolando la somma delle differenze al quadrato tra i valori predetti e quelli osservati, il metodo di *massima verosimiglianza* calcola la probabilità di osservare ciascun possibile valore, assumendo che un dato parametro sia vero. Il parametro che risulta associato alla probabilità più alta costituisce la stima di *massima verosimiglianza*.
A tale metodo non sono associate formule simili a quelle utilizzate per la stima dei minimi quadrati ma algoritmi che consentono di esaminare in modo iterativo più parametri fino a quando non viene identificato quello migliore. Nella prima fase si stabiliscono delle stime iniziali dei parametri; una serie di iterazioni produce in successione nuove stime e le confronta con quelle precedenti; le iterazioni continuano fino a quando le stime ottenute nel ciclo appena concluso differiscono da quelle ottenute nel precedente di una quantità inferiore a uno certo valore stabilito. Le distribuzioni campionarie delle stime di massima verosimiglianza sono note solo nel caso di grandi campioni quando sono normali.

non è più appropriata. L'espressione per la probabilità congiunta è chiamata *funzione di verosimiglianza* (*likelihood*, *L*):

$$L(u_1,u_2,...,u_i,u_n|d)=\prod_{i=1}^{n}P_i^{u_i}Q_i^{1-u_i}$$

Dato che P_i e Q_i sono funzioni di *d* e dei parametri degli item, anche la funzione di verosimiglianza rappresenta una funzione di tali parametri. L'applicazione di tale funzione può essere migliorata sottoponendola a trasformazione logaritmica[21]; utilizzando le proprietà dei logaritmi la funzione di verosimiglianza diviene funzione *log-verosimiglianza*:

$$\log_n L(u|d)=\sum_{i=1}^{n}[u_i\log_n P_i+(1-u_i)\log_n(1-P_i)]$$

dove **u** rappresenta il vettore delle risposte agli item. Il valore massimo di *d* prodotto dalla funzione per un soggetto è definito come *stima di massima verosimiglianza* di *d*.
Il problema dell'individuazione del valore massimo non è secondario. Il valore che massimizza la funzione può essere determinata utilizzando una procedura di ricerca automatica. Le procedure più efficienti sono quelle che, nel punto in cui la funzione raggiunge il massimo, la pendenza della funzione è zero. Così la stima di massima verosimiglianza può essere determinata utilizzando un metodo di approssimazione. L'utilità delle stime di massima verosimiglianza (*maximum likelihood estimes*, *MLE*) sta nel fatto che hanno proprietà asintotiche ben note; l'asintote si riferisce al numero di item all'aumentare del quale, la *MLE* di *d* $(\hat{d})$ è distribuita normalmente con media *d*. Ciò comporta che la distribuzione asintotica di $\hat{d}$ è centrata sul valore di *d*, quindi la *MLE* di $\hat{d}$ non è affetta da errore sistematico con un numero alto di item. E' possibile determinare per $\hat{d}$ l'errore standard e l'intervallo di confidenza. Accanto alla stima di massima verosimiglianza è utilizzabile anche la procedura Bayesiana che consente di risolvere i problemi presentati dall'approccio di massima verosimiglianza. Ricordiamo che le distribuzioni campionarie delle stime di massima verosimiglianza sono note solo nel caso di grandi campioni.

4.3.1.2 Stima dei parametri degli item

Nel descrivere le procedure per stimare *d* abbiamo assunto noti i parametri degli item. D'altra parte anche i parametri degli item devono essere stimati. Come abbiamo visto per stimare la capacità di un soggetto si assumono noti i parametri degli item e si applica la funzione di verosimiglianza alle risposte ottenute. Al contrario se si vogliono stimare i parametri degli item si assumono noti i valori individuali di *d* e si applica la funzione di verosimiglianza alle risposte ottenute:

$$L(u_1,u_2,...,u_j,...,u_N|d,a,b,c)=\prod_{j=1}^{N}P_j^{u_j}Q_j^{1-u_j}$$

dove
a, *b*, e *c* parametri degli item (modello con tre parametri)
L'applicazione della funzione di verosimiglianza per la stima dei parametri degli item, a differenza di quella per la stima della capacità, non richiede l'assunto d'indipendenza locale ma quello d'indipendenza delle risposte di *N* soggetti ad un item. Essendo noti i valori *d*, la stima dei parametri degli item è piuttosto semplice ed è confrontabile con la procedura descritta in precedenza; la differenza sta nel fatto che la funzione di verosimiglianza per i parametri degli item è multidimensionale.

4.3.1.3 Stima dei parametri di item e capacità

La situazione più difficile da risolvere, ma anche la più comune, è quella in cui devono essere

[21] Il vantaggio di tale trasformazione è quello di poter utilizzare, come abbiamo visto, le proprietà dei logaritmi.

stimati sia i parametri degli item che la capacità; in questo caso è necessario prendere in considerazione simultaneamente tutte le risposte a tutti gli item di tutti i soggetti.
La funzione di verosimiglianza con N soggetti che rispondono a n item, assumendo l'indipendenza locale, è

$$L(u_1,u_2,...,u_N|d,a,b,c)=\prod_{i=1}^{n}\prod_{j=1}^{N} P_{ij}^{U} Q_{ij}^{1-U}$$

dove

U u_{ij}

u_j modello di risposta del soggetto j agli n item

$\boldsymbol{d}$ vettore degli N parametri di capacità

$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$ vettori dei parametri degli item per gli n item

Considerando che:

a. il numero dei parametri degli item è n nel modello con un parametro, $2n$ nel modello con due parametri e $3n$ in quello con tre,
b. il numero di parametri di capacità è N,

il numero massimo di parametri da stimare è $3n+N$.
Prima di procedere con la stima è necessario affrontare il problema dell'indeterminatezza. Nella funzione di verosimiglianza vista in precedenza i parametri degli item e di capacità non sono determinati in maniera unica.
La funzione per l'item relativa al modello con tre parametri

$$P_i(d)=c_i+(1-c_i)\frac{e^{Da_i(d-b_i)}}{1+e^{Da_i(d-b_i)}} \qquad i=1,2,...,n$$

Se in tale equazione sostituiamo

$$d\to d^*=ad+\beta \qquad b\to b^*=ab+\beta \qquad a\to a^*=a/\alpha$$

la probabilità di una risposta corretta rimane invariata ovvero

$$P(d)=P(d^*)$$

Siccome i valori di α e β sono costanti e arbitrari, la funzione di verosimiglianza non avrà un massimo unico ovvero sarà indeterminata e quindi non consentirà di cercare il massimo valore di verosimiglianza. Tale problema non esiste nella stima di d quando i parametri degli item sono noti o nella situazione parallela in cui i parametri degli item sono stimati disponendo dei parametri di capacità.
Il problema dell'indeterminatezza può essere risolto scegliendo una scala arbitraria per i valori di capacità e per i valori di b, per esempio standardizzando gli N valori di capacità e gli n valori di difficoltà. Ciò consente di confrontare le stime dei parametri degli item effettuate su gruppi diversi.
Una volta eliminata l'indeterminatezza, è possibile determinare i valori dei parametri di capacità e degli item che massimizzano la funzione di verosimiglianza.
Una delle procedure che consente di fare ciò è la *stima congiunta della massima verosimiglianza* (*joint maximum likelihood estimation*) che viene eseguita in due fasi (*stage*):

1ª fase	○ scelta dei valori iniziali per il parametro di capacità per ciascun soggetto: *$\log_n$ (numero di risposte corrette / numero di risposte scorrette)* ○ standardizzazione di tali valori per eliminare l'indeterminatezza; ○ stima dei parametri dell'item, considerando noti i valori di capacità;
2ª fase	○ stima dei parametri di capacità, considerando noti i valori di dei parametri dell'item.

Tale procedura viene ripetuta secondo vari passaggi (*step*) fino a quando i valori delle stime non cambiano tra due successivi passaggi.
La procedura congiunta di massima verosimiglianza anche se concettualmente convincente presenta degli svantaggi:

a. non consente stime per capacità perfette o con punteggio zero;
b. non consente stime dei parametri per quegli item ai quali tutti i soggetti hanno risposto correttamente (o scorrettamente);

c. non produce stime consistenti[22] dei parametri di item e capacità per il modello con tre parametri.

E' possibile superare alcuni di tali problemi utilizzando un approccio alternativo che produce stime Bayesiane ottenute utilizzando distribuzioni a priori.

Il problema dell'inconsistenza delle stime congiunte di massima verosimiglianza è dovuta alla simultanea stima dei parametri degli item e di capacità. Tale problema però scompare se i parametri degli item possono essere stimati senza fare alcun riferimento ai parametri di capacità ma a specifiche distribuzioni dei parametri di capacità, considerando i soggetti estratti causalmente da una popolazione. Ne risultano *stime di massima verosimiglianza marginale* (*marginal maximum likelihood estimes*) che presenta positive proprietà asintotiche: le stime dei parametri degli item sono consistenti ovvero si avvicinano al valore del parametro all'aumentare del numero di soggetti.

Per poter ottenere la funzione di verosimiglianza marginale dei parametri degli item, è necessario approssimare la distribuzione di capacità. Per una buona approssimazione della distribuzione di capacità è importante disporre di un grande numero di soggetti ovvero la procedura di massima verosimiglianza marginale dovrebbe essere applicata solo in presenza di un campione ampio di soggetti.

Una volta stimati i parametri degli item con tale procedura, le stime dei parametri degli item possono essere utilizzate per stimare le capacità utilizzando il metodo visto in precedenza.

Oltre alle procedure di stima *di massima verosimiglianza*, *di massima verosimiglianza marginale* e *Bayesiana* esistono altre procedure, così riassumibili:

- Procedura di stima *congiunta di massima verosimiglianza* (Lord 1974, 1980), applicabile ai modelli con uno, due e tre parametri; i parametri di capacità e degli item sono stimati simultaneamente.
- Procedura di stima di *massima verosimiglianza marginale* (Bock & Aitkin 1981), applicabile ai modelli con uno, due e tre parametri; la stima della capacità avviene successivamente a quella degli altri parametri.
- Procedura di stima di *massima verosimiglianza condizionale* (Andersen 1972, 1973, Rasch 1960), applicabile solamente al modello ad un parametro; la funzione di verosimiglianza è condizionata sul numero di punteggi corretti.
- Procedure di stima *Bayesiana congiunta e marginale* (Mislevy 1986, Swaminathan & Gifford 1982, 1985, 1986), applicabile ai modelli ad uno, due e tre parametri; si definiscono distribuzioni a priori attribuite ai parametri degli item e di capacità, eliminando così alcuni dei problemi, come l'impropria stima dei parametri e la mancanza di convergenza, che si hanno con le procedure di massima verosimiglianza marginale e congiunta.
- Procedura di *stima euristica* (Urry 1974, 1978), applicabile principalmente ai modelli con due e tre parametri.
- Procedure basate su *analisi fattoriale non-lineare* (McDonald 1967, 1989), applicabile al modello con due parametri e ad un modello modificato con tre parametri in cui i valori *c* sono fissi.

4.3.1.4 Stima dei parametri per un modello semplice

Come abbiamo visto per il modello logistico con un parametro (modello *Rasch*), la risposta corretta è più probabile quando la capacità del soggetto supera la difficoltà dell'item. Tale modello assume che la probabilità di un soggetto di rispondere correttamente ad un item sia dovuto a due parametri:

- *capacità* del soggetto (d_j , posizione del soggetto *j* sul continuum),
- *difficoltà* dell'item (b_i , posizione dell'item *i* sul continuum).

Entrambi i parametri possono essere misurati dall'item; vediamo come:

- capacità del soggetto: *rapporto tra numero di risposte corrette* (x_j) *e il numero di risposte errate* $(n - x_j)$ *fornite dal soggetto* j *a tutti gli item*

$$d_j = \frac{x_j}{n - x_j}$$

dove

[22] Ricordiamo che una stima è detta "consistente" quando all'aumentare della dimensione del campione aumenta la probabilità di avvicinarsi al valore corrispondente della popolazione.

x_j numero di risposte corrette fornite dal soggetto *j* a tutti gli item

$n - x_j$ numero di risposte errate del soggetto *j*

n numero di item.

- <u>difficoltà dell'item</u>: *rapporto tra numero di risposte errate* $(N - y_i)$ *e numero di risposte corrette* (y_i) *fornite all'item* i *da tutti i soggetti del campione*

$$b_i = \frac{N - y_i}{y_i}$$

dove

y_i numero di risposte corrette fornite all'item *i* da tutti i soggetti del campione

$(N - y_i)$ numero di risposte errate all'item *i*

N numero di soggetti del campione.

Tali calcoli vengono effettuati dopo aver escluso dall'analisi i soggetti che hanno ottenuto i punteggi massimi e minimi e gli item che hanno registrato risposte costanti.

Disponendo della misura della capacità dei soggetti e della difficoltà degli item, la probabilità di risposta può essere definita come la differenza tra i due valori; quindi se la capacità del soggetto *j* (d_j) è maggiore della difficoltà dell'item *i* (b_i) allora la probabilità di rispondere correttamente all'item è maggiore di 0.5, ovvero

$$\text{se } (d_j - b_i) > 0 \text{ allora } P_i(d) > 0.5$$

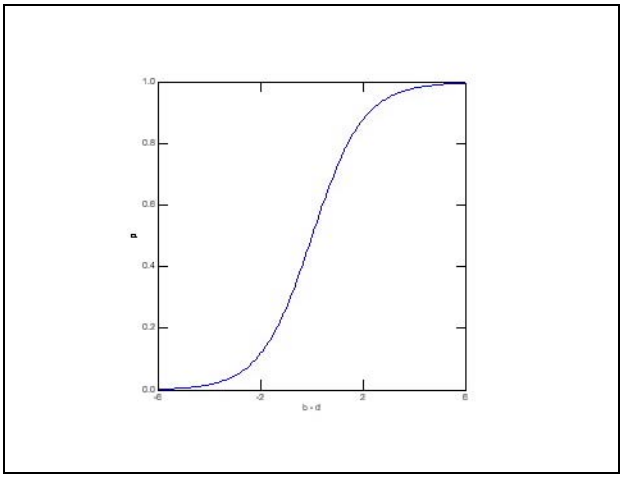
da cui

$$\text{se } (d_j - b_i) < 0 \text{ allora } P_i(d) < 0.5 \qquad \text{se } (d_j - b_i) = 0 \text{ allora } P_i(d) = 0.5$$

Ciò corrisponde al modello con un parametro definito dalla seguente equazione matematica

$$P_i(d) = \frac{e^{(d_j - b_i)}}{1 + e^{(d_j - b_i)}}$$

cui corrisponde la seguente curva:



Tali osservazioni sono sostenibili quando sia la *capacità* che la *difficoltà* sono misurate e sono esprimibili con la stessa unità di misura. Per poter ottenere una scala di misura con intervalli uguali la precedente equazione viene sottoposta a trasformazione logaritmica:

$$P_i(d) = \frac{e^{(\delta_j - \beta_i)}}{1 + e^{(\delta_j - \beta_i)}}$$

dove

δ_j $\log_n(d_j) \rightarrow$ *logit correct*

β_i $\log_n(b_i) \rightarrow$ *logit incorrect*

che corrispondono alle definizioni iniziali di capacità e di difficoltà. Quindi il *logit*

- della capacità del soggetto è il logaritmo naturale del rapporto tra il numero di risposte corrette e il numero di risposte errate fornite dal soggetto *j* per l'item che si trova all'origine

della scala ovvero che ha difficoltà zero;

- della difficoltà degli item è il logaritmo naturale del rapporto tra il numero di risposte errate e il numero di risposte corrette ottenute dall'item *i* per il soggetto che si trova all'origine della scala ovvero che ha capacità zero.

Le due quantità (d_j e b_i)[23] così calcolate e trasformate vengono riferite ad un'unica scala lineare che va da $-\infty$-a $+\infty$. Nella pratica i valori di difficoltà e quelli di capacità vanno da -4 a +4 *logit*:

- i valori negativi di capacità indicano soggetti con basse prestazioni,
- i valori negativi di difficoltà indicano item facili.

Per ciascuna stima è possibile calcolare i corrispondenti errori standard:

- Errore standard della capacità del soggetto:

$$E(d_j) = X\left[\frac{n}{x_j}(n - x_j)\right]^{1/2}$$

dove

x_j numero di risposte corrette fornite dal soggetto *j* a tutti gli item

$n - x_j$ numero di risposte errate del soggetto *j*

n numero di item.

- Errore standard della difficoltà dell'item:

$$E(b_i) = Y\left[\frac{N}{y_i}(N - y_i)\right]^{1/2}$$

dove

y_i numero di risposte corrette fornite all'item *i* da tutti i soggetti del campione

---

[23] Sono stati definiti ed elaborati altri algoritmi di stima della capacità e della difficoltà:

- Capacità del soggetto:

$$d_j = X(\delta_j) \quad \rightarrow \qquad d_j = X \log_n\left(\frac{x_j}{n - x_j}\right)$$

dove

n numero di item

X fattore di espansione di capacità, ovvero

$$\left[\frac{(1 + U/1.7^2)}{(1 - UV/1.7^4)}\right]^{1/2}$$

U varianza degli n valori β_i

V varianza degli N valori δ_j.

- Difficoltà dell'item:

$$b_i = Y\left(\beta_i - \frac{\sum \beta_i}{n}\right) \quad \rightarrow \qquad b_i = Y\left[\log_n\left(\frac{N - y_i}{y_i}\right) - \frac{\sum \log_n\left(\frac{N - y_i}{y_i}\right)}{n}\right]$$

dove

N numero di soggetti del campione

Y fattore di espansione di difficoltà, ovvero

$$\left[\frac{(1 + V/1.7^2)}{(1 - UV/1.7^4)}\right]^{1/2}$$

U varianza degli n valori β_i

V varianza degli N valori δ_j

$N - y_i$ numero di risposte errate all'item *i*
N numero di soggetti del campione.

Ciascun errore standard può essere interpretato come *indice di non-affidabilità*; il suo valore è maggiore per i valori che nella scala di misurazione stanno agli estremi in quanto in questi casi è minore l'informazione fornita. L'errore standard della misura della difficoltà $(E(b_i))$ dipende dalle capacità dei soggetti. Il valore soglia di $E(b_i)$ è 0.25; gli item che superano tale valore sono da considerare sospetti.

Se nel modello è stato compreso anche il parametro di discriminazione (la cui procedura di stima non è qui presentata) occorre tenere presente che gli item con valori di discriminazione negativi sono scartati dagli strumenti che misurano capacità in quanto indicano che all'aumentare della capacità del soggetto diminuisce la probabilità di rispondere correttamente all'item, situazione non accettabile.

Se i soggetti in un campione hanno una capacità che si situa all'inizio della scala, ovvero hanno una probabilità di successo di 0.5, la precisione della misura è massima.

Come per tutti i modelli di *IRT*, le stime di difficoltà degli item che si ottengono dopo la procedura di controllo dell'adattamento dei risultati del modello e della successiva eliminazione di soggetti e di item incoerenti, sono praticamente indipendenti dai soggetti particolari; ciò vuol dire che lo strumento può essere tarato anche con una sola applicazione.

Riassumiamo nel seguente schema quanto visto finora.

STIMA DEI PARAMETRI NELL'APPROCCIO LOGISTICO CON UN PARAMETRO

⇓

Stima dei valori dei parametri che producono la migliore curva che descrive la probabilità di un soggetto di rispondere correttamente ad un item:

⇓

Si escludono dall'analisi:

- soggetti che hanno ottenuto i punteggi massimi e minimi
- item che hanno registrato risposte costanti

1. <u>capacità del soggetto</u> (d_j, posizione del soggetto *j* sul continuum): rapporto tra:
   - *numero di risposte corrette* (x_j)
   - *numero di risposte errate* $(n - x_j)$

   fornite dal caso *j* a tutti gli item

$$d_j = \frac{x_j}{n - x_j}$$

2. <u>difficoltà dell'item</u> (b_i, posizione dell'item *i* sul continuum): rapporto tra:
   - numero di risposte errate $(N - y_i)$
   - numero di risposte corrette (y_i)

   fornite dall'item *i* da tutti i casi

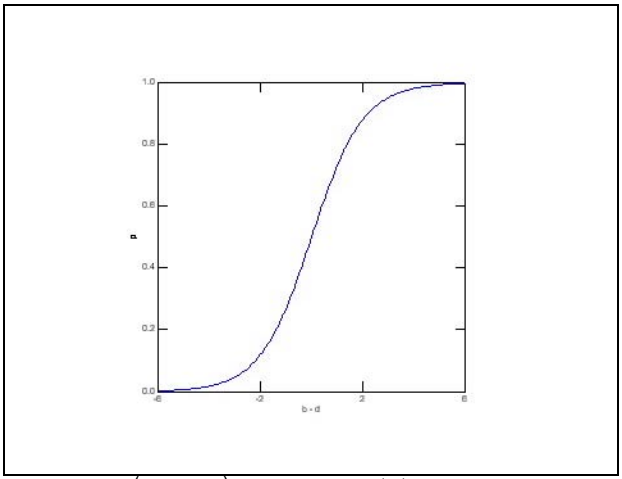
$$b_i = \frac{N - y_i}{y_i}$$

quanto maggiore è la differenza tra *capacità-soggetto* e *difficoltà-item* tanto maggiore è lo scarto quadrato che indica l'inaccettabilità di una risposta sbagliata (con $d_j > b_i$) o una risposta corretta (con $d_j < b_i$)

3. <u>probabilità di risposta</u>:

   (capacità del soggetto *j*) – (difficoltà dell'item *i*)

Modello $\rightarrow$ $$P_i(d) = \frac{e^{(d_j - b_i)}}{1 + e^{(d_j - b_i)}}$$



se $(d_j - b_i) > 0$ allora $P_i(d) > 0.5$

se $(d_j - b_i) < 0$ allora $P_i(d) < 0.5$

se $(d_j - b_i) = 0$ allora $P_i(d) = 0.5$

4. trasformazione di capacità e difficoltà su stessa unità di misura:

unica scala lineare per d_j e b_i → trasformazione logaritmica

↓

LOGIT

- della capacità del soggetto = *logit correct* = $\log_n(d_j) = \delta_j$
- della difficoltà dell'item = *logit incorrect* = $\log_n(b_i) = \beta_i$

↓

Valori in teoria: da $-\infty$ a $+\infty$

in pratica: da –4 a +4

- i valori negativi di capacità indicano soggetti con basse prestazioni
- i valori negativi di difficoltà indicano item facili

↓

Modello → $$P_i(d) = \frac{e^{(\delta_j - \beta_i)}}{1 + e^{(\delta_j - \beta_i)}}$$

5. errori standard → indici di *non-affidabilità*

a. errore standard della capacità del soggetto:

$$E(d_j) = X\left[\frac{n}{x_j}(n - x_j)\right]^{1/2}$$

b. errore standard della difficoltà dell'item:

$$E(b_i) = Y\left[\frac{N}{y_i}(N - y_i)\right]^{1/2}$$

$E(b_i)$:

- dipende dalle capacità dei soggetti
- valore soglia = 0.25
- se $E(b_i) > 0.25$ → item sospetto

dove

n	numero di item
N	numero di soggetti
x_j	numero di risposte corrette fornite dal soggetto *j* a tutti gli item
Y_i	numero di risposte corrette fornite all'item *i* da tutti i soggetti
$n - x_j$	numero di risposte errate del soggetto *j*
$N - Y_i$	numero di risposte errate all'item *i*
X	fattore di espansione di capacità
Y	fattore di espansione di difficoltà

4.3.2 L'adattamento del modello

Perché i modelli di *IRT* possano essere considerati validi, è necessario verificare il loro livello di adattamento ai dati. L'adattamento di un modello *IRT* ai dati indica che sono state raggiunte le caratteristiche desiderabili. Un modesto e debole adattamento del modello non produce parametri invarianti di item e di capacità. Quindi la riuscita delle specifiche applicazione di *IRT* non è legata solo alla possibilità di disporre di stime consistenti dei parametri ma soprattutto dalla possibilità di

valutare il livello di adattamento[24] tra dati e modello.
Il modello può essere considerato appropriato, per un particolare gruppo di dati, quando è in grado di prevedere o spiegare in modo adeguato i dati stessi. Alcuni autori raccomandano di valutare e verificare l'adattamento del modello sulla base di tre tipi di evidenza:

1. verifica della validità degli assunti del modello, utile nel selezionare modelli di *IRT* da utilizzare nella verifica del secondo e terzo tipo di evidenza;
2. verifica dell'invarianza dei parametri (item e capacità), essenziale in quanto tutte le applicazioni di *IRT* dipendono da tale caratteristica;
3. verifica dell'accuratezza delle previsioni del modello nell'utilizzare i dati reali o simulati; ciò richiede una valutazione del livello di spiegazione dei risultati da parte del modello *IRT* e della possibilità di comprendere la natura delle discrepanze tra modello e dati e delle conseguenze.

Nella scelta del modello più appropriato può essere utile verificare il livello di adattamento di più modelli e di confrontare i risultati con quelli ottenuti con dati simulati.

4.3.2.1 Verifica degli assunti

L'analisi dei principali assunti che stanno alla base dei modelli di *Item Response* consente di selezionare il modello migliore. Tale analisi in genere riguarda:

- l'*unidimensionalità* e l'*indipendenza locale* (per tutti i modelli),
- la *discriminazione* (modello con due e tre parametri),
- la *possibilità di rispondere a caso* (modello con tre parametri).

Di seguito vediamo alcuni tra i metodi di verifica utilizzabili tra i quali si noteranno alcuni utilizzati anche dal modello classico di *item analysis* e considerati validi anche in questo caso.

- **Unidimensionalità**

  Sono stati individuati moltissimi indici di valutazione dell'unidimensionalità (c'è chi ne ha contati ben 88!!); tra i pochi metodi giudicati più validi e soddisfacenti vi sono quelli basati sull'analisi fattoriale e l'analisi dei residui; di seguito vediamo alcuni tra i più interessanti.
  - *Plot* degli *eigenvalue* (dai più grandi ai più piccoli) dell'analisi fattoriale applicata alla matrice di correlazione tra gli item per verificare se è presente un primo fattore determinante.
  - Confronto dei *plot* degli *eigenvalue* di due matrici di correlazione tra item: una con dati reali e l'altra con dati casuali (con distribuzione normale, stessa dimensione campionaria e stesso numero di variabili dei dati osservati). Se nei dati osservati l'assunto di unidimensionalità è soddisfatto, i due grafici dovrebbero essere sostanzialmente simili.
  - Verifica dell'assunto di indipendenza locale investigando le matrici di covarianza o di correlazione calcolate sui soggetti raggruppati secondo diversi valori di capacità. L'assunto di unidimensionalità sarà soddisfatto (anche approssimativamente) se i valori al di fuori della diagonale sono piccoli e vicini a zero.
  - Adattamento al modello di analisi mono-fattoriale non-lineare della matrice di correlazione tra gli item e analisi dei residui.
  - Uso di un metodo di analisi fattoriale basato direttamente sull'*IRT*. Per spiegare il vettore delle risposte si assume una versione multidimensionale del modello con tre parametri basato sull'ogiva normale. La stima dei parametri del modello è complicata e richiede molto tempo ma i risultati ottenuti possono essere promettenti. Di particolare interesse è l'adattamento di una soluzione unidimensionale ai dati.
  - Verifica della presenza di particolari item che violano l'assunto per vedere se "funzionano" in maniera diversa. I valori *b* per tali item sono calibrati separatamente in sotto-gruppi e nel gruppo totale di item. Il contesto della taratura dell'item diviene trascurabile se gli assunti del modello vengono soddisfatti. Se il grafico dei valori di *b* tarati nei due contesti è lineare con una dispersione confrontabile con gli errori standard associati con le stime dei parametri dell'item, l'assunto di unidimensionalità è soddisfatto.
- **Indici di Equi-Discriminazione**
  - Analisi delle correlazioni (biseriali e punto-biseriali) tra item e punteggi totali secondo il modello classico di *item analysis*. Quando le correlazioni sono ragionevolmente omogenee, è possibile procedere alla selezione di

[24] L'adattamento dei risultati al modello è detto *fit* mentre il mancato adattamento è detto *misfit*.

un modello che assume item equidiscriminanti.

- **Minima Possibilità di Rispondere A Caso**
  - Verifica delle risposte dei soggetti con bassa capacità agli item più difficili; se i livelli di esecuzione sono vicini a zero, l'assunto è soddisfatto.
  - Verifica attraverso il *plot* delle regressioni dei punteggi item-strumento. L'assunto sarà soddisfatto quando gli item mostreranno *performance* vicino a zero per soggetti con punteggi bassi.
  - Analisi del livello di difficoltà dello strumento, dei limiti di tempo e della struttura degli item per verificare l'eventuale presenza della possibilità di rispondere a caso.
- **Altre particolari verifiche**
  - Rapporto tra la varianza del gruppo di item omessi e la varianza del gruppo di item con risposte scorrette; l'assunto è soddisfatto quando il rapporto è vicino a zero.
  - Confronto tra i punteggi dei soggetti cui si sono dati limiti di tempo specificati e quelli dei soggetti senza limiti di tempo. Grosse sovrapposizioni nell'esecuzione indicano che l'assunto è soddisfatto.
  - Verifica della percentuale di soggetti che completano l'esecuzione, la percentuale di soggetti che completano il 75% e il numero di item completati dall'80% dei soggetti. Se quasi tutti i soggetti completano quasi tutti gli item, la velocità di esecuzione non rappresenta un fattore importante.

4.3.2.2 Verifica dell'invarianza

I metodi di verifica dell'invarianza dei parametri del modello sono molti e, in ogni caso, piuttosto semplici. L'invarianza può essere studiata somministrando ai soggetti due o più gruppi di item; gli item di ciascun gruppo variano molto in difficoltà relativamente alla dimensione misurata. Per ciascun soggetto e per ogni gruppo di item vengono determinate le stime di capacità. Le due serie di stime di capacità per tutti i soggetti vengono messe a confronto attraverso un grafico. Se l'assunto che il punteggio atteso di capacità di ciascun soggetto è indipendente dagli item scelti è soddisfatto, tale grafico dovrebbe definire una retta con pendenza 1. L'errore di misurazione può produrre qualche dispersione dei punti intorno a tale retta. Il modello di *IRT* può non essere considerato adatto ai dati a disposizione quando non è possibile osservare una relazione lineare (con pendenza=1 e intercetta=0) o la dispersione eccede quella attesa dalla conoscenza degli errori standard delle stime di capacità.

1. **Invarianza delle stime dei parametri di capacità**. Confronto delle stime di capacità per diversi campioni di item (per identificare item difficili e facili, item che riflettono diverse categorie di contenuto all'interno del gruppo di item definiti). L'invarianza è soddisfatta quando le stime non differiscono molto tra loro.
2. **Invarianza delle stime dei parametri di item**. Confronto tra i modelli di stima di parametri degli item (valori *b*, valori *a* e/o valori *c*) ottenuti in due o più sottogruppi della popolazione cui è rivolto lo strumento. Quando le stime sono invarianti, il *plot* dovrebbe essere lineare con scarti che riflettono solo errori di campionamento.

4.3.2.3 Verifica delle previsioni del modello

La successiva verifica riguarda l'adattamento (*fit*) dei risultati al modello; tale verifica si basa su alcune considerazioni: se i valori di capacità dei soggetti e i valori di difficoltà degli item si estendono approssimativamente sullo stesso intervallo (sufficientemente ampio) della variabile da misurare allora sarà possibile tracciare una curva simile a quella rappresentata dall'*ICC*. Lo scostamento dei punti osservati da questa curva consente di determinare il grado di incoerenza rispetto al modello. La procedura seguita per tale verifica è iterativa con affinamenti successivi che depurano i dati da quelli incoerenti ovvero da quelli che, non adattandosi al modello, presentano un alto valore di *misfit*.
La valutazione dell'adattamento viene fatta attraverso la verifica delle previsioni del modello applicato. Di seguito vediamo quali sono le tecniche più utilizzate.

- Analisi dei residui e dei residui standardizzati dell'adattamento del modello ai dati.
- Confronti delle distribuzioni dei punteggi osservati e predetti ottenute assumendo corretti tutte

le stime dei parametri del modello. L'analisi dei risultati è fatto con metodi grafici o applicando particolari statistiche come il *chi-quadro*.

- Analisi degli effetti di particolari condizioni sperimentali come la posizione degli item, la velocità, noia, istruzioni, ecc.
- Diagramma di dispersione delle stime di capacità e dei corrispondenti punteggi. L'adattamento è accettabile quando la relazione si presenta forte con una dispersione intorno alla *ICC* che riflette l'errore di misurazione.
- Applicazione di una miriade di test statistici per determinare l'adattamento totale del modello, dell'item e individuale.
- Confronti dei parametri veri e stimati degli item e delle capacità con l'ausilio di metodi di simulazione computerizzati.
- Analisi della "robustezza" del modello attraverso l'uso di metodi di simulazione; per esempio è possibile studiare le implicazioni dell'adattamento dei modelli *IRT* unidimensionali a dati multidimensionali.

Tra i diversi approcci il più utilizzato è sicuramento il metodo che verifica le previsioni del modello attraverso l'analisi dei residui (detti anche scarti/errori e a volte "residui *raw*") degli item; vediamo come lo studio dei residui e/o dei residui standardizzati possa fornire preziose informazioni nella scelta del modello di *IRT*; dopo aver:

- scelto il modello di *IRT*,
- stimato i parametri di item e di capacità,
- calcolato le previsioni riguardanti le esecuzioni di vari gruppi di capacità, assumendo la validità del modello scelto,

si confrontano i risultati previsti con quelli osservati. Un residuo r_{ij} rappresenta la differenza tra l'esecuzione osservata e l'esecuzione attesa dell'item per un determinato sottogruppo di soggetti:

$$r_{ij} = x_{ij} - P_{ij}$$

dove

i	item
j	categoria dell'item (sottogruppo)
x_{ij}	proporzione osservata di risposte corrette per l'item i e la *j-esima* categoria di capacità
P_{ij}	proporzione attesa di risposte corrette ottenuta utilizzando il modello *IRT* adottato

Se si fa riferimento ad un unico soggetto allora:

x_{ij}	risposta data dal soggetto j all'item i
P_{ij}	corrispondente probabilità di risposta come prevista dal modello.

Per determinare la *proporzione attesa di risposte corrette* (P_{ij}) in ciascuna categoria di capacità si utilizzano le stime dei parametri del modello ipotizzato. In particolare si determina il valore *d* utilizzando il punto centrale della categoria di capacità come valore di capacità rappresentativo per la categoria; si calcola quindi la probabilità di una risposta corretta utilizzando tale valore.

Per identificare le *categorie di capacità*, il continuum corrispondente viene di solito suddiviso in intervalli di uguale dimensione (da 10 a 15). Gli intervalli non dovrebbero essere troppo piccoli per evitare che le statistiche risentano dell'instabilità soprattutto in presenza di piccoli campioni. D'altra parte, gli intervalli dovrebbero essere definiti in modo tale che i soggetti in ciascuna categoria risultino omogenei in termini di capacità.

L'uso degli scarti è limitato dal fatto che non tiene conto dell'errore campionario associato con il punteggio atteso all'interno di ciascuna categoria di capacità. Per tenere conto di tale errore campionario viene calcolato il *residuo standardizzato* (o *scarto standard*, z_{ij}) ottenuto dividendo i residui per i loro errori standard.

In pratica lo *scarto standard* serve per definire il calcolo del coefficiente di *misfit* (*t*):

$$z_{ij} = \frac{x_{ij} - P_{ij}}{\sqrt{\frac{P_{ij}(1-P)}{N_j}}}$$

dove

N_j numero di soggetti nella categoria di capacità *j*

$P_{ij}(1-P_{ij})$ varianza.

Quanto più i dati si approssimano al modello tanto più lo scarto standard sarà distribuito normalmente, con media 0 e varianza 1. Quindi per analizzare l'adattamento è possibile esaminare quanto tali scarti si approssimano ad una distribuzione normale oppure quanto i loro quadrati z_{ij}^2 si approssimano ad una distribuzione di *chi-quadro* con un grado di libertà[25].

Quanto maggiore è la differenza tra la capacità del soggetto e la difficoltà dell'item, tanto maggiore risulta lo scarto quadrato che indica l'inaccettabilità di una risposta sbagliata (con $d_j > b_i$) o di una risposta corretta (con $d_j < b_i$).

Di solito per verificare l'adattamento del modello vengono applicati anche test statistici, in genere il Q_1, statistica derivata dal *chi-quadro*. Il Q_1 per l'item *i*

$$Q_{1i} = \sum_{j=1}^{m} \frac{N_j(x_{ij} - P_{ij})^2}{P_{ij}(1-P_{ij})} \quad = \quad \sum_{j=1}^{m} z_{ij}^2$$

La statistica Q_1 è distribuita come il *chi-quadro* con $m-k$ gradi di libertà, dove *k* è il numero di parametri presenti nel modello *IRT* ed *m* è il numero delle categorie in cui è stato suddiviso il continuum. Se il valore osservato della statistica supera il valore critico (presentato nella tavola del *chi-quadro*), l'ipotesi nulla che la *ICC* si adatti ai dati viene rifiutata e deve essere trovato un nuovo modello di adattamento. E' possibile inoltre determinare:

- il grado di *misfit* di un soggetto sommando i quadrati degli scarti del soggetto *j* per tutti gli item *i* (z_j^2),
- il grado di *misfit* di un item sommando i quadrati degli scarti dell'item *i* per tutti i soggetti *j* (z_i^2).

Le medie normalizzate delle quantità z_j^2 e z_i^2 sono i due coefficienti *misfit* t_j e t_i.

Entrambi i coefficienti possono essere calcolati come di seguito:

- <u>*misfit* della capacità del soggetto</u>: $t_j = [\log_n(v_j) + v_j - 1] * [(n-1)/8]^{1/2}$

  dove

  n numero di item

  v_j $z_j^2/(n-1)$

- <u>*misfit* della difficoltà dell'item</u>: $t_i = [\log_n(v_i) + v_i - 1] * [(N-1)/8]^{1/2}$

  dove

  N numero di soggetti

  v_i $z_i^2/(N-1)$.

Non sempre nelle applicazioni si studia e si valuta in modo adeguato l'adattamento modello-dati e le conseguenze dei *misfit*. A causa delle poche applicazioni, non si conosce molto sul livello di adeguatezza dei modelli di *IRT*. Le poche applicazioni spesso utilizzano statistiche non appropriate allo studio della bontà di adattamento; conseguentemente le conclusioni risultano poco affidabili. In genere tali applicazioni hanno utilizzato il modello della *verifica dell'ipotesi* i cui test sono, come si sa, troppo sensibili alla dimensione del campione dei soggetti: con campioni di grosse dimensioni, è molto più probabile giungere alla conclusione di mancanza di adattamento modello-dati anche nei

[25] Nel caso in cui le risposte siano binarie è possibile ottenere per z_{ij}^2 una semplificazione dell'espressione: $z_{ij}^2 = \exp|d_j - b_i|$

casi di piccolo allontanamento dal modello. D'altra parte l'utilizzo di piccoli campioni non consente una significativa e utile stima di parametri a causa della presenza di grossi errori standard. A ciò va aggiunto il fatto che le distribuzioni campionarie di alcune statistiche di bontà di adattamento utilizzate non sono sempre note con precisione.
Vediamo di seguito la sintesi del procedimento descritto di verifica delle previsioni del modello.

VALUTAZIONE E VERIFICA DELL'ADATTAMENTO DEL MODELLO

⇓

VERIFICA DELL'ACCURATEZZA DELLE PREVISIONI DEL MODELLO APPLICATO

⇓

VALUTAZIONE DEL MISFIT

⇓

procedura iterativa:

1. Analisi dei residui (scarti/errori o residui raw) degli item

   Dopo aver:
   - scelto il modello di IRT,
   - stimato i parametri di item e di capacità,

   si confrontano i risultati previsti con quelli osservati.

   Residuo = (esecuzione osservata) – (esecuzione attesa)

   ↓

   $$r_{ij} = x_{ij} - P_{ij}$$

   Quanto più gli scarti approssimano la distribuzione normale tanto più i dati approssimano il modello

2. Residuo standardizzato

   scarto standard = residuo / errore standard

   $$z_{ij} = \frac{x_{ij} - P_{ij}}{\left[P_{ij}(1-P)\right]^{1/2}} \qquad \text{oppure} \quad z_{ij} = \frac{x_{ij} - P_{ij}}{\sqrt{\frac{P_{ij}(1-P)}{N_j}}}$$

3. Coefficiente di *misfit*
   - *Misfit* della capacità del soggetto:

     z_j^2 = somma dei quadrati degli scarti del soggetto *j* per tutti gli item

     media normalizzata di z_j^2 = $t_j = \left[\log_n(v_j) + v_j - 1\right] * \left[(n-1)/8\right]^{1/2}$
   - *Misfit* della difficoltà dell'item:

     z_i^2 = somma dei quadrati degli scarti dell'item *i* per tutti i soggetti

     media normalizzata di z_i^2 = $t_i = \left[\log_n(v_i) + v_i - 1\right] * \left[(N-1)/8\right]^{1/2}$

4. Test di verifica dell'adattamento del modello

   Per l'item *i* $Q_{1i} = \sum_{j=1}^{m} \frac{N_j\left(x_{ij} - P_{ij}\right)^2}{P_{ij}\left(1 - P_{ij}\right)}$ = $\sum_{j=1}^{m} z_{ij}^2$

   Distribuita come il chi-quadro con m-k gradi di libertà

   Se Q_{1i} > valore critico → adattamento mancato

dove

n	numero di item
N	numero di soggetti
N_j	numero di soggetti nella categoria di capacità *j*
P_{ij}	probabilità di risposta del soggetto *j* all'item *i* oppure proporzione attesa (secondo il modello) di risposte corrette per l'item i nella categoria di capacità *j*
$P_{ij}(1-P_{ij})$	varianza
k	numero di parametri presenti nel modello IRT
v_j	$z_j^2/(n-1)$

v_i	$z_i^2/(n-1)$
m	numero di categorie in cui è stato suddiviso il continuum
x_{ij}	risposta data dal soggetto j all'item i oppure proporzione osservata di risposte corrette per l'item *i* nella categoria di capacità *j*

4.3.2.4 Le funzioni informative

Secondo il modello probabilistico, ciascun item può essere caratterizzato e definito oltre che dai parametri visti, anche da una particolare funzione detta *funzione informativa* (*item information function*):

$$I_i(d) = \frac{[P'_i(d)]^2}{P_i(d)Q_i(d)} \qquad i = 1, 2, ..., n$$

dove

n	numero totale di item
d	capacità che influenza la risposta del soggetto ad uno strumento
i	item
$I_i(d)$	informazione fornita dall'item *i* alla capacità *d*
$P_i(d)$	*IRF*, *item response function*
$P'_i(d)$	derivata di P_i *(d)* rispetto a *d*
$Q_i(d)$	$1 - P_i(d)$

Tale equazione è applicabile ai modelli logistici (con uno, due o tre parametri). Nel caso del modello logistico con tre parametri essa è modificata nel modo seguente:

$$I_i(d) = \frac{2.89 \cdot a_i^2(1-c_i)}{\left(c_i + e^{1.7a_i(d-b_i)}\right)\cdot\left(1+e^{-1.7a_i(d-b_i)}\right)^2}$$

dove *a*, *b*, *c*, *d* sono naturalmente, rispettivamente, i parametri di discriminazione, difficoltà e del livello *pseudo-casuale* del'item e di capacità del soggetto.

Da tale equazione è possibile dedurre il ruolo che i parametri *b*, *a* e *c* hanno nella funzione informativa dell'item:

- l'informazione è maggiore quando il valore *b* è vicino a *d* di quando il valore *b* è lontano da *d*,
- l'informazione è generalmente maggiore quando il valore del parametro *a* è alto,
- l'informazione aumenta all'avvicinarsi a zero del parametro *c*.

Le funzioni informative dell'item, consentendo una valutazione dei singoli item, possono giocare un ruolo importante nello sviluppo di uno strumento con misure multiple, in quanto indicano il contributo che ogni item dà alla stima della capacità in punti diversi lungo il continuum di capacità. Tale contributo dipende in gran parte dalla forza di discriminazione di un item (pendenza di P_i); la posizione in cui tale contributo sarà realizzato è dipendente dalla difficoltà dell'item.

Dato che, in genere, il valore delle funzioni informative è minore quando $c>0$ di quando $c=0$, si potrebbe essere tentati di adottare i modelli con uno o due parametri. I valori delle funzioni informative che ne risultano saranno più grandi; comunque le curve relative alle funzioni informative per i modelli con uno o due parametri saranno utili solo quando le *ICC* da cui sono derivate si adattano ai dati. L'uso di *ICC* che non si adattano adeguatamente conducono a risultati scorretti.

In pratica, le funzioni informative consentono di selezionare quegli item che presentano il livello di informazione necessario per soddisfare le specifiche esigenze dello strumento che si intende costruire[26].

[26] Il modello probabilistico si presenta particolarmente adatto alla selezione di item per la misurazione di capacità; il metodo che consente di selezionare gli item è particolarmente potente grazie alle specifiche caratteristiche del modello quali

- l'invarianza dei parametri degli item, che consente di utilizzare campioni diversi per la messa a punto;
- la possibilità di misurare sulla stessa scala sia i parametri degli item che la capacità dei soggetti;

A tale proposito, Lord nel 1977 ha definito una procedura che consente di utilizzare le funzioni degli item come

Le funzioni informative di più item presentano la caratteristica di essere additive; quindi è possibile determinare una funzione informativa relativa ad un gruppo di item nel modo seguente:

$$I(d) = \sum_{i=1}^{n} I_i(d)$$

In altre parole gli item contribuiscono in modo indipendente alla funzione informativa del gruppo. Come abbiamo visto questa caratteristica non è riscontrabile in altri modelli di *scaling* per i quali non è possibile determinare per ciascun item le caratteristiche e il contributo all'affidabilità totale in modo indipendente dalle caratteristiche di tutti gli altri item.
Infine ricordiamo che la quantità di informazione fornita è inversamente correlata alla precisione con cui la capacità è stimata:

$$SE(\hat{d}) = \frac{1}{\sqrt{I(d)}}$$

dove
$SE(\hat{d})$ errore standard di stima

All'interno dell'*IRT*, l'errore standard di $\hat{d}$, $SE(\hat{d})$, che rappresenta la deviazione standard della distribuzione normale asintotica della stima di massima verosimiglianza della capacità per un dato valore vero di capacità *d*, svolge lo stesso ruolo dell'errore standard di misurazione nella teoria classica di misurazione.
Per poter confrontare più funzioni informative relative al la stessa capacità è possibile calcolare l'indice di efficienza relativa:

$$RE(d) = \frac{I_A(d)}{I_B(d)}$$

dove
$RE(d)$ efficienza relativa
$I_A(d)$ funzione informativa dello strumento *A*
$I_B(d)$ funzione informativa della strumento *B*
Se per esempio
$I_A(d) = 25.0$

---

elementi informativi per costruire scale che soddisfino le specificazioni considerate. Tale procedura impiega una banca di item, per ciascuno dei quali sono disponibili sia le stime dei parametri degli item secondo il modello prescelto che le funzioni informative. I passaggi suggeriti per tale procedura sono i seguenti:

1. definizione della forma della funzione informativa desiderata (*target information function*, *TIF*); per costruire una scala che
   - misura una capacità con un ampio *range*, la *TIF* dovrebbe essere abbastanza piatta; ciò consente di costruire una scala che presenta una precisione di stima di capacità uniforme per tutto il *range*;
   - richiede la definizione di un *cut-off* sulla scala di capacità che consenta di separare i soggetti misurati in due gruppi, la *TIF* dovrebbe essere molto appuntita vicino al punteggio di *cut-off*.
2. selezione dalla banca degli item con le funzioni informative che interessano;
3. calcolo della funzione informativa dell'intera scala;
4. ripetizione del precedente passaggio, eliminando item o inserendone altri fino al raggiungimento della funzione informativa di scala che più si avvicina alla *TIF*.

La procedura suggerita da Lord consente di costruire una scala che discrimina bene in una particolare regione del continuum di capacità; questo vuol dire che gli item possono essere selezionati in modo da massimizzare l'informazione di scala nella regione della capacità per i soggetti che devono essere misurati. E' il caso della costruzione di una scala che misura un rendimento; una tale scala dovrebbe avere item più semplici, ovvero che misurano *performance* più basse, quando utilizzata nel *pretest* e item più difficili, ovvero che misurano *performance* più alte, nel *post-test*. In ciascuna occasione, la precisione di misurazione sarà massimizzata nella regione di capacità dove i soggetti più probabilmente dovrebbero posizionarsi.
Ricordiamo però che il solo uso di criteri statistici per la selezione di item non assicura una scala con validità di contenuto; la sopravvalutazione dei criteri statistici è facile e comoda ma produce la conseguenza di non tenere in giusto conto il ruolo che il contenuto degli item ha nello sviluppo di una scala.

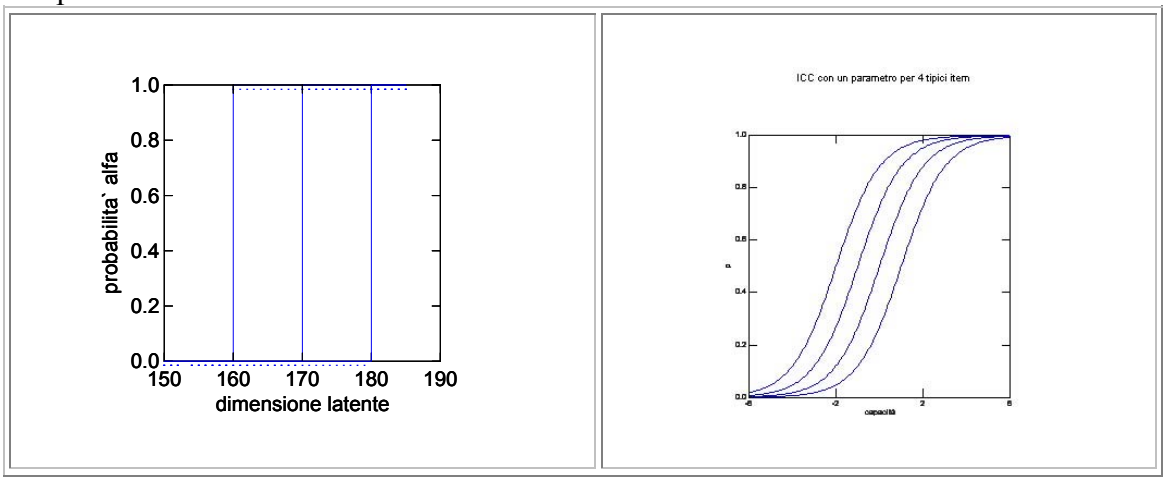
$I_B(d) = 20.0$
allora $RE(d) = 1.25$
ovvero a livello *d* lo strumento *A* funziona come se fosse 25% più lungo dello strumento *B*; in altre parole
- *B* dovrebbe essere allungato del 25% per produrre la stessa precisione di misurazione di *A* a livello *d*;
- *A* potrebbe essere accorciato del 20% per continuare a produrre stime di capacità a livello *d* possedendo la stessa precisione di *B*.

Tali conclusioni, riguardati l'allungamento e la riduzione, sono basate sull'assunto che gli item aggiunti o eliminati siano confrontabili nella loro qualità statistica agli altri item. In altre parole, l'efficienza relativa svolge un ruolo molto simile a quello svolto dal coefficiente Spearman-Brown nel modello additivo.

4.4 CONSIDERAZIONI FINALI

4.4.1 Confronto tra modelli deterministici e probabilistici

Le due seguenti figure presentano abbastanza chiaramente la differenza tra le due versioni di una scala cumulativa e in particolare i limiti del modello deterministico (tre item) rispetto a quello probabilistico (quattro item), soprattutto se si pensa che praticamente mai i modelli di *scaling* adattano perfettamente i dati.



Come appare evidente mentre nel caso del modello deterministico non è prevista una risposta negativa ad un item che ricade a sinistra del punto-soggetto, nel caso del modello probabilistico la probabilità di una risposta positiva a ciascun item da parte di un soggetto aumenta via via che il punto-soggetto si sposta verso le posizioni più alte (a destra) lungo la dimensione.
Una delle riflessioni che si possono fare a questo punto riguarda la relazione tra modello e dati: quanto più il ricercatore è in grado di rendere sempre più espliciti gli assunti sulla natura degli errori quanto più diviene possibile accettare grandi discrepanze tra i modelli di *scaling* perfetto, definiti a livello di modello, e gli stessi dati empirici. Sembrerebbe quindi che le versioni probabilistiche del modello cumulativo siano non solo più realistiche dell'approccio deterministico, ma più adatte a costruire scale anche a partire da dati che presentano errori.

4.4.2 Modelli cumulativi e multidimensionalità

Uno dei problemi che pongono le procedure di *scaling* basate sul modello cumulativo è dato dal fatto che non sono in grado di rilevare la struttura dimensionale sottostante un insieme di dati, ovvero nel caso in cui vi siano più fonti di variabilità.

Le misure di bontà di adattamento utilizzate per verificare il modello cumulativo possono indicare quando una dimensione singola risulta essere inadeguata a rappresentare i dati. Comunque esse non possono essere utilizzate per distinguere tra le diverse ragioni di un tale risultato (perdita della struttura, errore di misurazione, struttura dimensionale più complessa).
In generale gli approcci cumulativi funzionano meglio quando vengono applicati ai dati che si adattano un modello dimensionale nel quale entrambi gli insiemi di oggetti variano in modo sistematico rispetto ad un unico attributo.
Se il ricercatore ha il sospetto che:

- solamente uno dei due insiemi contiene una variazione sistematica,
- vi sono molte fonti di variabilità in un insieme di osservazioni,

può essere più proficuo orientarsi verso un altro approccio di *scaling*.
A tale proposito abbiamo visto, comunque, come sia i modelli deterministici che quelli probabilistici presentano vari tentativi di definire anche modelli multidimensionali.

5. CONFRONTO TRA MODELLI DI SCALING.
LA VALIDAZIONE DI UNA SCALA DI AUTOVALUZIONE DELL'AUTOSUFFICIENZA FISICA IN UNA POPOLAZIONE ANZIANA

L'esempio qui presentato riguarda la validazione di una particolare scala di autovalutazione dell'autosufficienza fisica negli anziani. Tale validazione ha riguardato un campione piuttosto ampio di anziani residenti in Toscana e in Molise, per un totale di 3389 soggetti[1]. L'obiettivo era quello di mettere a punto uno strumento che consentisse di misurare non tanto l'efficienza fisica oggettivamente valutata quanto il livello di autosufficienza nell'ambiente di vita e nelle attività quotidiane così come vissuto e percepito dall'anziano.

5.1 LA SCALA

La scala, messa a punto integrando gli item proposti dall'*Organizzazione Mondiale della Sanità* (1979), può essere considerata un test di *performance* per le capacità fisiche richieste nello svolgimento delle attività di base della vita quotidiana (batteria di item funzionali). Queste capacità possono essere fatte risalire a un unico fattore di idoneità fisica (*fitness*) a condurre una vita "piena". Per l'anziano, presumibilmente ritirato dal lavoro, è possibile restringere il significato della scala al *possesso di capacità fisiche* per lo svolgimento delle attività quotidiane; le capacità fisiche si manifestano nello svolgere, in piena autonomia, le funzioni essenziali di:

- cura della persona, che si identifica con l'igiene personale (lavarsi, fare il bagno, pettinarsi, farsi la barba);
- cura del proprio abbigliamento, che nella società moderna consiste soprattutto nel vestirsi e spogliarsi, in quanto la vera e propria cura degli abiti rappresenta un elemento di delega anche in situazioni di piena autonomia funzionale;
- preparazione del cibo: gli elementi connessi con quest'area possono essere delegati, ma l'incapacità completa fisica di prepararsi in caso di necessità un pasto caldo può rappresentare un grave disagio esistenziale;
- cura della casa: anche le faccende domestiche - pesanti e non - sono elementi delegabili il cui svolgimento è legato a fattori culturali e di genere; ma, come per la preparazione del cibo, l'incapacità a svolgerle diventa "handicap" in caso di perdita del familiare o della persona incaricata di queste funzioni;
- approvvigionamento dei generi di prima necessità, la "spesa": rappresenta un elemento a sé, per il quale valgono le stesse considerazioni fatte ai punti precedenti.

Presupposto di base, per ciascuna di queste capacità, è la maggiore o minore sicurezza nello spostarsi in casa e fuori nell'ambiente esterno con i propri mezzi, o con quelli del trasporto pubblico, il cui uso richiede una dose aggiuntiva di capacità fisica. Sono tutte capacità indispensabili per le necessità della vita quotidiana, da prendere in considerazione in assenza e in presenza di ostacoli (come pavimenti sconnessi, soglie, gradini, scale più o meno lunghe, isolamento geografico). Questi spostamenti sono stati visti anche come indicatori della capacità di utilizzare mezzi pubblici, condizione che rende possibili normali scambi e incontri sociali con familiari ed amici.
La costruzione e la validazione di un tale tipo di scala ha l'obiettivo di:

- costruire un indicatore sintetico di autosufficienza,
- rilevare e valutare con un'accuratezza nota i soggetti che presentano un livello critico di disabilità; ciò viene fatto suddividendo il punteggio in livelli critici (valori-soglia) che,

[1] Tale validazione ha coinvolto più ricerche. Per maggiori informazioni sulla metodologia di ricerca seguita vedi Tesi G., Antonini E., Maggino F., "La scala di autosufficienza", in *Gli anziani a casa. Uno studio a Dicomano*, 1993, Comune di Dicomano, Università di Firenze, Provincia di Firenze, USL 11, I.N.R.C.A.

facendo riferimento al carico assistenziale (tipo e frequenza dell'aiuto richiesto), consentono l'individuazione e la definizione di diverse aree di non autosufficienza fisica.

Gli item sono stati scelti in modo da coprire tutti gli aspetti di base dell'attività quotidiana; successivamente gli item sono stati selezionati e ordinati secondo il modello scalare legato al criterio della diversa difficoltà; ciò consente di ottenere una valutazione in cui le condizioni per superare un item sono "teoricamente" richieste per superare l'item precedente. Ogni item deve avere la capacità di dividere il gruppo in due sezioni, e tutti gli individui che hanno superato item più difficili si devono trovare compresi nel gruppo che ha superato l'item più semplice. In altre parole alla base della costruzione della scala vi è quindi un modello scalare di tipo cumulativo che, come sappiamo, richiede una verifica basata sullo scalogramma di Guttman. Una delle più interessanti applicazioni del modello deterministico, infatti, è quella che consente la costruzione di una scala per la misurazione di capacità fisiche quando è possibile ipotizzare che le abilità per le funzioni di ogni giorno dipendono da capacità comuni, individuabili nei vari aspetti dell'attività quotidiana. In genere gli scalogrammi utilizzati nelle indagini sulle capacità fisiche vengono costruiti con significato

- positivo: possesso delle capacità fisiche, capacità di accudire a se stessi o
- negativo: disabilità, dipendenza nello svolgimento delle attività quotidiane.

In una scala con significato positivo si attribuisce il punteggio alto alla risposta che denota il superamento della prova.

In questa sede si è proceduto alla verifica dell'adattamento non solo del modello additivo-deterministico ma anche di altri. Ciò ha consentito di mettere in rilevo particolari aspetti dell'attributo misurato e, soprattutto, della scala utilizzata; in particolare; schematicamente, l'analisi ha riguardato:

Verifica della	Attraverso la verifica del livello di adattamento del modello
➢ omogeneità degli item	additivo
➢ cumulabilità (scalabilità)	deterministico
➢ dimensionalità	- fattoriale - *multidimensional scaling*
➢ scalabilità multidimensionale	POSAC

Gli item individuati sono i seguenti (in ordine di somministrazione):

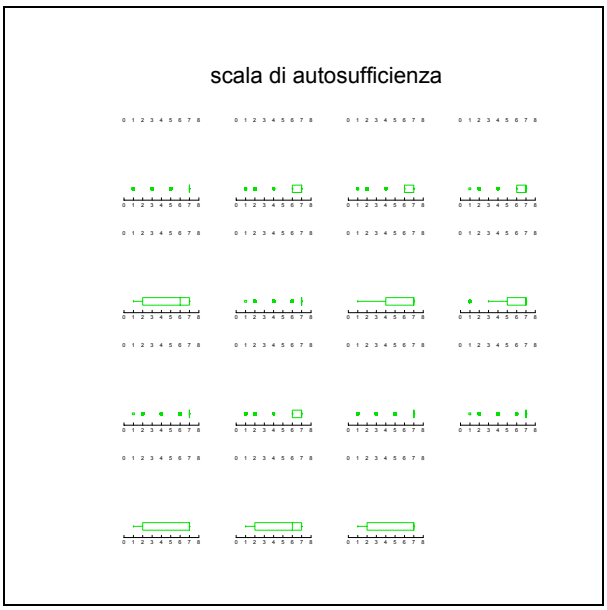
ITEM	Livelli di capacità misurati per ciascun item				
	1	2	3	4	5
	Senza difficoltà	Con difficoltà ma senza aiuto	Con aiuto per parte dell'azione	Con aiuto totale	Non lo fa per abitudine
1. Spostarsi per le stanze					
2. Uscire di casa					
3. Fare le scale					
4. Camminare per almeno 400 m					
5. Fare la spesa					
6. Lavarsi viso e braccia					
7. Fare il bagno o la doccia					
8. Vestirsi e spogliarsi					
9. Mangiare da solo					
10. Prepararsi un pasto caldo					
11. Usare il gabinetto					
12. Alzarsi e andare a letto					
13. Lavori domestici leggeri					
14. Lavori domestici pesanti					
15. Tagliarsi le unghie dei piedi					
	In alcuni item è stato espresso il codice 5 per identificare l'incapacità di eseguire una determinata attività per abitudine, tradizione familiare, culturale, ecc., e non per incapacità fisica (si pensi ai lavori domestici per gli uomini)				

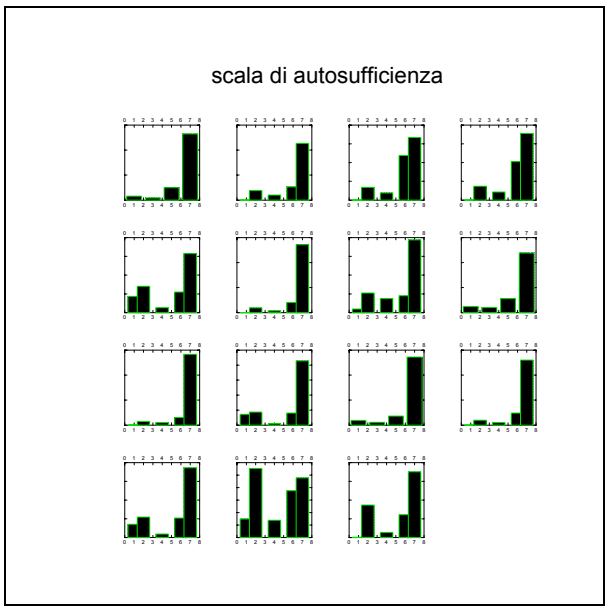
5.2 L'ANALISI

5.2.1 Analisi descrittiva dei singoli item

Di seguito possiamo osservare la distribuzione di frequenza relativa ai 15 item di autosufficienza.

ITEM	Livelli di capacità misurati per ciascun item					
	Senza difficoltà	Con difficoltà ma senza aiuto	Con aiuto per parte dell'azione	Con aiuto totale	Non lo fa per abitudine	Totale
1. Spostarsi per le stanze	2646	497	89	157		3389
2. Uscire di casa	2270	531	201	381	6	3389
3. Fare le scale	1669	1187	191	339	3	3389
4. Camminare per almeno 400 m	1777	1024	217	370	1	3389
5. Fare la spesa	1574	548	129	704	434	3389
6. Lavarsi viso e braccia	2720	408	73	187	1	3389
7. Fare il bagno o la doccia	1939	450	379	524	97	3389
8. Vestirsi e spogliarsi	2384	557	199	248		3389
9. Mangiare da solo	2831	308	107	142	1	3389
10. Prepararsi un pasto caldo	2140	403	53	435	358	3389
11. Usare il gabinetto	2727	363	112	187		3389
12. Alzarsi e andare a letto	2604	482	108	194	1	3389
13. Lavori domestici leggeri	1863	523	96	549	358	3389
14. Lavori domestici pesanti	955	752	277	1106	299	3389
15. Tagliarsi le unghie dei piedi	1757	613	139	869	11	3389





Osservando le distribuzioni di frequenza è possibile fare le prime valutazioni sulle performance rilevate secondo le affermazioni degli anziani. Osserviamo subito un risultato confortante per la salute degli anziani intervistati: la concentrazione delle risposte sui valori alti degli item indica una generale e positiva tendenza a svolgere le diverse attività senza difficoltà; tale concentrazione non appare però uniforme, ad indicare una possibile differenziazione degli item lungo il continuum di autosufficienza; inoltre, per alcuni item, si osservano distribuzioni con una forma irregolare di difficile interpretazione, in particolare per gli item 13, 14, 15.

5.2.2 Verifica dell'omogeneità

La definizione di autosufficienza qui adottata assume che tale caratteristica sia unidimensionale; conseguentemente la scala costruita assume che il punteggio totale sia monotonamente legato con la dimensione misurata. Per verificare ciò è necessario analizzare innanzitutto la consistenza interna del gruppo di item individuati. L'analisi della consistenza interna ha prodotto risultati piuttosto soddisfacenti; confrontando però i valori dello *split-half* per gli item suddivisi secondo due diversi metodi, *1ª metà-2ª metà* e *odd-even*, osserviamo subito come l'ordine degli item riflette in qualche modo una differenza tra gli item; infatti il primo metodo registra valori più bassi rivelando una diversa risposta dei soggetti ai due gruppi di domande.

INTERNAL CONSISTENCY DATA		
	HALF1-HALF2	ODD-EVEN
SPLIT-HALF CORRELATION	0.869	0.936
SPEARMAN-BROWN COEFFICIENT	0.930	0.967
GUTTMAN (RULON) COEFFICIENTE	0.926	0.960
COEFFICIENT *ALPHA*- ALL ITEM	0.943	
COEFFICIENT *ALPHA*	0.926 - 0.858	0.891 - 0.883

Item	Item-Total R	Item-reliability Index	Item-Total R Excl Item	Alpha Excl Item
1	0.820	0.611	0.799	0.939
2	0.859	0.880	0.834	0.936
3	0.857	0.814	0.834	0.937
4	0.851	0.838	0.826	0.937
5	0.706	1.082	0.635	0.943
6	0.805	0.623	0.781	0.939
7	0.785	0.979	0.741	0.938
8	0.862	0.775	0.842	0.937
9	0.766	0.552	0.740	0.940
10	0.660	0.958	0.586	0.944
11	0.827	0.652	0.806	0.938
12	0.834	0.671	0.813	0.938
13	0.713	1.049	0.647	0.942
14	0.663	0.926	0.592	0.943
15	0.782	0.993	0.737	0.939

5.2.3 Verifica della scalabilità

5.2.3.1 Verifica attraverso il modello deterministico

L'analisi della consistenza interna, come sappiamo, può condurre ad una valutazione finale ambigua perché uguali punteggi possono essere ottenuti con modelli di risposte (profili) diversi. Ciò vale soprattutto nel caso di item non dicotomici in cui sia importante non solo rilevare le differenze fra gli item ma anche pesare la "distanza", in termini di scalabilità, fra due item contigui. Quando è possibile assumere una gradualità degli item (per superare un item sono necessarie le stesse capacità necessarie per superare un item più facile più un livello di capacità ulteriore) è possibile validare la scala attraverso il confronto della distribuzione reale delle risposte dei soggetti del campione con un modello teorico di perfetta scalabilità tra item e di perfetta predicibilità di una prova sulla prova

seguente Per verificare ciò si procede al confronto ripetuto per tutti gli item; ciò consente una valutazione globale della predicibilità della scala fatta sulla base degli errori osservati nel confronto tra il modello teorico e quanto è stato osservato.
Nel nostro caso il calcolo degli errori di scalabilità (per ogni soggetto e per ogni item) è reso complesso dal fatto che le risposte ai vari item non sono dicotomiche, ma con più livelli (4 o 5) di difficoltà variabili da item a item. Così, per attribuire lo stesso peso ad ogni item, qualunque sia il numero di livelli presenti nelle risposte, è stato assegnato ai due livelli estremi sempre lo stesso valore (1 e 7). Prevedendo un massimo di 7 livelli per ogni item, i punteggi intermedi sono stabiliti secondo un criterio di simmetria. Di seguito è presentato lo schema di ricodifica dei singoli item (i nuovi codici sono in blu).

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	⇐ numero di identificazione degli item
Nuova codifica ⇓															Codici originari ⇓
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	1
5	6	6	6	6	6	6	5	6	6	5	6	6	6	6	2
3	4	4	4	4	4	4	3	4	4	3	4	4	4	4	3
1	2	2	2	2	2	2	1	2	2	1	2	2	2	2	4
	1	1	1	1	1	1		1	1		1	1	1	1	5
4	5	5	5	5	5	5	4	5	5	4	5	5	5	5	⇐ numero di livelli per ciascun item

Per ottenere un ordine gerarchico iniziale degli item è stato definito come *difficoltà crescente* il fatto che l'attività espressa da ogni item viene svolta da una percentuale decrescente di soggetti. La possibilità di trasporre questa progressiva difficoltà degli item così ricavata nella valutazione di singoli soggetti presuppone che in ciascun individuo le capacità fisiche richieste per il superamento di un item siano sufficienti per superare tutti gli item precedenti.
Dopo aver riunito in un unico codice le risposte relative alle modalità *senza difficoltà* e *con difficoltà ma senza aiuto* è stato ottenuto un riordinamento scalare degli item secondo la percentuale decrescente di soggetti che sono in grado di svolgere le attività relative all'item. I due riordinamenti (risposte originali e ricodificate riunendo le due modalità di risposta) coincidono per la maggior parte degli item. Di seguito vediamo una tabella nella quale a ciascun item è associato:

- l'ordine con cui è stato somministrato (colonna *a*),
- l'ordine scalare rispetto alle risposte non ricodificate (colonna *b*) e la frequenza di risposte tipo "1";
- l'ordine scalare rispetto alle risposte ricodificate (colonna *c*) e la frequenza di risposte tipo "1" e "2".

(a)	ITEM	*Senza difficoltà*		*Senza difficoltà e Con difficoltà ma senza aiuto*	
		Frequenza	(b)	Frequenza	(c)
1	1. Spostarsi per le stanze	2646	4	3143	1
2	2. Uscire di casa	2270	7	2801	8
3	3. Fare le scale	1669	13	2856	7
4	4. Camminare per almeno 400 m	1777	11	2801	9
5	5. Fare la spesa	1574	14	2122	14
6	6. Lavarsi viso e braccia	2720	3	3128	3
7	7. Fare il bagno o la doccia	1939	9	2389	11
8	8. Vestirsi e spogliarsi	2384	6	2942	6
9	9. Mangiare da solo	2831	1	3139	2
10	10. Prepararsi un pasto caldo	2140	8	2543	10
11	11. Usare il gabinetto	2727	2	3090	4
12	12. Alzarsi e andare a letto	2604	5	3086	5
13	13. Lavori domestici leggeri	1863	10	2386	12
14	14. Lavori domestici pesanti	955	15	1707	15
15	15. Tagliarsi le unghie dei piedi	1757	12	2370	13

E' stato adottato l'ordinamento ottenuto con la ricodifica (colonna *c*).
Tutti gli item hanno registrato coefficienti di riproducibilità superiori a 0.97 mentre l'intera scala ha prodotto un coefficiente di riproducibilità di 0.991. Vediamo di seguito i risultati nel dettaglio.

	Item	confrontato con	ERRORI 1 punto > 2 punti		RIPRODUCIBILITA' CR_{wi} (min richiesto 0.85)
1.	Spostarsi per le stanze		0	0	1 - 0 / 3389*6 = 1.000
9.	Mangiare da solo	1	662	346	1 - 346 / 3383*6 = 0.980
6.	Lavarsi viso e braccia	9	114	68	1 - 68 / 3387*6 = 0.997
11.	Usare il gabinetto	6	164	25	1 - 25 / 3124*6 = 0.999
12.	Alzarsi e andare a letto	11	607	107	1 - 107 / 3337*6 = 0.995
8.	Vestirsi e spogliarsi	12	125	9	1 - 9 / 3105*6 = 0.999
3.	Fare le scale	8	813	186	1 - 186 / 3355*6 = 0.991
2.	Uscire di casa	3	714	88	1 - 88 / 3384*6 = 0.996
4.	Camminare per almeno 400 m.	2	169	144	1 - 144 / 3355*6 = 0.993
10.	Prepararsi un pasto caldo	4	782	191	1 - 191 / 3387*6 = 0.991
7.	Fare il bagno o la doccia	10	555	404	1 - 404 / 3386*6 = 0.980
13.	Lavori domestici leggeri	7	537	366	1 - 366 / 3366*6 = 0.982
15.	Tagliarsi le unghie	13	578	388	1 - 388 / 3388*6 = 0.981
5.	Fare la spesa	15	451	285	1 - 285 / 3344*6 = 0.986
14.	Lavori domestici pesanti	5	428	268	1 - 268 / 3380*6 = 0.987
			totale= 2875		

Somma frequenze modali = 31856
Somma frequenze non-modali = 18978

$$CR_w \Rightarrow 1 - \frac{n_S}{n * p_{mm}} = 1 - \frac{2875}{3389 * 15 * 6} = 0.991$$

$$CS \Rightarrow 1 - \frac{\sum n_{ie}}{me} = 1 - \frac{2875}{18978} = 0.85$$

$$INP\% \Rightarrow \left[CR - \min(CR_i)\right] * 100 = (0.991 - 0.98) * 100 = 1.2$$

$$MMR \Rightarrow \frac{\sum nm_i}{N} = \frac{31856}{15 * 3389} = 0.63$$

dove

n_s	scarti negativi tra due item contigui
n	numero risposte (o numero di soggetti)
p_{mm}	punteggio_max - punteggio_min
nm_i	numero di risposte nella categoria modale (la più *scelta*) dell'item *i*
N	numero di risposte (*nitem* * *nsogg*)
n_{ie}	numero di errori dell'item *i*
me	errori marginali, somma di tutte le frequenze non-modali.

A questo punto è stato calcolato il coefficiente di predicibilità (*CP*) e il numero di previsioni realizzate per ciascun soggetto, ottenendo i seguenti risultati[2]:

[2] Ricordiamo che

$$CP_j = \sum \frac{pr_j}{pp_i}$$

dove

pr	previsioni realizzate

Coefficiente di predicibilità	Previsioni realizzate	Numero di soggetti
0.40	6	2
0.46	7	7
0.53	8	60
0.60	9	186
0.66	10	374
0.73	11	694
0.80	12	614
0.86	13	538
0.93	14	914

Gli item analizzati si sono rivelati nella loro applicazione pratica una scala con buone caratteristiche di scalabilità e di predicibilità.

5.2.3.2 Verifica attraverso il modello probabilistico

Come sappiamo il concetto di cumulabilità è presente anche nell'approccio probabilistico basato sull'*Item Response Theory*.
Proviamo a verificare se l'applicazione di tale modello, nella versione logistica con due parametri (difficoltà e discriminazione), produce risultati confrontabili con quelli ottenuti con il modello deterministico.
La preparazione della matrice dei dati per l'analisi ha richiesto l'applicazione dei seguenti passaggi:

a. disposizione dei dati in una matrice in cui le righe rappresentano i soggetti e le colonne gli item;
b. ordinamento della matrice per punteggi decrescenti e attribuzione del codice 1 alle risposte *corrette* e 0 alle risposte *scorrette* o (se significativo) alle *nulle*;
c. calcolo dei marginali di riga e colonna; eliminazione delle righe e delle colonne "complete" (ovvero quelle con risposte tutte corrette o con risposte tutte scorrette) e delle righe che presentano dei dati *missing*.

Per poter applicare il punto *b* si è proceduto alla ricodifica degli item; in particolare per dicotomizzare le risposte per ciascun item è stato seguito il seguente criterio[3]:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	⇐ item
Codifica originaria ⇓															Nuova codifica ⇓
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	1
5	6	6	6	6	6	6	5	6	6	5	6	6	6	6	0
3	4	4	4	4	4	4	3	4	4	3	4	4	4	4	
1	2	2	2	2	2	2	1	2	2	1	2	2	2	2	
	1	1	1	1	1	1		1	1		1	1	1	1	

Degli iniziali 3389 casi solo 2305 sono stati utilizzati per l'analisi per l'eliminazione di quelli che presentavano dati *missing*, il punteggio totale massimo o il punteggio totale minimo.

---

pp previsioni possibili.

[3] E', naturalmente, possibile individuare altri criteri di dicotomizzazione ugualmente o più validi.

ITEM	(1) P	(2) b_i	(3) $E(b_i)$	(4) a_i	(5) $E(a_i)$
1	0.8547	-1.3625	0.0334	1.4325	0.0616
2	0.6915	-0.5286	0.0220	1.9376	0.0826
3	0.4308	0.3005	0.0200	1.8563	0.0778
4	0.4777	0.1644	0.0188	2.0746	0.0871
5	0.3896	0.4166	0.0207	1.7484	0.0740
6	0.8868	-1.5360	0.0339	1.6181	0.0768
7	0.5479	-0.0795	0.0246	1.3809	0.0555
8	0.7414	-0.7235	0.0233	1.9235	0.0841
9	0.9349	-2.1318	0.0499	1.1709	0.0539
10	0.6351	-0.7268	0.0542	0.5079	0.0252
11	0.8898	-1.4847	0.0297	2.1165	0.1144
12	0.8364	-1.2195	0.0303	1.5820	0.0690
13	0.5150	0.0037	0.0301	1.0118	0.0430
14	0.1210	1.4091	0.0303	1.4450	0.0634
15	0.4690	0.1729	0.0253	1.2897	0.0535
Media item utilizzati	0.6281	-0.4884	0.0298	1.5397	0.0681
Deviazione standard	0.2253	0.9135	0.0099	0.4214	0.0205
Numero item utilizzati	15	15	15	15	15

(1) *P*
(2) Difficoltà dell'item *i*
(3) Errore standard della difficoltà dell'item *i*
(4) Discriminazione dell'item *i*
(5) Errore standard della discriminazione dell'item *i*
(6)

Notiamo subito come nessun item ha superato il valore soglia di 0.25 nell'errore standard di difficoltà.
A scopo esemplificativo di seguito è presentata anche la lista delle stime riguardanti una parte dei 2305 soggetti utilizzati nell'analisi. Come si può notare vi sono soggetti che pur avendo lo stesso punteggio totale registrano diversi punteggi sulla logistica di capacità.

```
Listing of estimated item-response abilities and their standard errors.
All data below are based on 15 usable items.

                      Total        Mean         IRT        Std.
            Case      Score       Score     Ability       Error
   2    *******Unusable Case****** zero or perfect total score
             4       8.0000      0.5333      0.4127      0.2945
   6    *******Unusable Case****** zero or perfect total score
             8      11.0000      0.7333      0.1902      0.2786
   9    *******Unusable Case****** zero or perfect total score
            10       8.0000      0.5333      0.4003      0.2931
  12    *******Unusable Case****** zero or perfect total score
            13      11.0000      0.7333      0.2713      0.2822
            14       6.0000      0.4000     -0.4386      0.2860
.....................................................................................
          3376      10.0000      0.6667      0.7825      0.3668
          3377       7.0000      0.4667      0.1312      0.2775
3378    *******Unusable Case****** zero or perfect total score
          3379      10.0000      0.6667     -0.8843      0.2912
          3380      14.0000      0.9333      0.1491      0.2777
          3383       4.0000      0.2667      1.2744      0.5177
3386    *******Unusable Case****** zero or perfect total score
          3388      12.0000      0.8000      1.1461      0.4764
          3389      11.0000      0.7333      0.2390      0.2805

Mean                 9.4217      0.6181      0.0076      0.3554
Std Dev              3.9217      0.2595      0.9940      0.1056
N cases                1584        1584        1584        1584
```

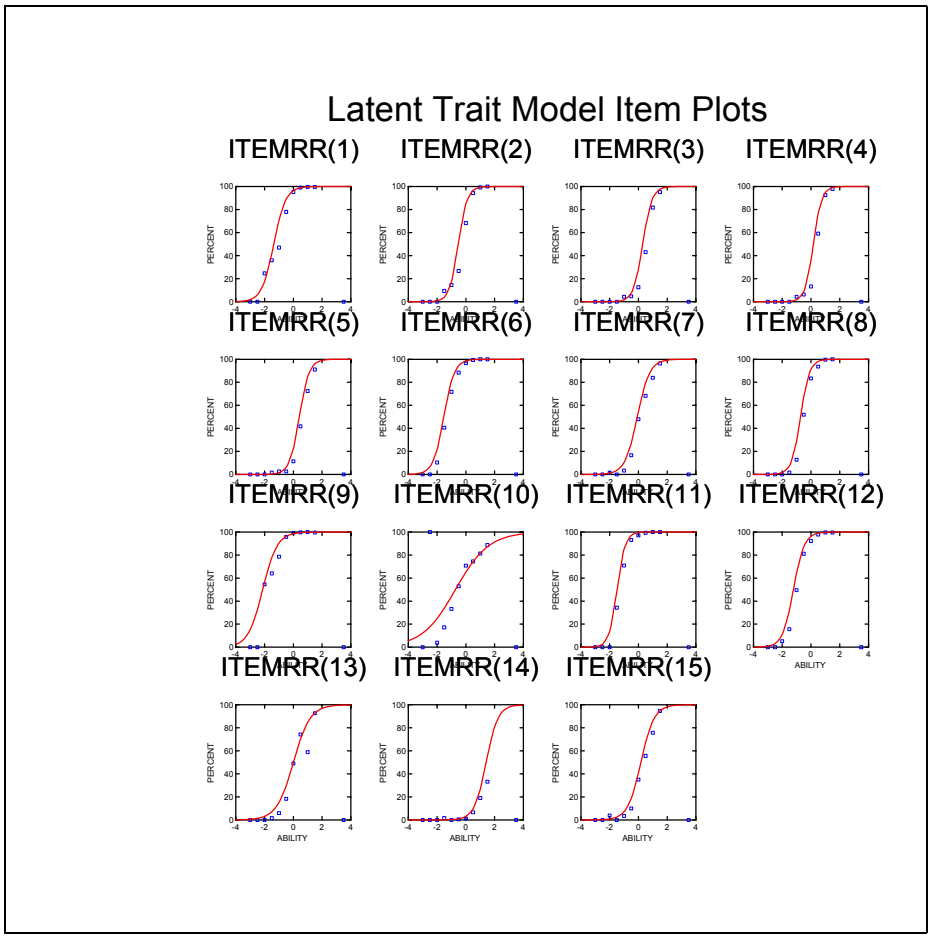
In sintesi i 2305 soggetti hanno registrato i seguenti risultati:

	Media	Deviazione standard
Punteggio medio	9.422	3.922
Media punteggi medi	0.628	0.261
Capacità media	0.000	1.000
Errore medio	0.357	0.106

Punteggio totale	Punteggio medio	Freq.	Freq. Cum.	%	% cum.
1	0.067	117	117	5.1	5.1
2	0.133	67	184	2.9	8.0
3	0.200	72	256	3.1	11.1
4	0.267	83	339	3.6	14.7
5	0.333	89	428	3.9	18.6
6	0.400	132	560	5.7	24.3
7	0.467	135	695	5.9	30.2
8	0.533	138	833	6.0	36.1
9	0.600	168	1001	7.3	43.4
10	0.667	151	1152	6.6	50.0
11	0.733	234	1386	10.2	60.1
12	0.800	270	1656	11.7	71.8
13	0.867	290	1946	12.6	84.4
14	0.933	359	2305	15.6	100.0

Ricordiamo che l'esame degli scarti standard (o dei loro quadrati) consente di identificare
- quali soggetti hanno dato una risposta poco plausibile,
- quali item non presentano una sufficiente affidabilità

che così possono essere eliminati dall'analisi.
Per fare ciò ci si dovrebbe riferire ai due coefficienti t_j e t_i che indicano il livello di incoerenza o *misfit*.
Come sappiamo la procedura di calcolo dei parametri ha carattere iterativo, in quanto ogni volta che le risposte relative ad soggetto o ad un item vengono considerate poco plausibili (alto *misfit*) il vettore di riga o di colonna corrispondente viene cancellato dalla matrice e si ripete l'intera procedura. Verifichiamo l'adattamento al modello logistico degli item utilizzando i grafici che mostrano la percentuale di punteggi *corretti* per ciascuno dei livelli di *capacità*. Ricordiamo che gli asterischi indicano per ciascun item la percentuale attesa di risposte corrette secondo il modello logistico e che i valori numerici delle proporzioni osservate (*P*) e attese (*E(P)*) sono presentati a destra.
Dall'osservazione di tali grafici si nota subito come l'item 10 presenti dei notevoli problemi di adattamento.



5.2.3.3 Confronto tra i risultati dei due approcci

Per confrontare lo scaling degli item ottenuto dalla applicazione dei due diversi modelli si è proceduto al calcolo delle associazioni tra i diversi parametri e coefficienti prodotti.
I risultati di tale confronto, relativamente agli item e ai soggetti, sono riassunti in due matrici presentate di seguito nella quale compaiono i valori delle associazioni calcolate per mezzo di tre coefficienti:

- *tau* di Kendall
- *rho* di Spearman
- *r* di Pearson.

Confronto dei risultati relativi agli item

	P	*bi*	*ebi*	*ai*	*eai*	*cr*	*irt*	*gutt*
r di Pearson								
P	1.00							
bi	-0.99	1.00						
ebi	0.40	-0.52	1.00					
ai	0.09	0.00	-0.74	1.00				
eai	0.24	-0.16	-0.55	0.95	1.00			
cr	0.45	-0.37	-0.12	0.52	0.59	1.00		
IRT	-0.81	0.84	-0.54	-0.04	-0.22	-0.43	1.00	
GUTT	-0.89	0.89	-0.29	-0.25	-0.36	-0.57	0.79	1.00
rho di Spearman								
P	1.00							
bi	-0.99	1.00						
ebi	0.57	-0.63	1.00					
ai	0.04	0.03	-0.63	1.00				
eai	0.18	-0.13	-0.50	0.98	1.00			
cr	0.44	-0.42	0.02	0.59	0.64	1.00		
IRT	-0.87	0.88	-0.64	-0.04	-0.19	-0.43	1.00	
GUTT	-0.88	0.88	-0.37	-0.20	-0.33	-0.58	0.79	1.00
tau di Kendall								
P	1.00							
bi	-0.96	1.00						
ebi	0.43	-0.47	1.00					
ai	0.03	0.01	-0.49	1.00				
eai	0.12	-0.09	-0.39	0.91	1.00			
cr	0.35	-0.31	0.02	0.41	0.44	1.00		
IRT	-0.68	0.71	-0.51	-0.01	-0.11	-0.33	1.00	
GUTT	-0.75	0.75	-0.29	-0.09	-0.18	-0.46	0.62	1.00

p: probabilità di rispondere correttamente all'item (P)

bi: parametro di difficoltà dell'item *i* (b_i)

ebi: errore standard del parametro di difficoltà $(E(b_i))$

ai: parametro di discriminazione dell'item (a_i)

eai: errore standard del parametro di discriminazione $(E(a_i))$

cr: coefficiente di riproducibilità (CR)

irt: ordinamento ottenuto con l'applicazione del modello logistico

gutt: ordinamento ottenuto con l'applicazione del modello deterministico

L'analisi di tali risultati ci consente innanzi tutto registrare un'alta concordanza tra gli ordinamenti degli item rispetto alla difficoltà ottenuti con i due approcci (alta correlazione tra le due serie ordinate *irt* e *gutt*). L'impressione avuta da questi dati è confermata anche confrontando direttamente le due graduatorie:

item ⇒		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ranghi ottenuti con l'applicazione dell'approccio	logistico	4	8	12	14	11	2	9	6	1	7	3	5	15	10	13
	deterministico	1	8	7	9	14	3	11	6	2	10	4	5	12	15	13

Proseguendo nel confronto i risultati ottenuti attraverso l'applicazione dei due modelli, è possibile notare come il coefficiente di riproducibilità (*cr*) presenta una correlazione piuttosto alta con il parametro di discriminazione (*a*) e una correlazione tendenzialmente negativa con il parametro di

difficoltà (*b*).

Confronto dei risultati relativi ai soggetti[4]

Mettendo a confronto le valutazioni che i due approcci fanno dei soggetti misurati, osserviamo la relazione esistente tra coefficiente di predicibilità e livello di capacità.

	pauto	*cp*	*tp*	*mp*	*dj*	*edj*
r di Pearson						
pauto	1.000					
cp	0.58	1.00				
tp	0.84	0.74	1.00			
mp	0.84	0.74	1.00	1.00		
dj	0.81	0.75	0.98	0.99	1.00	
edj	0.17	0.44	0.34	0.34	0.43	1.00
rho di Spearman						
pauto	1.00					
cp	0.68	1.00				
tp	0.84	0.80	1.00			
mp	0.84	0.80	1.00	1.00		
dj	0.80	0.79	0.99	0.99	1.00	
edj	0.29	0.48	0.47	0.47	0.48	1.00
tau di Kendall						
pauto	1.00					
cp	0.54	1.00				
tp	0.68	0.68	1.00			
mp	0.68	0.68	1.00	1.00		
dj	0.62	0.64	0.94	0.94	1.00	
edj	0.17	0.36	0.29	0.29	0.32	1.00
pauto: punteggio di autosufficienza (approccio deterministico)						
cp: coefficiente di predicibilità						
tp: punteggio totale						
mp: punteggio medio						
dj: capacità del soggetto (d_j)						
edj: errore standard della capacità $(E(d_j))$						

Osserviamo subito l'alta correlazione che sia il punteggio di autosufficienza (*pauto*) che il coefficiente di predicibilità (*cp*) registrano con la capacità registrata da ciascun soggetto secondo il modello *IRT*, rispettivamente 0.812 e 0.747.

E' possibile notare a questo punto come il modello deterministico (secondo l'approccio di Guttman) e quello logistico (secondo la teoria dell'*item response*), pur prendendo origine da considerazioni e ipotesi abbastanza differenziate, giungano ad una valutazione sostanzialmente confrontabile della scala e ad una misurazione sostanzialmente coerente e concorde sia degli item che dei soggetti.

5.2.4 Verifica della dimensionalità

I risultati finora osservati confermano la natura ordinale della caratteristica misurata. Alcuni aspetti emersi ci autorizzano ad ipotizzare che l'autosufficienza fisica, così com'è stata definita e com'è percepita, non sia perfettamente unidimensionale. In particolare si potrebbe ipotizzare la presenza di due componenti (fattori). A tal fine verifichiamo se il criterio fattoriale e quello di *scaling*

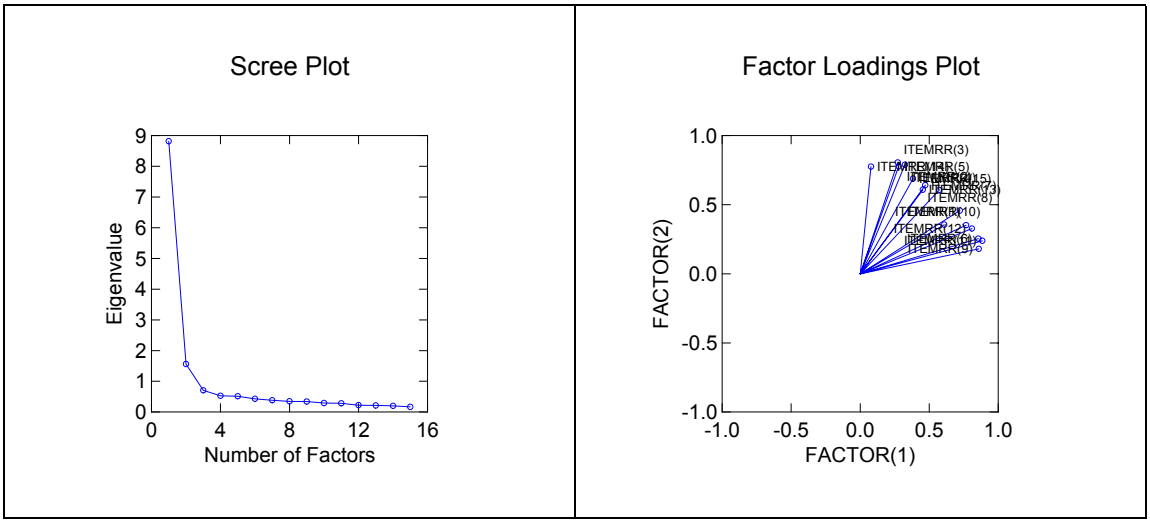
[4] Ricordiamo che per il modello logistico l'analisi viene effettuata solo sui soggetti che non presentano i punteggi estremi (massima e minima autosufficienza).

multidimensionale si adattano ai nostri dati.

Analisi fattoriale

L'applicazione dell'analisi fattoriale esplorativa sembra confermare la nostra ipotesi: i due fattori estratti spiegano quasi la stessa quantità di varianza per un totale di 70% di varianza totale.

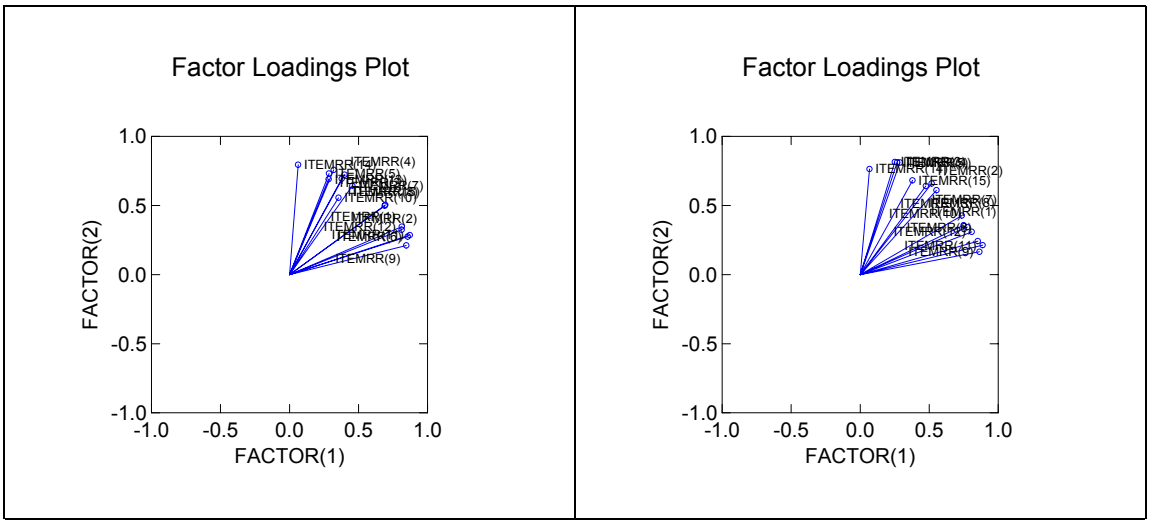
Rotated Loading Matrix (VARIMAX)		
ITEM 1	0.7675	0.3519
ITEM 2	0.5729	0.6064
ITEM 3	0.2719	0.8061
ITEM 4	0.3242	0.7895
ITEM 5	0.2785	0.7776
ITEM 6	0.8553	0.2545
ITEM 7	0.4717	0.6413
ITEM 8	0.7223	0.4580
ITEM 9	0.8592	0.1806
ITEM 10	0.6091	0.3556
ITEM 11	0.8849	0.2390
ITEM 12	0.8094	0.3275
ITEM 13	0.4534	0.6086
ITEM 14	0.0785	0.7757
ITEM 15	0.3830	0.6884
Varianza spiegata	5.5554	4.8247
Perc. Varianza spiegata	37.0362	32.1650



I due fattori estratti si riferiscono a due aspetti dell'autosufficienza tra loro ordinali: mentre il primo riguarda attività ed azioni che richiedono capacità minime (spostarsi per le stanze ed uscire di casa, lavarsi viso e braccia, vestirsi e spogliarsi, mangiare e prepararsi un pasto caldo da soli, usare il bagno, alzarsi e andare a letto), il secondo riguarda attività che richiedono capacità fisiche elevate (camminare e fare la spesa, fare il bagno o la doccia, tagliarsi le unghie dei piedi, fare lavori domestici).

In realtà i suggerimenti che ci pervenivano dai risultati precedenti facevano ipotizzare la presenza, o l'interferenza, di un fattore culturale accanto a quello riguardante strettamente le capacità fisiche. In particolare un indicatore di ciò poteva essere identificato con la variabile genere. Per questo motivo abbiamo provato a riapplicare il criterio fattoriale ai due gruppi separati.

Rotated Loading Matrix (VARIMAX)				
	maschi		femmine	
ITEM 1	0.812	0.324	0.749	0.357
ITEM 2	0.690	0.498	0.518	0.660
ITEM 3	0.321	0.758	0.251	0.815
ITEM 4	0.398	0.723	0.288	0.811
ITEM 5	0.286	0.732	0.267	0.812
ITEM 6	0.856	0.275	0.852	0.242
ITEM 7	0.455	0.644	0.476	0.641
ITEM 8	0.694	0.505	0.736	0.427
ITEM 9	0.846	0.211	0.864	0.165
ITEM 10	0.355	0.556	0.776	0.343
ITEM 11	0.871	0.286	0.887	0.213
ITEM 12	0.814	0.347	0.808	0.309
ITEM 13	0.284	0.691	0.552	0.611
ITEM 14	0.062	0.795	0.068	0.764
ITEM 15	0.378	0.694	0.379	0.682
Varianza spiegata	5.392	4.877	5.782	4.905
Perc. Varianza spiegata	35.944	32.515	38.545	32.702

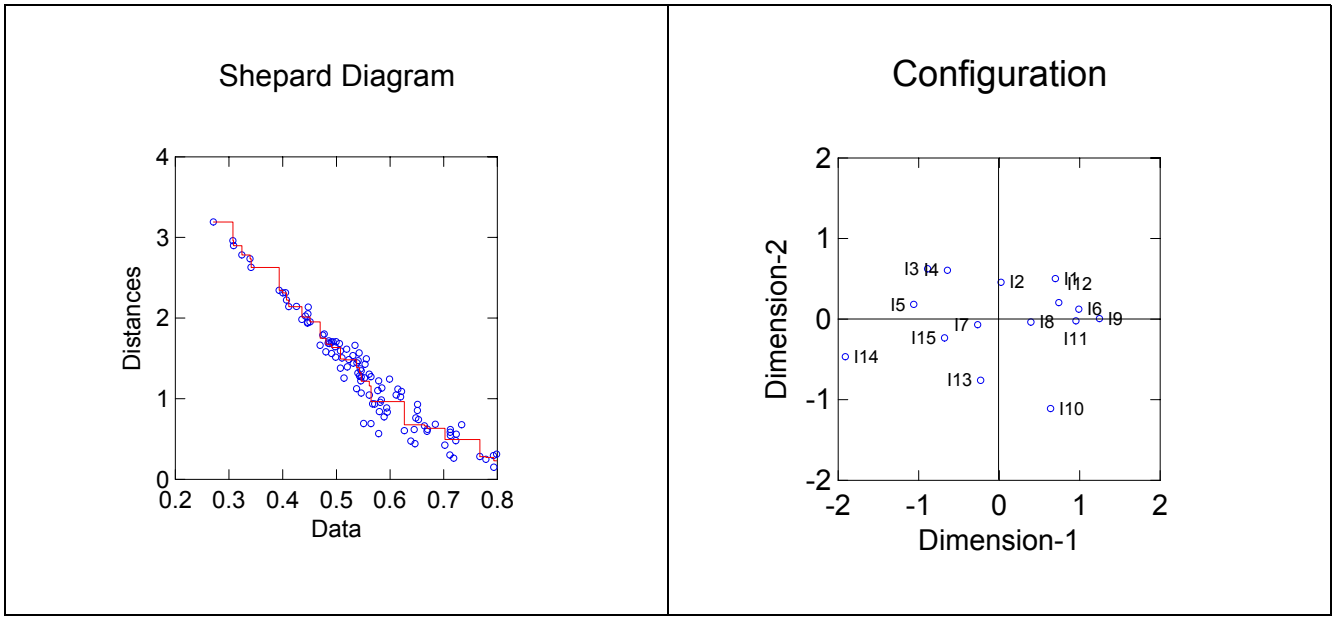


Come si può osservare la sostanziale uguaglianza tra le strutture fattoriali per maschi e femmine trova un'eccezione nell'item 8 ("vestirsi e spogliarsi") e nell'item 13 ("fare lavori domestici leggeri") che, il primo per i maschi e il secondo per le femmine, compaiono in entrambi i fattori; l'item che però rivela avere un comportamento molto legato al genere è il numero 10 ("prepararsi un pasto caldo") che per i maschi, al contrario di quanto succede per le femmine, definisce il fattore legato alle capacità elevate.

Il criterio di *scaling* multidimensionale

Anche se i dati sono del tipo stimolo-unico proviamo ad applicare il criterio di *scaling* multidimensionale; in tale applicazione la matrice di somiglianza sottoposta ad analisi è quella dei correlazione tra gli item; si assume una relazione monotona tra somiglianze e distanze (funzione di trasformazione monotona). I risultati rilevano un elevato adattamento tra matrice di somiglianza (osservata) e quella di distanze (calcolata).

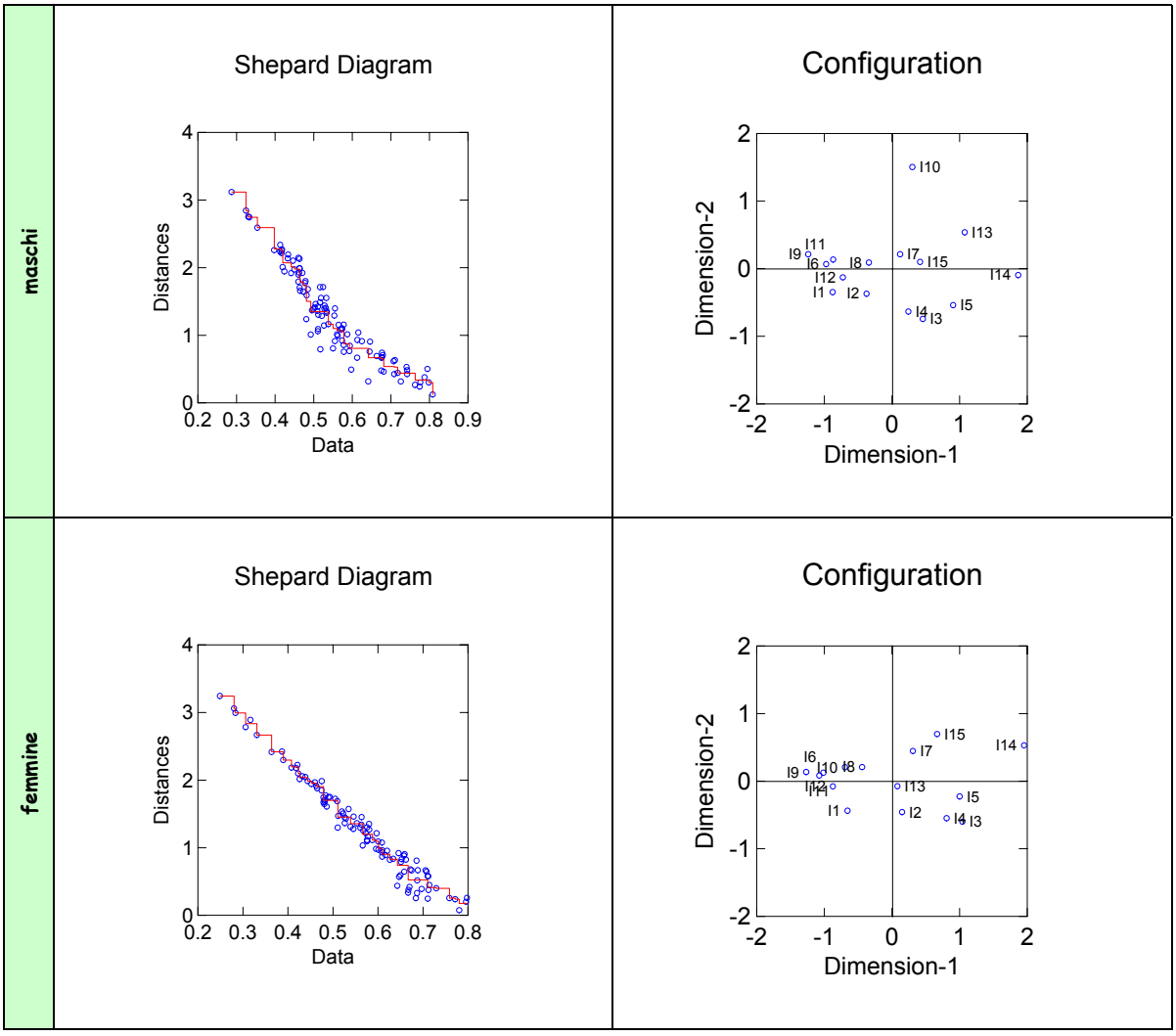
Monotonic Multidimensional Scaling Minimizing Kruskal STRESS (form 1) in 2 dimensions Stress of final configuration is: 0.08803 Proportion of variance (RSQ) is: 0.96222					
	Dim 1	Dim 2		Dim 1	Dim 2
ITEM 1	.70	.50	ITEM 8	.39	-.04
ITEM 2	.02	.46	ITEM 9	1.25	.00
ITEM 3	-.89	.62	ITEM 10	.64	-1.11
ITEM 4	-.64	.61	ITEM 11	.95	-.02
ITEM 5	-1.06	.18	ITEM 12	.74	.20
ITEM 6	.99	.12	ITEM 13	-.23	-.76
ITEM 7	-.27	-.07	ITEM 14	-1.91	-.47
			ITEM 15	-.68	-.23



I risultati di questa analisi risultano essere piuttosto interessanti. La configurazione finale conferma la struttura fattoriale precedentemente vista: rispetto alla *dimensione-1* osserviamo in corrispondenza della parte positiva gli item che prima avevamo osservato nel fattore 1 e in corrispondenza della parte negativa gli item che precedentemente avevamo osservato nel fattore 2; rispetto al criterio fattoriale qui appare più evidente il carattere ordinale/cumulativo degli item.
Osservando i risultati relativamente alla *dimensione-2* è possibile rilevare come gli item sembrano disporsi in modo più legato alle abitudini che alle capacità fisiche (si notino nella parte negativa gli item "prepararsi un pasto caldo" e quelli relativi ai lavori domestici). Avendo individuato nella variabile "genere" quella che può discriminare rispetto alla dimensione culturale, proviamo a ripetere l'analisi separatamente per maschi e femmine.

Monotonic Multidimensional Scaling Minimizing Kruskal STRESS (form 1) in 2 dimensions		
	Maschi	Femmine
Stress of final configuration	0.09985	0.06547
Proportion of variance (RSQ)	0.95271	0.98193

Coordinate in due dimensioni					
		Maschi		Femmine	
		Dim 1	Dim 2	Dim 1	Dim 2
ITEM	1	-.88	-.35	-.66	-.44
	2	-.37	-.37	.15	-.46
	3	.46	-.74	1.04	-.60
	4	.24	-.63	81	-.55
	5	.91	-.54	1.00	-.23
	6	-.97	.07	-1.01	.13
	7	.12	.21	.31	.45
	8	-.34	.09	-.44	.21
	9	-1.24	.21	-1.27	.13
	10	.30	1.50	-.69	.21
	11	-.87	.13	-1.07	.08
	12	-.72	-.13	-.87	-.08
	13	1.08	.54	.08	-.08
	14	1.87	-.09	1.95	.53
	15	.42	.10	.67	.70



Attraverso questa analisi appare abbastanza chiaro come per il gruppo dei maschi il significato della dimensione 2 sia quasi del tutto attribuibile all'item 13 ("fare lavori domestici leggeri") e, soprattutto, l'item 10 ("prepararsi un pasto caldo") mentre per il gruppo delle femmine la seconda dimensione sembra quasi non avere alcun significato, infatti rispetto a questa dimensione gli item risultano essere concentrati in pochi valori.

5.2.5 Nuove ipotesi

A questo punto si può dire che nonostante la verifica dimensionale abbia confermato la natura ordinale della caratteristica misurata, la presenza di due particolari item può far insorgere due diverse ipotesi:

- si tratta di item affetti da *bias*; in questo caso si può assumere che l'errore è prodotto da una differenza significativa tra gruppi (maschi e femmine) rispetto alla difficoltà relativa dell'item; in altre parole risulta comparativamente più difficile rispondere "correttamente" per un gruppo anziché per un altro;
- si tratta di item che misurano un'altra dimensione.

Per verificare la prima ipotesi utilizziamo la tecnica *TID (Transformed Item Difficulties)* mentre la verifica del livello di adattamento del modello di scalogramma multidimensionale (POSAC) consentirà provare la seconda.

5.2.5.1 Verifica della presenza di item affetti da bias

L'applicazione della tecnica detta *Transformed Item Difficulties (TID)* consente di individuare la presenza di item affetti da *bias* attraverso la verifica della presenza o l'assenza dell'interazione gruppi*item. Nel nostro caso i gruppi individuati sono maschi e femmine.
Dato il significato dei nostri item, si considera indice di difficoltà, *p*, la proporzione di soggetti che hanno riferito di non essere capaci a svolgere una determinata funzione. Per individuare il valore di *p* si è deciso di procedere ad un diverso accorpamento, rispetto a quanto fatto precedentemente, delle categorie utilizzate, ovvero:

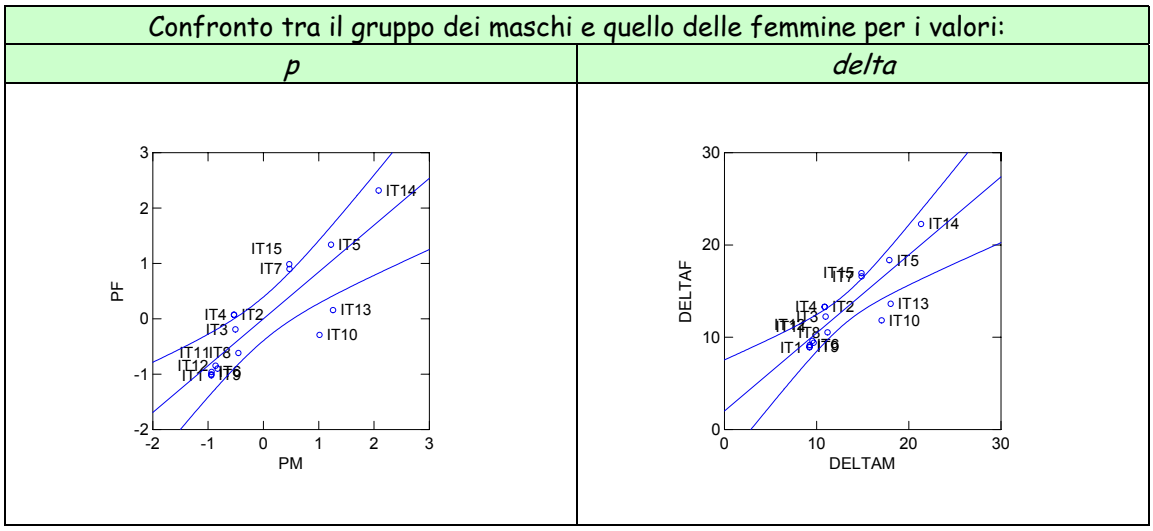
1-*p*	• Senza difficoltà • Con difficoltà ma senza aiuto
p	• Con aiuto per parte dell'azione • Con aiuto totale • Non lo fa per abitudine

Di seguito valori di *p* e di *1-p* per i due gruppi:

ITEM		maschi		femmine	
		1-p	p	1-p	p
11	Usare il gabinetto	1505	.132	.1585	.167
1	Spostarsi per le stanza	1530	.107	.1613	.139
2	Uscire di casa	1439	.198	.1362	.390
3	Fare le scale	1434	.203	.1422	.330
4	Camminare per almeno 400 m.	1440	.197	.1361	.391
5	Fare la spesa	1052	.585	.1070	.682
6	Lavarsi viso e braccia	1529	.108	.1599	.153
7	Fare il bagno o la doccia	1218	.419	.1171	.581
8	Vestirsi e spogliarsi	1422	.215	.1520	.232
9	Mangiare da solo	1530	.107	.1609	.143
10	Prepararsi un pasto caldo	1098	.539	.1445	.307
12	Alzarsi e andare a letto	1513	.124	.1573	.179
13	Lavori domestici leggeri	1044	.593	.1342	.410
14	Lavori domestici pesanti	862	.775	.845	.907
15	Tagliarsi le unghie	1219	.418	.1151	.601

I valori di *p* vengono quindi standardizzati (*valori delta*). Di seguito vediamo due diagrammi di

dispersione che mettono in relazione i valori *p* (primo diagramma) e i valori *delta* (secondo diagramma) per il gruppo di maschi e il gruppo delle femmine. Secondo questo approccio il livello di dispersione dei punti nel grafico così costruito è considerato una misura dell'interazione *gruppo*item*, una specie di coefficiente di correlazione inverso. La retta tracciata, che può essere considerata l'asse maggiore dell'ellisse descritta dai punti, serve come indice della relazione bivariata dei valori *p*/*delta* dei due gruppi. Esso diviene l'informazione base a partire dalla quale individuare la presenza di item *biased*. Per ciascuna retta sono determinati gli intervalli di confidenza dell'asse maggiore. Gli item che nel grafico appaiono al di fuori di tali intervalli possono essere giudicati "deviati".



Come si può osservare gli item, già segnalati come item deboli dall'analisi dello scalogramma, sono il 10 e il 13, risultati più "difficili" per i maschi, e il 15 e il 7 (per entrambi i gruppi) che per tutti richiedono capacità fisiche difficilmente rilevabili tra gli anziani.
Occorre però tener presente che in questo caso gli item risultati *biased* non sono di per sé non validi ma rivelano in realtà la probabile presenza di un'altra dimensione.

5.2.5.2 Verifica della scalabilità multidimensionale

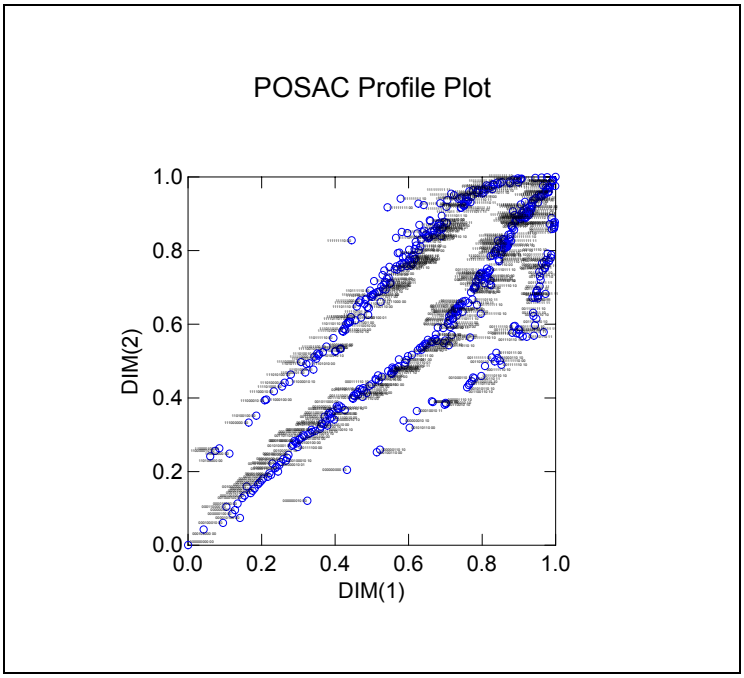
La misura del livello di adattamento del criterio di scalogramma multidimensionale consentirà di chiarire ulteriormente la presenza di due dimensioni. Per fare ciò, essendo la matrice dei dati molto grande, si utilizza la procedura *POSAC*; questa, come sappiamo, consente di determinare l'adattamento dei profili osservati in uno spazio bidimensionale.
Ricordiamo che l'obiettivo del *POSAC* è quello di verificare se è possibile assegnare due punteggi a ciascun profilo, in modo tale che sia possibile rappresentare la relazione tra due profili confrontando semplicemente i loro corrispondenti profili di coordinate. Si ottiene una rappresentazione perfetta quando i punteggi individuati descrivono perfettamente l'ordine e la confrontabilità dei profili originari.
Nel caso in cui non sia possibile ottenere per un certo scalogramma una perfetta rappresentazione in uno spazio bidimensionale, l'approccio *POSAC* definisce un criterio di bontà di adattamento che consente di stabilire qual è il tipo di collocazione bidimensionale che meglio descrive le relazioni d'ordine osservate tra i profili. A tal fine è stato definito un coefficiente, detto *coefficiente di corretta rappresentazione* (*CORREP*), basato sulla proporzione di profili rappresentati in modo corretto, tenendo conto della frequenza registrata da ciascun profilo. Il valore del coefficiente di corretta rappresentazione va da 0 a 1 (soluzione perfetta) ed è molto sensibile all'*approssimazione iniziale* utilizzata.

In altri casi il procedimento per definire l'approssimazione iniziale prevede che vengano eseguiti in successione i seguenti momenti:

a. calcolo della matrice dei coefficienti di debole monotonicità (*weak monotonicity coefficients*, *wm*);
b. identificazione dei due item (i_0 e j_0) che presentano la minore correlazione positiva (item estremi);
c. determinazione della posizione di ciascun profilo.

Reordered item weak monotonicity coefficients															
item	10	13	9	11	6	8	12	1	7	15	2	5	14	4	3
10	1.000														
13	0.874	1.000													
9	0.974	0.975	1.000												
11	0.964	0.977	0.987	1.000											
6	0.93	0.963	0.983	0.983	1.000										
8	0.889	0.921	0.986	0.990	0.985	1.000									
12	0.897	0.950	0.978	0.987	0.968	0.964	1.000								
1	0.890	0.963	0.955	0.965	0.968	0.956	0.961	1.000							
7	0.819	0.859	0.985	0.980	0.993	0.962	0.948	0.941	1.000						
15	0.804	0.847	0.980	0.985	0.976	0.968	0.956	0.939	0.918	1.000					
2	0.805	0.913	0.957	0.957	0.960	0.925	0.951	0.993	0.923	0.916	1.000				
5	0.787	0.858	0.986	0.975	0.978	0.944	0.958	0.976	0.899	0.872	0.974	1.000			
14	0.766	0.929	0.972	1.000	0.989	0.973	0.987	0.976	0.934	0.918	0.975	0.911	1.000		
4	0.724	0.849	0.958	0.964	0.967	0.933	0.959	0.991	0.880	0.869	0.976	0.931	0.944	1.000	
3	0.717	0.834	0.951	0.954	0.944	0.919	0.962	0.984	0.873	0.879	0.978	0.909	0.938	0.972	1.000

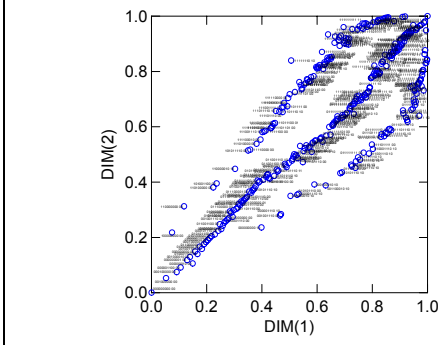
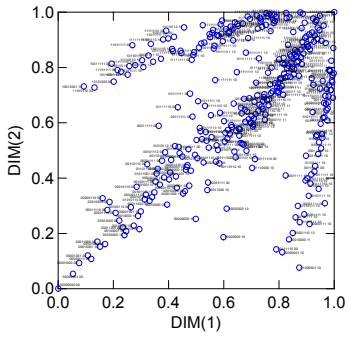
Final loss value	3350.8184
Proportion of profile pairs correctly represented	0.7701
Score-distance weighted coefficient	0.8520



Maschi															
Reordered item weak monotonicity coefficients															
item	10	9	11	6	8	12	13	1	7	15	2	5	14	4	3
10	1.000														
9	0.969	1.000													
11	0.977	0.989	1.000												
6	0.938	0.986	0.987	1.000											
8	0.899	0.988	0.994	0.988	1.000										
12	0.916	0.978	0.987	0.973	0.974	1.000									
13	0.783	0.978	0.975	0.961	0.906	0.940	1.000								
1	0.913	0.965	0.975	0.975	0.966	0.979	0.964	1.000							
7	0.827	0.987	0.978	0.993	0.963	0.944	0.811	0.945	1.000						
15	0.807	0.987	0.985	0.970	0.971	0.964	0.822	0.953	0.917	1.000					
2	0.859	0.964	0.967	0.969	0.948	0.967	0.935	0.996	0.929	0.941	1.000				
5	0.765	0.994	0.979	0.973	0.947	0.965	0.811	0.971	0.882	0.852	0.968	1.000			
14	0.759	0.974	1.000	0.990	0.968	0.984	0.912	0.981	0.904	0.895	0.970	0.862	1.000		
4	0.741	0.951	0.966	0.970	0.935	0.966	0.841	0.995	0.873	0.873	0.983	0.908	0.938	1.000	
3	0.714	0.944	0.952	0.950	0.922	0.968	0.815	0.994	0.883	0.881	0.983	0.886	0.916	0.973	1.000

Femmine															
Reordered item weak monotonicity coefficients															
item	15	7	8	10	6	11	9	13	12	14	1	5	2	4	3
15	1.000														
7	0.916	1.000													
8	0.965	0.961	1.000												
10	0.944	0.934	0.946	1.000											
6	0.982	0.993	0.982	0.958	1.000										
11	0.985	0.981	0.985	0.976	0.980	1.000									
9	0.972	0.983	0.985	0.986	0.980	0.986	1.000								
13	0.904	0.917	0.939	0.969	0.968	0.979	0.976	1.000							
12	0.947	0.950	0.953	0.939	0.963	0.988	0.978	0.961	1.000						
14	0.938	0.963	0.979	0.982	0.987	1.000	0.968	0.996	0.990	1.000					
1	0.923	0.936	0.948	0.936	0.960	0.956	0.946	0.969	0.941	0.969	1.000				
5	0.885	0.910	0.939	0.957	0.982	0.971	0.976	0.933	0.950	0.956	0.978	1.000			
2	0.896	0.920	0.914	0.929	0.956	0.954	0.954	0.940	0.938	0.977	0.990	0.977	1.000		
4	0.860	0.885	0.937	0.913	0.964	0.966	0.969	0.918	0.951	0.945	0.985	0.948	0.968	1.000	
3	0.873	0.858	0.918	0.909	0.937	0.958	0.959	0.904	0.956	0.954	0.972	0.926	0.972	0.969	1.000

	maschi	femmine
Final loss value	953.5039	255.3252
Proportion of profile pairs correctly represented	0.7365	0.8296
Score-distance weighted coefficient	0.8340	0.9518

Maschi	Femmine
	

Confrontando i risultati prodotti, sia numerici che grafici, sul campione totale e suddiviso per genere, notiamo subito come il migliore adattamento è quello ottenuto nel campione femminile ovvero quello che, secondo le nostre ipotesi è anche meno affetto dalla dimensione "culturale".
Può essere curioso a questo punto confrontare l'ordinamento degli item rispetto ai valori dei coefficienti di *weak monotonicity*.

	Ordinamenti degli item secondo i valori dei coefficienti di *weak monotonicity*		
	Totale	Maschi	Femmine
1. Spostarsi per le stanze	8	8	3
2. Uscire di casa	11	11	13
3. Fare le scale	15	15	15
4. Camminare per almeno 400 m	14	14	14
5. Fare la spesa	12	12	12
6. Lavarsi viso e braccia	5	4	5
7. Fare il bagno o la doccia	9	9	2
8. Vestirsi e spogliarsi	6	5	3
9. Mangiare da solo	3	2	7
10. Prepararsi un pasto caldo	1	1	4
11. Usare il gabinetto	4	3	6
12. Alzarsi e andare a letto	7	6	9
13. Lavori domestici leggeri	2	7	8
14. Lavori domestici pesanti	13	13	10
15. Tagliarsi le unghie dei piedi	10	10	1

Notiamo subito come l'ordinamento ottenuto sul campione totale e quello ottenuto sul campione di maschi sono abbastanza omogenei tra loro e molto diversi da quello ottenuto sul campione di femmine. In particolare vediamo come ciò riguarda gli item gli item 10, 15, 7 e 9.

5.2.6 Individuazione dei valori-soglia

Da questa scala è stato possibile ottenere non solo una valutazione dell'autosufficienza dei singoli soggetti del campione in termini di punteggio globale ma anche altri tre punteggi relativi alle aree di non autosufficienza identificate sulla base della tipologia delle attività che le caratterizzano ma soprattutto della diversità dei bisogni assistenziali che ne derivano.
Infatti, poiché uno dei principali obbiettivi dello studio era l'identificazione dei bisogni assistenziali, è stato analizzato il significato delle singole attività rilevate dalla scala. Il riordinamento degli item

secondo il modello della "scalabilità" ha messo in evidenza che le attività svolte senza difficoltà dal maggior numero di soggetti e il cui deterioramento è da considerarsi indicativo di una grave disabilità, riguardano le attività relative alla "cura di sé" che devono essere espletate più volte al giorno e in tempi non prevedibili. Al contrario le attività più "difficili" comprendono le attività da svolgersi in modo saltuario. In posizione intermedia si ritrovano attività che vengono svolte quotidianamente a tempi prevedibili. Cosí, il nuovo ordinamento delle attività permette di individuare tre aree di attività diverse per:

- periodicità temporale
- tipo di intervento richiesto.

IDENTIFICAZIONE DELLE AREE DI NON AUTOSUFFICIENZA			
(b)	(a)		autosufficienza nelle attività
1	3	Lavarsi viso e braccia	per la cura di sé
2	1	Spostarsi per le stanze	
3	2	Mangiare da solo	
4	4	Usare il gabinetto	
5	5	Alzarsi e andare a letto	
6	6	Vestirsi e spogliarsi	
7	10	Prepararsi un pasto caldo	
11	12	Fare lavori domestici leggeri	quotidiane
12	11	Fare il bagno o la doccia	
13	14	Portare una borsa della spesa	
14	13	Tagliarsi le unghie dei piedi	
8	8	Uscire di casa	saltuarie
9	9	Camminare per almeno 400 m.	
10	7	Fare le scale	
15	15	Fare lavori domestici pesanti	
(a) ordine di presentazione (b) ordine ottenuto con il criterio deterministico applicato sui dati non dicotomizzati (non presentato in questa sede)			

I primi 7 item, indicativi di un livello minimale di autosufficienza comprendono attività che sono essenziali alla cura di sé che devono essere svolte più volte al giorno, a intervalli brevi e/o imprevedibili. Carenze nella possibilità di compiere tali attività sono proprie degli individui in impellenti condizioni di bisogno che necessitano di interventi complessi, pressoché continui a domicilio e la cui messa in opera richiede, almeno in parte, personale specializzato. Tali servizi non possono prescindere da una valutazione dell'assistenza fornita dai familiari o dalla rete sociale. Infatti questi soggetti per la loro incapacità di vivere da soli, sono presenti nel nostro studio, dunque al loro domicilio, solo perché possono contare su adeguati supporti familiari o informali.

Gli item 11,12,13,14 identificano attività che devono essere svolte 1-2 volte al giorno. I deficit in tali attività possono essere colmati da una struttura familiare solida senza alcun ausilio esterno. Per questo, la decisione di fornire un servizio di assistenza deve derivare da una valutazione dei "supporti informali". Anche nei casi in cui tale assistenza si renda necessaria (per esempio: soggetti che vivono da soli o con un coniuge anziano) sarà comunque realizzabile con un impegno limitato perché un operatore può assistere più di un soggetto nel corso della giornata. In alcuni casi, il deficit è colmabile con servizi generali (pasti a domicilio, servizi di lavanderia ecc.) che forniscono l'intervento richiesto con strutture e personale generico.

Gli item 8,9,10,15 si riferiscono ad attività saltuarie che devono essere svolte con intervalli anche variabili di 2-7 giorni. I soggetti con deficit solo in quest'area sono spesso in grado di vivere autonomamente purché dispongano di un aiuto periodico, anche esterno, familiare o istituzionale.

Per rendere interpretabili i dati della scala sono stati individuati 3 livelli di autosufficienza nell'area generale; la validazione di ciascuna delle tre sottoaree è stato effettuato un confronto con i risultati ottenuti attraverso una rilevazione parallela effettuata mediante un questionario somministrato alla persona (generalmente un familiare) che presta la maggiore assistenza. Ciò ha consentito di individuare le soglie che nella scala continua dell'autosufficienza corrispondono ai vari livelli di assistenza. Del questionario sul carico assistenziale sono stati considerati, per questa validazione,

gli item relativi a:

- bisogno di aiuto fisico per la cura di sé
- bisogno di aiuto per i compiti domestici
- bisogno di sorveglianza diurna e notturna

Ai fini di una classificazione dei soggetti sulla base del bisogno di assistenza i gruppi sono definiti come:

1. autosufficienti
2. parzialmente disabili: necessitano di aiuto saltuario
3. gravemente disabili: necessitano di assistenza continuativa.

Per la definizione di queste soglie i punteggi dell'autosufficienza sono stati messi in relazione con le variabili del questionario che consentivano di valutare il carico assistenziale. Il punto mediano per ciascuna tipologia di assistenza rappresenta, in questo tipo di distribuzioni, l'indice di tendenza centrale più stabile. La trasformazione *logit* condotta per una verifica di questa prima analisi ha confermato la stabilità del punto mediano. Con la trasformazione *logit*, infatti, si ricava un punto centrale, ossia un valore analogo, come significato, alla mediana, ma rispetto a questa ottenuto dall'interpolazione della funzione $0.5\log\left(\frac{p_i}{1-p_i}\right)$, in cui p_i è la proporzione di soggetti con un dato punteggio o meno. Il *logit 50* rappresenta così un indice centrale più stabile della mediana, specialmente in presenza di vari soggetti a pari merito.

VALORE CENTRALE (50° CENTILE) PER LA VARIABILE ASSISTENZA DIURNA E NOTTURNA

		Frequenza dell'assistenza						
		Diurna e notturna			Notturna		Diurna	
Frequenza di intervento		+++	++	+	++	+	++	+
Punteggio di autosufficienza	Globale	27	40	55	57	62	59	75
	Cura di sé	29	52	68	70	69	71	80

+ saltuaria
++ periodica
+++ continua

In questo modo la similitudine fra mediana e *logit 50* e la vicinanza fra i quartili e la mediana concorrono a fissare i punti di soglia dei vari gruppi di autosufficienza definiti in termini di livelli di assistenza.

I punti di divisione sono stati ricavati dalle distribuzioni di frequenza dei punteggi in tabelle di contingenza con le variabili del questionario dell'assistenza.

SOGLIE FRA LE CATEGORIE DELL'AUTOSUFFICIENZA GLOBALE (CURA DI SÉ, ATTIVITA' QUOTIDIANE, ATTIVITA SALTUARIE) E DELLA SOLA CURA DI SÉ.

		GRUPPO			
		Dipendenza completa	Dipendenza parziale		Autosufficienza
Frequenza di intervento		continuo	periodico	saltuario+	no
Punteggio di autosufficienza	Globale	55	65		80
	Cura di sé	65	75		(*)

(*) La soglia fra dipendenza limitata e autosufficienza per la cura di sé non viene riportata perché presenta un valore troppo alto e instabile: la cura di sé è richiesta con continuità, e non è risolvibile con interventi saltuari. Gli alti valori delle soglie dell'autosufficienza nella cura di sé sono logicamente dovuti al fatto che una perdita di autonomia anche lieve in quest'area comporta un disagio che deve essere necessariamente colmato.

Il nostro studio mette chiaramente in evidenza, per esempio, che i 6 item che riguardano le attività della cura di sé permettono da soli di identificare una fascia di popolazione in stato di più impellente bisogno di aiuto fisico per la vita quotidiana.

E' importante sottolineare come i livelli di autosufficienza sono stati definiti sulla base dello studio del rapporto tra autosufficienza e supporti informali ovvero mediante l'identificazione delle modalità *assistenziali* che gli individui, non potendo attendere l'attuazione dei servizi istituzionali, hanno già messo spontaneamente in atto.

A. LOGARITMI E LOGIT

I logaritmi

I *logaritmi* originariamente sono stati sviluppati dai matematici per potere trattare con maggiore facilità grandi numeri in problemi complessi, Successivamente si è riscontrato che potevano essere utili nelle descrizioni matematiche di molti fenomeni naturali e anche psicologici e sociali. L'utilizzo dei logaritmi consente di sostituire operazioni quali la moltiplicazione, la divisione e elevazione a potenza con operazioni più semplici ovvero, rispettivamente, l'addizione, la sottrazione e la moltiplicazione. Le regole basilari della matematica dei logaritmi sono le seguenti:

- La moltiplicazione è sostituita dall'addizione: $\log_n(x*y)=\log_n(x)+\log_n(y)$
  es. il logaritmo di 6(.778) più il logaritmo di 2(.301) = 1.08 ovvero il logaritmo di 12.
- La divisione è sostituita dalla sottrazione: $\log_n(x/y)=\log_n(x)-\log_n(y)$
  es. il logaritmo di 6(.778) meno il logaritmo di 2(.301) = .477 ovvero il logaritmo di 3.
- L'elevazione a potenza è sostituita dalla moltiplicazione: $\log_n(x^a)=a*\log_n(x)$
  es. il logaritmo di 6(.778) volte 2 = 1.556 ovvero il logaritmo di 36.

Un altro vantaggio nell'utilizzo dei logaritmi è dato dalla possibilità di utilizzo nella definizione di scale di misura e nella loro rappresentazione grafica: uguali distanze (cicli logaritmici) vengono trasformate automaticamente in uguali rapporti (i seguenti intervalli espressi in forma logaritmica risultano uguali: 1-10, 10-100, 100-1000).
Per questi motivi i logaritmi vengono utilizzati anche in statistica.
Il vantaggio di tale trasformazione è quello di poter utilizzare le proprietà dei logaritmi:

$$\log_n(xy)=\log_n(x)+\log_n(y) \qquad \text{e} \qquad \log_n(x^a)=a*\log_n(x)$$

che semplifica notevolmente i calcoli.

I *Logit*

I *logit* sono unità matematiche o di probabilità logistica relativa a una data osservazione. In particolare un *logit* si ottiene calcolando il logaritmo naturale dell'*odds* di *p* che è dato dal rapporto tra *p* e il suo reciproco; nel nostro caso quindi il *logit* è dato dal logaritmo naturale del rapporto tra la probabilità di dare una risposta corretta (p_{ij}) e la probabilità di dare una risposta errata ($1-p_{ij}$):

$$\log_n\left[\frac{p_{ij}}{(1-p_{ij})}\right]$$

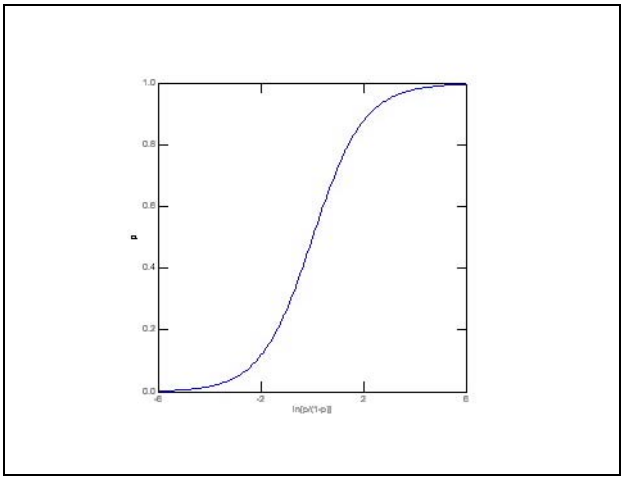
Il *logit* dà origine ad una scala ad intervalli e può essere sottoposto a qualsiasi trasformazione lineare. Si distribuisce simmetricamente intorno ad un valore centrale. Quando *p*=0.50, il *logit* sarà uguale a zero infatti:

$$\log_n\left[\frac{0.50}{0.50}\right]=\log_n(1)=0$$

Via via che la dicotomia si sposta verso una delle due direzioni, ovvero approssimando 0 o 1, i valori di *logit* si allontanano da 0; nella seguente tabella osserviamo la corrispondenza tra *p*, *1-p* e *logit* per alcuni valori di *p*:

p	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
1-p	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10
logit	-2.20	-1.39	-0.85	-0.41	0.00	0.41	0.85	1.39	2.20

Da notare che mentre i valori di *p* nella tabella sono equidistanti (infatti distano sempre 0.10), la distanti tra i corrispondenti valori dei *logit* non è sempre la stessa ma aumenta allontanandosi da 0.50; inoltre all'aumentare del valore assoluto del *logit*, la probabilità corrispondente si avvicina ai valori estremi, senza raggiungerli mai.
Non esiste alcun limite superiore o inferiore del *logit* ma quando *p* è uguale esattamente a 1 o a 0 il *logit* risulta indefinito. La distribuzione continua dei valori di probabilità trasformati in *logit* è la seguente:



Come si potrà notare la curva logistica e la curva normale standardizzata cumulata sembrano molto simili e in generale è possibile fare per entrambe le stesse deduzioni. Come si può notare la trasformazione logistica è praticamente lineare per una porzione di valori di *p* (in particolare tra 0.20 e 0.75); questo vuol dire che all'interno di tale intervallo il modello lineare di probabilità produce risultati molto simili a quello del modello logistico. Quando la probabilità si avvicina ai due estremi (0 e 1) la trasformazione è in modo evidente non-lineare. A tale proposito è possibile fare alcune osservazioni. Tale trasformazione appiattisce le probabilità molto elevate e molto basse; questo vuol dire che i *logit* risultano molto utili quando è necessario effettuare confronti tra proporzioni a livelli diversi. Sia dalla tabella che dal grafico appare evidente come i valori di *logit* siano simmetrici; infatti la trasformazione logistica di un *odds* (*p/(1-p)*) e del suo reciproco (*(1-p)/p)*) produce un *logit* di valore uguale ma di segno opposto.