Rilevazione e analisi statistica del dato soggettivo

Parte I

La costruzione e l'analisi del dato soggettivo

1. LA TEORIA DELLA MISURAZIONE DEL SOGGETTIVO

1.1 L'AFFIDABILITÀ: TEORIE PER LA STIMA E LA VALUTAZIONE DELL'ERRORE DI MISURAZIONE

1.1.1 La teoria classica

L'impossibilità di eliminare dalla misurazione l'errore è formalizzata nella *teoria classica della misurazione* (DeVellis, 1991; Nunnally, 1978; Spector, 1992) all'interno della quale sono definite le seguenti tre variabili:

- punteggio vero (t), ovvero il valore reale ma teorico (e quindi "atteso") che ciascun oggetto p possiede rispetto alla caratteristica e che, a causa degli errori di misurazione, non può essere osservato direttamente, ma può essere stimato attraverso la misurazione reale; per questo esso è concepito come una quantità ipotetica, non osservabile che non può essere direttamente misurata; in questo senso esso rappresenta il valore atteso per un certo oggetto;
- punteggio osservato (x), ovvero il valore realmente osservato per l'oggetto p, rispetto alla caratteristica, ottenuto attraverso la procedura di misurazione; esso rappresenta una stima del punteggio vero;
- errore (e), ovvero la deviazione del punteggio osservato dal punteggio vero; esso è inosservabile; l'errore di misurazione, qui considerato come casuale e non sistematico, non è una proprietà della caratteristica misurata ma è il prodotto della misurazione effettuata sull'oggetto; esso è correlato in modo inversamente proporzionale all'affidabilità: maggiore è la componente di errore, peggiore è l'affidabilità.

Naturalmente potendo disporre di una procedura di misurazione perfettamente affidabile e valida e quindi di un punteggio x esente da errore, i due punteggi t e x sono perfettamente equivalenti; in caso contrario ogni singola misurazione (x) viene considerata composta di due parti:

punteggio osservato = punteggio vero + errore di misurazione

ovvero

$$x = t + e$$

All'interno della teoria classica della misurazione questa rappresenta l'*equazione fondamentale* che consente di definire formalmente il concetto di errore di misurazione come differenza tra il punteggio osservato e il suo corrispondente punteggio vero: in pratica essa formalizza l'impossibilità da parte di un particolare punteggio osservato di eguagliare il punteggio vero a causa di *disturbi* casuali.

In realtà l'equazione alla base della teoria classica della misurazione necessita di un'ulteriore specificazione:

$$x = t + e + B$$

dove B indica *bias*, ovvero l'errore sistematico che, influenzando i punteggi osservati, li rende meno affidabili.

Dall'equazione fondamentale si ricava che

$$e = x - t$$

La teoria classica della misurazione è stata molto criticata, soprattutto nell'ambito delle scienze sociali. Tali critiche verranno approfondite in seguito.

Il modello classico di misurazione prevede alcuni postulati:

a. il valore atteso dell'errore di misurazione è uguale a zero ovvero $E(e)=0$
b. il valore atteso del *punteggio osservato* è uguale al valore atteso del *punteggio vero*; infatti

$$E(x) = E(t) + E(e)$$

ma essendo $E(e)=0$ allora $E(t)=E(x)=t$

c. la correlazione e la covarianza tra punteggio vero ed errore sono nulle

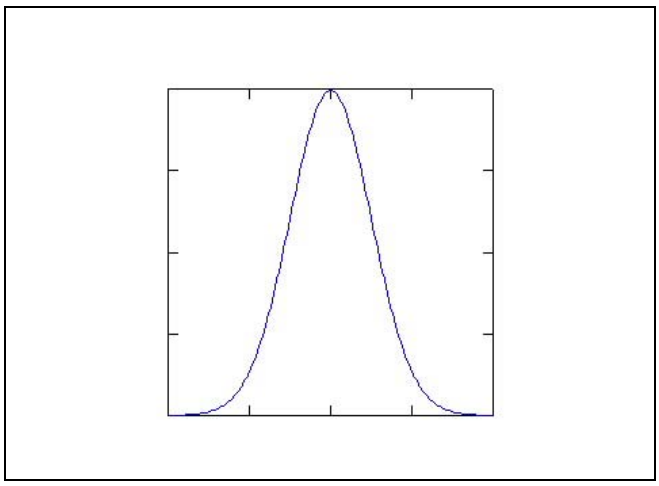
$$r_{te}=0 \quad \mathrm{cov}_{te}=0$$

Ciò deriva intuitivamente dall'idea che le dimensioni e i segni algebrici degli errori di misurazione non possono essere previsti dalle componenti dei punteggi veri delle osservazioni della variabile *x*;

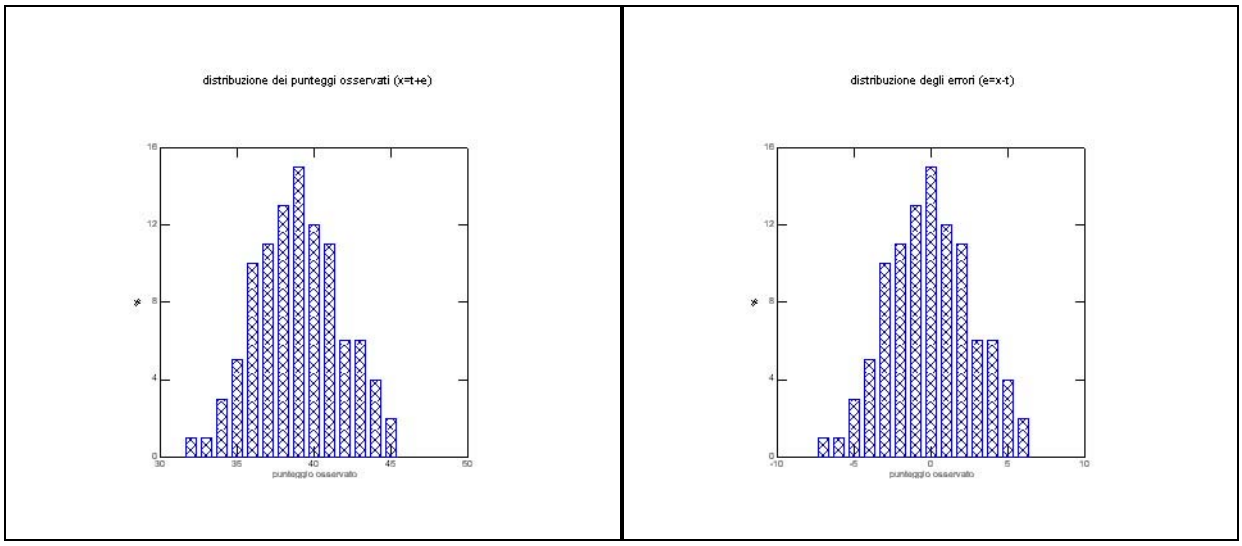
d. la varianza dei punteggi osservati è maggiore della varianza dei punteggi veri.

1.1.1.1 Distribuzione dell'errore

Potendo effettuare sullo stesso oggetto ripetute misurazioni della stessa caratteristica con la stessa procedura si può ipotizzare di ottenere, a causa della presenza dell'errore, risultati diversi che più o meno si avvicinano a quello vero.
La teoria classica della misurazione assume che la distribuzione di tutti i valori così registrati sia *normale* e che il valore vero sia quello che presenta la frequenza più alta o, meglio, che tale valore sia quello con la probabilità più alta di avvicinarsi a quello vero. In altre parole le misure rilevate si distribuiscono con maggiore frequenza intorno al valore del punteggio vero e simmetricamente al di sopra e al di sotto del punteggio vero (come nella seguente figura); in particolare gli errori positivi compensano gli errori negativi; maggiore è l'estensione della distribuzione dei punteggi ottenuti e l'oscillazione dei punteggi osservati intorno al punteggio medio (considerato stima del punteggio vero), maggiore è l'ampiezza dell'errore.



Analogamente si può assumere che all'aumentare del numero delle misurazioni, tali errori si annullino fino ad avere media zero. Se ne deduce che la distribuzione di frequenza, e quindi di probabilità, dell'errore di misurazione ha la stessa forma della distribuzione di frequenza (di probabilità) del punteggio osservato; conseguentemente *la funzione di distribuzione dell'errore e la funzione di distribuzione del punteggio osservato sono uguali*, come è possibile apprezzare dalle seguenti figure:



In realtà ciò può non essere sempre vero. Non necessariamente infatti gli errori in una direzione compensano gli errori nell'altra direzione, ovvero la distribuzione degli errori non è necessariamente simmetrica; si pensi, per esempio, alla variabile "peso dei bambini alla nascita", per la quale la distribuzione del punteggio osservato e degli errori è necessariamente asimmetrica; il discorso vale anche in campo sociale quando, per esempio, si misurano particolari sentimenti.

1.1.1.2 Valutazione e stima dell'affidabilità

Indichiamo

- la media e la varianza dei punteggi veri con rispettivamente μ_t e σ_t^2;
- la media e la varianza dei punteggi osservati con rispettivamente μ_x e σ_x^2;
- la media e la varianza dell'errori di misurazione con rispettivamente μ_e e σ_e^2;

Riscriviamo l'equazione fondamentale ($x=t+e$) in funzione delle corrispondenti varianze:

$$\sigma_x^2 = \sigma_{t+e}^2 = \sigma_t^2 + 2\mathrm{cov}_{te} + \sigma_e^2$$

Sapendo però che la correlazione e la covarianza tra il punteggio vero e l'errore è nulla $(\mathrm{cov}_{te} = 0)$:

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$$

ovvero la varianza dei *punteggi osservati* è uguale alla somma della varianza dei *punteggi veri* e di quella degli *errori*.

Dato ciò, il rapporto tra varianza dei punteggi veri e varianza dei punteggi osservati può essere definito come una valutazione del livello di *affidabilità* (indicato con rho_x) di *x* nel misurare *t*

$$\frac{\sigma_t^2}{\sigma_x^2} = rho_x$$

Sviluppiamo tale rapporto:

$$rho_x = \frac{\sigma_t^2}{\sigma_x^2} = \frac{\left(\sigma_x^2 - \sigma_e^2\right)}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

Sapendo che la varianza del punteggio osservato è maggiore di zero e che le varianze σ_t^2 e σ_e^2 rappresentano quantità non-negative e assumendo che almeno una di tali varianze è maggiore di zero, ne consegue che il valore di rho_x, prodotto dalla precedente equazione, presenta la proprietà secondo la quale i valori minimo e massimo sono rispettivamente:

- "0": totale mancanza di affidabilità: infatti quando tutta la varianza del punteggio osservato può essere attribuita all'errore, l'affidabilità è

$$1-\frac{\sigma_e^2}{\sigma_x^2}=1-\frac{1}{1}=0$$

- "1": totale affidabilità: infatti quando tutta la varianza del punteggio osservato può essere attribuita alla variazione nei punteggi veri, ovvero se non vi è alcun errore casuale nella misurazione, allora l'affidabilità è

$$1-\frac{\sigma_e^2}{\sigma_x^2}=1-\frac{0}{1}=1$$

Infine si può anche dire che la varianza dei punteggi veri è uguale al prodotto tra la varianza osservata e l'affidabilità della misura:

$$\sigma_t^2=\sigma_x^2 rho_x$$

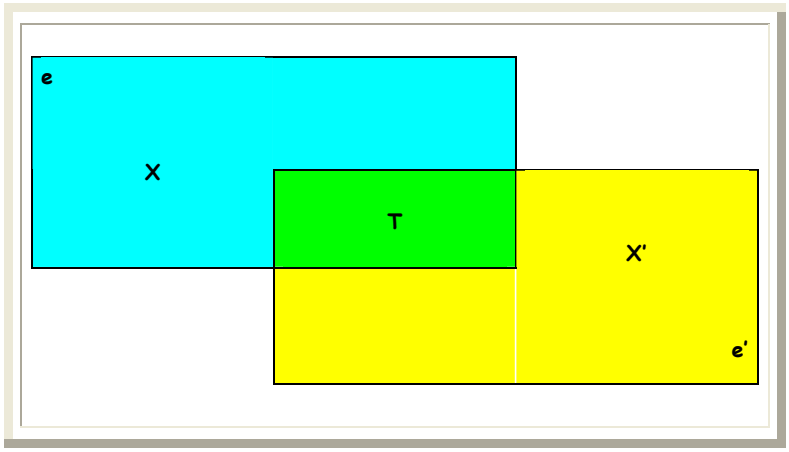
Quindi, se si conosce l'affidabilità e la varianza osservata di una misura, è possibile stimare la varianza del punteggio vero non osservato.
Il coefficiente di affidabilità rho_x, espresso in funzione delle varianze dei punteggi, rappresenta una quantità completamente astratta a causa dell'impossibilità di stimare la varianza del punteggio vero di un insieme di misurazioni. Il problema che si pone ora è quello di come stimare l'affidabilità attraverso i punteggi osservati.
In sintesi quindi la teoria classica suddivide la misurazione in due componenti ipotetiche tra loro non correlate: un punteggio latente vero ed un errore; ciò consente la valutazione dell'affidabilità: quando maggiore è la varianza del punteggio vero (tra oggetti: segnale) e minore è la varianza dell'errore (all'interno degli oggetti, rumore), migliore sarà la discriminazione tra oggetti. Da quanto detto finora, è possibile dedurre che l'affidabilità di un punteggio osservato è direttamente proporzionale alla correlazione tra *t* e *x* e inversamente proporzionale alla dimensione dell'errore (*e*). Siccome non è possibile conoscere il punteggio vero, non è possibile calcolare la correlazione r_{tx}; è possibile però effettuare una nuova misurazione (x') anch'essa stima di *t*. Le due misure ripetute sono distinte una dall'altra ma confrontabili.
La correlazione tra le due misure ripetute *x* e x' consente di stimare l'affidabilità di una misura. Le due misure ripetute (*x* e x') sono dette *parallele* se presentano:

1. uguali punteggi veri attesi per ciascun oggetto,
2. uguali varianze dell'errore o, in modo equivalente, errori standard di misurazione,

ovvero se risultano vere le seguenti relazioni

$$x=t+e \qquad x'=t'+e' \qquad t=t' \qquad \sigma_e^2=\sigma_{e'}^2$$

Ipotizziamo di calcolare la correlazione tra i due punteggi osservati $(r_{xx'})$; assumendo che gli errori di misurazione siano interamente indipendenti e quindi non correlati, il valore di tale correlazione dipende dalla parte comune ad entrambi e che può essere attribuita interamente al punteggio vero.
In definitiva quindi il livello di correlazione tra *x* e x' potrebbe rappresentare quanto riusciamo ad avvicinarci al punteggio vero *t* ovvero attraverso $r_{xx'}$ dovrebbe essere possibile ottenere una stima di r_{tx} e $r_{tx'}$.
Nella seguente figura, un rettangolo rappresenta il punteggio osservato *x* e il secondo l'altro punteggio osservato *x′*. L'area comune rappresenta quella porzione di entrambi i punteggi (dimensione dell'affidabilità) che è dovuta alla vera misurazione della caratteristica (*t*).



Verifichiamo ciò dal punto di vista formale. La correlazione tra misure parallele può essere espressa in funzione dell'errore, del punteggio vero e del punteggio osservato:

$$r_{xx'} = \frac{\sigma_{xx'}}{\sigma_x \sigma_{x'}} = \frac{\sigma_{(t+e)}\sigma_{(t+e')}}{\sigma_x \sigma_{x'}} = \frac{\sigma_t^2 + \sigma_{te} + \sigma_{te'} + \sigma_{ee'}}{\sigma_x \sigma_{x'}}$$

Avendo assunto che
- gli errori non sono correlati sia tra loro che con i punteggi veri,
- le deviazioni standard delle misure parallele sono uguali,

la correlazione tra due misure parallele è uguale al rapporto tra varianza dei punteggi veri e varianza dei punteggi osservati:

$$r_{xx'} = \frac{\sigma_t^2}{\sigma_x^2}$$

Tale risultato è importante in quanto consente di esprimere la varianza del punteggio vero inosservabile in termini di $r_{xx'}$ e σ_x^2, entrambi osservabili:

$$\sigma_t^2 = \sigma_x^2 r_{xx'}$$

ovvero la varianza del punteggio vero è uguale al prodotto tra varianza osservata e correlazione tra misure parallele. Ricordando che l'affidabilità è stata definita come $rho_x = \sigma_t^2 / \sigma_x^2$, ne segue che la stima dell'affidabilità non è altro che la correlazione tra misure parallele:

$$rho_x = \frac{\sigma_t^2}{\sigma_x^2} = \frac{\sigma_x^2 r_{xx'}}{\sigma_x^2} = r_{xx'}$$

Tale equazione rappresenta sicuramente un importante risultato nel tentativo di stimare l'affidabilità delle misure empiriche[1].

1.1.1.3 Schemi sperimentali per ottenere misure ripetute

Date tali premesse, per stimare l'affidabilità di una misurazione è necessario e fondamentale disporre di almeno due misure ripetute, dette misure parallele. Questo vuol dire che maggiore è il numero di misurazioni separate o parallele di un dato fenomeno, maggiore è l'accuratezza della

[1] Si può ulteriormente dire che la correlazione tra punteggi veri e punteggi osservati è uguale alla radice quadrata dell'affidabilità ovvero alla radice quadrata della correlazione tra misure parallele:

$$r_{tx} = \sqrt{rho_x} = \sqrt{r_{xx'}}$$

A partire dagli assunti della teoria classica della misurazione e la definizione di misure parallele si può dedurre che:

$$r_{xy} \le r_{tx} = \sqrt{rho_x} = \sqrt{r_{xx'}}$$

dove *y* rappresenta qualsiasi seconda misurazione. In altre parole, la correlazione tra una misura parallela e qualsiasi altra misura (per esempio una variabile criterio) non può essere superiore alla radice quadrata dell'affidabilità della misura parallela. Ciò significa che, la radice quadrata dell'affidabilità di una misura fornisce un limite superiore della sua correlazione con qualsiasi altra misura.

stima dell'affidabilità, come si è visto in precedenza.
Le misure ripetute devono però produrre punteggi confrontabili, ottenuti attraverso la definizione di uno schema sperimentale che può essere pianificato in modi diversi.
I diversi schemi sperimentali per la verifica dell'affidabilità della misurazione si differenziano tra loro rispetto al variare o meno dei seguenti elementi:

- strumento/i di misurazione utilizzati (o procedure di misurazione),
- rilevatore/i utilizzati,
- numero di rilevazioni effettuate,
- numero di momenti in cui avvengono le misure ripetute, ovvero la distanza temporale tra una misura e l'altra.

La combinazione dei diversi elementi produce i diversi schemi sperimentali.

SCHEMI SPERIMENTALI					STIMA DELL'AFFIDABILITA'			PROBLEMI
	numero di				TIPO DI VERIFICA	INDICATORE DI AFFIDABILITA'	APPROCCIO STATISTICO	
	strumenti	rilevatori	rilevazioni	momenti				
Test-retest *a*	uno	uno	due o più	uno	stabilità tra rilevazioni	coefficiente di stabilità (intra-rilevatore)	correlazione tra rilevazioni	procedure di rilevazione
Test-retest *b*	uno	uno	una	due o più	stabilità nel tempo	coefficiente di stabilità (intra-rilevatore)	correlazione tra momenti	verifica della stabilità nel tempo
Strumenti alternativi/paralleli	due o più	uno	una	uno	equivalenza tra strumenti	coefficiente di equivalenza	correlazione tra strumenti	verifica dell'equivalenza
Stabilità tra rilevatori	uno	più	una	uno	stabilità tra rilevatori	coefficiente di stabilità (infra-rilevatori)	correlazione tra rilevatori	procedura di rilevazione

Ricordiamo che le procedure per la stima dell'affidabilità per i diversi approcci sperimentali:

- conducono solo ad una approssimazione del valore di affidabilità (rho^2);
- comportano diversi assunti e producono risultati diversi;
- forniscono una stima campionaria del valore reale di affidabilità; diversi campioni possono condurre a stime anche molto diverse.

Test-retest

Il modello *test-retest* mira alla valutazione della stabilità della misurazione tra misurazioni diverse (*test-retest-a*) o lungo il tempo (*test-retest-b*); quest'ultimo è il modello sperimentale che presenta i maggiori problemi: sugli stessi oggetti viene effettuata una misurazione in più occasioni a distanza di un certo periodo di tempo. La procedura prevede in particolare le seguenti fasi:

a. misurazione di un oggetto con lo strumento *s*, nel tempo *t*, nel luogo *u*, secondo le modalità *m*, con il rilevatore *i*;
b. registrazione delle misurazione;
c. ripetizione delle fasi *a* e *b* nel tempo *t'*, rispetto allo stesso oggetto, nello stesso luogo *u* e con le stesse modalità *m*;
d. calcolo del coefficiente di correlazione tra le serie di punteggi ottenuti nelle diverse rilevazioni; tale coefficiente è detto *test-retest reliability coefficient* o *coefficiente di stabilità*; un'alta correlazione tra i punteggi ottenuti nelle diverse rilevazioni indica un'alta affidabilità.

Il modello *test-retest* si basa sull'assunto che la caratteristica da misurare non si modifichi e sia stabile lungo il tempo; in altre parole l'applicazione del modello *test-retest* è possibile solo per quei costrutti che si ipotizzano stabili nel tempo e risulta difficile quando si ipotizzano cambiamenti

negli oggetti che possono rendere i punteggi ottenuti nella prima rilevazione diversi dai punteggi della successiva rilevazione in modo non prevedibile e non necessariamente dipendente dall'errore del procedimento. Per poter interpretare i risultati in modo corretto, è necessario distinguere tra due tipi di stabilità:

- Stabilità a *breve termine* (o *breve periodo*), attribuibile a quei costrutti per la misurazione dei quali si assume che determinati fattori (come, nel caso della misurazione di individui, la *memoria*, intesa come tendenza a ricordare domande e risposte della precedente somministrazione, a ripetere errori o a fare le stesse considerazioni sugli item sui quali si sentono insicuri e incerti, ecc.) non influenzi il risultato. Nel caso in cui l'assunto non venga soddisfatto, l'osservazione di un'alta correlazione tra i punteggi ottenuti nelle due rilevazioni a breve termine non può essere attribuita necessariamente ad una alta affidabilità della misurazione ma anche ai pochi cambiamenti intervenuti negli oggetti nel breve intervallo di tempo.
- Stabilità a *lungo termine* (o *lungo periodo*), attribuibile solo a quei costrutti per la misurazione dei quali si assume che fattori individuali e naturali (crescita, apprendimento, modifiche di situazioni oggettive) non influenzino il risultato. Si tratta indubbiamente della stabilità più difficile da verificare. Nel caso in cui l'assunto non venga soddisfatto, la registrazione di una bassa correlazione tra i punteggi ottenuti nelle due rilevazioni a distanza di molto tempo non può essere attribuita necessariamente ad una minore affidabilità della misurazione ma ai possibili cambiamenti intervenuti negli oggetti[2].

Il principale problema di questo approccio è dovuto al fatto che è soggetto alle fluttuazioni nelle misurazioni prima-dopo dovute alla difficoltà di controllare sperimentalmente le condizioni nelle quali sono ottenuti i punteggi nelle due rilevazioni. Le variazioni, attribuite all'interazione *oggetto-occasione*, possono essere dovute a:

- *lo stato* dell'oggetto rispetto alla dimensione misurata che può cambiare lungo il tempo;
- *il cambiamento* dell'oggetto rispetto alla dimensione misurata; tale cambiamento può variare da oggetto ad oggetto;
- *la mancanza di indipendenza delle diverse misurazioni*, per esempio, nella misurazione di individui, la capacità di ricordare domande e risposte tra una somministrazione e l'altra; tale capacità varia da soggetto a soggetto;
- *le condizioni di rilevazione* che possono non essere costanti nelle due somministrazioni.

Inoltre, come abbiamo visto, l'effetto varia anche a seconda dell'estensione dell'intervallo di tempo trascorso tra le due rilevazioni (*problema della stabilità*).

Strumenti paralleli

Come abbiamo visto, il modello *test-retest* è di difficile applicazione pratica quando la caratteristica osservata non si presenta stabile negli oggetti lungo il tempo. Per poter isolare il fattore "tempo" è possibile adottare un diverso modello sperimentale che prevede una rilevazione in un unico momento ma con due o più strumenti considerati paralleli o equivalenti (*parallel forms* o *forme equivalenti*). In questo caso la stima dell'affidabilità viene fatta correlando i punteggi ottenuti con i

[2] Ciò richiede
- la teorizzazione della stabilità, a lungo e breve termine,
- l'identificazione di caratteristiche che nella maggior parte degli oggetti possono cambiare gradualmente nel tempo ma non in modo marcato.

Essa rappresenta un'importante questione che implica, per esempio nelle scienze sociali, la considerazione di problemi di sviluppo e declino di particolari capacità, di modifiche di atteggiamenti, ecc. La presenza di costrutti che possono cambiare nel tempo a livello individuale pone un ulteriore problema: la necessità di costruire strumenti che consentano la misurazione affidabile del cambiamento. Si tratta evidentemente di un altro problema di misurazione, sul quale torneremo più avanti.

due diversi strumenti definiti equivalenti. L'adozione di questo modello, consentendo un superamento dei problemi osservati nel precedente, presenta il vantaggio di omogeneità e uniformità di condizione sperimentale ma richiede che la rilevazione del primo strumento non modifiche o influenzi la rilevazione del secondo.

Assunti

La complessità e la difficoltà di applicazione di questo metodo stanno nella necessità di definire l'*equivalenza* tra due strumenti; due strumenti sono detti *paralleli* se è possibile registrare:

1. uguali punteggi osservati e uguali punteggi veri attesi per ciascun oggetto,
2. uguali varianze o, in modo equivalente, errori standard di misurazione $(\sigma_{x1}^2 = \sigma_{x2}^2)$,
3. uguali covarianze dei punteggi osservati e dei punteggi veri $(\sigma_{x_1x_2}^2 = \sigma_t^2)$,

dei due strumenti paralleli.

Non sempre gli strumenti possono essere considerati esattamente "paralleli" ma sono in grado di soddisfare assunti meno restrittivi che consentono, comunque, di stimare l'affidabilità degli strumenti. Di seguito sono sintetizzati a livello indicativo gli assunti di tali modelli

Modello	Strumenti paralleli (1, 2, 3, …)	Strumenti *tau-equivalenti* (1, 2, 3, …)
Assunti	- uguali punteggi veri attesi $\tau_1 = \tau_2 = \tau_3 = \ldots$ - uguali varianze $\sigma_{x1}^2 = \sigma_{x2}^2 = \sigma_{x3}^3 = \ldots$ - uguali covarianze osservate tra coppie di strumenti - uguali correlazioni registrate tra coppie di strumenti - uguali covarianze tra ogni strumento e un criterio - uguali correlazioni tra ogni strumento e un criterio	- uguali punteggi veri attesi $\tau_1 = \tau_2 = \tau_3 = \ldots$ - diverse varianze $\sigma_{x1}^2 \neq \sigma_{x2}^2 \neq \sigma_{x3}^3 \neq \ldots$ - uguali covarianze osservate tra coppie di strumenti - diverse correlazioni registrate tra coppie di strumenti - uguali covarianze tra ogni strumento e un criterio - diverse correlazioni tra ogni strumento e un criterio
Modello	Strumenti essenzialmente *tau-equivalenti* (i, j, k, …)	Strumenti con generici (i, j, k, …)
Assunti	- diversi punteggi veri attesi $\tau_i = \tau_j + c_{ij}$ - diverse varianze $\sigma_{xi}^2 \neq \sigma_{xj}^2 \neq \sigma_{xk}^3 \neq \ldots$ - uguali covarianze tra coppie di strumenti - diverse correlazioni registrate tra coppie di strumenti - uguali covarianze tra ogni strumento e un criterio - diverse correlazioni tra ogni strumento e un criterio	- diversi punteggi veri attesi $\tau_i = a_{ij}\tau_j + b_{ij}$ - diverse varianze $\sigma_{xi}^2 \neq \sigma_{xj}^2 \neq \sigma_{xk}^3 \neq \ldots$ - diverse covarianze tra coppie di strumenti - diverse correlazioni tra coppie di strumenti - diverse covarianze tra ogni strumento e un criterio - diverse correlazioni tra ogni strumento e un criterio

Stima dell'affidabilità

Se gli assunti di "parallelismo" possono essere soddisfatti allora è possibile stimare l'affidabilità dello strumento con questo modello; a tal fine è necessario dimostrare che la correlazione tra i due strumenti consente di stimare l'affidabilità; sappiamo che il coefficiente di correlazione può essere definito come il rapporto tra covarianza e il prodotto delle due deviazioni standard ovvero $r_{xy} = \sigma_{xy} / (\sigma_x \sigma_y)$; per questo possiamo scrivere

$$r_{x_1x_2} = \frac{\sigma_{x_1x_2}}{\sigma_{x_1}\sigma_{x_2}}$$

Sapendo che negli strumenti paralleli

$$\sigma_{x_1}^2 = \sigma_{x_2}^2 \qquad \text{e} \qquad \sigma_{x_1x_2} = \sigma_t^2$$

la precedente equazione diviene:

$$r_{x_1x_2} = \frac{\sigma_t^2}{\sigma_x^2} = rho_x$$

Quindi con strumenti paralleli è possibile stimare l'affidabilità correlando i punteggi ottenuti in un esperimento in cui entrambi gli strumenti siano stati somministrati allo stesso campione di soggetti. In questo caso il coefficiente di correlazione, detto *coefficiente di equivalenza*, indica quanto gli strumenti *paralleli* tendono a produrre gli stessi risultati[3]. I problemi che presenta questo approccio

[3] Il modello degli strumenti paralleli consente inoltre di stimare l'errore standard di misurazione. Ricordando che $\sigma_t^2 = \sigma_x^2 - \sigma_e^2$ e che $\sigma_e^2 = \sigma_x^2 - \sigma_t^2$ allora

sono attribuibili sia all'interazione *oggetto-occasione* che all'interazione *oggetto-contenuto dei due strumenti*.[4]

Misure ripetute con più rilevatori

Quando, nel tentativo di ottenere due misure, si utilizza lo stesso strumento ma con due diversi misuratori/rilevatori è necessario che il modello sperimentale venga definito in modo tale che consenta di verificare l'accordo tra i diversi osservatori. La messa a punto di una procedura di misurazione, infatti, non è solo un problema di taratura dello strumento ma richiede anche la "messa a punto" dei rilevatori, che non dovrebbero presentare differenze nelle misurazioni superiori a quelle che si otterrebbero per caso. In questo caso nel modello sperimentale si tengono costanti strumento, momento e rilevazione e si utilizzano più rilevatori; a seconda delle diverse modalità tali modelli assumono nomi diversi: *interindividual reliability*, *rater reliability*, *multi-judge reliability*, *interindividual costancy*, *intersubjectivity*, *coder/intercoder reliability* (Marradi, 1990). Per poter utilizzare i rilevatori è necessario prendere in considerazione alcuni elementi che nel caso della ricerca in campo sociale risultano essere particolarmente complessi. Infatti, per evitare si presentino problemi di confrontabilità delle misure ottenute, è importante che i rilevatori presentino particolari caratteristiche. Tali caratteristiche cambiano a seconda delle discipline e del tipo di misurazione che deve essere effettuata.

Combinazione tra diverse misure ripetute

I precedenti schemi sperimentali possono essere ridefiniti tenendo costanti solo due tra i quattro elementi; in questo caso si ottengono le seguenti combinazioni:

MODELLI SPERIMENTALI				*TIPO DI AFFIDABILITA' STIMATA*
Strumenti utilizzati	*Numero rilevatori*	*Numero rilevazioni*	*Numero momenti*	*verifica della*
uno	+	+	uno	Stabilità rilevazione-rilevatori
uno	uno	+	+	Stabilità rilevazioni nel tempo
+	+	uno	uno	Stabilità strumenti-rilevatori
+	uno	+	uno	Stabilità strumenti-rilevazioni
+	uno	uno	+	Stabilità strumenti nel tempo
uno	+	uno	+	Stabilità rilevatori nel tempo

La stima dell'affidabilità

Nei modelli più complessi è necessario poter distinguere tra più fonti di variabilità; questo vuol dire che nella maggior parte dei casi per validare una procedura di misurazione occorre distinguere tra:

- variabilità *intra-rilevatore* (errore nel rilevatore),
- variabilità *infra-rilevatori* (errore tra rilevatori),
- variabilità strumentale (errore strumentale),
- variabilità nelle applicazioni dello strumento (errore nel metodo di misurazione),

---

$$\sigma_e^2 = \sigma_x^2 - \sigma_x^2 rho_{x_1x_2} \qquad \sigma_e^2 = \sigma_x^2\left(1 - rho_{x_1x_2}\right)$$

Quindi l'errore standard è uguale alla radice quadrata della quantità posta a destra dell'equazione.

[4] Per poter controllare tali fonti di variazione, nel caso di misurazione di individui, è possibile pianificare l'esperimento secondo il metodo del *contro-bilanciamento*: si suddivide casualmente il campione di soggetti per la sperimentazione in due sottogruppi a ciascuno dei quali vengono somministrate le due forme:
I gruppo: forma *A* nella prima occasione, forma *B* nella seconda occasione;
II gruppo: forma *B* nella prima occasione, forma *A* nella seconda occasione.
In questo modo è possibile verificare se l'ordine di somministrazione influisce sul risultato.

cui va aggiunta la variabilità negli oggetti misurati.
Nei casi in cui occorre tener conto di più fonti di variabilità, la valutazione dell'affidabilità può essere affrontata con il modello dell'analisi della varianza; in questi casi è però necessario tenere presente che tale tipo di analisi richiede variabilità costante ovvero tutte le misure utilizzate devono avere uguale varianza, condizione molto difficile da soddisfare.
La variabilità presente (*intra* o *infra*) può risultare piccola se confrontata con l'alta variabilità degli oggetti; tale situazione non soddisfacendo la condizione richiesta falsa completamente i risultati dell'analisi della varianza.
Un'alternativa può essere quella di utilizzare la media del quadrato degli errori (media scarti al quadrato tra misurazioni).

1.1.2 La teoria della generalizzabilità

Secondo questa teoria (Bejar, 1983; Thompson, 2003), la rilevazione di qualsiasi caratteristica, definita operativamente da un indicatore, richiede l'utilizzo di più misure. Infatti, date le possibili fluttuazioni nelle misure, per ottenere misure stabili (e quindi affidabili) è necessario disporre di misure multiple, individuate a partire da una popolazione teorica di misure possibili, sono utilizzate per stimare la vera misura del concetto che interessa, controllando gli errori casuali. Il problema assume quindi la connotazione di significatività campionaria. Le misure infatti rappresentano un campione estratto dall'universo teorico di variabili, considerato infinitamente grande. Tale universo teoricamente non è osservabile e, conseguentemente, non è osservabile il punteggio e la misura reale dell'oggetto. Ciò significa che la misura reale può essere solo stimata. Il problema che deve essere affrontato a questo punto è la valutazione della quantità di errore presente in tale stima, come si fa per qualsiasi stima.
Le misure multiple devono soddisfare il requisito di indipendenza: ciascuna misurazione per ogni oggetto deve essere sperimentalmente indipendente dalle altre. L'indipendenza sperimentale garantisce l'assenza di correlazione tra gli errori scaturiti dalle misure multiple e la possibilità di stimare l'affidabilità.
Da tutto ciò si deduce che l'errore di misurazione è influenzato dalla dimensione del campione di misure multiple. Conseguentemente l'affidabilità della misura finale (e quindi dello strumento di misurazione), dedotta da un campione di misure multiple, dipende interamente

- dal numero di misure multiple: maggiore è il numero di misure multiple, minore è l'errore di misurazione, maggiore è l'affidabilità;
- dalla relazione tra le misure multiple e il concetto generale da misurare.

1.1.2.1 Stima del punteggio dell'universo

Utilizzando un campione di misure multiple, il primo obiettivo è quello di stimare il punteggio che si sarebbe ottenuto se fosse stato impiegato l'universo di misure. Tale stima, come qualsiasi stima, contiene un certo margine di errore.
E' possibile identificare principalmente due tipi di stima:

- *Proporzione di punteggi positivi/negativi*: come vedremo, il modo più semplice per definire una scala è quello dicotomico; in questi casi il modo più semplice per sintetizzare il punteggio di ciascun caso è quello di contare il numero di osservazioni "positive" o "negative" (dipende da ciò che si sta misurando) rilevato per il gruppo di misure multiple; tale numero convertito in proporzione rappresenta una stima *unbiased* della reale proporzione osservabile con l'intero universo di variabili. Ciò è vero anche quando le misure multiple sono tra loro eterogenee, purché il campionamento di variabili sia realmente casuale. Il significato di un punteggio così ottenuto dipende dalla validità di contenuto. La possibilità che su casi diversi vengano utilizzati misure multiple differenti conduce alla distinzione tra diversi disegni sperimentali di verifica:

- *nested*, quando a ciascun caso si applica un diverso gruppo di misure multiple casualmente estratto dall'universo,
- *crossed*, quando lo stesso campione casuale di misure multiple estratto dall'universo viene applicato a tutti i casi.

La semplicità dell'utilizzo della "proporzione" viene raggiunta sacrificando l'informazione che potrebbe essere utilizzata per migliorare la stima.

- *Stime di regressione*: le stime ottenute tramite regressione sono, dal punto di vista del calcolo, più complesse ma anche più precise in quanto utilizzano più informazioni. Le formule di regressione per la stima del punteggio dell'universo per i due schemi sperimentali:
  - *nested* $\mu_p = rho^2 X_{pI} - rho^2 \mu + \mu = rho^2 \left(X_{pI} - \mu\right) + \mu$
  - *crossed* $\mu_p = \xi rho^2 X_{pI} - \xi rho^2 \mu_I + \mu = \xi rho^2 \left(X_{pI} - \mu_I\right) + \mu$

  dove

μ_p	punteggio stimato per il caso *p*
X_{pI}	punteggio osservato per l'oggetto *p* per il gruppo *I* di misure
rho^2 , ξrho^2	coefficienti di generalizzabilità rispettivamente per lo schema *nested* e *crossed*
μ	punteggio medio complessivo per la popolazione di casi e l'universo di misure
μ_I	punteggio medio per la popolazione di casi rispetto al gruppo di misure *I*.

  La struttura della stima è la stessa in entrambi i casi. Ciascuna stima rappresenta una combinazione lineare del punteggio X_{pI} e le performance del gruppo al quale appartiene il caso *p*. Nel disegno *crossed*, dato che si utilizza lo stesso gruppo di variabili, la misura del risultato del gruppo è uguale a μ_I. Nel disegno *nested*, dato che per ogni oggetto si utilizza un insieme diverso di variabili, μ rappresenta la misura del gruppo.
  Tale stima assume che i punteggi osservati per ciascun caso abbiano distribuzione normale, con media corrispondente al punteggio vero individuale (punteggio dell'universo) e deviazione standard uguale per tutti i casi. Inoltre si assume che i punteggi veri (punteggi dell'universo) siano normalmente distribuiti.
  Le due equazioni non possono essere considerate stime effettuate secondo l'approccio Bayesiano in quanto non si dispone di alcun assunto sulla loro distribuzione.

1.1.2.2 Valutazione della generalizzabilità dei punteggi

Diversamente dalla teoria classica, che ammette solamente un coefficiente di generalizzabilità (*affidabilità*), nella teoria della generalizzabilità i punteggi possono disporre di molti coefficienti di generalizzabilità, a seconda dei fattori che, influenzando il procedimento di misurazione, vengono presi in considerazione. In ogni caso la definizione di generalizzabilità è la stessa del modello classico:

$$\frac{\text{var}\,ianza \cdot punteggio \cdot dell'universo}{\text{var}\,ianza \cdot punteggio \cdot osservato}$$

Se le diverse componenti della varianza corrispondenti alle diverse condizioni sono state precedentemente stimate, è possibile stimare la generalizzabilità dei punteggi sotto un dato insieme di condizioni utilizzando le stime delle componenti della varianza.
Nel caso più semplice, in cui un certo numero di casi è sottoposto ad una misura e la distinzione tra i due disegni sperimentali risulta ridondante, la generalizzabilità può essere stimata da:

$$\hat{\xi}^2_{rho} = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \hat{\sigma}^2_e}$$

dove

$\hat{\sigma}_p^2$ stima della varianza associata ai casi

$\hat{\sigma}_e^2$ stima della varianza ell'errore

$\hat{\xi}_{rho}^2$ coefficiente di generalizzabilità (quanto bene il punteggio osservato stima il punteggio dell'universo)

La stima delle due varianze, necessaria per poter stimare il coefficiente di generalizzabilità, è fatta attraverso l'analisi della varianza a due vie, nel modo seguente:

	Disegno	
	Nested	Crossed
$\hat{\sigma}_p^2$	$\frac{MS_p - \left(1 - \frac{n_i}{N_i}\right)MS_{i:p}}{n_i}$	$\frac{MS_p - \left(1 - \frac{n_i}{N_i}\right)MS_{ip}}{n_i}$
$\hat{\sigma}_e^2$	$\frac{\left(1 - \frac{n_i}{N_i}\right)MS_{i:p}}{n_i}$	$\frac{\left(1 - \frac{n_i}{N_i}\right)MS_{ip}}{n_i}$
$\hat{\xi}_{rho}^2$	$\frac{MS_p - \left(1 - \frac{n_i}{N_i}\right)MS_{i:p}}{MS_p}$	$\frac{MS_p - \left(1 - \frac{n_i}{N_i}\right)MS_{ip}}{MS_p}$

dove

MS_p quadrati medi per i casi

n_i numero di misure campionate

N_i numero di misure dell'universo

$MS_{i:p}$ quadrati medi per l'interazione casi – misure (disegno *nested*)

MS_{ip} quadrati medi per l'interazione casi – misure (disegno *crossed*)

Sono stati definiti anche altri coefficienti di generalizzabilità chiamati *rapporto segnale/rumore* (*signal-to-noise ratio*), basati sullo schema sperimentale *crossed*, in cui per ogni oggetto si utilizzano le stesse variabili. Secondo questo approccio (definito da Brennan e Kane nel 1977) il punteggio osservato può essere così definito:

$$X_{pI} = \mu + \pi_p + \beta_i + \pi\beta_i + e$$

dove

μ media complessiva nella popolazione di casi e nell'universo di misure

π_p effetto per il caso *p*

β_i effetto per la misura *i*

$\pi\beta_i$ effetto dell'interazione per l'oggetto *p* e l'item *i*

e errore casuale.

Si assume che ciascun effetto sia campionato in modo indipendente e che il valore atteso per ciascun effetto sia uguale a zero. Tale formulazione prevede la definizione di possibili errori di misurazione:

- se i punteggi osservati sono utilizzati per stimare i punteggi dell'universo, allora l'errore di misurazione (detto *rumore*) per il *p*-esimo caso è definito come la differenza (*delta*) tra il punteggio osservato e il punteggio dell'universo:

  $$\Delta_p = X_{pI} = \mu_p$$

  La varianza di Δ_p $\left(\sigma_{\Delta_p}^2\right)$, è detta *potenza del rumore*.

- se l'interesse riguarda il punteggio del caso non nel suo valore assoluto ma solo in relazione a

quello della popolazione, allora l'errore di misurazione è:

$$\delta_p = \left(X_{pI} - \mu_I\right) - \left(\mu_p - \mu\right)$$

dove il primo termine rappresenta la deviazione osservata per il *p*-esimo caso e il secondo rappresenta la deviazione vera; la loro differenza rappresenta l'errore di misurazione quando l'interesse è rivolto verso i punteggi assoluti. In questo caso la *potenza del rumore*, varianza di δ_p, è uguale a $\sigma^2_{\delta_p}$.

In entrambi i casi il *segnale* è definito come $\mu_p - \mu$ mentre la sua varianza è la potenza del segnale:

$$S = \varepsilon_p\left(\mu_p \mu\right)^2$$

dove ε_p indica l'attesa per i casi.

Il rapporto *segnale-rumore*, nel caso in cui l'interesse è rivolto verso i punteggi assoluti, sarà:

$$\lambda_1 = \frac{S}{\sigma^2_{\Delta_p}}$$

mentre nel caso in cui si è interessati alle deviazioni:

$$\lambda_2 = \frac{S}{\sigma^2_{\delta_p}}$$

La teoria della generalizzabilità è completata a livello logico dalla teoria classica; infatti, nelle applicazioni, le procedure di verifica dei modelli sono, in pratica, unificate.

1.1.3 La teoria del tratto latente

A differenza degli altri modelli, la *teoria del tratto latente*, nota anche come *Latent Trait Theory* fa una prima differenziazione tra variabili osservate e variabili latenti; queste ultime sono intese come costrutti teorici che non sono direttamente osservabili ma che hanno implicazioni per le relazioni tra le variabili osservate.

Secondo questa teoria, i concetti astratti (variabili latenti) non sono in grado di essere ridotti direttamente ad eventi o caratteristiche osservabili; per poter verificare empiricamente e direttamente un'ipotesi, definita sulla base di concetti astratti, è necessario definire le caratteristiche, possedute dai casi da studiare, osservabili e misurabili e che riflettono la natura dei concetti astratti considerati; tali caratteristiche misurabili empiricamente sono dette *indicatori* dei concetti, sviluppati in modo che siano fondati nel mondo empirico e che possano essere misurati.

La definizione di tali indicatori consente la verifica delle teorie e delle ipotesi rispetto a dati concreti. Riprendendo l'esempio precedente, il ricercatore sociale potrebbe, per ottenere una misura dello status sociale, utilizzare il numero di anni di scolarità per ciascun individuo; naturalmente non è difficile pensare ad altri possibili e alternativi indicatori per tale concetto.

Ipotesi che contengono concetti per i quali non si prevedono indicatori empirici sono destinate a rimanere speculazioni non verificate. Lo sviluppo di indicatori empirici adeguati costituisce un punto fondamentale nel processo della ricerca.

La definizione di tali indicatori non è però una questione sempre facile da affrontare; accanto alle teorie esplicite e alle ipotesi definite per spiegare il fenomeno sotto studio, vi è la necessità di verificare una seconda, spesso implicita, teoria. Secondo alcuni autori (come Blalock) la funzione di questa, detta anche *teoria ausiliaria*, è quella di specificare le relazioni tra il mondo teorico e il mondo empirico o, meglio, tra concetti astratti e i loro indicatori. Tali relazioni sono state indicate in molti modi in letteratura a seconda che si sia nell'ambito della filosofia della scienza, della sociologia o della psicometria (correlazioni epistemiche, regole di corrispondenza o definizioni operative); esse forniscono la base per specificare indicatori e verificare ipotesi astratte.

Un modo utile per rappresentare una serie di ipotesi astratte con la teoria ausiliaria necessaria per specificare le relazioni tra indicatori e concetti è quello di utilizzare la tecnica dei diagrammi

causali[5] in cui, in maniera semplice, è possibile rappresentare simultaneamente sia la teoria che si vuole verificare realmente (ipotesi) che la teoria ausiliaria per verificarla:

Costrutti:	X	$\Rightarrow$	Y		$\Rightarrow$	Z
Correlazioni epistemiche:	$\Downarrow$		$\Downarrow$	$\Downarrow$		$\Downarrow$
Indicatori:	x_1		y_1	y_2		z_1

Le frecce orizzontali rappresentano le relazioni tra i concetti che compongono la teoria reale, mentre le frecce verticali indicano le correlazioni epistemiche tra i concetti e gli indicatori che compongono la teoria ausiliaria. In questa rappresentazione la teoria ausiliaria risulta essere piuttosto semplice; infatti si assume che non vi sia alcun altro fattore che influenza gli indicatori al di fuori dei costrutti teorici.

La distinzione fatta tra i due livelli di analisi rappresenta un punto critico per comprendere i problemi fondamentali che si presentano nell'effettuare ricerche scientifiche soprattutto se si accetta l'impossibilità dello sviluppo di una scienza teorica in assenza di una osservazione adeguata dei concetti e delle loro interrelazioni.
L'approccio metodologico che consente di affrontare nella pratica tale teoria è quello dei *modelli ad equazioni strutturali*. Si può dire che la struttura teorica sottostante ai modelli ad equazioni strutturali sia la fusione di due tradizioni di ricerca: modelli a variabili latenti (psicologia) e modelli causali (sociologia).

Stima dell'affidabilità

L' affidabilità nell'ambito di questa teoria è definita dalla variazione osservata in ciascun indicatore x come composta da due fonti indipendenti:

1. variazione dovuta alla presenza della/e variabile/i latente/i (ξ), e condivisa dagli altri indicatori, e misurabile attraverso il peso fattoriale λ;
2. variazione specifica di ciascun indicatore (fattore unico, δ)[6].

Possiamo quindi dire che l'affidabilità di un indicatore nel misurare una variabile latente è rappresentata dalla proporzione di varianza di un indicatore attribuibile all'effetto della variabile strutturale; conseguentemente:

- ciascun coefficiente *lambda* al quadrato misura l'affidabilità dell'indicatore corrispondente,
- maggiore è la varianza dell'errore, minore è l'affidabilità.

Gli indicatori che misurano una variabile latente possono ottenere valori di *lambda* anche molto diversi tra loro, rivelando diversi livelli di accuratezza (affidabilità) nel misurare e rappresentare il concetto latente.
Tale approccio consente anche di verificare la validità della misurazione; in particolare le correlazioni di ciascun indicatore con le variabili latenti possono essere interpretate in termini di *validità fattoriale* o, meglio, di *composizione fattoriale* delle misure utilizzate; la registrazione della correlazione tra indicatori e costrutto consente di verificare la validità predittiva, di contenuto, di costrutto e convergente/divergente[7].

---

[5] Per essere precisi, i diagrammi causali – e le tecniche associate ad essi – in genere sono stati riservati ai modelli con alla base ipotesi di relazioni causali assunte *lineari*. Quindi $X \Rightarrow Y$ indica che Y è generato attraverso un processo causale che può essere rappresentato da $Y = a + bX$, dove a e b sono costanti. Quando le variabili sono in forma standardizzata, la funzione è ridotta a $Y = bX$ in quanto le medie delle variabili sono uguali a zero. La maggior parte dei modelli sono basate sull'assunto di linearità.

[6] Ciò ricorda quanto visto nell'ambito della teoria classica di misurazione secondo la quale il punteggio osservato è dato dalla somma di due componenti, punteggio vero ed errore.

[7] Infatti gli indicatori che risultano essere significativi rispetto ad una certa variabile latente tendono a correlare molto tra loro e meno con gli indicatori che definiscono gli altri costrutti latenti; in altre parole:

- gli indicatori che correlano molto tra loro riflettono lo stesso costrutto (validità convergente),
- gli indicatori che correlano molto poco tra loro riflettono costrutti diversi (validità discriminante).

1.2 LA VALIDITA': MODELLI E STRATEGIE DI VERIFICA

Uno dei momenti più difficili nel processo di sviluppo di una misura è quello dell'interpretazione dei punteggi ottenuti dalla misurazione. La registrazione di una significativa affidabilità rivela solo che lo strumento misura veramente qualcosa ma non dà alcuna informazione sulla natura di ciò che si sta misurando. La difficoltà deriva dal fatto che la validazione può essere verificata all'interno di un sistema di relazioni ipotizzate tra il costrutto di interesse e altri costrutti; tali relazioni possono essere di causa, effetto o di correlazione. Le misure empiriche sono utilizzate per verificare tali ipotesi; il supporto empirico per le ipotesi comporta la validità della misura. Le prove di validità implicano la verifica simultanea dell'ipotesi riguardante i costrutti e lo strumento (Campbell, 2001; Carmines, 1992; DeVellis, 1991; Ghiselli, 1964; Netemeyer, 2003; Nunnally, 1978; Spector, 1992).
Come si è già detto, molti costrutti nelle scienze sociali rappresentano astrazioni teoriche che non esistono nell'esperienza reale; per tale motivo gli indicatori che li rappresentano devono essere sottoposti a verifica di validità. La validazione di una misura può quindi essere vista come la verifica di una teoria. Il processo di validazione non può essere affrontato prima che non sia stato portato a termine quello di verifica dell'affidabilità.
Approfondiremo a questo punto le diverse **strategie** di verifica dei tre principali tipi di validità che più interessano la ricerca sociale e che corrispondono alle tre principali funzioni rispetto all'utilizzo di uno strumento di misurazione:

TIPO DI VALIDITA'	FUNZIONE
1. PREDITTIVA (*criterion validity*)	stabilire una relazione statistica con una particolare variabile
2. DI CONTENUTO (*content validity*)	rappresentare/rilevare un universo di contenuto specifico
3. DELLA STRUTTURA o TEORICA (*construct validity*)	misurare particolari tratti psicologici, rapporti sociali, ecc.

Anche se i tre tipi di validazione sono presentati separatamente nella pratica essi tendono ad essere complementari in quanto ciascuno di essi è sostenuto o riceve informazioni dall'altro.

La validità predittiva

Il modello statistico e la logica utilizzati per studiare e verificare la validità predittiva sono relativamente semplici, anche se dal punto di vista applicativo il procedimento può risultare piuttosto complesso.
Alla base del modello di verifica vi è il confronto tra la misura da validare ed altre misure dette *criteri*[8]. Tale approccio prevede e richiede la definizione e la verifica di una ipotesi riguardante le relazioni tra il costrutto ed altri costrutti. Un tale tipo di studio può riguardare sia strumenti sviluppati, allo scopo di valutare una teoria esistente (la situazione migliore), che strumenti sviluppati, allo scopo di valutare un costrutto che non fa parte di una teoria esistente.
L'indicatore operativo del grado di corrispondenza tra lo strumento e il criterio è di solito rappresentato dalla dimensione della loro correlazione; in altre parole il confronto tra misura e

[8] Tra i diversi tipi di criterio esistenti, secondo alcuni autori, è possibile distinguere le seguenti tipologie:
1. *product criteria*, rappresentati da variabili legate a particolari prestazioni e risultati di alcune specifiche attività (per esempio numero di lettere battute a macchina, numero di scatole ispezionate, ecc.);
2. *action criteria*, che riguardano l'attività stessa (per esempio misurazioni di particolari prestazioni: velocità o numero di errori nell'eseguire qualcosa, ecc.) e che sono utilizzabili quando il risultato di una certa attività non è rappresentato da un prodotto tangibile;
3. *subjective judgment criteria*, legati, a differenza delle precedenti tipologie oggettive, a giudizi di osservatori, le cui valutazioni soggettive possono essere tra loro correlate utilizzando metodi basati su confronti accoppiati, di ordinamento (*ranking* e *rating-scale*).

criteri è realizzato e misurato in termini di analisi della correlazione, estesa anche al caso multivariato: il livello di correlazione ottenuto specifica il grado di validità; per questo motivo in tale contesto il coefficiente di correlazione viene indicato spesso come *coefficiente di validità*: se i risultati statistici conducono a decisioni e scelte ragionevoli allora il gruppo di variabili in questione ha validità predittiva[9].
I complessi ambiti della ricerca sociale non consentono una stima accurata e, ragionevolmente, producono il più delle volte modeste correlazioni tra strumento e criterio esterno[10].
La misura e il criterio esterno possono essere misurati in momenti diversi; in relazione a ciò è possibile distinguere tra:

- criterio *post-diction*: quando il criterio esterno viene misurato dopo l'applicazione dello strumento da validare;
- criterio *concomitante*: quando il criterio esterno viene misurato contemporaneamente allo strumento da validare;
- criterio *predittivo*: quando il criterio esterno viene misurato prima dell'applicazione dello strumento da validare.

Tale distinzione riguarda solo l'applicazione dello strumento, in quanto la logica e le procedure di validazione nei tre tipi non sono diverse: la validità predittiva è determinata solamente dal grado di corrispondenza tra i dati, indipendentemente da quando sono stati rilevati. E' importante riconoscere che l'utilità scientifica e pratica della validazione predittiva dipende sia dalla misurazione del criterio che dalla qualità dello strumento di misurazione stesso.
E' importante inoltre che gli studi di validità presentino i seguenti due aspetti:

a. la teoria sulla quale sono basate le ipotesi deve essere solida;
b. il criterio utilizzato deve essere misurato in modo affidabile.

Come si è visto la logica alla base di tale validità è abbastanza semplice e chiara; è però altrettanto chiaro che le procedure di validazione predittiva presentano una limitata utilità in quanto in molte situazioni non esiste alcun criterio con cui valutare in modo ragionevole la misura. Più è astratto il concetto, più difficile è trovare un criterio appropriato per la valutazione della misura.
La validità predittiva diviene particolarmente importante quando a livello pratico la misura da validare ha funzioni *operative* e *decisionali* (assunzione di personale, creazione di servizi sulla base di previste necessità di un gruppo di cittadini, ecc.).

La validità di contenuto

Per la validità di contenuto è essenziale che la misura da validare sia accettata come indicatore che definisce l'universo di contenuto di riferimento.
Tale validità viene verificata misurando il grado di rappresentatività degli indicatori che compongono la misura: se gli indicatori rappresentano un buon campione dell'universo degli indicatori dell'ambito da indagare, ovvero riflettono uno specifico dominio di contenuto, lo strumento ha una buona validità di contenuto.
La verifica della validità di contenuto è particolarmente importante e deve essere effettuata al momento della costruzione dello strumento e quindi prima dell'analisi vera e propria; questa dovrebbe fornire solo informazioni aggiuntive su tale validità, ovvero dovrebbe fornire solo un

[9] Secondo un'altra strategia (*validità attraverso gruppi noti*), per tale validità si verifica l'ipotesi che determinati gruppi di soggetti ottengano punteggi più alti rispetto ad altri; la significatività della differenza tra i gruppi viene verificata calcolando la media dei punteggi ottenuti sullo strumento per ciascun gruppo e confrontando attraverso strumenti statistici le medie osservate (*t* di *Student*, analisi della varianza).

[10] Un esempio a riguardo può essere rappresentato da un test di maturità scolastica. Per poterlo completare occorre attendere la fine dell'anno. Relativamente a questo esempio occorre osservare che il profitto scolastico non dipende solo dalle capacità dello studente ma anche dalle caratteristiche degli insegnanti e della scuola: quindi anche per il criterio esterno si pongono problemi di validità e affidabilità. In questo caso si può procedere dimostrando la non-validità del criterio (confrontando i giudizi di più insegnanti sui risultati in una certa materia: se i giudizi divergono il criterio viene considerato non-valido).

sostegno e una giustificazione statistica.
Per verificare la validità di contenuto occorre controllare e valutare i seguenti standard:

- Qualità e della rappresentatività degli indicatori: per valutare ciò è necessario identificare un profilo dettagliato delle dimensioni e dei concetti che devono essere rilevate. Il profilo deve essere collegato direttamente con l'ipotesi di ricerca che deve guidare e condurre alla formulazione e alla costruzione dello strumento. Tale valutazione è comunque complicata dal fatto che è difficile, se non impossibile, valutare il campionamento del contenuto.
- Adeguatezza e sensibilità della costruzione degli indicatori: per valutare ciò è possibile utilizzare alcune verifiche:
    - un alto livello di accordo (consistenza interna) tra gli indicatori dovrebbe rilevare che gli indicatori tendono a misurare qualcosa in comune; tale verifica però non garantisce la verifica della validità della costruzione: infatti gli indicatori che concorrono ad un unico contenuto possono anche misurare caratteristiche diverse non correlate;
    - il confronto tra diverse applicazioni dello strumento (per esempio prima e dopo una certa sperimentazione) consente di verificare se lo strumento misura l'evoluzione di un certo *fenomeno*; tale verifica richiede comunque un modello complesso e di difficile applicazione nel campo delle scienze sociali;
    - un alto livello di correlazione tra misura da validare e altri indicatori consente di concludere che lo strumento misura ciò che si vuole misuri; in realtà la registrazione e l'osservazione di alti livelli di correlazione non è garanzia di validità di contenuto: gli strumenti confrontati potrebbero misurare male nello stesso modo la stessa dimensione.

Come si è visto tali standard non possono essere sempre giudicati e valutati in modo adeguato; inoltre, nonostante i tentativi di applicare metodi statistici di verifica, la validità di contenuto è stabilita principalmente con altri criteri; la statistica può aiutare ma la validità di contenuto riguarda principalmente le caratteristiche riguardanti il contenuto e il modo con cui è presentato. E' per questo che è più corretto mettere a confronto i giudizi di diversi esperti in materia. In ogni caso, indipendentemente dai risultati dell'analisi, la decisione finale sulla validità di un indicatore o di un insieme di indicatori spetta al ricercatore.

La validità teorica o della struttura

La validità della struttura è direttamente collegata alle teorie di riferimento per il ricercatore; in particolare essa riguarda il rapporto tra lo strumento e le dimensioni e i concetti da misurare e presuppone una solida base di preparazione, studio e sperimentazione sull'argomento. Da questo punto di vista un gruppo di variabili è valido se in esso trovano conferma le teorie di riferimento. Le teorie variano rispetto alla *specificità* e alla *grandezza* dell'area che gli indicatori devono ricoprire. Come sappiamo in alcuni casi il «dominio» è talmente piccolo che uno qualsiasi degli indicatori identificati è sufficiente a coprire l'area; maggiore è l'area da ricoprire, maggiore è la difficoltà nel definire quali indicatori appartengono all'area.
Lo strumento ritenuto valido nella struttura può essere utilizzato da coloro che accettano l'interpretazione teorica di ciò che esso rappresenta. Naturalmente il costrutto sottostante continua a rimanere un'entità teorica. La concezione della natura del costrutto e le ragioni delle relazioni con altri costrutti sono basate su una struttura teorica che può essere successivamente sostituita da una nuova struttura che ridefinisce il costrutto.
La verifica della validità della struttura è un problema che non può essere risolto statisticamente ma logicamente; comunque, anche se la verifica della validità teorica è essenzialmente di tipo logico, questa è affiancata da una verifica statistica. Tale verifica mira a determinare quanto gli indicatori tendono a misurare la stessa cosa o cose diverse; ciò rappresenta una condizione necessaria ma non sufficiente per verificare la validità teorica.
Gli strumenti statistici utilizzati per studiare la validità teorica possono essere espressi in termini di

- *consistenza interna*, ovvero sulla base della tendenza di misure diverse a correlare molto tra loro e ad essere influenzate allo stesso modo da trattamenti sperimentali;

- determinazione della correlazione della misura del costrutto con altre misure relative ad altri costrutti;
- *analisi fattoriale*.

Il prodotto finale di tale processo dovrebbe condurre ad un costrutto

a. ben definito in termini di osservazione,
b. ben rappresentato in termini di variabili osservabili,
c. eventualmente correlato con altri costrutti.

E' comunque molto importante tener presente che la validità di costrutto può essere sostenuta ma mai provata; è infatti possibile che successivi lavori e studi possano dare nuove interpretazioni dei risultati e verificare nuovi modelli.

Con il seguente schema cerchiamo di sintetizzare quali possono essere le finalità dei diversi approcci alla validità nel caso di indicatori soggettivi:

PROBLEMA	PROCEDIMENTO PER LA STIMA DELLA VALIDITA'	APPLICATO SOPRATTUTTO A
VALIDITA' PREDITTIVA E CONCORRENTE		
Come si correlano criterio e punteggio?	Somministrato lo strumento, dopo un certo periodo si raccolgono i dati relativi al criterio	1. strumenti per la selezione di soggetti; 2. strumenti usati in campo scolastico e clinico 3. strumento usato per valutazioni complesse
VALIDITÀ DI CONTENUTO		
Gli strumenti riescono ad esplorare e penetrare le situazioni che dovrebbero valutare?	Si analizzano gli item e le risposte richieste, in relazione all'oggetto che lo strumento dovrebbe valutare.	1. strumenti usati per valutare i programmi educativi. 2. Tecniche di osservazione per lo studio del comportamento abituale
VALIDITÀ DI COSTRUTTO		
In che modo interpretare i punteggi ottenuti? Lo strumento misura proprio le variabili che dovrebbero misurare?	Si elaborano delle ipotesi sul significato da attribuire ai punteggi dello strumento, stabilendo il grado secondo cui i punteggi alti debbano differire da quelli bassi o quali variabili possono alterarli. Ciascuna di queste ipotesi viene verificata singolarmente.	1. strumenti usati per valutare intelligenza o personalità sia a fini diagnostici e di orientamento individuale, che per ricerca e valutazione scolastica.

Verifica contemporanea della validità convergente e discriminante

Come abbiamo visto, la *validità*

- *convergente* è quella verificata dall'osservazione di un'alta correlazione osservata tra la misura da validare e altre misure di costrutti teoricamente legati al primo, e
- *discriminante* è quella verificata dalla osservazione di una modesta o nulla correlazione tra la misura da validare e altre misure di altri costrutti non legati a quello misurato.

Questi due tipi di validità possono essere studiati in relazione tra loro, quando è possibile fare le due ipotesi contemporaneamente. L'idea di base è che un costrutto debba essere correlato maggiormente con se stesso che con altri.

Nella validità convergente ci si aspetta un'alta correlazione tra misure di costrutti teoricamente legati. Idealmente dovrebbero registrare circa lo stesso livello di affidabilità. Siccome si ipotizza sempre un certo livello di errore che abbassi il livello di affidabilità, tali correlazioni osservate non raggiungono mai il massimo livello.

A metà del Novecento Cronbach ed altri elaborarono, nell'ambito della definizione degli Standard per lo sviluppo di misure in psicologia per l'*Associazione Americana di Psicologia*, un particolare approccio alla verifica della validità, la *rete nomologica* (Trochim, 2000); tale rete di verifica dovrebbe comprendere

a. la struttura teorica relativa a ciò che si cerca di misurare,
b. la struttura empirica relativa al come si cerca di misurare,
c. la specificazione dei legami tra le due precedenti strutture.

Tale rete è basata su un certo numero di principi che guida il ricercatore nella verifica; ciò che l'approccio cerca di fare è di collegare l'ambito concettuale/teoretico con quello osservativo. Se da una parte l'idea di rete nomologica può funzionare a livello filosofico, dall'altra essa non fornisce una metodologia pratica e utilizzabile per verificare realmente la validità.
Successivamente Campbell e Fiske (1959) hanno sviluppato un particolare approccio, definito *Multitrait-Multimethod Matrix* (*MTMM*), che ha dato un impulso verso una particolare metodologia di verifica della validità. Il *MTMM* consente di indagare simultaneamente la validità convergente e la validità discriminante; l'applicazione del *MTMM* richiede, come vedremo nel paragrafo successivo, che vengano misurati almeno due costrutti e che ciascuno di essi sia misurato con almeno due metodi distinti.

1.3 LA VALUTAZIONE SIMULTANEA DELLA VALIDITA' E DELL'AFFIDABILITA'

Secondo alcuni autori (Marradi, 1980) l'analisi della consistenza interna di un gruppo di indicatori è "una proprietà intermedia tra attendibilità e validità [...] e può essere misurata, dato che non fa riferimento a qualcosa di esterno alla matrice dei dati; può essere indizio di validità ma non certo una prova". Ciò ha indotto ad un certo scetticismo nei confronti di misure di affidabilità e validità basate sulla congruenza interna di batterie o altri gruppi di indicatori operativamente definiti in maniera troppo simile.
I primi a mettere in luce il fatto che nell'analisi della consistenza interna gli aspetti di affidabilità e di validità sono inestricabilmente legati sono stati Campbell e Fiske (1959; Campbell, 2001; Spector, 1992; Sullivan, 1981). Seguiamo il loro ragionamento; essi si sono chiesti se la consistenza interna di un gruppo di indicatori fosse dovuta alla prossimità semantica o piuttosto solo alla somiglianza delle loro definizioni operative[11]. Per affrontare tale problema sarebbe necessario utilizzare tecniche di rilevazione più raffinate.
Relativamente ad un determinato concetto, si dovrebbero scegliere quegli indicatori che possano essere registrati con definizioni operative quanto più possibile diverse l'una dall'altra (fonte delle informazioni, forma delle domande, tipo di risposte precodificate, strumenti e situazioni in cui registrare le risposte, ecc.). Se le correlazioni tra gli indicatori dello stesso concetto registrati con definizioni operative diverse fossero considerevolmente più alte delle correlazioni tra gli indicatori di concetti diversi registrati con definizioni operative simili, il problema sarebbe risolto. Si realizzerebbe in questo modo una validazione *convergente* perché indicatori formalmente simili ma legati semanticamente a concetti diversi risulterebbero scarsamente congruenti tra loro. Quindi il requisito necessario per verificare la congruenza tra indicatori deve derivare dalla loro vicinanza semantica e non dalla somiglianza tra le relative definizioni operative.
Tale approccio mette in crisi la capacità delle scienze sociali di conoscere il proprio oggetto senza alterarlo; la definizione operative appare più che un modo per registrare gli stati degli oggetti su una certa proprietà, un meccanismo che contribuisce a creare gli stati che registra.
Tale approccio ha condotto a ad un generale adeguamento alla richiesta di dare definizioni operative quanto più differenti degli indicatori dello stesso concetto; tale adeguamento richiede spesso una certa creatività metodologica e pone problemi pratici non sempre superabili.

[11] Ciò si pone per esempio quando in questionario le domande sono proposte in successione dallo stesso intervistatore e l'intervistato deve scegliere per ciascuna domanda una risposta in un limitato arco di alternative pre-codificate, favorendo il presentarsi del *response-set*.

1.3.1 Metodo di Campbell e Fiske

Come si è già detto, quello della validità rappresenta il problema più critico nella ricerca empirica[12] e riguarda il problema di come gli indicatori misurino il concetto astratto derivato dalla teoria; tale problema non consente di avere sicurezze e rimane in un qualche modo discutibile e incerto. Secondo Campbell e Fiske (1959; Campbell, 2001; Spector, 1992) è possibile confidare sulle caratteristiche delle misure di cui si dispone e su come esse correlano tra loro. Se si misura un certo tratto, o concetto astratto, con diverse metodologie tra loro molte diverse e se tali diverse procedure producono risultati che sono abbastanza simili, è possibile confidare nella validità delle misure, potendo affermare, in modo ragionevole e convincente, che ciascuno dei metodi produce una misura valida della caratteristica in questione. Con l'approccio definito da Campbell e Fiske i concetti di *affidabilità* e di *validità* trovano la seguente nuova definizione:

l'affidabilità rappresenta l'accordo tra due tentativi di misurare lo stesso tratto attraverso metodi tra loro il più possibile simili. La validità è rappresentata dall'accordo tra due tentativi di misurare lo stesso tratto attraverso metodi tra loro il più possibile diversi.

Tale definizione contiene le basi logiche della metodologia da loro proposta.

Nella teoria classica della misurazione, come abbiamo visto, si fa distinzione tra punteggi veri e punteggi osservati; sulla base degli assunti riguardanti tali punteggi è definito un coefficiente di affidabilità. Per poter calcolare i coefficienti basati sui metodi tradizionali è necessario, come sappiamo, poter disporre di almeno due misurazioni; in altre parole la valutazione dell'affidabilità di una misura è realizzabile attraverso diversi schemi sperimentali che alla base presentano la nozione di ripetibilità ovvero di applicazione di *metodi tra loro il più possibile simili*.
Con l'approccio proposto da Campbell e Fiske ci si sposta dalla nozione di pura affidabilità (attraverso metodi il più possibile simili tra loro) alla nozione di validità (metodi il più possibile diversi tra loro).
Per poter definire l'affidabilità e la validità, i due autori hanno ridefinito le nozioni di validità convergente e discriminante:

- metodi diversi di misurazione possono convergere verso la misurazione dello stesso tratto (*validità convergente*);
- stessi metodi di misurazione possono non correlare perché misurano tratti diversi (*validità discriminante*).

Il metodo da loro proposto può essere formalmente così presentato.
Il valore y_{ij} dell'indicatore *i*-esimo raccolto con il metodo *j*-esimo può essere decomposto in due componenti:

- una stabile (T_{ij}), corrispondente al punteggio vero nella teoria classica,
- una casuale (e_{ij}).

La risposta e le sue due componenti si legano tra loro nel modo seguente:

$$y_{ij} = h_{ij}T_{ij} + e_{ij}$$

dove

h_{ij} livello di relazione tra componente stabile (punteggio vero) e risposta.

Il punteggio vero può essere ulteriormente decomposto in tre componenti:

- una che rappresenta il punteggio sulla variabile che interessa F_i,
- una dovuta al metodo utilizzato M_j,

[12] In un certo senso tale problema è simile a quello dell'attribuzione di una etichetta a ciascun fattore nell'analisi fattoriale.

- una dovuta alla combinazione di metodo e tratto u_{ij};

dopo la standardizzazione ciò conduce alla seguente equazione:

$$T_{ij} = b_{ij}F_i + g_{ij}M_j + u_{ij}$$

dove

b_{ij} livello di relazione tra la variabile latente di interesse e il punteggio vero

g_{ij} effetto della componente metodo sul punteggio vero

Sapendo che tutte le variabili, eccetto i termini di disturbo, sono standardizzate e che il metodo e i fattori non sono correlati, i coefficiente h_{ij}, b_{ij} e g_{ij} indicano la forza delle relazioni tra le variabili nel modello; a tali coefficienti è stata data una speciale interpretazione:

- h_{ij} è chiamato *coefficiente di affidabilità*; il quadrato di tale coefficiente rappresenta una stima dell'affidabilità (*test-retest* nel senso della teoria classica);
- b_{ij} è chiamato *coefficiente di validità del punteggio vero* in quanto il quadrato di tale coefficiente rappresenta la varianza spiegata nel punteggio vero attribuita alla variabile cui siamo interessati;
- g_{ij} è chiamato *effetto del metodo* in quanto il quadrato di tale coefficiente rappresenta la varianza la varianza spiegata nel punteggio vero attribuita al metodo usato;
- la varianza di u_{ij} più g_{ij}^2 a volta è chiamata *invalidità* in quanto è la varianza spiegata nel punteggio vero che non è dovuta alla variabile di interesse.

1.3.2 Matrice Multi-Trait Multi-Method

Per poter stimare l'errore di misurazione come è stato definito qui è necessario misurare almeno tre tratti con tre diversi metodi. Tale disegno, introdotto da Campbell e Fiske, è chiamato *MultiTrait-MultiMethod* (*MTMM*) (Campbell, 1959, 2001; Sullivan, 1981) in quanto ciascun tratto di un numero di tratti (costrutti) è misurato con un numero di metodi differenti. Per analizzare i dati tratti dal disegno sperimentale *MTMM* è stato definito un modello causale. Oltre la distinzione tra punteggi veri e punteggi osservati, sono introdotti fattori latenti sia per il fattore "tratto" che per il fattore "metodo". Si assume che

- i fattori "tratto" siano tra loro correlati $(rho(F_1F_2))$,
- i fattori "metodo" non siano correlati tra loro,
- i fattori "metodo" non siano correlati con i fattori "tratto".

Il requisito principale per poter applicare la matrice *multitrait-multimethod* (*MTMM*) è quello di poter disporre di almeno di tre differenti caratteristiche, ciascuna delle quali misurata con metodi tra loro molto diversi[13]. Per poter procedere occorre quindi innanzi tutto identificare i tratti e i metodi.

La costruzione della matrice

- Identificazione dei tratti: supponiamo di voler studiare le valutazioni individuali riguardo ai due partiti politici misurando i seguenti tratti:
  a. valutazione del partito,
  b. ideologia politica (per esempio *liberale-moderata-conservatrice*),
  c. livello di coinvolgimento politico.
- Identificazione dei metodi: per misurare ciascuno dei tratti individuati è possibile utilizzare i medesimi tre metodi:

[13] Come per altri approcci qui proposti, le caratteristiche da misurare possono essere rappresentate da atteggiamenti, comportamenti e possono riguardare sia individui che aggregazioni quali istituzioni, organizzazioni, città, nazioni, ecc.

1. *scala di atteggiamento* o *questionario*;
2. *osservazione partecipata*, realizzata trascorrendo due o tre giorni con ciascun soggetto durante una campagna elettorale, registrando fedelmente qualsiasi accenno ad uno dei partiti; successivamente sarà possibile registrare il numero di affermazioni positive e negative date da ciascun soggetto e dirette a ciascun partito;
3. utilizzazione degli *informatori* che consiste nel chiedere ad amici e a parenti quali sono secondo loro le valutazioni riguardanti ciascun soggetto.

Tale quadro legittima l'utilizzo dell'approccio *MTMM*. Il primo passaggio dell'analisi finalizzata a stimare l'affidabilità, la validità e gli effetti dei metodi è quello di calcolare la matrice di correlazione[14] per le misure utilizzate. Tale matrice *MTMM* potrebbe essere rappresentata nel modo seguente:

METODI	⇒	1			2			3		
⇓	TRATTI⇒ ⇓	A	B	C	A	B	C	A	B	C
1	A	$r_{A_1A_1}$								
	B	$r_{B_1A_1}$	$r_{B_1B_1}$							
	C	$r_{C_1A_1}$	$r_{C_1B_1}$	$r_{C_1C_1}$						
2	A	$r_{A_2A_1}$	$r_{A_2B_1}$	$r_{A_2C_1}$	$r_{A_2A_2}$					
	B	$r_{B_2A_1}$	$r_{B_2B_1}$	$r_{B_2C_1}$	$r_{B_2A_2}$	$r_{B_2B_2}$				
	C	$r_{C_2A_1}$	$r_{C_2B_1}$	$r_{C_2C_1}$	$r_{C_2A_2}$	$r_{C_2B_2}$	$r_{C_2C_2}$			
3	A	$r_{A_3A_1}$	$r_{A_3B_1}$	$r_{A_3C_1}$	$r_{A_3A_2}$	$r_{A_3B_2}$	$r_{A_3C_2}$	$r_{A_3A_3}$		
	B	$r_{B_3A_1}$	$r_{B_3B_1}$	$r_{B_3C_1}$	$r_{B_3A_2}$	$r_{B_3B_2}$	$r_{B_3C_2}$	$r_{B_3A_3}$	$r_{B_3B_3}$	
	C	$r_{C_3A_1}$	$r_{C_3B_1}$	$r_{C_3C_1}$	$r_{C_3A_2}$	$r_{C_3B_2}$	$r_{C_3C_2}$	$r_{C_3A_3}$	$r_{C_3B_3}$	$r_{C_3C_3}$

I valori della matrice rappresentano coefficienti di correlazione; l'analisi di tale matrice richiede, a causa del diverso significato che le correlazioni assumono, la suddivisione delle correlazioni in quattro diversi raggruppamenti:

A. *Diagonale di affidabilità: same-trait same-method*, composto dalle correlazioni tra due tentativi di misurare la stessa caratteristica utilizzando lo stesso metodo; all'interno della matrice esse corrispondono ai valori $r_{A_1A_1}$, $r_{B_1B_1}$, $r_{C_1C_1}$, $r_{A_2A_2}$, $r_{B_2B_2}$, $r_{C_2C_2}$, $r_{A_3A_3}$, $r_{B_3B_3}$, $r_{C_3C_3}$ (valori su sfondo blu); tali correlazioni possono essere interpretate come coefficienti di affidabilità[15].

B. *Diagonali di validità: same-trait different-method*, composto dalle correlazioni tra la stessa caratteristica misurata con metodi diversi; all'interno della matrice esse corrispondono ai valori di $r_{A_2A_1}$, $r_{B_2B_1}$, $r_{C_2C_1}$, $r_{A_3A_1}$, $r_{A_3A_2}$, $r_{B_3B_1}$, $r_{B_3B_2}$, $r_{C_3C_1}$, $r_{C_3C_2}$ (valori in diagonale su sfondo verde) e che possono essere interpretati come coefficienti di validità convergente in quanto rappresentano le correlazioni tra le stesse caratteristiche misurate attraverso metodi tra loro molto diversi, coerentemente con la concettualizzazione di validità fatta da Campbell e Frisse: se si misura una particolare caratteristica, correttamente concettualizzata, in modi diversi e si

[14] Il tipo di correlazione scelta può influenzare i risultati; nel caso di variabili categoriche si suggerisce di utilizzare, per evitare grandi *bias*, le correlazioni policoriche e poliseriali.

[15] Tali valori possono essere prodotti da metodi quali il *test-retest* o lo *split-half* (v. *Parte II*).

ottengono gli stessi risultati, le procedure di misurazione risultano probabilmente valide. Tali valori dovrebbero essere statisticamente rilevanti e di grande dimensione. Per ciascun metodo il valore della validità dovrebbe essere maggiore di qualsiasi altro valore nella riga e nella colonna corrispondenti; in altre parole due misure dello stesso costrutto dovrebbero avere correlazioni più strette tra loro che con le misure di altri costrutti.

C. *Triangoli Different-trait same-method*, composto dalle correlazioni tra diverse caratteristiche misurate con lo stesso metodo; all'interno della matrice esse corrispondono ai valori di $r_{B_1A_1}$, $r_{C_1A_1}$, $r_{C_1B_1}$, $r_{B_2A_2}$, $r_{C_2A_2}$, $r_{C_2B_2}$, $r_{B_3A_3}$, $r_{C_3A_3}$, $r_{C_3B_3}$ (triangolo con sfondo rosso). Campbell e Friske chiamano tali triangoli *heterotrait-monomethod*. L'ordine di grandezza di tali correlazione (all'interno dei triangoli) dovrebbe essere lo stesso.

D. *Triangoli Different-trait different-method*, composto dalle correlazioni tra caratteristiche diverse misurate con metodi differenti; all'interno della matrice esse corrispondono ai valori $r_{A_2B_1}$, $r_{A_2C_1}$, $r_{B_2A_1}$, $r_{B_2C_1}$, $r_{C_2A_1}$, $r_{C_2B_1}$, $r_{A_3B_1}$, $r_{A_3C_1}$, $r_{A_3B_2}$, $r_{A_3C_2}$, $r_{B_3A_1}$, $r_{B_3C_1}$, $r_{B_3A_2}$, $r_{B_3C_2}$, $r_{C_3A_1}$, $r_{C_3B_1}$, $r_{C_3A_2}$, $r_{C_3B_2}$ (sfondo fucsia).

Ricapitolando, le categorie possono essere visualizzate nella matrice nel modo seguente:

a. *same-trait same-method* (sfondo blu);
b. *same-trait different-method* (sfondo verde);
c. *different-trait same-method* (sfondo rosso);
d. *different-trait different-method*: (sfondo fucsia).

L'interpretazione della matrice: i criteri

Per poter interpretare i valori all'interno della matrice *MTMM*, Campbell e Friske hanno individuato quattro criteri (Campbell, 1959, 2001; Sullivan, 1981):

A. *I coefficienti di affidabilità dovrebbero rappresentare i valori più alti della matrice*.

B. *I coefficienti di validità dovrebbero essere significativamente diversi da 0 e sufficientemente grandi da incoraggiare ulteriori analisi di validità*.
Nella matrice dell'esempio i valori di tali coefficienti vanno da .39 a .68; applicando un semplice test di significatività a tali correlazione dovrebbe essere possibile accertarsi del livello di significatività loro associato. Una volta osservato un livello soddisfacente di significatività sarà possibile esplorare gli altri criteri.

C. *Ciascun coefficiente di validità dovrebbe essere maggiore di tutte le correlazioni* different-trait different-method *presenti nella stessa riga o nella stessa colonna del coefficiente di validità*.
Nella matrice dell'esempio il valore del coefficiente di validità per *A1-A2* è .51. Il confronto rilevante è quindi tra .51 e i coefficienti che sono nella stessa riga o colonna del valore .51 nei due triangoli adiacenti. Tali quattro coefficienti sono .32, .29, .31 e .30, tutti più piccoli del coefficienti di validità rilevante.

D. *Ciascun coefficiente di validità dovrebbe essere maggiore del corrispondente coefficiente* different-trait same-method.
La motivazione di tale criterio è data dal fatto che, perché delle misure siano valide vi deve essere più varianza di tratto che varianza di metodo: se le caratteristiche sono tra loro veramente distinte a livello concettuale allora la maggior parte della varianza condivisa dovrebbe riflettere la varianza metodologica[16]. La stessa caratteristica misurata con metodi diversi dovrebbe riflettere principalmente la varianza di tratto che dovrebbe essere maggiore della varianza di metodi delle correlazioni *different trait-same method*.

[16] Strettamente parlando, le correlazioni *different trait-same method* dovrebbero comprendere più della varianza metodologica. Anche se i tratti possono essere concettualmente distinti, vi può essere tra loro probabilmente qualche relazione causale. In genere, si dovrebbe utilizzare l'approccio *M-M* con variabili che siano entrambe concettualmente distinte e che hanno tra loro solamente connessioni causali minori e non dirette. Se ciò risultasse vero, la varianza comune tra lo stesso concetto misurato attraverso diversi metodi dovrebbe essere maggiore della varianza comune tra concetti diversi misurati attraverso lo stesso metodo.

Nella matrice dell'esempio il valore del coefficiente di validità per la caratteristica *A* ovvero *A1-A2* è uguale a .51; tale valore dovrebbe essere maggiore delle correlazioni che coinvolgono *A1* o *A2* con altre caratteristiche ma con lo stesso metodo. Quindi .51 viene confrontato con le correlazioni tra *A1* e *B1*, *A1* e *C1*, *A2* e *B2* e *A2* e *C2*, rispettivamente .42, .38, .44 e .38: il criterio può dirsi soddisfatto.

E. *Si dovrebbe osservare lo stesso modello di correlazioni all'interno di ciascun triangolo*,
   - sia quelli con gli elementi che presentano correlazioni tra tratti diversi che utilizzano metodi diversi,
   - sia quelli che riflettono correlazioni tra tratti diversi utilizzando lo stesso metodo.

   Nella matrice dell'esempio nel triangolo in alto (che chiameremo triangolo 1-1) il modello, in ordine inverso di grandezza di correlazione, è

   A1-B1 (.42), *A1-C1* (.38) e *B1-C1* (.33).

   Tale modello di correlazioni dovrebbe riflettere le varie connessioni causali tra i tratti *A*, *B* e *C*. Tutte le correlazioni dovrebbero riflettere una parte della varianza comune di metodi, in tale triangolo, ma il modello di ineguaglianza dovrebbe riflettere anche la forza delle connessioni causali tra i tratti. Secondo il modello del triangolo 1-1 la correlazione tra *A* e *B* dovrebbero essere maggiore che non tra *A* e *C* o tra *B* e *C*.

Vediamo un esempio[17]:

METODO	⇒	1			2			3		
⇓	TRATTI ⇒ ⇓	A	B	C	A	B	C	A	B	C
1	A	(082)								
	B	0.42	(0.79)							
	C	0.38	0.33	(0.74)						
2	A	0.51	0.32	0.29	(0.69)					
	B	0.31	0.45	0.19	0.44	(0.84)				
	C	.030	0.25	0.39	0.38	0.32	(0.65)			
3	A	0.58	0.31	0.30	0.62	0.36	0.28	(0.89)		
	B	0.35	0.48	0.21	0.25	0.68	0.25	0.46	(0.75)	
	C	0.28	0.19	0.39	0.24	0.23	0.59	0.37	0.36	(0.68)

A. La correlazione più bassa nella diagonale di validità è statisticamente significativa a livello 0.05.

B. I valori di tali coefficienti vanno da 0.39 a 0.68; applicando un semplice test di significatività a tali correlazione dovrebbe essere possibile accertarsi del livello di significatività loro associato. Una volta osservato un livello soddisfacente di significatività sarà possibile esplorare gli altri criteri.

C. Il valore del coefficiente di validità $r_{A_1A_2}$ è 0.51. Il confronto rilevante è quindi tra 0.51 e i coefficienti che sono nella stessa riga o colonna del valore 0.51 nei due triangoli adiacenti. Tali quattro coefficienti sono 0.32, 0.29, 0.31 e 0.30, tutti più piccoli del coefficienti di validità rilevante.

D. Il valore del coefficiente di validità per la caratteristica *A* ($r_{A_1A_2}$) è uguale a 0.51; tale valore dovrebbe essere maggiore delle correlazioni che coinvolgono per la caratteristica *A* i metodi 1 e 2 con altre caratteristiche ma con lo stesso metodo. Quindi 0.51 viene confrontato con le correlazioni $r_{A_1B_1}$, $r_{A_1C_1}$, $r_{A_2B_2}$ e $r_{A_2C_2}$, rispettivamente 0.42, 0.38, 0.44 e 0.38: il criterio può dirsi soddisfatto.

E. Nella matrice dell'esempio nel triangolo in alto (che chiameremo triangolo 1-1) il modello, in

[17] I dati nella tabella non corrispondono ai dati ottenuti a partire da uno studio reale e sono presentati sono a titolo illustrativo.

ordine inverso di grandezza di correlazione, è

$r_{A_1B_1}$ (.42), $r_{A_1C_1}$ (.38) e $r_{B_1C_1}$ (.33).

Tale modello di correlazioni dovrebbe riflettere le varie connessioni causali tra i tratti *A*, *B* e *C*. Tutte le correlazioni dovrebbero riflettere una parte della varianza comune di metodi, in tale triangolo, ma il modello di ineguaglianza dovrebbe riflettere anche la forza delle connessioni causali tra i tratti. Secondo il modello del triangolo 1-1 la correlazione tra *A* e *B* dovrebbero essere maggiore che non tra *A* e *C* o tra *B* e *C*.

Si può dire che i dati presentati nell'esempio soddisfano tutti i criteri.

In particolare, i confronti da effettuare per verificare i <u>criteri **B** e **C**</u> sono presentati nella seguente tabella; essendo il valore del coefficiente di validità sempre maggiore di tutti i confronti rilevanti *different-trait*, *different-method* e *different-trait*, *same-method*, i criteri **B** e **C** sono stati soddisfatti in tutti i casi. All'interno di tutti i triangoli presenti nella tabella iniziale il modello è esattamente lo stesso: *A* e *B* evidenziano le correlazioni maggiori, successivamente *A* e *C* mentre *B* e *C* presentano le correlazioni più basse. Essendo ciò vero indipendentemente dalla combinazione dei metodi coinvolti nel confronto, è stato verificato anche il <u>criterio **D**</u>.

		Coefficienti di validità	Confronti per il criterio B				Confronti per il criterio C			
A. Criteri B e C	$r_{A_2A_1}$	0.51	0.32	0.29	0.31	0.30	0.42	0.38	0.44	0.38
	$r_{B_2B_1}$	0.45	0.32,	0.19	0.31	0.25	0.42	0.33	0.44	0.32
	$r_{C_2C_1}$	0.39	0.29	0.19	0.30	0.25	0.38	0.33	0.38	0.32
	$r_{A_3A_2}$	0.62	0.36	0.28	0.25	0.24	0.44	0.38	0.46	0.37
	$r_{B_3B_2}$	0.68	0.36	0.25	0.25	0.23	0.44	0.32	0.46	0.36
	$r_{C_3C_2}$	0.59	0.28	0.25	0.24	0.23	0.38	0.32	0.37	0.36
	$r_{A_3A_1}$	0.58	0.31	0.30	0.35	0.28	0.42	0.38	0.46	0.37
	$r_{B_3B_1}$	0.48	0.31	0.21	0.35	0.19	0.42	0.33	0.46	0.36
	$r_{C_3C_1}$	0.39	0.30	0.21	0.28	0.19	0.38	0.33	0.34	0.36

	Triangolo		Ordine
B. Criterio D	1-1		$r_{B_1A_1} \Rightarrow r_{C_1A_1} \Rightarrow r_{C_1B_1}$
	2-2		$r_{B_2A_2} \Rightarrow r_{C_2A_2} \Rightarrow r_{C_2B_2}$
	3-3		$r_{B_3A_3} \Rightarrow r_{C_3A_3} \Rightarrow r_{C_3B_3}$
	1-2	top	$r_{A_2B_1} \Rightarrow r_{A_2C_1} \Rightarrow r_{B_2C_1}$
	1-2	bottom	$r_{B_2A_1} \Rightarrow r_{C_2A_1} \Rightarrow r_{C_2B_1}$
	1-3	top	$r_{A_3B_1} \Rightarrow r_{A_3C_1} \Rightarrow r_{B_3C_1}$
	1-3	bottom	$r_{B_3A_1} \Rightarrow r_{C_3A_1} \Rightarrow r_{C_3B_1}$
	2-3	top	$r_{A_3B_2} \Rightarrow r_{A_3C_2} \Rightarrow r_{B_3C_2}$
	2-3	bottom	$r_{B_3A_2} \Rightarrow r_{C_3A_2} \Rightarrow r_{C_3B_2}$

A conclusione di tale analisi, e sulla base di tali risultati, è possibile affermare che le procedure di misurazione utilizzate sono molto probabilmente valide e che misurano probabilmente ciò che si è

cercato di misurare.
Naturalmente in molte situazioni empiriche non tutte le verifiche effettuate e presentate nell'ultima tabella sono soddisfatte dai dati, pur disponendo di misure valide. Ciò può dipendere da molti fattori quali l'osservazione di livelli diversi di affidabilità e validità a causa di fluttuazioni casuali nel campionamento degli indicatori e dei soggetti.
Campbell e Fiske non hanno però definito dei parametri di riferimento su cui basarsi per stabilire se i dati si avvicinano ai criteri proposti.
Secondo alcuni tale modello può essere applicato anche al caso in cui i metodi non sono molto differenti tra loro come nel seguente esempio in cui si studia la validità di una nuova misura del livello di soddisfazione del lavoro confrontandola con uno standard esistente. Il *Job Satisfaction Survey* è uno strumento suddiviso in nove sottoscale ideato per valutare individualmente il livello di soddisfazione e gli atteggiamenti verso vari aspetti del lavoro.
E' stato sviluppato specificamente per essere impiegato su lavoratori dipendenti nelle organizzazioni di servizi (ospedali, centri di igiene mentale, servizi sociali, ecc.). Nell'ambito della misurazione della soddisfazione del lavoro esiste una scala ben validata e largamente utilizzata chiamata *Job Descriptive Index* (*JDI*); cinque delle sottoscale di *JSS* (metodo 2) sono comprese anche nella *JDI* (metodo 1). Dopo aver somministrato entrambe le scale ad un centinaio di soggetti, e disponendo di dati rilevati con due scale considerate metodi distinti, è stato possibile applicare il modello di analisi *MTMM*.
Nella seguente matrice *MTMM* sono presentati i risultati ottenuti per tre sottoscale: soddisfazione rispetto a natura dei compiti (*A*), retribuzione (*B*) e supervisione (*C*). La matrice contiene le correlazioni tra i 6 punteggi.

METODI	⇒	JDI			JSS		
⇓	TRATTI ⇒ ⇓	A	B	C	A	B	C
JDI	A	( )					
	B	.27	( )				
	C	.31	.23	( )			
JSS	A	.66	.24	.24	( )		
	B	.33	.62	.34	.29	( )	
	C	.25	.27	.80	.22	.34	( )

I valori che risultano nell'esempio rivelano buone validità convergenti e discriminanti per la *JSS*. I valori di validità in diagonale sono tutti piuttosto alti (da .62 a .80) e sono i maggiori valori dell'intera matrice.
I valori delle corrispondenti correlazioni all'interno di ciascun triangolo sono abbastanza simili e sono piuttosto modesti e vanno da .22 a .34. Tali valori piuttosto bassi suggeriscono che le sottoscale valutano costrutti diversi.

1.3.3 Esempio

Anche se secondo molti ricercatori l'*achievement* (rendimento, realizzazione) rappresenta un concetto unidimensionale, alcuni hanno sottoposto tale assunto a verifica attraverso la matrice *MTMM* (Campbell, 1959, 2001; Sullivan, 1981).
Tale applicazione non è molto ortodossa in quanto non sono stati selezionati tratti che a priori si ipotizzavano non correlati (o solamente molto correlati) ma si è cercato di specificare molti tipi diversi di un unico tratto (*achievement motivation*), che possono o non possono essere molto correlati. Sono stati definiti sei tipi diversi di *achievement* e sono state individuate cinque diverse strategie di misurazione. A scopo illustrativo qui ne sono presentate solo tre. I tre tipi di *achievement* selezionati:

1. *status with expert*, impegno e accettazione di perizie e giudizi di esperti;
2. *tendenza ad accumulare*, che comprende ricompense materiali come motivazione primaria all'*achievement*;
3. *interesse per la perfezione*, che comprende alti standard di *achievement* intellettuale e culturale.

I tre metodi selezionati sono:

A. *self-rating*, in cui ciascun soggetto doveva valutare se stesso rispetto ad una serie di aggettivi polarizzati (tre per ogni tratto);
B. *simulation role selection*: dopo avere descritto dei ruoli, uno per ciascun tratto, veniva chiesto ai soggetti, di selezionarne uno per la simulazione; il titolo e la descrizione dei ruoli sono stati fatti in modo accurato in modo che ciascuno riflettesse solamente uno dei tratti;
C. *personality inventory* che comprendeva 38 item per ciascun tratto.

Nella seguente tabella sono presentate le correlazioni tra tratti e metodi:

METODI ⇒		1			2			3		
⇓	TRATTI ⇒ ⇓	A	B	C	A	B	C	A	B	C
1	A	(1.0)								
	B	-.16	(1.0)							
	C	.42	.07	(1.0)						
2	A	.30	-.09	.14	(1.0)					
	B	-.05	.34	.28	-.06	(1.0)				
	C	.15	-.19	.23	.19	-.04	(1.0)			
3	A	.52	.08	.33	.22	.12	.09	(1.0)		
	B	-.06	.69	.13	-.10	.38	.20	.17	(1.0)	
	C	.40	.02	.67	.11	.27	.18	.35	.22	(1.0)

I metodi sono piuttosto diversi, ma non è possibile stabilire se i tratti siano abbastanza differenti da validare le procedure di misurazione. Infatti non è stata utilizzata la matrice *MTMM* per validare le misure ma piuttosto per verificare l'unidimensionalità della motivazione all'*achievement*. Lo studio si conclude con una risposta negativa. Vediamo la tabella riguardante la verifica dei criteri 2, 3 e 4.

		Coefficienti di validità	Confronti per il criterio B				Confronti per il criterio C			
A. Criteri B e C	$r_{A_2A_1}$	.30	-.09	.14	-.05	.15	-.16	.42*	-.06	.19
	$r_{B_2B_1}$	.34	-.09	-.19	-.05	.28	-.16	.07	-.06	-.04
	$r_{C_2C_1}$	.23	.15	-.19	.14	.28*	.42*	.07	.19	-.04
	$r_{A_3A_2}$	.22	.12	.09	-.10	.11	-.06	.09	.17	.35*
	$r_{B_3B_2}$	.38	.12	.27	-.10	-.20*	-.06	-.04	.17	.22
	$r_{C_3C_2}$	.18	.11	.27*	.09	-.20*	.19*	-.04	.35*	.22*
	$r_{A_3A_1}$	.52	.08	.33	-.06	.40	-.16	.42	.17	.35
	$r_{B_3B_1}$	.69	.08	.02	-.06	.13	-.16	.07	.17	.22
	$r_{C_3C_1}$	.67	.40	.02	.33	.13	.42	.07	.35	.22

	Triangolo		Ordine
B. Criterio D	1-1		$r_{C_1A_1} \Rightarrow r_{C_1B_1} \Rightarrow r_{B_1A_1}$
	2-2		$r_{C_2A_2} \Rightarrow r_{C_2B_2} \Rightarrow r_{B_2A_2}$
	3-3		$r_{C_3A_3} \Rightarrow r_{C_3B_3} \Rightarrow r_{B_3A_3}$
	1-2	top	$r_{B_2C_1} \Rightarrow r_{A_2C_1} \Rightarrow r_{A_2B_1}$ **
	1-2	bottom	$r_{C_2A_1} \Rightarrow r_{B_2A_1} \Rightarrow r_{C_2B_1}$ **
	1-3	top	$r_{A_3C_1} \Rightarrow r_{B_3C_1} \Rightarrow r_{A_3B_1}$
	1-3	bottom	$r_{C_3A_1} \Rightarrow r_{C_3B_1} \Rightarrow r_{B_3A_1}$
	2-3	top	$r_{A_3B_2} \Rightarrow r_{A_3C_2} \Rightarrow r_{B_3C_2}$ **
	2-3	bottom	$r_{C_3B_2} \Rightarrow r_{C_3A_2} \Rightarrow r_{B_3A_2}$ **

I valori registrati dai coefficienti di validità, che in tale tabella corrispondono alle correlazioni tra *same-trait different-method*, vanno da .18 a .69. Qui, con *N*=155, essi sono risultati tutti statisticamente significativi, anche se notiamo come il secondo metodo (selezione del ruolo di simulazione) risulta avere i coefficienti di validità più bassi degli altri due metodi. Ciò è dovuto al fatto che si tratta del metodo che più si differenzia dagli altri, in quanto mentre il *self-rating* e il *personality inventory* rappresentano delle auto-descrizioni, la selezione del ruolo è in un certo qual modo più di una misura di comportamento.
Il secondo criterio[18] è soddisfatto in modo soddisfacente in quanto ben 33 dei 36 confronti sono rientrati nel criterio con solamente tre correlazioni *different-trait different-method* più grandi dei loro corrispondenti coefficienti di validità; come si può notare le tre eccezioni coinvolgono il metodo 2.
Il terzo criterio[19] è soddisfatto solo in 31 dei 36 confronti; le cinque violazioni hanno coinvolto il metodo 2, particolarmente con i metodi 2 e 3 che misurano il tratto C.
Quindi due delle tre violazioni del secondo criterio e tre delle sei violazioni del terzo criterio coinvolgono il metodo 2 che può essere legittimamente sospettato di non essere valido in particolare quando è riferito al tratto C.
Relativamente al quarto criterio[20] notiamo come per cinque dei nove triangoli, i tratti A e C correlano molto tra loro, poi i tratti B e C, e i tratti A e B correlano molto debolmente.
Le inversioni suggeriscono l'esistenza di una qualche interazione tra metodi e tratti rivelando che alcuni metodi possano essere migliori misure di alcuni tratti.
Con tali risultati, i ricercatori possono concludere che, dati gli errori di misurazione e di campionamento, le misure sono generalmente valide e che possono essere utilizzate per ulteriori analisi.

Il metodo messo a punto da Campbell e Fiske per la valutazione della matrice funziona abbastanza bene in circostanze come quelle presentate, in cui la validità per tutti i metodi è abbastanza netta. In

[18] Ciascun coefficiente di validità dovrebbe essere maggiore di tutte le correlazioni *different-trait different-method* che sono nella stessa riga o colonna del coefficiente di validità, nei triangoli composti dai valori in corsivo.

[19] Ciascun coefficiente di validità dovrebbe essere maggiore delle correlazioni *different-trait same-method* che coinvolgono la stessa variabile come coefficiente di validità.

[20] Lo stesso modello di correlazioni dovrebbe essere evidenziato all'interno di ciascuno dei triangoli.

altri casi le deviazioni dal caso ideale lo rendono piuttosto difficile da applicare e interpretare. Sono disponibili pochi metodi statistici per la valutazione delle matrici *MTMM*. Quello diventato recentemente più popolare è quello che utilizza il modello delle equazioni strutturali. Tale tecnica può essere abbastanza utile ma presenta alcuni seri limiti. Il recente sviluppo di modelli *direct product* si presentano promettenti, sebbene si osservino poche applicazioni. Dovendo sviluppare misure, comunque, l'analisi della matrice di per sé può dare delle interessanti tracce sull'esistenza o meno di problemi di validità potenziale in alcune scale.
Per esplorare l'utilità di tale approccio in maggiore dettaglio, considerando il modello sottostante la matrice *MTMM* è possibile utilizzare la tecnica degli indicatori multipli per

- rappresentare le diverse misure e le variabili astratte;
- esaminare la struttura delle relazioni così prodotte

ovvero è possibile applicare i principi della *path analysis*. Ciò consente di esaminare simultaneamente gli assunti sottostanti e di illustrare l'estensione dei modelli con indicatori multipli a situazioni più complesse.
Mentre Campbell e Fiske presentavano l'utilizzo degli indicatori multipli per valutare la validità, Costner nel 1969 presentava un approccio alla valutazione dell'affidabilità che utilizza gli indicatori multipli, ponendo l'attenzione sulla questione dell'errore di misurazione.

1.3.4 Vantaggi e svantaggi dell'approccio

Indubbiamente tale approccio fornisce una metodologia operativa per la valutazione della validità. In un'unica matrice è praticamente possibile esaminare simultaneamente sia la validità convergente che la validità discriminante. Nel cercare di includere metodi e tratti su un unico piano, i due ideatori hanno accentuato l'importanza della ricerca degli effetti del metodo di misurazione in aggiunta al contenuto della misurazione. In questo senso il *MTMM* fornisce una struttura rigorosa per valutare la validità.
Per molte ragioni, però, nonostante tali vantaggi, il *multitrait-multimethod* non ha trovato grandi applicazioni in quanto, richiede innanzitutto un disegno sperimentale di ricerca piuttosto complesso ed esteso (misurazione di più tratti attraverso diverse metodologie) non sempre pienamente realizzabile. Le decisioni finali, riguardanti la validità, sono sempre legate ad un giudizio individuale del ricercatore: ciò non sempre è apprezzato da chi ritiene che tale decisione sia presa "oggettivamente" per mezzo di un unico coefficiente statistico (Sullivan, 1981).

2. IL MODELLO PER LA DEFINIZIONE E LA COSTRUZIONE DEL DATO

2.1 LA NATURA DEI DATI: LA TEORIA DEI DATI DI COOMBS

Per procedere alla trasformazione dell'osservazione empirica in una informazione che può essere compresa, interpretata e successivamente analizzata (dato) è necessario fare riferimento ad una teoria che consenta di chiarire la natura dell'informazione (al fine di comprendere gli oggetti e di esaminarne le differenze) e conseguentemente le procedure analitiche più appropriate.
La teoria di riferimento per la definizione del dato soggettivo è quella di Coombs (1950, 1953, 1964; Flament, 1976; McIver, 1979), basata su un approccio geometrico.
Tutte le osservazioni empiriche possono essere rappresentate come confronti, più o meno espliciti[1] tra, almeno, due entità che possono essere definite come punti all'interno di uno spazio. Le posizioni relative dei due punti dipendono dal modo in cui gli analisti hanno scelto di interpretare il confronto tra due entità[2]; per ciascuna osservazione viene registrata quella porzione di osservazione che riassume il confronto tra le entità; questa può essere definita come una relazione geometrica di confronto tra i componenti di una coppia di punti.
La rappresentazione geometrica dei dati soddisfa l'obiettivo di ottenere un modello, infatti la nozione di coppie di punti è sufficientemente generale da comprendere qualsiasi tipo di osservazione. Inoltre molti tipi di confronti possono essere rappresentati da un numero sorprendentemente piccolo di relazioni geometriche tra punti.
Coombs (1953, 1964; Flament, 1976; McIver, 1979), ha sviluppato la teoria basata interamente sull'interpretazione geometrica dei dati. Egli afferma che due entità in un singolo dato possono variare secondo due criteri.

a. Due elementi in una coppia possono essere estratti da **insiemi**:
   - *diversi* (un consumatore e un prodotto, uno studente e una prova, ecc.);
   - *unici* (un consumatore *A* e un consumatore *B*, prodotto *A* e prodotto *B*, ecc.).

   Tale distinzione in genere può essere fatta in modo molto semplice considerando la natura degli oggetti; esistono però dei casi in cui due oggetti, pur appartenenti apparentemente allo stesso insieme, vengono considerati come appartenenti a due diversi insiemi (negli studi sociometrici è possibile distinguere tra soggetti che scelgono e soggetti che vengono scelti).
b. Il confronto tra entità di una coppia comporta una **relazione** di:
   - *dominanza*, quando un oggetto possiede un livello maggiore o minore di una determinata caratteristica (uno studente risponde correttamente a una domanda, ecc.);
   - *prossimità*, quando due oggetti corrispondono o coincidono tra loro a livelli diversi (il componente di un gruppo sceglie un altro per lavorare insieme, ecc.).

   La differenza tra i due tipi di relazione è di solito facilmente individuabile a partire dalla natura dell'osservazione empirica (uno studente risponde correttamente a più domande di un altro → dominanza; due studenti completano entrambi le stesse domande → prossimità); ma, in molti casi, la distinzione rimane, in ultima analisi, all'interpretazione delle osservazioni da parte

[1] Sappiamo che i confronti sono fondamentali per distinguere determinati oggetti dal contesto. «La mela è rossa» confronta «mela» con una serie di colori. Ciò vale anche a livello scientifico quando le osservazioni richiedono sempre dei confronti tra entità.

[2] Per l'osservazione «la mela è rossa» i punti «mela» e «rosso» saranno relativamente vicini all'interno dello spazio definito.

dell'analista[3] e non alle stesse osservazioni a conferma che i dati richiedono sempre un apporto creativo da parte del ricercatore.

I due criteri possono essere facilmente trasformati in rappresentazioni geometriche: le entità contenute in una singola osservazione sono sempre descritte e delineate come coppia di punti all'interno di uno spazio[4]. Se due elementi di una coppia appartengono

- a due diversi insiemi, lo spazio è detto *congiunto*,
- a uno stesso insieme, lo spazio è detto *oggetto*, o prende il nome dell'oggetto trattato (*soggetto*, *stimolo*, ecc.).

Se gli oggetti appartenenti alla coppia sono connessi da una relazione di dominanza ciò si riflette nell'ordine dei punti nello spazio: se uno domina l'altro il suo punto è collocato in una posizione più estrema lungo la dimensione.

La relazione di prossimità tra due oggetti è definita in termini di distanza tra punti: se due oggetti sono molto prossimi, la distanza tra due punti diviene più piccola o viceversa.

La combinazione tra i due criteri può produrre quattro diversi tipi di dati entro i quali rientrano tutte le osservazioni empiriche indipendentemente dalla loro natura sostanziale (Flament, 1976):

		Coppie di punti in osservazione	
		stesso insieme	diverso insieme
Relazione tra coppie di punti	dominanza	Stimulus comparison **a**	Single stimulus **b**
	prossimità	Similarities **c**	Preferential choice **d**

a. *Stimulus comparison* (confronto tra stimoli): le osservazioni sono rappresentate da coppie di elementi estratti dallo stesso insieme con una relazione di dominanza tra loro; tale combinazione si verifica quando oggetti simili sono confrontati tra loro sulla base di una proprietà comune, rappresentabile con una retta e le osservazioni possono essere adattate in termini di *ordinamento* di punti lungo una retta. Vediamo alcuni esempi:
   1. un'automobile ha un rapporto migliore tra chilometri percorsi e consumo di benzina rispetto ad un'altra: i punti rappresentano le automobili e la retta rappresenta i valori del rapporto;
   2. un esercizio richiede più tempo per essere eseguito di un altro: i punti sono gli esercizi e la retta rappresenta il tempo;
   3. un prodotto è più attraente di un altro: i punti rappresentano i prodotti e la retta rappresenta l'attrattiva.

b. *Single stimulus* (stimolo unico): le osservazioni sono rappresentate da coppie di oggetti estratti da insiemi diversi con tra loro una relazione di dominanza. Indipendentemente dal significato delle osservazioni il modello geometrico per questo tipo di dati comporta una relazione d'ordine tra ciascuna coppia di punti lungo la dimensione sottostante. Se un oggetto A ha un punteggio y sulla variabile x allora il punto dell'oggetto A domina y unità di un continuum corrispondente alla variabile x.[5] Esempi di questo tipo di dati sono

[3] Vediamo un seguente esempio:
   a. il livello di capacità di un soggetto è superiore a quello necessario per eseguire correttamente un compito ($\rightarrow$ dominanza),
   b. il livello di capacità di un soggetto coincide con quello necessario per eseguire correttamente un compito ($\rightarrow$ prossimità).

[4] Anche se tale spazio può essere multidimensionale, per semplicità nella presentazione si farà riferimento ad uno spazio unidimensionale.

[5] Superficialmente il precedente tipo di dati può sembrare molto simile a questo, infatti sia per entrambi l'informazione contenuta in una osservazione empirica comporta un ordinamento di una coppia di punti lungo una retta; tra i due

- intervistati e categorie ordinate su una scala di valutazione,
- studenti ed esercizi,
- lunghezza di un oggetto e graduazioni su una scala di misurazione.

Praticamente tutte le misure di tipo fisico ricadono in questa categoria di dati. In tali casi i due insiemi di punti sono:

- gli oggetti misurati,
- le unità che definiscono lo strumento di misurazione.

c. *Similarities* (somiglianze): le osservazioni sono rappresentate da coppie di oggetti estratti dallo stesso insieme con una relazione di prossimità tra loro; ciò comporta il concetto di somiglianza:
   - due stimoli sono giudicati più o meno simili (la prossimità tra loro aumenta o diminuisce),
   - due soggetti presentano o meno lo stesso comportamento,
   - ecc.

   Il confronto empirico tra due oggetti è adattato come *distanza* tra una coppia di punti. Non si valuta l'ordinamento dei punti lungo il continuum.

d. *Preferential choice* (scelta di preferenza): le osservazioni sono rappresentate da coppie di oggetti estratte da insiemi diversi con una relazione di prossimità tra loro; l'esempio più ovvio è quello dei dati di preferenza: più un certo soggetto preferisce un particolare stimolo maggiore è la prossimità esistente tra soggetto e stimolo. Geometricamente le prossimità sono rappresentate come distanze tra punti all'interno di uno spazio congiunto. L'aumento della prossimità tra un soggetto e uno stimolo corrisponde alla diminuzione della distanza tra il punto del soggetto e il punto dello stimolo. L'informazione contenuta in ogni singolo dato di questo tipo non comporta alcuna informazione riguardo all'ordinamento relativo del soggetto e dello stimolo all'interno dello spazio.

Di seguito vediamo schematicamente alcuni esempi di come osservazioni empiriche possano essere trasformate nei quattro tipi di dati. Per ciascun dato lo schema identifica

- la coppia di entità contenuta in quel dato,
- la relazione tra gli elementi della coppia,
- una possibile rappresentazione geometrica dei loro punti lungo la retta di riferimento.

Notare che

- per i dati di somiglianza e di scelte di preferenza sono presentati più modelli geometrici,
- in alcuni casi una singola osservazione può essere interpretata con più dati.

esistono però delle differenze fondamentali che riguardano l'informazione utilizzata per costruire la rappresentazione geometrica.

DATI	OSSERVAZIONE EMPIRICA	COPPIE DI PUNTI		RELAZIONE TRA PUNTI		POSSIBILE MODELLO GEOMETRICO
		1° punto	2° punto	azione	implica	
Confronto di stimoli	La squadra A ha vinto sulla squadra B ed ha perso con la C	Squadra A Squadra A	Squadra B Squadra C	"vincere" "perdere"	$>$ $<$	__B____A_C_
	Il cibo X è più salato del cibo Y	Cibo X	Cibo Y	"più salato"	$>$	____Y__X__
Stimolo-unico	Lo studente A risponde correttamente alla domanda 1	Studente A	Domanda 1	"risposte corrette"	$>$	_1__A___
	Il libro X pesa due etti	Libro X	Peso in etti	"pesi"	$> C$ e $< C$	_1_2&X___3
Somiglianza	I voti ottenuti dai deputati X e Y sono più simili di quelli dei senatori W e Z	Senatore X Senatore W	Senatore Y Senatore Z	"voti simili"	Distanze minori	_X_Y_W__Z _W__X_Y__Z _Z__Y_X__W
	La torta è più simile alla focaccia che al pane	Torta T Torta T	Focaccia F Pane P	"più simile"	Distanze minori	_F__T____P__ _P____F__T__
Scelta di preferenza	Al bambino A piacciono più i gelati delle carote	Bambino A Bambino A	Gelati G Carote C	"piacere"	Distanze minori	_G__A__C__ _A__G_C___ __C____A_G_
	La mela è rossa ma non verde o gialla	Mela M Mela M Mela M	Rosso R Verde V Giallo G	"essere" "non essere"	Distanze minori Distanze maggiori	_G_V___R_M __V_G__M_R_ _G_R_M___V__

Ricordiamo che i nomi dati a ciascuno dei tipi di dati sono stati scelti per convenzione; a ciascuna di tali tipologie in realtà può fare riferimento una varietà di osservazioni sostanzialmente differenti[6]. Tali nomi indicano semplicemente diversi tipi di relazioni geometriche ricavate dalle osservazioni empiriche. Essi non fanno alcun riferimento al processo attraverso il quale vengono generati i dati.

2.2 L'ORGANIZZAZIONE DEI DATI: LE MATRICI

Nel processo di definizione del dati è importante chiarire anche il tipo di matrice (Delli Zotti, 1985) in cui possono essere organizzati. Carrol, Arabie e Young[7] (Jacoby, 1991) hanno definito una classificazione delle diverse forme che può assumere le matrici dei dati, ciascuna delle quali si caratterizza per il numero di:

- ***way***, termine che si riferisce al *numero di dimensioni della matrice* e quindi il numero di indici utilizzati per l'identificazione degli oggetti; ciascuna *way* presenta un proprio numero di *livelli*, corrispondenti al numero di entità nell'insieme di oggetti; quindi con *way* si definisce la forma totale della matrice dei dati e i livelli specificano la dimensione della matrice; qualsiasi insieme di dati si presenta al minimo nella forma *two-way* in quanto un'osservazione implica sempre un confronto tra due oggetti;
- ***mode***, termine che si riferisce al *numero di oggetti rappresentati* dalle *way* della matrice, ovvero indica il numero di classi di entità; i *mode* determinano l'interpretazione degli oggetti.

Tenendo che è possibile identificare anche tipologie complesse di matrici e che, in genere il numero dei *mode* non può superare il numero delle *way*, vediamo alcuni esempi. Si parla di matrice

- *one-mode* quando l'insieme degli elementi che compaiono in riga è lo stesso di quelli che compaiono in colonna, ovvero nella matrice è contenuto un unico insieme di informazioni;

[6] Per esempio
- i dati *single stimulus* richiedono due insiemi di oggetti e il ricercatore può scegliere rispetto a quale far riferimento per l'analisi (è possibile decidere di ordinare soggetti sulla base delle capacità e le domande sulla base della difficoltà);
- i dati *preferential choice* non necessariamente fanno riferimento a preferenze concrete.

[7] Tale teoria viene spesso ricordata con le iniziali degli autori (*CAY*).

- *two-way two-mode* quando presenta un insieme di dati contiene confronti accoppiati tra *K* stimoli: due stimoli all'interno di ciascuna coppia (*two-way*) e gli stimoli, unici oggetti coinvolti nei confronti (*one-mode*); ciascun *way* avrà *K* livelli;
- *two-way two-mode* quando presenta in riga *N* soggetti e in colonna *K* variabili (la prima *way* sarà di *N* livelli e la seconda di *K* livelli); rappresenta la tipologia più nota e comune;
- *three-way three-mode* quando la terza *way* si riferisce alle ripetizioni; con *N* soggetti che rispondono a *K* item per *M* volte si ottiene una matrice *three-mode*:
  - soggetti (*N* livelli),
  - domande (*K* livelli),
  - momenti (*M* livelli);
- *three-way two-mode* quando si hanno osservazioni ripetute in *M* momenti di confronti accoppiati tra *K* stimoli;
- *four-way, four-mode* quando si hanno:
  - *N* (livelli) soggetti che valutano,
  - *K* (livelli) stimoli secondo,
  - *Q* (livelli) diversi attributi in ciascuna delle
  - *M* (livelli) diverse occasioni.

L'adeguatezza del tipo di matrice ad una particolare situazione dipende interamente dall'interpretazione che l'analista fa delle osservazioni piuttosto che dalla natura delle entità coinvolte nelle osservazioni empiriche. Inoltre il numero e il tipo di entità empiricamente distinguibili contenute nelle osservazioni possono o meno corrispondere alla forma e alla dimensione della matrice.

Riprendendo i concetti visti a proposito del *modello gerarchico* si può dire che l'organizzazione dei dati, nei casi in cui si assume una corrispondenza diretta tra *variabile non osservabile* e *indicatore*, richiede una matrice *two-way two-mode*; tale matrice si presenta come quella mostrata di seguito in cui ciascun indicatore misura un attributo diverso mentre le unità sono rappresentate da casi (individui, città, ecc.):

		ATTRIBUTI / VARIABILI					
		1	2	…	*j*	…	*k*
UNITA'	1	x_{11}	x_{12}	…	x_{1j}	…	x_{1k}
	2	x_{21}	x_{22}	…	x_{2j}	…	x_{2k}
	…	…	…	…	…	…	…
	i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{ik}
	…	…	…	…	…	…	…
	n	x_{n1}	x_{n2}	…	x_{nj}	…	x_{nk}

Per esempio un ricercatore che studia i collegamenti tra città può ottenere le seguenti varietà:

- numero di viaggi (o viaggiatori) tra *K* città (*two-way*, con *K* livelli, e *one-mode*, le città);
- numero di viaggi (o viaggiatori) tra *K* città in partenza e in arrivo (*two-way*, con *K* livelli, e *two-mode*, le città);
- numero di viaggi (o viaggiatori) tra *K* città in partenza e in arrivo per *Q* mezzi di trasporto su *M* diversi anni (*four-way*, *four-mode*, con *K* città di partenza, *K* città di arrivo, per *Q* mezzi di trasporto per *M* anni);
- numero di viaggi (o viaggiatori) tra *K* città in partenza e in arrivo per *QM* mezzi di trasporto negli anni (*four-way*, *three-mode*, con *K* città di partenza, *K* città di arrivo, per *QM* ripetizioni).

Con questo esempio risulta evidente come l'insieme dei dati sia trattato come un modello astratto delle osservazioni e come le caratteristiche dei dati siano interamente indipendenti dalle proprietà delle stesse osservazioni.
Questa teoria può essere considerata come un completamento dell'approccio proposto da Coombs[8]. Per questo può essere utile pensare ai dati facendo riferimento ad entrambe le teorie in quanto ciascuna di esse chiarisce aspetti diversi dell'informazione ottenuta a partire dalle osservazioni

[8] I dati *single stimulus* e di *preferential choice* possono produrre matrici con almeno *two-way* e *two-mode*. Le differenze tra loro riguardano la relazione di confronto tra i *mode*: relazione di dominanza nel primo caso e relazione di prossimità nel secondo. I dati *stimulus comparison* e *similarity* producono entrambe matrici *two-way one-mode*. Naturalmente per ciascuna di tali tipologie è possibile ottenere delle osservazioni ripetute: in questo caso il numero di *way* e *mode* aumenta coerentemente.

empiriche.

2.3 LA COSTRUZIONE DEL CONTINUUM: LE TECNICHE DI SCALING

Nella costruzione del dato soggettivo si pone il problema di definire e, in un certo senso, creare e generare il continuum lungo il quale posizionare gli oggetti o i soggetti relativamente alla caratteristica da misurare; tale procedimento, che richiama quanto definito a livello di teoria dei dati, è detto *scaling*.
In fase di identificazione del continuum occorre tenere presente che fino a questo punto esso è stato definito solo teoricamente; lo *scaling* può essere identificare secondo diverse modalità (Marradi, 1980):

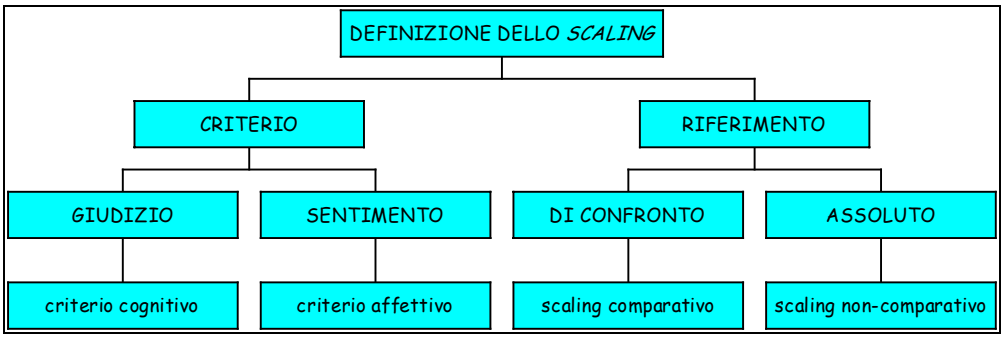
- classificazione: il continuum è definito da categorie che suddividono l'estensione del concetto;
- *scaling* discreto: il continuum è suddiviso in categorie discrete ovvero si individua un insieme di stimoli relativi ad un particolare attributo e alla loro collocazione lungo il continuum; uno dei problemi che sorgono nell'individuare punti discreti lungo il continuum è quello della possibilità di assumere distanze uguali tra i punti che definiscono le categorie (intervalli uguali)[9]; tale assunto consente di lo scaling discreto in termini metrici; spesso lo *scaling* discreto è definito come *underlying continuum*: nel definire le categorie discrete si assume che sotto tali categorie vi sia un continuum di risultati possibili; conseguentemente è possibile dire che tra gli elementi classificati nella categoria A vi sono alcuni più "vicini" alla categoria adiacente B rispetto ad altri si può affermare che tra queste due categorie vi è quindi un continuum che la classificazione ha riportato ad un numero determinato di categorie. Ciò ha delle conseguenze anche in fase di definizione dell'ampiezza delle categorie in quanto una individuazione delle categorie fatta in modo "grossolano" può occultare tale continuità.
- *scaling* continuo (metrico): le posizioni che individuano il continuum sono legate tra loro da proprietà metriche; in pratica lo *scaling* è quando tra due punti, per quanto vicini possano essere, esiste sempre la possibilità di individuarne un altro[10]; a tale proposito occorre osservare che una completa continuità richiederebbe una misurazione infinitamente precisa e la possibilità di definire tutti i valori lungo tale continuum. E' per questo che lo scaling "perfettamente" continuo è più una astrazione che una caratteristica osservabile nella realtà.[11]

Per quanto detto finora, si può affermare che la differenza tra le diverse modalità di identificazione del continuum sta nel livello di conoscenza sostanziale dell'effettiva ampiezza dei segmenti in cui è suddiviso il continuum o comunque nella capacità di stimare tale ampiezza da parte del ricercatore (Marradi, 1980).
Il procedimento di *scaling* richiede la definizione dei seguenti aspetti:

[9] La definizione della distanza tra i punti che definiscono i segmenti nei quali è stato suddiviso il continuum può essere stabilità attraverso il giudizio di esperti ma soprattutto della conoscenza sostanziale da parte del ricercatore della caratteristica studiata. Vedremo come alcuni modelli di *scaling* consentono di affrontare il problema dell'individuazione di tali punti.

[10] Il *magnitude scaling* (Lodge, 1981) rappresenta l'approccio più noto per la costruzione di continuum metrici e hanno alla base la teoria della psicofisica (Stevens, 1951, 1957).

[11] Ricordiamo che in aggiunta alla tecniche che verranno prese in considerazione, un approccio che può essere utile nell'individuazione di un continuum metrico è considerato quello proposta dall'analisi delle corrispondenze (Amaturo, 1989; Weller, 1990).



2.3.1 Criteri

Il *tipo* di riferimento si riferisce al tipo di valutazione (**criterio**) che viene sollecitata a livello soggettivo e legata al tipo di caratteristica studiata. E' possibile distinguere tra:

- *criterio cognitivo*: in questo caso si sollecita un giudizio, una conoscenza; l'obiettivo è quello di stabilire la relazione tra intensità *percepita* e intensità *reale* dell'attributo; questo vuol dire che in molti casi è anche possibile verificare il livello di correttezza e di accuratezza della reazione sollecitata; un tipico criterio cognitivo è quello che richiede la valutazione di *somiglianze*[12]*;*
- *criterio affettivo*: in questo caso si sollecita un sentimento, una sensazione, una preferenza, un interesse, una simpatia; l'obiettivo è quello di rilevare la relazione tra la caratteristica misurata e il caso; questo tipo di riferimento non consente di identificare e di definire un livello ed uno standard di correttezza e di accuratezza rendendo più complesso lo sviluppo di modelli.

2.3.2 Riferimenti

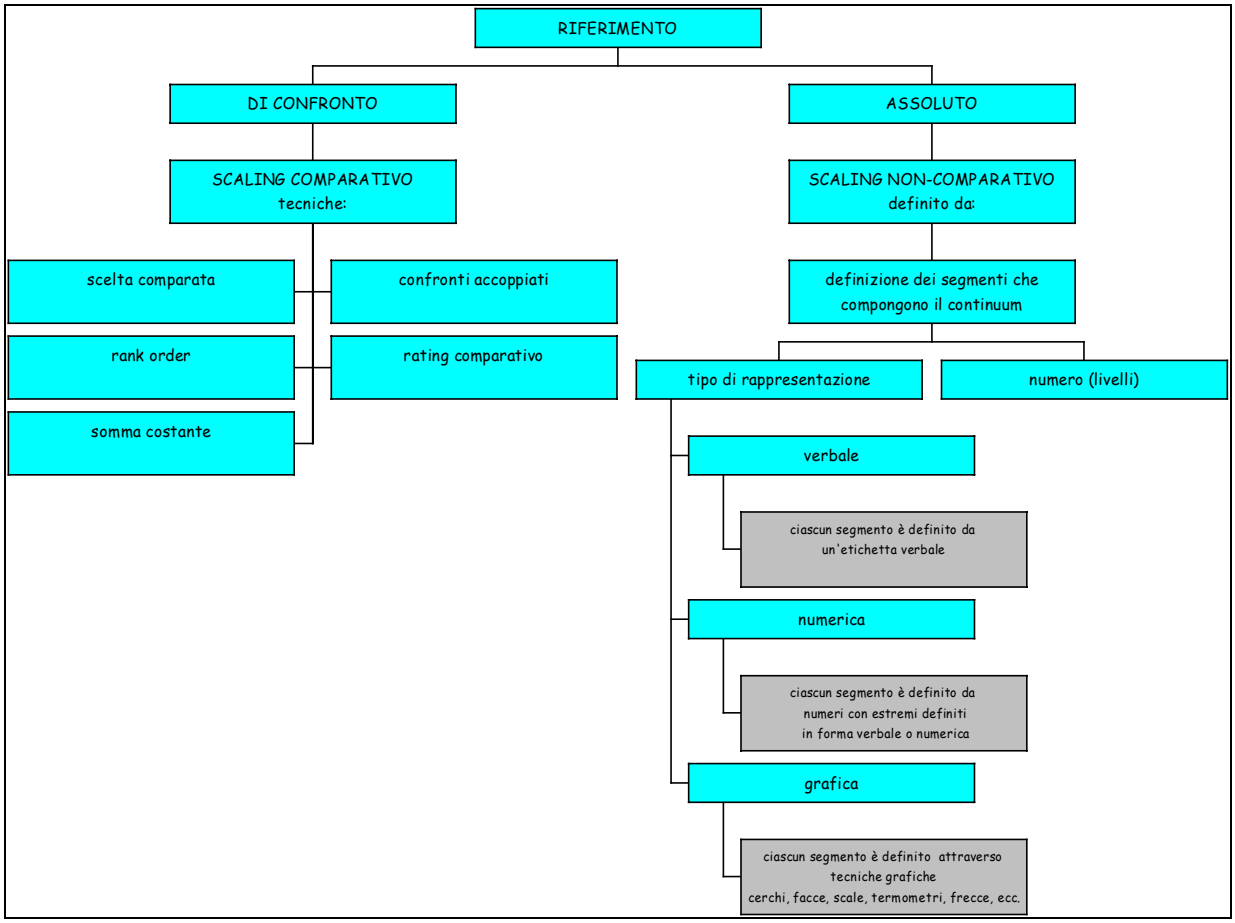
Il riferimento richiesto o proposto può essere **comparativo** o **assoluto**. In genere i riferimenti assoluti sono preferiti in quanto più velocemente applicabili e perché producono dati più facilmente interpretabili. Nella pratica però il riferimento comparativo consente di avere valutazioni più accurate.[13] I due riferimenti danno origine a due approcci diversi, rispettivamente lo *scaling* comparativo e lo *scaling* non-comparativo, applicabili entrambi con i due criteri precedentemente

[12] Il criterio della somiglianza può utilizzare, oltre al concetto di *somiglianza/differenza*, altri concetti considerati alternativi come *relatività/generalità*, *dipendenza/indipendenza*, *associazione/separazione*, *sostituibilità/insostituibilità*, *confusione/distinzione*, e così via.
Qualunque approccio si decida di adottare, e quindi del tipo di risposte richieste ai soggetti, è importante:

1. verificare se ciascun soggetto è nelle condizioni (sia in termini di capacità che di possibilità) di esprimere risposte che possano essere confrontabili con le altre. Si pensi a determinate condizioni in cui può essere difficile comprendere la corretta applicazione della somiglianza percentuale;
2. tenere conto che il concetto di somiglianza tra oggetti soprattutto in campo psicologico è piuttosto ambiguo e ingannevole e può condurre a diverse difficoltà di interpretazione dei risultati; i giudizi di somiglianza non sono dati dai soggetti sulla base di un concetto astratto di somiglianza assoluta; è quindi difficile ottenere giudizi oggettivi di somiglianza in quanto i criteri di valutazione, difficili da conoscere, cambiano da soggetto a soggetto. Anche se dal punto di vista matematico i concetti di "somiglianza" e di "differenza" sono inversamente correlati, in pratica si possono ottenere risultati diversi se si chiede allo stesso campione di soggetti di dare giudizi di somiglianza e di differenza sullo stesso insieme di oggetti.

[13] Occorre però sempre tenere presente che anche il riferimento assoluto di fatto si realizza nel soggetto attraverso una analisi e una interpretazione compativa. Infatti anche quando si richiedono valutazioni assolute, gli individui tendono a dare risposte sulla base di confronti o comunque a relativizzare rispetto a precedenti esperienze. Nella vita di tutti giorni ciò accade abitualmente; gli individui hanno sensazioni riguardanti la simpatia assoluta per un oggetto o un'attività, ma tali sentimenti sono influenzati dalla gamma di oggetti o attività disponibili.

visti.



2.3.2.1 Scaling comparativo

Lo *scaling comparativo*, che nella pratica si rivela molto versatile, richiede la presentazione al soggetto di due o più stimoli rispetto ai quali il soggetto effettua un confronto in termini

- relativi (il soggetto deve esprimersi sulla uguaglianza o differenza tra stimoli)
- ordinali (il soggetto deve esprimersi sul rango di ciascuno stimolo rispetto agli altri).[14]

Per questo motivo tale *scaling* è note anche come -metrico. I principali vantaggi dello *scaling* di confronto sono:

- possibilità di rilevare anche piccole differenze tra gli oggetti/stimolo; ciò è dovuto al fatto che si richiede di una scelta forzata tra loro;
- stessi punti di riferimento per tutti i soggetti che affrontano il procedimento di confronto con la conseguente semplicità di comprensione e facilità di applicazione;
- presenza di pochi assunti teorici;
- riduzione dell'effetto alone tra un giudizio e l'altro.

Il principale svantaggio delle scale di confronto riguarda la natura ordinale dei dati e l'incapacità di generalizzare gli oggetti/stimolo scalati.

Nella pratica tali scali si configurano in diverse tipologie.

Scelta comparata

Questo approccio richiede la definizione di una serie di stimoli (per esempio, aggettivi,

[14] Il confronto può avvenire anche tra uno stimolo ed un riferimento di altro tipo (il passato, le proprie aspettative, ecc.) e attraverso valutazioni del tipo *inferiore alla media, coincidente con la media, superiore alla media* oppure *mi sarei aspettato di più, più o meno uguale, mi sarei aspettato di meno* oppure *superiore, più o meno uguale, inferiore*.

affermazioni) da somministrare al soggetto che, rispetto al riferimento espresso, sceglie quelli che meglio descrivono una certa situazione, una certa figura, un certo personaggio, ecc. Il limite di questo approccio sta nel fatto che al termine della rilevazione per ciascuno stimolo si una valutazione dicotomica (scelto/non scelto) che non consente molte elaborazioni successive. Tipica applicazione di questa tecnica è l'*Adjective Check List*.

Confronti accoppiati

Due stimoli vengono presentati ad un soggetto che deve scegliere tra essi secondo un certo criterio definito. Questo può essere sia il riferimento cognitivo (per esempio, criterio della somiglianza tra stimoli) che quello affettivo (criterio della preferenza).[15]
Esistono delle varianti; per esempio è possibile dare al soggetto la possibilità di esprimere una risposta in termini monetari, di peso, ecc.; oppure di riferire una *risposta neutrale* (per esempio, *nessuna differenza tra gli stimoli* o *nessuna opinione in proposito*).
Una variante dei confronti accoppiati è quella che prevede il confronto non tra due ma tra tre stimoli per volta (metodo delle *triadi*). In questi casi occorre definire tutte le possibili triadi di oggetti.
Sia nei confronti accoppiati semplici che quello con triadi, quando il numero degli stimoli da confrontare è troppo elevato, diventa complicato ottenere valutazioni per tutte le coppie o le triadi possibili (ricordiamo che con *k* stimoli il numero di coppie da valutare è uguale a $k(k-1)/2$).
In determinati questi casi è possibile richiedere a ciascun soggetto di formare gruppi di oggetti relativamente a determinati aspetti (metodo dei *cluster*). In questi casi gli stimoli devono essere posti in categorie tra loro esclusive ed esaustive: gli oggetti appartenenti alla stessa categoria devono essere molto simili tra loro e poco con quelli delle altre categorie. La misurazione tra ciascuna coppia di oggetti viene ottenuta dal semplice conteggio del numero di volte in cui i due oggetti sono risultati nello stesso gruppo.
Se l'assunto di *transitività delle scelte*[16] può essere soddisfatto, è possibile convertire i dati dei confronti accoppiati in ranghi.
E' possibile determinare il numero di volte o la percentuale in cui ciascuno stimolo è stato preferito agli altri.
Vedremo come i dati ottenuti attraverso questa tecnica su un gruppo di stimoli possano essere trattati a livello di modelli di *scaling*.

Rank order

Dopo aver presentato simultaneamente a ciascun soggetto diversi stimoli, si richiede di metterli in ordine secondo il criterio definito (cognitivo o affettivo). Tale tecnica è più semplice e parsimoniosa della precedente e non richiede che venga soddisfatto l'assunto di transitività.
Perché la tecnica possa essere applicata nel modo più corretto, è necessario che

- il soggetto sia in grado di dare un ordine a tutti gli elementi ovvero sia nelle condizioni di conoscerli tutti;
- il numero degli stimoli non sia elevato in modo da mettere i soggetti nelle condizioni di poterli ordinare; a tale proposito ricordiamo che in presenza di *n* stimoli, ciascun soggetto deve prendere $n-1$ decisioni rispetto alle $n(n-1)/2$ richieste nei confronti accoppiati.

[15] Un particolare applicazione del concetto di confronto a coppie è quello, considerabile più oggettivo, che mira per esempio alla valutazione della quantità di comunicazione e interazione registrata tra individui, città, gruppi o altri elementi, per esempio traffico telefonico, volume di viaggi, ecc. Come vedremo, tali dati possono essere sottoposti ad analisi multidimensionale per valutare la presenza di una mappa sociometrica, in cui una grande distanza riflette minore interazione tra gli elementi associati, o di una mappa del flusso di comunicazione o di informazione.

[16] Per esempio, se l'oggetto *A* è preferito all'oggetto *B*, e l'oggetto *B* è preferito all'oggetto *C*, allora l'oggetto *A* è preferito all'oggetto *C*.

Rating comparativo

Rispetto al *rank order*, con il *rating* comparativo si richiede al soggetto di indicare per ogni stimolo un valore secondo il criterio indicato e in confronto con gli altri stimoli. Il rating comparativo può essere rappresentato anche in termini proporzionali o percentuali; in alcuni casi definire scale bipolari che vanno da 100% a -100%.

Somma costante

Secondo questo approccio, ciascun soggetto distribuire una certa somma di valore (punteggi, denaro, ecc.) tra gli stimoli utilizzando il criterio definito. La somma costante consente di identificare un continuum ordinale e offre la possibilità di discriminare tra gli stimoli in modo chiaro e veloce. Occorre fare particolare attenzione in fase di somministrazione in quanto se un soggetto, nell'attribuire i punteggi, non utilizza esattamente la somma assegnata (in eccesso o in difetto) rende inutilizzabili i dati ottenuti per l'analisi. Inoltre, come per le precedenti tecniche, l'uso di un numero elevato di oggetti può rendere il compito lungo, faticoso e confuso.

2.3.2.2 Scaling non-comparativo

Con il riferimento <u>assoluto</u> si produce uno *scaling* detto *non-comparativo* in quanto il soggetto valuta ogni stimolo indipendentemente dagli altri stimoli, attribuendo a ciascuno di questi una quantità , una posizione o una affermazione, definite in precedenza. Si assume che i dati prodotti dal procedimento rappresentino un continuum metrico (*scaling* metrico).
Esistono varie tecniche che realizzano lo *scaling* non-comparativo tutte, in genere, semplici da costruire e da somministrare[17].
Tali tecniche possono essere classificate rispetto a

- **tipo di rappresentazione** (verbale, numerica o grafica) del continuum (Aureli, 1977). La scelta del tipo di rappresentazione è molto legata alla modalità di rilevazione e di somministrazione (presenza o meno del rilevatore, utilizzo dello strumento cartaceo, telefonico, video, ecc.)
- **numero di segmenti.**

Successivamente sarà necessario procedere all'**attribuzione di valori ai segmenti** per rendere i dati utilizzabili per le successive analisi; per fare ciò occorre definire (come vedremo) un vero e proprio sistema di misurazione.

Tipo di rappresentazione

- ***Rappresentazione verbale***

In questo caso il continuum è suddiviso in segmenti ciascuno dei quali è definito da un'etichetta verbale che ne dovrebbe esplicitare il significato.
Nel caso in cui, per esempio, si richiede di utilizzare un criterio di valutazione che esprima un accordo, il continuum può essere suddiviso nel modo seguente:

molto favorevole - favorevole - indifferente - sfavorevole - molto sfavorevole

oppure

d'accordo - abbastanza d'accordo - non so - piuttosto in disaccordo - in disaccordo

Questo rappresenta il tipico *scaling* noto con il nome del ricercatore che l'ho per primo definito (Rensis <u>Likert</u>). In genere si richiede al soggetto di attribuire allo stimolo presentato (in genere una affermazione) uno dei livelli identificati a seconda del proprio livello di accordo. Tale tipo di *scaling* è in genere utilizzato per misurare dimensioni di personalità, opinioni, atteggiamenti, valori.

[17] In Maggino (2003) sono presentati i risultati di una ricerca che aveva l'obiettivo di valutare l'influenza che il tipi di rappresentazione, la polarità, il numero dei livelli del continuum hanno sui dati rilevati.

In fase di definizione dei segmenti in cui viene suddiviso il continuum verificare

a. l'ordine dei livelli,
b. la polarità dei livelli (dall'accordo al disaccordo → bipolare),
c. la simmetria dei livelli,
d. l'equidistanza tra i livelli.

Tali verifiche non sempre conducono a risultati chiari ed evidenti e sono molto legate al contesto linguistico e culturale nel quale si opera. Anche se non sempre necessario, spesso viene incluso all'interno di tale sequenza valutativa un punto neutrale (né in accordo né in disaccordo).
Vediamo di seguito alcuni esempi:

	Molto	Abbastanza	Poco	Per niente
Lo sport favorisce la solidarietà				

oppure

1. completamente in disaccordo 2. abbastanza in disaccordo 3. un po' in disaccordo				4. un po' d'accordo 5. abbastanza d'accordo 6. completamente d'accordo		
Lo sport favorisce la solidarietà	1	2	3	4	5	6

oppure

1. completamente in disaccordo 2. abbastanza in disaccordo 3. un po' in disaccordo	4. un po' d'accordo 5. abbastanza d'accordo 6. completamente d'accordo
Lo sport favorisce la solidarietà	☐

Il continuum può essere suddiviso verbalmente anche in termini di frequenza (*sempre*, *spesso*, *a volte*, *raramente*, *mai* oppure *mai, per poco tempo, per qualche tempo, per la maggior parte del tempo* oppure *una volta al giorno, due volte al giorno, ecc.*), utilizzata in relazione ad eventi, circostanze o comportamenti; in pratica si richiede quante volte o quanto spesso si verifica, o dovrebbe verificarsi, ciò che viene presentato.

- ***Rappresentazione numerica***

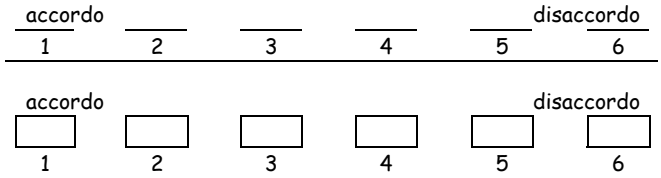
In questo caso il continuum viene suddiviso in segmenti ciascuno dei quali è definito in termini numerici. L'utilizzo di numeri dovrebbe aiutare a riconoscere la presenza di un *continuum*, a ricordare la gradualità delle valutazioni e a evitare i problemi d'interpretazione semantica propri delle rappresentazioni verbali.
Per esempio, si può chiedere ai soggetti di identificare con "0" il peggiore stato possibile e con "10" il miglior stato possibile relativamente ad una certa dimensione (soddisfazione di vita, del proprio lavoro, della situazione politica, ecc.) e quindi di indicare il valore che identifica il proprio stato. A volte lo *scaling* che utilizza una rappresentazione numerica è definito *rating*.
Nel caso in cui la rappresentazione è visualizzata, è necessario prestare particolare attenzione alla corrispondente rappresentazione del continuum in termini di:

- <u>orientamento</u>, che può essere orizzontale o verticale (secondo alcuni per esempio la scala verticale risulta essere più familiare alla maggior parte dei soggetti);
- <u>allineamento dei segmenti</u>, che possono risultare:
    a. *uniti*, facendo così diretto riferimento alla ipotizzata presenza del continuum; vediamo due esempi:

accordo					disaccordo
1	2	3	4	5	6

accordo					disaccordo
1	2	3	4	5	6

b. *separati*: a favore della visualizzazione di livelli separati vi è la considerazione secondo la quale i soggetti indicherebbero con maggiore chiarezza la loro posizione.

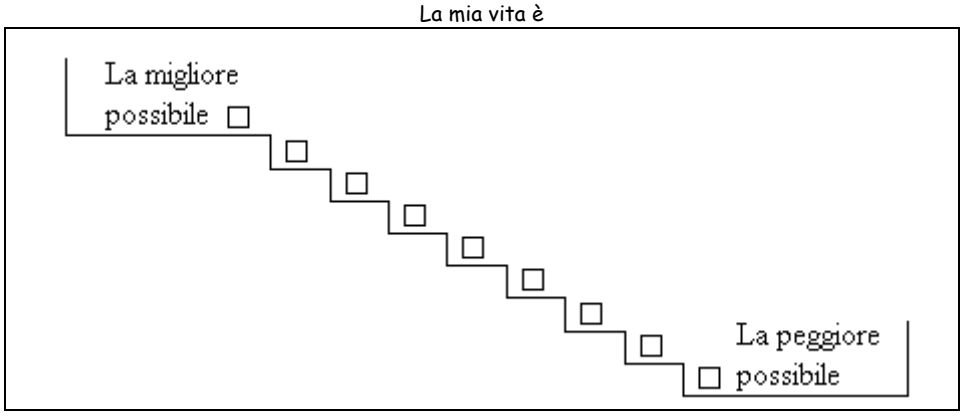


La rappresentazione numerica può essere utilizzata anche quando il continuum è suddiviso in termini di frequenza; in questo caso si rappresenta facendo riportando proporzionali o percentuali; in questi casi è possibile che si richieda al soggetto che la somma dei valori percentuali rilevati per una serie di stimoli consenta di ricomporre il totale.

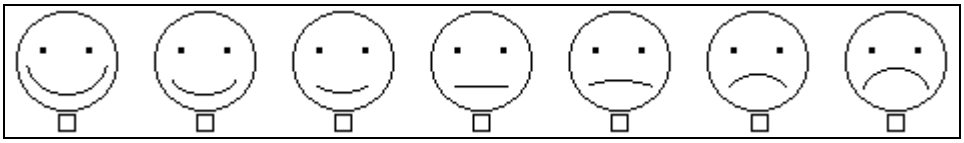
- ***Rappresentazione grafica***

Diversi sono i problemi che sorgono nella rappresentazione verbali (problemi d'interpretazione semantica) e numeriche (problemi di interpretazione del continuum in termini ordinali o quantitativi). Tali problemi si acuiscono nel passaggio di tali strumenti da una lingua all'altra (problemi di traduzione) e da un paese all'altro (problemi culturali). Tali problemi in molti casi possono essere superati facendo ricorso alla rappresentazione grafica che consentono varie soluzioni, a seconda del riferimento richiesto. La rappresentazione grafica consente di comunicare con maggiore chiarezza l'idea del continuum e di facilitare il compito dei soggetti che ricordano più facilmente i significati dei diversi livelli. Vediamo alcuni esempi.

- *Ladder scale:* il continuum è rappresentato come una scala a 9 o 11 pioli; è utilizzata usata per misurare, per esempio, il livello di soddisfazione per la propria vita; il soggetto deve indicare la posizione che meglio rappresenta la sua condizione di vita:

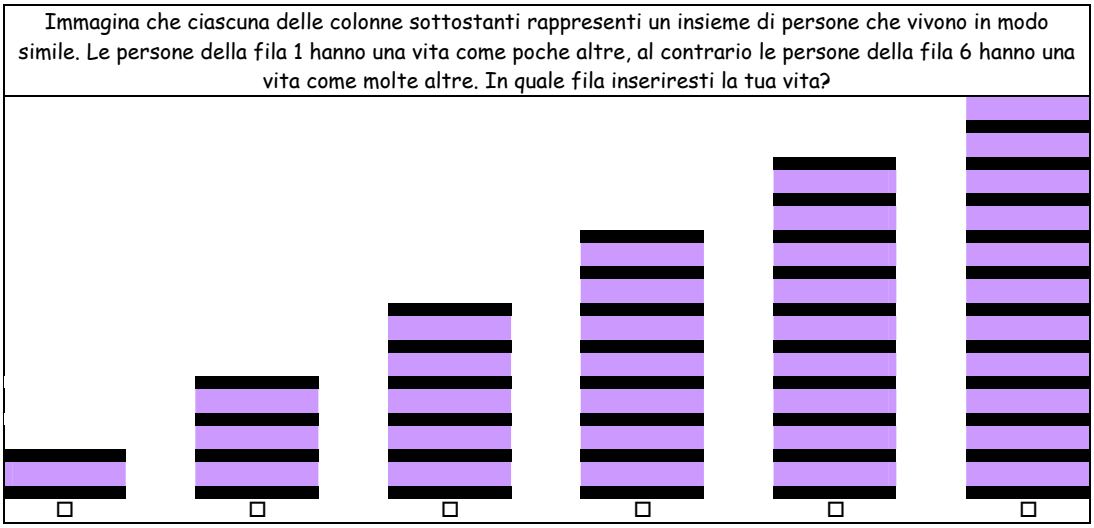


- *Faces scale*: il continuum con riferimento affettivo è rappresentato da 7 facce che si differenziano tra loro rispetto all'inclinazione della bocca; la diversa inclinazione esprime livelli diversi emotivi:

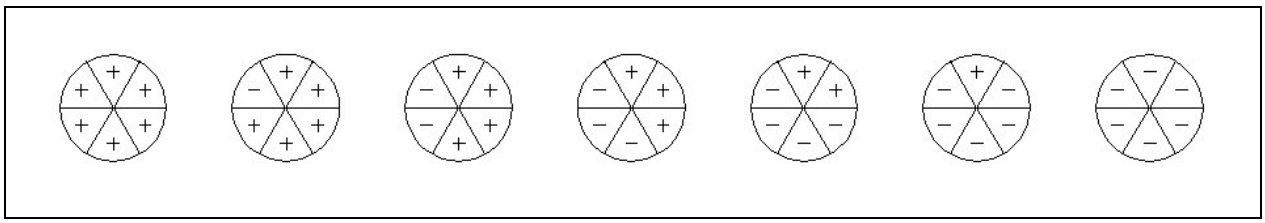


- *Weather scale*: il continuum è suddiviso in una serie di segmenti ognuno dei qauli è rappresentato da un evento atmosferico; la sequenza degli eventi dovrebbe richiamare il continuum di stato emotivo; vediamo alcune versioni:

- *Pile scale*: il continuum viene suddiviso in segmenti, ciascuno dei quali è rappresentato da una pila di diversa altezza; al soggetto si chiede di indicare la pila che meglio rappresenta un particolare riferimento:



- *Circle scale:* il continuum è suddiviso in segmenti ognuno dei quali è rappresentato da un cerchio suddiviso in spicchi che contengono un "+" o un "-"; la sequenza dei cerchi, ordinata in modo da contenere un numero decrescente di segni "+" ed un numero crescente di segni "-", viene utilizzata per richiamare riferimenti di tipo affettivo (sentimenti); per esempio, si richiede a ciascun soggetto di collocarsi, tra i sette cerchi, in quello che meglio rappresenta un proprio stato d'animo rispetto alla soddisfazione verso la propria vita. Come si può notare, si tratta di una sequenza simmetrica e bipolare, con la rappresentazione di una posizione intermedia o neutrale:



Ancoraggio

Sia nel caso della rappresentazione numerica che nel caso della rappresentazione grafica, i segmenti estremi possono o meno trovare una indicazione verbale; per questo si distingue tra *scaling*:

- *ancorato* (*anchoring scale*), quando viene data definizione verbale degli estremi del continuum, detti *agenti di ancoraggio*; la definizione di tali vincoli consente di facilitare il compito del soggetto; tra gli esempi di *scaling* grafico che utilizza l'ancoraggio vi è il **differenziale semantico**. Tale tecnica di *scaling*, utilizzata (come vedremo) per un approccio complesso alla misurazione sviluppato dal gruppo di C.E. Osgood, 1969, è definita da una coppia di aggettivi bipolari (per esempio: forte/debole, eccitato/calmo, caldo/freddo, veloce/lento, ecc.); tra i due aggettivi è posta la serie di segmenti (in genere cinque o sette): il soggetto reagisce allo stimolo ponendosi tra i due aggettivi utilizzando uno dei segmenti indicati.

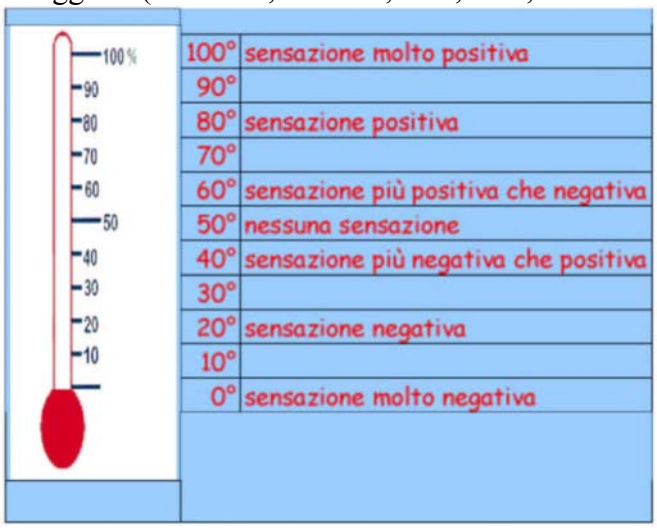
Pensando alla sua città, indichi una crocetta più o meno vicina all'aggettivo che pensa sia più adeguato a descrivere la sua idea della città.

silenziosa ☐☐☐☐☐☐☐ rumorosa

Il limite di tale approccio sta nella difficoltà di definire ed identificare coppie di aggettivi realmente bipolari.[18] Di seguito vediamo un altro esempio di *scaling* ancorato:

completamente in disaccordo	1	2	3	4	5	6	7	8	9	10	completamente d'accordo

Vediamo ora un altro esempio di suddivisione del continuum, anche questo con approccio misto, numerico e grafico, che presenta non solo un ancoraggio ma anche delle definizioni verbali di altri segmenti: il cosiddetto *termometro dei sentimenti* (*feeling thermometer*); in questo caso si istruisce l'intervistato a identificare con "0" il massimo di ostilità e con 100 il massimo favore verso l'oggetto (individui, concetti, idee, città, istituzioni, ecc.)[19].



- *auto-ancorato* (*self-anchoring scale*), quando non vengono definiti gli agenti di ancoraggio; tale

[18] E' possibile ovviare a tale problema utilizzando l'approccio conosciuto con il nome del ricercatore che l'ha sviluppato, Jan Stapel. Secondo tale approccio si individua un solo aggettivo rispetto al quale il soggetto indica la propria valutazione.

[19] Tale tecnica ha trovato molte applicazioni ed è ispirata al lavoro di Andrews F.M., Withey S.B. (1976) che rappresenta un punto di riferimento fondamentale per lo studio delle diverse tecniche di *scaling*.

approccio pur presentando degli indubbi vantaggi, può produrre punteggi non confrontabili tra loro in quanto gli ancoraggi definiti da ogni soggetto possono dipendere:

- dalle istruzioni che vengono date all'intervistato,
- dalle tendenze di risposta del soggetto (*response set*),
- dalle passate esperienze del soggetto.

Numero di segmenti

Non è sempre facile decidere qual è il numero ottimale di segmenti entro il quale deve essere suddiviso il continuum. Sappiamo però che per avere un buon livello di affidabilità della misurazione è sempre più vantaggioso utilizzare più livelli. E' infatti stato dimostrato che l'affidabilità delle scale è monotonamente legata, in senso positivo, al numero dei livelli; in altre parole, la capacità di misurare in maniera più o meno raffinata aumenta all'aumentare del numero delle posizioni definite. In altre parole, il livello di affidabilità aumenta in modo significativo al crescere dei livelli; tale aumento diviene però poco significativo, o nullo, se il numero dei livelli risulta troppo elevato. Nel caso si un numero elevato di segmenti, il soggetto può trovare difficile esprimere la propria posizione. In molti casi ciò può portare ad una polarizzazione delle risposte verso alcuni valori; si pensi, a tale proposito, al continuum espresso attraverso valori numerici che fanno riferimento alle percentuali (da 0 a 100); in questi casi i soggetti possono essere indotti ad utilizzare sono alcuni valori (5, 10, 20, 50, ecc.) vanificando l'idea di precisione della misurazione.
La scelta del numero di livelli deve inoltre tenere conto anche del corretto bilanciamento; una scala è bilanciata quando il numero di livelli positivi/favorevoli è uguale al numero di livelli negativi/sfavorevoli.
Associata alla scelta del numero dei livelli vi è la questione se sia preferibile un numero di posizioni pari o dispari. Un numero dispari consente di introdurre un livello intermedio cui attribuire un significato di "neutralità"; si ritiene che la definizione di una posizione mediana può rendere *confortevoli* le risposte. D'altra parte si obbietta che l'uso di un livello intermedio può introdurre i cosiddetti *response style*: il soggetto finisce per preferire la posizione intermedia per indecisione o per non volersi sbilanciare in una chiara posizione. Ciò può compromettere la valutazione delle differenze individuali in quanto risulterà difficile distinguere tra risposte che riflettono reali atteggiamenti neutrali e quelle che riflettono la non intenzione di esprimersi.

3. IL MODELLO PER L'ATTRIBUZIONE DEI VALORI: IL SISTEMA DI MISURAZIONE

3.1 LE REGOLE DI ATTRIBUZIONE

3.1.1 Genere di misurazione

Il genere fa riferimento alla *qualità dell'informazione rappresentata dai numeri* ovvero al modo in cui si sviluppa e si attribuisce significato alla misurazione. L'attribuzione del significato può avvenire a partire da un modello precedente, sulla base di esperienze, o a partire da considerazioni pratiche o di senso comune. In particolare è possibile distinguere tre diversi generi:

a. Misurazione fondamentale o misurazione per processo fondamentale: la misurazione è considerata *fondamentale* quando non fa riferimento a precedenti misurazioni ovvero quando:
   - per l'assegnazione di simboli non si richiede la misurazione di altre variabili (significato *operativo*);
   - i simboli riflettono leggi naturali che correlano quantità diverse della caratteristica (significato *costitutivo*).

   Esempi di caratteristiche misurate attraverso procedure fondamentali sono la lunghezza, la resistenza, il volume. La misurazione fondamentale rappresenta un esempio di costruzione e di verifica di teorie.
b. Misurazione per derivazione o misurazione derivata: la misurazione di questo tipo è basata su altre misurazioni, tra loro legate da una teoria più ampia; in particolare è costruita attraverso algoritmi applicati a misure fondamentali; un esempio tipico di misurazione derivata è la *densità* definita come il rapporto tra massa e volume; la legge di riferimento definisce il rapporto tra massa e volume come costante per qualsiasi quantità di una data sostanza e come diverso tra le diverse sostanze. Un altro esempio di misurazione derivata è la velocità definita come il rapporto tra la misura dello spazio percorso e quella del tempo trascorso.
c. Misurazione per definizione o misurazione per relazione: il significato è ottenuto attraverso una definizione arbitraria e, in genere, dipende da relazioni presenti tra le osservazioni e il concetto che interessa misurare. Tale misurazione è adottata ogni volta che, non potendo misurare direttamente la caratteristica, si misurano o si pesano altre variabili che si ritiene debbano essere correlate con tale concetto. A questa categoria appartengono molti degli indici e degli indicatori utilizzati nelle scienze sociali e psicologiche (*status socio-economico*, la *capacità di apprendimento*, ecc.), per le quali non sempre si dispone in partenza di un sistema teorico che consenta di introdurre nuovi concetti attraverso un procedimento fondamentale.

3.1.2 Criterio di misurazione

I criteri di misurazione definiscono il modo di rilevare l'informazione (Bruschi, 1999). Nella ricerca sociale è possibile distinguere tra diverse modalità di misurazione, principalmente

- come "risposta" a una "sollecitazione", come succede con le domande di un questionario,
- come rilevazione o registrazione di comportamenti o di eventi,
- come registrazione di eventi di varia natura.

A ciascuna di tali modalità è possibile applicare diversi criteri di misurazione, la cui applicazione richiede la definizione degli eventi e degli indicatori da osservare, l'accertamento della loro omogeneità (Bruschi, 1999):

1) criterio della frequenza, secondo il quale si conta il numero di risposte, comportamenti, eventi, secondo una definita variabile, quindi si rapporta tale numero a quello totale dei casi rilevati[1];
2) criterio della latenza, che si riferisce, per esempio, al tempo che trascorre tra lo stimolo e la risposta;
3) criterio della durata, che si riferisce, per esempio, al tempo in cui un singolo comportamento è mantenuto;
4) criterio dell'intensità, difficile da definire in quanto a volte si può sovrapporre al criterio della frequenza, adottato spesso come indicatore di intensità[2];
5) criterio della manifestazione, secondo il quale a ciascuno stato della proprietà da misurare corrisponde una variabile; tra le misurazioni ottenute secondo il criterio della manifestazione che utilizzano più variabili ricordiamo la scala per la misurazione dell'intensità e della magnitudo dei terremoti[3], basata sulla percezione di eventi, e la scala per la misurazione della forza del vento; nel caso in cui si utilizzino più variabili è necessario accertarsi che esse

---

[1] Una particolare evoluzione di tale criterio è quello converte il concetto di *frequenza* in quello di *probabilità* nelle due accezioni di:

- *probabilità congiunta*, ovvero probabilità che l'oggetto/evento *i* e l'oggetto/evento *j* si verifichino contemporaneamente,
- *probabilità condizionata*, ovvero probabilità che si verifichi l'oggetto/evento *i* posto che si sia verificato l'oggetto/evento *j*. In genere le matrici che si ottengono con questo approccio non sono simmetriche ma possono essere rese simmetriche attraverso particolari trasformazioni come la seguente:

$$\delta_{ij} = p_{ij} + p_{ji}$$

Naturalmente dopo una tale trasformazione l'indice di somiglianza non ha lo stesso significato delle probabilità che lo definiscono: dà solo una globale quantificazione della somiglianza tra i due oggetti.

Tale approccio è stato generalizzato e ha trovato applicazione in diverse discipline; in demografia, per esempio, è possibile utilizzare una matrice di probabilità di migrazione tra regioni; in studi sulla mobilità sociale o scolastica (in riga vi possono essere i titoli di scuola media superiore e in colonna i corsi di laurea prescelti). In questi casi, in cui l'asimmetria della matrice non è dovuta ad effetti casuali ma alla natura stessa delle informazioni, è possibile

- analizzare la matrice triangolare superiore o inferiore (per esempio analisi sulle emigrazioni o sulle immigrazioni),
- rendere simmetrica la matrice (analisi del flusso migratorio totale).

Le valutazioni possono avvenire sia rispetto ad attributi noti e definiti sia rispetto ad attributi non noti; in quest'ultimo caso l'obiettivo della misurazione potrebbe essere proprio quello di esplorare dimensioni non note. Quest'ultimo approccio trova la sua maggiore applicazione nelle indagini di mercato: gli oggetti, in questo caso, sono rappresentati dai prodotti che ciascun consumatore valuta quantificandone la somiglianza.

[2] Vedremo come in particolari casi i modelli che esamineremo potranno essere utilizzati per confrontare tra loro e "ordinare" gli indicatori; in un certo senso in questi casi l'oggetto da misurare è rappresentato dagli indicatori.

[3] La scala che misura l'intensità dei terremoti è stata elaborata dal sismologo Mercalli (1850-1914) e si basa sugli effetti prodotti:

Intensità	Indicatori
1° grado: Strumentale	Registrata solo dagli strumenti
2° grado: Leggerissima	Avvertita agli ultimi piani delle case
3° grado: Leggera	Avvertita da poche persone
4° grado: Sensibile	Avvertita da chi si trova in casa
5° grado: Sensibilissima	Oscillazione di oggetti mobili
6° grado: Forte	Caduta di oggetti e calcinacci
7° grado: Fortissima	Caduta di camini e lesioni nei fabbricati
8° grado: Rovinosa	Caduta di pareti interne
9° grado: Disastrosa	Crollo di alcuni fabbricati
10° grado: Distruttrice	Caduta di molti fabbricati
11° grado: Catastrofica	Distruzione completa dei fabbricati
12° grado: Grande Catastrofe	Distruzione di ogni opera umana

rappresentino la stessa proprietà[4];

6) criterio dell'assegnazione soggettiva, secondo la quale è il soggetto che stima l'intensità della presenza di una proprietà in un oggetto assegnandogli un valore.

3.2 IL SISTEMA DI CLASSIFICAZIONE

3.2.1 Tipo di misurazione

Il tipo di misurazione fa riferimento al modo di registrare la misurazione. A tale proposito ricordiamo come il carattere quantitativo della misurazione rappresenta uno dei temi maggiormente discussi. In realtà non si può essere sempre d'accordo nell'enfatizzazione della quantità e quindi del ruolo del numero nel processo di misurazione. Infatti non sempre l'utilizzo di simboli numerici indica una reale corrispondenza tra numero e presenza di una quantità.[5]
Per questo sarebbe più corretto parlare di misurazione in termini di registrazione dello *stato di ciascun caso sulla proprietà in questione, assegnando tale stato a una delle categorie di un elenco già predisposto* (Marradi, 1980).
La possibilità di misurare una certa caratteristica in termini quantitativi non rappresenta un problema quando lo stato di una caratteristica è percepito e misurato e può essere riferito in termini quantitativi e in particolare (Marradi, 1980; Velleman, 1993):

- come risultato di conteggi ovvero contando gli oggetti posseduti o relativi a ogni oggetto (conteggio);
- come misura composta da unità proprie delle scienze fisiche (per esempio lunghezza, peso) e sottoponibile ad operazioni aritmetiche (età, anzianità di lavoro, lunghezza della rete stradale di un paese, ecc.),
- ottenute attraverso manipolazioni matematiche dei precedenti tipi.

Mentre nel primo caso non si può parlare di misurazione vera e propria, negli altri due casi si deve

---

[4] In un'altra ottica si può parlare di *identificazione della fonte informativa* che può essere
- diretta, quando è rappresentata dall'oggetto stesso,
- indiretta, quando si ricorre a fonti esterne all'oggetto.

Quando, per esempio, si vuole misurare il livello di autosufficienza in un gruppo di anziani, la fonte
- diretta è rappresentata dalle misurazioni effettuate direttamente sui soggetti (svolgimento di prove),
- indiretta è rappresentata dalle risposte date dai soggetti relativamente alla loro capacità di fare o non fare determinate operazioni.

[5] Per tale motivo, può essere utile cercare di chiarire il rapporto tra *misurazione* e *quantificazione* (Nunnally, 1978). Infatti spesso si enfatizza il ruolo che il numero assume all'interno del processo di misurazione, finendo con il fare confusione tra misurazione e matematica, cui va aggiunta la statistica per il ruolo che gioca nel processo di validazione e di analisi delle misurazioni; è importante per questo fare una chiara distinzione tra le tre dimensioni:

- *Matematica*: disciplina astratta che opera su enti logici, definiti da proprietà non contraddittorie, senza aver la necessità di riferimenti empirici ovvero non necessariamente riguarda il mondo reale. I sistemi matematici sono puramente deduttivi, essendo composti da insiemi di regole per la manipolazione di simboli. Le quantità costituiscono solamente uno dei tipi di simboli presenti e gestiti in matematica; molta parte della matematica moderna tratta simboli che non necessariamente sono identificabili con numeri. Ciò vuol dire che qualsiasi insieme di regole internamente consistenti per la manipolazione di simboli può essere considerato legittimamente appartenente alla matematica.
- *Misurazione*: la misurazione riguarda direttamente il mondo reale e tratta sempre valori (di qualità o quantità) e la legittimazione di qualsiasi sistema di misurazione è determinata dai dati empirici (estratti dal mondo reale).
- *Statistica*: utilizza molti strumenti e concetti matematici ma solo con intenti strumentali anche nel caso in cui operi a livello astratto (statistica metodologica): la statistica opera su simboli che hanno riferimenti ad eventi empirici. In statistica il numero presenta un riferimento diretto con una situazione che, se anche simulata, può essere considerata reale.

parlare, come abbiamo visto, di misurazione derivata.

3.2.2 Livello di misurazione

Con "livelli di misurazione" ci si riferisce alle proprietà aritmetiche dei valori assegnati ai casi relativamente alla proprietà misurata (Caracciolo in Siegel, 1992) ovvero le relazioni matematiche tra gli elementi che definiscono il sistema di classificazione.
L'associazione tra gli elementi di un sistema empirico e quelli di un sistema numerico corrisponde a stabilire una funzione di relazione; tale funzione è detta "scala di misurazione"; è possibile identificare diverse scale a seconda dell'operazione di connessione permessa (Bruschi, 1999; Stevens, 1946; Velleman, 1993).

3.2.2.1 Scala nominale

La scala nominale corrisponde al tipo di misurazione più semplice (*classificazione*) e richiede la definizione di un insieme di categorie secondo il sistema di regole di attribuzione precedentemente definiti. La scala è definita da categorie che consentono di raggruppare elementi che presentano una relazione qualitativa, specificata tra le categorie (per esempio professione, titolo di studio, ecc.). Vediamo schematicamente quali sono le caratteristiche della scala nominale:

- *Definizione*: ogni oggetto viene classificato utilizzando simboli (numeri, lettere) ognuno dei quali corrisponde ad una modalità (categoria) del carattere misurato. L'insieme dei simboli costituisce la scala.
  La classificazione deve rispettare i seguenti criteri:
  a. le categorie devono essere definite prima di procedere alla classificazione degli oggetti e devono essere più di una[6];
  b. ogni oggetto deve essere attribuito a una categoria (esaustività dell'insieme delle categorie);
  c. nessun oggetto può essere attribuito a più di una categoria (mutua esclusività delle categorie);
  d. a ciascuna categoria può essere assegnato più di un oggetto;
  e. tutti gli oggetti assegnati alla stessa categoria presentano la stessa modalità dell'attributo misurato;
  f. l'attribuzione deve basarsi su un unico criterio (*fundumentum divisionis*).
- *Proprietà*: se le categorie identificate sono esaustive (ovvero riescono a definire tutte le possibili risposte) e si escludono a vicenda (ovvero ogni risposta può ricadere in un'unica categoria) la scala nominale presenta le seguenti proprietà:
  - simmetria $\Rightarrow$ se $a = b$ allora $b = a$
  - transitività $\Rightarrow$ se $a = b$ e $b = c$ allora $a = c$
  - riflessività.
- *Operazioni*: i simboli che identificano le diverse categorie della scala possono essere *interscambiati* senza alterare le informazioni essenziali (purché ciò venga fatto in modo sistematico e completo per tutte le categorie).

La scala nominale, pur essendo molto utilizzata soprattutto in ambito sociale, è però soggetta ad errori di definizione che possono produrre distorsioni in fase di misurazione che abbassano il livello di attendibilità; è il caso per esempio della mancata individuazione di una categoria che, in fase di misurazione, si rivela corrispondere a uno stato dell'oggetto ben distinto dagli altri; in questo caso si corre il rischio di ottenere un non-risultato (dato *missing*), un risultato ricondotto ad altra categoria o un risultato ricondotto ad una categoria rifugio (si pensi alla categoria *altro* utilizzata nelle domande

[6] Quando è possibile individuare e definire due categorie si parla di *scale dicotomiche*. In genere tali categorie riflettono posizioni del tipo *si/no*, *maschio/femmina*. Se i dati dicotomici possono essere significativamente rappresentati dai codici 0 e 1 si parla di *scale binarie* come quelle composte da posizioni del tipo *vero/falso*, *presenza/assenza* di una certa caratteristica, *superamento/fallimento* di una prova.

di questionari). Un altro possibile errore di definizione è quello dell'identificazione di categorie non adeguate allo studio.[7]

3.2.2.2 Scala ordinale

La scala ordinale corrisponde al tipo di misurazione chiamato *ordinamento*. Per poter ordinare occorre essere in grado di stabilire se un oggetto è *maggiore* (>), *minore* (<) o *uguale* (=) in relazione agli altri. In particolare l'ordinamento richiede che

a. gli oggetti siano ordinati, rispetto ad un determinato attributo, secondo un determinato criterio dal maggiore al minore;
b. non vi sia alcuna indicazione sulla quantità/livello dell'attributo posseduta/riferito in senso assoluto dall'oggetto;
c. non vi sia alcuna indicazione sulla distanza tra gli oggetti rispetto all'attributo.

Tale scala è definita da categorie che consentono di raggruppare elementi che presentano una relazione qualitativa, specificata tra le categorie (per esempio professione, titolo di studio, ecc.).
Sul piano pratico tale operazione che può essere di tre tipi:

a. assegnazione a ciascun oggetto del valore della categoria, appartenente ad una serie ordinata di categorie, alla quale si ritiene appartenga (classificazione ordinabile)[8];
b. assegnazione a ciascun oggetto del valore della categoria, definita a partire dalla suddivisione di un continuum (concettuale) in categorie discrete (continuum classificato)[9];
c. assegnazione a ciascun oggetto di un valore corrispondente alla posizione assunta in una graduatoria prodotta da una *performance*[10].

L'insieme dei valori identificati viene in genere indicato come scala ordinale. In un certo senso rappresenta un perfezionamento del criterio di misurazione visto in precedenza.
La scala ordinale presenta le seguenti caratteristiche:

- *Definizione*: gli oggetti classificati in una categoria non sono solo diversi da quelli classificati nelle altre ma tra loro è possibile stabilire relazioni del tipo *più grande-più piccolo*, *più difficile-*

[7] Si pensi alla proprietà *confessione religiosa* per la quale una possibile classificazione potrebbe essere: *cattolico*, *protestante*, *ortodosso*, *musulmano*, *buddista*, *brahmanista*, *confucista*, *scintoista*, *animista*, *altra confessione*, *senza religione*. Tale scala risulta essere squilibrata se utilizzata per classificare oggetti (individui, nazioni, ecc.) presenti su un territorio molto vasto e che comprenda più continenti, in quanto le prime tre modalità possono essere considerate specificazioni di una più ampia categoria (cristiano) che sta allo stesso livello di astrazione delle altre (Marradi, 1980). Tale scala è ugualmente squilibrata se si pensa di utilizzarla per esempio in ambiente europeo dove la specificazione delle altre categorie religiose potrebbe rientrare in un'unica categoria per la loro minore presenza sul territorio.

[8] Un esempio di tale tipo di classificazione è quella utilizzata per la variabile "titolo di studio": licenza elementare, licenza di scuola media inferiore, licenza di scuola media superiore, laurea.

[9] Un esempio di tale tipo di classificazione è quella utilizzata per la variabile "classi d'età": 0-5 anni, 6-10, 11-15, ecc.

[10] Rientrano in tale procedura:

- valori ordinabili (per esempio i voti riportati da un gruppo di individui in prove o esami). I seguenti voti riportati da un gruppo di studenti universitari rappresentano una serie di valori ordinati: 25 26 28 29 30 30
- graduatorie vere e proprie (in cui ciascun valore viene sostituito dal valore di graduatoria che esso occupa nella serie dei valori); ciascun voto della precedente serie può essere sostituito da quello di graduatoria, detto *rango* 1 2 3 4 5.5 5.5. Per procedere all'assegnazione, per ciascun caso, del valore di graduatoria (*rango*) occorre ordinare le osservazioni. A ciascuno dei punteggi originali ordinati viene associato il corrispondente valore di graduatoria; esso può andare da 1 a *n* (dimensione del campione). Anche nei casi in cui si assume la presenza di una distribuzione continua è possibile, soprattutto nel campo della ricerca sociale, osservare punteggi uguali (*ex-aequo*). I valori *ex-aequo* risultano avere la stessa posizione di graduatoria, uguale alla media dei ranghi che le singole osservazioni avrebbero se fossero state diverse. Quindi al termine di questa operazione, il punteggio originale di ciascuna osservazione è sostituito da un nuovo valore corrispondente alla posizione che occupa nella graduatoria; tale valore si chiama *rank* (*rango*).

Un numero troppo alto di valori parimerito è spesso indice di uno strumento di rilevazione imperfetto, poco affidabile, poco *raffinato*, non in grado di classificare e distinguere tra loro i vari punteggi.

più facile, ecc. In molti casi è possibile indicare tali relazioni con i simboli $<,>,=,\leq,\geq$. Il loro significato specifico dipende dalla natura della relazione che definisce la scala. I valori utilizzati nella scala ordinale indicano solamente le posizioni relative degli oggetti.

- *Proprietà*: tale scala presenta le seguenti proprietà:
  - transitiva (se $a > b$ e $b > c$ b allora $a > c$),
  - irriflessiva,
  - asimmetrica.
- *Operazioni*: sulle categorie ordinate identificate possono essere effettuate solo operazioni del tipo: $a = b$ o $a > b$ o $a < b$. Non è importante il valore del simbolo numerico che viene assegnato alle diverse categorie e quindi agli oggetti: l'importante è che il valore "più grande" venga assegnato all'elemento con il più alto livello e viceversa; se la sequenza dei simboli viene modificata per tutte le categorie, non vi è perdita di informazioni: se per esempio si inverte la serie dei simboli[11] l'ordine dei simboli della scala cambia ma non il significato.

Non è semplice costruire una scala ordinale in modo tale che soddisfi i criteri visti; in realtà i criteri garantiscono sempre l'ordinamento delle categorie, ovvero la costruzione della scala, ma non sempre la validità dei successivi utilizzi, per esempio il reale ordinamento degli oggetti. Ciò vale soprattutto nella misurazione di atteggiamenti, opinioni, ecc. espressi da individui.
Il pericolo di distorsioni è particolarmente presente soprattutto nel caso della definizione di categorie ordinate prodotte dalla suddivisione di un continuum (continuum classificato)[12]. In questo ambito la principale distorsione è data, in linea teorica, dal numero di categorie: minore è il numero di categorie maggiore è la possibilità che la reale posizione dell'oggetto lungo il continuum non sia rilevata; in altre parole "riducendo il numero delle categorie si accresce la distanza media tra la posizione di un oggetto e quella della categoria ad esso più vicina" (Marradi, 1980); in questo senso la massima distorsione si ottiene con l'utilizzo di due sole categorie (dicotomizzazione del continuum)[13]. D'altra parte sappiamo come la definizione di un numero molto grande di categorie non facilita la loro considerazione, comprensione e valutazione. Stabilire il numero ideale di categorie non è semplice e non prevede alcun criterio valido sempre.

3.2.2.3 Scala metrica

La scala metrica richiede una vera e propria *quantificazione*; ciò presuppone l'esistenza di un continuum lungo il quale posizionare e collocare l'oggetto misurato in corrispondenza della quantità della caratteristica posseduta dall'oggetto stesso; il valore numerico assegnato rappresenta la quantità e/o l'intensità di una certa proprietà posseduta da un oggetto. Conseguentemente, confrontando i valori numerici registrati da due oggetti è possibile stabilire:

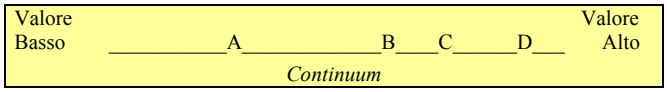
- se sono diversi,
- se uno è più grande dell'altro,
- quanto sono distanti.

Nel seguente esempio, la posizione assunta da un oggetto lungo il continuum consente di quantificare la sua distanza dagli altri oggetti, collocati lungo lo stesso continuum:

[11] Si inverte l'ordine di una serie assegnando ai singoli elementi ordinati da 1 a *n* i numeri da *n* a 1.

[12] Si pensi a tale proposito ad una tipica scala ordinale utilizzata per la misurazione di particolari opinioni: *molto d'accordo - parzialmente d'accordo - incerto - parzialmente in disaccordo - molto in disaccordo*. In questo caso gli individui classificati (o autoclassificati) nelle seconda categoria potrebbero nella realtà presentare livelli d'accordo diversi: alcuni più vicini alla prima, alcuni più vicini alla terza, altri in posizione centrale. Ciò che però risulta dalla misurazione è che tutti questi sono classificati nella seconda categoria.

[13] La dicotomizzazione del continuum aumenta anche gli errori di rilevazione: com'è facilmente intuibile un errore nel riportare il valore trasforma lo stato di un oggetto nel suo contrario.



In particolare, il *A* risulta avere una *quantità* di attributo considerevolmente più bassa rispetto a quella posseduta dagli altri oggetti; *B* e *C* hanno un punteggio simile e *D* presenta il punteggio più alto. Riportando gli intervalli è possibile stabilire con precisione le distanze tra gli oggetti.

Nel caso di scale metriche occorre fare una distinzione tra quelle che presentano una origine (corrispondente al valore zero) naturale o convenzionale. A partire da tale distinzione, la scala metrica può essere distinta tra scala a intervalli (quando l'origine della scala è convenzionale) e scala a rapporti (quando l'origine della scala è naturale).

Scala a intervalli

La scala metrica *ad intervalli* consente di

1. stabilire l'ordine degli oggetti misurati rispetto all'attributo,
2. conoscere la distanza tra gli oggetti misurati rispetto all'attributo;

Tale scala non dà alcuna informazione sulla grandezza assoluta di un oggetto misurato relativamente all'attributo[14].

- *Definizione*: presenta tutte le caratteristiche di una scala ordinale con in più la possibilità di stabilire la distanza tra due valori riportati su di essa, essendo nota la dimensione di tale distanza[15]. La scala è caratterizzata da un'unità di misura comune e costante che assegna un numero reale a ciascuna misurazione. La sua origine è arbitraria. Il rapporto tra due intervalli qualsiasi è dipendente dall'unità di misura.
- *Proprietà*: sui numeri associati alle misurazioni con questo tipo di scala è possibile effettuare operazioni aritmetiche. Tale scala consente di determinare non solo le equivalenze o le relazioni $(>,<,\leq,\geq,=)$ ma anche il rapporto esistente tra due qualsiasi intervalli:

  $$(a-b)=(c-d) \qquad (a-b)<(c-d) \qquad (a-b)>(c-d)$$

  Date queste proprietà è corretto calcolare la differenza tra due valori della scala (e sulle differenze calcolare i rapporti) ma non è corretto determinare il rapporto tra due valori[16].
- *Operazioni*: ogni cambiamento nei numeri associati con le posizioni delle unità misurate deve preservare non solo l'ordine ma anche le differenze relative tra gli oggetti. Se ciascuna misurazione effettuata su una scala viene moltiplicata per una costante positiva le informazioni non vengono alterate.

Scala a rapporti

Rispetto alla scala ad intervalli, la scala a rapporti è possibile stabilire un'origine non arbitraria (punto zero naturale). In particolare si parla di scala a rapporti quando si conoscono

1. l'ordine dei valori rispetto ad un attributo;
2. gli intervalli tra i valori;
3. la distanza dei valori registrati da uno zero razionale.

- *Definizione*: la scala ha tutte le caratteristiche di quella ad intervalli con in più la possibilità di

[14] Un esempio di scala considerata ad intervalli in modo scorretto, e che consente di fare alcune considerazioni, è quello dei voti scolastici dati su scala da 1 a 10. A seconda degli insegnanti si possono avere scale a quattro o cinque gradini (dal quattro all'otto - dal cinque all'otto); questi gradini vengono "allungati" con gradini intermedi (5+, 6½, 7-, 6/7) senza alcun significato oggettivo. Si comprende quindi la scorrettezza del trattamento quantitativo si fa di tali valori.

[15] Gli intervalli si intendono *uguali* ma possono essere anche definiti da funzioni matematiche, come nel caso della scala *a intervalli logaritmici*, nella quale i valori dei punti successivi (*a, b, c,* ecc.) sono definiti da rapporti successivi di grandezza corrispondenti (*a/b = b/c = c/d* ecc.), ovvero

$$\log_n(a)-\log_n(b)=\log_n(b)-\log_n(c)=\log_n(c)-\log_n(d)$$

[16] Non si può dire che 20°C rappresentano il doppio di 10°C, ma si può dire che "la differenza tra 30°C e 10°C rappresenta il doppio della differenza tra 30°C e 20°C".

stabilire un vero punto d'origine (zero); in questo senso la scala a rapporti rappresenta un caso particolare della scala a intervalli, in cui le distanze sono stabilite rispetto ad uno zero razionale piuttosto che rispetto a un valore adottato come unità di riferimento. Tale caratteristica consente di misurare la distanza e calcolare il rapporto tra due valori. Il rapporto tra due valori è indipendente dall'unità di misura.

- *Proprietà*: è possibile calcolare rapporti di equivalenza e di relazione. Al contrario delle scale a intervalli, per le scale a rapporti è corretto calcolare il rapporto tra due intervalli o tra due valori misurati su tale scala[17].

$$(a/b)=(c/d) \qquad (a/b)<(c/d) \qquad (a/b)>(c/d)$$

La presenza o l'assenza di un'origine non è comunque rilevante ai fini dell'applicazione delle procedure di analisi statistica. Sono infatti possibili le seguenti relazioni:

- *Operazioni*: i numeri utilizzati rappresentano veri numeri con uno zero reale, solamente l'unità di misura è arbitraria. Si può operare una trasformazione (per esempio moltiplicando i valori di scala per una costante) senza alterare le informazioni.

Occorre a questo punto dire che la differenza tra i due tipi di scala non ha grandi effetti rispetto all'analisi dei dati in quanto la maggior parte delle diverse tecniche di analisi statistica non tengono in nessun conto l'origine della scala.

Con il seguente schema cerchiamo di sintetizzare le proprietà dei livelli di misurazione visti:

		LIVELLO DI MISURAZIONE (SCALA)			
		Nominale	Ordinale	Ad intervalli	A rapporti
PROPRIETA' DELLE SCALE	Classificazione	Sì	Sì	Sì	Sì
	Ordinamento gerarchico	No	Sì	Sì	Sì
	Misurazione aritmetica delle distanze	No	No	Sì	Sì
	Misurazione proporzionale delle distanze	No	No	No	Sì
TIPO DI RELAZIONE	Equivalenza	Sì	Sì	Sì	Sì
	Relazione (maggiore di)	No	Sì	Sì	Sì
	Rapporto tra intervalli	No	No	Sì	Sì
	Rapporto tra due valori	No	No	No	Sì

I quattro livelli di misurazione appena discussi non esauriscono le categorie dei livelli di misurazione. E' possibile costruire una scala nominale che fornisca una parziale informazione sull'ordine (scala parzialmente ordinata). Allo stesso modo una scala ordinale può riferire una parziale informazione sulle distanze (scala metrica ordinata).
Come si è accennato, uno dei problemi più controversi nella ricerca sociale è rappresentato dalla possibilità di individuare caratteristiche misurabili *quantitativamente*. La diffusione dell'impiego di scale arbitrariamente definite ad intervalli, e quindi la possibilità di disporre di dati quantitativi, anche in ambito sociale è spiegabile infatti essenzialmente nella autorizzazione che ciò dà alla utilizzazione delle tecniche statistiche più sofisticate. La messa a punto di tecniche per la costruzione di unità di misura che rispondano correttamente a criteri quantitativi e che siano utilizzabili anche in ambito sociale è stata oggetto di molti studi soprattutto in ambito psicologico.

3.2.2.4 Livelli di misurazione espressi come funzioni matematiche

I livelli di misurazione possono essere visti come rappresentazioni (Bruschi, 1999; Velleman, 1993) espresse in termini di funzioni che consentono di trasformare le osservazioni empiriche in valori

[17] Si può dire che 10 metri rappresentano una misura doppia di 5 metri.

numerici.
Indichiamo con:

S un insieme di oggetti che variano rispetto a qualche attributo,
S_i un oggetto dell'insieme *S*,
M un insieme di numeri reali,
f la funzione di trasformazione, la sua natura determina il livello di misurazione,
$M(S_i)$ il numero assegnato a S_i secondo la relazione *f* che lega *M* e *S*.

- Livello nominale: identità tra categorie (a tutti gli oggetti all'interno di una data categoria viene assegnato uno stesso simbolo numerico):
  $f_n \Rightarrow (S_1 = S_2) \Rightarrow M(S_1) = M(S_2)$
  $f_n \Rightarrow (S_1 \neq S_2) \Rightarrow M(S_1) \neq M(S_2)$
  Notare che il simbolo "=" posto nelle equazioni non indica una uguaglianza matematica ma solo una corrispondenza tra simboli.
- Livello ordinale:
  $f_o \Rightarrow (S_1 = S_2) \Rightarrow M(S_1) = M(S_2)$
  $f_o \Rightarrow (S_1 < S_2) \Rightarrow M(S_1) \leq M(S_2)$
  $f_o \Rightarrow (S_1 > S_2) \Rightarrow M(S_1) \geq M(S_2)$
  La funzione preserva l'asimmetria empirica tra le categorie osservative; ciò vale soprattutto riguardo ai simboli "<" e ">" che non indicano una relazione quantitativa ma una costruzione del ricercatore.
- Livello a intervalli e a rapporti: la funzione rappresenta una funzione numerica specifica che lega *S* e *M*; il caso più semplice è quello che specifica una funzione lineare dalle osservazioni in valori numerici[18]:
  $f_i \Rightarrow M(S_1) = a + b(S_1)$
  dove *a* e *b* rappresentano coefficiente realmente e numericamente valutati.
  Se a=0 allora:
  $f_r \Rightarrow M(S_1) = b(S_1)$
  che si riferisce alla misurazione a rapporti la cui posizione dell'origine è significativa (ovvero è uguale a zero).

Dal punto di vista formale, ciò vuol dire che qualsiasi serie di numeri che soddisfa *f* fornisce ugualmente una buona misurazione dell'attributo considerato; in altre parole, non esiste una singola misurazione corretta per un determinato attributo: è sempre possibile trovare un'altro insieme di numeri/simboli che rappresentano ugualmente bene gli oggetti[19]. Anche se la natura di *f* diviene sempre più restrittiva via via che si passa dal livello nominale, al livello ordinale, al livello ad intervalli e a quello a rapporti, per qualsiasi variabile, indipendentemente dal livello di misurazione, esiste un numero infinito di misure possibili.
Il problema è chiarire il criterio attraverso il quale il ricercatore può scegliere tra le diverse misure *M* utilizzabili per una certa variabile. Uno dei criteri utilizzabili è "per convenzione"; nell'adottare tale criterio occorre però tener presente che è possibile assegnare agli stessi oggetti altri numeri/simboli per rilevare la stessa informazione.

[18] Con la misurazione ad intervalli o a rapporti è possibile utilizzare qualsiasi funzione specifica per passare dagli oggetti ai numeri; il polinomio $M(S_1) = a + (S_1)^b$ è perfettamente accettabile come variabile di livello a intervalli; la funzione lineare rimane tuttavia la forma adottata più semplice.

[19] E' importante sottolineare il fatto che il concetto di *livello* di misurazione è distinto e differente da quello di *accuratezza* di misurazione. Come si è visto, tutti gli strumenti di misurazione presentano limiti nella loro precisione. Ciò vuol dire che la trasformazione da oggetto a misura non è mai perfetta. Le imperfezioni che esistono nella funzione costituiscono l'errore di misurazione. Ovviamente la presenza dell'errore di misurazione non compromette l'esistenza di un particolare livello di misurazione.

3.2.2.5 Importanza del livello di misurazione

In genere si afferma che l'importanza di poter distinguere tra diversi livelli di misurazione è data dal fatto che per ciascun livello sono appropriate solo determinate operazioni matematiche.
In realtà tale argomento può essere più convenientemente visto in un'altra prospettiva. I ricercatori, cercando di spiegare le differenze tra gli oggetti empirici, sono più interessati alla variabilità nelle misure piuttosto che ai valori numerici delle misure. Ma la variabilità osservata può emergere da almeno due fonti:

a. dall'attributo che si cerca di misurare,
b. dai modi in cui i numeri/simboli sono assegnati agli oggetti, ovvero la natura della funzione che converte le osservazioni in numeri/simboli.

Come abbiamo già visto i livelli di misurazione differiscono tra loro rispetto alla seconda fonte di variabilità; quindi. Ciò vale
Quanto più la misurazione è precisa, tanto più la varianza osservata può essere attribuita alle differenze tra gli oggetti piuttosto che al sistema di misurazione.

4. LA CONDENSAZIONE DEGLI INDICATORI ELEMENTARI: VERIFICA DEL MODELLO DI SCALING

4.1 LE CONDIZIONI PER LA VERIFICA DEL MODELLO

La verifica del modello prevede la definizione di condizioni e di procedure che riguardano:
- la logica utilizzata per la verifica,
- i criteri per la valutazione dell'adattamento del modello,
- la procedura di raccolta dei dati e la definizione il campione.

4.1.1 Logica di verifica

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 ovvero osservando se le proprietà dello *scaling* specificate dal modello sono osservabili nei dati. In particolare:

dati - modello = residuo

Lo scarto tra dati e modello è attribuito alla presenza dell'errore; tale residuo viene ritenuto troppo elevato quando esso non può essere attribuito al caso ma rappresenta una reale divergenza tra valori attesi (modello) e osservati (dati). 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. Come sappiamo, il residuo può essere dovuto principalmente a due componenti:
- errore casuale o *disturbo* (*e*),
- errore sistematico o *bias* (*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 variabili).

4.1.2 Criteri per la 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. Nella definizione dei criteri per la verifica del modello di *scaling* occorre anche precisare come interpretare il *mancato adattamento del modello ai dati*; la discrepanza tra modello di scaling e dati definisce gli *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);
b. essi possono rappresentare semplicemente delle fluttuazioni che si verificano perché le singole

osservazioni sono affette da errori di misurazione, per la presenza di osservazioni che possono influenzare in maniera negativa il livello di adattamento (*outlier*), per la presenza dell'errore di campionamento o per altri fattori stocastici; in questo caso è possibile per esempio verificare la tecnica di *scaling* utilizzata in modo che produca un insieme di valori che meglio rappresenti la variabilità tra gli oggetti misurati[1].

Il problema del ricercatore è quello di comprendere 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*. In particolare è necessario definire il livello di tollerabilità, ovvero fino a che punto la discrepanza osservata è tollerata e quando invece induce a rigettare il modello.

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 di ciascuna particolare procedura.

4.1.3 Procedura di raccolta dei dati e definizione del campione

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.

Definizione del campione

In teoria la sperimentazione per la messa a punto di uno strumento dovrebbe essere condotta e riguardare l'universo (di casi e/o di indicatori). Per superare gli ostacoli che impediscono la rilevazione totale, 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[2].
- dimensione: è importante che la verifica del modello di *scaling* avvenga su campioni composti da un numero piuttosto alto di casi/oggetti in modo da minimizzare la fonte di errore dovuta all'errore di campionamento, sappiamo infatti che maggiore è la dimensione del campione, più precisa è la stima[3].

[1] 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*.

[2] Non sempre nella ricerca sociale tale requisito viene soddisfatto. La verifica dei modelli (il più delle volte per mancanza di sufficienti risorse) viene svolta su campioni di convenienza, costituiti dai soggetti più facilmente reperibili. Tale approccio però non consente di calibrare lo strumento rispetto ad una particolare popolazione.

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

Quindi, in pratica, le dimensioni di campionamento sono due ed è 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.

4.2 LA DIMENSIONALITA'

4.2.1 Interpretazione del concetto di dimensionalità

Di per sé le dimensioni non hanno una realtà sostanziale; la loro individuazione consente di semplificare le caratteristiche di un insieme di casi, per renderle più comprensibili a fini analitici; in questo senso una rappresentazione dimensionale di un insieme di casi è sempre un modello interpretativo dei casi e non una proprietà immutabile dei casi stessi (Netemeyer, 2003).

Variazione tra casi

Il concetto di dimensionalità può essere definito e chiarito facendo riferimento al concetto di "fonti di variazione"; in altre parole la dimensionalità di un gruppo di casi può essere definita come il *numero di fonti, separate e apprezzabili, di variazione che spiegano le differenze tra i casi*; ciò richiede alcune puntualizzazioni:

a. la dimensionalità di un insieme di casi è legata alle reali caratteristiche dei casi stessi; a volte la natura di tali caratteristiche è immediatamente visibile, a volte lo è meno;
b. anche se in molte situazioni le fonti di variabilità tra i casi sono note, la specificazione della dimensionalità spetta al ricercatore che deve specificare la struttura dimensionale (in termini di definizione e di numero delle dimensioni) prima del procedimento di misurazione; ciò succede ogni volta che i casi sono confrontati sulla base di criteri noti;
c. nell'ipotizzare il numero di fonti di variazione, se per ciascun punto si utilizza un numero di dimensioni
    - troppo piccolo, allora parte della variabilità tra i casi non può essere rappresentata ovvero incorporata nel modello,
    - troppo grande, allora il modello risulta ridondante;
d. la dimensionalità è di solito specifica di un contesto; il numero di dimensioni, assunte tra i casi, dipende dai modi in cui i casi sono esaminati; rispetto agli stessi casi un ricercatore può, per alcuni obiettivi, considerarli posizionati su una dimensione, per altri può ritenere più appropriata un posizionamento in uno spazio multidimensionale; ciò può essere fatto fino a quando si ritiene che la dimensionalità selezionata corrisponda a tutte le fonti rilevanti di variazione tra i casi; è però sempre il ricercatore che decide ciò che è rilevante.

Modello spaziale

La differenziazione tra i casi rispetto a ciascuna fonte di variabilità può essere rappresentata geometricamente individuando uno spazio definito da un numero d'assi corrispondente al numero di

dimensioni rilevate. I casi misurati sono rappresentati da punti la cui posizione in tale spazio è determinata dalle loro posizioni lungo ciascuno degli assi (coordinate). Se i casi variano solo rispetto ad una fonte di variazione, la rappresentazione richiede un unico asse e un'unica coordinata per ciascun caso; se i casi differiscono rispetto a due fonti di variazione, la rappresentazione richiede due assi e due coordinate per ciascun caso; e così via. In questo senso la dimensionalità di un gruppo di casi fa riferimento al numero minimo di coordinate richieste per posizionare in modo unico l'insieme dei punti che rappresentano i casi.
La rappresentazione geometrica risulta conveniente in quanto fornisce un semplice strumento per la visualizzazione delle differenze tra i casi. Tale rappresentazione diviene però problematica se le rette identificate definiscono uno spazio con un numero di dimensioni superiore a tre. In questi casi è comunque possibile costruire un modello geometrico che mostri parte dello spazio multidimensionale; tale procedura, però, fornirà una rappresentazione incompleta dei casi. A tale proposito occorre però dire che la rappresentazione reale del modello geometrico non è necessaria: per il corretto posizionamento reciproco dei casi nello spazio individuato sarà sufficiente la definizione delle coordinate per ciascun punto; in questo senso la raffigurazione fisica delle posizioni dei punti rappresenta un'informazione ridondante.
E' per questo necessario distinguere tra dimensioni fisiche e loro esistenza concettuale. Non esiste alcuna ragione per limitare la definizione del numero delle fonti di variabilità a tre solo per l'impossibilità di rappresentare fisicamente più di tre dimensioni. Vincolarsi alla rappresentazione geometrica delle dimensioni limita la comprensione della dimensionalità e quindi della variabilità[4].

4.2.2 Analisi e verifica della dimensionalità

Come abbiamo visto, la dimensionalità viene ipotizzata al momento della definizione del modello di *scaling*. Dopo la raccolta dei dati e prima della condensazione è molto importante procedere alla sua verifica che può far emergere anche una diversa dimensionalità. Per fare ciò si ricorre a modelli analitici che consentono di verificare qualsiasi livello di dimensionalità.
La verifica della dimensionalità consente di comprendere in quale modo effettuare la condensazione (Carmines, 1992). Nel caso in cui la variabile sia

- *unidimensionale*, la condensazione degli indicatori avviene in modo tale che a ciascun caso venga assegnato un valore che corrisponde geometricamente alla sua collocazione su una retta corrispondente all'attributo (spazio unidimensionale); ciò consente di valutare la relazione tra i punti/casi e tra questi e l'origine (quando presente e significativa)[5];
- *multidimensionale*, la condensazione degli indicatori avviene in modo tale che a ciascun caso venga assegnato un numero di valori corrispondente alla dimensionalità della caratteristica; tale procedimento geometricamente equivale alla collocazione dei casi misurati in uno spazio multidimensionale, in cui ciascuna dimensione corrisponde ad un aspetto; ciascun valore assegnato al caso corrisponde alla proiezione del punto su uno degli assi (dimensioni) dello spazio.

Nel caso in cui il modello di *scaling* abbia una ipotesi dimensionale debole è possibile procedere secondo una strategia esplorativa al fine di scoprire e identificare il numero delle dimensioni latenti necessarie a descrivere un insieme di osservazioni ovvero identificare la struttura dei dati. Tra tali strategie vi è quella detta *progressiva*. In questi casi l'analisi comincia verificando la

[4] A tale proposito può essere interessante, oltre che piacevole, leggere *Flatlandia* di Abbott; in tale divertente libro si mette in evidenza come l'incapacità di comprendere pienamente le basi della geometria limita la comprensione della varietà delle forme dimensionali.

[5] Ricordiamo che non sempre si può parlare di misurazione lungo un continuum come nel caso delle variabili discrete.

rappresentazione geometrica più semplice, ovvero quella unidimensionale. Se i dati risultano coerenti con il modello (ovvero la discrepanza tra modello e dati 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 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.
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*).
Tale strategia si presta però ad alcune critiche in quanto qualsiasi modello di *scaling* ipotizza una struttura dimensionale. A tale proposito sappiamo 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.
Un'altra strategia è quella che considera la multidimensionalità come una *rappresentazione multipla di unidimensionalità*. Per ciascuna dimensione ipotizzata si verifica un modello unidimensionale. 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.
Occorre a questo punto aggiungere e ricordare che l'insieme delle dimensioni così individuate attraverso strategie esplorative rappresenta semplicemente un sistema di coordinate che può essere utilizzato per posizionare un insieme di punti; in altre parole, il significato da attribuire a ciascuna dimensione non rappresentato dal risultato delle verifica ma fa parte dell'interpretazione che il ricercatore dà del risultato ottenuto. Il rischio è che il ricercatore si possa sentire forzato ad attribuire un significato ad una dimensione semplicemente perché la dimensione esiste, dando origine a punteggi che però non necessariamente risultano interpretabili.
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.

Il principale approccio analitico per la verifica del modello dimensionale è l'approccio fattoriale (presentato di seguito) che, più propriamente, ha l'obiettivo di analizzare e verificare l'esistenza di **strutture latenti** (Maggino, 2005). In alcuni casi è possibile anche usare, per tale verifica, la *cluster analysis* (Maggino, 2005). In alcuni casi la verifica della dimensionalità fa parte della verifica del modello di *scaling* (come nel *multidimensional scaling*) (Nunnally, 1978).

4.2.2.1 Modello fattoriale

La definizione dei modelli fattoriali è avvenuta nell'ambito della psicometria e della psicologia sperimentale. Alla base di tale modello vi era l'ipotesi secondo la quale esistono concetti ipotetici, quali l'intelligenza, la qualità della vita, ecc., non osservabili e misurabili direttamente, e che quindi rappresentano *fattori* o *dimensioni latenti*, misurati attraverso una o più variabili rilevate a loro volta tramite misure multiple. L'applicazione del modello, riducendo la complessità, consente di chiarire e verificare definizioni teoriche costituendo in questo senso uno strumento di verifica del significato di una o più variabili (valore euristico)[6] (Kim, 1989a, 1989b; Marradi, 1981).

[6] Spearman, uno dei primi a formulare un modello di misurazione multifattoriale all'inizio del '900, ha definito e

Tale approccio ha trovato, in seguito, nuovi sviluppi sia in ambito sociologico (con i lavori di Lazarsfeld e Rosemberg negli anni 50) che in ambito psicometrico (con i lavori di Heise e Bohrnstedt negli anni 70). Tale sviluppo ha condotto alla definizione di modelli che considerano contemporaneamente più variabili di interesse teorico non direttamente osservabili (*variabili latenti/fattori*) ciascuna delle quali misurata da più variabili osservate (*indicatori*). Ciascuna variabile latente può rappresentare anche un aspetto (dimensione) di un concetto più ampio.
L'obiettivo è quello di "registrare" gli effetti che le variabili latenti (variabili) hanno sugli indicatori che le misurano o, meglio, di stimare la relazione tra variabili latenti e tra queste e le variabili osservate (*relazioni strutturali*). Statisticamente ciò può essere studiato analizzando la covariazione esistente tra gli indicatori definiti per ciascuna variabile latente. La stima dei parametri strutturali richiede l'analisi della covariazione tra gli indicatori osservati e, conseguentemente, l'applicazione dei modelli ad equazioni strutturali applicati alle covarianze osservate; in altre parole, gli indicatori consentono di risolvere specifiche incognite presenti nelle equazioni simultanee che descrivono il modello. Quindi la semplificazione realizzata da tali modelli non coinvolge l'eliminazione di una delle dimensioni ma è rivolta alla stima dei parametri strutturali che legano le variabili latenti;
In generale, il modello ad indicatori multipli risulta di complessa e problematica applicazione in quanto:

- richiede per la sua analisi un'esplicita definizione di una relazione causa-effetto tra variabili misurate e non misurate; tale definizione non è sempre facile da stabilire soprattutto quando l'indicatore è solo 'un aspetto di', 'una parte di', 'correlato con';
- all'aumentare del numero degli indicatori misurati e delle variabili strutturali, diviene estremamente difficile, se non impossibile, prendere in considerazione tutte le logiche combinazioni di relazioni;
- se il modello è sovrastimato, diviene difficile capire in che modo procedere nella selezione di una stima tra quelle possibili.

Assunti

L'additività della varianza

Uno dei principali assunti sui quali si basa il modello fattoriale riguarda il concetto di additività della varianza; secondo questo concetto la varianza totale di ciascun indicatore x_i rappresenta la somma di tre componenti tra loro non correlate:

- varianza comune, che rappresenta quella porzione di varianza che è spiegata dalla presenza della variabile latente[7] (ξ) ed è misurata dalla correlazione che l'indicatore registra con altri indicatori della stessa variabile latente;
- varianza specifica, che rappresenta quella porzione di varianza non spiegata dalla presenza della variabile latente; essa non correla con nessun altro indicatore; insieme alla precedente componente va a comporre la *varianza attendibile*
- errore, che rappresenta quella porzione di varianza, non correlata con le precedenti; tale componente definisce la *varianza* detta *non attendibile*.

Essendo ciascun tipo di varianza espresso come porzione della varianza totale possiamo scrivere, per l'indicatore x_i :

postulato i concetti di "fattore generale" e "fattore specifico" per poter misurare l'intelligenza; in particolare egli ha teorizzato un modello che mirava a descrivere l'intelligenza di un individuo mediante il minor numero possibile di caratteri risultati maggiormente significativi; secondo tale teoria l'attività mentale può essere considerata come l'effetto dell'azione di un fattore generale, rappresentabile lungo un continuum lineare, e di un insieme di altri fattori particolari (memoria, conoscenza, ecc.).

[7] Qui viene utilizzata la notazione tipica del Lisrel (v. appendice a questo capitolo).

varianza totale	=	varianza comune	+	varianza specifica	+	errore
		varianza attendibile				
$\sigma^2_{x_i}$	=	$\sigma^2_{x_{ic}}$	+	$\sigma^2_{x_{is}}$	+	$\sigma^2_{x_{ie}}$

Inoltre, sapendo che, con dati standardizzati, la *varianza totale* è uguale a 1, possiamo scrivere:

varianza totale	=	varianza comune	+	varianza specifica	+	errore
		varianza attendibile				
1	=	$\sigma^2_{x_{ic}}$	+	$\sigma^2_{x_{is}}$	+	$\sigma^2_{x_{ie}}$

Nel modello fattoriale, e in termini di modello di misurazione, l'interesse è rivolto essenzialmente alla stima della varianza comune e non della varianza specifica che per questo viene considerata unitariamente all'errore in quella che viene detta varianza unica (o *unicità*, δ^2):

$$\delta^2_{x_i} = \sigma^2_{x_{is}} + \sigma^2_{x_{ie}}$$

I fattori comuni

La varianza comune di un indicatore solo raramente può spiegata da un'unica variabile latente; questo vuol dire che ciascun indicatore può essere descritto attraverso la combinazione di variabili latenti (dette *fattori comuni*):

indicatore = combinazione lineare di fattori comuni + errore

Riprendendo tale assunto in termini di varianza è possibile affermare che la varianza comune rappresenta la porzione di varianza che è spiegata dalla presenza di più variabili latenti (ξ_j) ed è per questo detta *comunanza* ($h^2_{x_i}$); a questo punto possiamo rappresentare la varianza totale nel modo seguente:

varianza totale	=	comunanza	+	varianza specifica	+	errore
		varianza attendibile				
$\sigma^2_{x_i}$	=	$h^2_{x_i}$	+	$\sigma^2_{x_{is}}$	+	$\sigma^2_{x_{ie}}$
1	=	$h^2_{x_i}$	+	$\sigma^2_{x_{is}}$	+	$\sigma^2_{x_{ie}}$

L'obiettivo del modello fattoriale è non solo quello di stimare la *comunanza* di ciascun indicatore ma anche di stimare quanto della comunanza è attribuibile alle diverse variabili latenti comuni al gruppo di indicatori. Ciò corrisponde alla stima per ciascun indicatore dei *factor loading* che lo legano ai fattori ($\lambda_{x_i\xi_j}$, dell'indicatore x_i rispetto alla variabile latente ξ_j). In particolare ciascun *factor loading* rappresenta la valutazione del contributo di ciascun fattore alla comunanza dell'indicatore, ovvero la valutazione dell'influenza della variabile latente su ciascun indicatore; esso può essere definito anche come peso che l'indicatore ha nel definire il fattore o come la misura della *saturazione* di ciascun indicatore rispetto al fattore. I *factor loading* vengono espressi con valori che vanno da

+1 massima saturazione di un indicatore in un determinato fattore, a

-1 massima saturazione di un indicatore in un determinato fattore ma in senso inverso.

Un valore di saturazione uguale a 0 indica che quel determinato indicatore non ha alcuna rilevanza rispetto al fattore. In pratica, i *factor loading* sono interpretati in termini di correlazione tra ciascun

indicatore ed il fattore. Sapendo che il quadrato della correlazione (coefficiente di determinazione) indica la porzione di varianza spiegata, il quadrato del *factor loading* indica la quantità della variabilità di un particolare indicatore spiegata dal corrispondente fattore.
Conseguentemente si può dire che la *comunanza* è data dalla somma dei *loading* al quadrato di un indicatore:

$$h_{x_i}^2 = \lambda_{x_i\xi_1}^2 + \lambda_{x_i\xi_2}^2 + \ldots + \lambda_{x_i\xi_m}^2$$

dove *m* rappresenta il numero di variabili latenti.

Di seguito proviamo a rappresentare l'assunto additivo della varianza appena visto:

varianza totale			
$\left(\sigma_{x_i}^2\right)$			
varianza comune		*varianza specifica*	*errore*
$\sigma_{x_{ic}}^2$		$\sigma_{x_{is}}^2$	$\sigma_{x_{ie}}^2$
communality (comunanza)		*unicità*	
$h_{x_i}^2$		$\delta_{x_i}^2$	
$\lambda_{x_i\xi_1}^2$	$\lambda_{x_i\xi_2}^2$ ..... $\lambda_{x_i\xi_m}^2$	$1 - h_{x_i}^2$	
varianza attendibile		*errore*	
$h_{x_i}^2 + \sigma_{x_{is}}^2$		$1 - \left(h_{x_i}^2 + \sigma_{x_{is}}^2\right)$	
$\sigma_{x_i}^2 = \sum_{j=1}^{m} \lambda_{x_i\xi_j}^2 + \delta_{x_i}^2$			
(*equazione fondamentale del modello dei fattori comuni*)			

Gli assunti alla base di tale modello possono essere a questo punto così riassunti:

a. le relazioni tra gli indicatori sono lineari;
b. la varianza totale negli indicatori è una funzione:
   - dei fattori (*comunanza*),
   - dei *disturbi* caratteristici di ciascun indicatore e degli errori di misurazione (*unicità*);
c. gli errori e i disturbi non sono correlati tra loro né con i fattori;
d. gli indicatori non sono casualmente correlati tra loro se non attraverso la reciproca relazione con i tratti latenti;
e. la struttura dei fattori osservati riproduce fedelmente la struttura delle dimensioni sottostanti.

Il modello esplorativo e il modello confermativo

A seconda dei vincoli che si pongono nella definizione del modello, è possibile individuare due approcci: il modello esplorativo e il modello confermativo.
Si parla di modello esplorativo quando nella definizione del modello si specifica il numero dei fattori comuni e si individuano gli indicatori ma non si specifica la struttura delle relazioni tra le variabili latenti e gli indicatori. Tale modello assume in particolare che:

- tutti i fattori comuni sono (o non sono) tra loro correlati,
- tutte le variabili osservate sono direttamente influenzate da tutti i fattori comuni,
- le unicità non sono tra loro correlate,
- le variabili latenti (fattori comuni) non sono correlate con i le unicità (fattori unici).

La stima dei parametri del modello richiede la definizione di altri assunti, generalmente arbitrari.

Tali caratteristiche rendono il modello esplorativo limitato nelle sue reali applicazioni a livello di modello di misurazione.
Si parla di modello confermativo quando la definizione del modello pone dei vincoli, sostanzialmente motivati, relativi al numero delle variabili latenti e degli indicatori e alle relazioni che legano fattori e variabili osservate. Le ipotesi sulla struttura possono derivare da particolari impostazioni teoriche, da precedenti ricerche, dalle caratteristiche del disegno sperimentale o dopo aver esaminato la matrice di correlazione. Tale approccio consente la verifica statistica di una struttura fattoriale ipotizzata al fine di determinare se il livello di adattamento del modello ai dati è statisticamente significativo, ovvero se i dati confermano il modello definito (Bartolomew, 1999; Bohrnstedt, 1994; Lazarsfeld, 1968; Long, 1993; Netemeyer, 2003).
Nella forma più semplice l'ipotesi comporta la specificazione del solo numero di fattori comuni, in modo non molto diverso da quello che viene fatto nell'approccio esplorativo, senza alcun riferimento alla relazione esistente (ortogonale o obliqua); in questi casi l'applicazione di un test di significatività o di altri criteri (come un coefficiente di affidabilità) è sufficiente per valutare l'adeguatezza della soluzione fattoriale.[8]
La verifica può però riguardare anche ipotesi più complesse che specificano a priori i vincoli che possono riguardare:

- il numero di fattori,
- la struttura fattoriale (quali indicatori definiscono ciascuno dei fattori e quindi quali variabili osservate sono influenzate da quali fattori comuni),
- la natura delle relazioni tra i fattori (ortogonali o obliqui),
- la dimensione dei valori dei *factor loading*,
- la presenza di correlazione tra determinati termini di errore.

La definizione di tali vincoli consiste nel decidere quali sono i parametri determinati (fissi) e quali quelli da determinare (liberi). Ciò porta alla distinzione tra:

- valori incogniti da risolvere, rappresentati dai parametri da stimare;
- valori noti, rappresentati da quelli che possono essere calcolati direttamente dai dati (medie, varianze, covarianze, correlazioni).

Successivamente alla definizione del modello occorre verificare se vi sono abbastanza quantità note per poter risolvere i parametri ignoti ovvero se il modello ha una soluzione per i valori incogniti dei parametri. L'*identificazione del modello* consiste appunto nel verificare la corrispondenza tra l'informazione che deve essere stimata (parametri liberi) e l'informazione utilizzata per tale stima (varianze e covarianze osservate relativamente alle ipotesi riguardanti la struttura del modello.
La differenza tra il numero di correlazioni/covarianze osservate e il numero di coefficienti proposti nel modello rappresenta i *gradi di libertà* (*gdl*), calcolati nel modo seguente:

$$gdl = \frac{(p)\cdot(p+1)-t}{2}$$

dove
p numero di indicatori
t numero di parametri liberi.

La prima parte dell'equazione calcola la dimensione non-ridondante della matrice di correlazione/covarianza (metà della matrice più la diagonale).
In questo caso, contrariamente ad altre definizioni di gradi di libertà utilizzate in altri ambiti dell'analisi statistica, non si utilizza il valore corrispondente alla dimensione del campione.
Tenendo presente il concetto di gradi di libertà, un modello (o sistema di equazioni) può essere *identificato*, *sottoidentificato*, *sovraidentificato*.

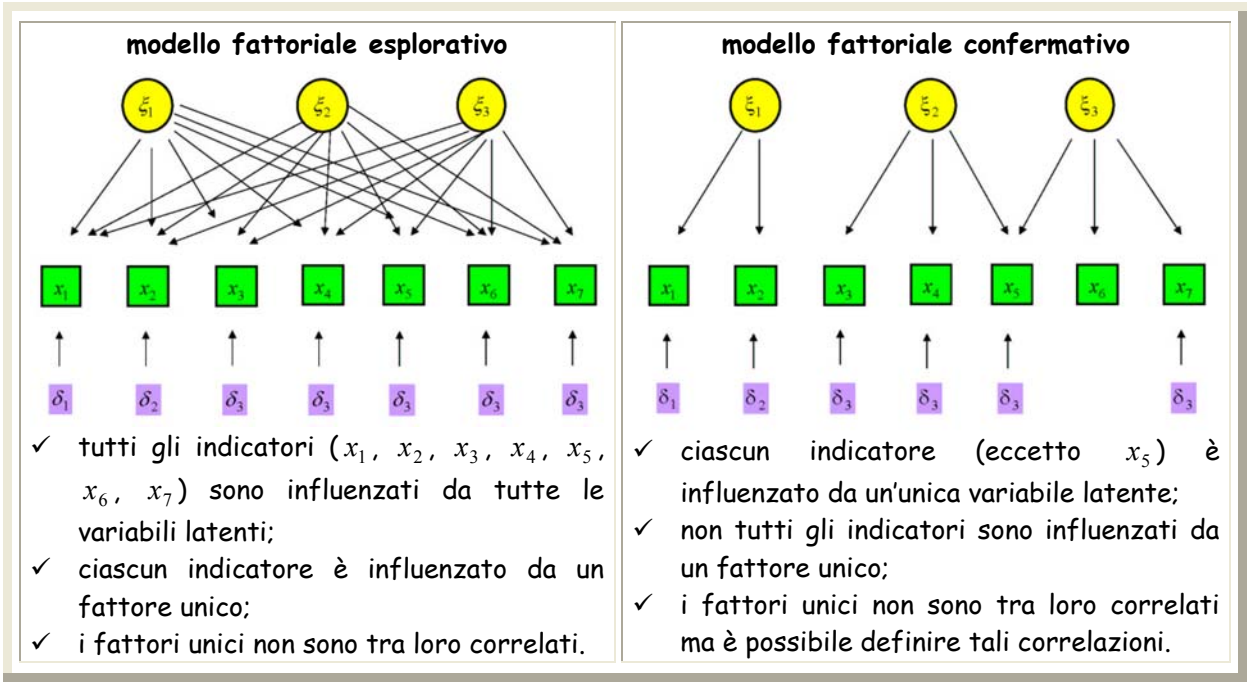
- Sistema identificato (zero gradi di libertà): è quello che contiene sufficienti informazioni per derivare una singola stima per ciascuna incognita; in altre parole, il numero delle quantità osservate è uguale al numero dei parametri incogniti. In questi casi per ciascun parametro libero

[8] Per una trattazione tecnica dell'approccio fattoriale sia esplorativo che confermativo si rimanda ad altre pubblicazioni.

è possibile ottener un valore attraverso una sola manipolazione (*just identified model*). Anche se fornisce un perfetto adattamento del modello, tale sistema non è interessante in quanto non è generalizzabile.

- Sistema sovraidentificato (numero di gradi di libertà positivo): è quello che presenta più informazioni di quelle necessarie per risolvere le incognite; in altre parole, il numero delle quantità osservate è superiore a quello dei parametri incogniti. In questi casi per ciascun parametro libero è possibile ottenere un valore corrispondente in molti modi (*overidentified model*). Tali modelli sono detti *confermativi* in quanto la loro applicazione è possibile quando è possibile definire un modello fattoriale.
- Sistema sottoidentificato (numero di gradi di libertà negativo) è quello che non presenta abbastanza informazioni per derivare stime non ambigue per le incognite. In altre parole, il numero di parametri incogniti è superiore a quello dei valori osservati. In un sistema *underidentified* non è possibile ottenere alcun valore per i parametri liberi, ovvero il modello non può essere stimato se non facendo nuovi assunti arbitrari. Tali modelli necessitano di un approccio fattoriale esplorativo al fine di comprendere meglio la struttura fattoriale sottostante e rendere il modello confermativo.

Con il seguente esempio proviamo a descrivere la differenza tra i due approcci. Poniamo di avere tre variabili latenti ξ_1, ξ_2, ξ_3 tra loro non correlate. Ciascuna di tali variabili influenza casualmente ciascuno degli indicatori x_1, x_2, x_3, x_4, x_5, x_6, x_7. Le tre variabili latenti vengono definite *fattori comuni* in quanto condividono l'influenza su uno o più indicatori. E' possibile quindi identificare altri fattori, detti *unici*, ciascuno dei quali influenza una sola variabile osservata (δ_1, δ_2, δ_3, δ_4, δ_5, δ_6, δ_7).



In entrambi gli esempi i fattori comuni (ξ_1, ξ_2, ξ_3) non sono tra loro correlati; nel caso confermativo sarebbe stato possibile anche assumere correlati due fattori comuni (per esempio ξ_1 e ξ_3), mentre nel caso esplorativo, l'alternativa sarebbe stata quella di definire tutti i fattori comuni correlati.

Nel procedere alla valutazione del modello fattoriale è necessario tenere presenti alcuni importanti elementi:

- *il livello di correlazione tra gli indicatori*: un alto livello di correlazione tra indicatori rappresenta un indice di buon livello di consistenza interna;
- *la dimensione del campione*: sono molti i suggerimenti che vengono dati per stabilire qualche criterio per definire la dimensione campionaria ottimale anche se nessuno di tali suggerimenti può assumere valore normativo generale;
- *la presenza di errori di misurazione correlati*: se gli indicatori condividono una varianza che va oltre quella spiegata dai fattori o l'esistenza di livelli non omogenei nelle correlazioni può essere spiegata con la presenza di errori di misurazione correlati, considerabili all'interno dell'approccio confermativo.

In generale la valutazione del modello utilizza gli approcci visti in precedenza ai quali si può aggiungere un importante metodo di verifica che coinvolge *factor loading*, varianza dei fattori e covarianza tra fattori, errore. In genere, tale verifica è fatta attraverso un approccio gerarchico finalizzato alla verifica, in successione, dell'adattamento del modello più restrittivo e dell'invarianza dei *factor loading* e delle covarianze tra campioni diversi.

Appendice.
IL LISREL: LA NOTAZIONE

Molti approcci di analisi statistica multivariata, visti e applicati in maniera autonoma, trovano molti elementi comuni di analisi tali da ricondurli ad un unico modello di analisi. In particolare, tale modello si basa su uno schema teorico prodotto dalla fusione di due modelli tradizionali di analisi che riguardavano discipline diverse:

- Psicometria, nell'ambito della quale è stato sviluppato l'approccio fattoriale[9], e sono stati definiti concetti, modelli e metodi di verifica della misurazione rispetto alla validità e alla affidabilità, mettendo a punto tecniche di messa a punto di strumenti di misurazione di dimensioni psicologiche attraverso scale unidimensionali e multidimensionali;
- Econometria, nell'ambito della quale sono stati sviluppati modelli riguardanti le relazioni causali tra le variabili in campo economico («modelli di equazioni simultanee») che presentavano il limite di non considerare il concetto di variabile latente e il problema dell'errore di misurazione, ritenendolo trascurabile in dati economici.
- Biometria, nell'ambito della quale sono stati sviluppati modelli di definizione di connessioni causali esistenti tra un gruppo di variabili; tali modelli cercano anche di quantificare l'impatto di ogni variabile su ognuna delle altre definite nel modello attraverso i *path coefficient* da cui è derivata la *path analysis*. Intorno agli anni '60 i lavori di Wright in biometria vengono proposti anche per le analisi sociologiche (Duncan) favorendo la diffusione della *path analysis* tra i sociologi.

All'inizio degli anni '70 tutte queste strade convergono nei lavori di Jöreskog che formula un modello generale riconducendo tutto ai modelli di equazioni strutturali.
La fusione, frutto di un seminario che ha visto lavorare congiuntamente un gruppo composto da econometristi, psicologi, sociologi e statistici, ha prodotto un approccio nel quale rientrano i modelli di misurazione, la *path-analysis*, l'analisi fattoriale, i modelli causali, la teoria delle equazioni strutturali e le stime di massima verosimiglianza. I modelli ad equazioni strutturali sono stati messi a punto da K.Jöreskog e inseriti in un programma chiamato *LISREL* che inizialmente il programma aveva sviluppato algoritmi per l'analisi fattoriale. Successivamente l'approccio presente nel programma ha subito un allargamento di applicazione andando oltre l'ambito dell'analisi fattoriale. Esso è quindi divenuto una procedura generale per i modelli basati su sistemi di equazioni strutturali, mantenendo la distinzione tra variabili latenti e variabili osservate.
Quindi *LISREL* da semplice termine per indicare un *software* è divenuto un termine utilizzato per intendere un approccio generale e una struttura di base nella quale posizionare metodi appartenenti a diversi approcci scientifici (analisi fattoriale, *path analysis*, analisi di strutture di covarianza, analisi di panel, ecc.) provenienti in genere dalla tradizione psicometrica, dalla tradizione econometrica, dai metodi della biologia e della sociologia.
Dall'approccio psicometrico è stato tratto il concetto di variabile latente[10], mentre dall'approccio

[9] Nata in ambiente psicometrico per risolvere il problema delle variabili latenti; l'analisi fattoriale cerca di scoprire se le correlazioni esistenti all'interno di un gruppo di variabili osservate consentono di spiegare un piccolo numero variabili latenti o fattori. I primi lavori in questo campo (K.Spearman) tentavano di definire e misurare l'intelligenza umana identificando una componente comune (fattore generale) e altre componenti uniche (riguardanti uno specifico tipo di misurazione adottato o gli errori di misurazione). Le successive applicazioni dell'analisi fattoriale ad altri campi non sempre sono stati sostenuti dagli statistici per la presenza di una certa arbitrarietà del metodo. Più recentemente sono stati prima teorizzati (Lawley, 1940) e poi realizzati in un algoritmo applicabile su computer (Jöreskog, 1967).

[10] A tale proposito è importante chiarire la distinzione tra

econometrico è stato tratto il modello di studio delle relazioni causali.
L'obiettivo dell'approccio *LISREL* è quello di rispondere a due questioni che direttamente riguardano il ricercatore sociale:

- la *misurazione*: nelle scienze sociali è molto difficile poter misurare le dimensioni che maggiormente interessano; ciò accade sia per la difficoltà di definire teoricamente i concetti teorici da misurare che per la difficoltà di definire adeguati strumenti di misurazione; il problema da risolvere è quindi quello del legame esistente tra variabili latenti e variabili osservate e della verifica della validità e affidabilità delle misure.
- la *causalità*: come abbiamo visto molta parte delle teorie scientifiche si fonda sulla definizione di modelli basati su relazioni causali. Ciò pone il ricercatore davanti al problema di dover disporre di metodi e strumenti per poter verificare l'esistenza dei legami.

Di fatti la definizione di un modello *LISREL* è composta di due parti:

- *modello di misurazione*, all'interno della quale si specifica come le variabili latenti sono misurate tramite le variabili osservate e serve per determinare la validità e l'affidabilità della misurazione;
- *modello strutturale*, all'interno del quale si specificano le relazioni causali tra le variabili latenti e serve per determinare gli effetti causali e ammontare della varianza non spiegata.

Per poter indicare in maniera univoca i diversi tipi di variabili e di parametri, il *LISREL* utilizza una particolare notazione che all'inizio può sembrare un po' complessa. Tale notazione attribuisce a ciascun tipo di variabile o parametro un simbolo, generalmente ripreso dall'alfabeto greco (di seguito riportato).

Alfabeto greco						
maiuscolo	minuscolo			maiuscolo	Minuscolo	
A	α	alfa		N	ν	nu
B	β	beta		Ξ	ξ	csi
Γ	γ	gamma		O	o	omicron
Δ	δ	delta		Π	π	pi
E	ε	epsilon		P	ρ	rho
Z	ζ	zeta		Σ	σ	sigma
H	η	eta		T	τ	tau
Θ	θ	theta		Y	υ	upsilon
I	ι	iota		Φ	ϕ	phi
K	κ	kappa		X	χ	chi
Λ	λ	lambda		Ψ	ψ	psi
M	μ	mu		Ω	ω	omega

A ciascuna lettera così utilizzata vengono attribuiti anche gli appositi indici per distinguere tra loro le diverse variabili o i diversi parametri che appartengono allo stesso tipo.
I simboli utilizzati possono essere raggruppati nelle seguenti categorie:

- variabili *osservate/misurabili* (età, reddito, ecc.) che, in quanto affette da *errori di misurazione*; sono legate alle variabili definite ma non coincidono con queste che quindi risultano di fatti non osservate;
- variabili *latenti*, intese come costrutti teorici (detti *fattori*) che per loro natura non sono direttamente misurabili (es. intelligenza, status sociale, ecc.); è possibile però definire delle variabili specifiche che risultano essere legate al più generale concetto teorico.

Elemento del modello LISREL			Descrizione	Notazione	
				Matrice	Elemento
Variabili	costrutto	Esogeno	Costrutto esogeno – variabile strutturale/latente esogena		ξ
		Endogeno	Costrutto endogeno – variabile strutturale/latente endogena		η
	indicatore	Esogeno	Indicatore di variabile strutturale esogena – variabile osservata		X
		Endogeno	Indicatore di variabile strutturale endogena – variabile osservata		Y
Errori stocastici			Errore stocastico associato alla variabile strutturale endogene (η)*		ζ
			Errore stocastico associato alla variabile osservata esogena (X)**		δ
			Errore stocastico associato alla variabile osservata endogena (Y)**		ε
Matrici	Modello strutturale	Beta	Relazioni tra costrutti endogeni(η)	B	β_{nn}
		Gamma	Relazioni tra costrutti esogeni (ξ) ed endogeni (η)	Γ	γ_{nm}
		Phi	Correlazioni tra costrutti esogeni (ξ)	Φ	ϕ_{mm}
		Psi	Correlazioni delle equazioni strutturali o costrutti endogeni (errori dei costrutti endogeni ζ)	Ψ	ψ_{nn}
	Modello di misurazione	Lambda-X	Parametro strutturale: corrispondenza di indicatori esogeni (relazione tra costrutto esogeno, ξ , e variabile osservata, X) – peso fattoriale	Λ_x	λ^x_{pm}
		Lambda-Y	Parametro strutturale: corrispondenza di indicatori endogeni (relazione tra costrutto endogeno, η , e variabile osservata, Y) – peso fattoriale	Λ_y	λ^y_{qn}
		Theta-delta	Varianze-covarianze tra errori di previsione degli indicatori di costrutti esogeni, δ	Θ_δ	θ^δ_{pp}
		Theta-epsilon	Varianze-covarianze tra errori di previsione degli indicatori di costrutti endogeni, ε	Θ_ε	θ^ε_{qq}
Equazioni	Modello strutturale		Relazioni tra costrutti esogeni ed endogeni	$\eta = \Gamma\xi + \beta\eta + \varsigma$	
	Modello di misurazione	esogeno	Specificazione degli indicatori per i costrutti esogeni	$X = \Lambda_x\xi + \delta$	
		endogeno	Specificazione degli indicatori per i costrutti endogeni	$Y = \Lambda_y\eta + \varepsilon$	
Indici delle matrici			Numero totale e indice dei costrutti esogeni	M	*m*
			Numero totale e indice dei costrutti endogeni	N	*n*
			Numero totale e indice degli indicatori di costrutti esogeni	P	*p*
			Numero totale e indice degli indicatori di costrutti endogeni	Q	*q*
* errori nell'equazione (aggregato di tutte le influenze su Y non esplicitate nel modello ** errori di misurazione nelle variabili *X* e *Y*					

Simbologia per la rappresentazione grafica del modello

Nella rappresentazione vengono inseriti gli elementi di base necessari per comprendere la struttura del modello:

- le variabili
- gli errori relativi alle variabili
- i legami esistenti tra le variabili (con frecce e con coefficienti di regressione o correlazione o covarianza).

La rappresentazione grafica dell'intero modello è costituita dalla ricostruzione grafica di più equazioni strutturali che hanno delle variabili in comune e deve seguire, secondo il *Lisrel*, una particolare notazione:

- Le variabili: quelle latenti sono rappresentate da un cerchio o un'ellisse, mentre le variabili osservate sono rappresentate da un quadrato o un rettangolo; gli errori stocastici sono rappresentati solo dalla lettera corrispondente.
- Il legame tra due variabili:
  - causale diretto, rappresentato da una freccia unidirezionale che si dirige in linea retta dalla variabile «causa» (indipendente) alla variabile «effetto» (dipendente),

- associazione (covariazione, correlazione), indicato da una freccia bidirezionale che collega le due variabili (spesso con un arco).

- La forza della relazione: indicata riportando il valore del coefficiente relativo
  - di regressione, se la freccia è unidirezionale (legame causale)
  - di correlazione o covariazione, se la freccia è bidirezionale.

  L'assenza del valore indica un valore del coefficiente uguale a 1 (come il caso dei coefficienti tra errori e relative variabili dipendenti).

  Se il parametro strutturale è espresso non con un valore numerico ma attraverso un simbolo, esso presenta due indici deponenti; nel caso di legami causali il primo si riferisce alla variabile «effetto» e il secondo alla variabile «causa»; nel caso di correlazioni l'ordine è indifferente.

5. GLI ASPETTI TECNICI DELLA CONDENSAZIONE

5.1 I CRITERI DI PONDERAZIONE

5.1.1 Metodi statistici

Correlazione

Come abbiamo visto l'attribuzione di pesi uguali in presenza di indicatori elementari correlati può introdurre un effetto di *double counting* nell'indicatore sintetico. Adottando il criterio della correlazione, il peso da attribuire a ciascun indicatore può essere determinato attraverso il livello di correlazione che l'indicatore registra con gli altri indicatori elementari coinvolti o con il punteggio totale non pesato. in particolare, tale peso può essere inversamente proporzionale al livello di correlazione in modo da attribuire meno importanza agli indicatori tra loro correlati. Nell'applicare tale approccio occorre tenere presente che esso richiede che vengono soddisfatti i tradizionali assunti parametrici (normalità delle distribuzioni, linearità delle relazioni tra indicatori). Nel caso estremo di perfetta collinearità tra gli indicatori, si può affermare che la variabile latente può essere rappresentata da un unico indicatore elementare.

Analisi delle Componenti Principali

Tale metodo è molto adatto quando la variabile latente è multidimensionale. Attraverso tale metodo è possibile, dopo aver individuato le componenti che spiegano la maggiore quantità di varianza esistente tra tutti gli indicatori elementari, individuare dei pesi (*component score*) da attribuire agli indicatori elementari nel calcolo dei punteggi di sintesi per ciascuna componente (Dunteman, 1983; Maggino, 2005). Ciascuna componente estratta individua un insieme di indicatori che presentano la più alta associazione. In pratica tale analisi consente di identificare gli indicatori che sono collineari in modo da poter calcolare un indicatore sintetico che sia in grado di catturare quanta più informazione possibile da tali indicatori. Il peso da attribuire agli indicatori elementari per il calcolo dei punteggi aggregati non può essere rappresentato dal *loading*, in quanto questo indica solo la validità di ciascun indicatore nel definire un concetto generale; è per questo necessario determinare il peso originale di ciascun indicatore nel definire ciascuna componente.
Attraverso questo metodo è possibile stimare tali pesi (detti *component score*) che non definiscono l'importanza degli indicatori elementari ma rilevano il contributo di ciascuno di questi alla definizione della componente cui si riferiscono, eliminando quella parte di contributo spiegato dalla correlazione tra indicatori. In generale gli *score* assumono valori più bassi dei *loading*; è possibile che alcuni indicatori, pur presentando valori di *loading* più alti di altri, registrino valori più bassi per gli *score*; ciò sta ad indicare che essi non danno un contributo *originale* alla definizione della componente ma si sovrappongono ad altri indicatori andando insieme a misurare la stessa dimensione.
Quando le componenti identificate riflettono perfettamente la struttura dimensionale esistente (e verificato attraverso l'analisi fattoriale) e ciascun indicatore ha *loading* significativi solo su una componente, i punteggi ottenuti su componenti ortogonali non sono correlati e conseguentemente sarà possibile considerare separatamente ciascuna dimensione.
Nell'adottare tale metodo occorre tenere presente che il peso che si ottiene ha un valore essenzialmente statistico legato al livello di variazione espresso da ciascun indicatore elementare.

Data Envelopment Analysis (DEA)

Tale metodologia non-parametrica fa parte de gli approcci sviluppati nell'ambito dell'economia e finalizzati allo studio e alla valutazione dell'efficienza/inefficienza nei processi di produzione.
L'obiettivo della *DEA* è quello di stimare la cosiddetta *efficiency frontier*, utilizzata per misurare le performance relative dei diversi casi (detti *Decision Making Unit*, DMA) in termini di distanza di ciascuno di questi da tale limite. L'insieme dei pesi deriva da tali confronti.
Dati gli obiettivi, la DEA può rappresentare un valido approccio alla individuazione dei pesi da attribuire agli indicatori elementari soprattutto nei casi in cui gli indicatori elementari si riferiscono a misure di capacità.
Tale approccio, sviluppato alla fine degli anni 70[1], misura l'efficienza nel caso di processi di produzione che presentano una struttura con diversi input e output, basandosi sulla programmazione lineare.

La realizzazione di tale approccio prevede l'esecuzione di due fasi:

Identificazione della *performance/efficiency frontier*, sulla base dei seguenti assunti:

- pesi positivi (maggiore è il valore dell'indicatore elementare, maggiore dovrà essere il valore del peso),
- nessuna discriminazione dei casi che risultano essere migliori in qualsiasi indicatore elementare (detto "dimensione") con conseguente uguale ordinamento,
- convessità della frontiera, ovvero possibilità di definire una combinazione lineare delle migliori performance.

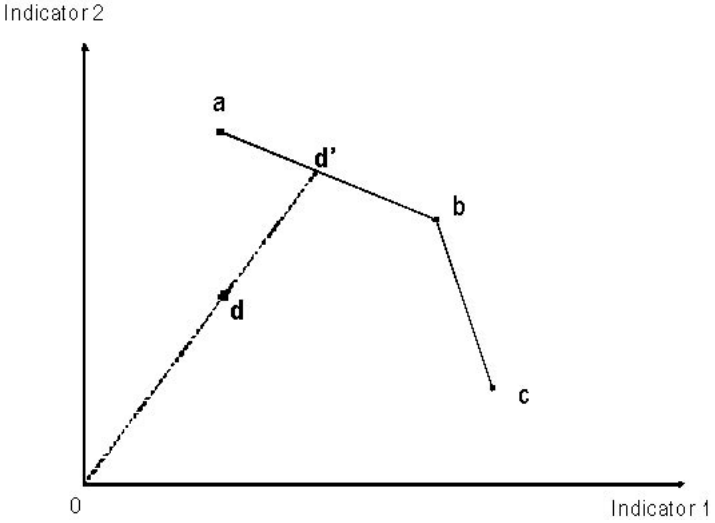
Misura della distanza di ciascun caso dalla *performance/efficiency frontier*, determinando sia la sua posizione che quella della frontiera.

Facciamo un esempio: poniamo di avere quattro casi (*a, b, c, d*) e due indicatori (1, 2). Nel sistema spaziale identificato dai due indicatori, le quattro unità si posizionano secondo i valori da loro registrati per i due indicatori elementari. I casi che hanno ottenuto le migliori performance sui due indicatori definiscono la *efficiency frontier* (retta che unisce i punti che individuano i casi *a*, *b* e *c*). Occorre a questo punto valutare le performance degli altri casi definendo due distanze: tra la posizione ottenuta e l'origine e tra la posizione teorica di migliore performance (proiezione del punto sulla frontiera) e l'origine. I casi che hanno ottenuto

- le migliori performance registreranno un valore di performance uguale a 1,
- performance peggiori registreranno un valore minore di 1.

Vediamo come viene calcolata la performance per il caso *d*:

$$performance = \frac{\overline{od}}{\overline{od'}}$$



[1] Il lavoro di riferimento è quello di Charnes, A., W.W. Cooper, and E. Rhodes (1978) "Measuring the Efficiency of Decision Making Units", *European Journal of Operational Research*, 2(6), pp. 429-444.

Per ciascun caso l'insieme dei pesi dipenderà dalla sua posizione rispetto alla *efficiency frontier*. Il limite rispetto al quale ciascun caso dovrà confrontarsi è quindi rappresentato dalla posizione che avrebbe se avesse ottenuto la performance migliore ideale (*d'*, nell'esempio).
La *efficiency frontier* può essere determinata anche dalla valutazione di esperti che definiscono per ciascun indicatore la performance ideale oppure la combinazione di valori preferibile per l'insieme degli indicatori elementari considerati (in maniera molto simile, come vedremo, a quanto avviene con l'approccio detto *Budget Allocation*).
I vantaggi nell'utilizzo di tale approccio risiedono nel fatto che:

- non richiede l'esplicita specificazione di una funzione matematica che descriva il modello di performance,
- può risultare utile nel portare alla luce relazioni che altre metodologie non riescono a identificare,
- è in grado di gestire molti indicatori elementari contemporaneamente,
- può essere utilizzato con qualsiasi tipo di misurazione,
- per ogni unità valutata è possibile identificare, analizzare e quantificare le fonti di inefficienza (ovvero rispetto a quale indicatore elementare si osserva le performance peggiore).

Un caso particolare di DEA è la procedura detta ***Benefit-Of-the-Doubt*** (BOD). Tale procedura consente per ciascun caso di enfatizzare o rendere prioritari quegli aspetti per i quali esso ottiene performance relativamente buone (identificazione individuale dell'obiettivo strategico o prioritario, *target*, rispetto alla *efficiency frontier*).
L'ottimizzazione del procedimento potrebbe condurre alla definizione di pesi nulli nel caso in cui non si pongano restrizioni alla definizione delle migliori performance individuali.
Per tale motivo l'utilizzo pratico di tale approccio richiede l'individuazione di restrizioni nella definizione delle frontiere individuali e conseguentemente dei pesi.

Unobserved Components Models (UCM)

L'idea alla base di questo approccio è che gli indicatori dipendono da variabili inosservate più un termine di errore. Per stimare la componente ignota è possibile fare luce sulla relazione tra l'indicatore sintetico e le sue componenti (indicatori elementari). I pesi ottenuti saranno determinati per minimizzare l'errore nell'indicatore sintetico, utilizzando il metodo di massima verosimiglianza (*maximum likelihood*).
Questo approccio ha qualche somiglianza con la logica della regressione lineare; la differenza sta nel fatto che la variabile dipendente qui è ignota.
Secondo questo modello, il punteggio osservato è uguale a

$$I_{iq} = \alpha_q + \beta_q \left(ph_c + e_{iq} \right) \text{dove}$$

ph_c	fenomeno ignoto
$q = 1,...,Q_c$	gruppo di indicatori, ciascuno dei quali misura un aspetto di ph_c
I_{iq}	punteggio osservato dal caso *i* per *q*
e_{iq}	termine di errore per il punteggio osservato dal caso *i*
α_q e β_q	parametri ignoti per posizionare ph_c su I_{iq}

Tale approccio richiede che siano soddisfatti molti assunti:

- l'errore rappresenta una variabile indipendente,
- l'errore comprende due fonti di incertezza: errore di misurazione e relazione imperfetta tra indicatori elementari e indicatore sintetico,
- l'indicatore sintetico ignoto (ph_c) è una variabile casuale con media zero e varianza unitaria,
- gli indicatori elementari devono essere ri-scalati in modo da avere valori tra zero e uno,

- ph_c e e_{iq} hanno distribuzione normale.

Utilizzando la media della distribuzione condizionale della componente non osservata, i pesi sono uguali a

$$w_{iq} = \frac{\sigma_q^{-2}}{1 + \sum_{q=1}^{Q_i} \sigma_q^{-2}}$$

dove w_{iq} rappresenta il peso per il caso *i* e l'indicatore *q*, w_{cq}

In pratica w_{iq} rappresenta una funzione

- decrescente rispetto alla varianza dell'indicatore *q* (maggiore è la varianza dell'indicatore, minore è la sua precisione e minore sarà il peso associato all'indicatore),
- crescente rispetto alla varianza degli altri indicatori.

La confrontabilità tra casi è assicurata dal fatto che il numero di indicatori elementari deve essere uguale per tutti i casi (nessun valore *missing*).

5.1.2 Metodi multi-criterio

Tra i modelli multi-criterio (*Multi-Attribute Models*) rientra una serie di metodi che consentono di prendere decisioni (di valutazione, di priorità, di selezione) tra diverse alternative disponibili. Tali metodi consentono di identificare l'importanza dei singoli indicatori elementari (chiamati attributi) sulla base di un certo numero di confronti combinati. I pesi prodotti sono meno sensibili agli errori che si evidenziano nelle valutazione dirette.
Tra tali metodi rientrano:

a. ***Multi-Attribute Decision Making*** (*MADM - Metodi Decisionali Multi-Criterio*): rappresenta un approccio che rientra nel più ampio settore del *Multiple Criteria Decision Making* (MCDM) che studio i meccanismi che vengono applicati nel prendere le decisioni tra diverse alternative disponibili che sono caratterizzate da molti attributi, di solito tra loro conflittuali (Yoon, 1995). Tra le tecniche utilizzate in questo ambito vi è l'**Analytic Hierarchy Process (AHP)** (*pairwise comparison of attributes*).
b. ***Multi-Attribute Compositional Models***, caratterizzati dall'approccio detto **Conjoint Analysis (CA)** (*comparison of attributes on different levels*); tale tecnica statistica, *decompositiva*, ha l'obiettivo di determinare quale combinazione di un numero limitato di attributi è preferita dal valutatore; mentre l'approccio AHP deriva l'"importanza" di una alternativa sommando il valore di ciascun item, la *conjoint analysis* procede in senso inverso, disaggregando le preferenze.

***Analytic Hierarchy Processes* (AHP)**

Tale tecnica consente di decomporre il problema legato ad una decisione in termini gerarchici, di incorporare nel processo valutativo aspetti sia qualitativi che quantitativi, di analizzare e confrontare più soluzioni, ciascuna delle quali è corredata da **pro** e **contro**.
L'AHP è considerata una metodologia *compensativa* in quanto le alternative che si rivelano efficienti rispetto a uno o più obiettivi possono compensare gli altri attraverso le loro performance.
L'AHP si basa su tre principi fondamentali:

- non sono ammessi attributi (indicatori elementari) che interagiscono o che sono interrelati (indipendenza dei criteri); le preferenze che si possono esprimere per le diverse alternative dipendono da attributi separati che possono essere indipendentemente sostenuti e a cui è possibile attribuire punteggi numerici;

- è possibile porre in gerarchia gli attributi (indicatori elementari) e calcolare il punteggio per un ciascun livello della gerarchia come somma pesata dei punteggi dei livelli più bassi; tale assunto non ammette attributi che abbiano una soglia;
- per ciascun livello è possibile calcolare punteggi a partire da confronti accoppiati; ciò risulta praticabile solo con un numero basso di indicatori; si pensi, infatti, che con 4 alternative i confronti sono 6 (4*3/2) mentre con 20 alternative, le coppie da considerare sono 190.

Il procedimento, che utilizza molti strumenti definiti nell'ambito dell'approccio di *scaling* di Thurstone (v. *Parte II*), si svolge secondo le seguenti fasi:

I fase: Identificazione e definizione del problema
1. Formulazione dell'obiettivo (decisione da prendere).
2. Identificazione dei soggetti coinvolti nella valutazione.
3. Definizione degli attributi (indicatori elementari).

II fase: confronto tra indicatori elementari
Durante tale processo il giudizio viene rilevato in modo sistematico utilizzando una delle seguenti note tecniche:
- confronti accoppiati (tra due indicatori si indica il più importante e di quanto)
- *ranking* (gli indicatori elementari vengono messi in ordine di importanza)
- *rating scale* (a ciascun indicatore si attribuisce un punteggio di importanza su una scala di rating "0-10" oppure "1-7").

In genere si utilizza la tecnica dei confronti accoppiati in cui bisogna indicare quanto il primo indicatore è più importante del secondo su una scala il cui *range* va da "1" (sono uguali) a "9" (il primo indicatore è 9 volte più importante del secondo). Sono considerati anche i valori inferiori a 1.

III fase: creazione della matrice di confronto
Si costruisce una matrice (**A**) in cui viene riportato il valore ***a*** che rappresenta l'importanza di un indicatore rispetto all'altro. In particolare

$$a_{i=j} = 1 \qquad a_{i \neq j} = k \qquad a_{j \neq i} = \frac{1}{k}$$

A =

		j			
		1	2	3	4
i	1	1	1/5	1/3	1/7
	2	5	1	3	5
	3	3	1/3	1	3
	4	7	1/5	1/3	1

A =

		j			
		1	2	3	4
i	1	1.000	0.200	0.333	0.143
	2	5.000	1.000	3.000	5.000
	3	3.000	0.333	1.000	3.000
	4	7.000	0.200	0.333	1.000

IV fase: standardizzazione della matrice
Si standardizzano i pesi ***a*** calcolando la somma di ciascuna colonna e dividendo ciascun valore per la corrispondente somma

$\overline{A} =$

		j			
		1	2	3	4
i	1	0.063	0.115	0.071	0.016
	2	0.313	0.577	0.643	0.547
	3	0.188	0.192	0.214	0.328
	4	0.438	0.115	0.071	0.109

Come si può notare nell'esempio, i valori della seconda riga sono più alti degli altri (eccetto colonna 1). Ciò indica incoerenza nei valori. Se i dati fossero stati coerenti le quattro colonne avrebbero riportato valori identici.

V fase: calcolo dei pesi per gli indicatori elementariI pesi rappresentano la media dei valori di ciascuna riga della matrice:

$\overline{A} =$

		j				*w*
		1	2	3	4	
i	1	0.063	0.115	0.071	0.016	0.066
	2	0.313	0.577	0.643	0.547	0.520
	3	0.188	0.192	0.214	0.328	0.231
	4	0.438	0.115	0.071	0.109	0.183

sapendo che con *n* indicatori elementari:

$$\sum_{i=1}^{n} w_i = 1$$

Occorre osservare che l'incoerenza dei giudizi, pur non facilmente evitabile, può essere in qualche modo valutata[2] (fase IV). Anche se in una matrice di dimensione QxQ bastano solamente $Q-1$ confronti per poter stabilire i pesi, se si vogliono evitare errori di giudizio il reale numero di confronti da realizzare è di $Q(Q-1)/2$.

Ciò può risultare costoso da un punto di vista del calcolo ma produce un insieme di pesi più coerenti. Inoltre la ridondanza consente di calcolare una misura degli errori di giudizio (**rapporto di incoerenza**). Bassi valori di tale misura (massimo 0.1 - 0.2) indicano bassa incoerenza.

Conjoint Analysis (CA)

Tale approccio, detto anche *multi-attribute compositional model* oppure *stated preference analysis*, rientra tra i modelli di *scaling* (v. *Parte II*) e presenta alla base un approccio statistico detto *decompositivo*.

La CA ha avuto origine nella psicologia quantitativa ed è utilizzata in molte scienze sociali e applicate, come il *marketing* (valutazione di nuovi prodotti o di pubblicità) o la ricerca operativa.

L'obiettivo è quello di determinare quale combinazione di un numero limitato di attributi è preferita dal valutatore (Hair, 1998; Louviere, 1988; Malhotra, 1993).

In pratica si richiede ad un gruppo di giudici (esperti) di fare una valutazione scegliendo (o ordinando) – secondo un criterio di preferenza – una serie di scenari alternativi ognuno dei quali rappresenta il profilo dell'indicatore sintetico. Ciascuno scenario rappresenta un insieme di valori per gli indicatori elementari. Successivamente, attraverso un procedimento di decomposizione, si mettono in relazione le singole componenti (i valori noti degli indicatori di tale scenario) e le valutazioni fatte.

Il valore assoluto (detto "livello") degli indicatori elementari può essere ricavato sia a livello individuale che a livello di gruppo.

Vengono stimate in successione

- una funzione di preferenza utilizzando l'informazione proveniente dalle preferenze espresse rispetto agli scenari,
- una probabilità della preferenza come funzione dei livelli degli indicatori che definiscono gli scenari alternativi:

$$pref_c = P(I_{1c}, I_{2c}, ..., I_{nc})$$

dove

I_{ic} livello dell'indicatore i $(i = 1, n)$ per il caso c

Dopo aver stimato tale probabilità (spesso utilizzando modelli di scelta discreta), è possibile utilizzare le derivate rispetto agli indicatori della funzione di preferenza come pesi per gli indicatori elementari nell'indicatore composito:

$$CI_c = \sum_{i=1}^{n} \frac{\partial P}{\partial I_{ic}} I_{ic}$$

Il rapporto rappresenta il peso.

L'idea è quella di calcolare il differenziale totale della funzione P nel punto di indifferenza tra stati alternativi. Risolvendo rispetto all'indicatore elementare, si ottiene il tasso marginale di sostituzione di I_{ic}. Il peso viene ad indicare quanto la preferenza cambia con il cambiamento dell'indicatore.

Tale approccio, rivelandosi *compensabile*, richiede una attenta valutazione sulla sua applicabilità a tutte le situazioni.

[2] E' possibile calcolare un indice di incoerenza utilizzando gli *eigenvalue* della matrice standardizzata.

5.1.3 Ricorso ad esperti

Come si è detto, è possibile determinare i pesi degli indicatori elementari anche ricorrendo alle opinioni di esperti. In questi casi, per raccogliere si utilizzano diverse tecniche tra le quali quella detta *Budget Allocation* (BAL).
Si chiede ad un gruppo di esperti di distribuire un certo "budget" di *N* punteggi tra gli indicatori elementari, attribuendo più budget a quegli indicatori la cui importanza si vuole mettere in evidenza.
Il procedimento può essere suddiviso in quattro fasi:
- selezione degli esperti che devono valutare;
- distribuzione del budget agli indicatori;
- calcolo dei pesi;
- iterazione del punto (b) fino al raggiungimento della convergenza (opzionale).

In alcuni casi il procedimento applicato è più semplice in quanto ricerca il consenso di un gruppo di esperti nel giudicare almeno il contributo relativo degli indicatori all'indicatore sintetico. Anche se i giudizi e le opinioni possono divergere, è importante poter giungere ad un reale consenso tra gli le persone che posseggono la conoscenza e l'esperienza in modo da assicurare un sistemi di pesi più adatto all'applicazione.
Tale approccio risulta essere ottimo nel caso di un numero massimo di 10-12 indicatori. Con un numero maggiore di indicatori, il lavoro richiesto agli esperti può risultare gravoso.

5.2 LE TECNICHE DI AGGREGAZIONE

5.2.1 Criteri per la scelta della tecnica

5.2.1.1 "Compensabilità" della tecnica

La tecnica di aggregazione compensa quando nella determinazione del punteggio sintetico, valori bassi in alcun indicatori vengono compensati da valori sufficientemente alti di altri indicatori.
Per comprendere meglio tale concetto, vediamo di seguito la rappresentazione di una classica tabella di aggregazione in cui è possibile osservare le diverse combinazioni che producono due indicatori elementari (A e B) – misurati rispettivamente con 4 e 3 livelli – quando vengono aggregati con una semplice somma:

		B		
		1	2	3
A	4	5	6	7
	3	4	5	6
	2	3	4	5
	1	1	3	4

Come si può osservare, a partire dal punteggio di sintesi non è possibile risalire al profilo dei punteggi degli indicatori elementari; infatti lo stesso punteggio aggregato è ottenuto da combinazioni diverse dei punteggi degli indicatori elementari.
Poniamo di avere un indicatore composito formato da alcuni indicatori elementari. Due casi, A e B, presentano diversi profili (relativamente agli indicatori elementari) ma uguali punteggi aggregati che evidentemente non è in grado di riflettere le loro diverse condizioni sociali:

A	[21,1,1,1]	6.00
B	[6,6,6,6]	6.00

Tale limite è prodotto dall'adozione della somma semplice (aggregazione additiva) che è inevitabilmente **compensativa**.
Ciò porta ad una attenta riflessione sull'importanza che assume la scelta della tecnica di aggregazione che può comportare una incoerenza rispetto alla determinazione (fatta di solito in termini di importanza) dei pesi e il reale significato che assumono quando si seleziona una o l'altra delle tecniche di aggregazione.
Per questo, se si vuole essere sicuri che i pesi continuino anche in sede di aggregazione a rappresentare una misura di importanza degli indicatori elementari, è più utile adottare procedure di aggregazione non compensative.

5.2.1.2 Omogeneità e confrontabilità dei livelli di misurazione

Indicatori metrici. Con questo tipo di indicatori è possibile scegliere tecniche di calcolo dei punteggi che prevedono calcoli aritmetici (somma, moltiplicazione, media); nel calcolo è importante prestare particolare attenzione al tipo di distribuzione osservata per ciascun indicatore elementare.
Indicatori ordinali (graduatorie) o con categorie ordinate. Il trattamento di questo tipo di indicatori è quello tipico delle variabili ordinali. Vi sono dei casi in cui si preferisce trattare tali dati come dati metrici. In questi casi, per poter assimilare gli indicatori ordinali a quelli metrici, si assume che tra le categorie vi sia equidistanza; la correttezza di tale assunto è molto discutibile e suscita molte perplessità. Ciò riguarda soprattutto il trattamento delle categorie ordinate. Si pensi a tale proposito alle tipiche categorie ordinate utilizzate nelle rilevazioni soggettive di diversi livelli di accordo, soddisfazione, valutazione, ecc., quando è molto difficile assumere l'equidistanza semantica tra tutte le posizioni.
A volte per superare tali problemi si può procedere alla dicotomizzazione delle scale ordinali utilizzate. Pragmaticamente, si preferisce adottare comunque le tecniche proprie per misure metriche in quanto la potenza statistica dovrebbe compensare eventuali distorsioni derivate dagli assunti.Indicatori categorici. Per poter combinare indicatori misurati attraverso categorie è necessario procedere osservando le distribuzioni combinate degli indicatori coinvolti (tavole di frequenze a due o più dimensioni). Tale procedimento consente di avere un quadro delle aggregazioni osservabili in riferimento a quelle teorizzate per misurare il concetto generale. Tali criteri devono considerare:
- l'*omogeneità* dei casi appartenenti ad una combinazione di categorie;
- la *differenziazione* tra i casi appartenenti a combinazioni diverse;
- l'*equilibrio* nella distribuzione dei casi tra le combinazioni identificate.

Si può fare l'esempio di un indicatore prodotto dalla sintesi di due indicatori elementari categorici (A e B) che hanno rispettivamente r e c categorie; la combinazione diretta di A e B produrrà un nuovo indicatore con $r * c$ categorie, corrispondenti a tutte le possibili combinazioni. In presenza di indicatori con un numero di categorie elevato, il metodo della semplice combinazione può risultare alquanto impraticabile, sia dal punto di vista pratico che dell'interpretazione.

5.2.2 Approccio lineare

Rappresenta l'approccio più utilizzato, utile quando tutti gli indicatori elementari presentano la stessa unità di misura; tale aggregazione può essere vista secondo due diverse prospettive: **additiva** e **cumulativa**[3]**.** Nel primo caso si assume che gli item contribuiscono alla definizione della variabile latente (e quindi al punteggio aggregato) nello stesso modo, nel secondo caso si assume che gli item

[3] Vedi a tale proposito i corrispondenti modelli di *scaling*.

contribuiscono alla definizione della variabile latente secondo una logica scalare (definiscono punti diversi del continuum). L'interpretazione dei punteggi aggregati è naturalmente diversa.
L'**aggregazione additiva classica** si basa sugli assunti di:

- *unidimensionalità:* l'insieme degli indicatori elementari combinati misura (sono correlati) con una sola variabile latente; in particolare gli indicatori elementari contribuiscono allo stesso modo, con lo stesso peso e la stessa importanza, alla descrizione della dimensione misurata; conseguentemente, gli indicatori sono selezionati sulla base della loro capacità nel discriminare tra unità e non tra livelli diversi del continuum che rappresenta la variabile latente;
- *monotonicità della relazione con la variabile latente:* ciascun indicatore elementare è monotonamente correlato al continuum della variabile latente; quindi, nel caso si misuri, per esempio, un atteggiamento, più favorevole (o sfavorevole) è l'atteggiamento del soggetto, maggiore (o minore) è il punteggio atteso per l'indicatore composito;
- *compensabilità* degli indicatori elementari;
- *omogeneità* dei livelli di misurazione degli indicatori elementari.

La base logica per verificare l'assunto di omogeneità è l'*analisi della consistenza interna* (si veda a tale proposito la parte relativa ai modelli di *scaling*) che mira a verificare quanto bene il gruppo di indicatori elementari descrive la variabile unidimensionale latente. Tale verifica si basa sull'analisi del livello di correlazione esistente in un gruppo di indicatori misurati con la stessa scala.
Per le sue caratteristiche e le restrizioni che impone, tale tecnica è adatta nel caso di sintesi di indicatori elementari che misurano insieme un'unica variabile latente.
Esiste un'altra versione di aggregazione additiva applicata quando non si assume la unidimensionalità ma l'**indipendenza** tra indicatori: in questo caso l'aggregazione additiva è ammessa se gli indicatori elementari sono indipendenti.
Tale assunto è particolarmente forte e implica che il rapporto di scambio tra due indicatori sia indipendente dai valori dei rimanenti indicatori. In altre parole, l'aggregazione additiva implica che tra i diversi aspetti (item) non vi sia alcun conflitto o sinergia. Tale assunto non è però sempre semplice da soddisfare e da verificare.
Tecnicamente, l'approccio lineare si esprime nel modo seguente:

lineare semplice	*lineare pesata*
$CI_c = \sum_{i=1}^{n} I_{ic}$	$CI_c = \sum_{i=1}^{n} w_i I_{ic}$

dove

CI_c	indicatore composito per il caso c
w_i	peso associato all'i-esimo indicatore elementare con $\sum_{i=1}^{n} w_i = 1$ e $0 \le w_i \le 1$
n	numero di indicatori elementari

Tale aggregazione può essere espressa anche in termini di media. Nel caso in cui i valori siano stati trasformati in numeri indice è consigliabile, in sostituzione della media aritmetica, occorre utilizzare la media geometrica

L'aggregazione lineare ha una variante nel caso in cui i valori per tutti gli indicatori elementari siano espressi come **ranghi**:

somma semplice	*media*	*indice relativizzato*
$CI_c = \sum_{i=1}^{n} rank_{ic}$	$\overline{CI}_c = \frac{CI_c}{n}$	$^*CI_c = \frac{CI_c - n}{mn - n}$

dove

CI_c indicatore composito per il caso *c*

n numero di indicatori elementari

m numero dei casi

$rank_{ic}$ rango dell'*i*-esimo indicatore elementare per il caso *c*

L'indice relativizzato varia tra 0 (migliori posizioni di graduatoria) e 1 (peggiori posizioni di graduatoria). Questo approccio è detto *M-ordinamento* e consente di esprimere un ordinamento dei casi che tenga conto di tutti gli indicatori elementari.
Un'altra variante è quella detta **threshold method**: si calcola la differenza tra il numero di indicatori i cui punteggi sono al di sopra o al di sotto di una soglia arbitraria (di solito la media).

$$CI_c = \sum_{i=1}^{n} \mathrm{sgn}\left[\frac{I_{ic}}{I_{EUi}} - (1 + p)\right]$$

dove

n numero di indicatori elementari

CI_c indicatore composito per il caso *c*

p valore soglia (scelto arbitrariamente sopra o sotto la media)

L'**aggregazione cumulativa** è adottata quando gli indicatori elementari non contribuiscono allo stesso modo, con lo stesso peso e la stessa importanza, alla descrizione della dimensione misurata. Ciascun indicatore elementare discrimina i casi in punti diversi del continuum della variabile latente; in altre parole gli indicatori sono selezionati sulla base della loro capacità di discriminare non solo tra unità ma anche tra livelli diversi del continuum che rappresenta la variabile latente.Perché il punteggio aggregato ottenuto con questo approccio abbia una giustificazione teorica, è necessario che vengano soddisfatti i seguenti assunti:

- *unidimensionalità*: l'insieme degli indicatori elementari esprimono la misura di una sola variabile latente;
- *relazione differenziata* di ciascun indicatore elementare con la variabile latente con conseguente
- *assenza di compensabilità* tra indicatori rilevabile, definibile in termini di *gradualità/scalabilità*: gli indicatori elementari devono essere scelti in modo tale che risultino essere discriminanti a livelli diversi della stessa dimensione; in altre parole deve essere possibile ordinare gli indicatori elementari secondo un livello crescente di intensità (capacità, disposizioni, difficoltà, ecc.); gli indicatori così selezionati presentano solo una parziale sovrapposizione di significato, consentendo di ottenere una *gradualità* della valutazione;*omogeneità*: tutti gli indicatori elementari sono rilevati con lo stesso tipo di livelli di classificazione;
- *esaustività*: l'insieme degli indicatori elementari deve rappresentare un inventario completo del dominio reale di una "dimensione" ovvero gli indicatori elementari devono ricoprire tutta la variabilità osservabile in modo da consentire una valutazione globale.

Tale approccio fa riferimento ai modelli di *scaling* detti cumulativi (v. *Parte II*).

5.2.3 Approccio geometrico

Tale approccio richiede il calcolo del prodotto degli indicatori elementari pesati e indicatori elementari misurati su scale a rapporti. Le funzioni moltiplicative non sono semplici da gestire; per tale motivo si cerca di semplificarne il trattamento trasformando i valori da aggregare in logaritmi (il prodotto di valori non è altro che la somma dei loro logaritmi). Tale procedimento deve comunque essere eseguito con una certa cautela.
Per comprendere meglio le caratteristiche dell'approccio geometrico, riprendiamo nella seguente tabella l'esempio utilizzato per spiegare il concetto di compensabilità:

		B		
		1	2	3
A	4	4	8	12
	3	3	6	9
	2	2	4	6
	1	1	2	3

Come si può osservare, se pur in misura minore, anche l'approccio moltiplicativo non consente, a partire dal punteggio di sintesi, di risalire al profilo dei punteggi degli indicatori elementari.
Ciò rende anche la tecnica moltiplicativa **compensativa** se pur in misura minore rispetto al precedente approccio e soprattutto rispetto agli indicatori con valori bassi. Riprendendo l'esempio usato in precedenza, vediamo quali sono per i due casi considerati i rispettivi punteggi aggregati:

A	[21,1,1,1]	2.14
B	[6,6,6,6]	6.00

Ammettendo la compensabilità, un caso con punteggi bassi su un indicatore avrà bisogno di un punteggio più alto sugli altri per "migliorare" il suo punteggio sintetico.
E' per questo che l'uso di questa tecnica può risultare molto utile nell'ambito, per esempio, di procedure valutative in quanto in questo modo si stimolano in modo diverso i casi (per esempio studenti) che per migliorare la propria valutazione totale si vedono costretti a migliorare le valutazioni relative a quei settori i cui indicatori producono i punteggi più bassi.

5.2.4 Approccio non compensativo

Come abbiamo visto, di solito si attribuiscono pesi maggiori agli indicatori elementari che sono considerati più significativi nel particolare contesto rappresentato dall'indicatore sintetico.
Nelle aggregazioni effettuate con le tecniche additive o geometriche, i pesi associati agli indicatori elementari finiscono con non indicare l'importanza dei corrispondenti indicatori. Ciò comporta, come abbiamo visto, una logica compensativa ovvero la possibilità di dare uno svantaggio su alcune variabili attraverso un vantaggio sufficientemente grande su altre variabili.
Tale implicazione rappresenta l'esistenza di una incoerenza teorica tra il modo in cui i pesi vengono utilizzati e il loro significato teorico. Quando

- si aggregano dimensioni molto diverse tra loro,
- i pesi sono interpretati in termini di "coefficienti di importanza",
- si ritiene che un aumento nelle performance in un ambito (per esempio la salute fisica o le relazioni interpersonali) non possa compensare una perdita o un peggioramento in altri ambiti,

la costruzione di indicatori compositi dovrebbero seguire un approccio non-compensativo. Ciò può essere fatto utilizzando il ***non-compensatory multi-criteria analysis*** (MCA).

Tale procedura cerca di risolvere, utilizzando una logica non-compensativa, i conflitti che possono emergere nei casi in cui, nel confrontare casi, vi sono indicatori particolarmente positivi per alcuni casi ed altri indicatori che sono particolarmente positivi per altri casi.

Poniamo di avere le seguenti condizioni:

- un insieme di indicatori elementari $\mathbf{G} = \{x_i\}$ con i=1, …, n
- un insieme di casi $\mathbf{M} = \{c\}$ con c=1, …, M
- la valutazione di ciascun caso c rispetto ad un indicatore elementare x_i è basata su una scala di misura ad intervallo o a rapporti,
- un valore maggiore di un indicatore elementare è preferibile rispetto ad uno minore (polarità positiva),
- un insieme di pesi (interpretati in termini di coefficienti di importanza) con $\sum_{i=1}^{n} w_i = 1 \quad w = \{w_i\}$ con i=1, …, n

Tali informazioni costituiscono la ***impact matrix***.
Il problema matematico è come utilizzare tali informazioni per ordinare i casi senza alcuna relazione di incomparabilità (*complete pre-order*).
Anche se è difficile eliminare completamente tutte le fonti di incertezza e di imprecisione, nel procedimento di aggregazione è comunque importante che sia possibile verificare le seguenti proprietà:

- intensità della preferenza (quanto il caso *a* è migliore del caso *b* rispetto all'indicatore *i*),
- numero di indicatori a favore di un dato caso,
- peso associato a ciascun indicatore,
- relazione di ciascun caso rispetto a tutti gli altri.

Il **procedimento matematico di aggregazione** può essere suddiviso in due passaggi:

(a) <u>confronto a coppie dei casi rispetto all'intero insieme di indicatori elementari</u>,
Si costruisce una matrice **E** di dimensione MxM (detta *outranking matrix*) in cui ogni elemento e_{jk} (con $j \neq k$) rappresenta il risultato del confronto tra il caso *j* e il caso *k*.

$$e_{jk} = \sum_{i=1}^{n} \left[w_i(\mathrm{Pr}_{jk}) + \frac{1}{2} w_i(In_{jk}) \right]$$

dove

$w_i(\mathrm{Pr}_{jk})$ peso dell'indicatore elementare *i* che presenta una relazione di preferenza

$w_i(\mathrm{Pr}_{jk})$ peso dell'indicatore elementare *i* che presenta una relazione di indifferenza.

In pratica, il punteggio e_{jk} rappresenta la somma di tutti i pesi di tutti gli indicatori elementari per i quali il caso *j* è migliore del caso *k*.
In tale procedimento occorre tenere presente che

$$e_{jk} + e_{kj} = 1$$

e che se due casi presentano entrambe buone performance per lo stesso indicatore, il peso sarà tra e_{jk} e e_{kj}.

Da notare che il procedimento di confronto a coppie è diverso da quello proposto *Analytical Hierarchy Processes* e dalla *Conjoint Analysis* (entrambi con logica compensativa); qui la questione è se l'indicatore *i* presenta un valore maggiore per il caso *a* o per caso *b*; in caso affermativo, è il peso dell'indicatore *i* ad entrare nel calcolo dell'importanza totale del caso *a* (coerentemente con la definizione di pesi come "misure di importanza").

(b) <u>ordinamento dei casi (*complete pre-order*)</u>.
Dopo aver effettuato tutti i confronti accoppiati si procede all'ordinamento sulla base di un algoritmo *Condorcet-Kemeny-Young-Levenglick* (CKYL) detto anche Condorcet-type of ranking procedure" secondo la logica del complete pre-order (ovvero alcuna relazione di incomparabilità).
Poniamo di avere tre casi (*A*, *B*, *C*) e di avere i loro punteggi *e*. E' possibile ora identificare tutte le possibili permutazioni nell'ordine delle tre unità (*ABC*, *ACB*, *BAC*, *BCA*, *CAB*, *CBA*) e calcolare per ciascuna di loro la somma ordinata dei punteggi, che per *ABC* sarà:

$$Y = e_{AB} + e_{AC} + e_{BC}$$

Si fa questo per tutte le permutazioni e si prende come *multi-criteria case ranking* quello che presenta il punteggio *Y* più alto.

Notare che tale procedimento è basato unicamente sui pesi e sul segno della differenza tra i valori ottenuti dai casi per un dato indicatore elementare (ovvero ignorando la dimensione della differenza).
Attraverso questo metodo un caso che risulta essere marginalmente migliore su molti indicatori occuperà una posizione migliore di un caso che decisamente migliore ma su pochi indicatori.
In questo sta la proprietà non compensativa dell'approccio: non può compensare le deficienze in alcune dimensioni grazie alle performance ottenute sulle altre.
Questo metodo di aggregazione ha il **vantaggio** di:
- superare i problemi prodotti dalle procedure additive e moltiplicative,
- trattare tutte le informazioni a livello ordinale (non premiando in questo modo le unità *outlier*),
- assicurare la confrontabilità degli indicatori elementari evitando qualsiasi manipolazione o normalizzazione dei dati.

Mentre ha lo **svantaggio** di:
- non considerare l'eventuale presenza di cicli/ranghi inversi nell'ordinamento finale (unità *a* prevale su *b*, *b* prevale su *c* e *c* prevale su *a*); tale svantaggio è lo stesso evidenziato, a livello di indicatori, dall'*Analytic Hierarchy Process*;
- non utilizzare informazioni riguardanti l'intensità di preferenza delle variabili (il metodo produce lo stesso valore di rango tra due unità indipendentemente dalla reale entità della differenza);
- presentare alti costi in termini di calcolo quando il numero dei casi è particolarmente alto (il numero delle permutazioni da calcolare aumenta in modo esponenziale).

5.3 LA VERIFICA DELLA ROBUSTEZZA

5.3.1 Analisi dell'incertezza e della sensibilità

Valutazione dell'incertezza

In generale, il migliore approccio per valutare le incertezze presenti è quello di ovvero un processo che consente di mettere a confronto (modello di valutazione) le diverse performance del punteggio prodotte da scelte alternative. I risultati di tale analisi riportano i *limiti di incertezza* ovvero il *range* all'interno del quale ricade il valore del punteggio per ciascun caso (maggiore è il *range* maggiore è l'incertezza).

Valutazione della sensibilità

Dopo aver verificato il livello di incertezza del punteggio per ciascun caso, è utile verificare quanto dell'incertezza riscontrata è attribuibile a ciascuna scelta fonte di incertezza (Edward, 1982).
In pratica si tratta di suddividere la varianza totale dell'output prodotta valutando il contributo individuale di tutte le fonti potenziali di incertezza. Il risultato di tale analisi è rappresentato dal *livello di sensibilità* di ciascuna delle fonti di incertezza. Il livello di sensibilità esprime (per un certo caso) quanta incertezza nel punteggio sarebbe ridotta se quella particolare fonte di incertezza fosse rimossa.
I risultati dell'analisi di sensibilità sono spesso presentati nella forma di uno *scatter-plot* che mostra in ordinata i valori del punteggio e in ascisse ciascuna fonte di incertezza.

5.3.1.1 Procedimento

Il procedimento di analisi dell'incertezza e della sensibilità (Nardo, 2005; Saisana, 2005; Saltelli, 2004; Tarantola, 2000) prevede le seguenti fasi:

a) individuazione delle fonti di incertezza;
b) specificazione della distribuzione di probabilità;
c) applicazione delle procedure per verificare la propagazione dell'incertezza e produrre una distribuzione di probabilità del modello(*uncertainty analysis*);
d) determinazione dell'intervallo di confidenza;
e) determinazione dei livelli di sensibilità di ciascuna fonte di incertezza (*sensitivity analysis*);
f) presentazione dei risultati.

(a) Individuazione delle fonti di incertezza

Possono essere considerate fonti di incertezza (*input factor*) il criterio di selezione degli indicatori elementari, il modello di definizione dell'errore di misurazione, le tecniche di trattamento dei dati (imputazione dei valori *missing*, tecniche di relativizzazione e di normalizzazione), criteri per la definizione dei pesi, tecnica di aggregazione.

(b) Specificazione della distribuzione di probabilità

Per poter procedere è necessario assegnare a ciascuna delle fonti di incertezza una distribuzione di probabilità. L'individuazione di tale distribuzione può scaturire da una precedente fase sperimentale oppure – più frequentemente – da una valutazione fatta dal ricercatore. Tale fase richiede un alto livello di competenza relativamente alla qualità dei dati. In genere per l'individuazione di tale distribuzione si fa riferimento al *range* dei possibili valori del fattore; con:

- *range* non superiore a 10 (incertezza bassa): si può adottare una distribuzione uniforme;
- *range* superiore a 10 (incertezza alta): si può adottare una distribuzione di probabilità dei logaritmi dei valori.

Nei casi in cui si considerino anche altri riferimenti, è possibile adottare altre distribuzioni (normale, gamma, beta, Poisson, Weibul, o distribuzioni discrete).

Nel caso esistano dubbi riguardo la distribuzione da adottare, è possibile assumere diverse distribuzioni per analizzarne gli effetti (*incertezza dell'incertezza*).

In generale, comunque, tenendo costanti media e varianza di tali possibili distribuzioni, l'effetto di tale scelta sulla valutazione dell'incertezza totale (ovvero sulla determinazione dell'intervallo di confidenza) è piuttosto limitato.

(c) Propagazione dell'incertezza

In genere per procedere alla analisi della propagazione dell'incertezza si ricorre ad approcci numerici,tra i quali ricordiamo la simulazione Monte Carlo che viene eseguita utilizzando uno dei seguenti approcci (procedimenti di campionamento casuale):

- *Simple Random Sampling* (SRS): a partire dalle distribuzioni specificate per ciascuna delle fonti di incertezza viene estratto un valore casuale e successivamente si calcola una stima; il procedimento viene ripetuto per un numero definito di iterazioni; si ottiene così una distribuzione di probabilità del modello; se la dimensione del campione finale è di poche migliaia, tale approccio si rivela meno efficiente del successivo;
- *Latin Hypercube Sampling* (LHS): la distribuzione di probabilità di ciascuna fonte di incertezza viene suddivisa in sezioni di ampiezza uguale; il numero delle sezioni è uguale al numero delle iterazioni; si estrae un numero casuale da ciascuna sezione che quindi viene esclusa dell'analisi successiva (in alternativa, si può individuare il valore mediano della sezione); i parametri (media e varianza) della distribuzione dei valori casuali ottenuti raggiungono abbastanza rapidamente una stabilità.

(d) Determinazione dell'intervallo di confidenza

Per ciascuna unità si determina l'intervallo di confidenza, presentando contemporaneamente la distribuzione ottenuta al precedente passaggio (valori ottenuti nelle diverse iterazioni), i suoi parametri (media, mediana, varianza) e il valore realmente registrato dal punteggio: maggiore è il *range*, maggiore è la differenza tra il valore osservato e la mediana dei valori casuali, maggiore è l'incertezza del valore per quella unità (valore detto "volatile").

(e) Determinazione dei livelli di sensibilità

Per la determinazione dei livelli di sensibilità è necessario scegliere un metodo (detto di *sensitivity analysis*); tale scelta dipende in gran parte dal livello di accuratezza desiderato nelle stime della misura di sensibilità (oltre che dal costo richiesto per il calcolo).
In tale fase occorre tenere presente che tutti i punteggi ottenuti sono considerati funzioni non-lineari dei fattori di incertezza in quanto si prendono in considerazione simultaneamente molti fattori di incertezza stratificati che interagiscono tra di loro.
In presenza di modelli non-lineari, occorre applicare tecniche di analisi di sensibilità robuste (*model-free variance based techniques*) che consentono di trattare i fattori di incertezza nel loro complesso, e non singolarmente; tali tecniche sono preferite anche perché sono semplici da interpretare, consentono di distinguere tra effetti principali (primo ordine) e effetti di interazione (secondo ordine e superiori).

(f) Presentazione dei risultati

La presentazione dei risultati prevede:
- la presentazione per ciascun caso della distribuzione dei punteggi ottenuti attraverso l'analisi di incertezza,
- il confronto del *range* di incertezza tra tutti i casi,
- la presentazione, per ciascun caso, del contributo (in termini percentuali) di ciascun fattore di incertezza sulla varianza totale.

Al termine di tale procedimenti di analisi è sicuramente possibile fare delle valutazioni che consentono di giungere alla definizione di una misura stabile.
Nel confermare la indubbia importanza della verifica della robustezza dei punteggi, è importante però non enfatizzare troppo questo tipo di analisi. Nel definire e scegliere i fattori di incertezza, infatti, si corre il rischio di selezionare fattori (e livelli dei fattori) che non sarebbero comunque "accettabili" e "plausibili" se si tentasse di applicarli alla situazione in cui si muove l'indicatore e al tipo di dati a disposizione.

5.4 LA VERIFICA DELLA CAPACITA' DI DISCRIMINARE

Tra le tecniche che consentono di verificare la capacità discriminante ricordiamo:
a. **test statistici di verifica dell'ipotesi**: per verificare la capacità discriminante dell'indicatore sintetico si può procedere al confronto delle performance di gruppi differenti definiti rispetto a particolari variabili di base (*comparative analysis*) e appartenenti a campioni probabilistici, utilizzando *test statistici di significatività* (Maggino, 2005);
b. **coefficienti di discriminazione**: una versione del coefficiente di discriminazione di un indicatore sintetico è quella definita nel caso in cui gli indicatori elementari presentano lo stesso numero di valori/categorie (da Ferguson in 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 dell'indicatore sintetico
N numero di casi
n numero di indicatori elementari che compongono l'indicatore sintetico
k numero di punteggi/valori/categorie per ciascun indicatore elementare
f_{ij} frequenza del j-esimo punteggio per l'i-esimo indicatore elementare

c. **tecniche per l'individuazione dei valori soglia**, delle quali ci occuperemo nei paragrafi successivi.

5.4.1 Verifica della selettività

Come si è detto, la verifica del modello di *scaling* legittima l'utilizzazione del punteggio rappresentato dall'indicatore sintetico – ottenuto dall'aggregazione di indicatori elementari – per posizionare ciascun caso misurato sul continuum che rappresenta la caratteristica rilevata.
E' necessario a questo punto individuare dei criteri che consentano di interpretare tale posizione rispetto alla caratteristica da misurare. L'esigenza di interpretazione dei punteggi è sia descrittiva che diagnostica.
A tale proposito, occorre notare che generalmente mentre l'interpretazione dei punteggi estremi appare piuttosto chiara, più difficile è l'interpretazione di tutti i punteggi intermedi; 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, il modello adottato è finalizzato alla misura della 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 cosa indicano i punteggi intermedi. La difficoltà deriva anche dal fatto che il punteggio ottenuto, così come la verifica del modello di *scaling*, non è indipendente dal campione utilizzato per la validazione.

5.4.1.1 Individuazione dei valori-soglia

L'obiettivo di tale analisi è quello di verificare la validità predittiva ovvero la capacità "diagnostica" o "di performance" dell'indicatore sintetico rendendo interpretabili i valori dell'intera distribuzione dell'indicatore.[4]
Per tale analisi non esiste una procedura standard ma quasi sempre si richiede l'utilizzo di criteri esterni. Nel caso in cui sia possibile disporre dei 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).
Il problema principale di tale tipo di analisi sta nell'incertezza e nella difficoltà di individuare punti della distribuzione che consentano realmente di discriminare tra popolazioni statistiche diverse.

[4] In genere i valori-norma possono essere espressi in forma di punteggi standard (in questo caso è possibile anche riferire i punteggi originali ad una distribuzione con media 500 ed deviazione standard 100), percentili, punteggi normalizzati (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).

In altre parole, utilizzando termini utilizzati in epidemiologia, è necessario individuare valori che non producano discriminazioni vere ovvero che non producano "falsi-positivi" e "falsi-negativi".
Per questo motivo in fase di costruzione e validazione dell'indicatore è necessario conoscere già informazioni riguardanti i casi che consentano di classificarli in modo corretto secondo criteri detti **golden standard**.
In questo modo attraverso una semplice tabella di contingenza sarà possibile valutare se il valore-soglia individuato è in grado di distinguere i casi che si vogliono selezionare da quelli che non si vogliono selezionare. Parlando per esempio in termini di

- performance positiva (ovvero da non selezionare e sui quali potrebbe, per esempio, non è necessario intervenire)
- performance negativa (ovvero da selezionare e sui quali potrebbe, per esempio, essere necessario intervenire),

è possibile costruire la seguente tabella di classificazione:

			Classificazione dei casi ottenuta con *golden standard*	
			Caso con performance	
			negativa	positiva
Classificazione dei casi prodotta dal *cut-point* individuato	Caso con performance	negativa	**TP**	**FP**
		positiva	**FN**	**TN**

In tale tabella i casi classificati dal *cut-point* dell'indicatore in maniera:

- **corretta**, quando la classificazione è in accordo con il *golden standard*; in particolare avremo casi:
  - **TP** (*true-positive*), ovvero quelli che il *cut-point* ha correttamente non selezionato,
  - **TN** (*true-negative*), ovvero quelli che il *cut-point* ha correttamente selezionato;
- **scorretta**, quando la classificazione non è in accordo con il *golden standard*; in questo caso avremo casi:
  - **FP** (*false-positive*), ovvero quelli che non sono da selezionare ma sono stati selezionati dal *cut-point*;
  - **FN** (*false-negative*), ovvero quelli che sono da selezionare ma non sono stati selezionati dal *cut-point*.

Le frequenze osservate in questa tabella consentono di calcolare due caratteristiche del valore-soglia (e conseguentemente dell'indicatore sintetico):

- la **sensibilità** (*Se*) ovvero probabilità che un caso con performance negativa risulti tale, → $Se = \frac{A}{A+C}$
- la **specificità** (*Sp*) ovvero probabilità che un caso con performance positiva risulti tale → $Sp = \frac{D}{D+B}$

Le due caratteristiche, *Se* e *Sp*, sono tra loro inversamente correlate in relazione alla scelta del valore di *cut-point*. Infatti, modificando quest'ultimo si può ottenere uno dei seguenti effetti:

- una diminuzione della *Se* cui corrisponde un aumento della *Sp*,
- un incremento della *Se* cui corrisponde una diminuzione della *Sp*.

In altre parole, l'adozione di una soglia che offre un'elevata *Se* comporta una perdita di *Sp* e viceversa.

Individuare il *cut-point* ottimale per un determinato indicatore sintetico e tentare di confrontare capacità selettive di indicatori diversi non è facile soprattutto se si tiene conto che:

a. è possibile scegliere un valore-soglia tale che risponda ad un predeterminato valore di *Se* o di *Sp*, ma non è detto che tale valore sia ottimale per gli obiettivi che si pongono;
b. la *Se* e la *Sp* associate ad un singolo valore-soglia non rappresentano reali descrittori della capacità selettiva di un indicatore sintetico (se si adottano altri valori-soglia la *sensibilità* e la *specificità* cambiano);
c. i valori predittivi non sono caratteristiche intrinseche dell'indicatore e quindi non possono essere utilizzati come descrittori esaurienti della capacità selettiva dell'indicatore in quanto tali valori dipendono, oltre che dalla *sensibilità* e dalla *specificità*, anche dalla *prevalenza* e *incidenza* della performance negativa nella popolazione studiata (all'aumentare della proporzione di casi con performance negative nel campione osservato, aumenta la proporzione di tali casi positivi classificati correttamente).

5.4.1.2 Identificazione dei migliori valori-soglia

Per superare le difficoltà viste e per esplorare la relazione tra la sensibilità e la specificità rispetto a diversi valori-soglia, al fine di individuare la migliore, è possibile utilizzare l'approccio noto come **ROC analysis** (*Receiver Operating Characteristic* o *Relative Operating Characteristic*)[5]. Tale approccio è largamente utilizzato in medicina, radiologia e psicologia, è stata recentemente introdotta in altri campi (*data mining*) e risulta interessante da utilizzare nell'ambito dell'analisi della capacità selettiva e discriminante di un indicatore sintetico per la possibilità che dà di studiare ed analizzare la relazione tra sensibilità e specificità per l'individuazione di valori-soglia (*cut-point* o *cut-off* o *operating-point*) realmente discriminanti.
Vediamo attraverso quali fasi procede tale analisi.

a. Costruzione della curva ROC
L'analisi ROC viene effettuata attraverso lo studio del rapporto e della funzione che lega

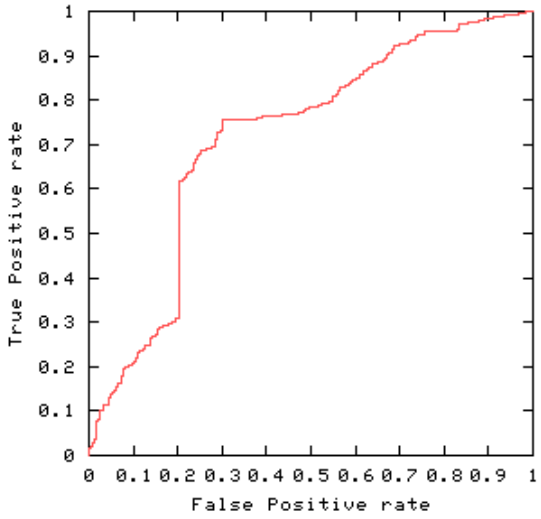
- probabilità di ottenere un risultato "allarme-vero" nel gruppo dei casi che necessitano di intervento (→ sensibilità → *hit rate* → **HR**),
- probabilità di ottenere un risultato "allarme-falso" nel gruppo dei casi che non necessitano di intervento (→ 1-specificità → *false alarm rate* → **FAR**).

Per studiare tale relazione si calcolano i due valori *rate* per ciascun *valore-soglia.* Definendo molti *valori-soglia* lungo tutto il continuum dei valori dell'indicatore composito è possibile ottenere una curva con risoluzione ottimale compatibilmente con il di dati disponibili.

[5] Il nome di tale procedimento deriva dal fatto che esso è nato durante la Seconda Guerra Mondiale per studiare e migliorare la ricezione dei radar e dei sonar. Peterson, W. W., Birdsall, T. G., and Fox, W. C. (1954). *The theory of signal detectability*. Institute of Radio Engineers Transactions, PGIT-4, 171–212.

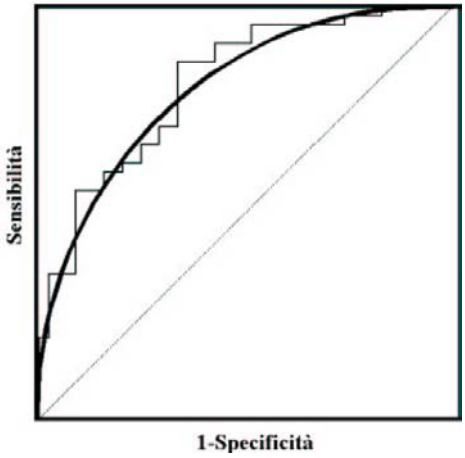
Si procede a questo punto alla costruzione di un grafico che riporta sull'asse
- *X* i valori **FAR**,
- *Y* i valori **HR**.

L'unione dei punti ottenuti riportando nel piano cartesiano ciascuna coppia genera una curva spezzata con andamento a scaletta (ROC *plot*).
Tale curva è sempre posizionata, come nell'esempio qui accanto, tra due punti



- (0,0) dove tutti sono classificati negativi (nessun allarme); rappresenta il punto in cui l'indicatore classifica tutti negativi (anche i positivi);
- (1,1) dove tutti sono classificati positivi; rappresenta il punto in cui l'indicatore classifica tutti positivi (anche i negativi).

Attraverso l'interpolazione, è possibile eliminare la "scalettatura" (*smoothing*) ed ottenere una curva (ROC *curve*) che rappresenta una stima basata sui dati osservati.
Accanto è rappresentata una curva ROC prima e dopo l'interpolazione.

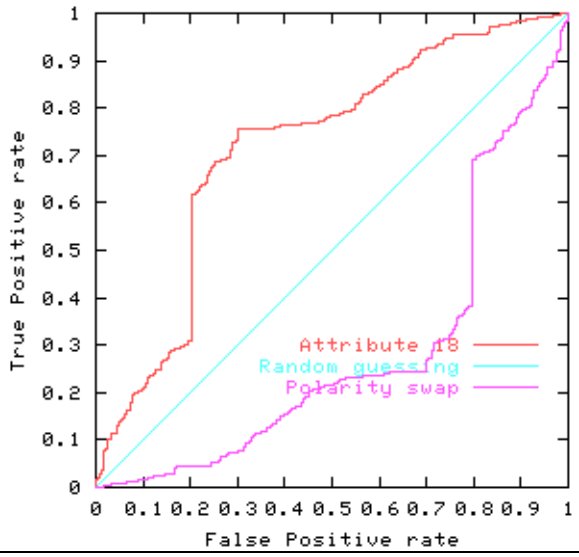


b. Valutazione della capacità discriminante

La capacità discriminante di un indicatore è valutata osservando
- la posizione della curva proporzionale e
- l'estensione dell'area sottesa alla curva ROC detta *Area Under Curve* (AUC).

L'interpretazione del valore di AUC può avvenire tenendo presente alcuni criteri adottati empiricamente che classificano i risultati secondo il seguente schema:
- AUC = 0.5 → indicatore non selettivo (nessuna capacità discriminante)
- $0.5 < AUC \leq 0.7$ → indicatore poco accurato
- $0.7 < AUC \leq 0.9$ → indicatore moderatamente accurato
- $0.9 < AUC < 1.0$ → indicatore altamente accurato
- AUC = 1.0 → indicatore perfettamente discriminante.

Quindi:



capacità discriminante		
buona	0.50 > AUC< 1.00	Perché un indicatore possa essere considerato discriminante è necessario che la corrispondente curva ROC sia situata al di sopra della *chance line* (**curva rossa**).
perfetta	AUC = 1.00	La AUC passa attraverso le coordinate {0;1} ed il suo valore corrisponde all'area dell'intero quadrato delimitato dai punti di coordinate (0,0), (0,1), (1,0) (1,1), che corrisponde ad una probabilità del 100% di una corretta classificazione. Si noti che, in tale caso limite i valori predittivi non dipendono più dalla prevalenza.
nulla (classificazione casuale)	AUC = 0.50	La AUC passa attraverso le coordinate (0;0) e (1,1). Tale retta (*chance line*) riflette i valori HR e FAR di un gruppo di valori-soglia senza alcun potere discriminante (**linea celeste**). In pratica tale curva si riferisce ad un indicatore che discrimina a caso.
inversa (classificazione peggiore del caso)		E' possibile che un indicatore produca una curva ROC che si situi al di sotto della *chance line*. Tale situazione rileva un indicatore negativamente correlato con la classificazione corretta (**curva fucsia**). In questo caso è necessario invertire la polarità dell'indicatore; ciò produrrà una rotazione della curva ROC.

L'area sottesa ad una curva ROC rappresenta un parametro fondamentale per la valutazione della *capacità selettiva* di un indicatore sintetico, in quanto costituisce una misura di accuratezza non dipendente dalla prevalenza ("*pure accuracy*").
Poiché AUC rappresenta una stima da popolazione campionaria finita, risulta quasi sempre necessario testare la *significatività della capacità discriminante* osservata, ovvero se l'area sotto la curva eccede significativamente il suo valore atteso di 0.5.
Tale procedura corrisponde a verificare se la proporzione dei veri positivi è superiore a quella dei falsi positivi. AUC può essere considerata una variabile normale, per cui si può costruire un test *z* nella seguente maniera:

$$z = \frac{AUC - 0.5}{\sqrt{\sigma^2}}$$

dove σ^2 rappresenta la varianza di AUC.

Se, ad esempio, il valore di *z* eccede il valore critico di 1.96, si può affermare che il test diagnostico presenta una *performance* significativamente superiore a quella di un test non discriminante, con $p < 0.05$. Se il test *z* risulta invece significativamente inferiore (curva ROC al di sotto della *chance line*), occorre invertire il criterio di classificazione, in quanto il marcatore evidenziato dal test presenta valori mediamente più elevati nella popolazione di cloro che registrano performance positive (evenienza di difficile riscontro).

c. Stima dell'*Area Under Curve* (AUC)

E' possibile stimare l'AUC ottenuta da un campione finito semplicemente connettendo i diversi punti del ROC *plot* all'asse delle ascisse con segmenti verticali e sommando le aree dei risultanti poligoni generati nella zona sottostante ("regola trapezoidale"). Tale approccio però può produrre risultati sistematicamente distorti per difetto.

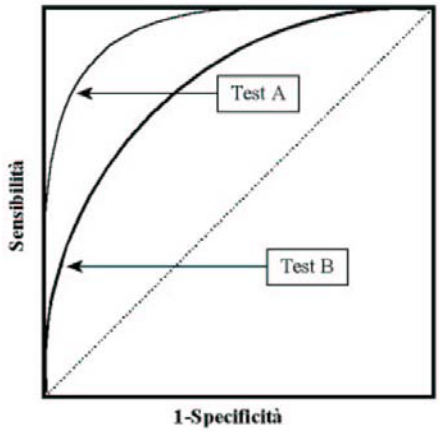
In questa sede non interessa analizzare i diversi metodi di stima dell'area "vera" e di interpolazione delle curve ROC; nella pratica per la stima della AUC è comunque possibile utilizzare i diversi algoritmi che diversi *package* statistici mettono a disposizione (R, SAS, SPSS, SYSTAT ecc.) oppure di altro software specifico per la valutazione delle curve ROC.

Confronto della capacità discriminante tra due indicatori

Il metodo più semplice per confrontare la capacità discriminante di due indicatori – assumendo una ipotesi bi-normale (*curve ROC proprie*) – è quello di valutare la differenza tra le rispettive aree sottese la curva ROC ovvero di confrontare le *accuracy* stimate mediante l'area sottesa alle corrispondenti curve RO.

figura qui accanto si mettono a confronto due indicatori (chiamati rispettivamente "test A" e "test B") mediante analisi ROC.

Risulta evidente la superiorità del test A la cui curva ROC teorica si trova interamente al di sopra di quella corrispondente al test B.



E' possibile applicare un test *z* per confrontare le due curve ROC indipendenti ovvero rapportando la differenza delle due aree all'errore standard di tale differenza.

Nel caso di indipendenza dei due indicatori, tale parametro viene facilmente stimato dalla radice quadrata della somma della varianza di ogni area:

$$z = \frac{AUC_1 - AUC_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

Nel caso i due indicatori non siano indipendenti (ovvero sono applicati sugli stessi casi), l'errore standard della differenza delle due aree viene a dipendere dalla correlazione esistente tra esse:

$$z = \frac{AUC_1 - AUC_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2r\sigma_1\sigma_2}}$$

Per la stima di *r* si può procedere calcolando prima il coefficiente di correlazione (o di cograduazione per dati ordinali) tra i due indicatori, separatamente per i due gruppi dei casi ("necessitano intervento", "non necessitano intervento") e poi la media dei due valori di correlazione precedentemente calcolati.

Individuazione dei valori-soglia ottimaliLa curva ROC può essere utilizzata per scegliere il migliore valore-soglia che corrisponde a quel valore che rappresenta il migliore compromesso (*trade off*) tra i costi prodotti

- dal fallire nell'identificare i positivi,
- dall'identificazione dei falsi allarmi.

In genere tali costi si assumono uguali ovvero si attribuisce la stessa importanza alla *Se* e alla *Sp* anche se è possibile differenziare il loro peso.

In generale, il punto sulla curva ROC più vicino all'angolo superiore sinistro rappresenta il miglior compromesso fra sensibilità e specificità.

Se il costo dell'errore di classificazione rappresenta semplicemente una somma dei "falsi positivi" e dei "falsi negativi", allora tutti i punti saranno posizionati su una linea retta ed avranno lo stesso costo. In questo caso se i due errori si si verificano con la stessa frequenza, la retta avrà una pendenza di 1 (= 45 gradi).
Il punto della ROC in cui identificare il migliore valore soglia è quello che si trova su una retta con pendenza 1 e il più vicino possibile all'angolo "0,1" (nord-ovest). E' possibile definire i seguenti valori:
alpha costo dei falsi positivi (*false alarm*)
beta costo dei positivi persi (*false negative*)
p proporzione dei casi positivi.
La media del costo di classificazione atteso al punto x,y nello spazio ROC è determinata nel modo seguente:

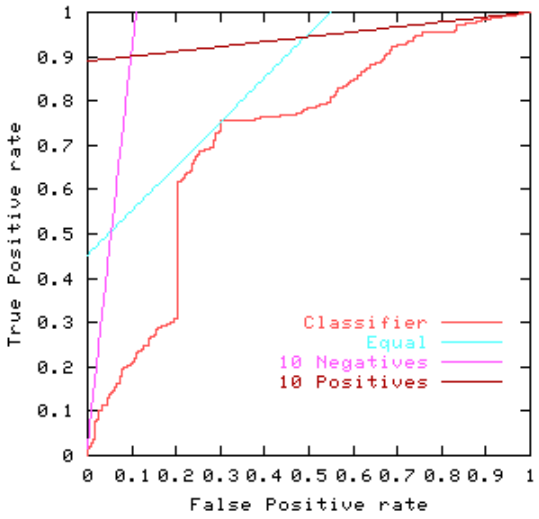
$$C = (1-p)alpha, x + p, beta(1-y)$$

Le linee con uguali costi, dette *isocost line*, sono parallele e rette. Il loro gradiente dipende da $alpha/beta$ e $(1-p)/p$. Se i costi sono uguali ($(alpha = beta)$ e la proporzione dei casi positivi è 50% $(p = 0.5)$, il gradiente è 1 e le *isocost line* sono a 45 gradi.
Di seguito sono presentati tre esempi.

Linea celeste: rappresenta una *isocost line*: i costi degli errori di classificazione dei casi positivi e negativi sono uguali.

Linea fucsia: corrisponde alla situazione in cui i costi dati dalla perdita dei casi negativi superano i costi dati dalla perdita dei casi negativi con un rapporto di 10 a 1 ovvero $p = 0.5$, $alpha = 1$ e $beta = 10$.

Linea marrone: corrisponde alla condizione operativa migliore in cui i costi di perdere i casi positivi superano dieci volte il costo di avere falsi allarmi.
In altre parole questa è la situazione in cui è più importante mantenere un alto tasso di veri positivi e in cui i casi negativi hanno poco impatto sui costi totali (per esempio, $p = 0.91$, $alpha = 1$ e $beta = 1$).



L'esempio dimostra la tendenza naturale della ROC ad operare vicino agli estremi nel caso in cui o i due tipi di costi molto diversi tra loro o nei casi in cui i casi da classificare sono molto influenzati da un tipo di costo a scapito dell'altro. Comunque in questi casi estremi vi saranno comparativamente pochi dati. Ciò rende il calcolo della ROC più soggetto a fluttuazioni statistiche: se si vuole raggiungere lo stesso livello di significatività statistica occorrono più dati.

ROC e tasso di errore

A partire da una ROC è possibile ricavare il tasso di errore (*error rate*, *ER*); ponendo i costi di errore di classificazione uguali tra loro e unitari

$$alpha = beta = 1$$

allora

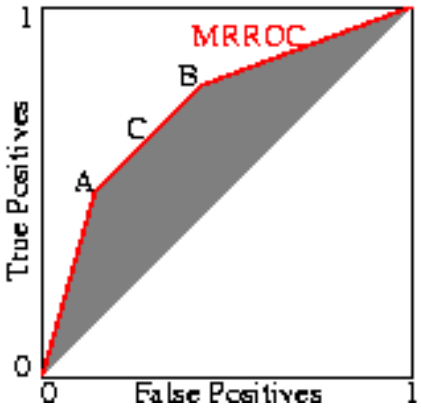
$$ER = (1-p)x + p(1-y)$$

I punti che compongono la ROC con tasso uguale di errore rappresentano linee rette. Nel semplice caso in cui si osserva un numero di casi positivi uguale a quello di casi negativi, le linee si posizionano a 45 gradi (parallelamente alla *chance line*).

"Maximum Realisable" ROC (metodo **Scott**)[6]

Tale metodo consente di individuare da due valori-soglia un terzo (composito) le cui performance (in termini di ROC) sono descritte da una linea che collega le performance delle due di partenza. Vediamo un esempio.

Dopo aver individuato a caso due punti, A e B, si vuole determinare un valore-soglia il cui tasso di falsi positivi rispetto al tasso di veri positivi giaccia su una linea che sia a metà strada tra i punti A e B. Il nuovo punto darà casualmente per metà delle volte la risposta data da A e per l'altra metà la risposta data da B.
La cosiddetta "*Maximum Realisable ROC*" è rappresentata dalla linea rossa.
Procedendo iterativamente sarà possibile individuare il punto migliore di discriminazione.



Notare che in una curva ROC esistono in genere due segmenti di scarsa o nulla importanza ai fini della valutazione della capacità discriminante. Essi sono rappresentati dalle frazioni di curva sovrapposte all'asse delle ascisse e all'asse delle ordinate. Infatti, i corrispondenti valori possono essere scartati in quanto esistono altri valori di *cut-off* che forniscono una migliore *Sp* senza perdita di *Se* o, viceversa, una migliore *Se* senza perdita di *Sp*.

Nella pratica la selezione del *cut off ottimale* rappresenta una decisione molto più complessa che deve tener conto sia della situazione in cui verrà utilizzato l'indicatore sintetico che dell'esame comparativo delle conseguenze pratiche derivanti dall'ottenimento di risultati allarme-falso e non-allarme-falso in quella particolare situazione contingente.

[6] Scott M.J.J., M. Niranjan, R.W. Prager, *Realisable Classifiers: Improving Operating Performance on Variable Cost Problems*, Cambridge University Department of Engineering, http://www.bmva.ac.uk/bmvc/1998/pdf/p082.pdf