04.04.07
uso dei modelli in statistica
abbiamo detto k le procedure k sono state descritte fanno riferimento a dei modelli statistici, ovvero a una rappresentazione del fenomeno secondo certe caratteristiche che è in grado da un lato di prevedere quello che accade nel fenomeno stesso e dall'altro lato di rappresentare con un certo margine di errore le caratteristiche peculiari di quel fenomeno.
Quindi vi dicevo che i modelli sono un'altra rappresentazione del fenomeno rappresentazione che ha degli scopi principali. Quello di essere maggiormente fedele al fenomeno stesso e quindi di avere meno margine di errore in questa soluzione rappresentativa e quello di prevedere quindi il comportamento del fenomeno. L'aspetto primario che ho appena descritto, cioè quello della previsione, è l'aspetto peculiare dei modelli a struttura predeterminata. Attraverso questi modelli in effetti è possibile valutare attraverso la significatività il margine di errore che ci permette di utilizzare le variabili che abbiamo considerato del fenomeno per prevedere il comportamento della variabile dipendente.
Modelli come generalizzazione delle procedure: GLM e regressione logistica

Invece l'aspetto peculiare di altri tipi di modelli descrittivi, esplorativi, abbiamo visto l'analisi fattoriale, la cluster analisi e l'analisi discriminante, è quello di essere il + vicino possibile, il + adeguato possibile nella rappresentazione delle caratteristiche del fenomeno stesso.
Non sono gli unici modelli statistici che avrete modo di incontrare nei vostri studi di psicologia, ma sono quelli che + frequentemente compaiono negli studi psicologici.
Per quello che riguarda i modelli a struttura predeterminata, cioè quei modelli che vogliono prevedere il comportamento di una certa variabile dipendente, ci siamo occupati di modelli non parametrici x' quelli parametrici sono tutti riconducibili al GLM che avete già affrontato nel 1° anno di corso. In particolare abbiamo analizzato cosa succede in alcuni modelli che studiano variabili dipendenti non gaussiane; si occupano di var categoriche e quindi che possono avere solo alcuni determinati valori e tali valori sono rappresentativi di categorie e il + delle volte non sono nemmeno ordinabili secondo una gerarchia.
Ciò che distingue i modelli a struttura predeterminata dai modelli a struttura stimata è proprio la presenza nei 1° della significatività. Il fatto che questi modelli mirano tutti a prevedere il comportamento della variabile dipendente ci fa soffermare anche su di un'altra caratteristica dei modelli stessi ovvero il fatto di essere generalizzabili, cioè di utilizzare il campione per fare riferimento ad 1 popolazione generale. I risultati, quindi le conclusioni che provengono da questo tipo di modelli a struttura predeterminata sono risultati che fanno riferimento alla statistica inferenziale e quindi che permettono di dimostrare delle ipotesi sperimentalmente costruite.
Altra caratteristica è proprio il fatto che l'ipotesi di partenza di questo tipo di modelli è posta a priori e quindi si suppone che la relazione tra le variabili che il modello descrive, sia di un certo tipo PRIMA di raccogliere i dati e si scommette attraverso la significatività su questa relazione. La significatività del risultato finale del test permette di accettare la scommessa o di rifiutarla perché permette di dire che la mia ipotesi sperimentale è accettabile per un margine di errore del 5% oppure considera l'ipotesi sperimentale come non accettabile perché questo limite del 5% è stato superato.

Esempi di domande:
Caratteristiche e differenze tra i modelli descrittivi a struttura predeterminata e i modelli inferenziali a struttura stimata.

1. i modelli a struttura predeterminata sono generalizzati, mentre a struttura stimata si riferiscono al campione preso in esame e solo a quel campione
2. i modelli a struttura predeterminata sono basati su ipotesi a priori riguardo alla struttura e quindi alle relazioni tra le variabili che vanno ad indagare, mentre nei modelli a struttura

stimata le relazioni fra le variabili vengono via via costruite e decise in base al modello stesso

3. nei modelli a struttura predeterminata c'è un calcolo che indica la probabilità di errore dimostrando il modello come adeguato, mentre nei modelli a struttura stimata questo tipo di calcolo viene usato diciamo così al contrario, cioè c'è tutta una serie di procedure per vedere se il modello è adeguato in base al fatto che sia uguale ai dati sperimentali. Ed effettivamente i modelli a struttura stimata hanno come ulteriore caratteristica che non hanno il calcolo della significatività e quindi dell'oscillazione casuale dei risultati che ottengono.
4. i modelli a struttura predeterminata sono utili per fare inferenza, mentre quelli a struttura stimata sono semplicemente descrittivi ed esplorativi rispetto ai dati. Sono due strumenti diversi, con scopi diversi e caratteristiche diverse per indagare il fenomeno in studio. Abbiamo fatto l'esempio della differenza tra regressione lineare e l'analisi fattoriale ed abbiamo visto che l'analisi fattoriale sullo stesso campione produce risultati diversi a seconda (l'abbiamo visto anche per la cluster analisi, o per l'analisi discriminante) quindi vengono prodotti risultati diversi a seconda delle variabili che vengono prese in considerazione. Questo perché proprio quando tecniche descrittive vogliono dare una prospettiva, una delle tante, di come si comporta quel fenomeno in quel determinato campione. Quindi i risultati ottenuti con modelli a struttura predeterminata sono comunque riferibili solo a quel campione con quelle caratteristiche.
5. la variabilità del numero dei parametri non viene mai fornita nei modelli a struttura stimata e questo è collegato al fatto che non c'è mai una stima dell'oscillazione casuale dei risultati che si ottengono.

Caratteristiche e aspetti dei modelli non parametrici (in realtà le diapositive devono servire + come punto di riferimento, non viene chiesto tutti i tipi affrontati a lezione dei test non parametrici, non serve che li sappiate tutti a memoria, ma dovete sapere che fra i modelli non parametrici quelli + importanti sono le tavole di contingenza, i modelli log lineari e la regressione logistica. La differenza tra questi 3 tipi di modello è che mentre i primi due sono misure di associazione, quindi sono in grado di vedere l'associazione di variabili a 2 a 2, i modelli regressivi della regressione logistica (ci sono diversi tipi di modelli regressivi: binomiale, multinomiale, ecc) sono molto simili al GLM ovvero non ci dicono semplicemente se le variabili che andiamo a studiare sono legate fra loro, ma sono strutturati in modo di vedere la variabile dipendente in funzione di alcuni fattori, e quindi di vedere le variabili in un rapporto casuale fra di loro. Le variabili indipendenti, i fattori, sono quelle che dovrebbero, secondo il nostro modello inferenziale creare delle modifiche sulla variabile dipendente. Se il nostro modello è adeguato, allora la relazione che viene spiegata dalle variabili indip, dai fattori, dovrebbe giustificare la funzione che abbiamo utilizzato, quindi la regressione logistica è in grado di studiare l'effetto di alcune variabili indipendenti sulla variabile dipendente. Quindi non si parla più di associazione, ma si parla di relazione causale fra le variabili. Gli aspetti peculiari della regressione logistica sono:

1. innanzitutto la struttura del modello di riferimento, che ha una struttura logistica ed è quindi basata sul calcolo dei logaritmi; utilizza questa struttura logaritmica per poter linearizzare la relazione fra i fattori e quindi per rendere lineare la funzione che va poi a spiegare il comportamento della variabile dipendente.
2. Questa trasformazione viene attuata perchè attraverso ciò riesco a quantificare anche la variabile dipendente, che abbiamo detto deve essere una variabile qualitativa dicotomica. La variabile qualitativa dicotomica, in quanto tale, per le proprie caratteristiche non sarebbe quantificabile, ma con la trasformazione logaritmica sì.
3. la trasformazione logaritmica, chiaramente, non viene fatta sui valori grezzi della variabile, ma sulle probabilità associate a tali valori, o, ancora meglio, sull'odds dei valori ovvero sul rapporto tra la probabilita di un evento fratto la probabilità di non evento.

4. La funzione che lega questa variabile e quind la struttura del modello di riferimento è di tipo logaritmico. Quindi il logaritmo della variabile dipendente, altrimenti detto logit è visto attraverso una funzione che è in grado di vedere i fattori come combinati linearmente fra di loro, quindi l'effetto di ciascun fattore viene sommato all'effetto degli altri fattori in questo modo.
5. Ciascun fattore, ciascuna variabile dipendente viene spiegata nella funzione come un parametro b moltiplic per un coefficiente x e tutte le variabili, quindi tutti i fattori siano essi qualitativi o quantitivi vengono resi dicotomici dalla funzione. Nel caso in cui la variabile sia qualitativa dicotomica il fattore spiega solo 1 dei 2 valori, l'altro valore viene spiegato dalla costante; nel caso in cui invece il fattore sia quantitativo si assume come costante la media del fattore quantitativo, e invece il parametro b di quel fattore è relativo all'aumento di 1 unità rispetto alla media. Quindi il valore basale si riferisce all'aumento di1 unità del fattore quantitativo.

In questo caso, invece, vengono fatte vedere tutte le trasformazioni della variabile dipendente in logaritmo dell'odds. La variabile dipendente dicotomica, per fare un esempio: se volete studiare l'influenza del sesso sulla presenza o meno di una patologia, se io penso che essere F sia più influente dell'essere M nell'aumento della probabilità di una patologia, allora il mio fattore relativo al sesso di appartenenza, sarà relativo all'essere F, mentre il valore basale è relativo all'essere M. ciò si traduce nel risultato del b, quindi il b mi dice quanto l'essere F cambia la probabilità di avere la patologia rispetto alla probabilità della non patologia. Quindi il parametro b è riferito alla condizione della variabile qualitativa che noi poniamo come decisiva nel cambiamento. Se attribuiamo alle F il valore 1 e ai M il valore 0, automaticamente spss calcola la probabilità relativa ai soggetti che noi abbiamo posto come 1. se noi mettiamo alle F 1 significa che noi pensiamo che quel valore del fattore genere sia determinante nelle modifiche della probabilità dell'evento che andiamo a studiare e quindi della variabile che andiamo a studiare. Poniamo invece il valore 0 dei M nella costante e quindi a partire dal valore della costante, quanto l'essere F aumenta la probabilità di avere la malattia o di non averla? E questo è il senso del parametro b e della variabile qualitativa.

Nel caso, invece, in cui la variabile qualitativa non sia dicotomica, la variabile viene divisa in tante sottovariabili dicotomiche, quindi viene dicotomizzata suddividendola in tante sottovariabili, se ad es ha 3 valori possibili, chiaram queste sottovariabili dicotomiche rappresentano la combinazione tra i valori, quindi nella prima variabile avremo il primo valore come 0, il secondo come 1 e il terzo valore viene posto invece nella costante. L'altro sottotipo di questa variabile pone come primo valore 1 e il secondo valore come 0 e il terzo valore è ancora una volta posto nella costante. Quindi 1 variabile a 3 valori viene suddivisa in 2 sottovariabili, nella prima variabile il 1° valore è valutato 1, il 2° 0 e il 3° 0, nella seconda variabile il 1° valore è valutato 0, il secondo 1 e il terzo sempre 0, non viene presa in considerazione la terza variabile perché essa viene comunque sempre assunta come costante (b0). Quindi l'informazione che viene portata dall'ultimo valore viene sempre considerato nella costante, quindi è 0. Quindi quando la variabile quantitativa ha + di 2 valori viene suddivisa in un numero n di variabili dicotomiche, quindi per poter interpretare i risultati si crea una tabella che mostra come vengono ricodificate le variabili.

Gli aspetti sui logaritmi sono assolutamente esplicativi e non vi verranno chiesti. Ma vi potrebbe sicuramente essere chiesto: alla fine della regressione logistica si ha un risultato che indica che quel fattore è influente sulla variabile dipendente e anche il senso di questa influenza. Da che cosa è dato questo senso? Quindi: qual è l'indice nella regressione logistica che dice quanto quella particolare (Regressione logistica: x poter utilizz… questa slide spiega la …) condizione del fattore è determinante sul valore della variabile dipendente. Questa risposta l'abbiamo nell'odds ratio detto anche esponenziale di b, quindi l'esponenziale relativo al fattore b che andiamo ad analizzare è dato dal rapporto tra l'odds del parametro b della condizione 1 (nell'esempio di prima l'odds dell'avere la malattia essendo F risp all'odds, sempre del genere, b1 essendo M). Nel caso in cui questo rapporto sia = 1 ci troviamo nella condizione dell' H0, perché quando l'odds è 01 significa che non

c’è nessun cambiamento nella probabilità di avere la malattia sulla probabilità di non averla essendo M rispetto alla probabilità di avere la malattia sulla probabilità di non averla essendo F.
Dopo di che la regressione logistica è caratterizzata come procedura da tutta una serie di tests che vanno ad analizzare la bontà del modello. In questo caso la bontà del modello viene studiata ancora 1 volta attraverso una dimostrazione di somiglianza tra il modello ed i dati sperimentali. Se a noi interessa la semplice relazione tra la variabile dipendente e i fattori, tutto questo aspetto sulla bontà del modello diventa marginale, perché quello che ci interessa di più vedere nella regressione logistica è se i parametri b che siamo andati ad indagare, sono significativi e in che senso lo sono. Quindi tutta questa parte sulla valutazione della bontà del modello è + simile agli aspetti descrittivi dei modelli, perché effettivamente viene valutata tramite dei parametri come il goodness of fit o la statistica Wald e così via, che sono tutti riferiti alla probabilità che il modello sia adeguato rispetto ai dati sperimentali.
Come viene valutata la bontà del modello?
Ci sono diversi parametri, il goodness of fit è uno di questi, c’è anche lo pseudo erre squared oppure la statistica Wald, che sono però parametri che hanno più un valore descrittivo rispetto al modello. L’aspetto invece inferenziale è dato dalla significatività dei parametri b. nel caso in cui i parametri non siano significativi vuol dire che la loro influenza sulla variabile dipendente non è abbastanza forte da provocare su di essa un cambiamento sufficientemente alto rispetto al cambiamento dovuto al caso.
Quindi come vengono calcolati i parametri?
Abbiamo anche parlato dei vari metodi di approssimazione che utilizza la regressione logistica, come tutti i modelli non ha come obiettivo quello di essere totalmente affidabile rispetto ai dati sperimentali, quindi assolutamente speculare rispetto ai dati sperimentali, ma vuole creare un modello che sia in grado di rappresentare questi dati in modo sintetico, attraverso appunto la relazione tra le variabili, e quindi per questo vengono utilizzati dei metodi approssimativi, cioè che via via, attraverso un numero di operazioni, scelgono il modello migliore.
Quali sono questi metodi di approssimazione?
Si può procedere attraverso un metodo a blocchi nel caso in cui consideri l’apporto alla significatività dovuto da tutte le variabili insieme. Quindi a blocchi considera tutte le variabili insieme quanto ciascuna di queste variabili contribuisce ad aumentare la significatività del modello. Se ci sono delle variabili che non sono influenti rispetto alla bontà del modello allora vengono escluse.
Nel metodo stepway o a passi, si può procedere in due modi: o in avanti o all’indietro. Quindi partendo in avanti considerando una variabile alla volta all’interno del modello, valutando poi se questa, cioè la forza di questa variabile che è stata inserita nel modello aumenta la bontà del modello oppure no. Nel caso in cui non la aumenti questa viene esclusa e si passa alla variabile successiva. Quindi le variabili vengono inserite 1 alla volta a seconda del loro apporto, della loro modifica nella bontà del modello. Nel metodo all’indietro, viceversa, si inseriscono tutte le variabili e poi se ne toglie 1 alla volta e si verifica man mano se togliendo una variabile la bontà del modello cambia significativamente oppure no.
Chiaramente tutti questi metodi di approssimazione dovrebbero essere scelti a priori dallo sperimentatore. Lo sperimentatore individua i fattori, individua la variabile dipendente e, non a caso, va a scegliere il metodo di approssimazione, ma a priori dovrebbe dire “voglio identificare l’effetto delle variabili indipendenti, dei fattori, tutti insieme, ad es. nel metodo a blocchi, sulla variabile dipendente”, oppure “voglio verificare una su quell’altra l’apporto di ciascuna variabile indipendente, di ciascun fattore, sulla variabile dipendente. Quindi il processo, il metodo, di approssimazione dovrebbe essere scelto a priori.
Può essere anche chiesto di farvi raccontare un esempio di ricerca, (ovviamente se portate la tesina e avete fatto la prima prova, difficilmente vi verrà chiesto un esempio, ma se non avete superato la prima prova potrebbe capitare) quindi comunque, visto che i libri sono densi, specialmente su questa 2° parte, di ricerche, cercate di pensare praticamente a come esemplificare i concetti,

potrebbe anche esservi chiesto di fare un esempio di regressione logistica, cioè ad es. ipotesi di ricerca nella quale può essere applicata la regressione logistica.

ANALISI LOG LINEARE

L'analisi log lineare è una misura di associazione e quindi ha come scopo quello di studiare la relazione tra + variabili quantitative che possono avere anche + di 2 valori, non sono + distinte, le variabili, in dipendente e indipendente perchè abbiamo detto la frequenza delle celle, che rappresenta la relazione fra quelle variabili, diventa la variabile dipendente, mentre i fattori sono considerate tutte le variabili prese singolarmente. La struttura del modello dell'analisi log lineare ancora 1 volta trasforma la frequenza di cella (la relazione fra le variabili è data dalla frequenza di cella) nel suo logaritmo. Questo per vederlo in funzione di alcuni parametri combinati ancora una volta tra loro in modo lineare quindi sommati fra loro.

Il punto di partenza per questa trasformazione è 1 tavola di contingenza multidimensionale nella quale ciascuna cella rappresenta la relazione tra le variabili studiate. Quindi ad es la cella 120 rappresenta l'incrocio tra tipo di personalità, tipo di terapia ed esito della terapia. In particolare questa frequenza di cella a seconda che la guardi rispetto al totale di colonna o rispetto al totale di riga rappresenterà una relazione oppure un'altra. Se la guardo rispetto al totale di colonna, questa relazione sarà riferita al tipo di personalità rispetto alla terapia, se invece la guardo rispetto al totale di riga sarà riferita all'esito della terapia rispetto al tipo di terapia. In ogni modo la frequenza di cella viene vista sempre in relazione ai totali a cui si riferisce.

Qual è lo scopo dell'analisi log lineare? Misurare l'associazione tra le variabili

Qual è la funzione che viene utilizzata come modello di riferimento nell'analisi log lineare?

È una funzione logaritmica che trasforma la frequenza di cella nel suo logaritmo data quella combinazione lineare di fattori che vengono presi in esame

Come può essere strutturato il modello? quali tipi di modello possono essere studiati?

Ci sono diversi modelli nell'analisi log lineare. Si parla di modello gerarchico quando a termini di ordine superiore vengono accostati termini di ordine inferiore.

Nel caso in cui siano analizzate dal modello tutte le possibili combinazioni di fattori allora ci troviamo nel caso del modello saturo. Chiaramente, il modello saturo, rappresentando tutte le possibili relazioni tra le variabili, sarà perfettamente adeguato rispetto ai dati sperimentali. Difficilmente è interessante il modello saturo, + spesso sarà interessante vedere attraverso il minimo numero di combinazioni possibili se queste effettivamente spiegano il comportamento dei dati sperimentali oppure no. E quindi + spesso siamo interessati a verificare la bontà del modello non saturato dove solo alcune combinazioni tra le variabili sono considerate e non tutte.

Quindi il modello gerarchico è un modello che può essere saturo o non saturato. Attraverso quali parametri viene valutata la bontà del modello?

Abbiamo visto che si può valutare tramite la stima dei parametri, quindi attraverso il calcolo, in questo caso non si parla di b, ma si parla di lambda, però comunque si riferisce ai parametri che caratterizzano quel fattore e si confrontano quindi, per valutare, per stimare questi parametri, si confrontano quindi le frequenze attese, quelle create dal modello, rispetto alle frequenze osservate. (Come viene valutata la discrepanza?)Tale differenza tra le frequenze attese create dal modello e le frequenze osservate, produce i residui che ci danno la stima di quanto il modello sia adeguato rispetto ai dati sperimentali. Sui residui vengono calcolati lambda. Quindi già dalla valutazione dei parametri lambda ho un'idea di quanto il modello sia adeguato a rappresentare i dati. Se tutti i parametri lambda, quindi se tutti i fattori, sono significativi significa che il modello è buono. Se invece la maggior parte dei fattori non è significativa probabilmente significa che il modello che abbiamo utilizzato non è adeguato rispetto ai dati sperimentali.

C'è tutta una serie di tests per valutare la bontà del modello che però ancora una volta si basano sulla non falsificaz dell'H0 in modo simile a quelli che abbiamo visto nella regressione logistica, e quindi il goodness of fit test e il likelyhood ratio test sono più indirizzati a verificare la probabilità che il modello sia legato ai dati sperimentali. Quindi anche in questo caso sono più dei valori di tipo descrittivo. Non sono dimostrativi.

Un altro punto importante è che dopo la trasformazione logaritmica delle frequenze, le frequenze di cella e tutti i parametri vengono trasformati in punti z. questo perché siano giustamente confrontabili tra loro e perché sia più semplice il processo di falsificazione dell'ipotesi nulla.
Oltre a questi 2 modelli non parametrici abbiamo poi affrontato ed analizzato altre 2 tecniche descrittive che però fanno parte dei modelli a struttura stimata il cui scopo è dunque + descrittivo. La cluster analysis e l'analisi discriminante .
Pima però riguardiamo l'Analisi fattoriale: gli aspetti + importanti che possono essere oggetto d'esame riguardano innanzitutto lo scopo: quali sono gli scopi dell'analisi fattoriale? E quindi che cosa significa trovare i fattori, cosa rappresentano i fattori e cosa sono i fattori nell'analisi fattoriale. Quali sono i metodi per creare questi fattori? Quindi il metodo della massima verosimiglianza e il metodo delle componenti principali.
Quali sono i limiti dell'analisi fattoriale?
Anche gli assunti e quindi:
Da che tipo di variabili si assume e quindi si parte? Solo variabili gaussiane.
Tra i limiti→Problematicità delle sorgenti delle variabili: non sono il punto di partenza delle variabili ma sono i fattori stessi che sono creati matematicamente in modo tale da spiegare la variabilità delle variabili. Quindi il fatto di chiamarle sorgenti potrebbe essere ingannevole e potrebbe far pensare che attraverso questi fattori si spiega la totalità delle relazioni fra le variabili. Non si spiega
→Problematicità di assegnazione del significato a queste variabili. Può essere dato, e viene fatto abitualmente, ma è comunque un significato arbitrario e quindi comunque l'analisi fattoriale per come è strutturata rimane una tecnica solitamente esplorativa o descrittiva, sebbene ci siano dei tentativi di trasformarla in un'analisi invece di tipo inferenziale. Questo è un punto di vista. Alcuni autori invece sostengono che nonostante l'analisi fattoriale sia una tecnica descrittiva sia possibile attraverso opportune modifiche, trasformarla in tecnica inferenziale, il polo di BO non lo condivide, per loro l'analisi fattoriale va usata solamente come tecnica descrittiva.
Il vantaggio di questa tecnica, in quanto tecnica descrittiva, è quello di darci una visione sintetica delle relazioni tra le variabili che si stanno studiando e quindi di ridurre l'informazione in modo tale da non perderne troppa e quindi i fattori sono in un certo senso, ulteriori variabili create matematicamente sulla base delle relazioni tra le variabili di partenza e ci permettono di riassumere l'informazione di tutte le variabili di partenza, ancora una volta perdendo una certa quantità. Un altro vantaggio della tecnica descrittiva è che in questo modo troviamo sempre un numero n di fattori che viene creato dal modello, ma non sappiamo quanto possa oscillare questo numero e quindi è una descrizione che ha una stima dell'errore, dell'oscillazione casuale del risultato trovato.
Qual è il punto di partenza dell'analisi fattoriale? Il punto di partenza è la matrice di correlazione tra le variabili.
Da questa matrice di correlazione vengono poi estratti i fattori attraverso diversi metodi. (nella tesina ad esempio se qualcuno la vuole portare come domanda a scelta, potrebbe confrontare all'interno dell'analisi fattoriale questi due diversi metodi: massima verosimiglianza e componenti principali.)
Metodi di estrazione
Cosa sono gli autovalori e gli autovettori?
Come decidere quanti fattori tenere? Anche in questo caso, proprio perché la tecnica è descrittiva, ci sono diversi criteri di scelta, che vanno specificati. Alcuni autori confrontano diverse analisi fattoriali sullo stesso campione utilizzando diversi criteri, confrontando i diversi criteri di scelta. Si può decidere di scegliere una decina di fattori perché ad esempio è, nell'esempio del questionario, il numero delle sottoscale. Quindi si potrebbe far corrispondere le sottoscale del questionario con i fattori che vengono estratti nell'analisi fattoriale. Oppure si può decidere un limite di perdita di informazione, quindi voglio scegliere quei fattori che almeno mi spieghino il 70% (di solito) della varianza. E quindi dell'informazione. Oppure scelgo quei fattori che hanno il valore, l'autovettore maggiore o uguale a 1, che hanno varianza > o = 1.

Questo perchè sono i fattori che almeno spiegano l'informazione apportata da una variabile perché sapete che il primo punto le variabili o gli item del questionario vengono tutte trasformate in punti z, quindi la loro varianza, per tutti, è uguale a 1.
Perché si scelgono i fattori che hanno almeno varianza uguale a 1? Perché almeno spiegano una delle variabili considerate o degli item di un questionario.
Scopi dell'analisi fattoriale.
L'utilizzo dell'analisi fattoriale per la validazione di questionari è soprattutto legato alla capacità di questa tecnica nell'individuazione della struttura del questionario stesso e quindi attraverso l'analisi fattoriale io sono in grado di vedere quali sono gli item legati fra loro e se possono essere accorpati, se l'informazione dovuta a questi item può essere accorpata in 1 unico fattore e quindi se questo fattore può essere legato ad una sottoscala di riferimento.
Scopo delle rotazioni ortogonali: ovviare ad 1 dei difetti che sono stati riscontrati nell'analisi fattoriale ovvero i fattori che vengono individuati nell'analisi fattoriali sono creati in modo da essere indipendenti fra loro. Quando noi ruotiamo i fattori siamo in grado di vedere quali sono rispetto all'asse x e y, anche quali sono tutte le possibili combinazioni e quando facciamo delle rotazioni oblique siamo in grado di considerarli come correlati fra loro. E quindi viene a essere superato il difetto criticato nell'analisi fattoriale, ovvero quello di considerare i fattori come indipendenti fra loro. Perché, effettivamente, quando si studiano le variabili ad es in 1 questionario si pensa anche, parlando di sottoscale, che in qualche modo queste sottoscale, queste variabili studiate siano correlate e non indipendenti e invece attraverso un'analisi fattoriale semplice i fattori non rispecchiano più la correlazione fra le variabili. Invece la rotazione permette di considerare anche la correlazione fra le variabili. …allo stesso modo è in grado, attraverso la rotazione, di modificare la relazione, cioè di vedere come attorno all'asse xy che è un … modificazione geometrica, le due variabili cambiano nella loro relazione. E quindi le rotazioni permettono di considerare …… x renderlo + adeguato rispetto alla relazione effettiva che c'è fra gli item o fra le variabili. Le rotazioni possono essere scelte: si può scegliere di farle oppure no.
In spss vi viene chiesto di specificare se volete la rotazione, nel metodo che volete, oppure se non volete eseguirle. Quindi questo vi rende + consapevoli del fatto che a seconda della modalità della scelta che lo sperimentatore intende fare rispetto alle variabili da considerare, rispetto al tipo di estrazione e rispetto a tutti i metodi che sono in grado di affinare la rappresentazione alla realtà possiamo avere risultati molto diversi tra loro. Infatti, solitamente, quando si presenta un'analisi fattoriale relativa alla validazione di un questionario, ad es. non si da solo 1 risultato possibile, ma se ne danno + di 1. Anche negli articoli potete trovare x es l'estrazione fatta con il metodo della massima verosimiglianza e l'estrazione con le componenti principali che differenza ha prodotto. Questo perché uno sperimentatore in questo modo… rimane una tecnica descrittiva, non è che dando tutte le possibili combinazioni l'analisi che state facendo diventi più valida, comunque rimane 1 tecnica esplorativa, la tendenza rimane quella di cercare di renderla il + affidabile possibile. Quindi dando tutti i risultati o alcuni dei risultati possibili, si pensa che questa tecnica sia più affidabile. Per questo ad es negli articoli troverete spesso che l'analisi fattoriale viene presentata attraverso uno o più metodi di rappresentazione. In realtà, dal mio punto di vista come metodologo, il suo aspetto primario è quello descrittivo e non serve, non dico non serve a nulla, ma serve a ben poco e sicuramente non serve a renderla più affidabile, il fatto di rappresentare tutti i risultati possibili. Sarebbe forse + utile accettare il fatto che essa sia una tecnica descrittiva ed offrire una delle prospettive e quindi essendo consapevoli che è soltanto una delle prospettive.
Va sotto il nome della riduzione dei dati perché nell'analisi fattoriale, per vostro interesse, ce ne sono altri di tests che permettono di ridurre i dati, l'informazione dei dati, che sono altrettanto validi e sono l'analisi delle corrispondenze e l'ordinal scale.
Ad es nel caso di un questionario sul mobbing le rotazioni possono anche non essere usate. Si può scegliere tra vari tipi di rotazione: Varimax è il + utilizzato (che minimizza il numero delle variabili

che hanno alte correlazioni con i fattori) o se no avete le rotazioni oblique, gli altri sono molto meno usate.
Oltre all'analisi fattoriale abbiamo anche parlato di altre due tecniche descrittive molto usate in psicologia, che sono la Cluster analisi e analisi discriminante
Lo scopo principale è quello di suddividere il campione in sottogruppi.
L'analisi discriminante si differenzia dalla cluster analisi perché come assunto ha il fatto che ci serve come punto di partenza una variabile che ci classifica già il nostro campione in sottogruppi. Quindi è necessaria nell'analisi discriminante una variabile che almeno in una parte del campione, almeno x alcuni soggetti abbia già questa classificazione. Attraverso questa variabile poi si è in grado di calcolare una funzione discriminante che associa la variabile stessa ad una serie di fattori.
Qual è lo scopo principale e qual è lo scopo secondario? Nell'es di ieri abbiamo visto che oltre al fatto di suddividere i gruppi, si può anche analizzare l'influenza che alcune variabili indipendenti, che sono appunto i fattori, hanno sulla classificazione in gruppi, quindi quanto le variabili indipendenti sono utili per discriminare nei gruppi che già conosciamo della variabile di raggruppamento. E quindi questo è un fine secondario, diciamo così.
Quali sono gli assunti dell'analisi discriminante? Ovvero quali sono quelle caratteristiche che permettono di fare 1 analisi discriminante?

Innanzitutto i fattori devono essere a distribuzione gaussiana, dopo di che devono avere,devono essere scarsamente correlati fra loro, e quindi devono avere bassi indici di correlazione. Quando esistono questi indici di correlazione devono essere costanti nei diversi gruppi. Quindi non solo i fattori non devono essere correlati fra loro, ma anche le medie e le deviazioni standard dei fattori non devono essere correlate fra loro.
Qual è il procedimento dell'analisi discriminante?
Fase di addestramento nella quale vengono calcolati i parametri sulla base della variabile di raggruppamento, segue la fase di analisi che è in grado di applicare quei parametri calcolati sulla variabile di raggruppamento per classificare i nuovi casi che non avevano alcuna classificazione.
Il tutto si basa sulla funzione discriminante che vede la variabile di raggruppamento in funzione di parametri sommati fra loro e quindi di una combinazione lineare fra i fattori considerati. Ciascun parametro viene poi, come in tutte le funzioni lineari, moltiplicato ad un coefficiente. Tale funzione viene applicata a ciascun soggetto e quindi è in grado di dare un punteggio discriminante per ciascun soggetto. Questo punteggio del soggetto viene poi utilizzato per calcolare i centroidi ovvero le medie di ciascun gruppo nella funzione discriminante. Quindi per ogni soggetto oltre al punteggio discriminante viene anche calcolata la probabilità di appartenere ad un gruppo oppure ad un altro, ma abbiamo visto vengono calcolati anche i centroidi, ovvero le medie di ciascun gruppo nella funzione discriminante. La distanza tra queste 2 medie è la distanza massima data dal modello. Se noi dividiamo in 2 questa distanza troviamo il punto soglia che permette di dividere i due gruppi. Quando ci sono quindi 2 gruppi, vengono calcolate tante funzioni discriminanti quanti sono il numero dei gruppi meno 1. E quindi allo ss modo ci sarano più soglie a seconda del numero dei gruppi.
Questi sono i 2 utilizzi principali dell'analisi discriminante.
Cluster analisi
La differenza è che non occorre avere 1 variabile di raggruppamento sulla base della quale classificare il campione in gruppi. Poi ci sono diversi cluster che permettono di lavorare sia con variabili ordinali, quindi comunque numeriche però a livello più basso della scala, che con variabili gaussiane. Quindi ci sono diversi tipi di cluster analisi, tutti questi tipi però tendono a creare dei gruppi che siano il massimamente omogenei al loro interno e il più possibile lontani tra loro e quindi che abbiano minima la varianza e massima la distanza delle medie che vengono chiamate centroidi. Una volta individuati i gruppi, essi possono essere utilizzati per altre analisi (quindi come scopo secondario la cluster analisi ha il fatto che può, una volta individuati i gruppi e i sottogruppi del campione, utilizzarli per analisi successivi)

Abbiamo detto che ci sono diversi tipi di analisi dei cluster: Analisi gerarchica, non gerarchica (che solitamente viene chiamata “k means” o “k medie”) che invece è basata su un numero prefissato di gruppi a differenza della cluster gerarchica che invece individua un numero n-1 di gruppi, perchè parte a raggruppare i casi in coppie e dopo raggiunge alla fine il raggruppamento di tutti i casi sotto un unico cluster. Quindi nell’analisi cluster gerarchica spetta al ricercatore scegliere la classificazione ottimale, quella che meglio massimizza la differenza tra i gruppi e minimizza la differenza all’interno dei gruppi. Questa classificazione chiaramente viene fatta però secondo gli assunti, secondo le esigenze del ricercatore. Mentre invece nella cluster non gerarchica, k means, il ricercatore dice quanti gruppi vuole trovare, dopo di che il test di analisi individua come raggruppare attorno a questo numero n di gruppi, i casi. Le distanze nella cluster non gerarchica sono comunque distanze euclidee perché l’assunto della non gerarchica è che i fattori siano a distribuzione gaussiana.
Quindi la prima differenza è che nella non gerarchica le variabili devono essere gaussiane. Un’altra differenza è nel numero di cluster che in un caso deve essere scelto secondo le esigenze quindi una volta ottenuti i risultati il ricercatore sceglie quello che è meglio utilizzare secondo i suoi dati, nell’altro caso invece a priori il ricercatore dice quanti cluster vuole.
Invece abbiamo visto anche un altro processo sempre della cluster, che è abbastanza nuovo rispetto ai primi due, quindi che non troverete nei libri, ma avete visto solo a lezione. Non si pretende comunque che lo conosciate; all’esame sarà praticamente impossibile che venga chiesto, insomma.è il metodo che riesce a trovare il numero ottimale di gruppi attraverso diversi criteri di scelta e che quindi permette di analizzare, di trovare questi gruppi sia in funzione di variabili quantitative gaussiane che di variabili categoriche. Si chiama Twostep cluster.
Correlazione
Ci sono diversi indici di correlazione, parametrici e non parametrici. Soprattutto perché tali indici potevano essere efficaci come ripasso all’analisi fattoriale che si basa tutta sulla matrice di correlazione e quindi in realtà queste slides sono più di ripasso, difficilmente vi verrà chiesto.
La misura della forza dell’effetto (effect size) è stata ripresa, l’avevamo già vista nella prima parte del corso, perché la forza dell’effetto viene utilizzata per la metanalisi (tecnica statistica basata sul calcolo dell’effetto medio che permette di confrontare fra loro ricerche diverse, ma riguardanti lo stesso argomento e serve un po’ da analisi riassuntiva di tutte le ricerche che riguardano lo stesso argomento che hanno prodotto risultati diversi.
Qual è lo scopo della metanalisi? Come procede la metanalisi? Qual è il parametro che misura la relazione tra le varie ricerche? Questo parametro è il parametro che misura la forza dell’effetto. Effettivamente per ogni ricerca viene calcolata la forza dell’effetto e moltiplicata all’effetto medio. Quindi in realtà il procedimento utilizzato nella metanalisi è un procedimento molto semplice dal punto di vista statistico e invece l’aspetto + interessante riguarda da un lato i limiti della ricerca metanalitica e dall’altro le possibilità, la ricchezza del metodo metanalitico. Limiti: non può essere omnicomprensiva di tutti gli studi fatti sull’argomento in quanto alcuni non sono stati pubblicati o di alcuni non è possibile calcolare la forza dell’effetto o non è possibile averne accesso. Quindi, in realtà, tutte le ricerche metanalitiche, basate sulla metanalisi si riferiscono ad un numero limitato di ricerche sull’argomento anche se si cerca di considerarne il numero più elevato possibile.
Un altro limite è: Quali variabili considerare in una ricerca metanalitica? Chiaramente anche come accorpare variabili molto diversetra loro che comunque riguardano lo stesso argomento. E per questo, per tutti questi limiti, comunque la tecnica metanalitica rimane 1 analisi di tipo descrittivo perché inevitabilmente, sebbene si basi su un indice molto affidabile come la forza dell’effetto che è in grado di misurare quanto quella variabile abbia effetto sulla variabilità totale, indipendentemente dalla numerosità campionaria, sebbene quindi sia un indice molto potente, molto utile però le ricerche e quindi il campione di riferimento per questo tipo di analisi è comunque un campione che ha un bias, un errore di partenza molto molto alto. E quindi rimane, per questo motivo, una ricerca di tipo descrittivo.

Non ne abbiamo parlato molto a lezione. La forza dell'effetto: ci sono diversi indici anche nella metanalisi per calcolarlo e ciascun indice si riferisce al tipo di variabile considerate. C'è una formula che riguarda le variabili gaussiane che chiaramente però non va bene per studi che hanno utilizzato variabili non gaussiane, che non hanno media e varianza. L'effect size viene calcolato in modo diverso, ci sono formule diverse a seconda delle variabili che che vanno ad indagare. In questo caso, nella formula più semplice per calcolare la forza dell'effetto, si valuta la differenza fra le medie dei 2 gruppi diviso la radice quadrata della varianza. In questo modo si rapporta quanto sia + forte la differenza fra i gruppi rispetto alla variabilità, alla varianza totale. Però tale forza viene calcolata in modo diverso a seconda delle variabili. Quindi non è tanto importante che sappiate questa formula, ci sono diversi modi per calcolarla. L'importante è che sappiate che la forza dell'effetto misura l'influenza della variabile rispetto alla variabilità totale misurata sui gruppi (non la varianza teorica).

La significatività non è in grado di darci la direzione, non bisogna chiederlo né nell'ipotesi, né nell'interpretazione dei risultati.
Per calcolare la numerosità basta individuare il numero minimo di soggetti necessari, non bisogna prender il numero massimo perchè si aumenta il rischio di trovare delle differenze significative sempre, anche se esse in realtà sono minime. Si può arrivare a delle fallacie, a delle conclusioni errate. Il campione deve essere composto dal numero minimo sufficiente per verificare, per dimostrare la differenza che noi riteniamo clinicamente influente sulla variabile dipendente, specificando anche i criteri di inclusione e di esclusione. Quindi che vada a giustificare quelle che sono le caratteristiche che mi permettono di scegliere un soggetto rispetto ad un altro e quindi i criteri di inclusione ed esclusione dei soggetti nei gruppi.
Chi non ha superato la prima prova deve sapere:
A quali variabili corrispondono i tests statistici: gaussiane→ param, non gauss→ non param.
Significatività e potenza del test e interpretazione dei loro risultati.
Limiti di falsificazione come e dove sono distribuiti in una distribuzione gaussiana.
Se variabili correlate fra loro e disegno within subjects si applica → t-test x prove ripetute o anova o GML per prove ripetute
Distinzione tra variabile dipendente e variabile indipendente.
Errore α e β
Come distinguere le variabili a seconda della distribuzione (gaussiane e non gaussiane) o del loro ruolo nel disegno sperimentale (variabili dipendenti e indipendenti)