

PSP6075525 - Testing psicologico (matr. dispari)

Caso studio del 24-01-22

Istruzioni iniziali

- Si avvii una nuova sessione di R (o RStudio).
- Si crei un nuovo script di R e lo si salvi come `cognome_nome.R`.
- Si effettui il download del file di dati dell'esame `dati_esame.Rdata` disponibile presso la pagina moodle del corso e lo si carichi nell'ambiente di lavoro di R.
- Si crei un nuovo documento di testo (mediante LibreOffice Writer, Microsoft Word o software analogo) e lo si salvi come `cognome_nome.doc`. Il file dovrà contenere le risposte ai quesiti d'esame accompagnati dai comandi di R, dai risultati ottenuti e dai grafici prodotti. Le risposte dovranno essere inserite in ordine, rispettando il numero del quesito a cui si riferiscono. Alla fine, il file dovrà essere convertito in formato non modificabile (PDF: `cognome_nome.pdf`) ed inviato al docente utilizzando la procedura "Consegna documento" disponibile presso la pagina Moodle del corso. Nel caso di utilizzo di **R-markdown** per la compilazione dinamica di documenti di testo, sarà necessario inviare il file sorgente `.Rmd` unitamente al file PDF generato. Si ricorda di riportare chiaramente Nome, Cognome e Matricola all'interno dei file contenenti le soluzioni finali (`.pdf`, `.R`, `.Rmd`).
- La valutazione della prova sarà effettuata utilizzando primariamente il file `cognome_nome.pdf`: si raccomanda pertanto la chiarezza nella scrittura delle risposte e la correttezza nel riportare i comandi e gli output di R. Il file `cognome_nome.R` dovrà essere allegato al file `cognome_nome.pdf` solo per un controllo aggiuntivo (pertanto non verrà primariamente valutato).

Caso studio

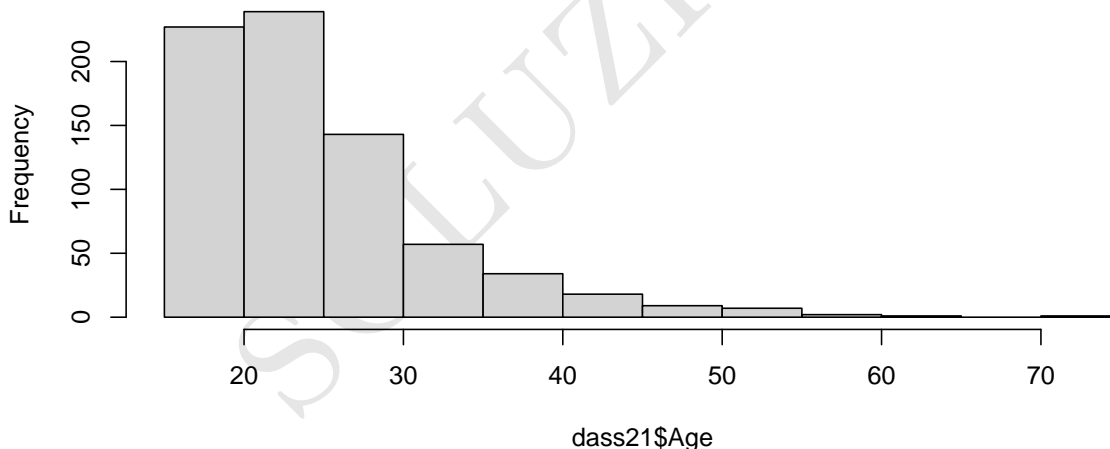
Il caso studio si riferisce alla valutazione del test DASS-21 (*Depression Anxiety Stress Scale*) usato per la valutazione della depressione, ansia e stress. Il test è composto da tre dimensioni (D: depressione; S: stress; A: ansia) quantificate mediante 21 item. I dati si riferiscono ad uno studio¹ che ha coinvolto 738 partecipanti (di cui 374 di genere maschile) di nazionalità britannica. Gli item (frequenza di comportamenti) sono stati rilevati su scale ordinali a 4 livelli (1: “Did not apply”, ..., 4: “Applied most of the time”). Alcuni di questi item sono i seguenti: *How often...Life was meaningless*, *Unable to become enthusiastic*, *Felt scared without reason*, *Difficult to relax*. L'obiettivo dell'analisi è quello di studiare la dimensionalità complessiva del test DASS-21.

1. Si individuino il numero di unità statistiche e si commenti il tipo di dato a disposizione.

Il numero di unità statistiche è pari a $n = 738$ (maschi: $n = 364$; femmine: $n = 374$). I dati a disposizione consistono nelle risposte su scala ordinale ai $p = 21$ item del test.

2. Si valuti l'omogeneità delle unità statistiche rispetto alla variabile **Age**. Si rappresenti tale variabile graficamente e si calcolino delle statistiche di sintesi. Infine, si rimuovano dal dataset le righe corrispondenti alle unità per le quali la variabile **Age** sia maggiore o uguale al 90% percentile.

```
hist(dass21$Age, main = "")
```



```
summary(dass21$Age)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
18.00	20.00	23.00	25.29	28.00	72.00

```
q90 = quantile(x = dass21$Age, probs = 0.90)
print(q90)
```

```
90%
35
```

¹ Gomez, R., Stavropoulos, V., & Griffiths, M. D. (2020). Confirmatory factor analysis and exploratory structural equation modelling of the factor structure of the Depression Anxiety and Stress Scales-21. *PloS one*, 15(6), e0233998.

```
sum(dass21$Age>=q90) / NROW(dass21)

[1] 0.1097561

iid = dass21$Age>=q90 #units to be removed
dass21 = dass21[!iid,]
```

Il dataset presenta 81 unità statistiche (pari all'11% del numero di unità complessivo) con età superiore a 35 anni (il quantile corrispondente al 90% percentile della distribuzione). Le statistiche di sintesi evidenziano che il campione non è omogeneo per età: questa varia tra 18 e 72 anni, con il cinquanta per cento delle osservazioni tra i 20 e i 28 anni.

3. Si valuti la coerenza interna della scala D rispetto alla variabile **Gender**. Gli item della scala D sono i seguenti: DASS_Q3 DASS_Q5 DASS_Q10 DASS_Q13 DASS_Q16 DASS_Q17 DASS_Q21.

```
jjd = paste0("DASS_Q",c(3,5,10,13,16,17,21))
coef_alpha(dass21[dass21$Gender=="M",jjd])

[1] 0.8455167

coef_alpha(dass21[dass21$Gender=="F",jjd])

[1] 0.9005602
```

La coerenza interna della scala D è pari a 0.845 per il gruppo **Gender=M** e 0.902 per il gruppo **Gender=F**. Ciò indica che la scala D riesce a ben separare la varianza del misurando da quella dell'errore di misurazione.

4. Si definisca e si adatti ai dati un modello fattoriale confermativo a tre variabili latenti secondo la seguente assegnazione:

D: DASS_Q3 DASS_Q5 DASS_Q10 DASS_Q13 DASS_Q16 DASS_Q17 DASS_Q21
 A: DASS_Q2 DASS_Q4 DASS_Q7 DASS_Q9 DASS_Q15 DASS_Q19 DASS_Q20
 S: DASS_Q1 DASS_Q6 DASS_Q8 DASS_Q11 DASS_Q12 DASS_Q14 DASS_Q18

Il modello è composta da $q = 3$ variabili latenti per $p = 21$ variabili osservate secondo l'equazione lineare

$$\Sigma_{y_{21 \times 21}} = \Lambda_{21 \times 3} \Phi_{3 \times 3} \Lambda_{21 \times 3}^T + \Theta_{\delta_{21 \times 21}}$$

L'adattamento ai dati $\mathbf{S}_{y_{21 \times 21}}$ deve essere fatto mediante stimatori DWLS per dati ordinali. Il modello necessita di 44 parametri da stimare (21 coefficienti fattoriali, 21 varianze d'errore, 2 varianze delle variabili latenti) su un totale di $p(p+1)/2 = 231$ parametri totali. Poiché la stima è effettuata mediante DWLS, vi sono parametri aggiuntivi da stimare (c.d. *thresholds* degli item) e che si riferiscono alle soglie continue associate alle categorie di risposta.

```
#ricodifichiamo ciascuna variabile osservata in termini di 'ordered factor'
for(j in 3:23){
  dass21[,j] = factor(dass21[,j],ordered=TRUE)
}

mD = paste0("D =~ ", paste(paste0("DASS_Q",c(3,5,10,13,16,17,21)),collapse=" + "))
mA = paste0("A =~ ", paste(paste0("DASS_Q",c(2,4,7,9,15,19,20)),collapse=" + "))
mS = paste0("S =~ ", paste(paste0("DASS_Q",c(1,6,8,11,12,14,18)),collapse=" + "))

vars_ord = names(dass21)[3:23]
```

```
model1 = paste(mD,mA,mS,sep = " \n ")
fit1 = lavaan::cfa(model = model1,data = dass21,ordered = vars_ord,estimator="DWLS")
```

5. Si interpreti il risultato del modello adattati al punto 4 anche mediante l'utilizzo di indici di adattamento complessivo (si suggerisce l'utilizzo dei coefficienti standardizzati nell'interpretazione della soluzione fattoriale).

```
print(fit1)
```

lavaan 0.6-7 ended normally after 84 iterations

Estimator	DWLS
Optimization method	NLMINB
Number of free parameters	87
Number of observations	657

Model Test User Model:

Test statistic	328.874
Degrees of freedom	186
P-value (Chi-square)	0.000

```
res1 = lavaan::inspect(fit1,what="std.all")
Xout = cbind(res1$lambda,diag(res1$theta))
colnames(Xout)=c("lambda_D","lambda_A","lambda_S","diag(ThetaDelta)")
print(Xout)
```

	lambda_D	lambda_A	lambda_S	diag(ThetaDelta)
DASS_Q3	0.7449312	0.0000000	0.0000000	0.4450775
DASS_Q5	0.6460897	0.0000000	0.0000000	0.5825680
DASS_Q10	0.7947867	0.0000000	0.0000000	0.3683142
DASS_Q13	0.7946189	0.0000000	0.0000000	0.3685809
DASS_Q16	0.7744575	0.0000000	0.0000000	0.4002156
DASS_Q17	0.7962420	0.0000000	0.0000000	0.3659987
DASS_Q21	0.8038929	0.0000000	0.0000000	0.3537562
DASS_Q2	0.0000000	0.3835662	0.0000000	0.8528770
DASS_Q4	0.0000000	0.6386704	0.0000000	0.5921002
DASS_Q7	0.0000000	0.5176666	0.0000000	0.7320213
DASS_Q9	0.0000000	0.7065338	0.0000000	0.5008101
DASS_Q15	0.0000000	0.8056017	0.0000000	0.3510058
DASS_Q19	0.0000000	0.5655003	0.0000000	0.6802094
DASS_Q20	0.0000000	0.7506359	0.0000000	0.4365457
DASS_Q1	0.0000000	0.0000000	0.09859417	0.9902792
DASS_Q6	0.0000000	0.0000000	0.68448184	0.5314846
DASS_Q8	0.0000000	0.0000000	0.74407957	0.4463456
DASS_Q11	0.0000000	0.0000000	0.68952004	0.5245621
DASS_Q12	0.0000000	0.0000000	0.69895655	0.5114597
DASS_Q14	0.0000000	0.0000000	0.51054814	0.7393406
DASS_Q18	0.0000000	0.0000000	0.63490487	0.5968958

```
res1$psi
```

D	A	S
---	---	---

```
D 1.000
A 0.726 1.000
S 0.775 0.902 1.000
```

```
fitMeasures(fit1,fit.measures = c("RMSEA","CFI","chisq","df","npar"))
```

```
rmsea    cfi    chisq    df    npar
0.034    0.995 328.874 186.000 87.000
```

Globalmente il modello adattato evidenzia buoni indici di CFI e RMSEA. La struttura fattoriale della scala D è ben formata dagli item a disposizione, con coefficienti fattoriali di magnitudine sufficientemente elevata. La scala A, invece, presenta coefficienti di magnitudine moderata, con qualche item (es.: DASS_Q2, DASS_Q7) di magnitudine medio-bassa. La scala S, infine, presenta coefficienti di magnitudine accettabile, ad eccezione degli item DASS_Q14 e DASS_Q1. Le varianze d'errore per ciascun item sono contenute ad eccezione dell'item DASS_Q1. La matrice $\Phi_{3 \times 3}$ evidenzia che le variabili latenti sono correlate tra loro.

6. Sulla base della stessa assegnazione item-variabile latente usata per il modello precedente, si definisca un nuovo modello CFA di tipo bifattoriale a variabili latenti incorrelate. Nota: per fissare a zero tutte le correlazioni tra le variabili latenti si può utilizzare il comando `orthogonal=TRUE` all'interno della funzione `cfa(..)`. Inoltre, all'interno di quest'ultima, si consiglia l'aggiunta del parametro `optim.force.converged=TRUE`.

Il modello è composta da $q = 4$ variabili latenti per $p = 21$ variabili osservate secondo l'equazione lineare

$$\Sigma_{y_{21 \times 21}} = \Lambda_{21 \times 4} \mathbf{I}_{4 \times 4} \Lambda_{21 \times 4}^T + \Theta_{\delta_{21 \times 21}}$$

L'adattamento ai dati $\mathbf{S}_{y_{21 \times 21}}$ deve essere fatto mediante stimatori DWLS per dati ordinali. Rispetto al modello precedente, il modello bifattoriale è meno parsimonioso rispetto a quello precedente.

```
options(warn=-1)
mG = paste("G =~", paste(vars_ord,collapse=" + "))
model2 = paste(mD,mA,mS,mG,sep = " \n ")
fit2 = lavaan::cfa(model = model2,data = dass21,ordered = vars_ord,
  estimator="DWLS",orthogonal=TRUE,optim.force.converged=TRUE)
```

7. Si valuti mediante l'utilizzo di indici di adattamento complessivo se la soluzione bifattoriale (punto 6) sia superiore o meno a quella a tre variabili latenti correlate (punto 4). Si scelga, dopo opportune argomentazioni, il modello fattoriale finale che meglio si adatta ai dati.

Il modello è composta da $q = 4$ variabili latenti per $p = 21$ variabili osservate secondo l'equazione lineare

$$\Sigma_{y_{21 \times 21}} = \Lambda_{21 \times 4} \mathbf{I}_{4 \times 4} \Lambda_{21 \times 4}^T + \Theta_{\delta_{21 \times 21}}$$

L'adattamento ai dati $\mathbf{S}_{y_{21 \times 21}}$ deve essere fatto mediante stimatori DWLS per dati ordinali. Rispetto al modello precedente, il modello bifattoriale è meno parsimonioso.

```
Yout = rbind(fitMeasures(fit2,fit.measures = c("RMSEA","CFI","chisq","df","npar")),
  fitMeasures(fit1,fit.measures = c("RMSEA","CFI","chisq","df","npar")))
rownames(Yout)=c("mod2","mod1")
print(Yout)
```

```
rmsea    cfi    chisq    df    npar
mod2 0.02783430 0.9968292 253.3835 168 105
mod1 0.03421906 0.9946943 328.8739 186 87
```

Il modello bifattoriale presenta indici di adattamento pressoché equiparabili a quelli del modello a tre fattori. Tuttavia, rispetto a quest'ultimo, il modello è meno parsimonioso e presenta difficoltà interpretative rispetto alla nuova variabile latente G. Per tali ragioni si sceglie di considerare il modello a tre fattori latenti per le successive analisi.

8. Sulla base del modello scelto al punto 7, si valuti se il modello può essere migliorato e se ne giustifichi il perché.

Il modello a tre fattori latenti presenta indici di adattamento molto elevati, ragione per cui effettuare un miglioramento basato su indici di modifica risulterebbe privo di senso. Al contrario, un miglioramento plausibile è quello che si basa sull'eliminazione degli item che presentano bassi coefficienti fattoriali. Ci si riferisce, in particolare, a DASS_Q1 con $\lambda_{1,3} = 0.098$. L'eliminazione di quest'ultima variabile permette di definire un modello più parsimonioso del precedente e che si adatta ai dati ancora in maniera adeguata.

```
mS = paste0("S =~ ", paste(paste0("DASS_Q",c(6,8,11,12,14,18)),
                             collapse=" + ")) #si esclude l'item Q1
model1b = paste(mD,mA,mS,sep = " \n ")
fit1b = lavaan::cfa(model = model1b,data = dass21,ordered = vars_ord,estimator="DWLS")

fitMeasures(fit1b,fit.measures = c("RMSEA","CFI","chisq","df","npar"))
```

rmsea	cfi	chisq	df	npar
0.031	0.996	269.559	167.000	83.000

9. Sulla base dei risultati ottenuti al punto 8, si calcoli l'attendibilità delle scale D, A, S e se ne interpreti il risultato.

```
semTools::reliability(fit1b,what = "omega",return.total = TRUE)
```

	D	A	S	total
omega	0.8822343	0.775014	0.7887148	0.9235311

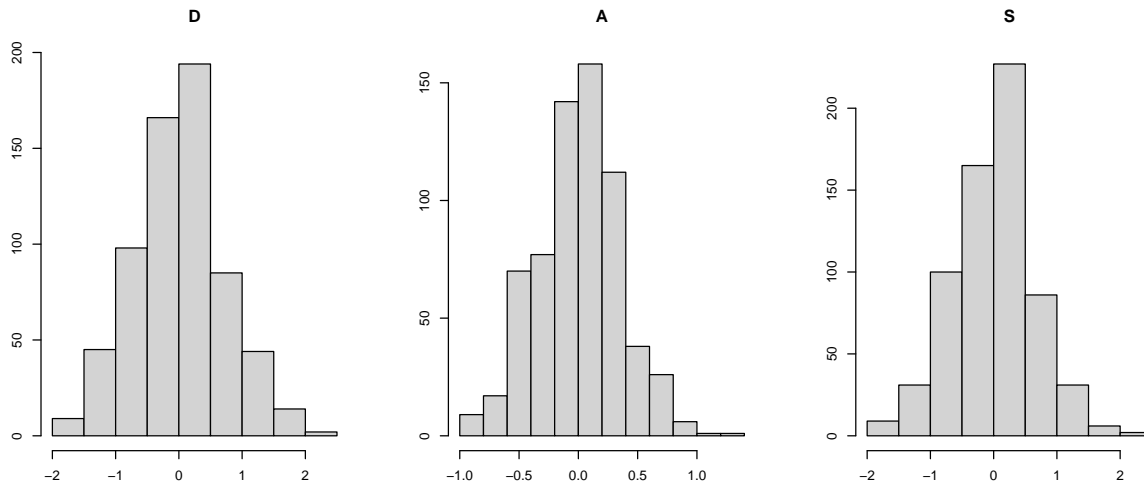
L'indice ω per l'attendibilità di un modello CFA adattato ai dati, evidenzia una buona attendibilità delle tre sottoscale, con più del 70% della variabilità osservata da attribuire alla presenza dei misurandi latenti (più che all'errore di misurazione).

10. Si calcolino i valori predetti dal modello finale a livello delle unità statistiche (c.d. *factor scores*) e li si rappresenti graficamente. Cosa possiamo dire circa la loro forma distributiva? Si fornisca una spiegazione basata sul razionale del modello CFA.

I valori latenti predetti dal modello $\hat{\eta}_i = \mathbb{E}[\eta_i | \mathbf{y}_i]$ possono essere calcolati in diversi modi. Un modo è quello di usare uno stimatore lineare basato su una procedura di regressione. In *lavaan* tale metodo è implementato mediante `lavPredict(...,method='regression')`.

```
Xpred = lavaan::lavPredict(object = fit1b,newdata = dass21,type = "lv",
                           method = "regression")

par(mfrow=c(1,3))
hist(Xpred[,1],main="D",ylab="",xlab="")
hist(Xpred[,2],main="A",ylab="",xlab="")
hist(Xpred[,3],main="S",ylab="",xlab="")
```



```
summary(Xpred)
```

D	A	S
Min. : -1.542581	Min. : -0.807473	Min. : -1.506969
1st Qu.: -0.439482	1st Qu.: -0.208447	1st Qu.: -0.416766
Median : 0.023305	Median : 0.016259	Median : 0.032862
Mean : 0.009239	Mean : 0.003417	Mean : 0.005874
3rd Qu.: 0.426678	3rd Qu.: 0.224891	3rd Qu.: 0.400045
Max. : 2.221488	Max. : 1.309894	Max. : 2.339077

I punteggi fattoriali presentano distribuzione simmetrica e centrata sullo zero. Ciò dipende dal modello CFA adattato ai dati: questo infatti non modella le medie dei fattori latenti ($\tau = 0$).