Slide 1

More about the Chi-square test

Besides giving estimates for $\mathbf{\Lambda}$, $\mathbf{\Phi}$ and Ψ, Maximum Likelihood Provides a Goodness of Fit test.

- Use Likelihood Ratio Test, i.e.,

$$
\begin{aligned}
-2\log\frac{\text{L(MLE of restricted model)}}{\text{L(MLE of unrestricted model)}} & \\
&= 2\{log(L(\mathbf{S})) - log(L(\hat{\underline{\Sigma}}))\} \\
&= n\{trace\hat{\underline{\Sigma}}^{-1}\mathbf{S} - \log|\hat{\underline{\Sigma}}^{-1}\mathbf{S}| - p\}
\end{aligned}
$$

- IF the model fits the data well, this statistic should be small!
- Its distribution is asymptotically distributed χ^2 with degrees of freedom = ($\frac{p(p+1)}{2}$ - number of unique parameters in model).
- We can determine if the the statistic is "small" enough by comparing to the χ^2 distribution and obtaining a p-value.

Slide 2

More about the Chi-square test

- In general, the Hypothesis being tested is
  H_0: $\mathbf{\Sigma} = \mathbf{\Sigma}(\mathbf{\Theta})$ (your model)
  H_A: $\mathbf{\Sigma} = \mathbf{S}$ the saturated model (or just-identified model)
- Since the degrees of freedom for the saturated model are 0, this means it fits the data perfectly
- So we are comparing $\mathbf{\Sigma}(\mathbf{\Theta})$ to a model that fits the data perfectly. Thus if it is not significantly different than the model that fits perfectly, it means it is pretty good
- Thus we are looking for the model where we DO NOT REJECT the H_0 (i.e. find a big p-value)
- Note, the chi-square test has been proven to be **asymptotically** valid even when the data is not normally distributed (Amemiya and Anderson, 1985). Note: this is an asymptotic result.
- FROM KLINE, page 209-210 Satorra Bentler correction to the Chi-square statistic when the data is non-normal. This correction is available in Mplus but not in AMOS.

Slide 3

Factor score estimation - Overview

Ch 7 of Skrondal and Rabe-Hesketh (2004) gives details for "latent scoring" and "latent classification"

- Once the number of factors and factor structure are identified, researcher often wants to create a score, a **factor score estimate** to represent each factor, which then can be used in subsequent analysis.
- Using principal component analysis we can estimate the first q components simply by taking $\mathbf{\Lambda}'\mathbf{x}$. That is, create a weighted sum of the p observed variables, weighting them by centered columns of $\mathbf{\Lambda}$.
- If a common factor model is assumed then we will create a factor score estimate of the true underlying factors... It should be emphasized that this score is an error-prone measurement of the latent variable, i.e.

$$\hat{f}_i = f_i + \delta_i$$

  That is, the score is expected to contain measurement error, although we would expect it to be more reliable than any one of the measures that went into it.
- NOTE: Individual level data is required (not just correlation matrix) in order to get scores for each individual.
- It should also be noted that an alternative approach to creating a factor score estimate is to explicitly model the latent variables in subsequent analysis as will be done in structural equation modeling.

Slide 4

Calculating factor score estimates - Simple sum score

The most common method used in all different literatures is to create a simple (equally weighted) sum score of items which load significantly on a particular factor

- Since this method does not take into account varying factor loading, it would not be expected to be most efficient method
- Turning that argument around, it can also be claimed that since the factor loadings are likely sample specific, we should not rely on them too much when making the score so that the score is more generalizable
- Use Cronbach's alpha to estimate reliability, recall it assumes $x_{ij} = f_i + \epsilon_{ij}$ that all loadings are equal (i.e. $\lambda_j = 1$ for all j). Raykov (1997-2001) shows Cronbach's alpha is a lower bound for actual reliability of "congeneric" measures, i.e. measures where the factor loadings are not all the same.
- Another argument in favor of this technique is that each score only includes items which significantly loaded on that factor

Slide 5

Calculating factor score estimates - Regression weights

- If we assume $\mathbf{f}$ is normally distributed and $\mathbf{x}|\mathbf{f}$ is distributed normally, then

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{f} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi} & \boldsymbol{\Phi\Lambda}' \\ \boldsymbol{\Lambda\Phi} & \boldsymbol{\Phi} \end{pmatrix}\right)$$

  thus the best predictor of the underlying factor given the model is correct is

$$E(\mathbf{f}|\mathbf{x}) = \boldsymbol{\Phi\Lambda}'(\boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi})^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

  and so the factor score estimate is

$$\hat{f}_i = \hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}}'(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}})^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$$

  or sometimes it is calculated slightly different as

$$\hat{f}_i = \hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Lambda}}'\mathbf{S}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$$

In general each factor score depends on all p observed variables. Also, in general, $Corr(\hat{f}_m, \hat{f}_{m'}) \neq Corr(f_m, f_{m'})$.

The square of the estimated correlation between $\hat{\mathbf{f}}$ and $\mathbf{f}$ can be used as an estimate of reliability

Slide 6

Calculating factor score estimates - Bartlett; Anderson and Rubin

Make no assumption about the distribution of the underlying factors, in fact, treat them as fixed unknown quantities, then given estimates $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ we can consider

$$\mathbf{x}_i = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\Lambda}}\mathbf{f}_i + \boldsymbol{\epsilon}$$

with $\widehat{Var}(\boldsymbol{\epsilon}) = \hat{\boldsymbol{\Psi}}$ to be a regression model with $\mathbf{f}_i$ as the q unknown parameter and $\hat{\boldsymbol{\Lambda}}$ as the observed regressors matrix. Thus the generalized least squares estimator is (Bartlett factor score):

$$\hat{f}_i = (\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}})^{-1}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$$

Anderson and Rubin, 1956 is a modification of the Bartlett method so that $Corr(\hat{f}_m, \hat{f}_{m'}) = \mathbf{I}$

- Bartlett (1937) The statistical conception of mental factors. *British Journal of Psychology*, **28**, 97-104.
- Anderson, TW and Rubin, H. (1956) Statistical inference in factor analysis. *Proceedings of the third Berkley Symposium on Mathematical Statistics and Probability*, vol 5, pp 111-150.

Slide 7

Factor score weights - PCA - Unrotated factors

```
 proc factor data = a rotate = promax
scores output = factscores;
 var gaelic english history math algebra geometry; run;

                              The FACTOR Procedure
                     Initial Factor Method: Principal Components

                     Scoring Coefficients Estimated by Regression

             Squared Multiple Correlations of the Variables with Each Factor

                                Factor1         Factor2

                              1.0000000       1.0000000

                          Standardized Scoring Coefficients

                                       Factor1         Factor2

                       gaelic          0.24217         0.39376
                       english         0.25231         0.25671
                       history         0.18923         0.56658
                       math            0.26959        -0.36944
                       algebra         0.27188        -0.33023
                       geometry        0.24853        -0.31370
```

Slide 8

Factor score weights - ML - Unrotated factors

```
                              The FACTOR Procedure
                      Initial Factor Method: Maximum Likelihood

                     Scoring Coefficients Estimated by Regression

             Squared Multiple Correlations of the Variables with Each Factor

                                Factor1         Factor2

                             0.82064705      0.53269908

                          Standardized Scoring Coefficients

                                       Factor1         Factor2

                       gaelic          0.19686         0.39032
                       english         0.17164         0.22457
                       history         0.10929         0.32620
                       math            0.35087        -0.34613
                       algebra         0.29270        -0.22151
                       geometry        0.16980        -0.09914
```

Slide 9

Factor score weights - ML - Rotated factors

The FACTOR Procedure
Rotation Method: Promax (power = 3)

Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with Each Factor

Factor1	Factor2
0.79341037	0.69325853

Standardized Scoring Coefficients

	Factor1	Factor2
gaelic	0.06728	0.40662
english	0.09425	0.27754
history	0.00367	0.29857
math	0.44031	0.03178
algebra	0.34664	0.07124
geometry	0.19206	0.06085

Slide 10

Genetic Concerns Questionnaire - EFA - ML

```
proc factor data = genetictest scree
method = ml rotate = promax nfact=4 flag = .3; var c1-c20; run;
```

The FACTOR Procedure
Initial Factor Method: Maximum Likelihood
Prior Communality Estimates: SMC

C1	C2	C3	C4	C5	C6	C7
0.41029924	0.69555063	0.69345231	0.54651245	0.51414191	0.58527718	0.62437419

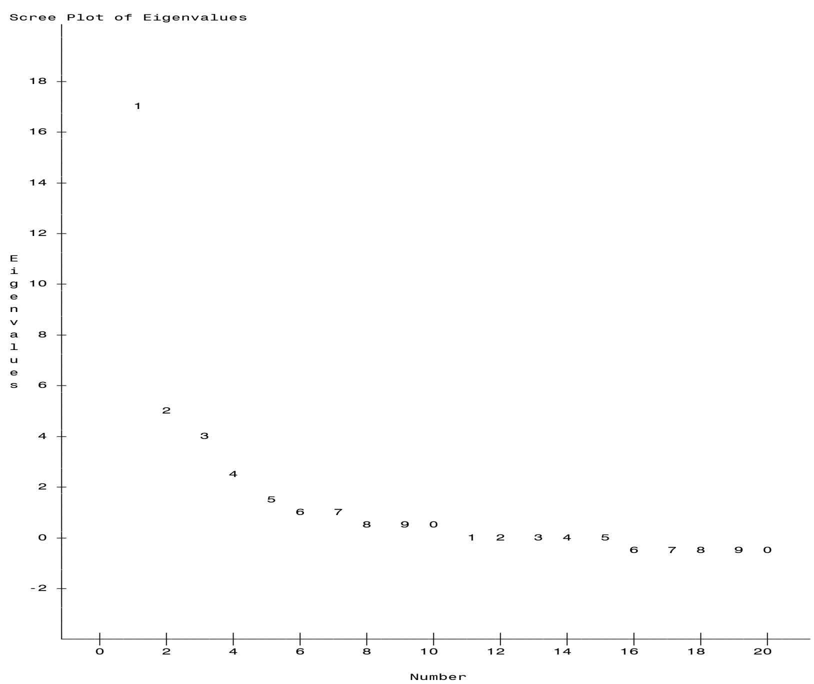
C8	C9	C10	C11	C12	C13	C14
0.47598962	0.60194636	0.63636979	0.39761689	0.52851826	0.67491498	0.59745581

C15	C16	C17	C18	C19	C20
0.67083904	0.58465071	0.52457147	0.64761079	0.69479515	0.61797719

Preliminary Eigenvalues: Total = 30.372251 Average = 1.51861255

	Eigenvalue	Difference	Proportion	Cumulative
1	16.7672687	11.9963216	0.5521	0.5521
2	4.7709471	1.0027379	0.1571	0.7091
3	3.7682092	1.0293909	0.1241	0.8332
4	2.7388183	1.3083265	0.0902	0.9234
5	1.4304918	0.3590838	0.0471	0.9705
6	1.0714080	0.2230647	0.0353	1.0058
7	0.8483433	0.1408710	0.0279	1.0337
8	0.7074723	0.1696340	0.0233	1.0570
9	0.5378383	0.2754239	0.0177	1.0747
10	0.2624143	0.0221818	0.0086	1.0833
11	0.2402325	0.1761596	0.0079	1.0912
12	0.0640730	0.0939153	0.0021	1.0934
13	-0.0298423	0.1060550	-0.0010	1.0924
14	-0.1358972	0.1120360	-0.0045	1.0879
15	-0.2479332	0.0857999	-0.0082	1.0797
16	-0.3337331	0.0373563	-0.0110	1.0687
17	-0.3710894	0.0893819	-0.0122	1.0565
18	-0.4604714	0.0903286	-0.0152	1.0414
19	-0.5507999	0.1546994	-0.0181	1.0232
20	-0.7054993		-0.0232	1.0000

Slide 11



Slide 12

Genetic Concerns Questionnaire - EFA - ML

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2
C1	C1	57 *	-13
C2	C2	-4	77 *
C3	C3	77 *	0
C4	C4	15	53 *
C5	C5	-18	71 *
C6	C6	23	56 *
C7	C7	73 *	-9
C8	C8	14	26
C9	C9	-1	38 *
C10	C10	70 *	3
C11	C11	-4	36 *
C12	C12	63 *	-4
C13	C13	72 *	11
C14	C14	35 *	27
C15	C15	7	76 *
C16	C16	59 *	11
C17	C17	55 *	4
C18	C18	47 *	7
C19	C19	43 *	16
C20	C20	78 *	-4

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	190	2126.5455	<.0001
HA: At least one common factor			
H0: 2 Factors are sufficient	151	818.6337	<.0001
HA: More factors are needed			

Slide 13

Genetic Concerns Questionnaire - EFA - ML

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2	Factor3
C1	C1	51 *	-13	12
C2	C2	1	77 *	-5
C3	C3	86 *	1	-11
C4	C4	16	52 *	2
C5	C5	-19	68 *	6
C6	C6	23	54 *	5
C7	C7	83 *	-8	-15
C8	C8	-12	19	50 *
C9	C9	-4	36 *	10
C10	C10	66 *	3	7
C11	C11	-6	35 *	4
C12	C12	57 *	-4	11
C13	C13	70 *	11	5
C14	C14	31 *	25	11
C15	C15	10	77 *	-2
C16	C16	65 *	12	-8
C17	C17	30	-2	50 *
C18	C18	9	-2	73 *
C19	C19	-3	5	91 *
C20	C20	61 *	-8	33 *

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors HA: At least one common factor	190	2126.5455	<.0001
H0: 3 Factors are sufficient HA: More factors are needed	133	593.5235	<.0001

Slide 14

Genetic Concerns Questionnaire - EFA - ML

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2	Factor3	Factor4
C1	C1	51 *	-4	13	-16
C2	C2	0	77 *	-4	7
C3	C3	85 *	-1	-11	4
C4	C4	17	68 *	6	-24
C5	C5	-19	63 *	8	10
C6	C6	23	45 *	3	21
C7	C7	82 *	-2	-15	-4
C8	C8	-11	28	55 *	-19
C9	C9	-11	5	1	73 *
C10	C10	66 *	6	7	-3
C11	C11	-12	8	-3	63 *
C12	C12	57 *	-7	9	7
C13	C13	69 *	7	4	12
C14	C14	26	-3	4	62 *
C15	C15	10	68 *	-2	17
C16	C16	67 *	13	-7	-4
C17	C17	28	-12	48 *	20
C18	C18	7	-3	74 *	4
C19	C19	-4	2	90 *	7
C20	C20	62 *	-2	35 *	-12

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors HA: At least one common factor	190	2126.5455	<.0001
H0: 4 Factors are sufficient HA: More factors are needed	116	454.4120	<.0001

Slide 15

Genetic Concerns Questionnaire - EFA - ML

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2	Factor3	Factor4	Factor5
C1	C1	32 *	1	12	-16	32 *
C2	C2	-6	80 *	-7	8	8
C3	C3	91 *	-5	-8	3	-3
C4	C4	8	71 *	4	-23	10
C5	C5	-2	60 *	11	10	-40 *
C6	C6	24	44 *	4	21	-6
C7	C7	76 *	-2	-12	-5	11
C8	C8	1	23	59 *	-20	-28
C9	C9	-7	4	3	73 *	-8
C10	C10	43 *	13	5	-3	45 *
C11	C11	-16	10	-4	64 *	9
C12	C12	29	1	5	8	57 *
C13	C13	61 *	9	6	11	15
C14	C14	31 *	-6	6	62 *	-3
C15	C15	5	71 *	-5	18	9
C16	C16	67 *	12	-4	-5	0
C17	C17	24	-14	50 *	20	7
C18	C18	1	-4	75 *	4	7
C19	C19	-15	2	92 *	7	12
C20	C20	58 *	-4	39 *	-12	5

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	190	2126.5455	<.0001
HA: At least one common factor			
H0: 5 Factors are sufficient	100	374.4601	<.0001
HA: More factors are needed			

Slide 16

Genetic Concerns Questionnaire - EFA - ML

```
By default if the number of factors is not designated SAS it uses
the proportion criterion looking for $>100\%$ cumulative of the
reduced correlation matrix variance explained.

6 factors will be retained by the PROPORTION criterion.


NOTE: 6 factors will be retained by the NFACTOR criterion. ERROR:
Communality greater than 1.0. NOTE: The SAS System stopped
processing this step because of errors. NOTE: PROCEDURE FACTOR
used:
      real time           0.03 seconds
      cpu time            0.03 seconds
```

NOTE: Mplus had problems giving results for these data, it only converged when q=2

Slide 17

Genetic Concerns Questionnaire - EFA - ML

```
1 factor
              Akaike's Information Criterion               784.3265
              Schwarz's Bayesian Criterion                 219.4148

2 factor
              Akaike's Information Criterion              553.68549
              Schwarz's Bayesian Criterion                 51.91098

3 factor
              Akaike's Information Criterion              356.51308
              Schwarz's Bayesian Criterion                -85.44725

4 factor
              Akaike's Information Criterion              246.24617
              Schwarz's Bayesian Criterion               -139.22299

5 factor
              Akaike's Information Criterion              195.46089
              Schwarz's Bayesian Criterion               -136.84011
```

Slide 18

Genetic Concerns Questionnaire - EFA - ML

Use the options "reorder" to get SAS to arrange output

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2	Factor3	Factor4
C3	C3	85 *	-1	-11	4
C7	C7	82 *	-2	-15	-4
C13	C13	69 *	7	4	12
C16	C16	67 *	13	-7	-4
C10	C10	66 *	6	7	-3
C20	C20	62 *	-2	35 *	-12
C12	C12	57 *	-7	9	7
C1	C1	51 *	-4	13	-16
C2	C2	0	77 *	-4	7
C15	C15	10	68 *	-2	17
C4	C4	17	68 *	6	-24
C5	C5	-19	63 *	8	10
C6	C6	23	45 *	3	21
C19	C19	-4	2	90 *	7
C18	C18	7	-3	74 *	4
C8	C8	-11	28	55 *	-19
C17	C17	28	-12	48 *	20
C9	C9	-11	5	1	73 *
C11	C11	-12	8	-3	63 *
C14	C14	26	-3	4	62 *

Squared Multiple Correlations of the Variables with Each Factor

Factor1	Factor2	Factor3	Factor4
0.90613605	0.84330261	0.88215144	0.77540283

AND her are the associated Cronbach alphas using only the items
with loading over .4

.87	.82	.79	.73