# Factor Analysis

**Final Project** **April 2019**

## Overview

*What is factor analysis?*

- Multivariate statistical technique used in statistical modelling and exploratory data analysis in order to determine the most effective model for a given context
- Factor analysis is an unsupervised learning technique

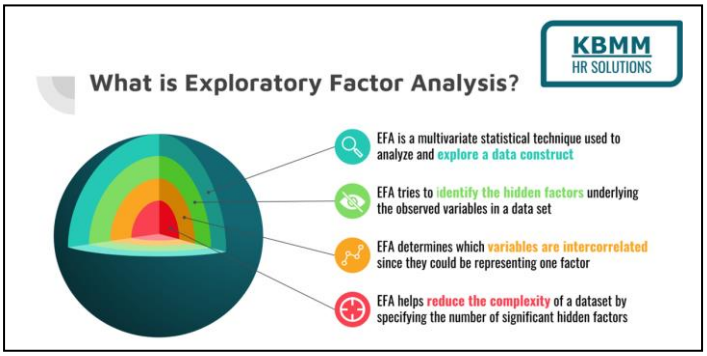
### Types of Factor Analysis

*Exploratory Factor Analysis*

- Used to reduce the number of relevant variables in a model and assess a data construct's dimensionality
- Determines which measured variables are highly intercorrelated as these variables could very well just be reflecting one hidden factor
- Reduces the complexity of the data and generates the most precise predictive models

*Confirmatory Factor Analysis*

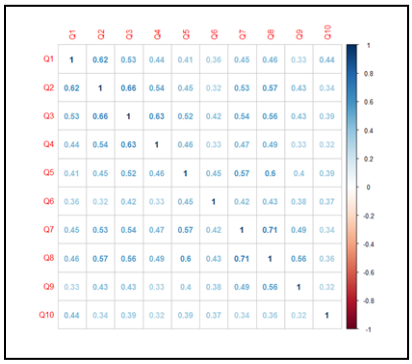
- Can specify the exact number as well as indicate, in particular, which measured variable is related to which underlying factor




## Assumptions

*Each of the below need to be satisfied in order to perform factor analysis*

- All data must be metrical- based on a numeric scale, and there should be no outliers
- Large sample size- rule of thumb is larger than 200 observations
- Minimum correlation between variables of at least 0.30. Can be tested using a correlation matrix (below)



- Homogeneous sample – if this assumption is violated, the sample size will increase as the number of variables increases

Reliability analysis is conducted to test homogeneity. We can test this by looking at the Coefficient Alpha (Cronback's Alpha) and the Average Split-Half Reliability of our data. This tests the internal consistency, or reliability, of our data.

With set1 equal to the data of desired variables- if the below r-code is greater than 0.8- the sample can be considered homogeneous.

```
alpha(set1)

splitHalf(set1)
```

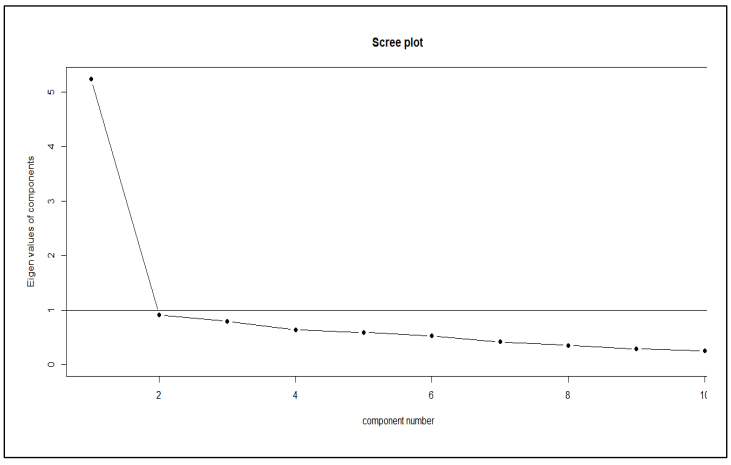
## Data

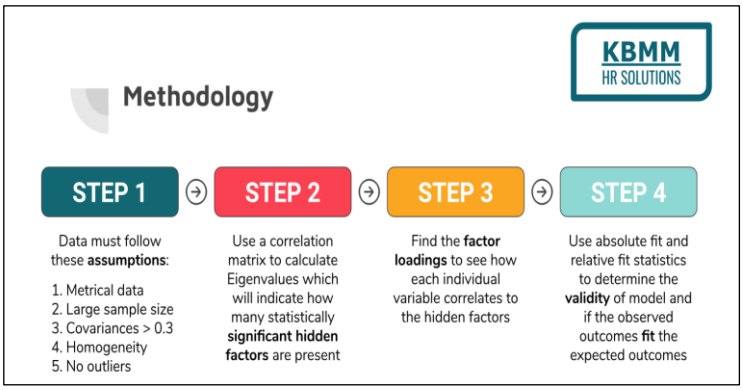
Our data consists of the responses to the ten-question survey, each question serving as an individual variable. There is also a score variable which is the sum of responses to the ten questions. The gender variable is as follows: 1=male, 2=female, 3=other, and 0=not answered. Lastly, the age variable corresponds to the respondent's age in years. After data cleaning we had 3241 observations, and for our purposes the responses to the ten questions were our main focus. The five number summary is shown below.

| | n | mean | sd | median | min | max | Q1 | Q3 | missing_ob |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3241 | 2.319346 | 1.087879 | 2 | 1 | 4 | 1 | 3 | FALS |
| 2 | 3241 | 2.236655 | 1.071731 | 2 | 1 | 4 | 1 | 3 | FALS |
| 3 | 3241 | 2.233570 | 1.079007 | 2 | 1 | 4 | 1 | 3 | FALS |
| 4 | 3241 | 1.944462 | 1.036376 | 2 | 1 | 4 | 1 | 3 | FALS |
| 5 | 3241 | 2.238198 | 1.115562 | 2 | 1 | 4 | 1 | 3 | FALS |
| 6 | 3241 | 3.099352 | 0.965454 | 3 | 1 | 4 | 2 | 4 | FALS |
| 7 | 3241 | 2.204875 | 1.068916 | 2 | 1 | 4 | 1 | 3 | FALS |
| 8 | 3241 | 2.297748 | 1.086061 | 2 | 1 | 4 | 1 | 3 | FALS |
| 9 | 3241 | 2.469608 | 1.157271 | 2 | 1 | 4 | 1 | 4 | FALS |
| 10 | 3241 | 2.521135 | 1.189304 | 3 | 1 | 4 | 1 | 4 | FALS |

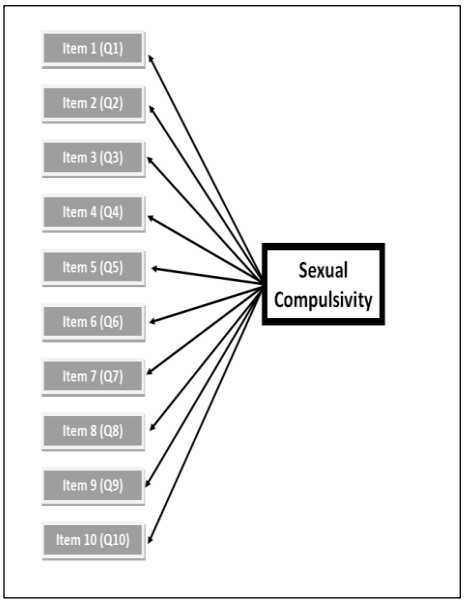
## Results

- From the correlation matrix eigenvalues will be calculated
- Eigenvalues are numeric indicators of the amount of variance explained by each factor
- Eigenvalues above the cutoff of 1 are considered statistically significant and relevant factors in our data. These eigenvalues above the cutoff can be seen visually in a **Scree Plot** (below)




- After isolating the exact number of factors, one can create the EFA model. This model can determine how each survey question (item) relates (correlates) to the underlying factors by reading the **factor loadings**
- Factor loadings are essentially correlation results on a scale from -1 to 1
- One can also see how each individual observation relates to the factors through the **factor scores**
- This is done on a spectrum where high positive values reflect strong expression of the underlying factor, while high negative values reflect a strong lack of expression of the factor with 0 indicating neutral expression of the factor.
- The model's validity will be tested using **relative fit** and **absolute fit** statistics



## Advantages

- Ability to simplify models
    - Through factor analysis, one can reduce both the number of dimensions and variables in a data set as the inclusion of irrelevant additional measures in a model leads to unnecessary noise and bias that will hinder the generalizability of the results
    - This will lead to reduction in the complexity of the data to generate the most precise predictive models

## Disadvantages

- Only describes the number of factors that produce the best-fitting mode
    - cannot actually be used to identify what the hidden factors represent
- Factor analysis is inherently subjective
    - Therefore, the researcher is free to interpret the results of factor analysis in any way they see fit (be careful with bias here)

## Next Steps

*Confirmatory Factor Analysis*

- Hypothesis test for the existence of a **relationship** between observed variables and their hidden constructs

*Principal Component Analysis*

- Further modify our variables to a reduced set that **correlates significantly** to the construct in question

## Works Cited


Brussow, Jennifer. “Factor Analysis in R.” Datacamp, C. Ismay & B. Robins. Datacamp.https://www.datacamp.com/courses/factor-analysis-in-r

Child, D. (1990). The essentials of factor analysis, second edition. London: Cassel Educational Limited.

Joreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis, Psychometrika, 34, 183-202.

Kalichman & Rompa. *Sexual Compulsivity Scale*. 1995. (Data file and code book). Openpsychometrics.org. Accessed April 5, 2019.

Salkind, Neil J. *Encyclopedia of Research Design.* Thousand Oaks, CA: SAGE Publications, Inc., 2010. *SAGE Research Methods*. Web. 19 Apr. 2019, doi: 10.4135/9781412961288.

Statistics Solutions. (2013). Confirmatory Factor Analysis. Retrieved from http://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/confirmatory-factor-analysis/

Statistics Solutions. (2013). Exploratory Factor Analysis. Retrieved from https://www.statisticssolutions.com/factor-analysis-sem-exploratory-factor-analysis/

“What Is a Factor Score?” *Psychology.okstate.edu*, psychology.okstate.edu/faculty/jgrice/factorscores/fs_q.html.