Quantitative Data Analysis: A Companion for Accounting and Information Systems Research

Teaching Materials

Created by Willem Mertens, Amedeo Pugliese & Jan Recker
What these materials are about
Offering a guide through the essential steps required in quantitative data analysis

1. **Introduction**
2. **Comparing Differences Across Groups**
3. **Assessing (Innocuous) Relationships**
5. **Nested Data and Multilevel Models: Hierarchical Linear Modeling**
6. **Analyzing Longitudinal and Panel Data**
7. **Causality: Endogeneity Biases and Possible Remedies**
8. **How to Start Analyzing, Test Assumptions and Deal with that Pesky p-Value**
9. **Keeping Track and Staying Sane**
Part 1: Exploring Data and Testing Assumptions
There are three kinds of lies: lies, damned lies, and statistics.

Statistics are no substitute for judgment.

Benjamin Disraeli

Henry Clay
1. Exploring Data
   - Structuring data
     - Basics
     - Variable types
   - Cleaning data and eliminating outliers
   - Visualising data

2. Understanding data
   - Distributions, means and standard deviations
   - Models and significance
   - Correlations and differences

3. Testing assumptions
   - Independence
   - Homoscedasticity
   - Normality
     - Skew and kurtosis
     - Transformations

4. Scales and factors
   - Basics
   - PCA/EFA vs. CFA
### Structuring data

1. Exploring data

- One row per case, one variable per column

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Role</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>F</td>
<td>Student</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>F</td>
<td>Professor</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>M</td>
<td>Admin</td>
<td></td>
</tr>
</tbody>
</table>

- Depends on unit of analysis (e.g. person)
### Structuring data

1. Exploring data

- Nested data

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Role 1</th>
<th>Role 2</th>
<th>Role 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>F</td>
<td>Student</td>
<td>Tutor</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>Tutor</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>53</td>
<td>F</td>
<td>Professor</td>
<td>Head of School</td>
<td>Supervisor</td>
</tr>
<tr>
<td>53</td>
<td>F</td>
<td>Head of School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>F</td>
<td>Supervisor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>M</td>
<td>Admin</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Structuring data

1. Exploring data

- Recoding data: variable types
  - Categorical variables
    - Nominal (e.g. role)
    - Dichotomous (e.g. gender)
    - Ordinal (e.g. hierarchical level)
  - Continuous variable
    - Interval (e.g. degrees): 5-10 = 15-20
    - Ratio (e.g. weight): 0 is nothing, 10 = 2*5

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Gender</th>
<th>Role 1</th>
<th>Role 2</th>
<th>Role 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>19</td>
<td>1</td>
<td>Student</td>
<td>Tutor</td>
<td>-</td>
</tr>
<tr>
<td>Person 2</td>
<td>53</td>
<td>1</td>
<td>Professor</td>
<td>Head of School</td>
<td>Supervisor</td>
</tr>
<tr>
<td>Person 3</td>
<td>27</td>
<td>2</td>
<td>Admin</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Cleaning data and eliminating outliers

1. Exploring data

- Cleaning data
  = Taking out *unreliable* (not inconvenient) cases
    - Missing data (or listwise/pairwise)
    - Extreme tendencies (e.g. all 6/all 1)
    - Improbable response time (e.g. outliers)
    - Inconsistent responses (e.g. age < tenure)

≠ Introducing bias
  - Consistent application of rules
  - Mindful of hypotheses and method (IV/DV)

- Consider power and credibility
Cleaning data and eliminating outliers

1. Exploring data

- Eliminating outliers
  - Outliers are highly improbable or erroneous values
    - They can influence statistics --> introduce bias
    - They affect generalizability
    - The decision to exclude depends on the RQs
  - How to find outliers
    - Box-plots
    - Histograms
    - Scatter plots
    - z-scores <-3.29 or >3.29 (see slide 16)
Visualising data

1. Exploring data

- Histograms

Histogram showing the age distribution of senior league bowls players.
Visualising data

1. Exploring data

- Box plots

"Boxplot vs PDF" by Jhguch - Wikipedia
Visualising data
1. Exploring data

- Scatter plots

![Scatter plot of Age of senior league bowls players vs # points scored](image)
Distributions, means and standard deviations

2. Understanding data

- Frequency distributions

<table>
<thead>
<tr>
<th>Age of senior league bowls players</th>
<th>Population of senior bowls players</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>n</td>
</tr>
<tr>
<td>66</td>
<td>n</td>
</tr>
<tr>
<td>67</td>
<td>n</td>
</tr>
<tr>
<td>68</td>
<td>n</td>
</tr>
<tr>
<td>69</td>
<td>n</td>
</tr>
<tr>
<td>70</td>
<td>n</td>
</tr>
<tr>
<td>71</td>
<td>n</td>
</tr>
<tr>
<td>72</td>
<td>n</td>
</tr>
<tr>
<td>73</td>
<td>n</td>
</tr>
<tr>
<td>74</td>
<td>n</td>
</tr>
<tr>
<td>75</td>
<td>n</td>
</tr>
<tr>
<td>76</td>
<td>n</td>
</tr>
<tr>
<td>...</td>
<td>n</td>
</tr>
<tr>
<td>83</td>
<td>n</td>
</tr>
</tbody>
</table>
Distributions, means and standard deviations
2. Understanding data

- Probability distributions - e.g.: normal distribution

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 0
\]

\[
s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}} = 1
\]

Normalization:

\[
z = \frac{X - \bar{X}}{s}
\]

\[\Rightarrow X = s^* z + \bar{X}\]
Models and significance

2. Understanding data

- Models
  - Attempt to explain/summarise data
  - Vary in how well they “fit” the data
    - E.g.: mean is a model; s illustrates fit
  - Fit

- Significance
  - Hypothesis testing involves comparing two models (H₀ vs. H₁)
  - Comparing models is done using test statistics:
    variance explained by the model/variance not explained by the model
  - If the probability of observing this test statistic, or anything more extreme, is smaller than .05/.01/.001, then we conclude statistical significance (i.e. H₁ explains the data better than H₀)

Significance ≠ importance
Non-significance does not say anything about H₀
Example of a model/hypothesis test: difference between means = $t$-test

Population of senior bowls players

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$
Example of a model/hypothesis test: difference between means = $t$-test

$$t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

- Values depend on degrees of freedom (df): the number of values that are free to vary when calculating the statistic. For $t$-tests this is $n-1$; the example shown is for large samples ($n>100$).
Correlations and differences

2. Understanding data

**Example of a model/hypothesis test: correlation**

- **Correlations and differences**

  - **Understanding data**

  - **Age of senior league bowls players**

  - **# points scored**

  - **$H_0$**

  - **$r$**

  - **Age of senior league bowls players**

  - **# points scored**
Independence
- Data was collected from independent sources
- Variable measurements were independent (e.g. regression)

Homoscedasticity/homogeneity of variance
- Variance is equal in different (sub-)samples

Normality
- Sampling distribution/errors/data follow a normal distribution --> have limited skew and kurtosis
Independence

3. Testing assumptions

- Data was collected from independent sources
  - No repeated measures
  - No mutual influence between participants
  - No nested structures (see HLM module)

- Variable measurements were independent
  - No priming, framing, context or other question order effects
  - In regression-based models:
    - Variables are unrelated to external (exogenous) variables
    - Errors are independent
Hososcedasticity/homogeneity of variance
3. Testing assumptions

- One variable, multiple groups (e.g. t-test): spread of values is equal across different groups
  - Visual test: scatter- or boxplot
  - Statistical test: Levene’s test for equality of variance
    - When significant ($p < .05$): no homo-scedascity (i.e. heteroscedascity)

Levene’s test will usually be significant in large samples; use other tests (e.g. Hartley’s $F_{max}$)
Homo-scedasticity/homogeneity of variance

3. Testing assumptions

- Two variables (e.g. regression): spread of errors/residuals is equal across different values of $x$
In many statistical tests

- Sampling distribution is normally distributed
  --> test normality of sample
  - Visually testing normality of (sub-)sample data
    - Histograms (see slide 10)
    - Q-Q plots: theoretical vs. actual quantiles

"Normal normal qq" by Skbkekas - Wikipedia
Normality

3. Testing assumptions

- Statistical tests for normality of (sub-)sample data
  - Compute descriptives including skew and kurtosis
  - Convert skew and kurtosis to z-scores, e.g.:

\[
Z_{\text{skewness}} = \frac{\text{skewness} - 0}{\text{SE}_{\text{skewness}}} \Rightarrow \left| \frac{\text{skewness}}{\text{SE}_{\text{skewness}}} \right| \text{ must be } \leq 1.96
\]

![Warning]

Increase to 2.58 in larger samples and do not use in very large samples \((n > 200)\)

- Shapiro-Wilk test: significant \((p < .05)\) when NOT normal
In regression-based models
- Errors/residuals, not indicators need to be normally distributed
- Same visual principles as Q-Q plot apply

Please note: in this case, both graphs do not represent the same data.
What if assumptions are violated?

3. Testing assumptions

- Correct data
  - Exclude outliers
  - Transform data, e.g.:
    - Log-, square root and reciprocal (1/x) transformations shorten the right tale (i.e. correct positive skew)
    - The same transformations applied to the reverse score (score – highest score + 1) correct for negative skew

  ! The same transformation has to be applied to variables that are compared directly

- Turn to tests that are robust against violations or to non-parametric tests, e.g.
  - Mann–Whitney U for group comparisons
  - Kendall's tau for dependence between two variables
Scales and factors - basics

4. Scales and factors

- Scales are sets of indicators that measure the same latent variable / factor
  ≠ response scales!

- E.g. To aid me in my teaching, overall, I feel PowerPoint ... is:
  - Easy to Learn
  - Easy to manipulate
  - Clear to interact with
  - Flexible to interact with
  - Difficult to master (reverse scored)
  - Very cumbersome (reverse scored)

Ease of use
Visualisation of scale with three indicators measuring one latent variable / factor:
Run PCA with no restriction on the number of factors and with a scree plot.

Decide how many factors to retain based on eigenvalues, scree plot and $R^2$

- Separate mountain from scree
- Eigenvalue $> 1$
  - Eigenvalue: proportion of variance explained by factor (sum = # variables)
- Cumulative $R^2 > .6$
Run PCA again
  - Restrict the number of extracted factors
  - Rotate factors orthogonally or oblique based on theory (or trial and error/inspection of the component correlation matrix)
  - Study the component matrix (orthogonal) or pattern matrix (oblique) to interpret factors and exclude indicators when
    - Loading is small (< .4/.7) on all factors
    - Loadings are high for multiple factors (> .4/.7)
    - Difference between loadings on different factors < .2
  - Run PCA again after each exclusion
Once a stable solution has been reached, evaluate reliability and unidimensionality of scales

- Inter-item correlation when # indicators for factor is 2
  - Should be significant
- Chronbach’s Alpha when # indicators for factor is > 2
  - Should be higher than .7
  - “Alpha if item deleted” should be lower than Alpha
    - If not: exclude item and run PCA again
End of Part 1