WORKSHOP 1: INTRODUCTION TO SPSS 1

SPSS Basics:

MINJEONG PARK, GRADUATE ACADEMIC ASSISTANT
MOHAMMAD Zubair, GRADUATE ACADEMIC ASSISTANT

UBC Library Research Commons
Background

• Have you used SPSS before?

• Have you worked with other data analysis software?

• Have you conducted any quantitative data analysis before?
Outline

• Overview of quantitative research
• Software for statistical models
• SPSS environment
• Data preparation in SPSS
• Data management in SPSS
• Descriptive statistics in SPSS
• Statistical models in SPSS
Overview of quantitative research

- **Research purpose/questions** (e.g., association, group difference, prediction, change)
- **Research design** (e.g. Cross-sectional, Longitudinal)
- **Data collection** (e.g., continuous, categorical, ordinal data)
- **Statistical analysis**
  - Manipulating data (e.g., recode data, select cases)
  - Applying statistical models (e.g. correlation, t-test, ANOVA, regression)

→ **Software programs** for statistical analysis help you to conduct statistical analysis
Software for statistical analysis

- SPSS
- R
- Stata
- SAS
- Matlab
- Minitab
- Mplus
- HLM7
- ...

Software selection can be dependent on…
- Convention in disciplines
- Availability
- Characteristics of software
## Software for statistical analysis

<table>
<thead>
<tr>
<th></th>
<th>SPSS</th>
<th>Stata</th>
<th>R &amp; R studio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Price</strong></td>
<td>Gradpack (Base version)</td>
<td>Stata IC</td>
<td>Non-commercial purpose</td>
</tr>
<tr>
<td></td>
<td>34.49 USD* (6 month)</td>
<td>125 USD* (annual)</td>
<td>(Free)</td>
</tr>
<tr>
<td></td>
<td>Free trial available</td>
<td>30 days money back option</td>
<td></td>
</tr>
<tr>
<td><strong>Availability in UBC</strong></td>
<td>Koerner Rm 217A, 218A</td>
<td>Koerner Rm 217A, 218A</td>
<td>Koerner Rm 217A, 218A</td>
</tr>
<tr>
<td></td>
<td>Woodward Rm B25</td>
<td>Woodward Rm B25</td>
<td>Woodward Rm B25</td>
</tr>
<tr>
<td><strong>Interface</strong></td>
<td>clicking button + syntax</td>
<td>clicking button + syntax</td>
<td>syntax/command</td>
</tr>
<tr>
<td><strong>Data management</strong></td>
<td>Basic data management</td>
<td>Basic data management</td>
<td>Basic data management</td>
</tr>
<tr>
<td></td>
<td>options</td>
<td>options</td>
<td>options + more data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>management options</td>
</tr>
<tr>
<td><strong>Statistical models</strong></td>
<td>Well-established statistical models</td>
<td>Well-established statistical models</td>
<td>Well-established statistical models + recent statistical models are updated</td>
</tr>
<tr>
<td><strong>Data visualization</strong></td>
<td>Basic plots, graphs, charts</td>
<td>Basic plots, graphs, charts</td>
<td>Basic plots, graphs, charts + more visualization options</td>
</tr>
<tr>
<td><strong>Discussion forum</strong></td>
<td>Stack Overflow</td>
<td>Statalist</td>
<td>Stack Overflow</td>
</tr>
<tr>
<td></td>
<td>Cross Validated</td>
<td></td>
<td>Cross Validated</td>
</tr>
<tr>
<td></td>
<td>Nabble</td>
<td></td>
<td>Nabble</td>
</tr>
</tbody>
</table>
SPSS environment
SPSS environment

SPSS environment is composed of 3 main windows:

- **Data Editor window**
  
  (Data View + Variable View)

- **Output window**

- **Syntax window**
SPSS environment

- **Data Editor - Data View**: present whole “data”
SPSS environment

- **Data Editor - Variable View**: present information of all “variables”
SPSS environment

- **Data Editor - Variable View**: “Key” information of variables
SPSS environment

- **Data Editor - Variable View**: “Key” information of variables

Meaning of values in variable
E.g. 0 = male, 1 = female

Numbers indicating missing data
SPSS environment

- **Data Editor - Variable View**: “Key” information of variables

Type of data in variable
- **Scale** = continuous data
- **Nominal** = categorical data
- **Ordinal** = ordinal data

For more information about scale, nominal, and ordinal options,
https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/
SPSS environment

- **Drop-down menu in Data Editor:**

![SPSS Example data.sav [DataSet1] - IBM SPSS Statistics Data Editor](image-url)
SPSS environment

- **Output**: present “history of your analysis” and all “outputs”
SPSS environment

- **Syntax**: write “syntax”
Data preparation and management in SPSS
Data preparation with SPSS using example data

• Cross-sectional design:
  A set of variables measured from each person in one time point

• A set of variables:
  - Gender (Male = 0, Female = 1)
  - Age (range 10 – 80)
  - Marital status (Married, common law = 1, Widow, divorce, separate = 2, Single, never married = 3)
  - Employment (no job = 1, part time = 2, full time = 3)
  - Quality of life_total (range 0 – 20)
  - Distress_total (range 0 – 20)
  - Self-esteem items (range 0 – 3)

• Missing data are coded as 999

Data preparation in SPSS:
Importing the data & Looking at descriptive statistics
Data import/entry

• Open data in “SPSS format” (.sav):
  File > Open > Data > Select SPSS data file
• Import data in “different formats” (spreadsheet, text etc.):
  File > Open > Data…
Data import/entry

- Enter your data in SPSS *Data Editor – Data View*:

![SPSS Data Editor](image)

- **Data View**
- **Variable View**
Checking information of variables in Variable View

- A set of variables:
  - Gender (Male = 0, Female = 1)
  - Age (range 10 - 80)
  - Marital status (Married, common law = 1, Widow, divorce, separate = 2, Single, never married = 3)
  - Employment (No job = 1, Part time = 2, Full time = 3)
  - Quality of life_total (range 0 - 20)
  - Distress_total (range 0 – 20)
  - Self-esteem items (range 0 – 3)

- Missing data is coded as 999

E.g., Meaning of values

Click it
Editing value label

- **Employment**
  - No job = 1
  - Part time = 2
  - Full time = 3

Variable View > Values > Add
Recode to a new variables

Old variables – employment (3 levels)

New variable – employment2 (2 levels)

Transform > Recode in different variables

Add old and new values

Transform > Recode in different variables

Old and new values
Compute a variables

New variable – Esteem_Total

Description: total score of 10 esteem questions

Label: Total esteem score

Transform > compute variable

Type “Target Variable”
Compute total esteem score
Descriptive statistics in SPSS
Descriptive statistics

• Descriptive statistics provide a summary of your data

• Purpose of looking at descriptive statistics:

  (1) Check whether valid data are loaded properly
      E.g., unexpected values (e.g., 999, -2) in “Age” variable (range 10-80)

  (2) Explore data
      E.g., potential group differences, associations between variables

  (3) Sample description
      E.g., % of gender, mean and standard deviation of quality of life score
Descriptive statistics

Descriptive statistics in SPSS:
Descriptive statistics - Frequencies

- Frequencies for “categorical data”
Descriptive statistics - Frequencies

- Frequencies for "categorical data" – Descriptive statistics

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Gender</th>
<th>Marital status</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valid: 200</td>
<td>Valid: 200</td>
</tr>
<tr>
<td></td>
<td>Missing: 0</td>
<td>Missing: 0</td>
</tr>
<tr>
<td>Minimum: 0</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Maximum: 1</td>
<td>80</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid: male</td>
<td>91</td>
<td>45.5</td>
<td>45.5</td>
<td>45.5</td>
</tr>
<tr>
<td>Valid: female</td>
<td>109</td>
<td>54.5</td>
<td>54.5</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marital status</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid: married, common law</td>
<td>63</td>
<td>31.5</td>
<td>31.5</td>
<td>31.5</td>
</tr>
<tr>
<td>Valid: widow, divorce, separate</td>
<td>65</td>
<td>32.5</td>
<td>32.5</td>
<td>64.0</td>
</tr>
<tr>
<td>Valid: single, never married</td>
<td>72</td>
<td>36.0</td>
<td>36.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Employment</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid: no job</td>
<td>58</td>
<td>29.0</td>
<td>29.0</td>
<td>29.0</td>
</tr>
<tr>
<td>Valid: part time</td>
<td>75</td>
<td>37.5</td>
<td>37.5</td>
<td>66.5</td>
</tr>
<tr>
<td>Valid: full time</td>
<td>67</td>
<td>33.5</td>
<td>33.5</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>
Descriptive statistics - Frequencies

- Frequencies for “categorical data” - Bar plots

Gender

- Male
- Female

Marital status

- Married
- Widow
- Single

Employment

- No job
- Part-time
- Full-time
Descriptive statistics - Frequencies

- Frequencies for "continuous data"
Descriptive statistics - Frequencies

• Frequencies for “continuous data” – Descriptive statistics

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Age</th>
<th>Quality of life</th>
<th>Distress</th>
</tr>
</thead>
<tbody>
<tr>
<td>N Valid</td>
<td>196</td>
<td>200</td>
<td>198</td>
</tr>
<tr>
<td>Missing</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Mean</td>
<td>44.20</td>
<td>11.45</td>
<td>10.68</td>
</tr>
<tr>
<td>Minimum</td>
<td>13</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>80</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
Descriptive statistics - Frequencies

- Frequencies for “continuous data” – Histograms

![Histogram of Age](image1)
![Histogram of Quality of life](image2)
![Histogram of Distress](image3)
Descriptive statistics - Graphs

- **Scatter/Dot plots**: Graphs > Legacy Dialogs > Scatter/Dot…

  → *Useful to explore associations between variables*
Descriptive statistics - Graphs

- Scatter plots: output
Statistical models in SPSS
### Statistical models in SPSS

- **Statistical models we are covering today…**

<table>
<thead>
<tr>
<th>Data type</th>
<th>Continuous</th>
<th>Categorical</th>
<th>Ordinal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inference</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Descriptive statistics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inferential statistics</td>
<td>Independent sample T-test</td>
<td>Independent one-way ANOVA</td>
<td>Pearson correlation</td>
</tr>
</tbody>
</table>


Independent sample T-test

• **Independent T-test** compares *means* between *two groups*.

• It is often used to see whether there is *group difference* in *continuous data* between *two groups* (e.g., gender, treatment vs. control).

• Example

<table>
<thead>
<tr>
<th>Gender</th>
<th>Quality of life</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Male</td>
<td>9</td>
</tr>
<tr>
<td>Male</td>
<td>15</td>
</tr>
<tr>
<td>Male</td>
<td>18</td>
</tr>
<tr>
<td>Female</td>
<td>19</td>
</tr>
<tr>
<td>Female</td>
<td>16</td>
</tr>
<tr>
<td>Female</td>
<td>9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

• Model assumptions
  
  (1) Independence, (2) Normality, (3) Equal variance
Independent sample T-test

• Assumption checking

(1) Data are independent of one another (Independence)
   → Independence is checked by data collection (sampling design)

(2) Data in each group should be close to normal distribution (Normality)
   → Checked by histogram, QQ plots, normality test etc.

“Analyze → Descriptive → Explore”
Independent sample T-test

• Assumption checking

(2) Data in each group should be close to normal distribution (Normality)

*Note.* Independent sample T-test is robust to the violation of normality
Independent sample T-test

• Assumption checking

(3) Variances are equal between groups (**Equal variance; homogeneity**)

→ Checked by comparing standard deviations of groups
→ Checked by Levene’s test (given by default in SPSS)
Independent sample T-test

Analyze > Compare Means > Independent-Sample T Test …
Independent sample T-test

- **Output:** Test for equal variance assumption

<table>
<thead>
<tr>
<th>Levene's Test for Equality of Variances</th>
<th>t-Test for Equality of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Sig.</td>
</tr>
<tr>
<td>Equal variances assumed</td>
<td>2.490</td>
</tr>
<tr>
<td>Equal variances not assumed</td>
<td>-1.762</td>
</tr>
</tbody>
</table>

- **Conclusion:**
  Variances of male group and female group are **not significantly different**

Note. Given alpha level = 0.05
### Independent sample T-test

- **Output:** Results of independent T-test

<table>
<thead>
<tr>
<th>Levene’s Test for Equality of Variances</th>
<th>t-test for Equality of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of life</td>
<td></td>
</tr>
<tr>
<td>Equal variances assumed</td>
<td>2.490</td>
</tr>
<tr>
<td>Equal variances not assumed</td>
<td>-1.738</td>
</tr>
<tr>
<td></td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>.084</td>
</tr>
<tr>
<td></td>
<td>-1.079</td>
</tr>
<tr>
<td></td>
<td>.621</td>
</tr>
<tr>
<td></td>
<td>-2.303</td>
</tr>
<tr>
<td></td>
<td>.145</td>
</tr>
<tr>
<td></td>
<td>-1.762</td>
</tr>
<tr>
<td></td>
<td>197.875</td>
</tr>
<tr>
<td></td>
<td>.080</td>
</tr>
<tr>
<td></td>
<td>-1.079</td>
</tr>
<tr>
<td></td>
<td>.612</td>
</tr>
<tr>
<td></td>
<td>-2.286</td>
</tr>
<tr>
<td></td>
<td>.128</td>
</tr>
</tbody>
</table>

- **Conclusion:**
  There was *no statistically significant difference* in level of quality of life between males and females, $t(198) = -1.738, p = 0.084$.

Note. Given alpha level = 0.05
Independent sample one-way ANOVA

• Independent sample one-way ANOVA compares *means* between *more than two groups*

• It is often used to see whether there are *group differences* in *continuous data* between *more than two groups*

• Example

<table>
<thead>
<tr>
<th>Marital status</th>
<th>Quality of life</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Married</td>
<td>8</td>
</tr>
<tr>
<td>Married</td>
<td>11</td>
</tr>
<tr>
<td>Widow/Sep</td>
<td>12</td>
</tr>
<tr>
<td>Widow/Sep</td>
<td>9</td>
</tr>
<tr>
<td>Single</td>
<td>17</td>
</tr>
<tr>
<td>Single</td>
<td>16</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

• Model assumption:
  1. Independence, 2. Normality, 3. Equal variance
Independent sample one-way ANOVA

• Assumption checking

(1) Data are independent of one another (Independence)
→ Independence is checked by data collection (sampling design)

(2) Data in each group should be close to normal distribution (Normality)
→ Checked by histogram, QQ plots, normality test etc.

“Analyze → Descriptive → Explore”
Independent sample one-way ANOVA

- Assumption checking

(3) Variances are equal between groups (**Equal variance; homogeneity**)
  → Checked by comparing standard deviations of groups
  → Checked by Levene’s test (option in one-way ANOVA in SPSS)
Independent sample one-way ANOVA

- Analyze > Compare Means > One-Way ANOVA …
Independent sample one-way ANOVA

• **Output:** Test for equal variance assumption

<table>
<thead>
<tr>
<th>Quality of life</th>
<th>Levene Statistic</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.983</td>
<td>2</td>
<td>197</td>
<td>.376</td>
</tr>
</tbody>
</table>

• **Conclusion:**
Variances of married, widow, and single groups are *not significantly different*

Note. Given alpha level = 0.05
**Independent sample one-way ANOVA**

- **Output:** Overall group difference (omnibus test results)

- **Conclusion:** There was *statistically significant group differences* in level of quality of life between martial status groups, \( F(2, 197) = 19.827, p <0.001. \)

Note. Given alpha level = 0.05
Independent sample one-way ANOVA

• **Output:** Which groups differ? (post hoc test results)

<table>
<thead>
<tr>
<th>Marital status (i)</th>
<th>Marital status (j)</th>
<th>Mean Difference (i-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>married, common law</td>
<td>widow, divorce, separate</td>
<td>4.381*</td>
<td>.712</td>
<td>.000</td>
<td>2.70 - 6.06</td>
</tr>
<tr>
<td></td>
<td>single, never married</td>
<td>1.423</td>
<td>.695</td>
<td>.104</td>
<td>-2.22 - 3.06</td>
</tr>
<tr>
<td>widow, divorce, separate</td>
<td>married, common law</td>
<td>-4.381*</td>
<td>.712</td>
<td>.000</td>
<td>-6.06 - -2.70</td>
</tr>
<tr>
<td></td>
<td>single, never married</td>
<td>-2.958*</td>
<td>.689</td>
<td>.000</td>
<td>-4.59 - -1.33</td>
</tr>
<tr>
<td>single, never married</td>
<td>married, common law</td>
<td>-1.423</td>
<td>.695</td>
<td>.104</td>
<td>-3.06 - .22</td>
</tr>
<tr>
<td></td>
<td>widow, divorce, separate</td>
<td>2.958*</td>
<td>.689</td>
<td>.000</td>
<td>1.33 - 4.59</td>
</tr>
</tbody>
</table>

* The mean difference is significant at the 0.05 level.

• **Conclusion:**

The level of quality of life for married group was *significantly higher* than widow group (p < 0.001).

Single group showed *significantly higher* level of quality of life than widow group (p < 0.001)
Pearson’s correlation

- Pearson’s correlation is used to examine associations between variables (represented by continuous data) by looking at the direction and strength of the associations.

Example

- Distress → Quality of life → Self-esteem

 Checking outlier

→ “Graphs > Legacy Dialogs > Scatter/Dot…”

Research interest

Data type
Pearson’s correlation

- Analyze > Correlate > Bivariate
Pearson’s correlation

- **Output:**

![Pearson's Correlation Table]

- **Conclusion:**

There were statistically significant negative correlations between quality of life and distress ($r = -0.708$, $p < 0.001$) and between self-esteem and distress ($r = -0.685$, $p < 0.001$).

There was statistically significant positive correlation between quality of life and self-esteem ($r = 0.660$, $p < 0.001$).
Ordinary least squares linear regression

• Ordinary least squares (OLS) or Linear regression is used to **explain/predict** the phenomenon of interest (**continuous data**) of interest

• Example

  **Simple OLS regression**
  - Dependent V (Quality of life)
  - Explain/Predict?
  - IV 1 (Distress level)

  **Multiple OLS regression**
  - Dependent V (Quality of life)
  - Explain/Predict?
  - IV 1 (Distress level)
  - IV 2 (Self-esteem)
  - IV 3 (Gender)

• Model assumptions
  1. Independence
  2. Normality
  3. Equal variance
  4. Linearity
Ordinary least squares linear regression

- Analyze > Regression > Linear ...

Distress
Esteem_total
Gender
Age
Ordinary least squares linear regression

- Assumption checking (2) – (4)

Analyze > Regression > Linear…
Ordinary least squares linear regression

• Assumption checking

(2) Errors are normally distributed (Normality)
→ Checked by histogram or QQ plots of the errors
Ordinary least squares linear regression

- Assumption checking

(3) Error variances are equal across all the predicted values

(Equal variance)
Ordinary least squares linear regression

• Assumption checking

(4) A linear model is appropriate (**Linearity**)

→ DV is explained by a linear combination of the IVs
→ If the linear model is appropriately specified, residuals are randomly scatter around the residual value of 0
Ordinary least squares linear regression

• Output

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.745</td>
<td>.555</td>
<td>.546</td>
<td>2.932</td>
</tr>
</tbody>
</table>

**ANOVA**

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>4</td>
<td>506.964</td>
<td>58.992</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>189</td>
<td>8.594</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>193</td>
<td>8.594</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: Quality of life
b. Predictors: (Constant), Age, Gender, Esteem_total, Distress

• Conclusion:

Approximately, 55% of the variability in the quality of life was explained by the variables in the regression model. The overall regression model significantly explained the quality of life.
Distress and self-esteem significantly predicted the level of quality of life.

We would expect -0.416 points decrease in quality of life for every one point increase in distress, assuming all the other variables are held constant.
Summary

• SPSS environment
• Data preparation in SPSS
• Data management in SPSS
• Descriptive statistics in SPSS
• Statistical models in SPSS

→ Try your own quantitative analysis in SPSS!
RESEARCH COMMONS: AN INTERDISCIPLINARY RESEARCH-DRIVEN LEARNING ENVIRONMENT

- Literature review
- Systematic Review Search Methodology
- Citation Management
- Thesis Formatting
- Nvivo Software Support
- SPSS Software Support
- R Group
- Multi-Disciplinary Graduate Student Writing Group
SPSS SERVICES BY RESEARCH COMMONS

• Workshops

• One-on-one Consultation

Request form to book SPSS consultation:

THANK YOU!

QUESTION, COMMENT, IDEAS
FEEDBACK