Statistics for a Single Measure (Univariate)

(Click icon for audio)
Taming the Statistical Beast
Walkup’s First Laws of Statistics

Law No. 1
Everything correlates with everything, especially when the same individual defines the variables to be correlated.

Law No. 2
It won’t help very much to find a good correlation between the variable you are interested in and some other variable that you don’t understand any better.

Law No. 3
Unless you can think of a logical reason why two variables should be connected as cause and effect, it doesn’t help much to find a correlation between them. In Columbus, Ohio, the mean monthly rainfall correlates very nicely with the number of letters in the names of the months!

Evidence

Good afternoon, Mrs. Ferngreen.
Is your son Johnny home?

Oh, hello officer. Please come in.
I was just formulating a null hypothesis.

He: My little Johnny's a good boy.

It says here he stole a car last Monday,

Yes, but he said he only missed his bus and needed a way home.

and he stole another car last Tuesday,

It wasn't his fault; the driver left his keys in the ignition.

and another one last Wednesday,

Well, the driver was careless and forgot to take his ignition wires out of the dashboard.

and a fourth one last Thursday.

Ummm, maybe Johnny isn't a very good boy after all.

A dramatization of the philosophy of statistical testing. The null hypothesis always gets the benefit of the doubt, and continues to be accepted until evidence to the contrary becomes statistically overwhelming (i.e., if the null hypothesis were true, the observed results would have a very slim chance of occurring).
Descriptive Analysis

Transformation of raw data into a form that make them easy to understand and interpret; rearranging, ordering, and manipulating data to generate descriptive information
Analysis Begins with Tabulation

COUNTING NOSES
Tabulation

- Tabulation: Orderly arrangement of data in a table or other summary format
- Frequency table
- Percentages
Frequency Table

• Numerical data arranged in a row-and-column format that shows the count and percentages of responses or observations for each category assigned to a variable

• Pre-selected categories
# Sample Frequency Table

<table>
<thead>
<tr>
<th>Category label</th>
<th>Code</th>
<th>Number</th>
<th>Percentage</th>
<th>Adjusted percentage</th>
<th>Adjusted cumulative percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under $5,000</td>
<td>1</td>
<td>25</td>
<td>6.3</td>
<td>6.9</td>
<td>6.9</td>
</tr>
<tr>
<td>$5,000–$9,999</td>
<td>2</td>
<td>40</td>
<td>10.0</td>
<td>11.1</td>
<td>18.0</td>
</tr>
<tr>
<td>$10,000–$14,999</td>
<td>3</td>
<td>60</td>
<td>15.0</td>
<td>16.7</td>
<td>34.7</td>
</tr>
<tr>
<td>$15,000–$19,999</td>
<td>4</td>
<td>70</td>
<td>18.0</td>
<td>19.4</td>
<td>54.1</td>
</tr>
<tr>
<td>$20,000–$24,999</td>
<td>5</td>
<td>90</td>
<td>22.5</td>
<td>25.0</td>
<td>79.1</td>
</tr>
<tr>
<td>$25,000–$34,999</td>
<td>6</td>
<td>40</td>
<td>10.0</td>
<td>11.1</td>
<td>90.2</td>
</tr>
<tr>
<td>$35,000–$49,999</td>
<td>7</td>
<td>25</td>
<td>6.3</td>
<td>6.9</td>
<td>97.1</td>
</tr>
<tr>
<td>$50,000 and over</td>
<td>8</td>
<td>10</td>
<td>7.5</td>
<td>2.8</td>
<td>99.9</td>
</tr>
<tr>
<td>No answer</td>
<td>9</td>
<td>40</td>
<td>10.0</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>400</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample SPSS Frequency Output

<table>
<thead>
<tr>
<th>Value Label</th>
<th>Value</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cum Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>liked it very much</td>
<td>1.00</td>
<td>36</td>
<td>6.7</td>
<td>6.7</td>
<td>6.7</td>
</tr>
<tr>
<td>liked it</td>
<td>2.00</td>
<td>169</td>
<td>31.3</td>
<td>31.6</td>
<td>38.3</td>
</tr>
<tr>
<td>neither liked/disliked</td>
<td>3.00</td>
<td>191</td>
<td>35.4</td>
<td>35.7</td>
<td>74.0</td>
</tr>
<tr>
<td>disliked it</td>
<td>4.00</td>
<td>90</td>
<td>16.7</td>
<td>16.8</td>
<td>90.8</td>
</tr>
<tr>
<td>disliked it very much</td>
<td>5.00</td>
<td>49</td>
<td>9.1</td>
<td>9.2</td>
<td>100.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>540</td>
<td>100.0</td>
<td></td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Valid cases 535  Missing cases 5
SPSS Histogram Output

Histogram of Purchase Price

- Std. Dev. = 8.04
- Mean = 134
- N = 721.00
Base

- Number of respondents or observations (in a row or column) used as a basis for computing percentages
- Possible bases
  - All respondents
  - Respondents who were asked question
  - Respondents who answered question
- Be careful with multi-response questions
Measures of Central Tendency

• **Mode** - the value that occurs most often
• **Median** - midpoint of the distribution
• **Mean** - arithmetic average
  – $\mu$, population
  – $\bar{x}$, sample
Measures of Dispersion or Spread for Single Measure

- Range
- Inter-quartile range
- Mean absolute deviation
- Variance
- Standard deviation
Low Dispersion

Value on Variable

Frequency
High Dispersion

Value on Variable

Frequency

- 5
- 4
- 3
- 2
- 1
Range as a Measure of Spread

• Difference between smallest and largest value in a set
• Range = largest value – smallest value
• Inter-quartile range = 75\textsuperscript{th} percentile – 25\textsuperscript{th} percentile
Deviation Scores

Differences between each observed value and the mean

\[ d_i = x_i - \bar{x} \]
Average Deviation

$$\frac{\sum (X_i - \bar{X})}{n} = 0$$
Mean Squared Deviation

$$\sum \left( X_i - \bar{X} \right)^2 \over n$$
Variance

Population

\( \sigma^2 \)

Sample

\( S^2 \)
Variance

\[ S^2 = \frac{\sum(X - \overline{X})^2}{n - 1} \]
Variance

- Variance is given in squared units
- Standard deviation is the square root of variance

\[ S = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n - 1}} \]
### Summary: Central Tendency and Dispersion

<table>
<thead>
<tr>
<th>Type of Scale</th>
<th>Measure of Central Tendency</th>
<th>Measure of Dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Mode</td>
<td>None</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Median</td>
<td>Percentile</td>
</tr>
<tr>
<td>Interval or ratio</td>
<td>Mean</td>
<td>Standard deviation</td>
</tr>
</tbody>
</table>
Distributions for Single Measures
Symmetric Distribution
Skewed Distributions
Normal Distribution

• Bell shaped
• Symmetrical about its mean
• Mean identifies highest point
• Almost all values are within ±3 standard deviations
• Infinite number of cases—a continuous distribution
Normal Distribution
Standard Normal Curve

• The curve is bell-shaped or symmetrical
• About 68% of the observations will fall within 1 standard deviation of the mean
• About 95% of the observations will fall within approximately 2 (1.96) standard deviations of the mean
• Almost all of the observations will fall within 3 standard deviations of the mean
Normal Curve: IQ Example
Standardized Normal Distribution

- Area under curve has a probability density $= 1.0$
- Mean $= 0$
- Standard deviation $= 1$
Standardized Normal Curve
Standardized Normal is $Z$ Distribution
Standardized Scores

Used to compare an individual value to the population mean in units of the standard deviation

\[ z = \frac{x - \mu}{\sigma} \]
Linear Transformation of Any Normal Variable into a Standardized Normal Variable

\[ z = \frac{x - \mu}{\sigma} \]

Sometimes the scale is stretched

Sometimes the scale is shrunk
Data Transformation

- Data conversion
- Changing the original form of the data to a new format
- More appropriate data analysis
- New variables
Data Transformation

Summative Score = VAR1 + VAR2 + VAR 3
Index Numbers

• Score or observation recalibrated to indicate how it relates to a base number
• CPI - Consumer Price Index
Recap

• Count/frequency data
• Measures of central tendency
• Dispersion from central tendency
• Data distributions
  – Symmetric versus skewed
  – (Standardized) normal
• Data transformation