Descriptive Analysis

• The transformation of raw data into a form that will make them easy to understand and interpret; rearranging, ordering, and manipulating data to generate descriptive information
Type of Measurement

Nominal

- Two categories
- More than two categories

Type of descriptive analysis

- Frequency table
  - Proportion (percentage)
  - Frequency table
    - Category proportions (percentages)
    - Mode

Type of Measurement

Ordinal

Type of descriptive analysis

- Rank order
  - Median
**Type of Measurement**
- Interval

**Type of descriptive analysis**
- Arithmetic mean

**Type of Measurement**
- Ratio

**Type of descriptive analysis**
- Index numbers
  - Geometric mean
  - Harmonic mean
Tabulation

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

Frequency Table

- The arrangement of statistical data in a row-and-column format that exhibits the count of responses or observations for each category assigned to a variable
### Central Tendency

<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</td>
</tr>
<tr>
<td>deviation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Cross-Tabulation

- A technique for organizing data by groups, categories, or classes, thus facilitating comparisons; a joint frequency distribution of observations on two or more sets of variables
- Contingency table- The results of a cross-tabulation of two variables, such as survey questions
Cross-Tabulation

• Analyze data by groups or categories
• Compare differences
• Contingency table
• Percentage cross-tabulations

Base

• The number of respondents or observations (in a row or column) used as a basis for computing percentages
Elaboration and Refinement

• Moderator variable
  – A third variable that, when introduced into an analysis, alters or has a contingent effect on the relationship between an independent variable and a dependent variable.
  – Spurious relationship
    • An apparent relationship between two variables that is not authentic.

Quadrant Analysis

Two rating scales

4 quadrants two-dimensional table

Importance-Performance Analysis)
Data Transformation

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

Summative Score = VAR1 + VAR2 + VAR3
Collapsing a Five-Point Scale

- Strongly Agree
- Agree
- Neither Agree nor Disagree
- Disagree
- Strongly Disagree

- Strongly Agree/Agree
- Neither Agree nor Disagree
- Disagree/Strongly Disagree

Index Numbers

- Score or observation recalibrated to indicate how it relates to a base number
- CPI - Consumer Price Index
Calculating Rank Order

• Ordinal data
• Brand preferences

Tables

• Bannerheads for columns
• Studheads for rows
Charts and Graphs

- Pie charts
- Line graphs
- Bar charts
  - Vertical
  - Horizontal
Line Graph

Bar Graph
Computer Programs

- SPSS
- SAS
- SYSTAT
- Microsoft Excel
- WebSurveyor
Microsoft Excel - Data Analysis

The Paste Function Provides Numerous Statistical Operations
Computer Programs

- Box and whisker plots
- Interquartile range - midspread
- Outlier
Interpretation

• The process of making pertinent inferences and drawing conclusions
• concerning the meaning and implications of a research investigation

Research Methods

William G. Zikmund

Univariate Statistics
Univariate Statistics

• Test of statistical significance
• Hypothesis testing one variable at a time

Hypothesis

• Unproven proposition
• Supposition that tentatively explains certain facts or phenomena
• Assumption about nature of the world
Hypothesis

• An unproven proposition or supposition that tentatively explains certain facts or phenomena
  – Null hypothesis
  – Alternative hypothesis

Null Hypothesis

• Statement about the status quo
• No difference
Alternative Hypothesis

• Statement that indicates the opposite of the null hypothesis

Significance Level

• Critical probability in choosing between the null hypothesis and the alternative hypothesis
Significance Level

- Critical Probability
- Confidence Level
- Alpha
- Probability Level selected is typically .05 or .01
- Too low to warrant support for the null hypothesis

The null hypothesis that the mean is equal to 3.0:

$$H_o : \mu = 3.0$$
The alternative hypothesis that the mean does *not* equal to 3.0:

\[ H_1 : \mu \neq 3.0 \]
Critical values of $\mu$

Critical value - upper limit

$$= \mu + ZS_{\bar{X}} \quad \text{or} \quad \mu + Z \frac{S}{\sqrt{n}}$$

$$= 3.0 + 1.96 \left( \frac{1.5}{\sqrt{225}} \right)$$

$\text{Health Economics Research Method 2003/2}$

---

Critical values of $\mu$

$$= 3.0 + 1.96(0.1)$$

$$= 3.0 + .196$$

$$= 3.196$$

$\text{Health Economics Research Method 2003/2}$
Critical values of $\mu$

Critical value - lower limit

$$= \mu - Z S \frac{1}{\bar{x}} \quad \text{or} \quad \mu - Z \frac{S}{\sqrt{n}}$$

$$= 3.0 - 1.96 \left( \frac{1.5}{\sqrt{225}} \right)$$

Critical values of $\mu$

$$= 3.0 - 1.96(0.1)$$

$$= 3.0 - 0.196$$

$$= 2.804$$
Region of Rejection

Hypothesis Test $\mu = 3.0$
Type I and Type II Errors

<table>
<thead>
<tr>
<th>Null is true</th>
<th>Accept null</th>
<th>Reject null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct-no error</td>
<td>Type I error</td>
<td></td>
</tr>
<tr>
<td>Null is false</td>
<td>Type II error</td>
<td>Correct-no error</td>
</tr>
</tbody>
</table>

State of Null Hypothesis in the Population

<table>
<thead>
<tr>
<th>Decision</th>
<th>Accept Ho</th>
<th>Reject Ho</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ho is true</td>
<td>Correct--no error</td>
<td>Type I error</td>
</tr>
<tr>
<td>Ho is false</td>
<td>Type II error</td>
<td>Correct--no error</td>
</tr>
</tbody>
</table>
Calculating $Z_{obs}$

$$Z_{obs} = \frac{x - \mu}{S_x}$$

Alternate Way of Testing the Hypothesis

$$Z_{obs} = \frac{\bar{X} - \mu}{S_{\bar{X}}}$$
Alternate Way of Testing the Hypothesis

\[ Z_{obs} = \frac{3.78 - \mu}{S_{\bar{X}}} = \frac{3.78 - 3.0}{.1} = \frac{0.78}{.1} = 7.8 \]

Choosing the Appropriate Statistical Technique

- Type of question to be answered
- Number of variables
  - Univariate
  - Bivariate
  - Multivariate
- Scale of measurement
t-Distribution

- Symmetrical, bell-shaped distribution
- Mean of zero and a unit standard deviation
- Shape influenced by degrees of freedom
Degrees of Freedom

- Abbreviated d.f.
- Number of observations
- Number of constraints

Confidence Interval Estimate
Using the t-distribution

\[ \mu = \bar{X} \pm t_{c.l.} \frac{S}{\sqrt{n}} \]

or

\[
\begin{align*}
\text{Upper limit} & = \bar{X} + t_{c.l.} \frac{S}{\sqrt{n}} \\
\text{Lower limit} & = \bar{X} - t_{c.l.} \frac{S}{\sqrt{n}}
\end{align*}
\]
Confidence Interval Estimate Using the t-distribution

\[ \mu = \bar{X} \pm t_{c.l.} \frac{s}{\sqrt{n}} \]

\( \bar{X} = 3.7 \)
\( S = 2.66 \)
\( n = 17 \)
upper limit = 3.7 + 2.12(2.66\sqrt{17})
= 5.07

Lower limit = 3.7 - 2.12(2.66\sqrt{17})
= 2.33
Suppose that a production manager believes the average number of defective assemblies each day to be 20. The factory records the number of defective assemblies for each of the 25 days it was opened in a given month. The mean $X$ was calculated to be 22, and the standard deviation, $S$, to be 5.
$H_0 : \mu = 20$

$H_1 : \mu \neq 20$

\[ S_{\bar{X}} = S \div \sqrt{n} \]
\[ = 5 \div \sqrt{25} \]
\[ = 1 \]
Univariate Hypothesis Test Utilizing the t-Distribution

The researcher desired a 95 percent confidence, and the significance level becomes .05. The researcher must then find the upper and lower limits of the confidence interval to determine the region of rejection. Thus, the value of $t$ is needed. For 24 degrees of freedom ($n-1, 25-1$), the $t$-value is 2.064.

Lower limit:

$$
\mu - t_{c.l.} \frac{S}{\bar{X}} = 20 - 2.064 \left( \frac{5}{\sqrt{25}} \right) \\
= 20 - 2.064(1) \\
= 17.936
$$
Upper limit:

$$\mu + t_{c.l.} S_{\bar{X}} = 20 + 2.064 \left( \frac{5}{\sqrt{25}} \right)$$

$$= 20 + 2.064(1)$$

$$= 20.064$$
Testing a Hypothesis about a Distribution

• Chi-Square test
• Test for significance in the analysis of frequency distributions
• Compare observed frequencies with expected frequencies
• “Goodness of Fit”

\[ x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \]
Chi-Square Test

$x^2 = \text{chi-square statistics}$

$O_i = \text{observed frequency in the } i^{\text{th}} \text{ cell}$

$E_i = \text{expected frequency on the } i^{\text{th}} \text{ cell}$

Chi-Square Test

Estimation for Expected Number for Each Cell

$$E_{ij} = \frac{R_i C_j}{n}$$
Chi-Square Test
Estimation for Expected Number for Each Cell

\[ R_i = \text{total observed frequency in the } i^{th} \text{ row} \]
\[ C_j = \text{total observed frequency in the } j^{th} \text{ column} \]
\[ n = \text{sample size} \]

Univariate Hypothesis Test
Chi-square Example

\[ X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \]
Univariate Hypothesis Test
Chi-square Example

\[ X^2 = \frac{(60 - 50)^2}{50} + \frac{(40 - 50)^2}{50} \]

\[ = 4 \]

Hypothesis Test of a Proportion

\( \pi \) is the population proportion
\( p \) is the sample proportion
\( \pi \) is estimated with \( p \)
Hypothesis Test of a Proportion

\[
H_0 : \pi = 0.5 \\
H_1 : \pi \neq 0.5
\]

\[
S_p = \sqrt{\frac{(0.6)(0.4)}{100}} = \sqrt{\frac{0.24}{100}} = \sqrt{0.0024} = 0.04899
\]
\[ Z_{obs} = \frac{p - \pi}{S_p} = \frac{.6 - .5}{.04899} = 2.04 \]

Hypothesis Test of a Proportion:
Another Example

\[ n = 1200 \]
\[ p = .20 \]
\[ S_p = \sqrt{\frac{pq}{n}} \]
\[ S_p = \sqrt{\frac{.2(.8)}{1200}} \]
\[ S_p = \sqrt{\frac{.16}{1200}} \]
\[ S_p = \sqrt{.000133} \]
\[ S_p = .0115 \]
Hypothesis Test of a Proportion: Another Example

\[ n = 1200 \]
\[ p = .20 \]
\[ S_p = \sqrt{\frac{pq}{n}} \]
\[ S_p = \sqrt{\frac{(.2)(.8)}{1200}} \]
\[ S_p = \sqrt{\frac{.16}{1200}} \]
\[ S_p = \sqrt{.000133} \]
\[ S_p = .0115 \]

\[ Z = \frac{p - \pi}{S_p} \]
\[ Z = \frac{.20 - .15}{.0115} \]
\[ Z = \frac{.05}{.0115} \]
\[ Z = 4.348 \]

The Z value exceeds 1.96, so the null hypothesis should be rejected at the .05 level. Indeed it is significant beyond the .001.