Univariate data analysis

- Measures of central tendency
  - Mean, median, mode
- Measures of spread
  - Variance, standard deviation, standard error
- Measures of skew

Variance and standard deviation

- The variance ($\sigma^2$) of a distribution is the mean of the squares of the deviations of each observation from the mean
  
  $$\sigma^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n - 1}$$

- The standard deviation ($\sigma$) of a distribution is the square root of the variance
  
  $$\sigma = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n - 1}}$$

Step 1 – look at the data

- Examine the distribution
  - histogram
  - scatter plot
- Identify outliers
Step 2 – Describe the data

• What is the most effective way to present the information contained in the data?
  – Table, histogram, line graph, box plot, etc.
• Median vs. mean
  – Income and age at marriage

Skewed distributions

Bimodal distributions
Box plot: a convenient way to describe the central tendency and spread of a distribution

![Box plot diagram]

Descriptive tables

Table 3: Descriptive Statistics and z-Tests: Unique-Mean Comparisons by Report Type for Variables Used in the Short-term Analyses (Y = 1992) (N = 914)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.24</td>
<td>0.21</td>
<td>0.24</td>
<td>0.21</td>
</tr>
<tr>
<td>Proportional survival (N = 1992)</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>Proportion of those in need (N = 1992)</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Proportion of those in over 65 (N = 1992)</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Proportion of those in over 75 (N = 1992)</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Proportion of those in over 85 (N = 1992)</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Proportion of those in over 95 (N = 1992)</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Proportion of those in over 100 (N = 1992)</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Source: Data from the National Longitudinal Mortality Study.

* Significant at the 0.05 level.
** Significant at the 0.01 level.
Recoding

Easier to interpret?

Too much formatting