Lecture 1: Review and Exploratory Data Analysis (EDA)

Sandy Eckel  
seckel@jhsph.edu  
Department of Biostatistics,  
The Johns Hopkins University, Baltimore USA  

21 April 2008

Course Information I

Course website:  
http://www.biostat.jhsph.edu/~seckel/biostat2/  
Office hours  
- For questions and help  
- When? I’ll announce this tomorrow  
Homework  
- Three assignments  
- Follow-up on material from class  
Written exam  
- When: Wednesday 21 May, 10.00 - 12.00  
- Where: Multimedia classroom (C wing, 2nd floor), Mannerheimintie 172/Kytosuontie 9

Course Information II

21 April 2008 to 21 May 2008  
08.30 - 12.30 Monday, Tuesday, Thursday, Friday  
- 08.30 - 10.15 Lecture  
- 10.15 - 10.30 Break  
- 10.30-12.30 Informal lecture, class exercise or computer lab  
Activities for the second half of class will vary; also time for questions!

Class goals

Biostat I  
- Numbers and probability  
- Sampling distributions and inference  
- Statistical models and association / causality

Biostat II  
- Developing scientific questions  
- Translating questions into regression models  
- Interpreting results of regression  
- Critiquing the literature
Issues and recurring themes

- Populations are complicated... statistical techniques may not capture all of the nuances
- Natural laws will not perfectly predict outcomes
- Signal-to-Noise: Comparing a trend to its variability
- Bias-Variance trade-off: Unadjusted vs. adjusted estimates
- Population vs. sample

What is Biostatistics?

Biostatistics is the use of data to describe and make inferences about a scientific problem
- Remember the “Bio” in Biostatistics!
- Biostatistics has limitations: you can’t have it all

Types of Biostatistics

1. Descriptive statistics
   - Exploratory data analysis (EDA): often not in literature
   - Summaries: “Table 1” in a paper
   - Goal: to visualize relationships, generate hypotheses

2. Inferential statistics
   - Confirmatory data analysis
   - Methods section of a paper
   - Goal: quantify relationships, test hypotheses

Exploratory Data Analysis (EDA)

- ALWAYS look at your data!
- If you can’t see it, then don’t believe it!
- EDA allows us to:
  1. Visualize distributions and relationships
  2. Detect errors
  3. Assess assumptions for confirmatory analysis
- EDA is the first step of data analysis
EDA methods (One-Way)

- Ordering: Stem-and-Leaf plots
- Grouping: frequency displays, distributions; histograms
- Summaries: summary statistics, standard deviation, box-and-whisker plots

Stem-and-Leaf Plots I

Age in years (10 observations):
25, 26, 29, 32, 35, 36, 38, 44, 49, 51

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>5 6 9</td>
</tr>
<tr>
<td>30-39</td>
<td>2 5 6 8</td>
</tr>
<tr>
<td>40-49</td>
<td>4 9</td>
</tr>
<tr>
<td>50-59</td>
<td>1</td>
</tr>
</tbody>
</table>

Stem-and-Leaf Plots II

- The age interval is the “stem”
- The observations are the “leaves”
- Rule of thumb:
  - The number of stems should roughly equal the square root of the number of observations
  - Or the stems should be logical categories

Stem-and-Leaf Plots III

Some statistical programs print output like this:

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>2*</td>
<td>5 6 9</td>
</tr>
<tr>
<td>3*</td>
<td>2 5 6 8</td>
</tr>
<tr>
<td>4*</td>
<td>4 9</td>
</tr>
<tr>
<td>5*</td>
<td>1</td>
</tr>
</tbody>
</table>

where 2* means 20-29.
Stem-and-Leaf Plots IV

Output may also be shown like this:

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>5 6 9</td>
</tr>
<tr>
<td>3*</td>
<td>2</td>
</tr>
<tr>
<td>3.0</td>
<td>5 6 8</td>
</tr>
<tr>
<td>4*</td>
<td>4</td>
</tr>
<tr>
<td>4.0</td>
<td>9</td>
</tr>
<tr>
<td>5*</td>
<td>1</td>
</tr>
</tbody>
</table>

where 3* means 30-34 and 3. means 35-39.

Frequency Distribution Tables

- Shows the number of observations for each range of data
- Intervals can be chosen in ways similar to stem-and-leaf displays

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>3</td>
</tr>
<tr>
<td>30-39</td>
<td>4</td>
</tr>
<tr>
<td>40-49</td>
<td>2</td>
</tr>
<tr>
<td>50-59</td>
<td>1</td>
</tr>
</tbody>
</table>

Cumulative Frequency Distribution Tables

Show the frequency, the relative frequency, and cumulative frequency of observations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>3</td>
<td>3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>30-39</td>
<td>4</td>
<td>7</td>
<td>0.4</td>
<td>0.7</td>
</tr>
<tr>
<td>40-49</td>
<td>2</td>
<td>9</td>
<td>0.2</td>
<td>0.9</td>
</tr>
<tr>
<td>50-59</td>
<td>1</td>
<td>10</td>
<td>0.1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

This table shows an empirical distribution function obtained from a sample

The true distribution function is the distribution of the entire population

Histograms

Picture of the frequency or relative frequency distribution

Note: Graphs are generally better to use in presentations than tables. They allow your audience to visualize a trend quickly.
Summary Statistics

- Percentiles
- Measures of central tendency
- Measures of dispersion or variability

Percentiles

The $r^{th}$ percentile, $P_r$, is the value that is greater than or equal to $r$ percent of a sample of $n$ observations or less than or equal to $(100-r)$ percent of the observations.

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Quartile</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{25}$</td>
<td>$Q_1$</td>
<td>$\frac{n+1}{4}$th observation</td>
</tr>
<tr>
<td>$P_{50}$</td>
<td>$Q_2$</td>
<td>$\frac{n+1}{2}$th observation</td>
</tr>
<tr>
<td>$P_{75}$</td>
<td>$Q_3$</td>
<td>$\frac{3(n+1)}{4}$th observation</td>
</tr>
</tbody>
</table>

Calculating quartiles I

From the age data:

25, 26, 29, 32, 35, 36, 38, 44, 49, 51

with $n=10$

$Q_2 = \text{median} = \text{average of 5th and 6th observations} = \frac{35 + 36}{2} = 35.5$

Remember to order your data!

Calculating quartiles II

$Q_1 = \text{median of lower half of data} = \text{third smallest value} = 29$

$Q_3 = \text{median of upper half of data} = \text{third largest value} = 44$

Note: If $n$ is odd, include the median in the upper and lower half of the data.
Measures of Central Tendency

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>( \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} )</td>
</tr>
<tr>
<td>Median</td>
<td>Middle observation</td>
</tr>
<tr>
<td>Mode</td>
<td>Most frequent observation</td>
</tr>
</tbody>
</table>

From the age example the mean is:
\[ \frac{25+26+29+32+35+36+38+44+49+51}{10} = 36.5 \]

The mode is more helpful for categorical data, i.e. the most frequent age interval is 30-39 and it has 4 observations.

Measures of spread: Range

- Range = max-min
- The difference between the maximum and minimum values

From age example:
- Max = 51, Min = 25
- Range = 51-25 = 26

Measures of spread: Variance

Variance = Expected value of the squared deviation of the observations from the true mean
\[ \sigma^2 = E[(X - \bar{X})^2] \]

Sample variance = Average of the squared deviation of the observations from the sample mean
\[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \]

Sample variance from age example = 82.9

Standard deviation

Standard deviation = Square root of the variance
\[ \sigma = \sqrt{E[(X - \bar{X})^2]} \]

Sample standard deviation = Square root of the sample variance
\[ s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \]

From the age data: \( s = \sqrt{82.9} = 9.1 \)

Note: The units of the variance are years\(^2\), while the units of the standard deviation are years.

Interpretation: The standard deviation gives an idea of how much observations differ from the mean.
Box-and-whisker plots I

- Box-and-whisker plots display quartiles
- Some terminology:
  - Upper Hinge = $Q_3$ = Third quartile
  - Lower Hinge = $Q_1$ = First quartile
  - Interquartile range (IQR) = $Q_3 - Q_1$
    Contains the middle 50% of data
  - Upper Fence = Upper Hinge + 1.5 * (IQR)
  - Lower Fence = Lower Hinge - 1.5 * (IQR)
  - Outliers: Data values beyond the fences

“Whiskers” are drawn to the smallest and largest observations within the fences

Box-and-whisker plots II

![Boxplot of Age](image)

- IQR = 44-29 = 15
- Upper Fence = 44 + 15*1.5 = 66.5
- Lower Fence = 29 - 15*1.5 = 6.5

Pairwise EDA

2 Categorical Variables

- Frequency table
- Stratified stem-and-leaf plots
- Side-by-side box plots
- Scatterplot

1 Categorical, 1 Continuous Variable

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>Gender</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-29</td>
<td>Female</td>
<td>1</td>
</tr>
<tr>
<td>30-39</td>
<td>Male</td>
<td>2</td>
</tr>
<tr>
<td>40-49</td>
<td>Female</td>
<td>1</td>
</tr>
<tr>
<td>50-59</td>
<td>Male</td>
<td>0</td>
</tr>
</tbody>
</table>

| Total        | 5      | 5     | 10    |

It looks like the men tend to be younger than women in this example.
1 Categorical and 1 Continuous Variable I

Stratified Stem-and-Leaf plots

<table>
<thead>
<tr>
<th>Age Interval</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs.</td>
<td>Obs.</td>
</tr>
<tr>
<td>20-29</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>30-39</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>40-49</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>50-59</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

1 Categorical and 1 Continuous Variable II

Side-by-Side Box Plots

Boxplot of Age by Gender

Allows us to compare the distribution of the continuous variable (age) across values of the categorical variable (gender)

2 Continuous Variables

Scatterplot

Age by Height

Scatterplots visually display the relationship between two continuous variables

EDA: What to notice

- Shape
- Center
- Spread
Common Distribution Shapes

- Symmetrical and bell shaped
- Positively skewed or skewed to the right
- Negatively skewed or skewed to the left

Other Distribution Shapes

- Bimodal
- Reverse J-shaped
- Uniform

Measures of Center

- Mode: Peak(s)
- Median: Equal areas point
- Mean: Balancing point

Skewness

- Positively skewed
  - Longer tail in the high values
  - Mean > Median > Mode
Skewness II

Negatively skewed
- Longer tail in the low values
- Mode > Median > Mean

Symmetric
- Right and left sides are mirror images
- Left tail looks like right tail
- Mean = Median = Mode

EDA: What to notice

Outliers
- Values that are “far” from the bulk of the data
- Outliers can influence the value of some statistical measures
- Age example

<table>
<thead>
<tr>
<th>Data</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>36.5</td>
</tr>
<tr>
<td>With 80-year-old added</td>
<td>40.5</td>
</tr>
</tbody>
</table>

Take Home Message

- Look at your data FIRST!
- Happy exploring!