Descriptive Statistics

Kari Lock
Harvard University
Department of Statistics
Rigorous Research in Engineering Education
12/3/09
Before making inferences about the population, we need to describe the data observed in our sample.

- **Visual displays**: graphs, plots
- **Summary statistics**: numerical summaries of the data

The appropriate descriptive techniques depend on the type of data collected.
Visualization
One Categorical Variable

Pie Chart

Pie Preferences
- Pumpkin 30%
- Cherry 23%
- Pecan 8%
- Coconut 3%
- Cream 3%

Goal: display the number or percent of people in each category

Bar Chart

Birthdays of Students by Month

Number of Students

Month
- Jan
- Feb
- Mar
- Apr
- May
- Jun
- Jul
- Aug
- Sep
- Oct
- Nov
- Dec
• Goal: Visually display the entire *distribution* of observed numbers. Location, spread, shape and outliers are all visible.
Visualization
One Quantitative Variable

• Symmetric:

• Skewed:

• Outlier – values that fall outside an overall pattern.
Distribution

• If a histogram is theoretically known, it is smoothed over with a distribution. This distribution represents what the hypothetical histogram would look like if we were to take many samples.
Visualization
One Categorical, One Quantitative

Side-by-Side Boxplots

Exam Scores

Males
Females

Gender

Compare the numerical distributions of two (or more) different groups
Visualization
One Categorical, Two Quantitative

Side-by-Side Boxplots

- Often it’s not obvious whether two groups are significantly different. That’s what statistics is for!
Visualization
Two Quantitative

• Goal: Visualize the relationship between two quantitative variables.
Note: Both variables have to be measured on the same set of individuals.
• **Proportion**: Report the proportion of observations that fall in each category.
Summary Statistics
One Quantitative Variable

**Mean** – the “average”: add everything up and divide by the sample size.

**Median** – the “middle” point at which 50% of the data are below and 50% are above.

**Standard Deviation** – A measure of the average distance from the mean. Indicates how “spread out” the data are.

**Variance** = (Standard Deviation)^2
Summary Statistics
One Quantitative Variable

Lower standard deviation (0.33):

Higher standard deviation (1.32):

Same mean
Summary Statistics
Two Quantitative Variables

• **Correlation (r)** a measure of linear association between two numeric variables.

• Correlation runs between -1 and 1. Values closer to -1 and 1 represent stronger relationships, values closer to 0 are weaker.
  – Positive Correlation: as one variable goes up the other goes up
  – Negative Correlation: as one variable goes up the other goes down
  – Zero Correlation means no linear association
Correlation

$r = .85$

Diagram showing scatter plots with correlation coefficients:
- $r = +1$
- $r = -1$
- $r = 0$
- $r = .7$
- $r = .95$
- $r = -.4$
Correlation Cautions

CORRELATION DOES NOT IMPLY CAUSATION
Correlation Cautions

- Correlation only measures *linear* associations:

\[ r = 0 \]
Outliers

- Correlation can be highly affected by outliers:

  $r = .78$

  $r = .17$
Outliers

• Mean and standard deviation are also highly influenced by outliers:

• Mean = 3. Standard Deviation = 0.21:

• Mean = 3.8, Standard Deviation = 1.86:
Outliers

• ALWAYS plot your data to see if you have outliers

• If you do have extreme outliers, you should first check that you haven’t made an error entering the data

• If the outliers are legitimate, you should run your analyses with and without the outliers to see how much the outliers influence the results