Introduction to Biostatistics for Clinical Researchers

University of Kansas
Department of Biostatistics
&
University of Kansas Medical Center
Department of Internal Medicine
Schedule

Friday, November 19 in 1025 Orr-Major
Friday, December 3 in 1023 Orr-Major
Friday, December 10 in 1023 Orr-Major
Friday, December 17 in B018 School of Nursing

Possibility of a 5th lecture, TBD
All lectures will be held from 8:30a - 10:30a
Materials

- PowerPoint files can be downloaded from the Department of Biostatistics website at http://biostatistics.kumc.edu

- A link to the recorded lectures will be posted in the same location
Instructor

- Jo A. Wick, PhD
  - Assistant Professor of Biostatistics
  - Email: jwick@kumc.edu
  - Office: 5028C Robinson
  - Phone: (913) 588-4790
Biostatistical Support

- Robinson, 5th floor
- (913) 588-4703
- Use the Project Registration link at top of page
Project Registration

- Routing form for the Department
  - Grant development
  - Experimental design
  - Database development
  - Data management
  - Data analysis
  - Abstract or publication help
Part I: Describing Data
Topics

- What role does statistics play in clinical research?
- Types of data: continuous, binary, categorical, time-to-event
- Numerical summary measures for continuous data
- Visual summary measures for continuous data
- Sample data versus population level data
What Role Does Biostatistics Play in Clinical Research?
Data is Everywhere

- Data is utilized and summarized frequently in the literature

- From *Archives of Internal Medicine* article, 2010:
  - **Background:** This study aimed to assess the efficacy of an intensive exercise intervention strategy in promoting physical activity (PA) and improving hemoglobin A₁c (HbA₁c) level and other modifiable cardiovascular risk factors in patients with type 2 diabetes mellitus (T2DM).
  - **Results:** . . . Compared with the control group, supervised exercise produced significant improvements in physical fitness; HbA₁c level (−0.30% [−0.49% to −0.10%]; *P* < .001); systolic (−4.2 mm Hg [−6.9 to −1.6 mm Hg]; *P* = .002) and diastolic (−1.7 mm Hg [−3.3 to −1.1 mm Hg]; *P* = .03) blood pressure; high-density lipoprotein (3.7 mg/dL [2.2 to 5.3 mg/dL]; *P* < .001) and low-density lipoprotein (−9.6 mg/dL [−15.9 to −3.3 mg/dL]; *P* = .003) . . .

Data is Everywhere

- Data is utilized and summarized with statistics frequently in popular media

- From CNN.com, Wednesday, November 17, 2010:
  - “Post-traumatic stress disorder (PTSD) affects more than the mind. The disorder may damage blood vessels and increase the risk of dying early, according to new research presented today. . . The vets with PTSD . . . had more than double the risk of dying during the 10-year study compared to their peers who didn’t have the disorder. . . A separate analysis involving heart scans . . . found that men and women with PTSD had more calcium buildup in their arteries than vets without post-traumatic stress. . . Among veterans with similar degrees of calcium buildup, those who had PTSD were 48 percent more likely to die of any cause during the study and 41 percent more likely to die from heart disease compared to those without PTSD . . .”
Data is Everywhere

- Data is utilized and summarized with statistics frequently in popular media

- From NBC News, November 16, 2010:
  - “One third of U.S. patients dying of cancer end up getting costly but futile treatment in hospitals . . . The United States spent $2.3 trillion on healthcare in 2008—$7,681 per resident, accounting for 16 percent of gross domestic product. Studies have shown about 31 percent of this . . . is spent in hospitals.”
Data Provides Information

- Good data can be analyzed and summarized to provide useful information

- Bad data can be analyzed and summarized to provide incorrect/harmful/non-informative information
Steps in a Research Project

- Planning and design of study
- Data collection
- Data analysis
- Presentation
- Interpretation

*Biostatistics CAN play a role in each of these steps*
Biostatistics Issues

- Planning/design of studies
  - Primary question(s) of interest
    - Quantifying information about a single group?
    - Comparing multiple groups?
  - Sample size
    - How many subjects total?
    - How many in each of the groups to be compared?
  - Selecting study participants
    - Randomly chosen from “master list”?
    - Selected from a pool of interested persons?
    - Take whomever shows up?
  - If group comparison of interest, how to assign to groups?
Biostatistics Issues

- Data collection
- Data analysis
  - What statistical methods are appropriate given the data collected?
  - Dealing with variability (both natural and sampling related)
    - Important patterns in data are obscured by variability
    - Distinguish real patterns from random variation
  - Inference: using information from the single study coupled with information about variability to make statement(s) about the larger population or process of interest
Biostatistics Issues

- Presentation
  - What summary measures will best convey the “main messages” in the data about the primary (and secondary) research questions of interest?
  - How best to convey/rectify uncertainty in estimates based on the data?

- Interpretation
  - What do the results mean in terms of practice, the population, etc.?
1954 Salk Polio Vaccine Trial

School Children

Vaccinated
N = 200,745
82 cases

Placebo
N = 201,229
162 cases

Design: Features of the Polio Trial

- Comparison group
- Randomized
- Placebo controls
- Double blind

**Objective:** the groups should be equivalent except for the factor (vaccine) being investigated
Analysis Question

- There were almost twice as many polio cases in the placebo compared to the vaccine group—could the results be due to chance?
Such Great Imbalance by Chance?

- Polio cases
  - Vaccine: 82
  - Placebo: 162

- Statistical methods tell us how to make these probability calculations
Binary Data

- Binary (dichotomous) data
  - Yes/No
  - Polio: yes/no
  - Cure: yes/no
  - Sex: male/female (or as yes/no, “Is subject male?”)
Categorical Data

- Categorical data places observations into mutually exclusive categories

- **Nominal** categorical data: no inherent order to categories
  - Race/ethnicity
  - Country of birth
  - Religious affiliation

- **Ordinal** categorical data: ordered categories
  - Income level categorized
  - Level of agreement from strongly disagree to strongly agree (“Likert-scale data”)
Continuous Data

- Continuous data consists of finer measurements (and more information)
  - Blood pressure, mmHg
  - Weight, lbs (kg, oz, etc.)
  - Height, ft (cm, in, etc.)
  - Age, years (months)
  - Income level, dollars/year
Time-to-Event Data

- Data that is a hybrid of continuous and binary data
  - Whether an event occurs (yes/no) and time to the occurrence (or time to last follow-up in the absence of event)
Different Methods for Different Types of Data

- To compare the number of polio cases in the two treatment arms of the Salk Polio vaccine, you could use . . .
  - Fishers’ Exact Test
  - Chi-Square Test

- To compare blood pressures in a clinical trial evaluating two blood pressure-lowering medications, you could use . . .
  - Two-sample t-Test
  - Wilcoxon Rank Sum Test
Continuous Data: Numerical Summary Measures & Sample Estimates versus Population Parameters
Summarizing and Describing Continuous Data

- Measures of the center of data
  - Mean
  - Median

- Measure of data variability
  - Standard deviation (variance)
  - Range
Sample Mean: The Average or Arithmetic Mean

- Five systolic blood pressures (mmHg) \((n = 5)\)
  - 120, 80, 90, 110, 95

- Can be represented with mathematical notation
  - \(x_1 = 120, x_2 = 80, \ldots, x_5 = 95\)

- The sample mean is easily computed by adding up the five values and dividing by five—in statistical notation the sample mean is frequently represented by a letter with a line over it

\[
\bar{x} = \frac{120 + 80 + 90 + 110 + 95}{5} = 99 \text{ mmHg}
\]
Notes on Sample Mean

- Generic formula representation

\[ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \]

- In the formula we use the summation sign, \( \Sigma \), which is simply shorthand for “add up all the observations”

\[ \sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n \]
Notes on Sample Mean

- Also called the sample average or arithmetic mean

- Sensitive to extreme values
  - One data point could make a great change in the sample mean
Sample Median

- The median is the middle number (also called the 50th percentile)
  - Other percentiles can be computed but are not measures of central tendency

80 90 95 110 120
Sample Median

- The sample median is not sensitive to extreme values
  - Example: If 120 became 200, the median would remain the same

- What would happen to the sample mean?
Sample Median

- If the sample size is an even number, a “middle value” does not exist

- Depending on who you ask . . .

\[
\begin{align*}
80 & \quad 90 & \quad 95 & \quad 110 & \quad 120 & \quad 125 \\
M & \quad & \quad & \quad & \quad & \quad \\
M &= \frac{95 + 110}{2} = 102.5 \text{ mmHg}
\end{align*}
\]
Describing Variability

- Sample variance ($s^2$)
- Sample standard deviation ($s$ or $SD$)
- The sample variance is the average squared deviation of observations from the sample mean

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$
Describing Variability

- The sample standard deviation is the square root of the sample variance

\[ s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \]

- \( s \) is an “average distance” of observations from the sample mean reported in the same units as the data (unlike the variance, which is units-squared)

- If it is an average, why is it not divided by \( n \)?
Describing Variability

- Recall, the five systolic blood pressures (mm Hg) with sample mean \( \bar{x} = 99 \) mmHg
  
  \[ \begin{align*}
  &120, 80, 90, 110, 95 \\
  \sum_{i=1}^{5} (x_i - \bar{x})^2 = (120 - 99)^2 + (80 - 99)^2 + (90 - 99)^2 \\
  &+ (110 - 99)^2 + (95 - 99)^2 \\
  &= 21^2 + (-19)^2 + (-9)^2 + 11^2 + (-4)^2 \\
  &= 1020 \text{ mmHg}^2
  \end{align*} \]
Describing Variability

- Sample variance

\[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} = \frac{1020}{4} = 255 \text{ mmHg}^2 \]

- Sample standard deviation

\[ s = \sqrt{s} = \sqrt{255} = 15.97 \text{ mmHg} \]
Notes on $s$

- The bigger the $s$, the more variable the data.
- $s$ measures spread about the sample mean.
- $s$ can only equal zero if all observations are identical.
- The units of $s$ are the same as the units of the data (easily interpreted).
- $s$ is the best sample estimate of the population standard deviation, $\sigma$ (“sigma”).
Population versus Sample

- **Sample:** a subset of a larger group from which information is collected to learn about the larger group
  - Example: A sample of blood pressure readings of $n = 5$ 18-year-old male college students in the United States

- **Population:** the entire group for which information is wanted
  - Example: The blood pressure of all 18-year-old male college students in the United States
Random Sampling

- For studies, it is optimal (but not always possible) for the sample providing the data to be representative of the population under study.

- Simple random sampling provides a representative sample (theoretically)
  - A sampling scheme in which every possible sub-sample of size $n$ from a population is equally likely to be selected.
Population versus Sample

- The sample summary measures (mean, median, standard deviation) are called statistics
  - They are just estimates of their population counterparts

- Assuming the sample is representative of the population from which it came, these sample estimates should be “good” estimates of true quantities

Population

Mean: $\mu$ ("mu")
Standard deviation: $\sigma$ ("sigma")

Sample

Mean: $\bar{x}$
Standard deviation: $s$
Population versus Sample

- Because in most instances we are unable to measure every unit in the population, we will never know the population mean $\mu$.

- To gather information about $\mu$, we draw a sample from the population.

- We calculate the sample mean, $\bar{x}$.

- How close is the sample mean to the truth, $\mu$?

- Statistical theory allows us to estimate how close $\bar{x}$ is to $\mu$ using other information from our sample.
The Role of Sample Size on Sample Estimates

- Increasing the sample size increases the “goodness” of sample statistics as estimates for the population counterparts
  - The sample mean based on a random sample of 1,000 observations is a “better” estimate of the true population mean than one based on a random sample of 100
  - The same logic applies to estimates of the population variance and standard deviation
The Role of Sample Size on Sample Estimates

- Increasing sample size does not dictate how sample estimates from two different representative samples of different sizes will compare in value.

- A researcher cannot systematically decrease (or increase) the value of the sample mean by taking larger samples.
The Role of Sample Size on Sample Estimates

- Extreme values, both large and small, are actually more likely in larger samples
  - The smaller and larger extremes in larger samples balance each other out
  - This balancing act tends to keep the mean in a “steady state” as sample size increases
S: Why Do We Divide by n-1 instead of n?

- N - 1 is called the “degrees of freedom” of the variance or standard deviation
- The sum of the deviations is zero
- The last deviation can be found once we know the other $n - 1$
- Only $n - 1$ of the deviations has the “freedom” to vary
- The term degrees of freedom accompanies many other areas of statistics and is not always $n - 1$
Why Use $s$ As a Measure of Variation

- Why not use the range, $R$?
  - $R = \max(x_i) - \min(x_i)$

- What happens to the sample maximum and minimum as sample size increases?
  - The maximum tends to increase and the minimum tends to decrease—extreme values are more likely with larger samples
  - This will tend to increase the range systematically with increased sample size
Visualizing Continuous Data: Histograms
Pictures of Data: Continuous Variables

- Histograms
  - Means, medians, and standard deviations do not tell the whole story
  - Differences exist in the shape of distributions
  - Histograms are a way of displaying the distribution of a set of data by charting the number (or percentage) of observations whose values fall within pre-defined numerical ranges
How to Make a Histogram

- Consider the following data collected from the 1995 Statistical Abstracts of the United States
  - For each of the 50 United States, the proportion of individuals over 65 years of age has been recorded
### How to Make a Histogram

<table>
<thead>
<tr>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>12.9</td>
<td>IN</td>
<td>12.8</td>
<td>NE</td>
<td>14.1</td>
<td>SC</td>
<td>11.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AK</td>
<td>4.6</td>
<td>IA</td>
<td>15.4</td>
<td>NV</td>
<td>11.3</td>
<td>SD</td>
<td>14.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AZ</td>
<td>13.4</td>
<td>KS</td>
<td>13.9</td>
<td>NH</td>
<td>11.9</td>
<td>TN</td>
<td>12.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>14.8</td>
<td>KY</td>
<td>12.7</td>
<td>NJ</td>
<td>13.2</td>
<td>TX</td>
<td>10.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>10.6</td>
<td>LA</td>
<td>11.4</td>
<td>NM</td>
<td>11.0</td>
<td>UT</td>
<td>8.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td>10.1</td>
<td>ME</td>
<td>13.9</td>
<td>NY</td>
<td>13.2</td>
<td>VT</td>
<td>12.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>14.2</td>
<td>MD</td>
<td>11.2</td>
<td>NC</td>
<td>12.5</td>
<td>VA</td>
<td>11.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td>12.7</td>
<td>MA</td>
<td>14.1</td>
<td>ND</td>
<td>14.7</td>
<td>WA</td>
<td>11.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FL</td>
<td>18.4</td>
<td>MI</td>
<td>12.4</td>
<td>OH</td>
<td>13.4</td>
<td>WV</td>
<td>15.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td>10.1</td>
<td>MN</td>
<td>12.5</td>
<td>OK</td>
<td>13.6</td>
<td>WI</td>
<td>13.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HI</td>
<td>12.1</td>
<td>MI</td>
<td>12.5</td>
<td>OR</td>
<td>13.7</td>
<td>WY</td>
<td>11.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>11.6</td>
<td>MO</td>
<td>14.1</td>
<td>PA</td>
<td>15.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>12.6</td>
<td>MT</td>
<td>13.3</td>
<td>RI</td>
<td>15.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## How to Make a Histogram

<table>
<thead>
<tr>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
<th>State</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL</td>
<td>12.9</td>
<td>IN</td>
<td>12.8</td>
<td>NE</td>
<td>14.1</td>
<td>SC</td>
<td>11.9</td>
</tr>
<tr>
<td>AK</td>
<td>4.6</td>
<td>IA</td>
<td>15.4</td>
<td>NV</td>
<td>11.3</td>
<td>SD</td>
<td>14.7</td>
</tr>
<tr>
<td>AZ</td>
<td>13.4</td>
<td>KS</td>
<td>13.9</td>
<td>NH</td>
<td>11.9</td>
<td>TN</td>
<td>12.7</td>
</tr>
<tr>
<td>AR</td>
<td>14.8</td>
<td>KY</td>
<td>12.7</td>
<td>NJ</td>
<td>13.2</td>
<td>TX</td>
<td>10.2</td>
</tr>
<tr>
<td>CA</td>
<td>10.6</td>
<td>LA</td>
<td>11.4</td>
<td>NM</td>
<td>11.0</td>
<td>UT</td>
<td>8.8</td>
</tr>
<tr>
<td>CO</td>
<td>10.1</td>
<td>ME</td>
<td>13.9</td>
<td>NY</td>
<td>13.2</td>
<td>VT</td>
<td>12.1</td>
</tr>
<tr>
<td>CT</td>
<td>14.2</td>
<td>MD</td>
<td>11.2</td>
<td>NC</td>
<td>12.5</td>
<td>VA</td>
<td>11.1</td>
</tr>
<tr>
<td>DE</td>
<td>12.7</td>
<td>MA</td>
<td>14.1</td>
<td>ND</td>
<td>14.7</td>
<td>WA</td>
<td>11.6</td>
</tr>
<tr>
<td>FL</td>
<td>18.4</td>
<td>MI</td>
<td>12.4</td>
<td>OH</td>
<td>13.4</td>
<td>WV</td>
<td>15.4</td>
</tr>
<tr>
<td>GA</td>
<td>10.1</td>
<td>MN</td>
<td>12.5</td>
<td>OK</td>
<td>13.6</td>
<td>WI</td>
<td>13.4</td>
</tr>
<tr>
<td>HI</td>
<td>12.1</td>
<td>MI</td>
<td>12.5</td>
<td>OR</td>
<td>13.7</td>
<td>WY</td>
<td>11.1</td>
</tr>
<tr>
<td>ID</td>
<td>11.6</td>
<td>MO</td>
<td>14.1</td>
<td>PA</td>
<td>15.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL</td>
<td>12.6</td>
<td>MT</td>
<td>13.3</td>
<td>RI</td>
<td>15.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How to Make a Histogram

- Break the data range into mutually exclusive, equally sized “bins”
  - Here, each is 1% wide
- Count the number of observations in each bin

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
<th>Class</th>
<th>Count</th>
<th>Class</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>[4-5)</td>
<td>1</td>
<td>[9-10)</td>
<td>0</td>
<td>[14-15)</td>
<td>7</td>
</tr>
<tr>
<td>[5-6)</td>
<td>0</td>
<td>[10-11)</td>
<td>4</td>
<td>[15-16)</td>
<td>4</td>
</tr>
<tr>
<td>[6-7)</td>
<td>0</td>
<td>[11-12)</td>
<td>10</td>
<td>[16-17)</td>
<td>0</td>
</tr>
<tr>
<td>[7-8)</td>
<td>0</td>
<td>[12-13)</td>
<td>12</td>
<td>[17-18)</td>
<td>0</td>
</tr>
<tr>
<td>[8-9)</td>
<td>1</td>
<td>[13-14)</td>
<td>10</td>
<td>[18-19)</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \bar{x} = 12.7\%; s = 2.1\% \]
How to Make a Histogram

- Draw the histogram
- Label scales
Pictures of Data: Histograms

- Suppose we have blood pressure data on a sample of 113 men
- Sample mean, $\bar{X}$: 123.6 mmHg
- Sample median, $M$: 123.0 mmHg
- Sample standard deviation, $s$: 12.9 mmHg
Histogram of systolic blood pressure for a sample of 113 men. Each bar spans a width of 5 mmHg on the horizontal axis. The height of each bar represents the number of individuals with SBP in that range.
Another histogram of systolic blood pressure for 113 men. In this graph, each bar has a width of 20 mmHg, resulting in a total of four bars. This makes it hard to characterize the shape of the distribution.
Yet another histogram of the same BP information on 113 men. Here, the bin width is 1 mmHg, perhaps giving more detail than is necessary.
Another way to present the data in a histogram is to label the y-axis with relative frequencies as opposed to counts. The height of each bar represents the percentage of individuals in the sample with BP in that range. The bar heights should sum to 1.
Intervals

- How many intervals (bins) should a histogram have?
  - There is no perfect answer
  - Depends on the sample size
  - Rough rule of thumb: \# intervals \( \approx \sqrt{n} \)

<table>
<thead>
<tr>
<th>( n )</th>
<th>Number of Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>( \sim 3 )</td>
</tr>
<tr>
<td>50</td>
<td>( \sim 7 )</td>
</tr>
<tr>
<td>100</td>
<td>( \sim 10 )</td>
</tr>
</tbody>
</table>
Visualizing Continuous Data: Stem and Leaf Plots, Box Plots
Blood Pressure Example

- Suppose we took another look at our random sample of 113 men and their blood pressure measurements.

- Another common tool for visually displaying continuous data is the stem and leaf plot.

- Very similar to a histogram:
  - Looks like a histogram placed on its side.
  - Allows for easier identification of individual values in the sample.
Stem and Leaf: BP for 113 Males

8. |  9
9* |  
9. |  9
10* | 11334
10. | 566777899
11* | 111223333344444
11. | 556666667779
12* | 00000000111223344
12. | 5566677778888999999
13* | 000112222334
13. | 5677789
14* | 0000112222
14. |  67
15* |  0122
Stem and Leaf: BP for 113 Males

Stems

8.  | 9
9*  |
9.  | 9
10* | 11334
10. | 566777899
11* | 111223333344444
11. | 556666667779
12* | 00000000111223344
12. | 5566677778888999999
13* | 000112222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122
Stem and Leaf: BP for 113 Males

8. | 9
9* | |
9. | 9
10* | 11334
10. | 566777899
11* | 111223333344444
11. | 55666667779
12* | 00000000111223344
12. | 556667777888899999
13* | 000112222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122
Stem and Leaf: BP for 113 Males

8. | 9
9* |
9. | 9
10* | 11334
10. | 566777899
11* | 111223333344444
11. | 55666667779
12* | 00000000111223344
12. | 5566677778888999999
13* | 000112222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122
Stem and Leaf: BP for 113 Males

8. | 9
9* | 
9. | 9
10* | 11334
10. | 566777899
11* | 11122333334444
11. | 55566666779
12* | 00000000111223344
12. | 5556677778888999999
13* | 000112222334
13. | 5677789
14* | 0000112222
14. | 67
15* | 0122
Blood Pressure Example

- Another common visual display tool is the boxplot
  - Gives good insight into the distribution shape in terms of skewness and outlying values
  - Very nice tool for easily comparing distributions of continuous data in multiple groups—can be plotted side by side
Boxplot: BP for 113 Males
Boxplot: BP for 113 Males

Boxplot of Systolic Blood Pressure
Sample of 113 Men
Boxplot: BP for 113 Males

Boxplot of Systolic Blood Pressure
Sample of 113 Men

75th Percentile, $P_{75}$

25th Percentile, $P_{25}$
Boxplot: BP for 113 Males

Largest observation, max
Smallest observation, min

Boxplot of Systolic Blood Pressure
Sample of 113 Men
Hospital Length of Stay for 1,000 Patients

- Suppose we took a representative sample of discharge records from 1,000 patients discharged from a large teaching hospital in a single year

- How could we visualize this data?
Histogram: Length of Stay
Boxplot: Length of Stay

Hospital LOS (Days) for 1,000 Patients
Boxplot: Length of Stay

Hospital LOS (Days) for 1,000 Patients
Boxplot: Length of Stay

Hospital LOS (Days) for 1,000 Patients

- Largest
- Non-Outlier
- Smallest
- Non-Outlier
Boxplot: Length of Stay

Hospital LOS (Days) for 1,000 Patients

Large Outliers
### Stem and Leaf: Length of Stay

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11</td>
<td>(298)</td>
</tr>
<tr>
<td>0</td>
<td>22</td>
<td>(269)</td>
</tr>
<tr>
<td>0</td>
<td>44</td>
<td>(140)</td>
</tr>
<tr>
<td>0</td>
<td>66</td>
<td>(95)</td>
</tr>
<tr>
<td>0</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0000000000000000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0111111111111111</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
Side by Side Distribution Comparison

- Side-by-side histograms of LOS for female and male patients in sample
Side by Side Distribution Comparison

- Side-by-side boxplots of LOS for female and male patients in sample
Sample *versus* Population: Sample Distribution *versus* Underlying Population Distribution
Sample Distribution

- In research, samples are taken from a larger population
- If the sample is random, the sample characteristics will imperfectly mimic the population characteristics
- The characteristics include the mean, median, standard deviation, and shape of the distribution
Blood Pressure Example

- Histogram of BP values for random sample of 113 men
Blood Pressure Example

- Histogram of BP values for random sample of 500 men
Blood Pressure Example

- Histogram of BP values for male population
The Histogram and the Probability Density

- The *probability density* is a smooth idealized curve that shows the shape of the distribution in the population.

- This is generally a theoretical distribution that we can never see: we can only estimate it from the distribution presented by a representative (random) sample from the population.

- Areas in an interval under the curve represent the percentage of the population within that interval.

- The distributions shown are indicative of a symmetric, bell shaped distribution for blood pressure measurements in men.
Length of Stay Example

- Histogram of LOS values for 100 patients
Length of Stay Example

- Histogram of LOS values for 500 patients
Length of Stay Example

- Histogram of LOS values for all patients
Common Shapes of the Distribution

- Some shapes of data distributions

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symmetrical and bell shaped</td>
<td>Positively skewed or skewed to the right</td>
<td>Negatively skewed or skewed to the left</td>
</tr>
</tbody>
</table>
Shapes of the Distribution

- Some possible shapes for frequency distributions

A  B  C
Bimodal  Reverse J-shaped  Uniform
Distribution Characteristics

- **Mode**: peak(s)
- **Median**: point at which the area under the curve is halved
- **Mean**: center of gravity
Shapes of Distributions

- **Symmetric**
  - Mean = Median = Mode
Shapes of Distributions

- *Right skewed* (positively skewed)
  - Long right tail
  - Mean > Median
Shapes of Distributions

- *Left skewed* (negatively skewed)
  - Long left tail
  - Mean < Median
Part II: Describing Data
Topics

- The Normal Distribution
- Means, variability, and the normal distribution
- Calculating standard normal (z) scores
- Means, variability, and z-scores for non-normal distributions
The Normal Distribution
The Normal Distribution

- The *normal distribution* is a theoretical probability distribution that is perfectly symmetric about its mean (median and mode) and has a “bell”-like shape.
The Normal Distribution

- The normal distribution is also called the “Gaussian distribution” in honor of its inventor Carl Friedrich Gauss
The Normal Distribution

- There are an infinite number of normal distributions
- Each is uniquely defined by two quantities: a mean ($\mu$) and standard deviation ($\sigma$)
The Normal Distribution

- There are an infinite number of normal distributions.
- Each is uniquely defined by two quantities: a mean (μ) and standard deviation (σ).
The Normal Distribution

- There are an infinite number of normal distributions

- Each is uniquely defined by two quantities: a mean ($\mu$) and standard deviation ($\sigma$)

- This function defines the normal curve for any given ($\mu$, $\sigma$)

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
The Normal Distribution

- Areas under a normal curve represent the proportion of total values described by the curve that fall in that range.
The Normal Distribution

- This shaded area represents the proportion of values (observations) between 0.5 and 2 following a normal distribution with \( \mu = 0 \) and \( \sigma = 1 \)

- The shaded area is approximately 29% of the total area under the curve
The Normal Distribution

- The normal distribution is a theoretical distribution: no real data will truly be normally distributed (at the sample or population level)
  - Example: The tails of the normal curve are infinite
Normal Distribution

- However, some data approximates a normal curve pretty well.

- Here is a histogram of the BP of the 113 men with a normal curve superimposed (curve has the same mean and standard deviation as the sample).
  - Mean 123.6 mmHg, standard deviation 12.9 mmHg
Normal Distribution

- Data does not approximate a normal distribution

- Here is a histogram of the hospital length of stay of 500 patients with a normal curve superimposed (curve has same mean and standard deviation as the sample of 500 patients)
  - Mean 5.1 days, standard deviation 6.4 days
Variability in the Normal Distribution: Calculating Standard Normal Scores
The standard normal distribution has a mean of 0 and standard deviation of 1.
The Empirical Rule for the Normal Distribution

- 68% of the observations fall within one standard deviation of the mean
The Empirical Rule for the Normal Distribution

- 95% of the observations fall within two standard deviations of the mean (within 1.96 s)
The Empirical Rule for the Normal Distribution

- 99.7% of the observations fall within three standard deviations of the mean
## The Empirical Rule for the Normal Distribution

<table>
<thead>
<tr>
<th>Z</th>
<th>Within 1 SD of the mean</th>
<th>More than 1 SD above the mean</th>
<th>More than 1 SD below the mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>68.27%</td>
<td>15.87%</td>
<td>31.73%</td>
</tr>
<tr>
<td>2.0</td>
<td>95.45%</td>
<td>2.28%</td>
<td>4.55%</td>
</tr>
<tr>
<td>2.5</td>
<td>98.76%</td>
<td>0.62%</td>
<td>1.24%</td>
</tr>
<tr>
<td>3.0</td>
<td>99.73%</td>
<td>0.13%</td>
<td>0.27%</td>
</tr>
</tbody>
</table>
The Empirical Rule for the Normal Distribution

<table>
<thead>
<tr>
<th>Z</th>
<th>Within Z SDs of the mean</th>
<th>More than Z SDs above the mean</th>
<th>More than Z SDs above or below the mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>68.27%</td>
<td>15.87%</td>
<td>31.73%</td>
</tr>
<tr>
<td>2.0</td>
<td>95.45%</td>
<td>2.28%</td>
<td>4.55%</td>
</tr>
<tr>
<td>2.5</td>
<td>98.76%</td>
<td>0.62%</td>
<td>1.24%</td>
</tr>
<tr>
<td>3.0</td>
<td>99.73%</td>
<td>0.13%</td>
<td>0.27%</td>
</tr>
</tbody>
</table>
## The Empirical Rule for the Normal Distribution

<table>
<thead>
<tr>
<th>$Z$</th>
<th>Within $Z$ SDs of the mean</th>
<th>More than $Z$ SDs above the mean</th>
<th>More than $Z$ SDs above or below the mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>68.27%</td>
<td>15.87%</td>
<td>31.73%</td>
</tr>
<tr>
<td>2.0</td>
<td>95.45%</td>
<td>2.28%</td>
<td><strong>4.55%</strong></td>
</tr>
<tr>
<td>2.5</td>
<td>98.76%</td>
<td>0.62%</td>
<td>1.24%</td>
</tr>
<tr>
<td>3.0</td>
<td>99.73%</td>
<td>0.13%</td>
<td>0.27%</td>
</tr>
</tbody>
</table>
The Empirical Rule for the Normal Distribution

- What about other normal distributions with other mean/standard deviation combinations?

- The same properties apply

- In fact, any normal distribution with any mean and standard deviation can be transformed to a standard normal curve
Transforming to a Standard Normal: Standardizing

- The standard normal curve (blue) and another normal with $\mu = -2$ and $\sigma = 2$
Transforming to a Standard Normal: Standardizing

- To center at zero, subtract the mean of –2 from each observation under the red curve
Transforming to a Standard Normal: Standardizing

- To change the spread, divide each new observation by the standard deviation of 2
Transforming to a Standard Normal: Standardizing

- To change the spread, divide each new observation by the standard deviation of 2
Transforming to a Standard Normal: Standardizing

- This process is called standardizing or computing $z$-scores.

- A $z$-score can be computed for any observation from any normal curve.

- A $z$-score measures the distance of any observation from its distribution’s mean in units of standard deviation (e.g., $z = 2$ means the original observation is 2 standard deviations above the original mean).

- The $z$-score can help assess where the observations fall relative to the rest of the observations in the distribution.

$$z = \frac{obs - mean}{SD}$$
Blood Pressure Example

- Histogram of BP values for random sample of 113 men suggest BP measurements approximated by a normal distribution

\[ \bar{x} = 123.6 \text{ mmHg}; s = 12.9 \text{ mmHg} \]
Blood Pressure Example

- Using the sample data, estimate the range of blood pressure values for “most” (95%) of men in the population

- For normally distributed data, 95% will fall within 2 standard deviations of the mean

\[
\bar{x} \pm 2s
\]
\[
123.6 \pm 2(12.9)
\]
\[
(97.8, 149.4)
\]

- This is just an estimate using the best guesses from the sample for mean and standard deviation of the population
Blood Pressure Example

- Suppose a man comes into your clinic, gets his blood pressure measured, and wants to know how he compares to all men.

- His blood pressure is 130 mmHg.

- What percentage of men have blood pressures greater than 130 mmHg?

- Translate to z-score: \[ z = \frac{130 - 123.6}{12.9} \approx 0.5 \]

- Question is akin to “What percentage of observations under a standard normal curve are 0.5 standard deviations or more above the mean?”
Blood Pressure Example

- Could look this up in a normal table (in any stats book or online)
- Could also use the normal function in Excel
Blood Pressure Example

- Typing “=NORMDIST(0.5)” gives the proportion of observations less than 0.5 standard deviations from the mean:

- The NORMDIST function allows you to also specify a mean and standard deviation, making it unnecessary to compute $z$.
Blood Pressure Example

- For $z = 0.5$, roughly 69% of observations fall below $0.5s$ from the mean.
Blood Pressure Example

- For $z = 0.5$, roughly $100\% - 69\% = 31\%$ of observations fall above $0.5\sigma$ from the mean.
Blood Pressure Example

- Approximately 31% of all men have blood pressures greater than our subject with a blood pressure of 130

- What percentage of men have blood pressures more extreme; i.e., farther than 0.5σfrom the mean of all men in either direction?
Blood Pressure Example

- What we want (pictures, pictures, pictures!)
Blood Pressure Example

- By symmetry of the normal curve, 31% of observations are above 0.5s and 31% are below -0.5s

- A total of 62% of men are farther than 0.5s from the mean in either direction
Standard Normal Scores and Variability in Non-Normal Data
Why Do We Like the Normal Distribution So Much?

- The truth is there is nothing special about the standard normal distribution or z-scores
  - These can be computed for observations from any sample or population of continuous data values
  - The score measures how far an observation is from its mean in standard units of statistical distance
Why Do We Like the Normal Distribution So Much?

- However, unless the population or sample has a well-known, well-behaved distribution (i.e., approximately normal), we may not be able to use the mean and standard deviation to create interpretable intervals or to measure “unusuality” of individual observations.
Length of Stay Example

- Random sample of 500 patients
  - Mean length of stay: 4.8 days
  - Median length of stay: 3 days
  - Standard deviation: 6.3 days
Length of Stay Example

- Histogram of sample data
Constructing Intervals

- Suppose I wanted to estimate an interval containing roughly 95% of the values of hospital length of stay in the population

- The distribution is right skewed, so we cannot appeal to the properties of the normal distribution

- Mean ± 2SD
  - 4.8 ± 2(6.3)
  - This yields an interval from -7.8 to 17.4 days . . .
Length of Stay Example

- Histogram of sample data
Constructing Intervals

- We would need to estimate the interval from the histogram or by finding sample percentiles.
Constructing Intervals

- Excel can compute percentiles for you using the PERCENTILE function
  - Enter the data and the percentile number you seek:
  - “=PERCENTILE(data,2.5)” gives 1
  - “=PERCENTILE(data,97.5)” gives 23.475

- Based on this sample data we estimate that 95% of discharged patients had LOS between 1 and 24 days
Constructing Intervals

- What percentage of patients had LOS greater than five days?

- Wrong approach: \[ z = \frac{5 - 4.8}{6.5} \approx 0.03 \]

- Assuming normality, this would suggest that nearly 50% of the patients had LOS greater than five days.

- According to percentiles, five days is the 75th percentile:
  - Only 25% of the sample have LOS over five days.
Next Lecture

- Friday, December 3 in 1023 Orr-Major from 8:30a - 10:30a

- Topics include
  - Sampling variability
  - Confidence intervals
  - $P$-values
  - One- and Two-sample $t$-tests
  - ANOVA
References and Citations

Lectures modified from notes provided by John McGready and Johns Hopkins Bloomberg School of Public Health accessible from the World Wide Web: http://ocw.jhsphs.edu/courses/introbiostats/schedule.cfm