Statistical Methods
Chapter 1: Overview and Descriptive Statistics

Walid Sharabati
Purdue University
January 15, 2014
General Introduction
- Statistics studies data, population, and samples.
- Descriptive Statistics vs Inferential Statistics.

Descriptive Statistics
- Pictorial and tabular methods
  - Stemplot, dotplot, histogram, boxplot.
- Numerical measures
  - Measures of Location: Mean and Median.
  - Measures of Variability: Range, Variance, and IQR.

Inferential Statistics
- Draw conclusions about a certain population parameter.
  - Confidence Intervals.
  - Hypothesis Testing.
What does statistics study?

Statistics is a mathematical science pertaining collection, presentation, analysis and interpretation of data.

- **Population**: a well-defined collection of objects.
- **Sample**: a subset of the population.
- **Variable**: characteristics of the objects.
- **Observation**: an observed value of a variable.
- **Data**: a collection of observations.

statistics → study data → understand the population
About Variable

What is variable?
Characteristics of a population of interest whose values vary.

A variable can be
- Categorical
  - *e.g.* $x = \text{gender of a person (male, female)}$
- Numerical
  - Discrete variable: *e.g.* $x = \# \text{ of students in a class}$
  - Continuous variable: *e.g.* $x = \text{height of a student}$
Types of Data

Data come from making observations either on a single variable or simultaneously on two or more variables.

- **Univariate data**: observations on a single variable
- **Bivariate data**: observations on two variables
  - e.g. \((x, y) = (\text{height, weight}) \text{ of a student}\)
- **Multivariate data**: observations on more than two variables
  - e.g. \((x, y, z) = (\text{height, weight, gender}) \text{ of a student}\)
How to study data?

What is Statistics?

- Data collection
  - Sampling methods, experimental design.
- Data analysis, presentation & interpretation
  - **Descriptive statistics**
    - summarize and describe features of data
      - Visual methods: dotplot, pie chart, histogram.
      - Numerical methods: measures of location (mean, median) and variation (range, variance)
  - **Inferential statistics**
    - make inference about the population from samples
      - Point estimate, confidence intervals, hypothesis testing.
Descriptive Statistics: Visual Methods

- Stem-and-leaf display
- Dotplot
- Histogram
- Boxplot
Stem-and-leaf Display

Example 1

The number of touchdown passes thrown by each of the 31 teams in the National Football League in 2000 is given below:
{14, 29, 22, 18, 20, 15, 6, 9, 32, 18, 19, 18, 23, 28, 37, 21, 14, 19, 21, 20, 16, 22, 33, 28, 12, 18, 22, 14, 33, 21, 12}

What does the data tell?
The tens digits called **stems** are arranged as a column to the left. The ones digits are listed to the right of each stem and are called **leaves**.

```
<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>69</td>
</tr>
<tr>
<td>1</td>
<td>4858984962842</td>
</tr>
<tr>
<td>2</td>
<td>920381102821</td>
</tr>
<tr>
<td>3</td>
<td>2733</td>
</tr>
</tbody>
</table>
```

What can we say about the data set now?
Most teams had 10 – 29 touchdown passes.
Refined Stem-and-leaf Display

When too many leaves are lumped into a few stems, splitting the stem helps reveal more information about the distribution of data. We can further "refine" the above stem-and-leaf display by splitting each stem into two parts: low and high.

Stem-and leaf displays of touchdown passes

0H|69
1L|44242
1H|85898968  Stem: Tens digits
2L|203110221  Leaf: Ones digits
2H|988
3L|233
3H|7

What can we say about the data set now?
Most teams had 15 – 24 touchdown passes.
### Example 2

Suppose we also have data from the 1998 season. We can compare the numbers of touchdown passes in the 1998 and 2000.

<table>
<thead>
<tr>
<th>Year 1998</th>
<th>Year 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0H 69</td>
</tr>
<tr>
<td>321</td>
<td>1L 22444 Stem: Tens digits</td>
</tr>
<tr>
<td>987776665</td>
<td>1H 5688889 Leaf: Ones digits</td>
</tr>
<tr>
<td>44331110</td>
<td>2L 00111223</td>
</tr>
<tr>
<td>8865</td>
<td>2H 889</td>
</tr>
<tr>
<td>332</td>
<td>3L 233</td>
</tr>
<tr>
<td>3H 7</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>4L</td>
</tr>
</tbody>
</table>

- The peaks of the two seasons are slightly different.
- For both seasons, most teams had 15 – 24 touchdown passes.
- The shapes of the data distributions are similar.
Summary: Stem-and-leaf Display

- How to make a stem-and-leaf display?
  1. Select one or more leading digits for the stem values *(any value appropriate)*. The trailing digits become the leaves.
  2. List possible stem values in a vertical column.
  3. Put the leaf for each observation besides the corresponding stem.
  4. Indicate the units for stems and leaves.

- What can a stem-and-leaf display tell?
  - Typical value
  - Symmetry of distribution
  - Peaks
  - Outliers

Stem-and-leaf display is suitable for a data set with *a moderate size*. 
Example 3

O-ring temperatures ($F^\circ$) for test firings or actual launches of the shuttle rocket engine. \{84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31\}
How to make a dotplot?
1. Represent each obs by a dot above the corresponding location on a measurement scale.
2. Stack dots vertically when a value occurs more than once.

What can a dotplot tell?
- Location of typically values
- Spread of data set
- Extreme values
- Gaps between values

Dotplot is a nice display of data when a data set is reasonably small or has only a few distinct values.
What if a data set is large?

Use Histogram

For different types of data, we construct histograms differently.

- Histogram for discrete data
- Histogram for continuous data
- Histogram for categorical (qualitative) data, also known as Bar-graph
Histogram for Discrete Data

- Frequency (Count)
  In a discrete data set, frequency of a value $c$ is the number of occurrences of $c$ in the data set.

- Relative frequency
  The relative frequency of a value $c$ is

  \[
  \text{relative frequency of a value } c = \frac{\text{frequency of } c}{n}
  \]

  where $n$ is the total number of observations in the data set.

*If we list frequencies of a data set in a table, it is called frequency distribution/table.*
How to create a histogram for a discrete data set?

1. Determine the distinct values $c_1, c_2, c_3, \ldots, c_r$ in the data set.

2. Calculate the relative frequency for each $c_j$, $j = 1, 2, \ldots, r$:

   \[
   \text{relative frequency of } c_j = \frac{\text{number of occurrences of } c_j}{n}
   \]

3. Mark the $c_j$’s on a horizontal scale, draw a rectangle whose height is the relative frequency of $c_j$, where $(j = 1, 2, \ldots, r)$.

The area of the rectangle is proportional to the relative frequency.
Example 4

100 married couples between 30 and 40 years of age are studied to see how many children each couple have. Table below is the frequency table of this data set.

<table>
<thead>
<tr>
<th>Kids</th>
<th># of couples</th>
<th>Relative Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11</td>
<td>0.11</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>0.22</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>0.24</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>0.30</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Professor Sharabati (Purdue University)**

**Overview and Descriptive Statistics**

**Statistical Methods**

17 / 40
Histogram of Example 4

Histogram of kids data

Percentage (%) vs. Kids

Professor Sharabati (Purdue University)
How to create a histogram for a continuous data set?

1. Divide the measurement axis into a suitable number of class intervals/classes
   - To ensure that each observation falls into exactly one interval
   - Intervals could be either equal class width or unequal class width

2. Calculate relative frequency for each interval

3. Draw a rectangle above each interval with a height \textit{densities}:
   \[
   \text{rectangle height} = \frac{\text{relative frequency of the class interval } I_j}{\text{class interval width}}
   \]

The area of the rectangle is proportional to the relative frequency. For unequal class width histograms, the total area of all rectangles is 1.
Example 5

Adjusted energy consumption during a particular period for a sample of 90 gas-heated homes are recorded.

We divide the class intervals as follows:

<table>
<thead>
<tr>
<th>Class</th>
<th>[1, 3)</th>
<th>[3, 5)</th>
<th>[5, 7)</th>
<th>[7, 9)</th>
<th>[9, 11)</th>
<th>[11, 13)</th>
<th>[13, 15)</th>
<th>[15, 17)</th>
<th>[17, 19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq.</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>21</td>
<td>25</td>
<td>17</td>
<td>9</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Relative freq.</td>
<td>0.011</td>
<td>0.011</td>
<td>0.122</td>
<td>0.233</td>
<td>0.278</td>
<td>0.189</td>
<td>0.100</td>
<td>0.044</td>
<td>0.011</td>
</tr>
</tbody>
</table>
Histogram Shapes

Histograms have a variety of shapes, the shape of a histogram conveys important information about the distribution of data.

- **Unimodal**: Single peak
- **Bimodal**: Two peaks
- **Multimodal**: Two more peaks
- **Symmetric**: Left $\approx$ right
- **Positively skewed**: Right tail stretching out
- **Negatively skewed**: Left tail stretching out
Histogram Shapes
Visual displays give us the shape of data distribution, typical values. Numerical measures give us quantitative measures instead.

- Measures of location
  - Mean
  - Median
  - Trimmed mean
  - Quartiles

- Measures of variability
  - Variance
  - Standard deviation

- Another visual display of data: Boxplot.
**Sample mean** of a sample of size \( n \) \( \{x_1, x_2, \ldots, x_n\} \) is the arithmetic mean of all obs in the data set and is denoted by \( \bar{x} \):

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

- **Interpretation of** \( \bar{x} \): measures location/center of a sample.
- \( \bar{x} \) takes every individual obs into account and weigh them equally.

**Population mean** is "average/center" point of a population, and is usually denoted by \( \mu \).

- Use sample mean \( \bar{x} \) to estimate and make inferences about the usually unknown population mean \( \mu \).
Sample Mean

Example 6

The following sample contains weights (lbs) of basses in a specific lake:
\( \{ x_1 = 1.22, x_2 = 1.51, x_3 = 1.34, x_4 = 1.60, x_5 = 0.98, x_6 = 1.71, x_7 = 1.82, x_8 = 1.04, x_9 = 1.10, x_{10} = 0.85, x_{11} = 1.08 \} \)

The mean weight of this sample is:
\[
\bar{x} = \frac{1.22 + 1.51 + \cdots + 1.08}{11} = 1.30
\]

Suppose we catch another bass in the lake and it weighs 12.52 lbs.
\( \{ x_1 = 1.22, x_2 = 1.51, x_3 = 1.34, x_4 = 1.60, x_5 = 0.98, x_6 = 1.71, x_7 = 1.82, x_8 = 1.04, x_9 = 1.10, x_{10} = 0.85, x_{11} = 1.08, x_{12} = 12.52 \} \)

The mean weight of this sample becomes:
\[
\bar{x} = \frac{1.22 + 1.51 + \cdots + 12.52}{12} = 2.23
\]

Drawback: Sample mean is very sensitive to outliers. Alternative measure: Median
**Measure of Location: Median**

- **Sample median** of a sample of size \( n \) \( \{x_1, x_2, \ldots, x_n\} \) is the middle value of the sample, denoted by \( \tilde{x} \). It is obtained by:
  1. Order the \( n \) obs from smallest to largest \( \{x(1), x(2), \ldots, x(n)\} \).
  2. The median is then:

\[
\tilde{x} = \begin{cases} 
  x(\frac{n+1}{2}) & \text{when } n \text{ is odd} \\
  \frac{x(\frac{n}{2}) + x(\frac{n}{2}+1)}{2} & \text{when } n \text{ is even}
\end{cases}
\]

- **Interpretation of** \( \tilde{x} \): the value in the middle of the sample
- **Note:** only one or two obs in the middle are needed to calculate the median.

- **Population median** is the middle point in a population, and is usually denoted by \( \mu \).
- Use sample median \( \tilde{x} \) to estimate and make inferences about the usually unknown population median \( \mu \).
Sample Median - Example 6

Before we caught the huge bass, we had \( n = 11 \) obs in the sample:

1. Order the data set from smallest to largest:
   \[ x_1 = 0.85, x_2 = 0.98, x_3 = 1.04, \ldots, x_6 = 1.22, \ldots, x_{11} = 1.82 \]
2. \( n \) is odd, so \( \tilde{x} = x_{\left(\frac{11+1}{2}\right)} = x_6 = 1.22 \)

Comparing \( \bar{x} = 1.30 \) and \( \tilde{x} = 1.22 \), the difference is not big.

Now after we caught the 12.52-lb fish, our sample size becomes \( n = 12 \), and median:

1. Order the data set from smallest to largest:
   \[ x_1 = 0.85, x_2 = 0.98, x_3 = 1.04, \ldots, x_6 = 1.22, x_7 = 1.34, \ldots, x_{11} = 1.82, x_{12} = 12.52 \]
2. \( n \) is even, so \( \tilde{x} = \frac{x_6 + x_7}{2} = \frac{1.22 + 1.34}{2} = 1.28 \)

Median is clearly not severely affected.
Measures of Location: Trimmed Mean

- $\bar{x}$ is sensitive to outliers
- $\tilde{x}$ is very insensitive to outliers

A **trimmed mean** is a compromise between these two.

- Give the number $\alpha$, where $0 < \alpha < 1$, the $100\alpha\%$ trimmed mean is computed by eliminating the smallest and largest $100\alpha\%$ of the sample and averaging what remains.

- See details in your textbook (page 32).
Measures of Variability

**Data set 1**

\{-0.20, -0.10, -0.01, 0, 0.01, 0.10, 0.20\},
Sample mean: \( \bar{x}_1 = ? \)

**Data set 2**

\{-10000, -2000, -100, 0, 100, 2000, 10000\},
Sample mean: \( \bar{x}_2 = ? \)
Measures of Variability: Variance

To compute sample variance for a sample \( \{x_1, x_2, \ldots, x_n\} \)

1. Calculate the sample mean \( \bar{x} \)
2. Calculate the deviations of each obs from \( \bar{x} \): \( x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x} \)
3. Sample variance: 
   \[
   s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}
   \]
4. Sample standard deviation: 
   \[
   s = \sqrt{s^2}
   \]

- **interpretation:** average magnitude of the deviation from the sample mean
- **population variance** \( \sigma^2 \) and **population std dev** \( \sigma \) as a measure of variability of the population.
- \( s^2/s \) could be used to estimate or make inferences about \( \sigma^2/\sigma \).
Why do we use $n - 1$ as the divisor to calculate sample variance?

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

- **Population variance**
  \[ \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \]

- Compensate for underestimating of $\sigma^2$
- $n - 1$ degree of freedom (df)
Properties of $s^2$

A working formula for $s^2$

$$s^2 = \frac{S_{xx}}{n-1}, \quad S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Properties of $s^2$

Let $\{x_1, x_2, \ldots, x_n\}$ be the sample and $c$ be any nonzero constant.

- If $y_1 = x_1 + c$, $y_2 = x_2 + c$, $\ldots$, $y_n = x_n + c$, then $s_y^2 = s_x^2$.
- If $y_1 = cx_1$, $y_2 = cx_2$, $\ldots$, $y_n = cx_n$, then $s_y^2 = c^2 s_x^2$ and $s_y = |c| s_x$. 
Let us look at the data sets that have the same mean.

Data set 1:
\( \{-0.20, -0.10, -0.01, 0, 0.01, 0.10, 0.20\} \), \( \bar{x}_1 = 0 \)

\[
 s_1^2 = \frac{0.20^2 + 0.10^2 + 0.01^2 + 0 + 0.01^2 + 0.10^2 + 0.20^2}{7 - 1} = 0.017
\]

Data set 2:
\( \{-10000, -2000, -100, 0, 100, 2000, 10000\} \), \( \bar{x}_2 = 0 \)

\[
 s_2^2 = \frac{10000^2 + 2000^2 + 100^2 + 0 + 100^2 + 2000^2 + 10000^2}{7 - 1} = 3.0 \times 10^7
\]
Measures of Location: Quartiles

Median splits the sample into lower/upper sub-sample

Quartiles further split the lower and upper sub-sample

- $Q_1 =: \text{median of the lower sub-sample (lower fourth)}$
- $Q_2 =: \text{median of the entire sample}$
- $Q_3 =: \text{median of the upper sub-sample (upper fourth)}$
- $IQR = Q_3 - Q_1$ (Inter Quatile Range or fourth spread)
Still use our bass example, rank the 11 obs:
\{x(1) = 0.85, x(2) = 0.98, x(3) = 1.04, x(4) = 1.08, x(5) = 1.10, x(6) = 1.22, x(7) = 1.34, x(8) = 1.51, x(9) = 1.60, x(10) = 1.71, x(11) = 1.82\}

\[Q_1 = \frac{x(3) + x(4)}{2} = 1.060\]
\[Q_2 = \bar{x} = 1.22\]
\[Q_3 = \frac{x(8) + x(9)}{2} = 1.555\]

\[IQR = 1.555 - 1.060 = 0.495\]
Boxplot

**Five-number summary**

smallest, lower fourth ($Q_1$), median ($Q_2$), upper fourth ($Q_3$), largest

Boxplot shows center, spread, symmetry and outliers.

1. Draw a horizontal axis, find $Q_1$, $Q_2$ and $Q_3$
2. Place a rectangle (left edge at $Q_1$, right edge at $Q_3$) above the axis
3. Put a line segment inside the rectangle at the location of $Q_2$.
4. Draw whiskers out from each end of the rectangle to the smallest and largest observations.
What’s outlier?

- **mild outlier**: any obs farther than $1.5IQR$ from the nearest quartile
- **extreme outlier** more than $3IQR$ from the nearest quartile

To draw boxplot that show outliers

1. Drawing a whisker out from the rectangle to the smallest and largest obs that are not outliers.
2. Plot mild outliers by solid dots, plot extreme outliers with circles. (optional)
Example 7

Pulse width data with sample size $n = 25$:

\{5.30, 8.20, 13.80, 74.10, 85.30, 88.00, 90.20, 91.50, 92.40, 92.90, 93.60, 94.30, 94.80, 94.90, 95.50, 95.80, 95.90, 96.60, 96.70, 98.10, 99.00, 101.40, 103.70, 106.00, 113.50\}
Solution

\[ Q_1 = 90.2 \]
\[ Q_2 = \bar{x} = 94.8 \]
\[ Q_3 = 96.7 \]
\[ IQR = 6.5 \]
\[ 1.5IQR = 9.75 \]
\[ 3IQR = 19.5 \]

- Extreme outliers are: 5.30, 8.20, 13.80
- Mild outliers are: 74.10, 113.5
Distribution Shapes, Boxplots and Measures of Location

(a) Uniform

(b) Bell-shaped

(c) Right skewed

(d) Left skewed