Statistical Methods

Chapter 1: Overview and Descriptive Statistics

- 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

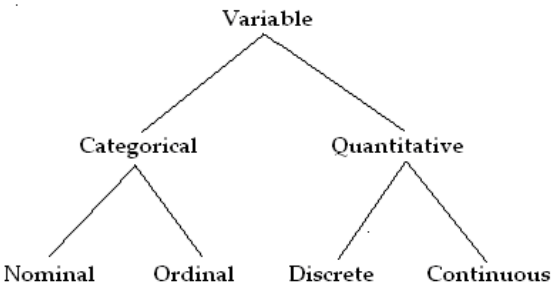
About Variable

What is variable?

Characteristics of a population of interest whose values vary.

A variable can be

- Categorical
  - *e.g.* $x =$ *gender of a person (male, female)*
- Numerical
  - Discrete variable: *e.g.* $x =$ *# of students in a class*
  - Continuous variable: *e.g.* $x =$ *height of a student*



Types of Data

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

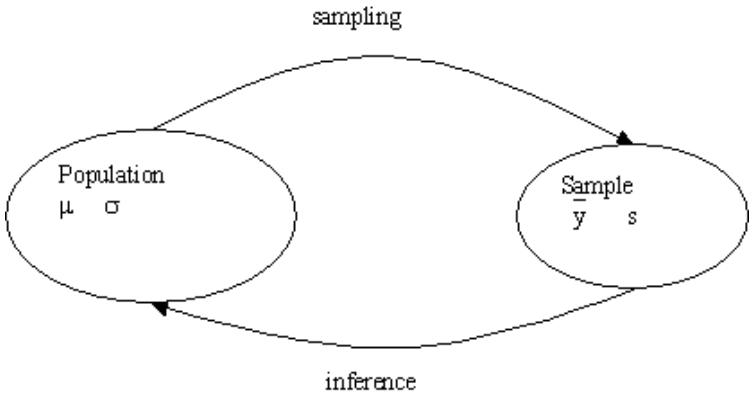
- Univariate data: observations on a single variable
- Bivariate data: observations on two variables *e.g.* (x, y) =*(height, weight) of a student*
- Multivariate data: observations on more than two variables *e.g.* (x, y, z) = *(height, weight, gender) of a student*

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.

Inferential Statistics and Probability Theory



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**.

```
0|69                    0|69
1|4858984962842         1|2244456888899
2|920381102821          2|001112223889
3|2733                  3|2337
```

- 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.

```
0H|69
1L|44242
1H|85898968
2L|203110221
```

```
2H|988
3L|233
3H|7
```

- What can we say about the data set now? Most teams had $15 - 24$ touchdown passes.

Compare Data by Stem-and-leaf Display

Example 2
Suppose we also have data from the 1998 season. We can compare the numbers of touchdown passes in the 1998 and 2000.

```
Year 1998|  |Year 2000
        7|0H|69
      213|1L|44242
987776665|1H|85898968
 44331110|2L|203110221
     8865|2H|988
      332|3L|233
         |3H|7
       11|4L|
```

- 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.

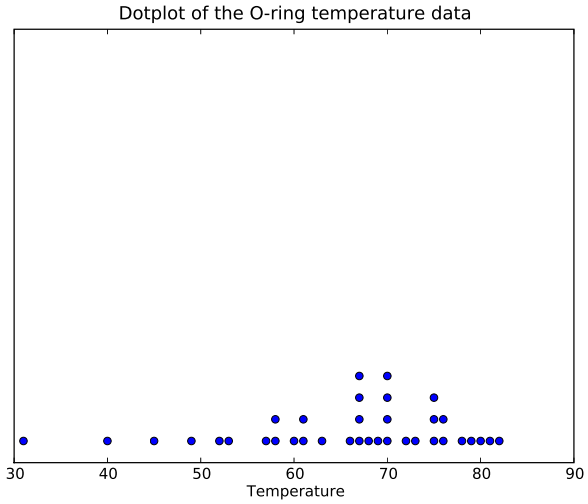
Summary: Stem-and-leaf Display

- 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.

Dotplot

Example 3
O-ring temperatures (F°) for test firings or actual launches of the shuttle rocket engine. $\{84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31\}$



Summary: Dotplot

- 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.

Histogram

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****.*

Constructing Histogram for Discrete Data

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.

Histogram for Discrete Data

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.

Kids	# of couples	Relative Freq
0	11	0.11
1	22	0.22
2	24	0.24
3	30	0.30
4	11	0.11
5	1	0.01
6	0	0.00
7	1	0.01
	100	1.00

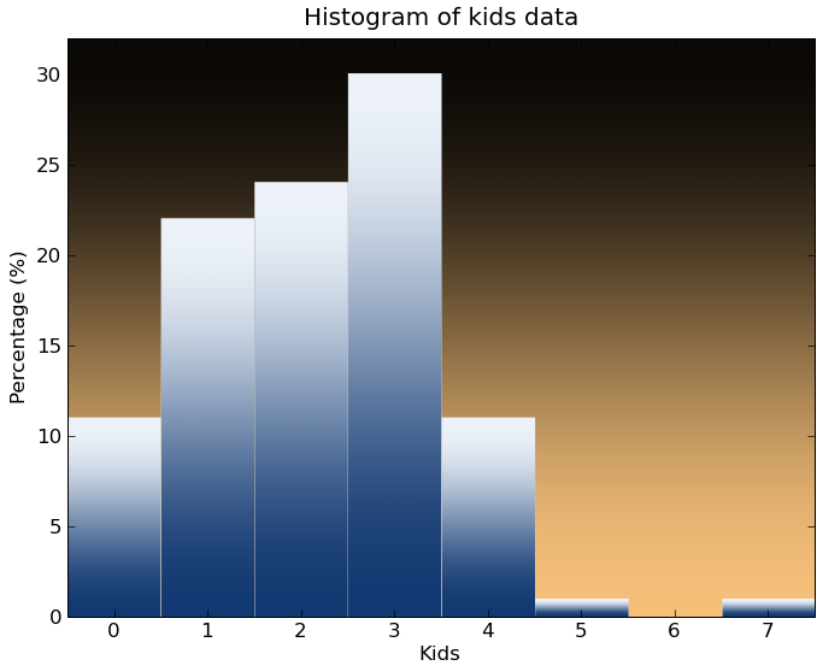
Histogram of Example 4

Histogram for Continuous Data

How to create a histogram for a continuous data set?

1. Divide the measurement axis into a number of **class intervals/classes** such that each obs falls into exactly one interval. Denote these intervals by: $I_1, I_2, \ldots, I_r$.
   - To ensure that each obs falls into exactly one interval, we may use intervals in the form: $I_1 = [a_1, a_2)$, $I_2 = [a_2, a_3), \ldots$
   - We may use I_j's of the same interval length, this is called **equal class width**; we may also use I_j's of different interval lengths, this is called **unequal class width** $j = 1, 2, 3, \ldots, r$.
2. Calculate relative frequency for each interval I_j, $j = 1, 2, 3, \ldots, r$.
3. Draw a rectangle above each I_j.
   - For equal class width case, rectangle height = relative frequency.
   - For unequal class width case: rectangle height $= \frac{\text{relative frequency of the class interval } I_j}{\text{class interval width}}$, the resulting rectangle heights here are called **densities**.

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



Histogram for Continuous Data

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:

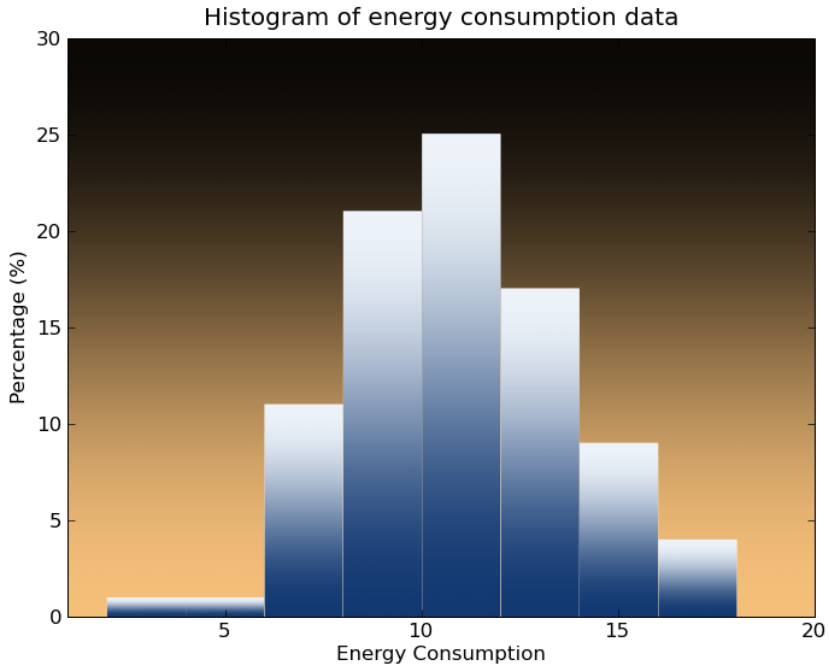
Class	[1, 3)	[3, 5)	[5, 7)	[7, 9)	[9, 11)	[11, 13)	[13, 15)	[15, 17)	[17, 19)
Freq.	1	1	11	21	25	17	9	4	1
Relative freq.	0.011	0.011	0.122	0.233	0.278	0.189	0.100	0.044	0.011

Histogram of Example 5

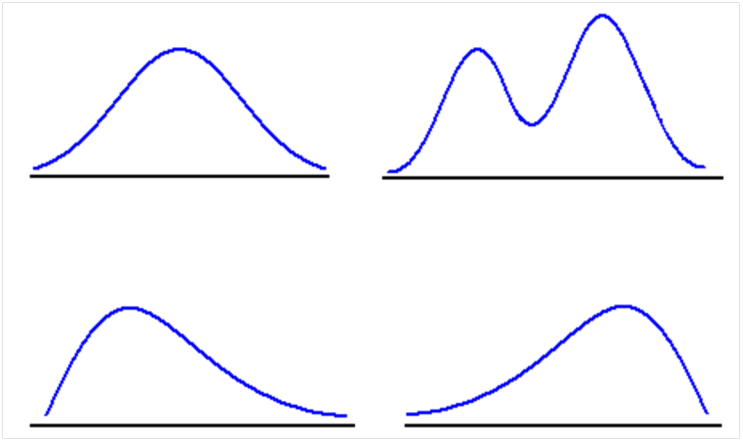
Histogram Shapes

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



Descriptive Statistics: Numerical Measures

Visual displays give us general ideas about 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.

Measure of Location: Mean

- **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 μ.
- Use sample mean $\bar{x}$ to estimate and make inferences about the usually unknown population mean μ.

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 that to calculate sample median, only one or two obs in the middle are needed.
- **Population median** is the middle point in a population, and is usually denoted by $\tilde{\mu}$.
- Use sample median $\tilde{x}$ to estimate and make inferences about the usually unknown population median $\tilde{\mu}$.

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 $\tilde{x} = x_{(\frac{11+1}{2})} = x_{(6)} = 1.22$

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, while $\tilde{x}$ is very insensitive to outliers, two extremes.
- A **trimmed mean** is a compromise between these two.
- Give the number α, where $0 < \alpha < 1$, the $100\alpha\%$ trimmed mean is computed by eliminating the smallest and largest $100\alpha\%$ in the sample and then calculate the average over the obs left in the sample.
- See details in your textbook (page 28).

Measures of Location: Quartiles

- Median separates the sample into two parts: lower sub-sample and upper sub-sample. n odd: $\{x_{(1)}, \ldots, x_{(\frac{n+1}{2})}\}$ and $\{\{x_{(\frac{n+1}{2})}\}, \ldots, x_{(n)}\}$ n even: $\{x_{(1)}, \ldots, x_{(\frac{n}{2})}\}$ and $\{\{x_{(\frac{n}{2}+1)}\}, \ldots, x_{(n)}\}$
- **Quartiles** divide the lower and upper sub-samples into two parts:
  - **1st Quartile:** $Q_1 =$ median of the lower sub-sample, also called the lower fourth
  - **2nd Quartile:** $Q_2 =$ median of the entire sample
  - **3rd Quartile:** $Q_3 =$ median of the upper sub-sample, also called the upper fourth
  - **Inter Quartile Range:** $IQR = Q_3 - Q_1$, also called fourth spread

Quartile Example

Still use our bass example, rank the 11 obs: $\{x_{(1)} = 0.85, x_{(2)} = 0.98, x_{(3)} = 1.04, x_{(4)} = 1.08, x_{(5)} = 1.10, x_{(6)} = 1.22, x_{(7)} = 1.34, x_{(8)} = 1.51, x_{(9)} = 1.60, x_{(10)} = 1.71, x_{(11)} = 1.82\}$

$$Q_1 = \frac{x_{(3)} + x_{(4)}}{2} = 1.060$$
$$Q_2 = \tilde{x} = 1.22$$
$$Q_3 = \frac{x_{(8)} + x_{(9)}}{2} = 1.555$$
$$IQR = 1.555 - 1.060 = 0.495$$

Measures of Variability

Data set 1
$\{-0.20, -0.10, -0.01, 0, 0.01, 0.10, 0.20\}$, Sample mean: $\bar{x}_1 = 0$

Data set 2
$\{-10000, -2000, -100, 0, 100, 2000, 10000\}$, Sample mean: $\bar{x}_2 = 0$

Two data sets have the same means, but obviously second one is more spread out. So we need numeric measures of such variability too. **Variance** is one of such measures.

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. s^2 is the average sum of squares of the deviations: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

- **interpretation:** average magnitude of the deviation from the sample mean
- Sometimes we also use **sample standard deviation**: $s = \sqrt{s^2}$
- Similarly, we also have **population variance** σ^2 and **population std dev** σ as a measure of variability of the population.
- s^2/s could be used to estimate or make inferences about σ^2/σ.

The Divisor $n - 1$

Why do we use $n - 1$ **as the divisor to calculate** s^2**?**

- We hope s^2 can be a good estimate of σ^2, ideally, we want to calculate s^2 as:

$$s^2 = \frac{\sum (x_i - \mu)^2}{n}$$

- μ is something unknown from the population, a replacement of μ is $\bar{x}$, but obs in a sample tend to be closer to the sample mean $\bar{x}$, resulting a relatively smaller sum of squares, so we use $n - 1$ instead of n as the divisor to compensate for this.
- $n-1$ is called **degree of freedom**. This is because s^2 is based on n deviations $x_1 - \bar{x}, \ldots, x_n - \bar{x}$, but since $\sum(x_i - \bar{x}) = 0$, any $n - 1$ deviations will be enough.

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 $s_y = |c| s_x$.

Sample Variance

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: $\{-10000, -2000, -100, 0, 100, 2000, 10000\}$, $\bar{x}_2 = 0$

$$s_2^2 = \frac{10000^2 + 2000^2 + 100^2 + 0 + 100^2 + 2000^2 + 10000^2}{7-1} = 3.0 \times 10^7$$

Boxplot

Boxplot is very useful in describing several of a data set's important features such as: center, spread, symmetry and outliers.

1. Draw a horizontal axis, find Q_1, Q_2 and Q_3 and calculate IQR.
2. Place a rectangle above the axis, with the left edge at Q_1, right edge at Q_3.
3. Place a vertical 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 obs.

Boxplot With Outliers

We can also draw boxplots that show outliers.

- Any obs farther than $1.5IQR$ from the nearest quartile is a **mild outlier**
- Any obs farther than $3IQR$ from the nearest quartile is a **extreme outlier**

To draw boxplot that show outliers, we modify the boxplot by:

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)

Boxplot

Example 7

This is example 1.18 in your textbook (page 37) Pulse width data, $n = 25$:

```
{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}
```

We have:

$$Q_1 = 90.2, \quad Q_2 = \tilde{x} = 94.8, \quad Q_3 = 96.7, \quad IQR = 6.5$$

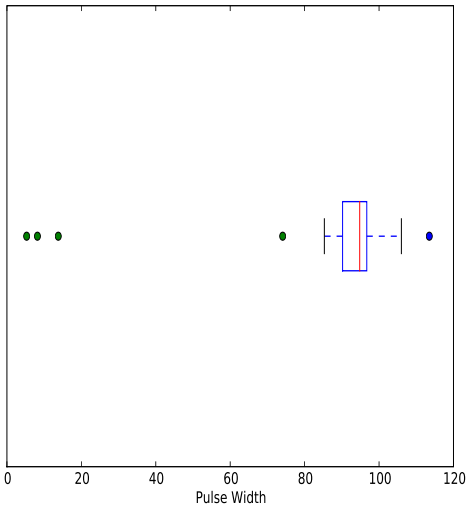
$$1.5IQR = 9.75, \quad 3IQR = 19.5$$

So, the extreme outliers are:

$$5.30,\ 8, 20,\ 13.80$$

The mild outliers are:

$$74.10,\ 113.5$$



Boxplot of Example 7

Distribution Shapes, Boxplots and Measures of Location