Displaying Quantitative Data

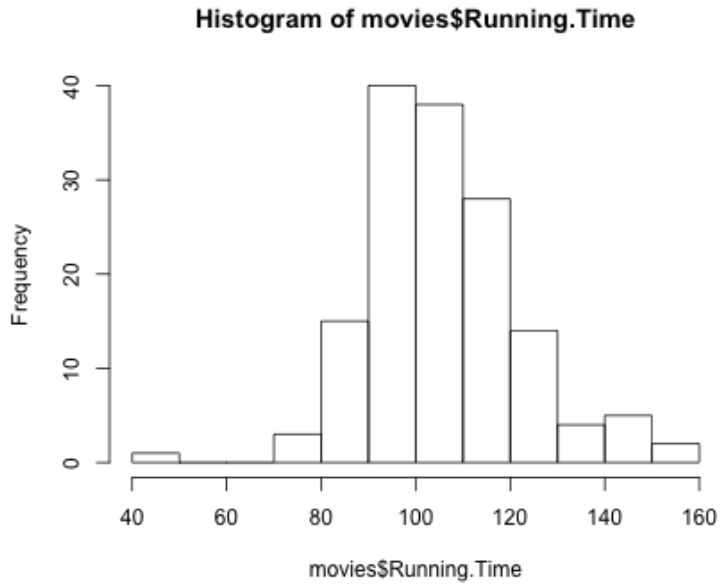
September 11, 2013

Jason Bryer (jason@bryer.org)
epsy530.bryer.org

Histograms

- First introduced by Karl Pearson, a histogram is a graphical representation of the distribution of data.
- The interval of data are divided into bins (on the x-axis) and the y-axis is simply a count of data points within that bin.

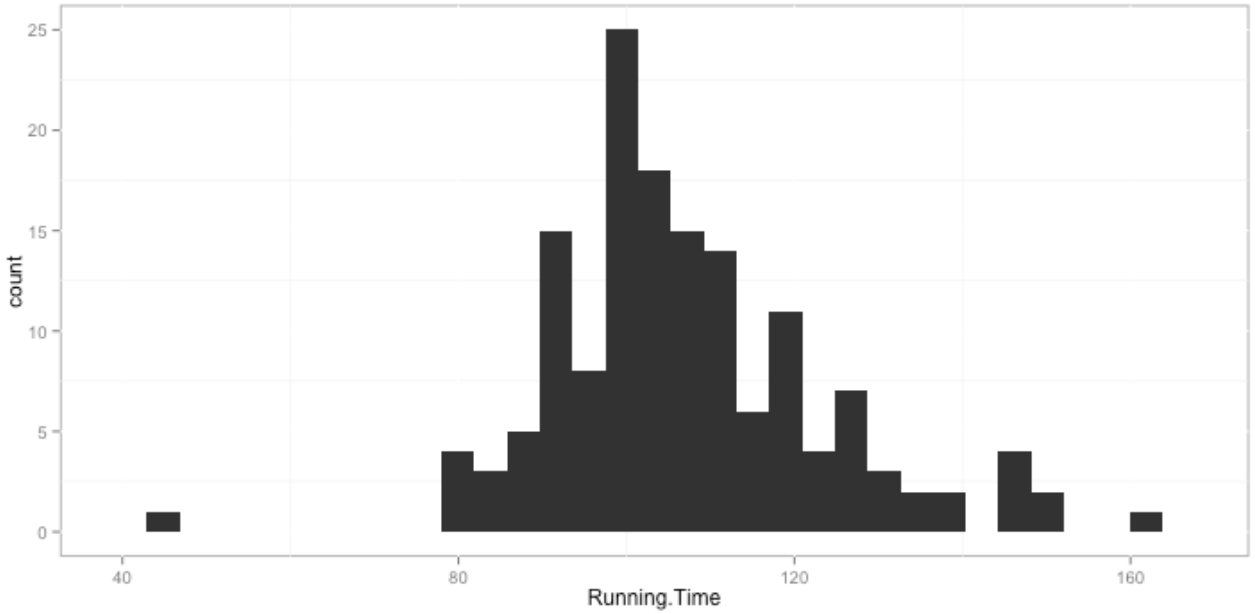
```
hist(movies$Running.Time)
```



Histograms (with ggplot2)

```
ggplot(movies, aes(x = Running.Time)) + geom_histogram()
```

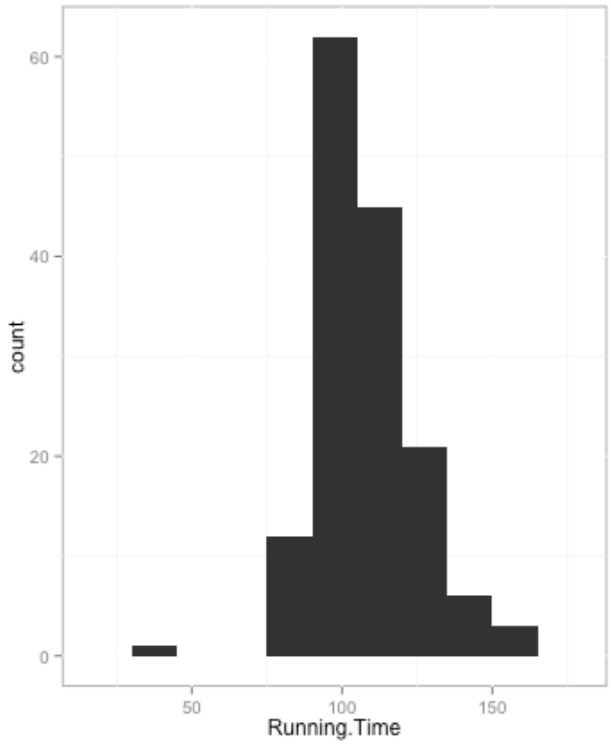
```
stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Histograms (with ggplot2): Binwidths

- Different bin widths tell different stories.
- Choose the width that best shows the important features.
- Presentations can feature two histograms that present the same data in different ways.
- A gap in the histogram means that there were no occurrences in that range.

```
ggplot(movies, aes(x=Running.Time)) +
    geom_histogram(binwidth=15)
```



Outliers

```
movies[movies$Running.Time < 50, ]
```

```
       Title Running.Time
97 Hubble 3D           43
```

Stem and Leaf Plot

```
stem(movies$Running.Time)
```

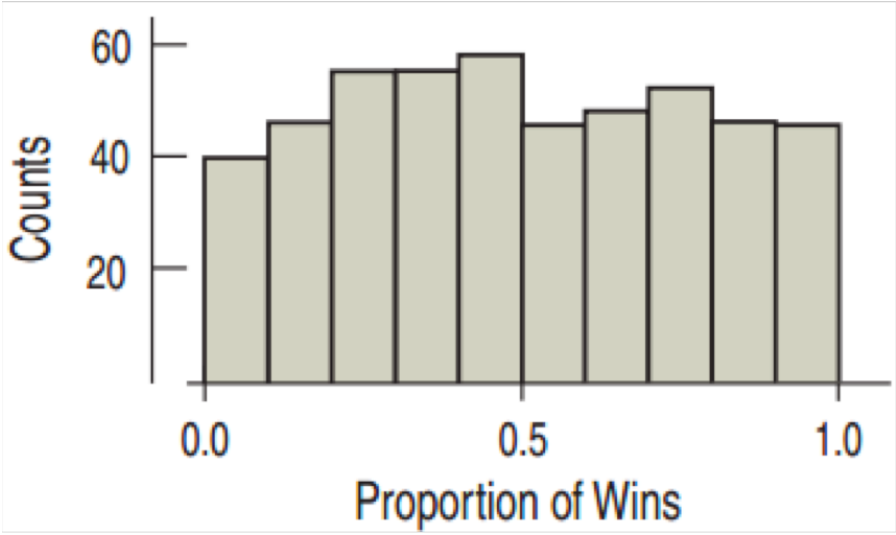
```
The decimal point is 1 digit(s) to the right of the |

 4 | 3
 5 |
 6 |
 7 | 9
 8 | 00122578889
 9 | 000000111222233455567778888888899
10 | 00000000000000112222233333344455556667777778889999
11 | 000111122222334555667778888
12 | 000001234557778899
13 | 0338
14 | 07888
15 | 02
16 | 0
```

Modes

A Mode of a histogram is a hump or high-frequency bin.

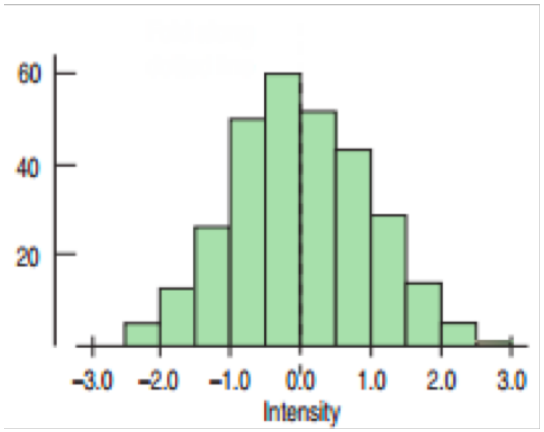
Uniform Distributions

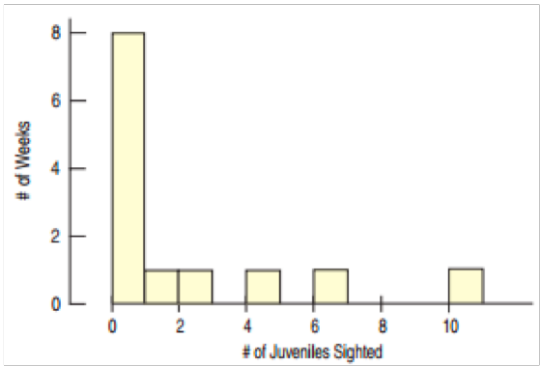
- Uniform Distribution: All the bins have the same frequency, or at least close to the same frequency.
- The histogram for a uniform distribution will be flat.

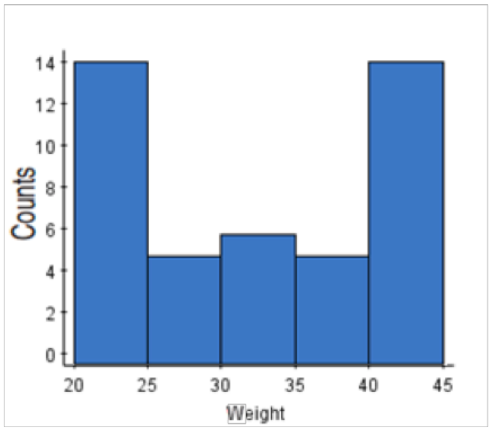


Symmetry

The histogram for a symmetric distribution will look the same on the left and the right of its center.



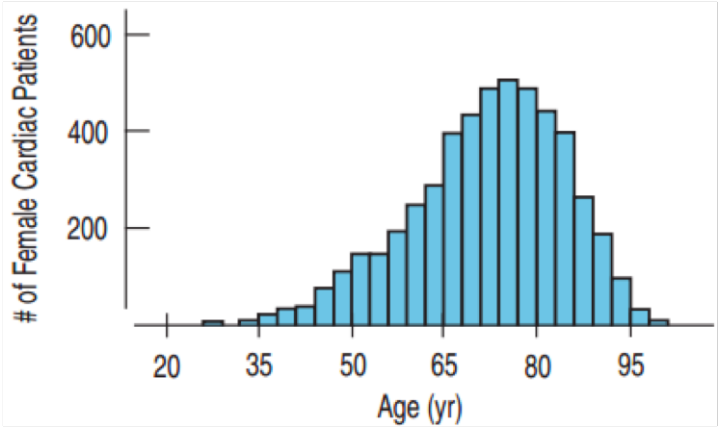




Skewness

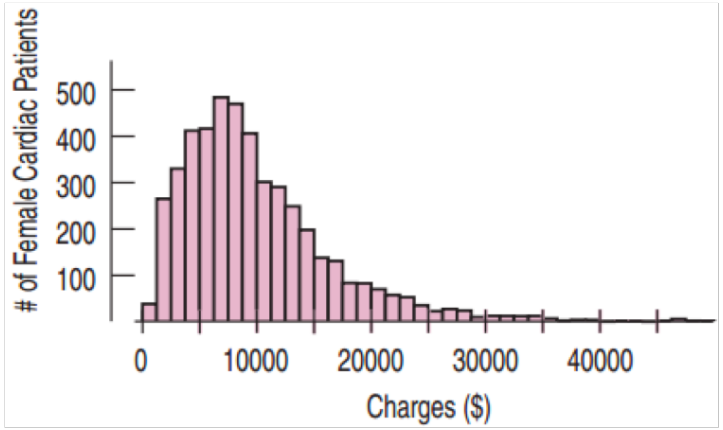
A histogram is skewed left if the longer tail is on the left side of the mode.

Negatively Skewed



A histogram is skewed right if the longer tail is on the right side of the mode.

Positively Skewed



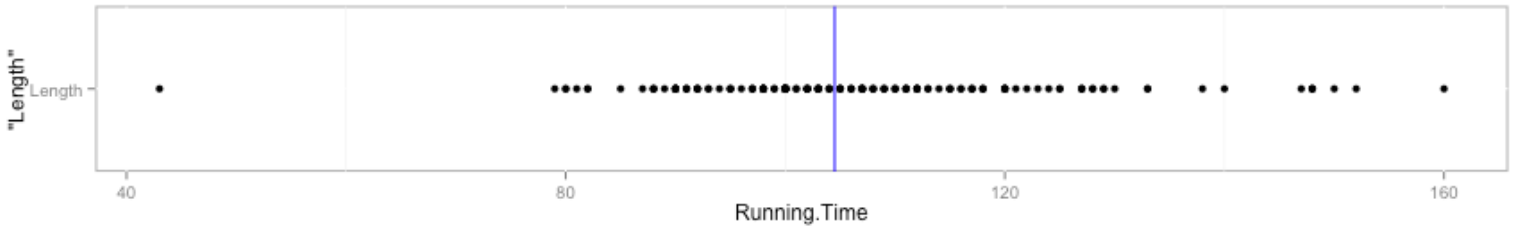
Center: Median

- Median: The center of the data values
- Half of the data values are to the left of the median and half are to the right of the median.

```
(mediantime <- median(movies$Running.Time))
```

```
[1] 104.5
```

```
ggplot(movies, aes(x=Running.Time, y='Length')) + geom_point() +
    geom_vline(xintercept=mediantime, color='blue')
```



Center: Mean

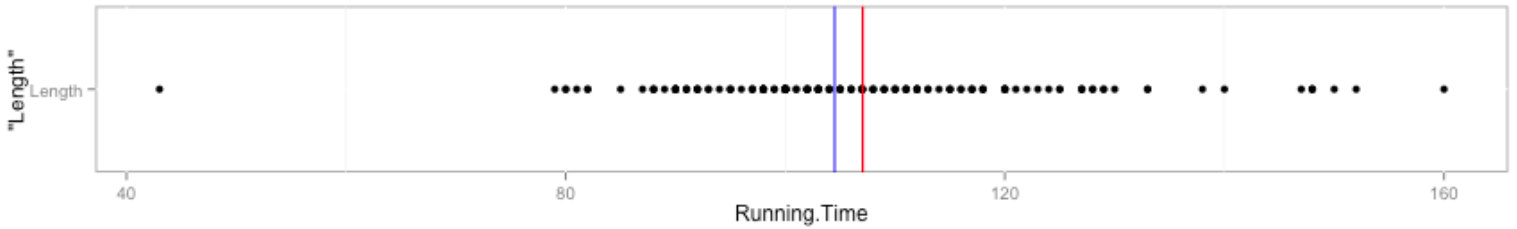
- The Mean is what most people think of as the average.

$$\bar{y} = \frac{\sum y}{n}$$

```
(meantime <- mean(movies$Running.Time))
```

```
[1] 107.1
```

```
ggplot(movies, aes(x=Running.Time, y='Length')) + geom_point() +
    geom_vline(xintercept=mediantime, color='blue') + geom_vline(xintercept=meantime, color='red')
```



Median vs. Mean

- For symmetric distributions, the mean and the median are equal.
- The balancing point is at the center.
- The tail "pulls" the mean towards it more than it does to the median.
- The mean is more sensitive to outliers than the median.
- The mean is larger than the median since it is "pulled" to the right by the outlier
- The median is a better measure of the center for data that is skewed.

Why Use the Mean?

- Although the median is a better measure of the center, the mean weighs in large and small values better.
- The mean is easier to work with.
- For symmetric data, statisticians would rather use the mean.
- It is always ok to report both the mean and the median.

Spread

- Locating the center is only part of the story
- Are the data all near the center or are they spread out?
- Is the highest value much higher than the lowest value?
- To describe data, we must discuss both the center and the spread.

Range

- The range is the difference between the maximum and minimum values. Range = Maximum – Minimum
- The range is sensitive to outliers. A single high or low value will affect the range significantly.

Percentiles and Quartiles

- Percentiles divide the data in one hundred groups.
- The **nth percentile** is the data value such that n percent of the data lies below that value.
- For large data sets, the median is the 50th percentile.
- The median of the lower half of the data is the **25th percentile** and is called the **first quartile** (Q1).
- The median of the upper half of the data is the **75th percentile** and is called the **third quartile** (Q3).
- The **Interquartile Range** (IQR) is the difference between the upper quartile and the lower quartile: IQR = Q3 – Q1

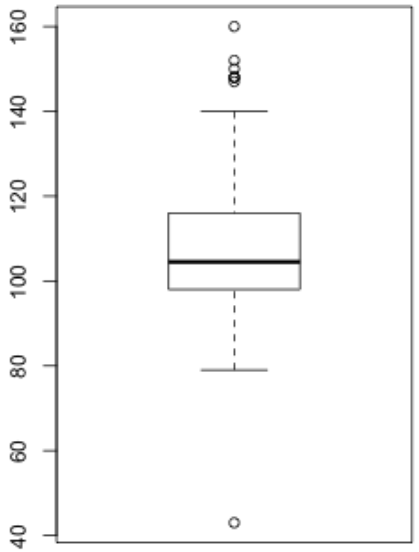
```
summary(movies$Running.Time)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     43      98     104     107     116     160 
```

Boxplots

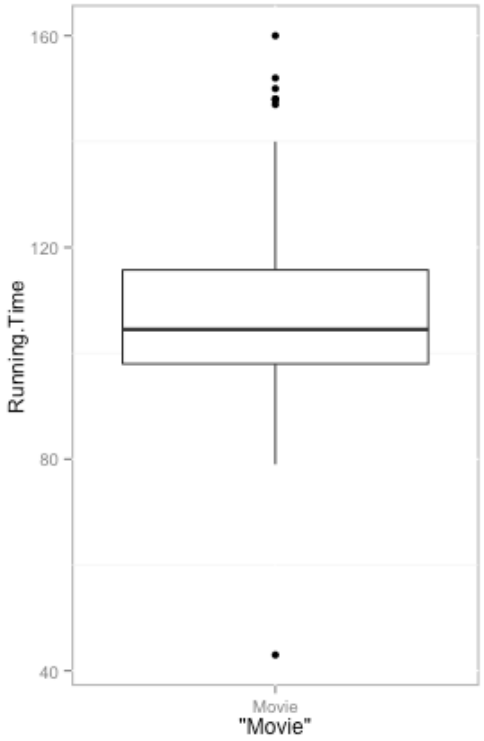
- A Boxplot is a chart that displays the 5-Point Summary and the outliers.
- The Box shows the Interquartile Range.
- The dashed lines are called fences, outside the fences lie the outliers.
- Above and below the box are the whiskers that display the most extreme data values within the fences.
- The line inside the box shows the median.

```
boxplot(movies$Running.Time)
```



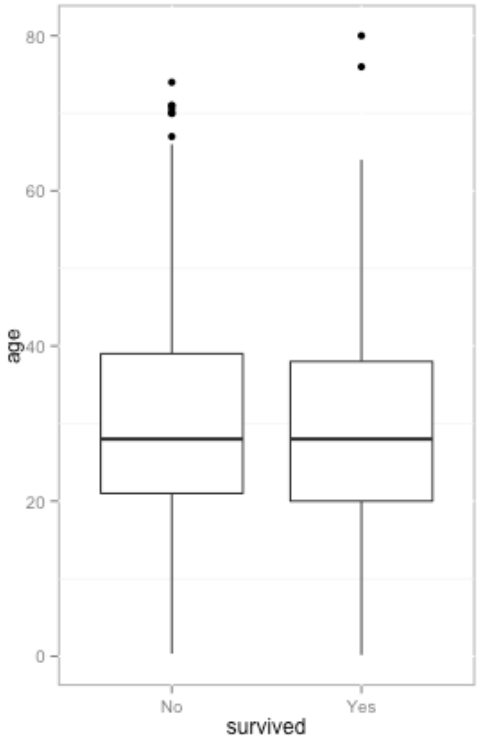
Boxplots

```
ggplot(movies, aes(x = "Movie", y = Running.Time)) + geom_boxplot()
```



Boxplots

```
ggplot(titanic, aes(x = survived, y = age)) + geom_boxplot()
```



Variance

$$S^2 = \frac{\sum (y - \bar{y})}{n - 1}$$

- The variance is a measure of how far the data is spread out from the mean.
- The difference from the mean is: $y - \bar{y}$
- To make it positive, square it.
- Then find the average of all of these distances, except instead of dividing by n, divide by n – 1.
- Use S^2 to represent the variance.
- The variance will mostly be used to find the standard deviation s which is the square root of the variance.

Standard Deviation

$$S^2 = \sqrt{\frac{\sum (y - \bar{y})}{n - 1}}$$

- The variance's units are the square of the original units.
- Taking the square root of the variance gives the standard deviation, which will have the same units as y.
- The standard deviation is a number that is close to the average distances that the y values are from the mean.
- If data values are close to the mean (less spread out), then the standard deviation will be small.
- If data values are far from the mean (more spread out), then the standard deviation will be large.