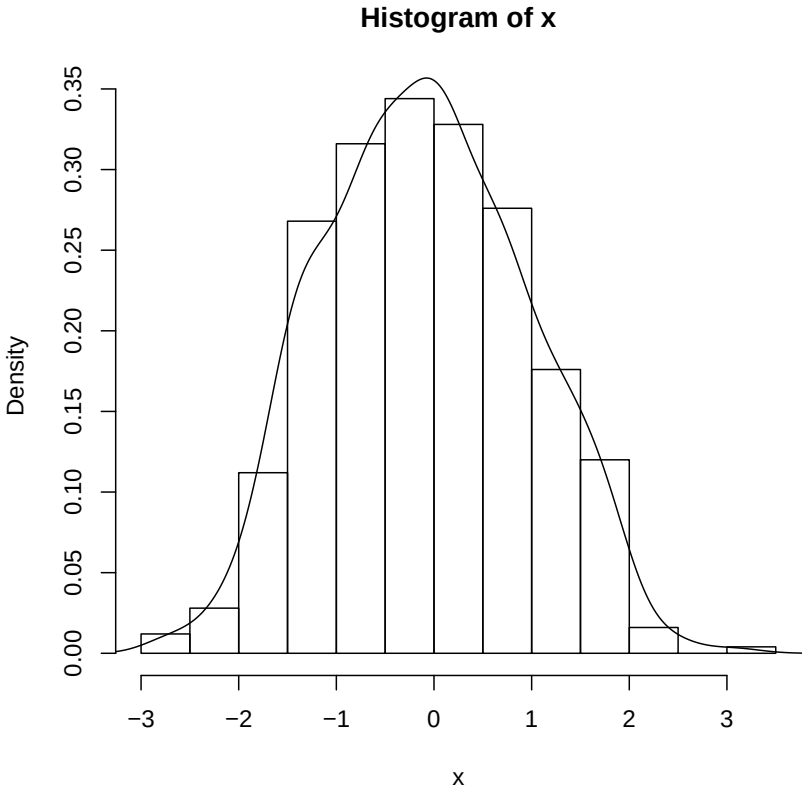
Math 324 Lecture 4
Dr Bolstad

http://math324sfsu.bmbolstad.com

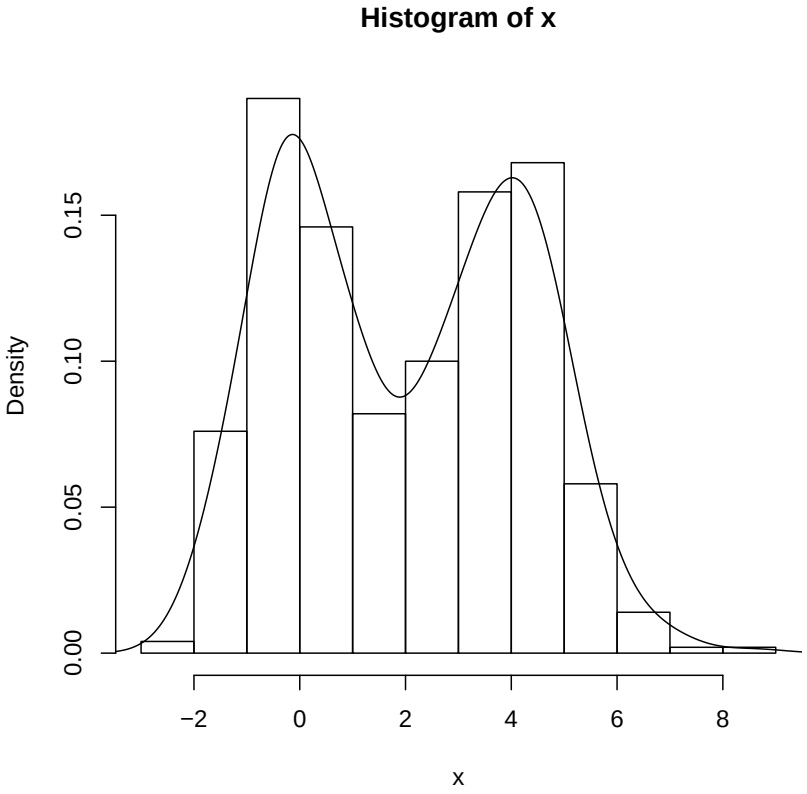
What are we looking for when we look at these plots?

We are looking at the *shape*, *center* and *spread* of the data. We are also looking for observations that seem to deviate from the rest of the data in some way. We call these observations *outliers*.

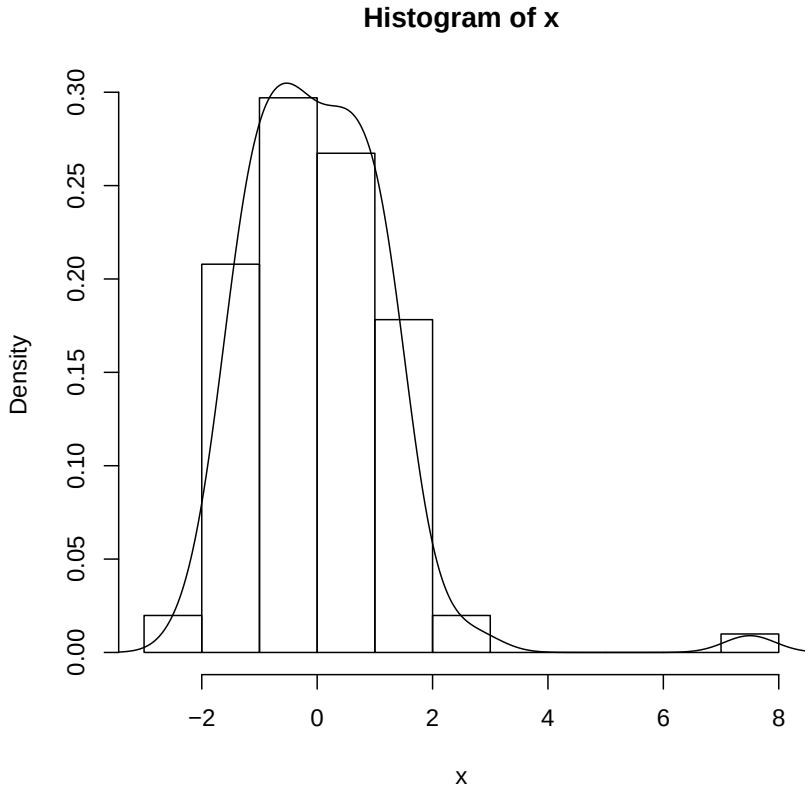
Unimodal centered at 0. basically symmetric



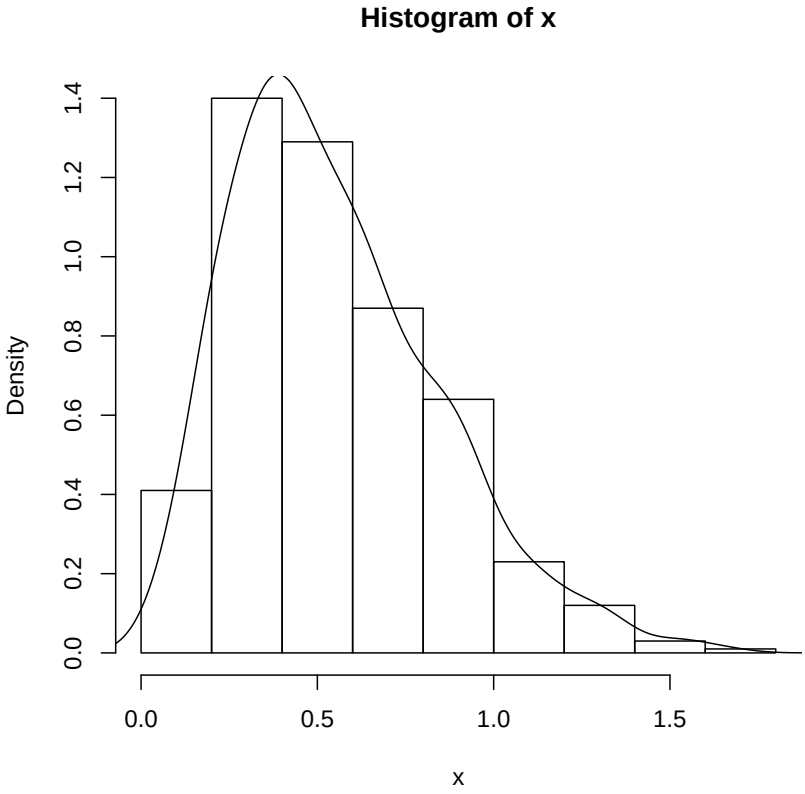
Bimodal, with modes at 0 and 4, centered at about 2



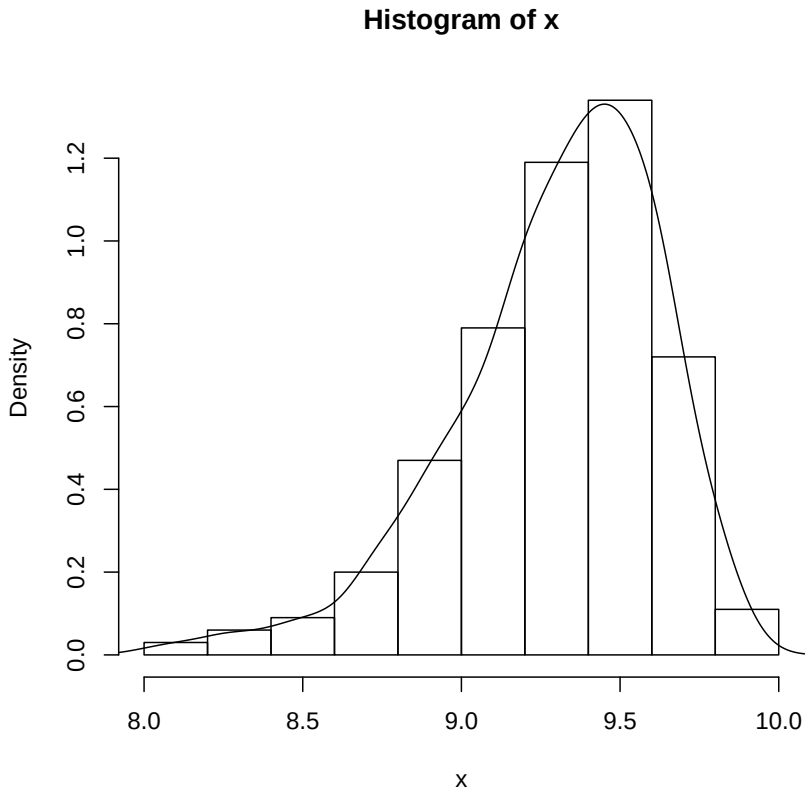
Most of the data is symmetric and centered around 0, but we have an few "outliers"



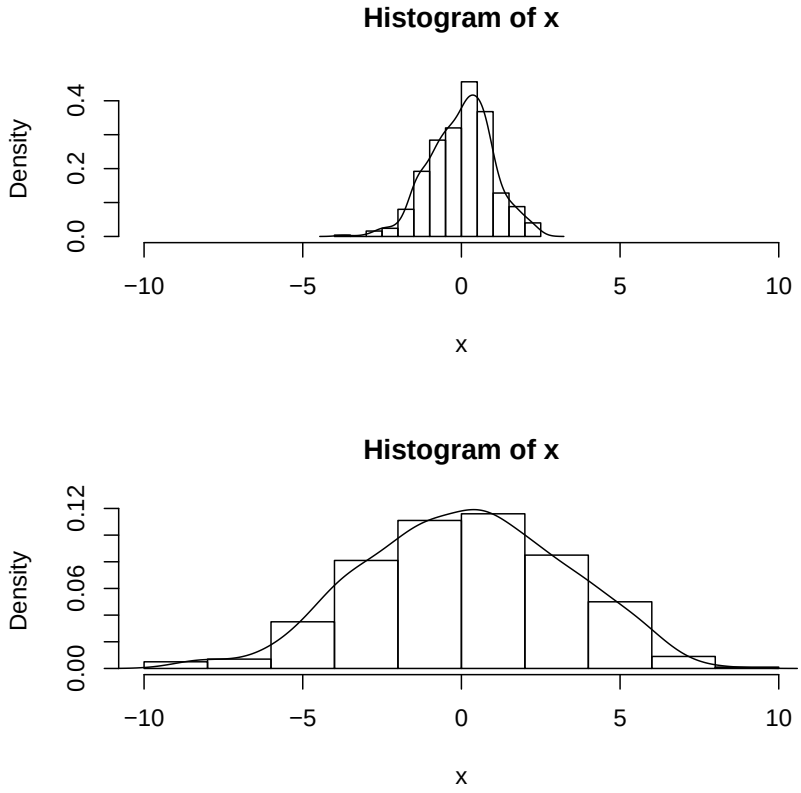
Not symmetric, skewed to the right (positively skewed)



Not symmetric, skewed to the left (negatively skewed)



Two datasets, both are centered around 0, but one is much more spread out then the other.



Summary Statistics

We have talked about visually assessing the distribution of data using histograms and stem-plots. In particular we are often interested in the location and the spread of the data. Rather than doing it graphically it is also possible to use numerical summary values to assess the center, location and spread.

Mean

The mean, sometimes called the average, is a commonly used method of measuring the center. In words it is the value that is given by the *sum of*

the value for all the observations divided by the number of observations. In mathematics we write

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

where $x_1, x_2, \ldots, x_n$ are the observed data values.

Example

Suppose we have the following data

```
5.3, 4.5, 3.2, 1.9, 6.5
```

the mean is given by

```
(5.3 + 4.5 + 3.2 + 1.9 + 6.5)/5 = 21.4/5 = 4.28
```

Median

The median is the center of the distribution. In words it is the *data value that is larger than half the data and smaller than the other half.* The median is computed by the following steps

1. Arrange all the data values so that they are ordered from smallest to largest.
2. If the number of observations, n is odd then the $\frac{n+1}{2}$th observation in the sorted list is the median
3. If the number of observations, n is even then the median is the average of the $\frac{n}{2}$th and $\frac{n}{2} + 1$th observations in the sorted list

Example 1

Suppose we have the following data

```
9, 3, 5, 7, 12
```

first sorting the data

```
3, 5, 7, 9, 12
```

we have 5 data values so the median is the 3rd value

```
7
```

Example 2

Suppose we have the following data

```
100, 105, 98, 63, 102, 101
```

first sorting the data

```
63, 98, 100, 101, 102, 105
```

we have 6 data values so the median is the average of the 3rd and 4th values in the list

```
(100 + 101)/2 = 100.5
```

Which one to use: mean or median?

If the distribution from which the data came from is symmetric, then the mean and median are close together. If the data is skewed then the mean will be further out in the longer tail the median. The median is robust, this means that it is not affected by outliers. For example, suppose we are interested in the income distribution of people in the USA. Most people earn relatively moderate incomes, a few people earn extremely large incomes (eg Bill Gates). If we take the mean we will get a much larger estimate of the center of the income distribution, than if we take the median.

Other measures of location

Rather than using the mean or median, sometime a *trimmed mean* is used. This is a compromise between the median which is very insensitive to outliers and the mean which is very sensitive to outliers. A 10% trimmed mean is computed to removing the lowest 10% and highest 10% of data values and then computing the average of the remaining data values.

Besides looking at the median, which divides the data exactly in half, we might also look for the data values where 25% is lower and 75% is higher an d also the value where 75% is lower and 25% is higher. These are known as the *lower quartile* (LQ) and *upper quartile* (UQ) respectively. Together, with the median it allows us to divide the data into four parts. You can think of the lower quartile as being the median of the data values to the left of the median and the upper quartile as being the median of the data values to the right of the median.

Standard deviation and variance

The *standard deviation*, and a closely related quantity, the *variance* are the most commonly used measures of the variability of data. In words the variance is the *average of the sum of the squares of the deviations from the mean.* Mathematically, the variance is

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

the standard deviation is the square root of the variance. ie $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$. Why do we square the deviations? Because we would get 0 if we summed the deviations. Why is the denominator $n-1$? Because it makes the estimate *unbiased* (we will talk about this more later in the class). Why is the standard deviation used more than the variance? Because it is in the same units of measurement as the data, whereas the variance will be in the squared units.

An alternative formula for the variance is

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

The standard deviation is always positive. It only equals zero when there is no spread in the data. The standard deviation is not resistant to outliers.

Range

The *range* is given by the distance between the largest data value and the smallest data value.

IQR

The *interquartile range* is a more robust measure of variability. It is given by the difference between the upper and lower quartile. ie

$$IQR = UQ - LQ$$

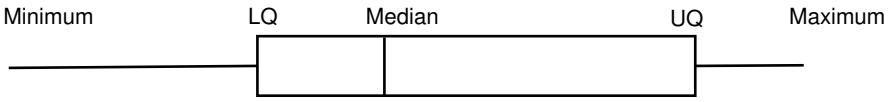
Five number summary

The five number summary consists of the following five numbers reported in the following order: minimum, LQ, median, UQ, maximum.

Boxplot

A *boxplot*, sometimes called a box and whisker plot, is a graphical version of the five number summary. it is created in the following manner.

1. Draw a box that spans the LQ and UQ.
2. Divide the box by drawing a line at the median
3. Extend a lines out from the ends of the box to the minimum and maximum.



A common modification is to use the IQR to establish a criterion for outliers. In particular, any observation that is more than 1.5 times the IQR above the UQ or more than 1.5 times the IQR below the LQ. A modified boxplot is to then drawn such that the whiskers (lines) are only extended out as far as the observations that are not outliers. Outliers are marked as individual points.

An example

Suppose we have the data

```
0.8 0.9 1.4 0.6 0.2 2.0 0.9 3.9 2.4 1.0
```

sorting the data gives us

```
0.2 0.6 0.8 0.9 0.9 1.0 1.4 2.0 2.4 3.9
```

The median is

```
(0.9 + 1.0)/2 = 0.95
```

the LQ is given by `0.8` and the upper quartile is `2.0`. The IQR is `2.0 - 0.8 = 1.2`. 1.5 times the IQR above the UQ is `3.8` and 1.5 times the IQR below the LQ is `-1.0`.

The resulting boxplot