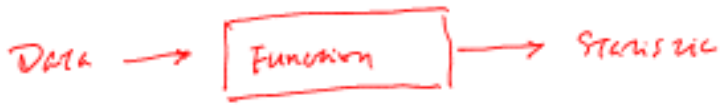
Statistics. What is it? Or, what are they?

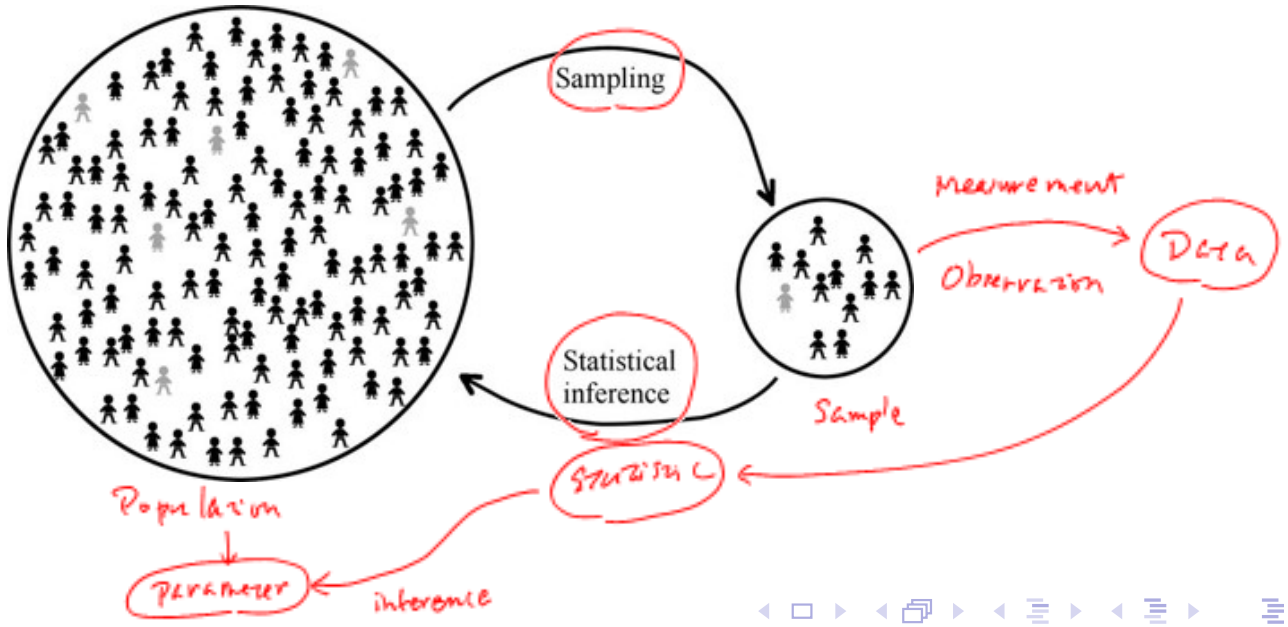
Statistics is the science of learning from data.

A statistic is a measurement constructed from data which intersa parameter of population.



Population and sample.

A **population** is the complete collection to be studied; it contains all subjects of interest. A **sample** is a part of the population of interest, a subset selected by some means from the population. In statistics we sample subjects from a large population and use the information obtained from the sample to infer characteristics about the general population. Thus the upper arrow can be viewed as "sampling" while the lower arrow is "statistical inference."



Parameter and statistic.

A **parameter** is a numerical value that describes a characteristic of a population, while a **statistic** is a numerical measurement that describes a characteristic of a sample. We will use a statistic to infer something about a parameter.

EXAMPLE:

- The true average height of the population is a parameter. It would be too expensive and time-consuming to measure the height of all individuals in the population.
- Instead we draw a random sample of 12 individuals and measure the height of each of them. The average of those individuals is our statistic.

Statistical data analysis is concerned with methods for making inferences about population parameters based on sample statistics.

Categorical data.

Categorical data can be grouped into categories based on some qualitative trait. The resulting data are merely labels or categories, and examples include gender (male and female) and ethnicity (e.g., Caucasian, Asian, African).

1. **Nominal.** When there is no natural ordering of the categories we call the data nominal—gender, race, smoking status (smoker or non-smoker), or disease status.
2. **Ordinal.** When the categories may be ordered, the data are called ordinal variables—judge pain (e.g., none, little, heavy) or income (low-level income, middle-level income, or high-level income).

Quantitative data.

Quantitative data are numerical measurements where the numbers are associated with a scale measure rather than just being simple labels.

1. **Discrete.** Discrete quantitative data are numeric data variables that have a finite or countable number of possible values—household size or the number of kittens in a litter.
2. **Continuous.** The real numbers are continuous with no gaps; physically measurable quantities like length, volume, time, mass, etc., are generally considered continuous.

Using data in R: Tenderness of pork.

The data file "cooling.txt" is from the study of two different cooling methods for pork meat. The first line of the file is called a "header," representing the variable names of data set. To read this type of data set, we need to use the following command:

role of '<-'

```
databox <- read.table(file.choose(), header=T)
```

The argument "file.choose()" allows us to choose the file interactively from a dialog box presented to the user. Once it is read, we can view "databox" in use, specify a particular variable name in header, for example, "tunnel", and set it as a new variable "x"

reference to each column name

```
databox
x = databox$tunnel
x
```

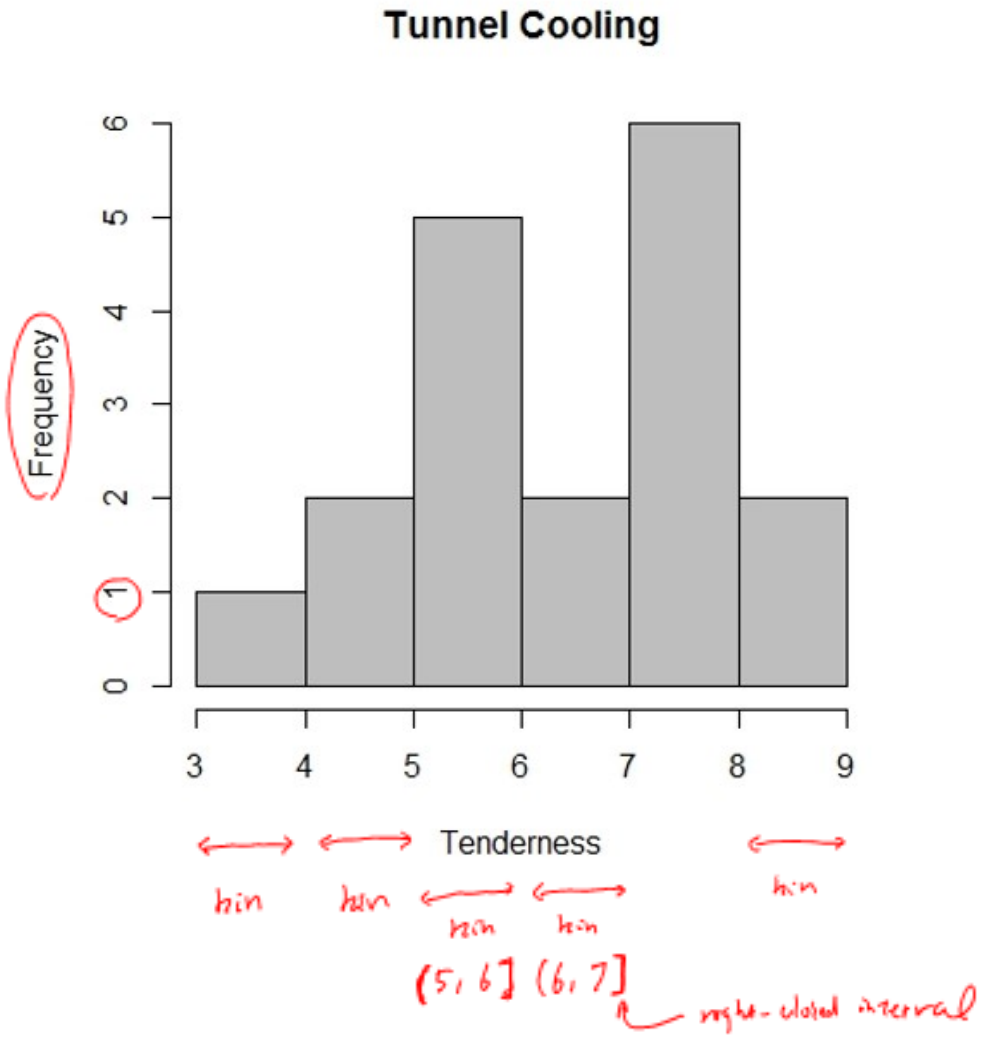
attach(databox)

x = tunnel

Histogram: Visualizing quantitative data.

By grouping the quantitative continuous data into bins, we can count the number of observations that fall into each bin and the resulting bins, and their related relative frequencies give the distribution of the quantitative continuous variable.
A **histogram** allows us to graphically summarize the shape of distribution; e.g., symmetric, skewed, and number of modes in the data. The **relative frequency histogram** can be used to compare the distributions from different populations since the relative frequency histogram has the inherent feature that areas for each bar in the histogram are proportional to the probability.

Example: Tenderness of pork.



Example: Tenderness of pork, continued.

Histograms and relative frequency histograms are both produced with the `hist()` function. By default the `hist()` function automatically groups the quantitative data vector into bins of equal width and produces a frequency histogram. We can make a frequency plot or a relative frequency plot by specifying either "freq=T" or "freq=F" option, respectively. The number of bins is controlled by "breaks" option. Also, by default the bins are closed at the right end. The option "right=F" makes bins left-closed and right-open [a,b).

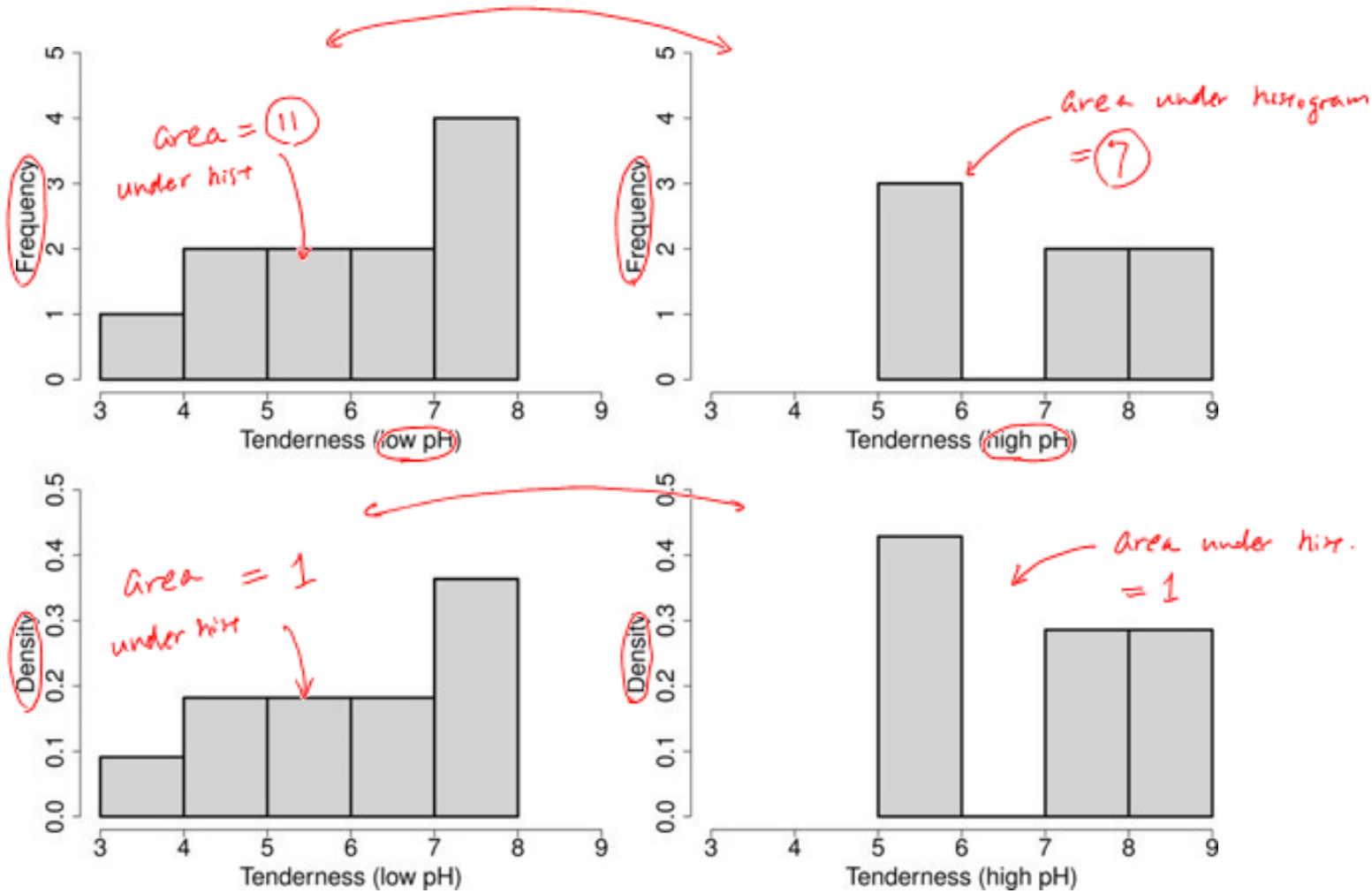
```
hist(x)
hist(x, freq=F)
hist(x, breaks=10)
hist(x, right=F)
hist(x, col="gray", main="Tunnel Cooling")
```

Example: Tenderness of pork, continued.

Two different cooling methods for pork meat were compared in an experiment with 18 pigs from two different groups: low or high pH content. After slaughter, each pig was split in two and one side was exposed to rapid cooling while the other was put through a cooling tunnel. After the experiment, the tenderness of the meat was measured.

Notice that the shapes for the low- and high-pH groups do not change from the histograms to the relative frequency histograms. The relative frequency histograms make it easier to compare the distributions in the low- and high-pH groups since the two groups have different numbers of observations.

Example: Tenderness of pork, continued.



Example: Tenderness of pork, continued.

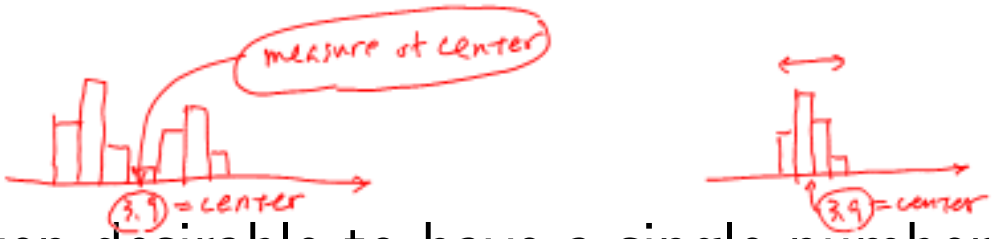
The data "databox" contains the variable "ph" of either "low" or "high" which allows us to select the variable "x" with each attribute.

```
ph = databox$ph
x[ph=="low"]
x[ph=="high"]
```

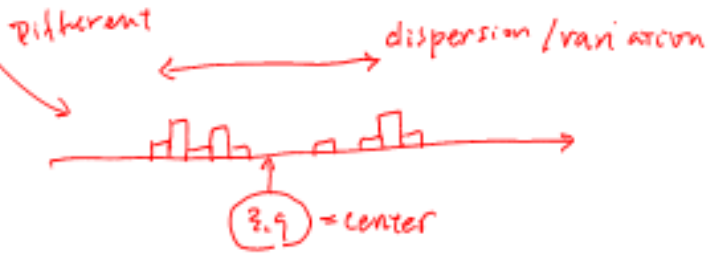
The following commands produce multiple histograms and lower seen in the previous page by first declaring 2 by 2 plots.

```
par(mfrow=c(2,2))
hist(x[ph=="low"], xlab="Tenderness (low pH)")
hist(x[ph=="high"], xlab="Tenderness (high pH)")
hist(x[ph=="low"], freq=F, xlab="Tenderness (low pH)")
hist(x[ph=="high"], freq=F, xlab="Tenderness (high
pH)")
```

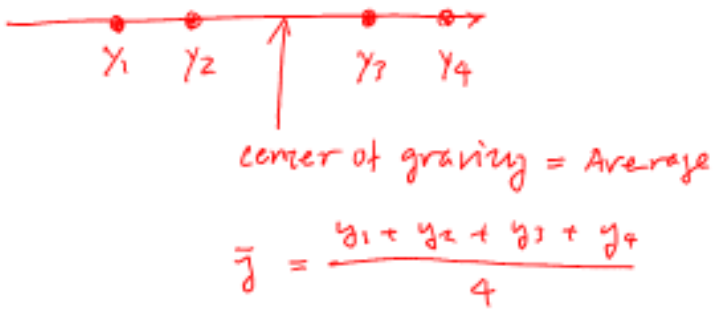
Measures of center and variation.



It is often desirable to have a single number to describe the values in a dataset, and this number should be representative of the data. It seems reasonable that this representative number should be close to the "middle" of the data such that it best describes all of the data, and we call any such number a measure of **central tendency**. Very different sets of data can have the same central tendency. Thus a single representative number is insufficient to describe the distribution of the data, and we are also interested in how closely the central tendency represents the values in the dataset. The **dispersion** represents how widely the data are "spread out."



Sample mean. = Average

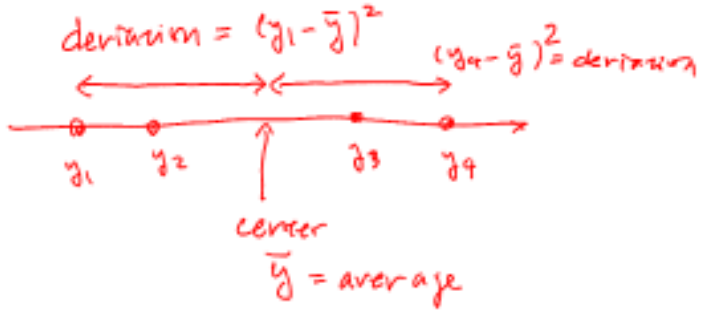


Let $y_1, \ldots, y_n$ denote the quantitative observations in a sample of size n from some population. The sample mean is defined as

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

and is calculated as a regular average. We add up all the observations and divide by the number of observations.

Sample standard deviation.



The **sample standard deviation** is a measure of dispersion for quantitative data and is defined as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_n-\bar{y})^2}{n-1}}$$

Loosely speaking, the standard deviation measures the "average deviation from the mean" observed in the sample; i.e., the standard deviation measures how far away from the center we can expect our observations to be on average.

Sample variance.

The **sample variance** is denoted s^2 and is simply the sample standard deviation squared:

one degree of freedom is used by calculation of $\bar{y}$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{1.1}$$

current degree of freedom

The sample variance is roughly the average of the squared deviations. It would be the average if we divided the sum in (1.1) by n instead of $n-1$. The variance of the population (not the sample, but the population) is $\sigma^2 = \sum_{i=1}^{n}(y_i - \mu)^2/n$, which requires knowledge about the true mean of the population, μ. We divide by $n-1$ in (1.1) in order to take uncertainty about the estimate of μ into account.

Example: Tenderness of pork.

The mean of the tunnel cooling data is

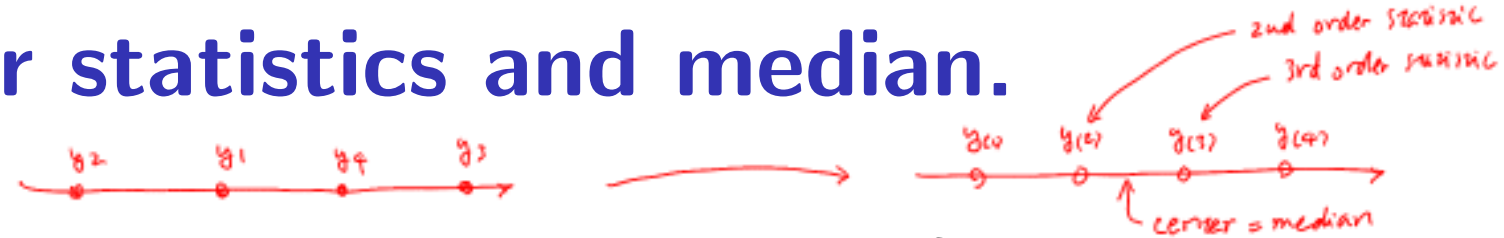
$$\bar{y} = \frac{3.11 + 4.22 + \cdots + 8.67}{18} = 6.382$$

The standard deviation becomes

$$s = \sqrt{\frac{(3.11 - 6.382)^2 + (4.22 - 6.382)^2 + \cdots + (8.67 - 6.382)^2}{18 - 1}}$$
$$= 1.527$$

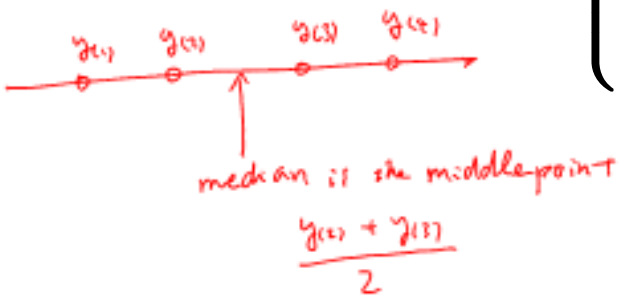
Thus the mean tenderness for tunnel cooling is 6.382 and the corresponding standard deviation is 1.527 units on the tenderness scale.

Order statistics and median.



We can order the observations $y_1, \ldots, y_n$ from lowest to highest and we use the following notation to represent the set of ordered observations: $y_{(1)}, \ldots, y_{(n)}$. Thus $y_{(1)}$ is the smallest value of $y_1, \ldots, y_n$, $y_{(2)}$ is the second smallest, etc. The **median** of n numbers is a measure of the central tendency and is defined as the middle number when the numbers are ordered. If n is even then the median is the average of the two middle numbers:

$$\text{Median} = \begin{cases} y_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd; that is, } n = 3, 5, 7, 9, \cdots \\ \dfrac{y_{(n/2)} + y_{(n/2+1)}}{2} & \text{if } n \text{ is even. that is, } n = 2, 4, 6, 8, \end{cases}$$



Example: Tenderness of pork.

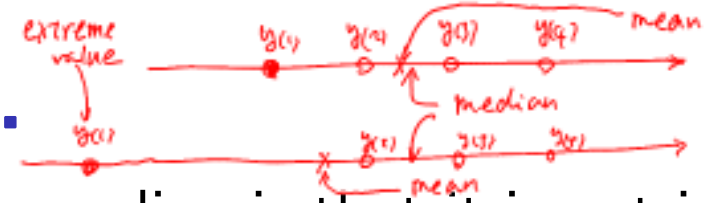
If we order the 18 measurements for tunnel cooling from the pork tenderness data according to size we get

3.11 4.22 4.33 5.11 5.56 5.78 5.78 6.00 6.44
6.78 7.11 7.22 7.33 7.44 7.56 8.00 8.44 8.67

such that $y_{(1)} = 3.11$, $y_{(2)} = 4.22$, etc. There is an even number of observations in this sample, so we should take the average of the middle two observations to calculate the median

$$\text{Median} = \frac{6.44 + 6.78}{2} = 6.61$$

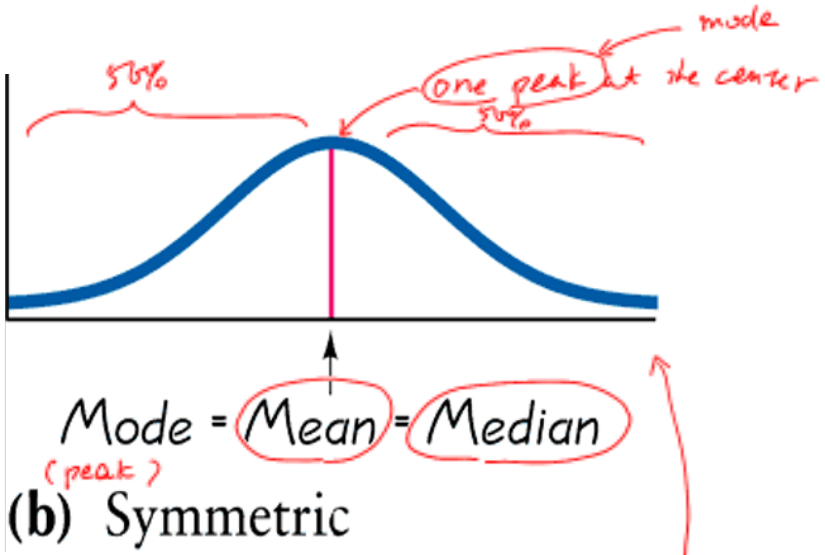
Mean or median.

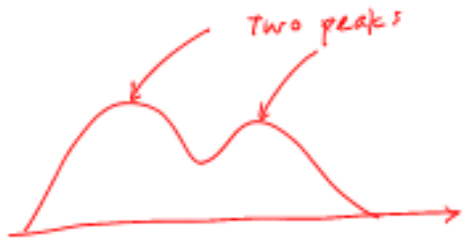


- One advantage of the median is that it is not influenced by extreme values in the dataset. Only the two middle observations are used in the calculation, and the actual values of the remaining observations are not used.
- The mean on the other hand is sensitive to all values in the dataset since every observation in the data affects the mean, and extreme observations can have a substantial influence on the mean value.
- Generally the mean is used for symmetric quantitative data, except in situations with extreme values, where the median is used.
- The mean value has some very desirable mathematical properties that make it possible to prove useful results within statistics, and inference methods naturally give rise to the mean value as a parameter estimate.

Shape of distribution.

The peaks of a distribution often represent its modes. When there is only one peak, the distribution is **unimodal**. When the distribution has two peaks, we call it **bimodal**. When the shape is folded and the left and right folds are each other's mirror images, the distribution is **symmetric**.

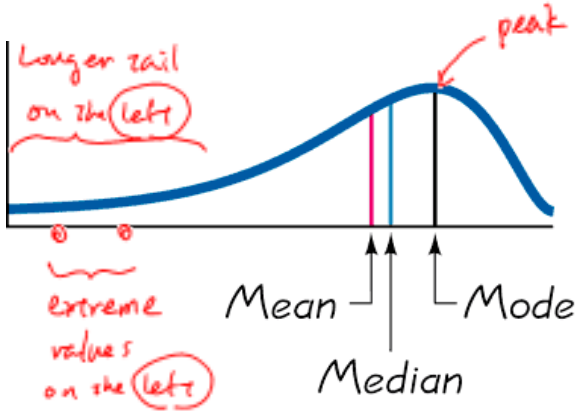




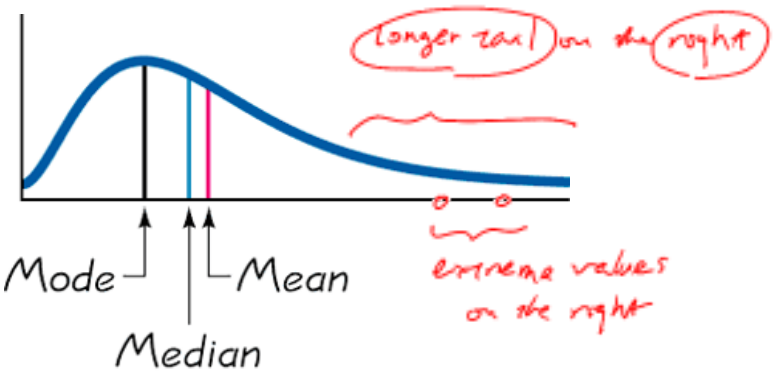
When the distribution is unimodal and symmetric, its mean, mode, and median will have the same values.

Shape of distribution, continued.

When the shape of the distribution is asymmetric, we can see whether the distribution is **skewed to the left** or **skewed to the right**.



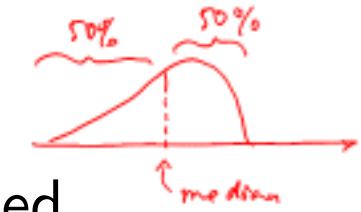
(a) Skewed to the Left (Negatively)



(c) Skewed to the Right (Positively)

When the curve is skewed to the left, we expect that the mode has the largest value followed by the median and then the mean. Similarly, when the shape of the distribution is skewed to the right, the mean has the highest value followed by the median and then the mode.

Quartiles and inter-quartile range.

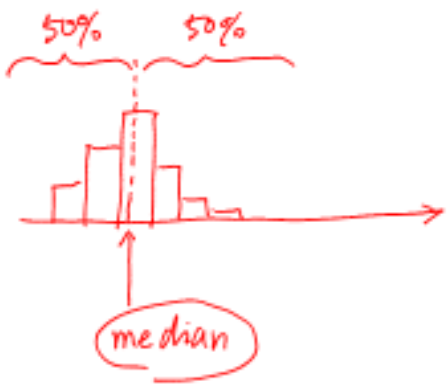


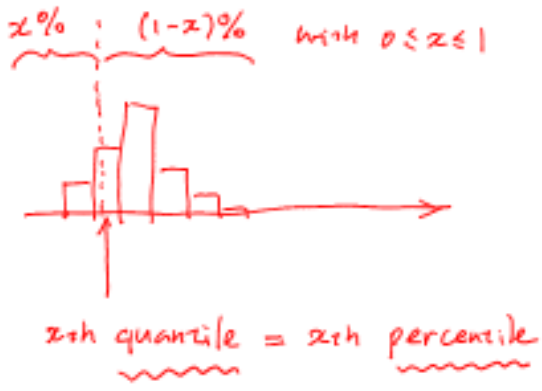
We divide the ordered data into 100 essentially equal-sized subcollection of points such that the **x-th quantile** is defined as the cut-off point where x% of the sample has a value equal to or less than the cut-off point. For example, the 40th quantile splits the data into two groups containing, respectively, 40% and 60% of the data. The **first quartile** Q1 is defined as the 25th quantile, and the **third quartile** Q3 is defined as the 75th quantile, so the first quartile, the median, and the third quartile split the data into 4 groups of equal size. Another measure of variation is the **inter-quartile range** (IQR), and it is calculated as follows:

$$\text{IQR} = \text{Q3} - \text{Q1}$$

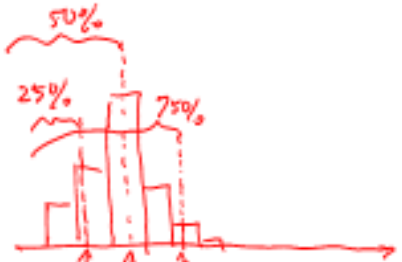
The advantage of the IQR over the standard deviation is that the IQR is not as sensitive to extreme values because the IQR is based on the middle 50% of the observations.

What is quantile? Extend a notion of median → Percentile





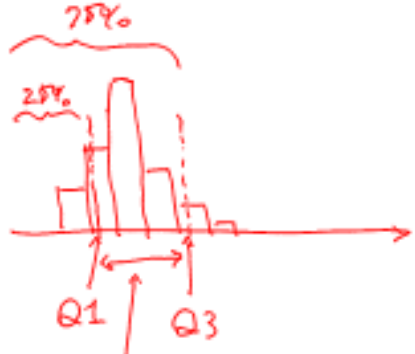
What is quartile?



1st quartile or lower quartile

2nd quartile = median

3rd quartile or upper quartile



IQR (Interquartile Range) = Q3 - Q1

Summary statistics: Tenderness of pork.

To calculate sample statistics of the variable "x", use the summary command. To find out the data size, use the length command.

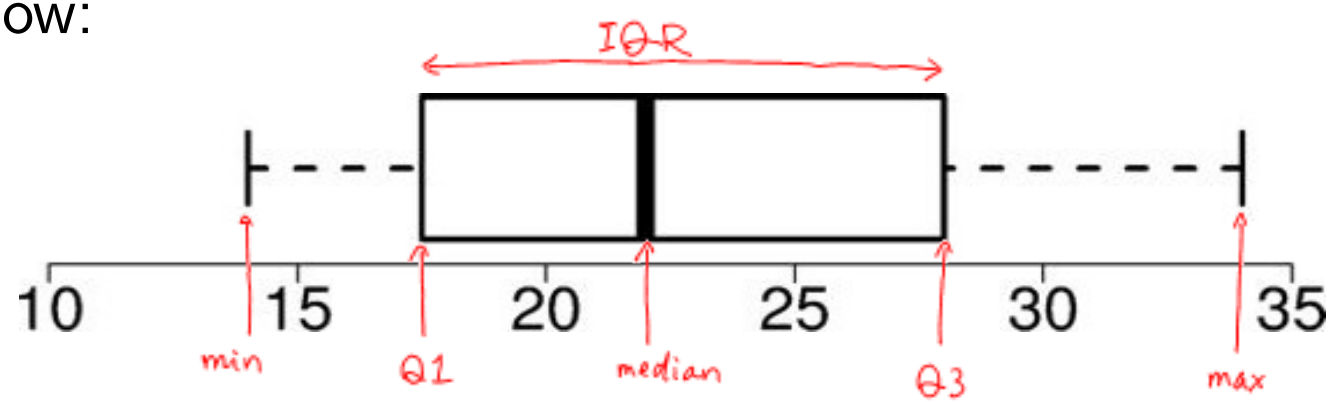
```
summary(x)
length(x)
```

We can use the mean(), median(), quantile(), IQR(), sd() functions to individually calculate the mean, median, quantile, inter-quartile range, standard deviation for the variable "x".

```
mean(x)
median(x)
quantile(x,1/4)
quantile(x,3/4)
IQR(x)
sd(x)
```

Boxplot.

A **boxplot** (also called a **box-and-whiskers plot**) summarizes the data graphically by plotting the following five summaries of the data: minimum, first quartile, median, third quartile, and maximum, as shown below:



The middle 50% of the data are represented by a box and the median is shown as a fat line inside the box. Two whiskers extend from the box to the minimum and the maximum value.

Boxplot: Visualizing summary statistics.

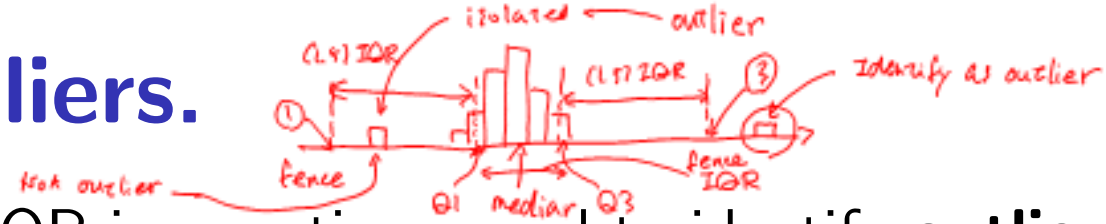
Data can be directly entered into the variable "y" by using c() to directly "combine" data values and set them to a variable "y":

```
y = c(14,16,19,22,26,30,34)
summary(y)
```

A boxplot is a mere visualization of summary statistics, starting from the minimum, quartiles, and maximum. A standard boxplot is vertically displayed. In order to make it horizontally, "horizontal=T" option is

```
boxplot(y)
boxplot(y,horizontal=T)
```

Outliers.



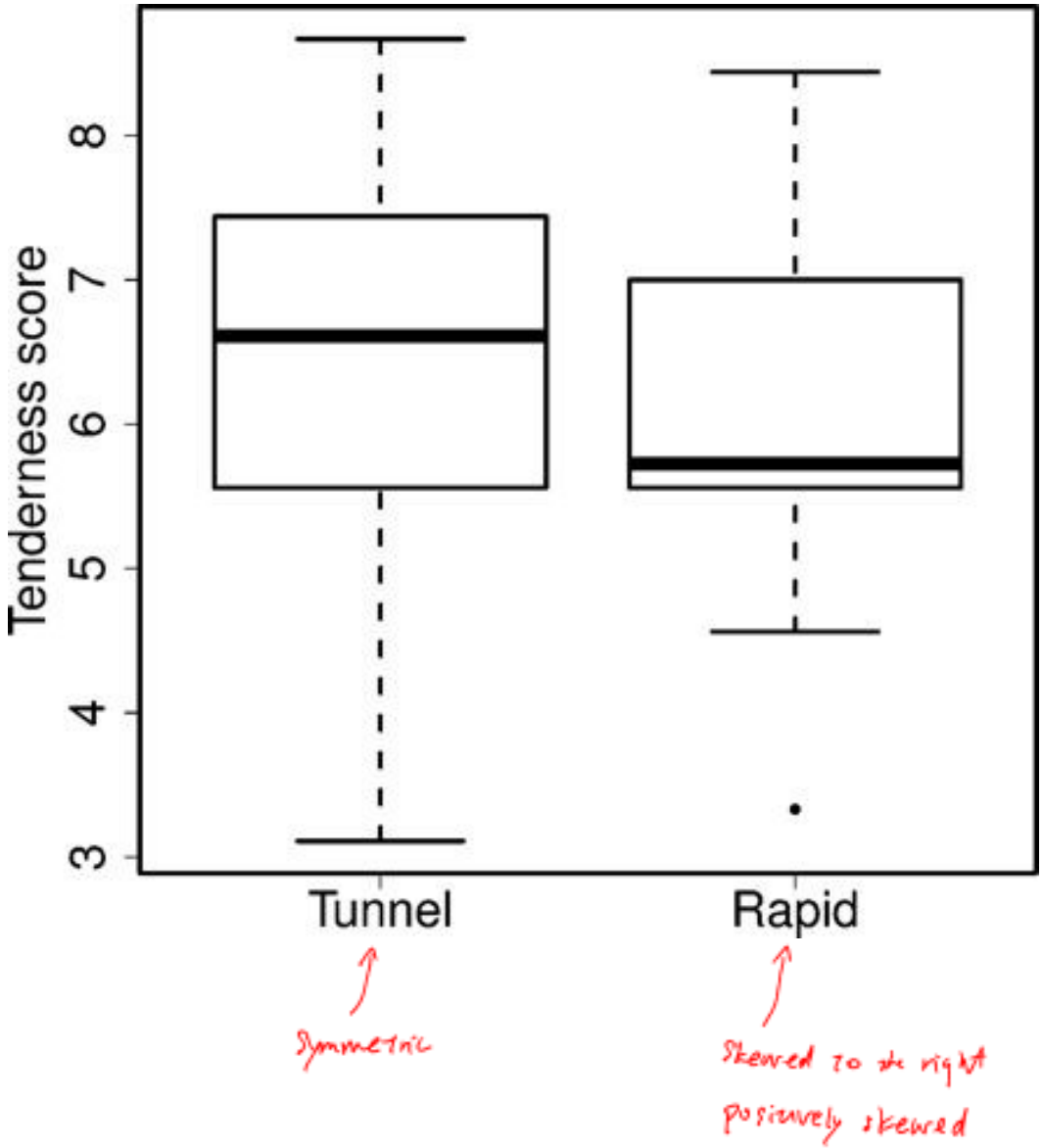
The IQR is sometimes used to identify **outliers**—observations that differ so much from the rest of the data that they appear extreme compared to the remaining observations. As a rule-of-thumb, an outlier is an observation that is smaller than (1.5)IQR under the first quartile or larger than (1.5)IQR over the third quartile; i.e., anything outside the following interval:

$$[Q1 - (1.5)IQR, Q3 + (1.5)IQR] \tag{1.2}$$

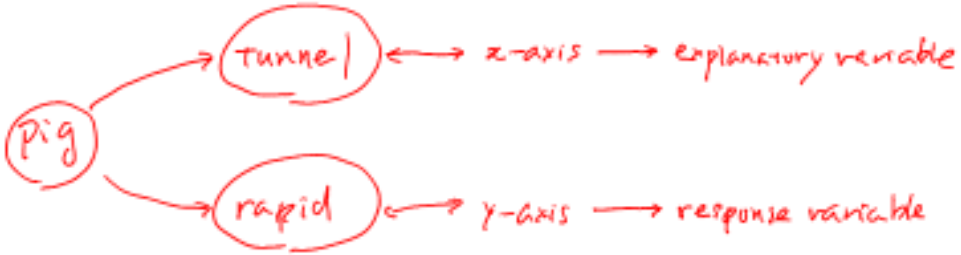
It is often critical to identify outliers and extreme observations as they have an enormous impact on the conclusions of statistical analysis. In a **standard boxplot**, outliers are plotted as individual points and where the minimum and maximum summaries are replaced by the smallest and largest observations that are still within the interval (1.2).

Example: Tenderness of pork.

From the standard boxplots we see that the distribution of values for tunnel cooling is fairly symmetric whereas the distribution of the observations from rapid cooling is highly skewed. By placing boxplots from two samples next to each other we can also directly compare the two distributions: the tenderness values from tunnel cooling generally appear to be higher than the values from rapid cooling although there are a few very small values for tunnel cooling. We can also see from the boxplot that there is a single outlier for rapid cooling. It is worth checking the dataset to see if this is indeed a genuine observation.


Tenderness score
3
4
5
6
7
8
Tunnel
Rapid
Symmetric
Skewed to the right
Positively skewed

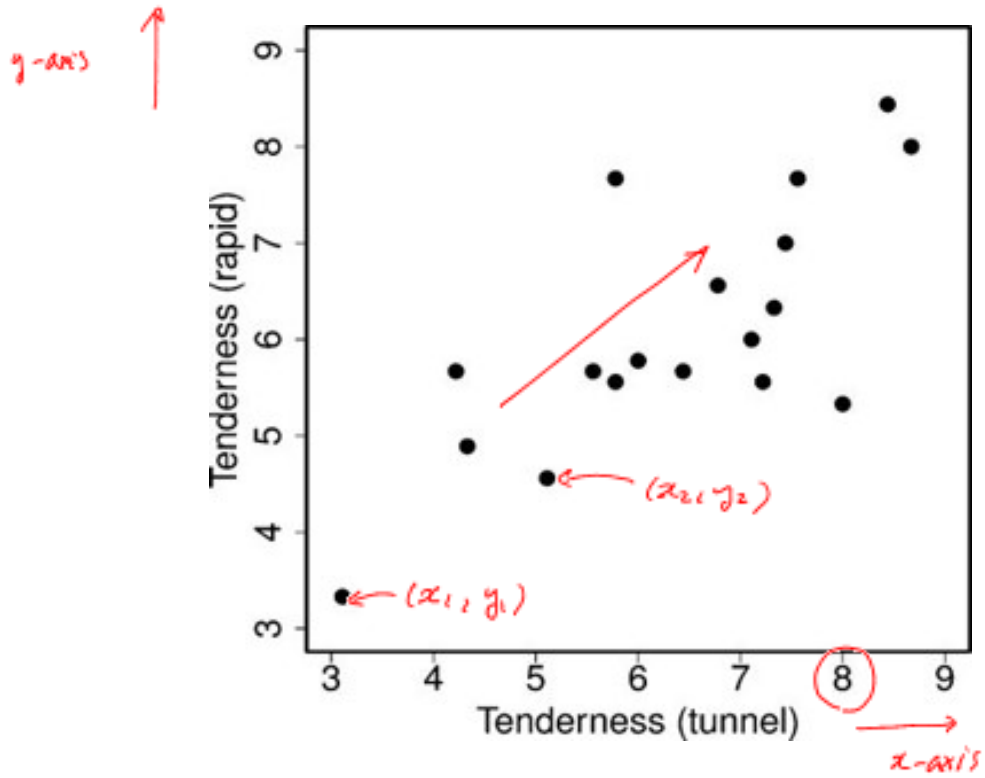
Scatter plot.



The relationship between two quantitative variables can be illustrated with a **scatter plot**, where the data points are plotted on a two-dimensional graph. Scatter plots provide information about the relationship between the variables, including the strength of the relationship and the direction (positive or negative), and make it easy to spot extreme observations. If one of the variables can be controlled by the experimenter then that variable might be considered an **explanatory variable** and is usually plotted on the x-axis, whereas the other variable is considered a **response variable** and is plotted on the y-axis. If neither the one or the other variable can be interpreted as an explanatory variable then either variable can be plotted on either axis and the scatter plot will illustrate only the relationship but not the causation between the two variables.

Example: Tenderness of pork.

There is some relationship of tenderness between the rapid and tunnel cooling methods for the combined data of low- and high-pH groups. Scatter plots are extremely useful as tools to identify relationships between two quantitative continuous variables.



Example: Tenderness of pork.

The following command will generate the boxplots seen in the earlier page.

```
attach(databox)
par(mfrow=c(1,1))
boxplot(tunnel, rapid, names=c("Tunnel","Rapid"),
ylab="Tenderness score")
```

The following command will generate the scatter plot in the previous page.

```
plot(tunnel, rapid, xlab="Tenderness (tunnel)",
ylab="Tenderness (rapid)")
detach(databox)
```

It produces a scatter plot, and we can add additional information to the plot by specifying the labels for the x-axis and the y-axis with the xlab and ylab. At the end you should detach "databox".

Example: Pulse rates of females.

The data file "FHEALTH.csv" is from U.S. Department of Health and Human Services, National Center for Health Statistics Third National Health and Nutrition Examination Survey. The csv file can be read into R using

```
databox <- read.csv(file.choose())
```

Before doing anything else, we have to declare the data frame "databox" in use by the following command:

```
attach(databox)
```

Now we can use the variable PULSE. To see what variables are available, use names function as follows:

```
names(databox)
```

Example: Pulse rates of females, continued.

To calculate summary statistics of the variable PULSE, use the summary command. In addition you want to obtain measures of variation, standard deviation and IQR. To find out the data size, use the length command.

```
summary(PULSE)
sd(PULSE)
IQR(PULSE)
length(PULSE)
```

We can observe that the sample mean is slightly higher than the median.

Visualizing data in R.

Histograms and relative frequency histograms are both produced with the hist() function. By default the hist() function automatically groups the quantitative data vector into bins of equal width and produces a frequency histogram. We can make a frequency plot or a relative frequency plot by specifying either "freq=T" or "freq=F" option, respectively. The number of bins is controlled by "breaks" option. If the "breaks" option is not entered, then R will try to determine a reasonable number of bins.

```
hist(PULSE)
hist(PULSE,breaks=10)
hist(PULSE,freq=F)
hist(PULSE, col="gray", main="Pulse Rates of Females")
```

We find that the shape of distribution is skewed to the right, and use it to explain the different values in the mean and the median.

Visualizing data in R, continued.

Horizontal and vertical boxplots are produced by the boxplot() function. By default, R creates the modified boxplot. The boxplot can be made horizontal by including "horizontal=T" option.

```
boxplot(PULSE)
boxplot(PULSE, col="green", horizontal=T, main="Pulse
Rates of Females")
```

In either of the boxplots we find two outliers identified, and further observe a longer tail on the right side, which reinforces the earlier observation of right-skewed shape of distribution. When you are done with the data frame "databox" you should detach it at the end.

```
detach(databox)
```

Practice problem 1.

Discuss your findings regarding "Material Strength" (Data file strength.csv).

1. Compute the mean and standard deviation for each supplier.
2. Construct boxplots for the data.
3. Which supplier appears to provide material that produces lenses having reliably low strength?

Practice problem 2.

Discuss your findings regarding "Tree Diameters" (Data file tree.csv).

1. Calculate the median and the IQR of the data.
2. Does the first quartile Q1 represent 25% of data exactly? How about the median or the third quartile? No!
3. Construct a relative frequency histogram and a boxplot.
4. Are there any outliers? No.
5. How can you characterize the shape of distribution?

Symmetric or skewed to the left.

Practice problem 3.

Discuss your findings regarding "Diesel Generator Demands" (Data file diesel.csv).

1. Calculate the mean and the median of the successful demands between failures.
2. Calculate the IQR and standard deviation.
3. Construct a relative frequency histogram and a boxplot for the data.
4. Describe the shape of distribution. Right-skewed
5. Are there any outliers? 3 outliers
6. Which measure of center appears to best represent the center of the data? Median is better.