Organization of Data

MATH MATH 1441, BCIT

Statistics for Food Technology

January 9, 2017

Introductory Concepts in Statistics I

- A population is the complete collection of all measurements or data that are being considered.
- A census is the collection of data from every member of the population.
- A sample is a subcollection of members selected from a population.

Introductory Concepts in Statistics I

- A population is the complete collection of all measurements or data that are being considered.
- A census is the collection of data from every member of the population.
- A sample is a subcollection of members selected from a population.

Introductory Concepts in Statistics I

- A population is the complete collection of all measurements or data that are being considered.
- A census is the collection of data from every member of the population.
- A sample is a subcollection of members selected from a population.

Introductory Concepts in Statistics II

- A voluntary response sample or self-selected sample is one in which the respondents themselves decide whether to be included.
- A random sample is one in which each member has the same probability of being selected.
- A stratified random sample is one in which random samples from subgroups are drawn and proportionally combined to form the complete sample.

Introductory Concepts in Statistics II

- A voluntary response sample or self-selected sample is one in which the respondents themselves decide whether to be included.
- A random sample is one in which each member has the same probability of being selected.
- A stratified random sample is one in which random samples from subgroups are drawn and proportionally combined to form the complete sample.

Introductory Concepts in Statistics II

- A voluntary response sample or self-selected sample is one in which the respondents themselves decide whether to be included.
- A random sample is one in which each member has the same probability of being selected.
- A stratified random sample is one in which random samples from subgroups are drawn and proportionally combined to form the complete sample.

Introductory Concepts in Statistics III

- A parameter is a numerical measurement describing some characteristic of the population.
- A statistic is a numerical measurement describing some characteristic of a sample.

Introductory Concepts in Statistics III

- A parameter is a numerical measurement describing some characteristic of the population.
- A statistic is a numerical measurement describing some characteristic of a sample.

Introductory Concepts in Statistics IV

- Quantitative (or numerical) data consist of numbers representing counts or measurements.
- Categorical (or qualitative) data consist of names or labels that are not numbers representing counts or measurements.

Introductory Concepts in Statistics IV

- Quantitative (or numerical) data consist of numbers representing counts or measurements.
- Categorical (or qualitative) data consist of names or labels that are not numbers representing counts or measurements.

Introductory Concepts in Statistics V

- Discrete data result when the data values are quantitative and the number of values is finite or countable.
- Continuous data result when the data values are quantitative and the number of values is infinite and not countable.

Here is an example for infinite discrete outcomes (this is rare). Roll a die until you roll a six. There are infinitely many ways to do this, but the data is not continuous.

Introductory Concepts in Statistics V

- Discrete data result when the data values are quantitative and the number of values is finite or countable.
- Continuous data result when the data values are quantitative and the number of values is infinite and not countable.

Here is an example for infinite discrete outcomes (this is rare). Roll a die until you roll a six. There are infinitely many ways to do this, but the data is not continuous.

Introductory Concepts in Statistics VI

- Blinding is when the subject doesn't know whether they are receiving a treatment or a placebo.
- The placebo effect occurs when an untreated subject reports an improvement in symptoms because of their participation in the study.
- An experiment is double-blind when it is blind and the experimenter also doesn't know whether they are applying a treatment or a placebo.

Introductory Concepts in Statistics VI

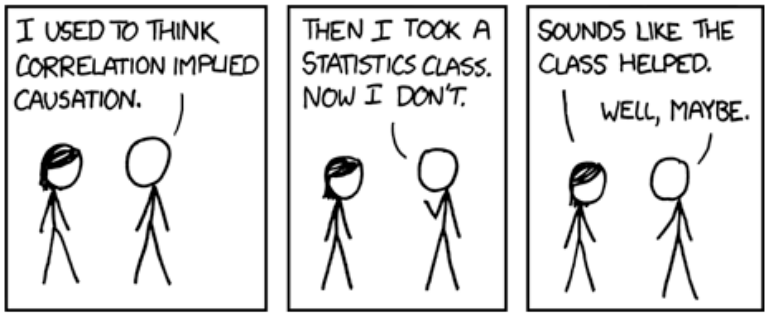
- Blinding is when the subject doesn't know whether they are receiving a treatment or a placebo.
- The placebo effect occurs when an untreated subject reports an improvement in symptoms because of their participation in the study.
- An experiment is double-blind when it is blind and the experimenter also doesn't know whether they are applying a treatment or a placebo.

Introductory Concepts in Statistics VI

- Blinding is when the subject doesn't know whether they are receiving a treatment or a placebo.
- The placebo effect occurs when an untreated subject reports an improvement in symptoms because of their participation in the study.
- An experiment is double-blind when it is blind and the experimenter also doesn't know whether they are applying a treatment or a placebo.

Correlation Does Not Imply Causation

Confounding occurs in an experiment when the investigators are not abloe to distinguish among the effects of different factors.

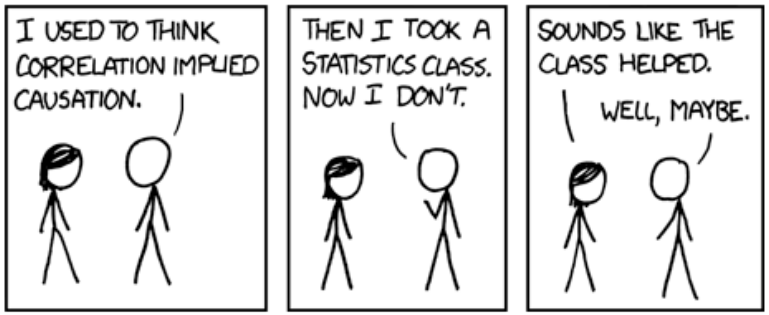


Examples:

1. astrological sign and IQ in elementary school
2. soft drinks and obesity
3. birth control pills and thrombosis

Correlation Does Not Imply Causation

Confounding occurs in an experiment when the investigators are not abloe to distinguish among the effects of different factors.

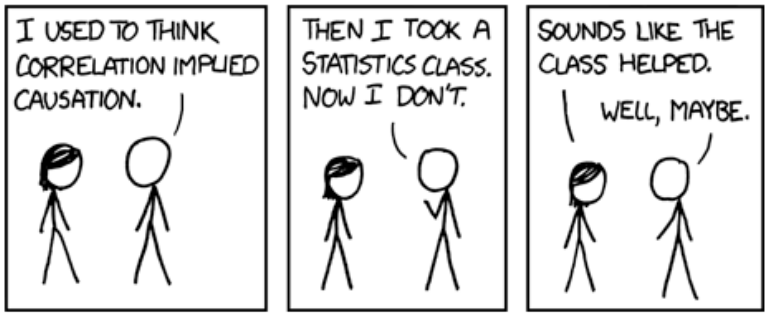


Examples:

1. astrological sign and IQ in elementary school
2. soft drinks and obesity
3. birth control pills and thrombosis

Correlation Does Not Imply Causation

Confounding occurs in an experiment when the investigators are not abloe to distinguish among the effects of different factors.



Examples:

1. astrological sign and IQ in elementary school
2. soft drinks and obesity
3. birth control pills and thrombosis

Modes of Data Presentation

The three modes of data presentation.

- textual
- tabular
- graphical

Textual

Example for textual data presentation.

Philippine Stock Market

The Philippine Stock Exchange composite index lost 7.19 points to 2,099.12 after trading between 2,095.30 and 2,108.47. Volume was 1.29 billion shares worth 903.15 million pesos (16.7 million dollars). The broader all share index gained 5.21 points to 1,221.34. (From: Freeman dated March 17, 2005)

Types of Graphs

Four types of graphs.

- Line plot
- Pie chart
- Bar plot
- Multi-bar graph
- Histogram

Line Plot

Usually suitable for data on a timeline. Here is an example. Rainy days in Vancouver.

	Q1	Q2	Q3	Q4
2012	55	43	13	65
2013	53	41	27	35
2014	45	38	18	54
2015	49	19	25	53
2016	61	28	27	69

Data for Line Plot Example

Here is the file `fs01.csv`. On a Windows computer, open the program called Notepad and paste the data into it. Then save as `fs01.csv`. You can then open this file in R Studio, for example (but also in other statistical software, such as minitab or excel).

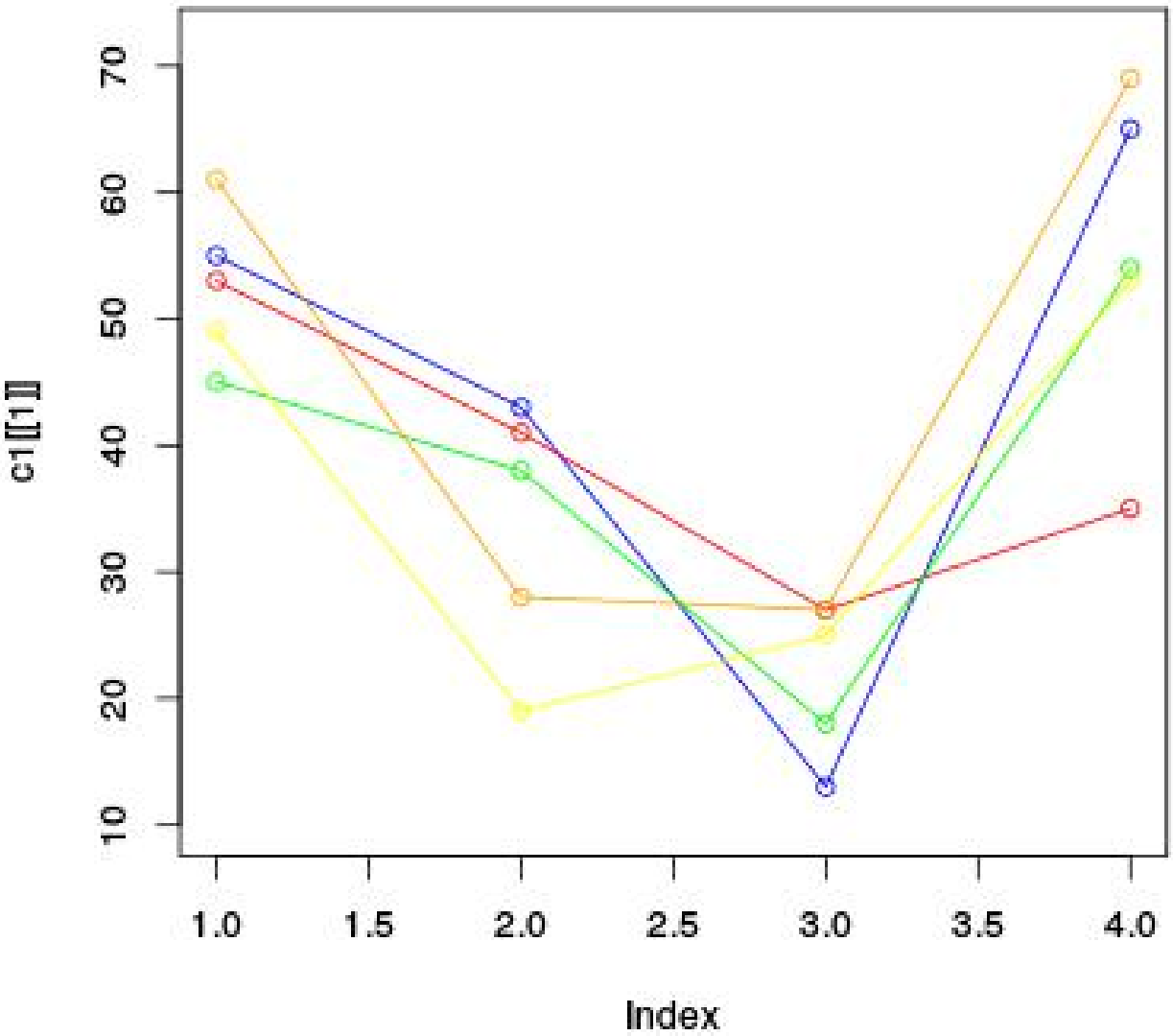
```
year,quarter,dor
2012,Q1,55
2012,Q2,43
2012,Q3,13
2012,Q4,65
2013,Q1,53
2013,Q2,41
2013,Q3,27
2013,Q4,35
2014,Q1,45
2014,Q2,38
2014,Q3,18
2014,Q4,54
2015,Q1,49
2015,Q2,19
2015,Q3,25
2015,Q4,53
2016,Q1,61
2016,Q2,28
2016,Q3,27
2016,Q4,69
```

Line Plot Example By Year

In R Statistics, use the file `fs01.csv` and the following code.

```
a<-read.table("fs01.csv",sep=",",header=TRUE)
c1<-subset(a,year=="2012",select=c(dor))
c2<-subset(a,year=="2013",select=c(dor))
c3<-subset(a,year=="2014",select=c(dor))
c4<-subset(a,year=="2015",select=c(dor))
c5<-subset(a,year=="2016",select=c(dor))
ylima<-min(a[[3]])-3
ylimb<-max(a[[3]])+3
plot(c1[[1]],type="o",ylim=c(ylima,ylimb),col="blue")
lines(c2[[1]],type="o",ylim=c(ylima,ylimb),col="red")
lines(c3[[1]],type="o",ylim=c(ylima,ylimb),col="green")
lines(c4[[1]],type="o",ylim=c(ylima,ylimb),col="yellow")
lines(c5[[1]],type="o",ylim=c(ylima,ylimb),col="orange")
```

Line Plot Example By Year



Pie Chart

Pie charts are often used for tabular representations of categorical data. Consider the following data set `fs02.csv` and the corresponding chart on the next slide.

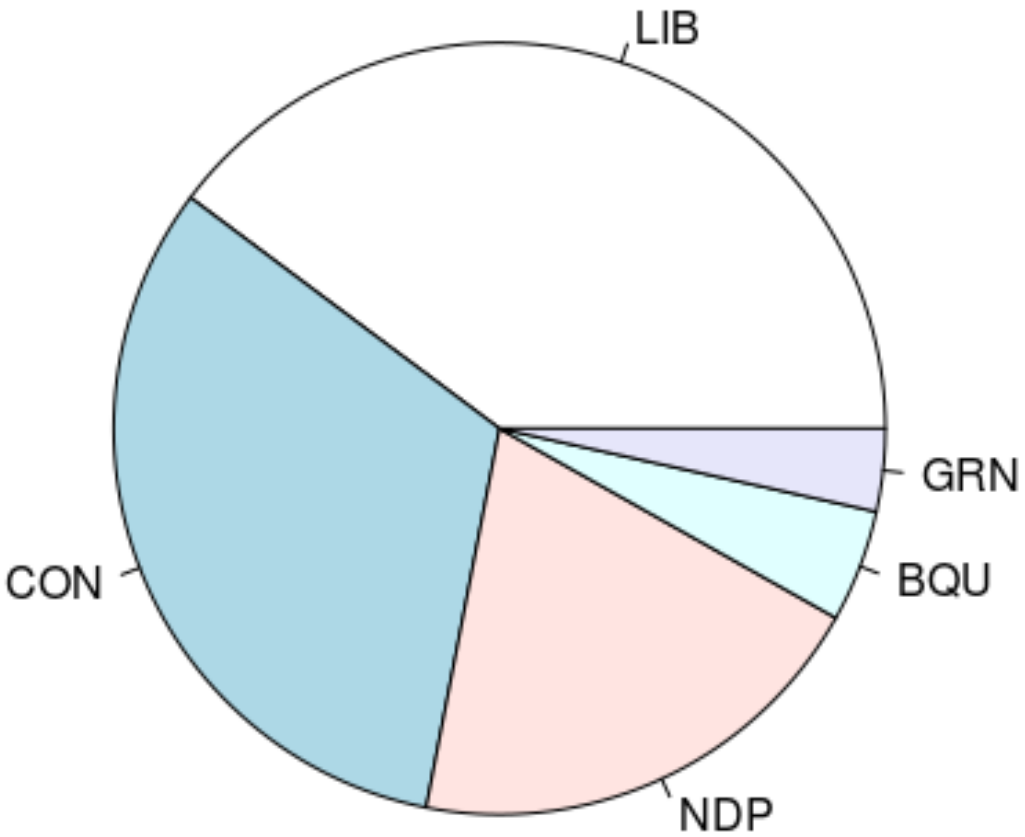
```
party,votes
LIB,6943276
CON,5613614
NDP,3470350
BQU,821144
GRN,602944
```

Pie Chart Graph

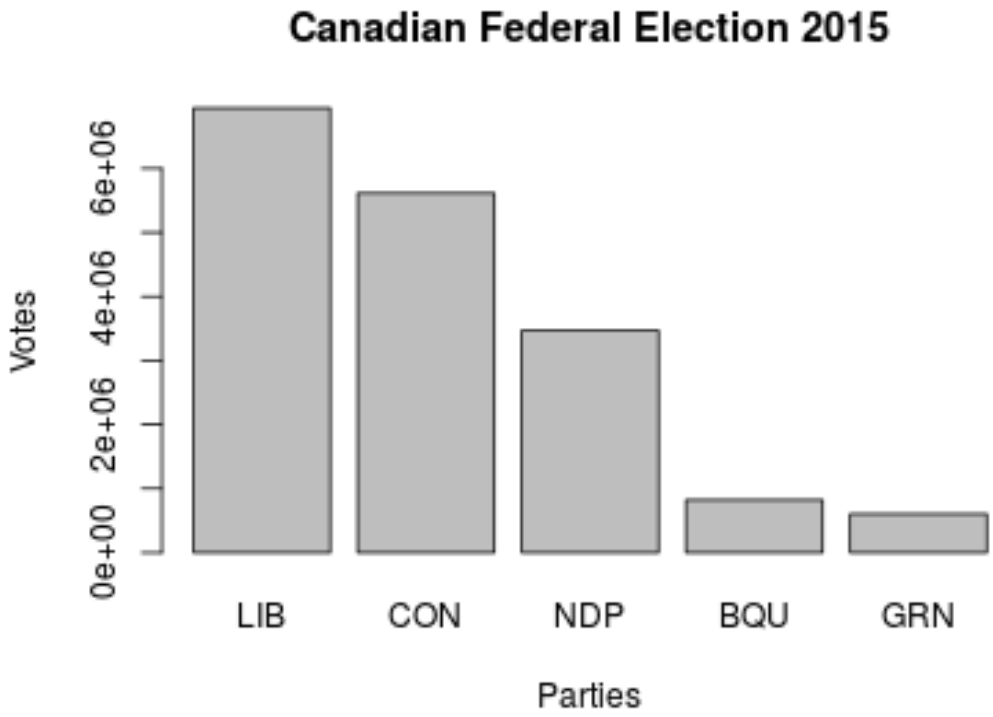
We are using the following R commands:

```
d<-read.table("fs02.csv",sep=",",header=TRUE)
pie(d[[2]],label=d[[1]])
```

However, pie charts are not recommended for visualizing statistical data. Bar plots make differences between data points more clear.



Bar Plot Graph



We have used the following R command:

```
barplot(d[[2]], main="Canadian Federal Election 2015",
xlab="Parties",ylab="Votes",names.arg=d[[1]])
```

Multi-Bar Graph

Use the following R commands for a multi-bar graph of tabular data:

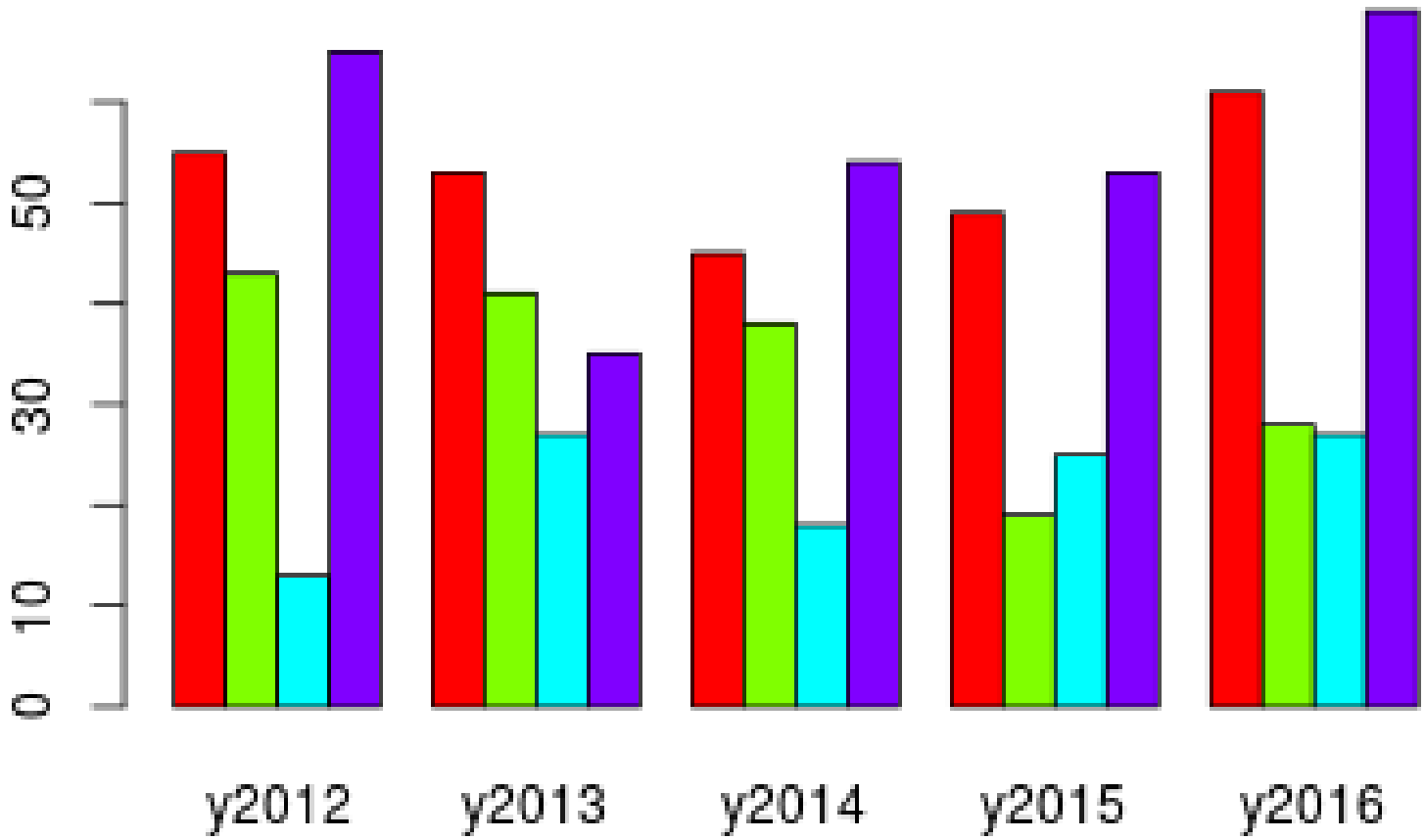
```
e<-read.table("fs03.csv",sep=",",header=TRUE)
barplot(as.matrix(e),beside=TRUE,col=rainbow(4))
```

Here is how to organize the data in the file `fs03.csv` for these commands:

```
y2012,y2013,y2014,y2015,y2016
55,53,45,49,61
43,41,38,19,28
13,27,18,25,27
65,35,54,53,69
```

These are the rainy days in Vancouver as in `fs01.csv`, but organized in tabular form.

Multi-Bar Graph



Histogram

We use histograms for continuous data, bar plots for discrete data. Use the R command

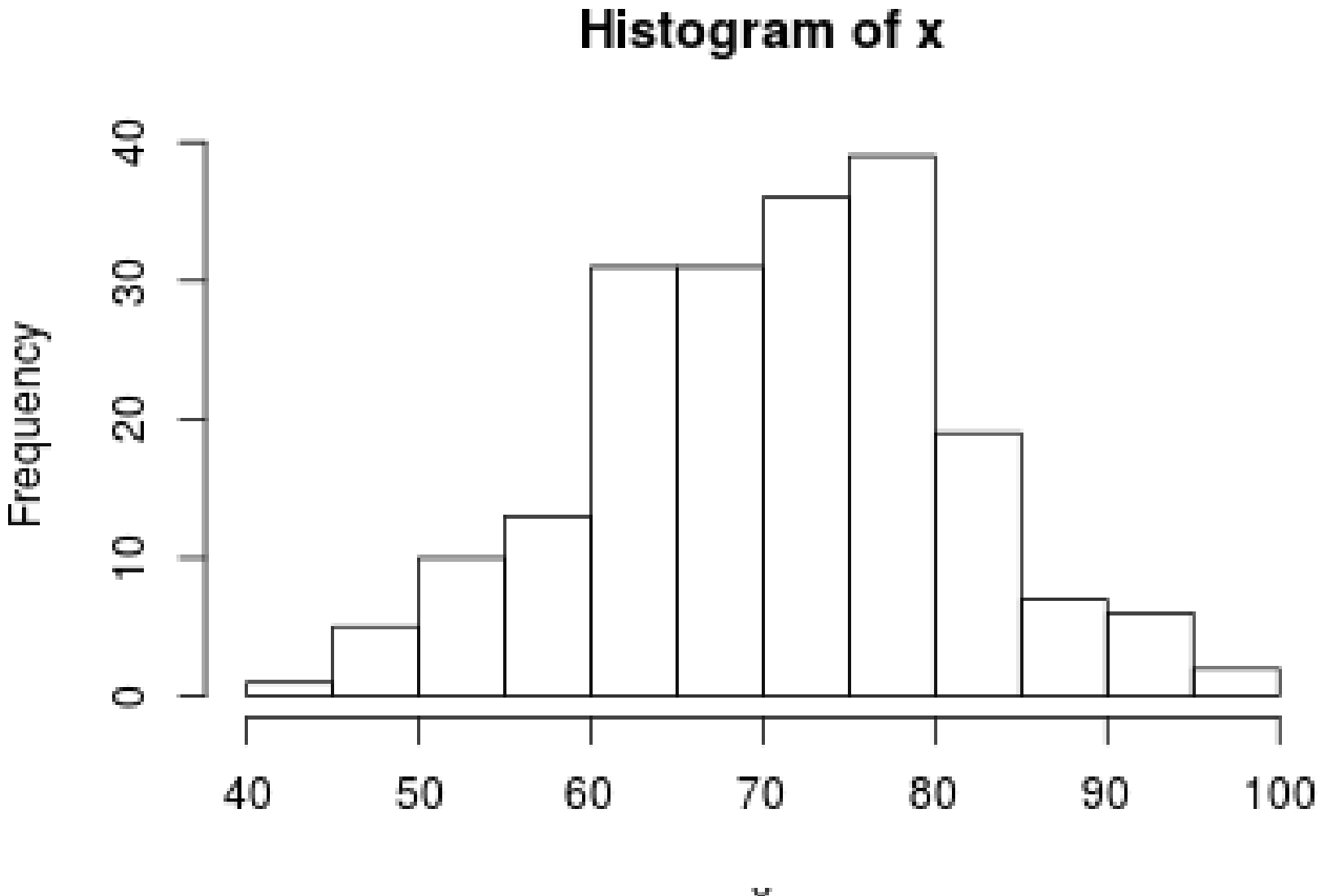
```
x<-round(rnorm(200,71,11)*100)/100
```

for the grades of 200 students in a course. Then use

```
hist(x)
```

for the histogram. There are different suggestions for what the ideal length of intervals is for a histogram. $\sqrt{n}$ is one useful rule of thumb, where n is the sample size (number of data points in the data set).

Histogram



Measures of Centre: Mean

$$\text{mean} = \frac{\sum x}{n} \tag{1}$$

where n is the number of data points in your quantitative data set and x is your random variable. Often, we write mu for the mean of a population and $\bar{x}$ for the mean of a sample.

Example: Find the mean of the following five counts for Chips Ahoy chocolate chip cookies: 22 chips, 22, chips, 26 chips, 24 chips, and 23 chips.

Frequency Distributions

Often, data is provided in the form of a frequency distribution. For example, when I asked a class of statistics students about the number of countries they had visited in their lifetime, the response was as follows (given as an R command),

```
cn<-c(5,4,7,3,6,4,3,4,2,4,4,2,4,3,2,4,4)
```

A more intelligible way to display the data is to provide a frequency distribution.

```
> table(cn)
cn
2 3 4 5 6 7
3 3 8 1 1 1
```

There are 3 people who have been to 2 countries, 3 people who have been to 3 countries, 8 people who have been to 4 countries, 1 person who has been to 5 countries, and so on.

Calculating the Mean from a Frequency Distribution

If you have a frequency distribution (usually of a sample, so we will call the mean $\bar{x}$), the mean is

$$\bar{x} = \frac{\sum (f \cdot x)}{\sum f} \tag{2}$$

For the example in the last slide,

$$\bar{x} = \frac{2 \cdot 3 + 3 \cdot 3 + 4 \cdot 8 + 5 \cdot 1 + 6 \cdot 1 + 7 \cdot 1}{3 + 3 + 8 + 1 + 1 + 1} \approx 3.82 \tag{3}$$

In R, you can also simply use the command `mean`. Notice that `mean(x)` and `sum(x)/length(x)` will give you the same number.

Measures of Centre: Median

The median is the value in the middle. If there is an even number of data points, the median is the mean of the two data points in the middle. To find the median, sort the data points. For example, the numbers of countries visited are

2	2	2	3	3	3	4	4	4	4	4	4	4	4	5	6	7

The value in the middle is the number 4, which is also the median of the data. It is quite similar to the mean, which is approximately 3.82.

Difference Between Mean and Median

Imagine we had one more student in the class who was a world traveler. She had visited 112 countries! The mean is now

$$\bar{x} = \frac{\sum x}{n} = \frac{177}{18} \approx 9.83 \tag{4}$$

A mean of 9.83 is no longer a good summary of the data. Let's see if the median does better.

2	2	...	4	4	4	4	4	...	6	7	112

The new median is $(4 + 4)/2 = 4$, which is a much better summary of the data, pretty much ignoring the outlier.

Measures of Centre: Mode

The mode of a data set is the value that occurs with the greatest frequency. In the numbers of countries visited example the mode is clearly 4. To be precise, the mode of a data set is itself a set of numbers. A data set can have one mode, as in our example, but if more values are repeated the same number and a maximum number of times, they are all modes. If no data point is repeated, the data set has no mode.

Measures of Centre: Midrange

The midrange is the midpoint between the maximum and the minimum data points. It is very sensitive to outliers! For example,

$$\text{midrange} = \frac{7+2}{2} = 4.5 \tag{5}$$

without the outlier, and

$$\text{midrange} = \frac{112+2}{2} = 57 \tag{6}$$

with the outlier in the numbers of countries visited example.

Measures of Dispersion: Motivation

Have a look at these two different data sets.

```
x1<-c(12,12,12,12,12,12,11,12,12,13,12,12,12,12)
```

and

```
x2<-c(15,10,14,7,17,15,11,18,12,12,15,9,7,6)
```

The mean of both data sets is 12. The median of both data sets is 12. However, something about these two data sets is different. `x2` is more dispersed than `x1`, which means that the data is spread out more. There is more variation in `x2`. We try to capture this variation by finding measures of dispersion.

Measures of Dispersion: Range

One very simple measure of dispersion is the range, which is just the lowest value subtracted from the highest value.

$$\text{range of } \texttt{x1} = 13 - 11 = 2 \tag{7}$$

$$\text{range of } \texttt{x2} = 18 - 6 = 12 \tag{8}$$

The problem is outliers: they would change the range significantly while many other data points would be ignored. Another possibility is to count up the difference between data points and the mean. Can you guess what the problem of this measure would be?

Measures of Dispersion: Absolute Value of Deviation

The difference between data points and the mean always sums to zero! That is not helpful. If we want to make this measure of dispersion more useful, we need to sum the absolute value of deviation

$$\text{mean absolute deviation} = \frac{\sum |x - \bar{x}|}{n} \tag{9}$$

For our examples `x1` and `x2`, the mean absolute deviations are 2 and 44, respectively. Although at first glance this measure of dispersion looks useful, it makes for very complicated calculations that can be simplified by choosing a different way to make all the distances between data points and mean positive: not the absolute value, but the square of the distance.

Measures of Dispersion: Variance

The variance is calculated as follows,

$$\text{variance of a population } = \sigma^2 = \frac{\sum (x - \mu)^2}{n} \tag{10}$$

Something odd happens when we take the variance of a sample. If we were to use equation (10) to calculate the sample variance for all possible samples of a population, the mean of these sample variance would not equal the population variance. This means that in this case the sample variance would be a biased estimator of the population variance. We don't want that! To correct for this problem and define a sample variance which is an unbiased estimator of the population variance, we introduce Bessel's correction and define

$$\text{variance of a sample } = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \tag{11}$$

Measures of Dispersion: Standard Deviation

One disadvantage of the variance is that it is not an intuitive measurement of dispersion. If we take the square root of the variance, then we get something similar to the absolute value of deviation, which tells us approximately how far on average the data points are from the mean. We call this measurement the standard deviation

$$\text{standard deviation of a population } = \sigma = \sqrt{\frac{\sum (x-\mu)^2}{n}} \quad (12)$$

$$\text{standard deviation of a sample } = s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} \quad (13)$$

Why we still sometimes prefer the variance will become clear on the next slide. The standard deviation, whether with or without Bessel's Correction, is a biased estimator!

Bessel's Correction

	$\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$	$\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$	$\frac{\sum(x-\bar{x})^2}{n}$	$\frac{\sum(x-\bar{x})^2}{n-1}$
2 and 2	0.00	0.00	0.00	0.00
2 and 3	0.50	0.71	0.25	0.50
2 and 8	3.00	4.24	9.00	18.00
3 and 2	0.50	0.71	0.25	0.50
3 and 3	0.00	0.00	0.00	0.00
3 and 8	2.50	3.54	6.25	12.50
8 and 2	3.00	4.24	9.00	18.00
8 and 3	2.50	3.54	6.25	12.50
8 and 8	0.00	0.00	0.00	0.00
mean	1.33	1.89	3.44	6.89

The population standard deviation is:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}} = 2.62$$

The population variance is:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{n} = 6.89$$

Calculating the Variance I

It is easiest to calculate the variance using statistical software. In R Studio, for example,

```
> var(x1)
[1] 0.1538462
> var(x2)
[1] 14.76923
```

and

```
> sd(x1)
[1] 0.3922323
> sd(x2)
[1] 3.843076
```

Calculating the Variance II

When you do have to calculate the variance by hand, it is helpful to use the following shortcut formula,

$$s^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} \tag{14}$$

because you do not have to keep entering the mean, which may contain numerous significant digits.

Variance for a Frequency Distribution I

Consider the data set `x3`

```
4,4,2,2,4,3,3,3,1,4,4,1,2,2,2,4,2,1,2,1,1,2,3,3,2,3,3,3
```

We can summarize the data in a frequency distribution

```
> table(x3)
x3
1 2 3 4
5 9 8 6
```

Variance for a Frequency Distribution II

Remember that in equation (2) we calcluated the mean using a formula for the frequency distribution,

$$\bar{x} = \frac{\sum (f \cdot x)}{\sum f} \tag{15}$$

We can do the same for the variance and the standard deviation,

$$s^2 = \frac{\sum \left(f \cdot x^2\right) - \frac{\left(\sum f \cdot x\right)^2}{\sum f}}{\sum f - 1} \tag{16}$$

Remember that $n = \sum f$.

Measures of Position: Quartiles I

Sometimes we want to know where a data point approximately ranks in relation to other data points. For a more finely-grained measure of position, we use percentiles. For a more coarsely-grained measure of position, we use quartiles. Let's use again the R command

```
x<-round(rnorm(200,71,11)*100)/100
```

for the grades of 200 students in a course. On the next slide, you can see the random numbers generated when I just now ran this command in R Studio.

Measures of Position: Quartiles II

```
[1] 58.52 74.38 80.79 84.24 86.61 70.92 86.98 76.02 70.67 66.10
[11] 62.01 63.41 76.33 68.60 73.50 64.15 85.85 86.22 76.80 71.12
[21] 78.29 52.77 81.25 57.98 66.09 92.43 71.19 65.26 96.96 55.92
[31] 71.87 70.31 66.84 69.42 67.90 66.46 69.87 72.35 76.83 58.30
[41] 61.90 57.93 74.90 97.23 87.41 74.86 77.69 63.41 53.55 78.95
[51] 78.76 71.04 68.63 70.10 77.72 94.69 64.18 76.67 70.97 83.96
[61] 70.93 75.89 65.19 60.34 64.89 81.38 65.59 72.89 74.22 64.68
[71] 54.10 84.13 79.10 59.91 74.13 60.49 72.70 68.50 87.30 75.63
[81] 83.24 71.80 75.54 64.11 77.46 82.05 74.20 72.45 75.03 53.60
[91] 54.20 65.16 81.77 63.27 57.38 83.93 72.36 63.62 73.02 72.18
[101] 54.66 84.89 58.05 70.27 80.31 76.43 70.66 71.31 86.39 77.85
[111] 73.52 68.07 44.34 62.52 81.15 70.20 76.16 86.35 64.60 85.13
[121] 61.21 65.25 72.94 61.48 90.48 80.50 108.81 57.91 73.53 65.53
[131] 58.08 78.47 75.61 51.90 76.72 70.57 65.18 90.92 86.01 68.36
[141] 78.16 54.97 81.10 75.30 52.39 68.64 82.96 71.82 80.44 59.15
[151] 100.15 54.56 52.91 67.48 75.07 61.07 71.14 58.55 84.35 67.56
[161] 94.91 78.32 70.50 75.73 67.25 71.49 62.55 68.54 59.55 63.01
[171] 65.63 83.72 70.64 82.58 71.13 69.20 77.55 74.76 72.95 61.53
[181] 73.04 84.79 64.35 85.49 78.86 56.27 74.11 97.87 72.58 92.96
[191] 72.99 66.93 78.41 69.93 67.88 80.88 70.84 69.55 74.69 89.32
```

Measures of Position: Quartiles III

Let's say you are student number 111, and your score is 73.52%. The students are divided up into four groups of approximately the same size. The first quartile Q_1 is the score which divides the first group (with the lowest scores) from the second group. The second quartile Q_2 is the median and divides the second group from the third group. The third quartile Q_3 is the score which divides the third group from the fourth group (with the highest scores). To calculate the quartiles you have to rank the data and find the corresponding scores, just as you did with the median. Or you use statistics software to check out the summary.

```
> summary(x)
Min. 1st Qu. Median Mean 3rd Qu. Max.
44.34 65.19 71.84 72.34 78.42 108.80
```

Measures of Position: Percentiles

Percentiles work just like quartiles, using the number 100 instead of the number 4. Student number 111 turned out to be in the third group because her score was better than the median but worse than the third quartile. If we sort the data with the R Studio command

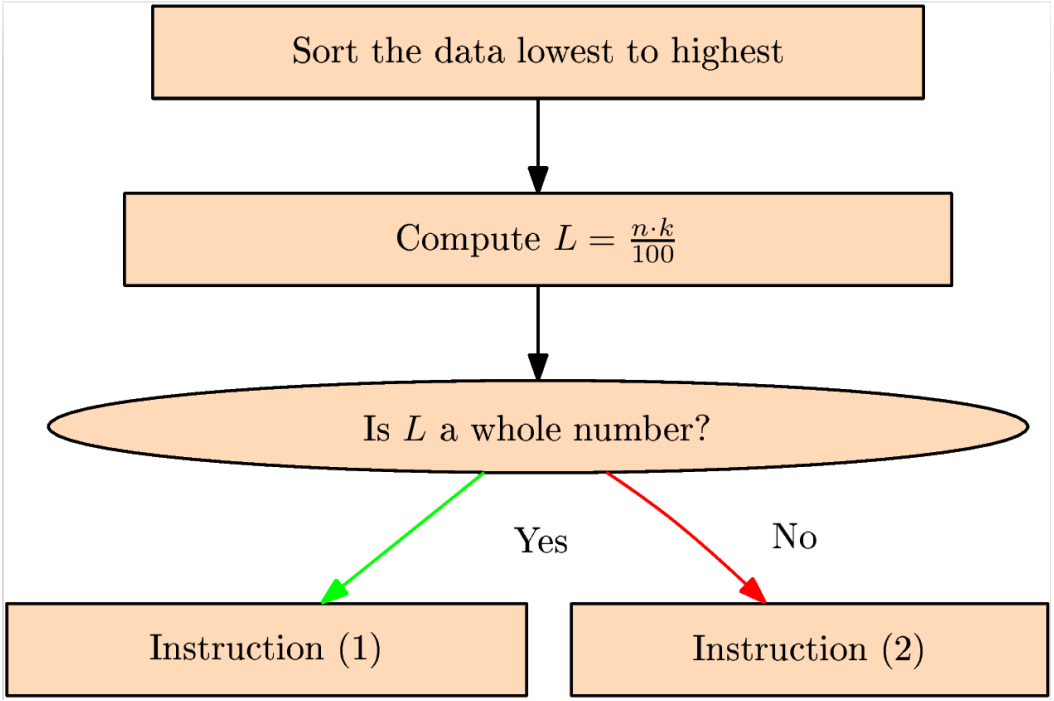
```
sort(x)
```

we discover that student number 111 is in 115th position (counting from the bottom), which is the 58th percentile. To find the k-th percentile use

$$\frac{n \cdot k}{100} \tag{17}$$

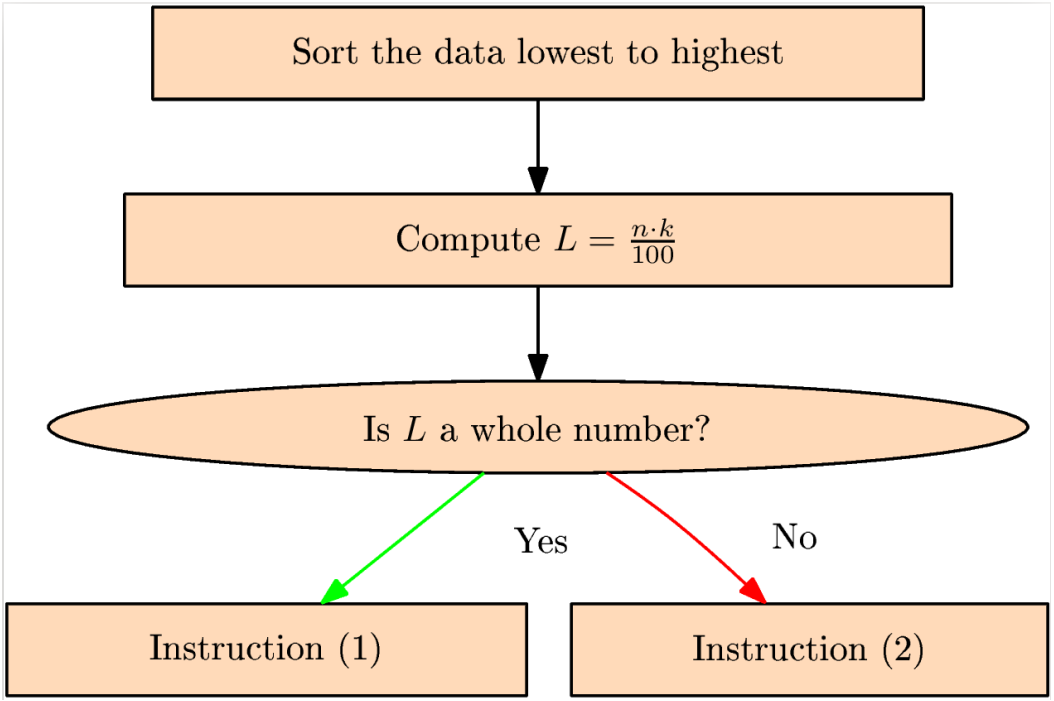
For example, the 90th percentile P_{90} is the mean between the 180th and the 181st data point ranked from the bottom (in our case, 85.93%). As a reference point, $P_{50} = Q_2 =$mean.

Percentiles Flow Chart I



Instruction (1): The value of the k-th percentile P_k is midway between the L-th value and the next value in the sorted set of data. Find P_k by adding the L-th value and the next value and dividing the total by 2.

Percentiles Flow Chart II



Instruction (2): Change L by rounding it up to the next larger whole number. The value of P_k is the L-th value, counting from the lowest.

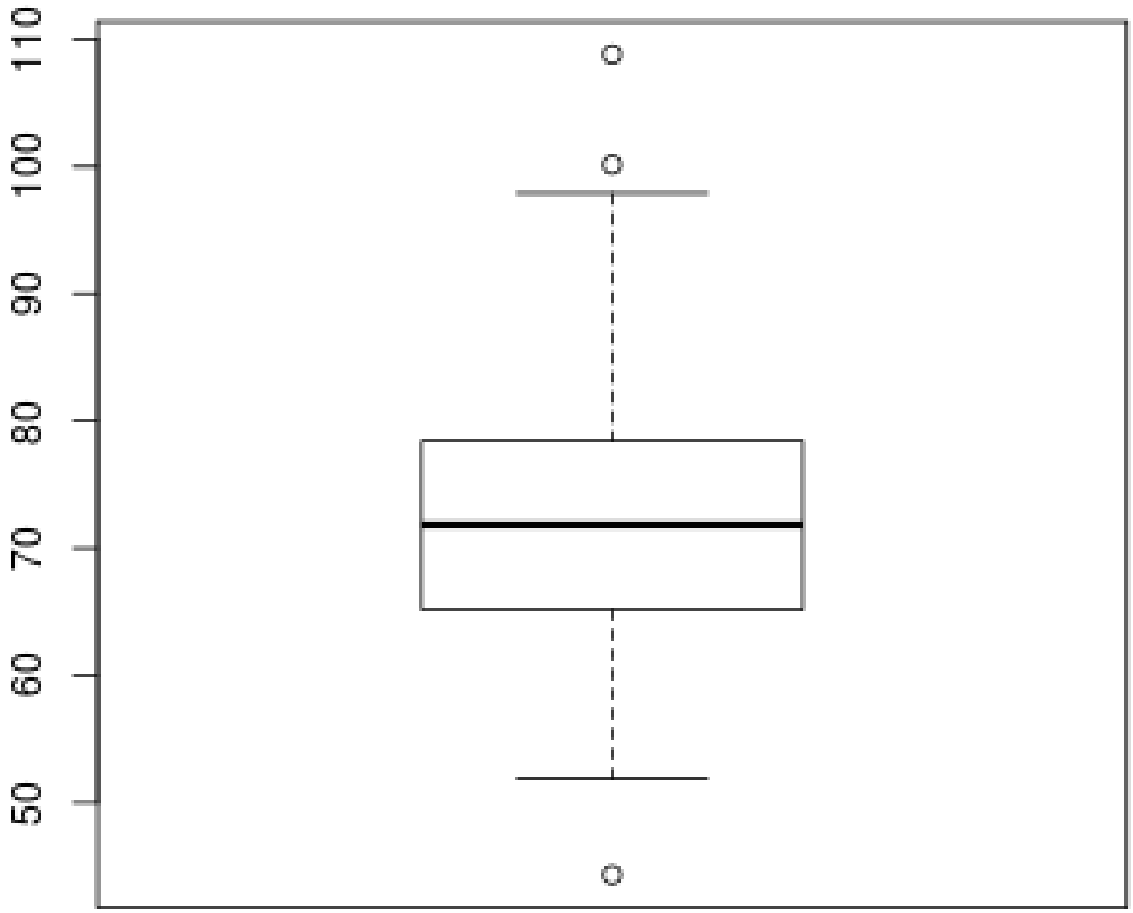
Measures of Dispersion: Midquartile

The interquartile range is a measure of dispersion which is not as vulnerable to outliers as the range. The interquartile range is visually best represented in a box-and-whiskers display. For the 200 students, the box-and-whiskers display is generated by the following R Studio command,

```
boxplot(x,range=0)
```

`range=0` means that the plot will not pay attention to outliers. The default is `range=1.5`, which will show some outliers.

Box-and-Whiskers Display with Outliers



End of Lesson

Next Lesson: Elementary Probability.