STAT*2060DE
Summary Notes for Unit 1: Descriptive Statistics

For each section of the course I'll put up summary notes of the required material. At the end of these notes there will be suggested textbook readings, suggested textbook questions, and suggested course manual questions.

Start each section by reading the notes I put up. Then move on to the suggested questions and the textbook readings.

An Example of a Type of Problem we will Encounter

A consumer suspects that a beverage company is underfilling pop bottles. To investigate this, they buy 10 two-litre bottles, and find the bottles contain an average of 1.97 litres of pop. Is the company not properly filling the bottles?

This can't be answered with only the given information, but we'll learn how to make a start on it later in the course.

Some definitions:

Individuals or **units** or **cases** are the objects on which a measurement is taken.
(The pop bottles)

A **population** is the set of individuals or units of interest to an investigator.
(All the two-litre pop bottles of this type.) Note: We sometimes use the term *population* to represent the objects themselves (the pop bottles) and sometimes use the term to represent the measurements on the objects (the amount of pop in each bottle of this type). But this subtlety is not likely to cause too much confusion.

A **parameter** is a numerical characteristic of a population.
(For example, the average amount of pop in all two-litre bottles of this type). In the vast majority of practical situations, we will not know the exact value of a parameter. What is the average weight of adult Canadian males? What is the average number of songs on iPods owned by Ontario students in Grade 9? It is, effectively, impossible to know the answers to these two questions. There is some true value of the parameter of interest, but we cannot know it for these problems. We will regard population parameters are fixed, usually unknown quantities, and we will estimate them using *sample statistics*.

A **sample** is a subset of individuals or units selected from the population.
(The ten bottles that were selected.) We will learn about different types of sampling. Overall, we hope to obtain a sample that is *representative* of the population.

A **statistic** is a numerical characteristic of the sample.
The sample mean of the 10 bottles (1.97 litres) is a statistic. For any given sample, we will be able to calculate the value of the statistic or statistics of interest.

In **statistical inference** we attempt to make statements about a population based on sample data. For example, is there strong evidence that the company is underfilling the bottles? In other words, we try to estimate population parameters using sample statistics.

The use and understanding of some statistical inference techniques is one of the fundamental aspects of the course. But before we get there, we will have to learn to speak the same language. We will begin with a discussion of descriptive statistics, move on to the concepts of probability, probability distributions, sampling distributions, and eventually get back to learning some inference techniques. Let's start with a discussion of descriptive statistics.

The variable of interest above is **quantitative** (there is a natural numerical measurement scale). For quantitative variables, numerical summaries like means and medians are meaningful. For example, the question "what is the average age of University of Guelph students?" makes sense, and would have a meaningful answer. Age is a quantitative variable.

There are also **qualitative** or **categorical** variables. These are variables that can be classified into groups, but there is no natural numerical measurement scale. e.g. Brand of cola, program a student is in, make of car a student owns. For qualitative variables, numerical summaries like means and medians do not have any real meaning. For example, the question "what is the average program a University of Guelph student is in?" does not make any sense.

Methods of obtaining data:

1. Published source.
2. Designed Experiment.
3. Observational Studies
4. Surveys

The most interesting statistical problems involve investigating relationships between variables.

Example #1. Fifty people sign up for a new LSAT prep course. 25 are randomly assigned to an online version of the course, and 25 are assigned to the regular classroom version. The 50 students then write the LSAT exam, and their scores are compared.

This is an **experiment** because the researchers applied a condition (the different courses) to the experimental units.

The **experimental units** are the individuals on which the experiment is performed (could be stores, people, types of machinery, etc.)

A **treatment** is a condition applied to the experimental units (drug, marketing strategy, price reduction, etc.)

The **response variable** is the LSAT score.
(a response variable is the outcome of the experiment or observational study)

The **explanatory variable** is the type of prep course (online vs live).

(an explanatory variable explains or *possibly* causes changes in the response variable).

Randomized experiments allow us to make the strongest conclusions about the effect of the explanatory variable. If we find strong evidence of a difference between the groups in a well-designed experiment, then we have strong evidence that the difference was *caused* by the different treatments. Our conclusions in randomized experiments can be stronger than those from observational studies. But it is not always possible to conduct experiments. For example, if we wished to assess the effect of cocaine use by pregnant women on the birth weight of the baby, it would impossible to carry out an ethical experiment. We could not ethically assign some pregnant women to a cocaine-use group, and another to a no cocaine group. So experiments are sometimes impossible to carry out in a given scenario, or may be overly costly to be of practical use.

Example #2. A certain stats professor takes attendance at every class. At the end of a stats course, the professor finds that students who attended class more frequently tended to get a higher mark on the final exam.

The response variable is mark on the final exam. The explanatory variable is class attendance.

This is an **observational study**, since the professor observed and measured, but did not actively impose any condition (did not force certain students to come to class).

Since this is an observational study, and not an experiment, we cannot state any conclusions of a causal nature. Even though a relationship was found between class attendance and grade on the final, it is entirely conceivable that this was simply the result of the fact that good students will tend to come to class more often, and good students will tend to do well on the final exam. The "quality of student" is a lurking variable.

Example #3. A research polling group telephoned 1000 Canadians and asked if they supported the Federal government supplying funds for the recent auto manufacturer bailout. This is a survey (a type of observational study).

Surveys can suffer from some potentially severe problems. Often, some people are systematically left out of the survey. For example, people who do not have a landline telephone are often left out of telephone surveys. It is likely that the opinions of those who do not own a landline telephone differ from those who do, in which case the results of the survey may not be representative of the entire population of interest. This is sometimes called **selection bias**. Another problem is **nonresponse**, in which some people who are selected may refuse to participate. For example, people often hang up when confronted with a telephone survey. This often biases the results, but the extent and direction of the bias can be difficult to determine. For example, do Republicans refuse to participate more often or less often than Democrats? This is a tough question to answer. Statisticians *sometimes* have ways for adjusting for certain types of bias, but it can be tricky business. There are a host of other possible issues, including the facts that people sometimes lie, and sometimes say things that are different from the reality of the situation. We should think of these issues when confronted with the results of any survey. Surveys can be useful, but we should interpret the results carefully.

Types of Sampling

There are different ways of going about getting a sample from some larger population. We could just haphazardly pick individuals from the population, but this would likely leave us with a very biased sample, and is not often a good idea. The main idea is that we are attempting to obtain a sample that is *representative* of the population. We want to minimize any possible biases. We frequently do this by properly incorporating randomization in our sampling design.

For this course, the main type of sampling we will be using is **Simple Random Sampling**. In simple random sampling, every possible sample of size n has the same chance of being selected in the sample.

For example, suppose I wish to know a little bit more about the 1000 students in one of my courses. I may have access to the information on all the students. For example, suppose I wish to know the average semester level of all 1000 students. This is information I have access to, and I could get the average semester level of all students quite easily. But suppose I wish to know what proportion of students have taken data management in high school. This is not information I have access to. I may wish to conduct a survey to get some idea of what proportion of students have taken data management. If I do this survey by email, I may not wish to contact everyone, as it would be too time consuming. So I may wish to only contact 50 students or so (we'll talk more later about why we might choose a certain sample size). For me to obtain a Simple Random Sample, I would have to randomly sample the 50 students from the 1000 students, such that every possible group of 50 students has the same chance of being picked. I would typically carry out the sampling using a computer. We'll talk more about this later.

In SRS, each member of the population has the same chance of being selected in the sample. An individual's chance of being selected in a SRS is $\frac{n}{N}$ (The sample size divided by the population size). In the above example, since I'm picking 50 students from 1000, each student would have a 50/1000 = .05 chance of being picked in the sample.

For most of our statistical methods later in the course, we will assume we have a simple random sample from the population of interest.

There are many other types of sampling. Another common method is a **stratified random sample**. In this type of sampling, the population is broken up into different strata (groups), and a simple random sample is conducted within each stratum. This ensures that the different strata are properly represented in the sample. For example, suppose we wish to conduct to a sample investigating adult Canadians' opinions on the current Prime Minister. We could carry out a simple random sample of Canadian households (or at least, conduct a phone survey on a simple random sample from households on a telephone listing). This wouldn't be a terrible way of going about it, but it does leave open the possibility that some areas are underrepresented in the sample. If we divide the country into the different provinces (strata), then conduct a simple random sample within each province, we could ensure that each province is adequately represented in the sample.

In **Cluster Sampling**, the population is grouped into clusters, and then clusters are randomly sampled. For example, we may divide a city into blocks, randomly select several blocks (clusters), then sample every household in each selected block (cluster).

Or we may wish to conduct a sample of every 10^{th} person that walks in the door of a business. This **1-in-*k*** type of sampling can be useful and easy to carry out. But there can be problems if there is some kind of cyclical or seasonal effect (the sample may be always hitting the high or low point of a cycle).

In **Voluntary Response Samples** individuals choose to participate by responding to a general appeal. For example, a radio talk show may ask listeners to call in and give their opinion on Barack Obama. A web site may have a survey on opinions about OHIP. Course evaluations are also a type of voluntary response sample. Only those choosing to participate end up becoming part of the sample. Those people with extreme views ("love the prof", "hate the prof") are more likely to participate than those who have views more toward the middle of the spectrum ("this prof did an okay job"). One must be very wary of conclusions drawn from voluntary response sample (although the biases are much more extreme for the radio talk show than for course evaluations.).

Example.
A electronics-store manager wants to estimate the average amount customers will spend at their store on boxing day. They find that the mean amount spent by the first 10 customers of the day is $1100.

This sample of 10 individuals is most likely severely biased. It is not a simple random sample from the population. This is NOT a simple random sample, as the manager simply looked at the first 10 customers. The chances are very good that the first customers were first in line because they intended to buy big ticket items, which can be greatly reduced in price on boxing day. Or possibly my thinking is flawed and the bias is in the other direction. I simply do not know for certain. When we do not randomize, we will not know the extent of the bias in our sample.

Eventually we want to learn some theory and methods of statistical inference. These methods will rely strongly on obtaining a good sample from the population. But we've got a long way to go before we reach statistical inference. First, we will go over some of the basics of descriptive statistics.

Descriptive Statistics

Descriptive statistics involves using plots and numerical summaries to illustrate a data set. If we had the data for an entire population at our disposal, we would use only descriptive statistics and would not have to use our inference techniques. The statistics world would be pretty easy, but I'd likely be out of a job.

Plots for Qualitative Variables

If we have qualitative (categorical) data, then we want to display the proportion of the observations that are within each category. We do this using:

- Bar Graphs
- Pareto diagrams
- Pie Charts

Example from the text.
Inspectors at an automobile assembly line found that 70 of the new cars produced on a given day had the following types of defect:

Type of Defect	Frequency	Relative Frequency	% Relative Frequency
Chrome	2	2/70 = .029	2.9%
Dents	25	25/70 = .357	35.7%
Paint	30	30/70 = .429	42.9%
Upholstery	10	10/70 = .143	14.3%
Windshield	3	3/70 = .043	4.3%

The **frequency** is the number of observations in a category.

The **relative frequency** is the proportion of observations in a category.

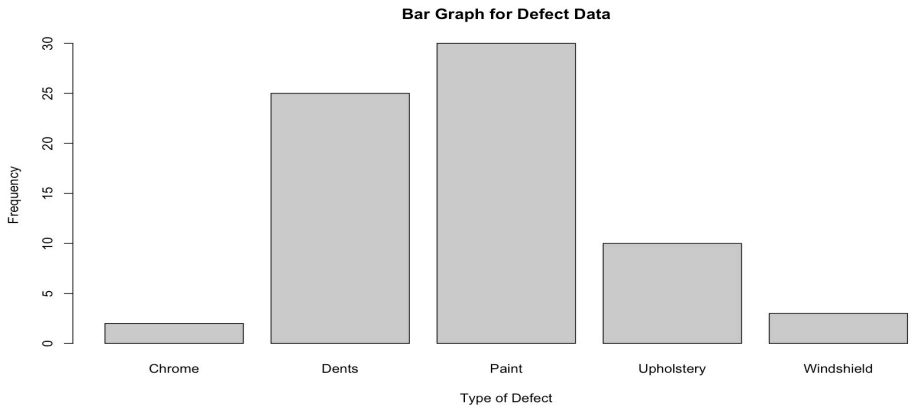
$$relative\ frequency = \frac{frequency}{n}$$

The **percent relative frequency** is the relative frequency expressed as a percentage.

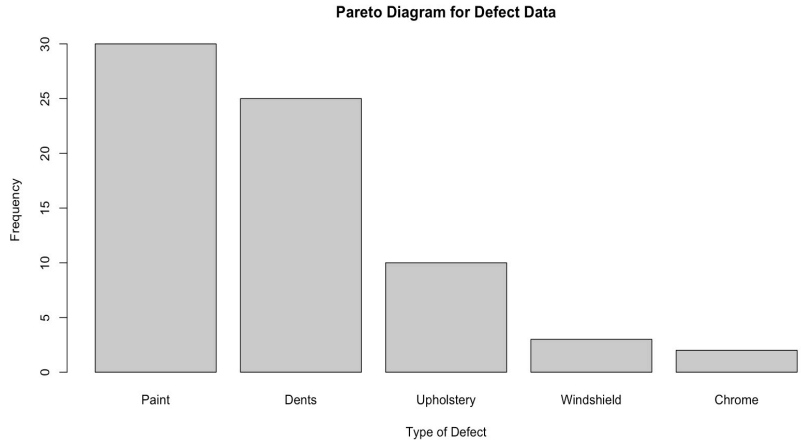
$$percent\ relative\ frequency = \frac{frequency}{n} * 100\,\%$$

A **Bar Graph** plots the categories on the *x* axis, and the frequency, relative frequency, or percent relative frequency on the *y* axis.

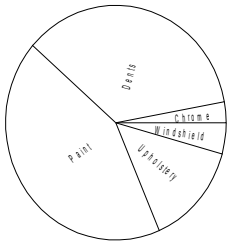
Bar Graph for the defect data:



In a bar graph, the categories are ordered in whatever way you wish to display them. Sometimes it may be a little easier to interpret the plot if we put the frequencies in ascending or descending order. A *Pareto diagram* is a bar graph in which the frequencies are sorted from largest to smallest:



If the categories include all possible outcomes, a pie chart may be informative.



In a pie chart, the area of a section corresponds to the proportion of observations in that section.

Graphs for Quantitative Variables:

- Histograms
- Stemplots (stem-and-leaf displays)
- Dot plots
- Boxplots
- Timeplots

Example. Text 2.32. Delivery time (in days, from order to delivery) for a make-to-order firm.

32, 33, 39, 43, 44, 49, 49, 50, 50, 51, 51, 54, 56, 59, 63, 64, 64, 65, 68,71, 73, 82, 86, 95, 102

A histogram is a very common way of displaying quantitative data. These days we rely almost exclusively on computer software to create histograms for us, so we won't go too much into the details. But let's look at the basics. To create a histogram, first we create a **frequency table**. To do this, choose

an appropriate number of classes (groupings of the variable). Depending on the sample size, this could vary from just a few classes to many thousand. Then choose appropriate class boundaries, and then count the number of observations within each class.

For example, in the above data, suppose we wish to create a histogram with 4 classes. We usually want to have a few more, but let's keep things simple here. Since our observations vary from a minimum of 32 to a maximum of 102, we'll need to span 102-32 = 70 units with our 4 classes. A class width of 20 units sounds reasonable enough:

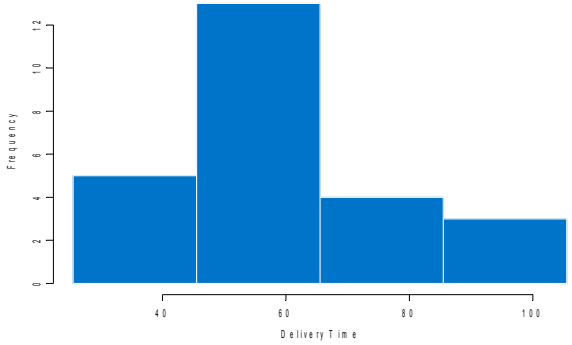
Class Boundaries	Frequency	Relative Frequency	% Relative Frequency	Cumulative Frequency	% Relative Cumulative Frequency
25-45	5	5/25 = .20	20.00%	5	20
46-85	13	13/25 = .52	52.00%	18	72
66-85	4	4/25 = .16	16.00%	22	88
86-105	3	3/25 = .12	12.00%	25	100

We have a couple of new terms here. The **cumulative frequency** of a class is the number of observations in that class and any lower class. For example, the cumulative frequency of the third class is 5 + 13 + 4 = 22, since there were 5 observations in the first class, 13 in the second, and 4 in the third.

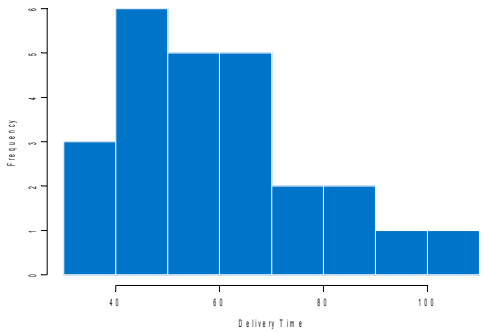
The **percent relative cumulative frequency** of a class is the percentage of observations in that class and any lower class. For example, the third class has a percent relative cumulative frequency of 22/25*100% = 88%.

A **histogram** is a plot of the class frequencies, relative frequencies, or percent relative frequencies against the class boundaries (or class midpoints).

Histogram of delivery times:



Four classes is not very many, and does not give us a very good idea of the distribution of data values. If we use computer software, we can easily choose more. Here is a histogram of the same data, using 8 classes:



The choice of classes in the table above is somewhat arbitrary. We want to choose enough classes to give us a reasonable picture of the shape of the distribution of our data, but not so many that there are too few observations in each class. The more observations we have, the more classes we can have. Computer packages (Excel, S-Plus, Minitab, etc.) often do a reasonable job picking a number of classes and class boundaries, but you can always change the default values, playing around until you get a pretty little histogram.

Stemplots.

Stemplots, also known as stem-and-leaf displays, give us a plot that is similar to a histogram but they are easier to construct by hand and they retain the exact data values.

To construct a stemplot:

1. Split each observation into a stem and a leaf.
2. List the stems in a column.
3. Write out the leaves next to their corresponding stem.

Consider again the data from the previous example.

32, 33, 39, 43, 44, 49, 49, 50, 50, 51, 51, 54, 56, 59, 63, 64, 64, 65, 68,71, 73, 82, 86, 95, 102

We could split each observation in the stems (the tens column – 3, 4, 5, ..., 10), and the leaves (the ones column). The leaf is the last digit of the number.

Stem Leaf	Stem Leaf	.....	Stem Leaf	Stem Leaf
3|2	3|3		9|5	10|2

```
 3| 239        (this represents the numbers 32, 33, and 39)
 4| 3499       (this represents the numbers 43, 44, 49, and 49)
 5| 0011469
 6| 34458
 7| 13
 8| 26
 9| 5
10| 2
```

Note that this plot is similar to a histogram turned on its side. These stemplots are easy to construct by hand, especially for smaller data sets. They can be useful, but are not used as often these days as in the past.

There are some different variants of the stemplot, including splitting stems and back-to-back stemplots. See the text for more information. A stemplot must include some type of legend informing the reader how to interpret the stem and the leaf. For example, a stemplot of:

.0012, .0014, .0016, .0016, .0018, .0022, .0029, .0031, .0038, .0052, .0052, .0067

from the computer package R looks like:

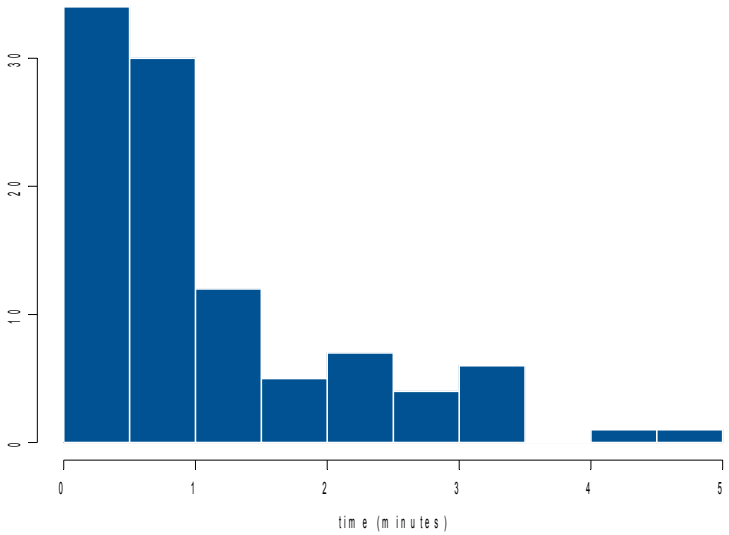
The decimal point is 3 digit(s) to the left of the |

```
 1 | 24668
 2 | 29
 3 | 18
 4 |
 5 | 22
 6 | 7
```

Without the line telling us that the decimal point is 3 digits to the left of the |, we would have no idea if the data set was 02, 04, 06, ..., or .2, .4, .6..., or .02, .04, .06,... etc.

Histograms and stemplots allow us to have a look at the distribution of the data. The **distribution** of a variable tells us what values it takes on, and how often it takes on these values.

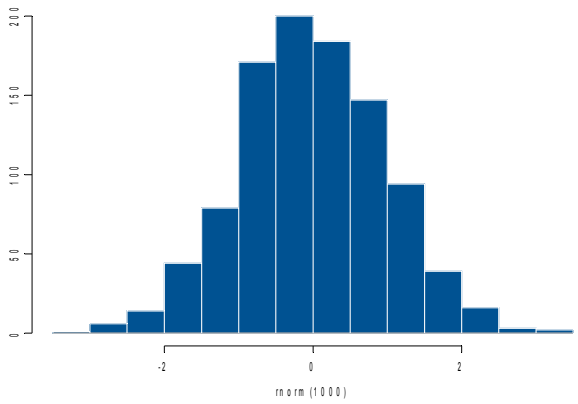
A histogram of customer waiting times at a call centre:



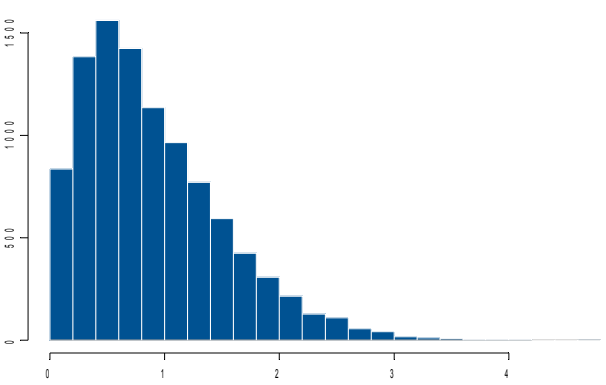
We are very often interested in:

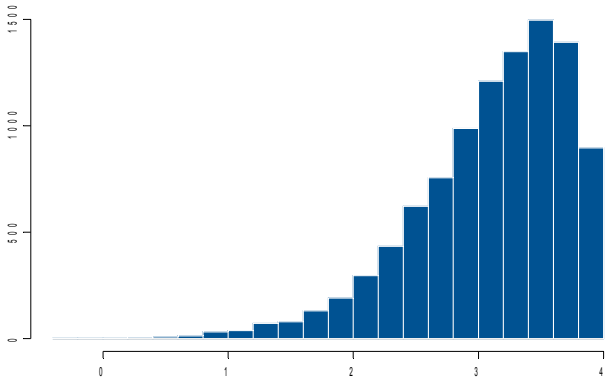
- The centre of the distribution (we'll use numerical measures such as the mean and the median as descriptive measures of centre)
- The variability of the observations (we'll use numerical measures such as the variance and standard deviation as descriptive measures of variability)
- The shape of the distribution. Is the distribution symmetric (left and right sides mirror images)? Is it skewed to the right (stretched out to the right side)? Skewed to the left?
- Outliers. Are there any outliers? Outliers are points that fall far from the overall pattern of observations. These can pose some problems for our descriptive measures and our statistical inference techniques.

An approximately symmetric distribution:



The distribution looks the same on the left and right sides.
A distribution that is skewed to the right: a distribution that is skewed to the left:





We very often see right-skewed data for things like salaries, housing prices, or time-to-event data. Left-skewness is not seen as often.

Numerical Measures

Let n represent the number of observations in a sample.

Suppose we have a sample of size $n = 4$: -4, 12, 18, -2

We label these values $x_1, ..., x_4$

That is; $x_1 = -4$, $x_2 = 12$, $x_3 = 18$, $x_4 = -2$

Summation notation

$\sum_{i=1}^{n} x_i$ means add up the x values from x_1 to x_n

Since we almost always want to deal with all the observations, we'll simplify the notation a bit:

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{4} x_i = \sum x_i = -4 + 12 + 18 + (-2) = 24$$

$$\left(\sum x_i\right)^2 = 24^2 = 576$$

$$\sum x_i^2 = (-4)^2 + 12^2 + 18^2 + (-2)^2 = 488$$

Note that this is NOT equal to $\left(\sum x_i\right)^2$.

Measures of Centre

(arithmetic) mean. This is just the ordinary average: $\bar{x} = \frac{\sum x_i}{n}$

The **Median** M is the value that falls in the middle when the data are ordered from smallest to largest:

- If n is odd, M is the middle value.
- If n is even, M is the average of the two middle values.

The **Mode** is the most frequently occurring observation.

Example. The Dow Jones Industrial Average (Sometimes called DJIA, or simply the DOW), is a composite of the stock prices of 30 large American Companies. Its performance is often given in the media as an indicator of the performance of the U.S. stock markets (although there are much better measures). A simple random sample of 5 companies from the DOW yielded stock prices of:

98.31 67.80 60.50 80.19 31.25

$$\bar{x} = \frac{\sum x_i}{n} = \frac{98.31 + 67.80 + 60.50 + 80.19 + 31.25}{5} = 67.61$$

To find the median, first order the data from least to greatest: 31.25 60.50 67.80 80.19 98.31

There are 5 observations, so the median is the third value in the ordered list (67.80)

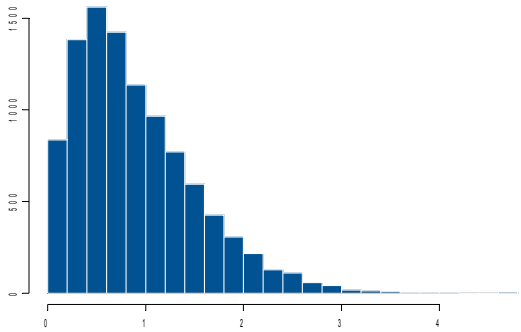
Properties of the Mean and Median

The mean uses every value in the data. The median uses the *ranks* of the every value, but only only the middle values are used in the calculation. Because of this, the mean uses more information, but is more sensitive to extreme values in the data.

For example, suppose we changed the above data set to:

31.25 60.50 67.80 80.19 980.31

The largest observation has been changed to 980.31 from 98.31. What effect does this have on the mean and median? The new mean is $\bar{x} = \frac{\sum x_i}{n} = \frac{980.31 + 67.80 + 60.50 + 80.19 + 31.25}{5} = 244.01$, which is a great deal different 67.61 (the mean of the original data). However, the median is unchanged at 67.80. The median is not nearly as sensitive to extreme observations as the mean. For right-skewed distributions:



The values in right tail will have more of an effect on the mean that the median. For these data sets, the mean will be greater than the median. For left-skewed data sets, the mean will be less than the median. If a distribution is perfectly symmetric, the mean and median will be equal.

Although the mean uses more information than the median, and is often our measure of choice in our statistical inference techniques, the mean can sometimes give a misleading measure of centre. When there is strong skewness, outliers or other extreme values, the median is often preferred as a descriptive measure of centre. We often see the median as the preferred measure of centre for housing prices, salaries, and time-to-event data.

Measures of Variability

The variability of a sample or population data set is often a very important measure. Consider a pop bottler putting 2L of pop into a bottle. They want to put an *average* of 2L of pop into each bottle, but they would also like the variability to be as little as possible. If there is a lot of variability, then some bottles will be overfilled, and some will be underfilled. This will not leave customers very happy. Or consider a manufacturer of bolts that have a 1 cm diameter. The manufacturing process must result in very little variance in the diameter of the bolts, otherwise many of them will not fit.

The simplest measure of variability is the range: Range = Maximum - Minimum

Consider a simple example of a sample of size $n = 4$: -4, 12, 18, -2. Range = 18 - (-4) =22. The range is a simple measure of variability, but it is not typically of much use. We could construct many different data sets, with very different variance, that all have the same range.

The best measurements of variability are based on deviations from the mean:

		(deviations)	*(absolute value of deviations)*	
i	x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$
1	-4	-4 - 6 = -10	10	$(-10)^2$
2	12	12 - 6 = 6	6	6^2
3	18	18 - 6 = 12	12	12^2
4	-2	-2 – 6 = -8	8	$(-8)^2$

Every observation has a deviation associated with it. Note that the deviations sum to 0. For any set of data: $\sum (x_i - \bar{x}) = 0$

The *mean absolute deviation* = (10+6+12+8)/4 = 9. This is the average distance from the mean. It is a reasonable measure of the variability of a data set, and has a nice simple interpretation. However, its mathematical properties are not as "nice" as those of the variance and standard deviation, so we will not use it very often.

We will base our measures of variability on the squared distance from the mean.

A common method of measuring dispersion is the sample variance: $s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$

The sample variance is the average squared distance from the mean. Why do we divide by n-1 and not n? Well, it has to do with the fact that we have sample data, and thus we are using the sample mean to estimate the population mean. In so doing, we lose something we call a **degree of freedom**. This term isn't overly important for us to know right now. Just know that the sample variance uses n-1 in the denominator. Using n-1 in the denominator provides us with a better estimate of the variance.

An equivalent calculation formula for s^2: $s^2 = \dfrac{\sum x_i^2 - \dfrac{(\sum x_i)^2}{n}}{n-1}$

These two formulas are equivalent. They will result in exactly the same value. The second one is often easier to use when calculating the variance by hand (on the calculator, without using programmed functions).

Note that the units of the sample variance are the square of the units of the original variable. For example, if we are measuring the weight of bags of potato chips in grams, then the sample variance will have units of grams^2. To get back to our original units, we deal with the square root of the variance.

The sample standard deviation s of a data set is the nonnegative square root of the variance.

Example. Consider the simple data set given above: -4, 12, 18, -2.

$$s^2 = \frac{(-4-6)^2 + (12-6)^2 + (18-6)^2 + (-2-6)^2}{4-1} = 114.6667 \quad , \quad s = \sqrt{114.6667} = 10.70825 .$$

N.B. Although it is important to know meaning of the variance and standard deviation, **I strongly recommend that you learn to use your calculator's pre-programmed functions to calculate them**. Most calculators worth $10 or more will calculate the standard deviation for you. If you are always using the above formulas to calculate the standard deviation, you will be wasting a whole lot of your time. For assignments, learn to use your calculator, or Excel, or any of a variety of other statistical software. For the final exam, make sure you know how to do this on your calculator. **It is a massive waste of time to calculate this from scratch this every time for the whole course**. Look up how to calculate the sample standard deviation on your calculator. If you no longer have the manual, google it. Test it out for a simple data set and make sure you know how to do it properly. Trust me on this one.

Note that the variance and standard deviation are nonnegative. They both have a minimum value of 0, and are only exactly equal to 0 if every single observation in a data set is the same. The larger the variance or standard deviation, the more variable the data set.

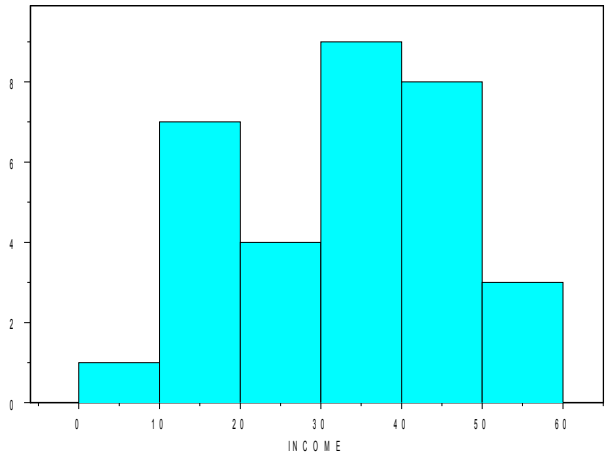
Interpretation of the Standard Deviation

The standard deviation is the square root of the average squared distance from the mean. Can we simplify this meaning? Not really. The *mean absolute deviation* is the average distance from the mean, which is a much simpler interpretation. However, as mentioned above, the mean absolute deviation does not have the nice mathematical properties that the standard deviation does. The standard deviation will be a little bit bigger than the mean absolute deviation. How much bigger depends on the data set. If you want to think of the standard deviation as the average distance from the mean, that won't be quite true, but you won't be too far off the mark.

Very rough guideline:
For many data sets the standard deviation will be in the neighbourhood of: Range/6 to Range/4.

Example. Operating income for NFL teams in 2004 (millions):

This represents ALL NFL teams in 2004. If we're investigating the operating incomes of NFL teams, then this is our entire population. Most times we'll be dealing with sample data though, so I'm going to pretend this is a sample.



Max = 54.3, Min = 7.8 . Range = 54.3 – 7.8 = 46.5.

sample mean = 32.425. sample median = 34.85. mean absolute deviation = 9.525.

sample variance = 172.6484, sample standard deviation = 13.13957.

You cannot calculate these number from the histogram; you would need the exact data values. But you should be able to *roughly approximate* these values from a histogram.

Measures of Relative Standing

In many situations, it is not the raw data score that is important, but how the score relates to other observations. For example, consider a person's score on an LSAT or MCAT exam. The raw score is not nearly as important as your score *relative to other writers of the test*. We'll look at two main measures of relative standard here: *z*-scores and percentiles.

Sample *z*-score: $z = \frac{x - \bar{x}}{s}$

A *z*-score measures the distance of an observation from the mean, in standard deviations.

The lowest operating income ($7.5M) in the NFL data above has a z-score of:
(7.5-32.425)/13.13957 = -1.89. This observation is 1.89 standard deviations below the mean.

The z-score of Bob's height among Canadian males is approximately 2.9. Canadian males have a mean height of approximately 174 cm, with a standard deviation of approximately 7. Hence Bob is about 194 cm tall.

A student who got 48 on the final exam of STAT*2060 W08 had a z-score of approximately -1 (One standard deviation below the mean).

As a rough guideline, if the distribution is mound-shaped:

Approximately 68% of *z*-scores will lie between -1 and 1.
[in other words, approximately 68% of the observations lie within 1 standard deviation of the mean]

Approximately 95% of *z*-scores will lie between -2 and 2.

All or almost all *z*-scores will lie between -3 and 3.

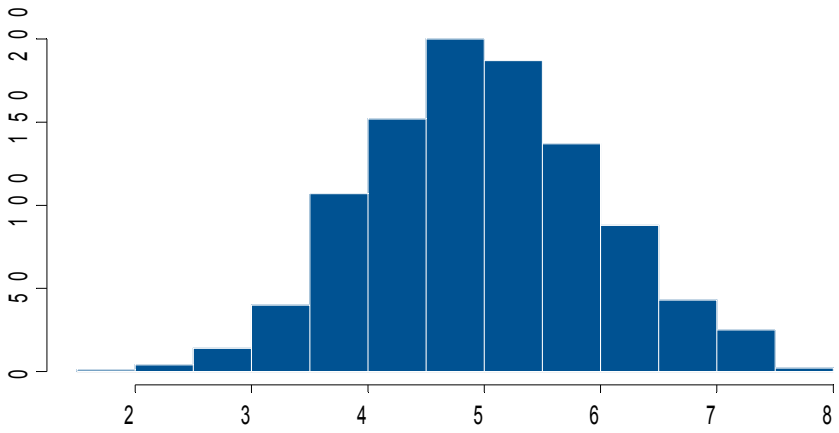
This is called the **empirical rule**. It is a *rough guideline* for mound shaped distributions.
Note that almost all *z*-scores lie between -3 and 3, a *z*-score outside of that range may be considered an outlier. Also note that the Range/6 to Range/4 rough guideline for the standard deviation discussed above is based on the empirical rule.

Percentiles

The p^{th} percentile is the value of the variable such that *p*% of the sorted data values are at or below this value.
(This definition is a little loose, but it gives the gist of the meaning)

Quartiles:

The first quartile (Q_1) is the 25th percentile
The second quartile is the 50th percentile (also known as the median)
The third quartile (Q_3) is the 75th percentile.



The quartiles split this distribution into quarters. The median would be the point along the *x* axis that splits the distribution, with 50% of the observations to the left. The median would be approximately 5. The first quartile would be the value with approximately 25% of the observations to the left. This would be approximately 4.2 (ballpark). The third quartile would be approximately 5.8.

We do have rules for the quartiles that are a little bit more specific:

Q_1 is the median of the observations to the left of the median in the ordered list.
Q_3 is the median of the observations to the right of the median in the ordered list.

NOTE: The above is not the only definition for quartiles, as different software and textbooks may define it differently. For large data sets, there is typically little difference between the different definitions, but the numbers can be different. We'll go with this definition for consistency's sake.

Other important quantities:

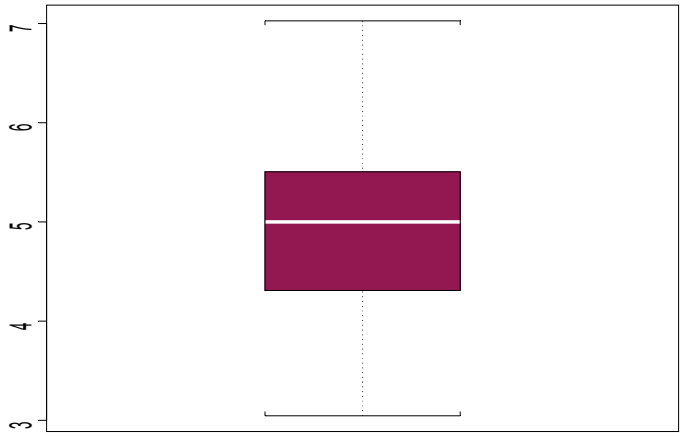
The interquartile range is defined as $IQR = Q_3 - Q_1$, and is a measure of the variability of the observations.

The *IQR* is not as sensitive as the variance and standard deviation (not as affected by extreme observations). It is used mainly as a **descriptive** measure of variability, and not for the inference procedures that we will use later in the course.

Boxplots

Boxplots are another plot for quantitative variables. Although they can give some indication of the shape of the distribution, they are not as effective as histograms for this purpose. The main purpose of boxplots is in comparing two or more distributions.

Boxplots look like:



A boxplot is made up of:

- A box extending from Q_1 to Q_3
- A line through the centre of the box marking the median
- Lines (whiskers) extending from the box to the largest and smallest observations (to a maximum length of 1.5*IQR). Any observation outside of these values will be considered **outliers**.
- Outliers (extremely large or extremely small observations) are marked individually outside the whiskers.

Example.

Find Q_1, Q_3, and draw a boxplot for the following data.

112 114 120 126 132 141 142 147 189

Median = 132.

Q_1 is the median of the values to the left of 132. This is (114+120)/2 = 117

Q_3 is the median of the values to the right of 132. This is (142+147)/2 = 144.5

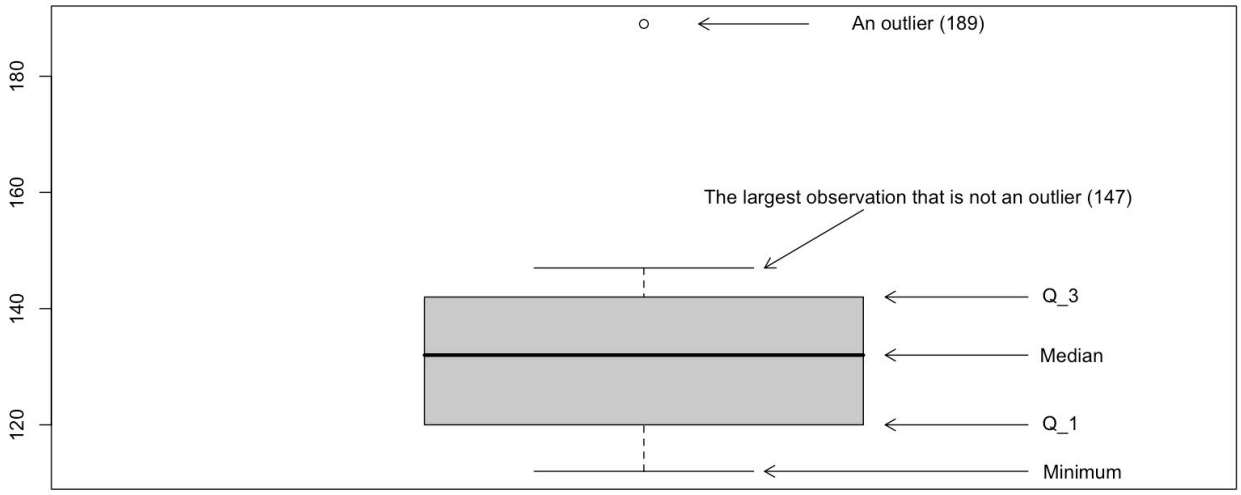
IQR = 144.5-117 = 27.5.

Using our guideline for outliers:

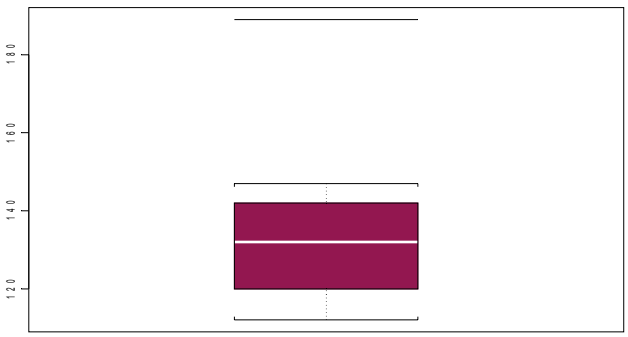
Any observations greater than 144.5 + 1.5(27.5) = 185.75 is a suspected outlier. 189 falls into this category. So, the upper whisker of our boxplot will stop at the largest observation that is less than 185.75. This is 147. 189 will be drawn in individually.

On the low end, 117-1.5(27.5) = 75.75. No observations in our data set are less than 75.75. So, the whisker will stop at the lowest observation, 112.

Annotated boxplot from the statistical package R for the above data:



Plot from the statistical package S-Plus:



Note that outliers may be drawn in using dots, lines, asterisks, etc.

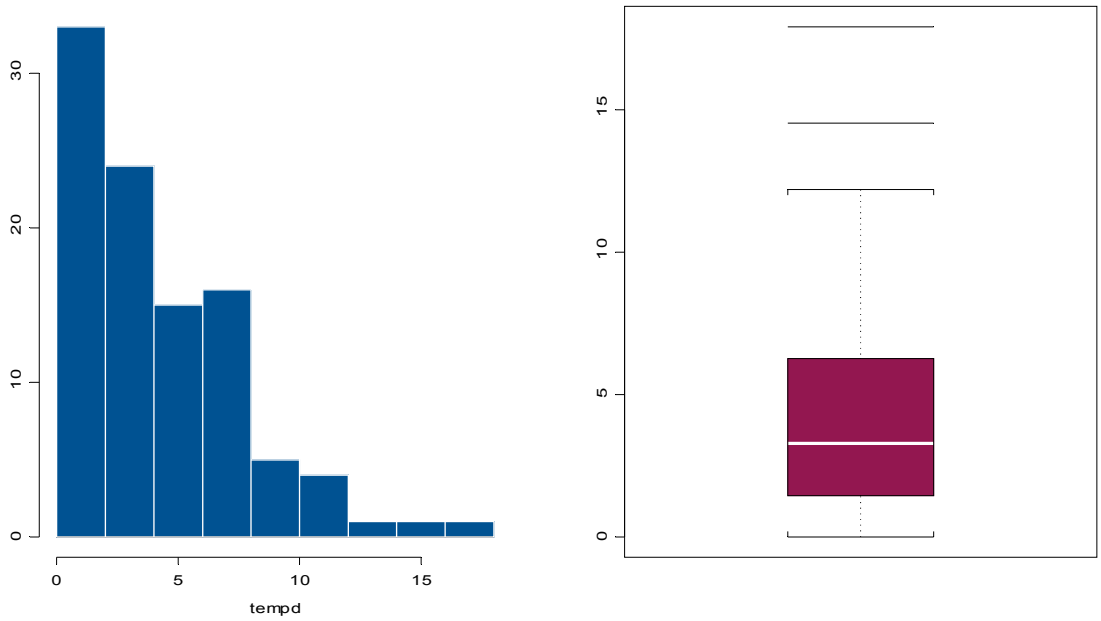
Boxplots can of course be a pain to draw by hand. I may get you to construct boxplots using a computer program, but your main job as far as boxplots go will be to properly interpret them.

Histogram, boxplot and stem-and-leaf display for a simulated data set that is skewed to the right

Decimal point is at the colon

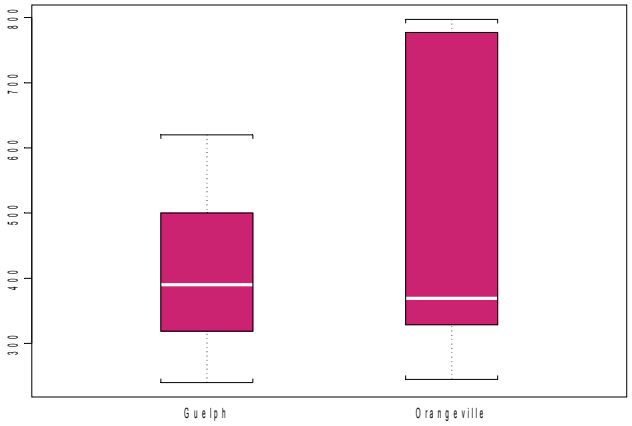
0 : 000111112225667889
1 : 00113336666689
2 : 00001244456669
3 : 02233466678
4 : 123555
5 : 01125589
6 : 00024466779
7 : 245568
8 : 167
9 : 67
10 : 457
11 :
12 : 02
13 :
14 : 5

High: 17.91077



All 3 plots above illustrate the skewness of the distribution. Boxplots can give us an idea of the shape of the distribution, but histograms and stemplots usually give us a better picture of the shape. Boxplots are most useful for *comparing* two or more distributions.

Example. Listing prices for detached houses in Guelph and Orangeville
(Sept 2008 – random sample of 10 listed on mls.ca for each city)



The samples from both locations have a median in the 350-400 range, but there is some indication that the variance in prices in Orangeville is greater than that of Guelph. We can't say anything definitively from samples of size $n = 10$, but there is some indication of that effect. Boxplots allow for easy visual comparison of these effects.

Linear Transformations

Example. Suppose we have a set of measurements in $US: $57, $42, $89, $121. These result in a sample mean and standard deviation of $\bar{x}=77.25, s=35.141$

Suppose we need to add a $12 US shipping cost to each, and then convert to $CAN.

On August 30 2010, the exchange rate was approximately $1US = $1.04CAN.

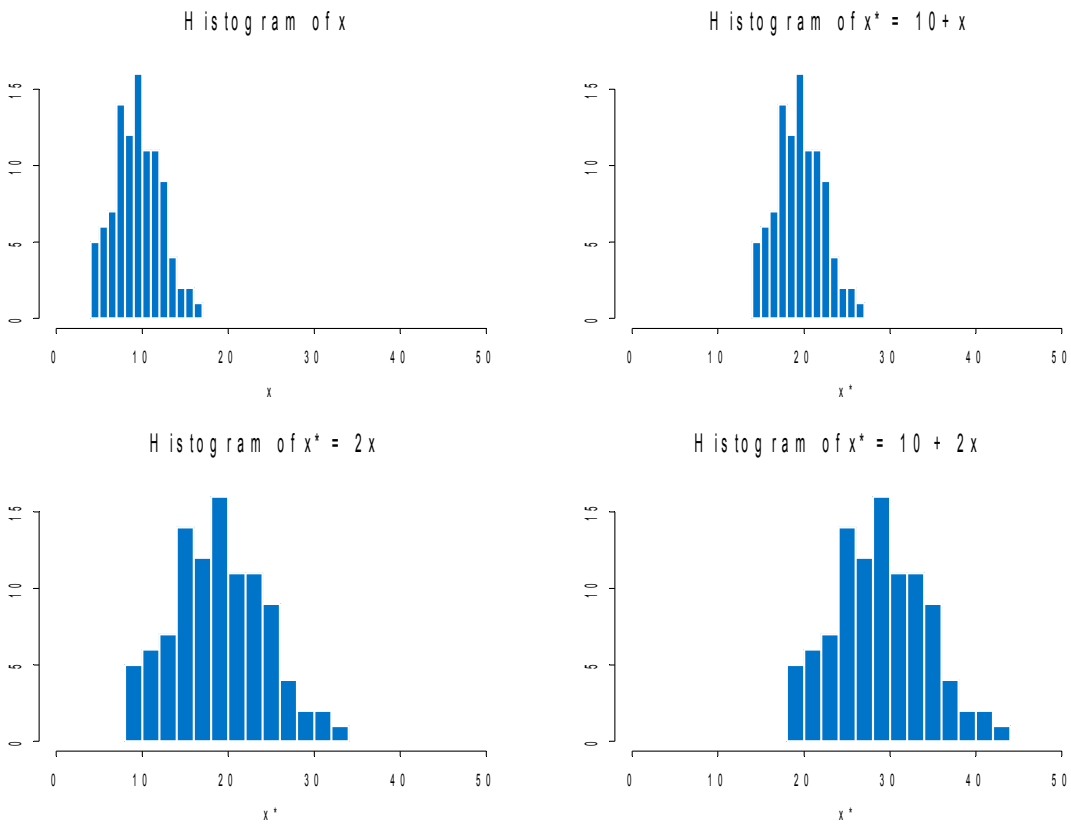
Our cost, in $CAN, is thus $1.04(12+x)$

$$x^{*}=12.48+1.04x$$

where x is the cost (without shipping) in $US and x^* is our cost in $CAN, including shipping.

This is an example of a linear transformation: $x^* = a + bx$. These transformations come up frequently. It's a good idea to understand the effect of these transformations on the summary statistics.

Consider the following histograms, representing some linear transformations.



Note that in the above plots, adding a constant changes the measures of position (mean, median, quartiles), but **does not change the variability**. If we add 3, say, to every observation, then the

distribution will be shifted 3 units to the right, but there will be no difference in the variability.

Multiplying by a constant changes both the measures of position and the variability.

The effect of a linear transformation on the summary statistics: $\bar{x}^* = a + b\,\bar{x}$

This holds true of other measures of position as well (median, quartiles): $M_{x^*} = a + b\,M_x$

But the additive constant *a* does not affect the measures of variability:

$s_{x^*} = |b|\,s_x$, $IQR_{x^*} = |b|\,IQR_x$, $s^2_{x^*} = b^2\,s^2_x$

For example. From above, the original cost in $US had summary stats of $\bar{x} = 77.25, s = 35.141$

Adding shipping and converting to $CAN, we had the linear transformation: $x^* = 12.48 + 1.04\,x$

So our costs in $CAN have a mean of 12.48+1.04(77.25) = $92.82. Since the additive constant does not change our standard deviation, the costs in $CAN have a standard deviation of 1.04(35.141) = 36.54664.

Suggested Textbook Readings: Textbook: Chapter 1 (all), Chapter 2 (all)

Suggested Textbook Questions:

Chapter 1: 1.3, 1.6, 1.8, 1.13, 1.17, 1.21, 1.27.
Section 2.1. Describing Qualitative Data.(Page 38): 2.3, 2.7, 2.13.
Section 2.2 Describing Quantitative Data (Page 48): 2.19, 2.21, 2.31.
Section 2.4 Numerical Measures of Central Tendency (Page 60): 2.37, 2.39, 2.43, 2.52.
Section 2.5 Numerical Measures of Variability (Page 68): 2.57, 2.59, 2.63, 2.65.
Section 2.6 Interpreting the Standard Deviation (Page 75): 2.71, 2.73, 2.75, 2.77, 2.85.
Section 2.7 Numerical Measures of Relative Standing (Page 80): 2.89, 2.91, 2.97, 2.101.
Section 2.8 Methods for Detecting Outliers (Page 87): 2.103, 2.105, 2.107, 2.109.
Chapter 2 Supplementary Exercises (Page 103): 2.129, 2.131, 2.133, 2.134, 2.155, 2.159

Suggested Course Manual Questions: Unit 1 (All).
But learn to use your calculator to automatically calculate the standard deviation.