

Data Analysis and Statistical Methods

Statistics 651

<http://www.stat.tamu.edu/~suhasini/teaching.html>

Lecture 2 (MWF)

Suhasini Subba Rao

A representative sample

- When making a confidence statement (inference) about a population based on a sample we need to ensure that the sample is somehow representative of the data.
- For example, if we want to make a confidence statement about the mean height of students at A&M (the population is all students at A&M) based on a sample containing only females. It is likely that this sample will be biased.
- This sample is NOT a representative sample of students at A&M.
- Female students form a subpopulation of the population of all students. The sample is representative sample of female students, rather than the population of all students.

- A 'representative sample' has nothing to do with sample size.
- A simple random sample (SRS) is an example of a representative sample. This is where every individual in the population has an equal chance of being selected. No subpopulation is excluded.
- Using a SRS strategy, there is always a chance that an individual will be selected more than once.

We briefly return to an SRS, at the start of Lecture 3, where give an example.

- For surveys, implementing a true SRS is usually not feasible; people do not want to be interviewed twice. However, if the population size is sufficient large as compared to the sample size, the chance an individual is sampled twice is extremely small.

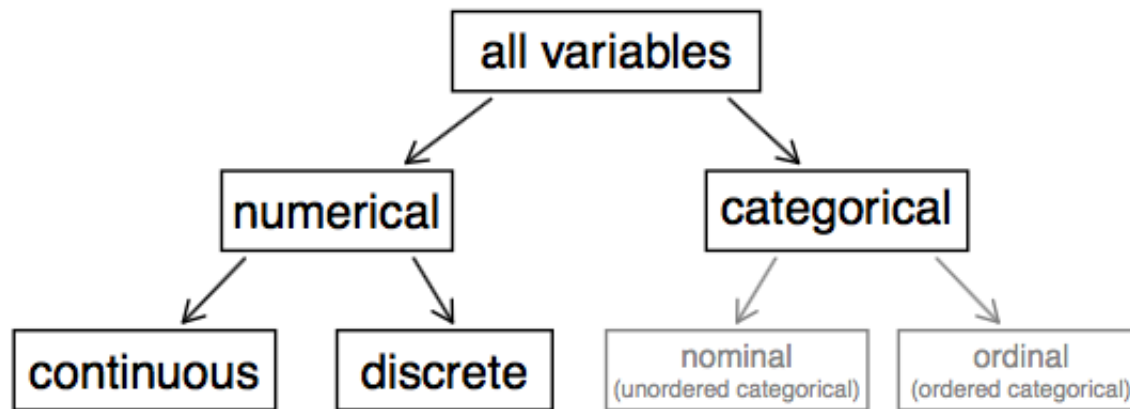
- In this case, usually a rule of thumb a sample size of at **most** 5% of the population size is thought is yield
- Designing an experiment in a good way is extremely important, but something we shall not cover in this course.
- In this course we will mainly assume that the sample is simple random sample.

Samples, Populations and Variables

- The population and sample are made up of *individuals* (these are not necessarily human), these can be people, companies, animals, a chemical etc.
- A variable is a characteristic in the individual that we are interested in. For example, for people it could be height, blood pressure, ethnicity or mother tongue.

The characteristic of interest varies from individual to individual it is natural to call it a **variable**. We will learn later that since it is variable it is 'random'.

Different types of variables



Different variables in an M&M bag

- In a bag of M&Ms we may be interested in the main colour, number of M&Ms, weight of bag, type of M&M (chocolate or peanut) etc.

bag no.	majority colour	number of M&Ms	weight of bag	type
1	blue	18	2.2 ounces	chocolate
2	brown	19	2.3 ounces	chocolate
3	red	12	2.1 ounces	peanut

- **Types of Variables** From the above we can see that variables come in several different types:
 - Numerical continuous: eg. weight (2.2 ounces)
 - Numerical discrete: eg. the number of M&Ms in a bag (18)

- Binary: eg. Type (chocolate/peanut)
- Categorical: eg. Majority colour (blue/brown/red/green)

- Numerical variables always have a meaningful ordering. Beware of categorical variables disguised as a numerical variable. For example, the number of a bus is not a numerical variable but a categorical variable.

- In statistics we treat different types of variables in different ways.

- There are two types of Numerical variables, numerical discrete and numerical continuous. There is an interesting connection between these two variables. Numerical discrete variables “become” numerical continuous variables when averaged. For example, the number of children in a family is discrete but the average number of children in a family is continuous.

- Therefore in this course we will treat numerical continuous and numerical discrete variables in the much same way (with a few exceptions).
 - In more advanced courses (such as STAT652), where more sophisticated models are used. Numerical discrete and numerical continuous variables will be treated differently.
-
- During the course we will consider different methods for treating different types of variables.

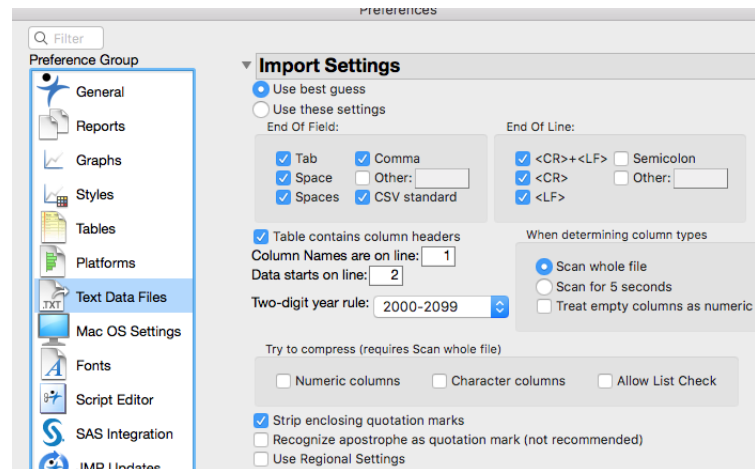
Examples of variables

What type of variables are the following:

- The gender of a randomly chosen person (we can use M/F or 0/1)?
- The number of a randomly selected bus?
- The make of bicycle of a randomly chosen person?
- The number of bicycles owned by a randomly chosen person?
- The height of person?
- Whether a random selected person responds to a drug?
- The predictions of Paul the octopus (win or lose).

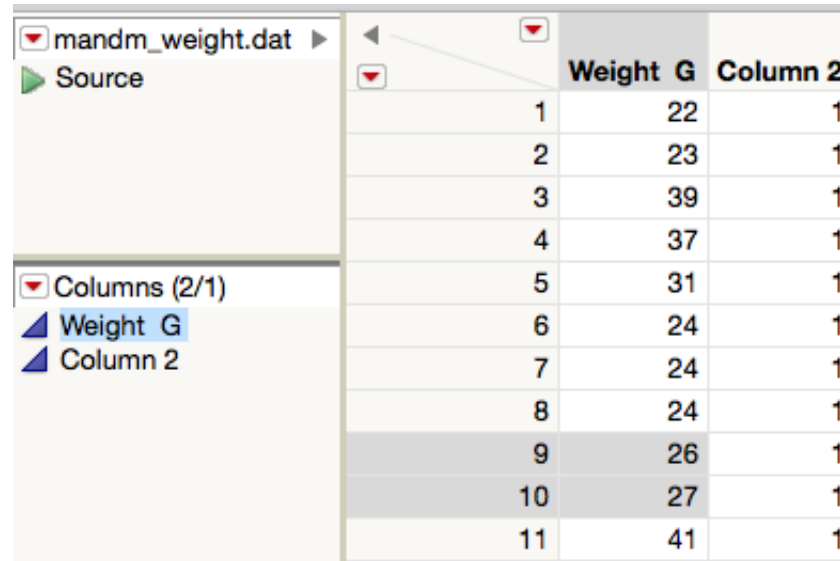
Opening data in JMP

- Open JMP
- JMP > Preferences > Text Data Files > Import Setting **check** the **Use best guess** box. This displays the data in the correct way (recognizes spaces or commas as a new column etc,). A screen shot is given below of what needs to be ticked.



- **Data on hard drive** Go to File > Open. Then you will see a Finder or File Manager. Select file and and press open.
- **Data on internet** Go to File > New (an empty spread sheet will pop up) > File > Internet Open...A window will pop up.
- Paste desired url in the pop-up window.
- You should see the data in a JMP spreadsheet.

The data in JMP



		Weight G	Column 2
1		22	1
2		23	1
3		39	1
4		37	1
5		31	1
6		24	1
7		24	1
8		24	1
9		26	1
10		27	1
11		41	1

- The symbol on the left indicates how JMP reads each variable.
- The blue right angle triangle mean JMP reads the variables are continuous numerical.

- You can change the “type” of variable by clicking on the symbol/triangle. It can be changed to
 - ordinal (numerical discrete; data with an ordering such as ratings)
 - nominal variable (which is another name for categorical).
- Ensuring the type variable is correctly is specified in JMP is important for using the appropriate statistical procedure.

Statistical Analysis comes in three stages

- (1) Data description. When starting a data analysis first use a graphical method to represent the data (Chapter 3, Ott and Longnecker). I.e. histograms, pie charts, line graphs, line and whisker plots etc.
- (2) Summary statistics, average (mean), median, variance, quantiles etc. This describes the data set (which can be large) in a few numbers, it also gives us an idea about the spread of the data.
- (3) Quantative techniques (this will be the main focus of the course, Chapter 3-11, Ott and Longnecker). We can evaluate an average, but what does this average tell us about of the true population average (usually called population mean)? How close is the sample average to the population average? We will be finding out a few weeks from now.

Histograms

- An important plotting tool for depicting the “distribution” or numerical data is the histogram.
- It is a visual aid which gives information on the spread of data, which outcomes are more likely and the shape of the spread.
- Though simple to define, it has many deep mathematical properties.
- Example Consider the data 22, 41, 23, 39, 37, 31, 24, 24, 26, 27.

The **Range** The smallest interval which contains all the data for this example it is $[22, 41]$.

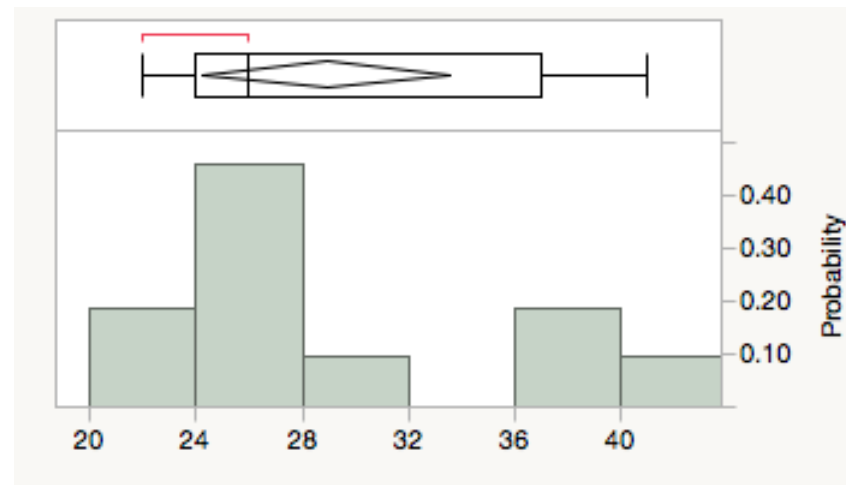
The range is partitioned into **bins** (usually, but not always, into equal parts). The **relative frequency** is the proportion of observations in each bin. The size of each bin is called the **binwidth**.

Plotting a Histogram

- Example using binwidth 4.

interval	[20-24]	[25-29]	[30-34]	[35-39]	[40-44]
count	4	2	1	2	1
percent/relative frequency	40%	20%	10%	20%	10%

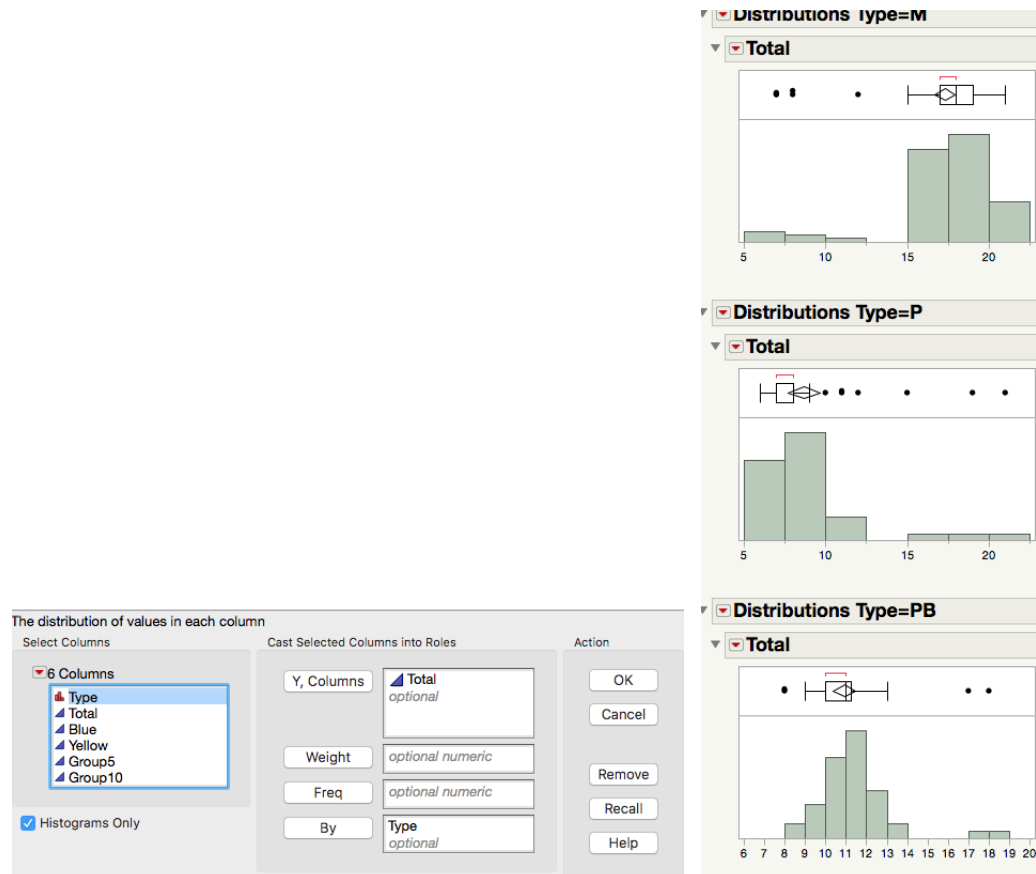
- The Histogram in JMP:



Plotting a Histogram in JMP

- Once data is loaded into JMP.
- Analyze > Distribution. A window will pop up with variable. Highlight and (double) click on variable you want to plot. Press OK.
- You can adjust the histogram by selecting red arrow next to the variable and going to Histogram option.
- To get counts on the y-axis choose Count Axis. To get proportions/relative frequency (percentage of data on the y-axis) choose Prob Axis. You can change the bin width by selecting Set Bin Width.
- If you click on a block in the histogram, it will be highlighted as a striped block. The data which contributes to that block will be highlighted on the corresponding spreadsheet.

- To make comparisons between *different* subgroups, highlight the factor variable click on By.

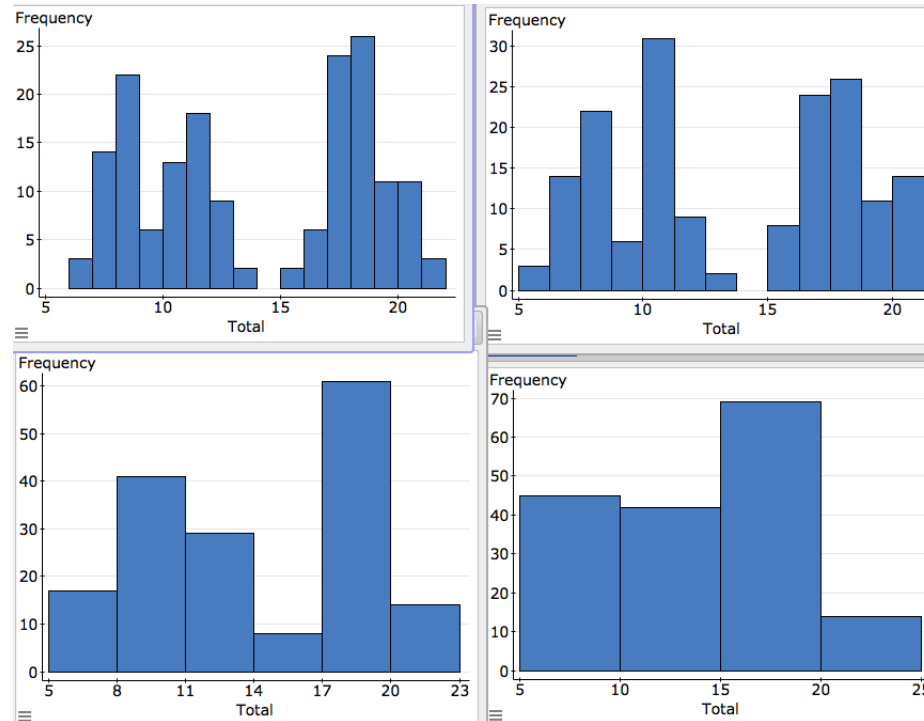


Features in a histogram

We can use the histogram to observe the following features:

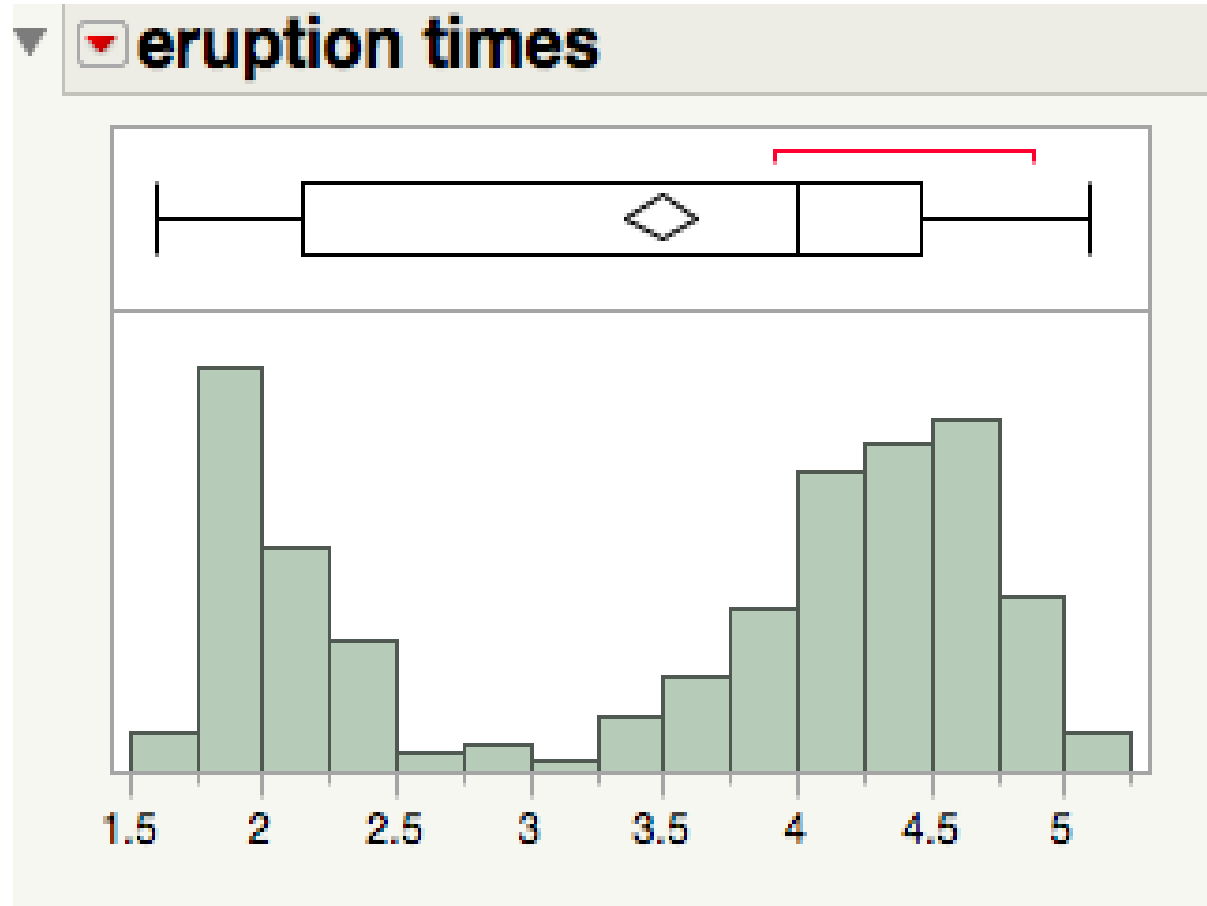
- Outcomes that are most frequent.
- If the sample is a composition of multiple populations (more of this later), these can be seen with multiple modes in the histogram.
- The spread of the data, is it concentrated or spread out.
- Most statistical software packages have a default method for selecting the bin width size. These usually give a good description of the data. But there are situations where you may have to manually change the bin width.

The distribution of M&Ms and bin width



Observe how different bin widths can change your perception of the same data set.

Eruption times of Old Geyser



Using a histogram to compare populations

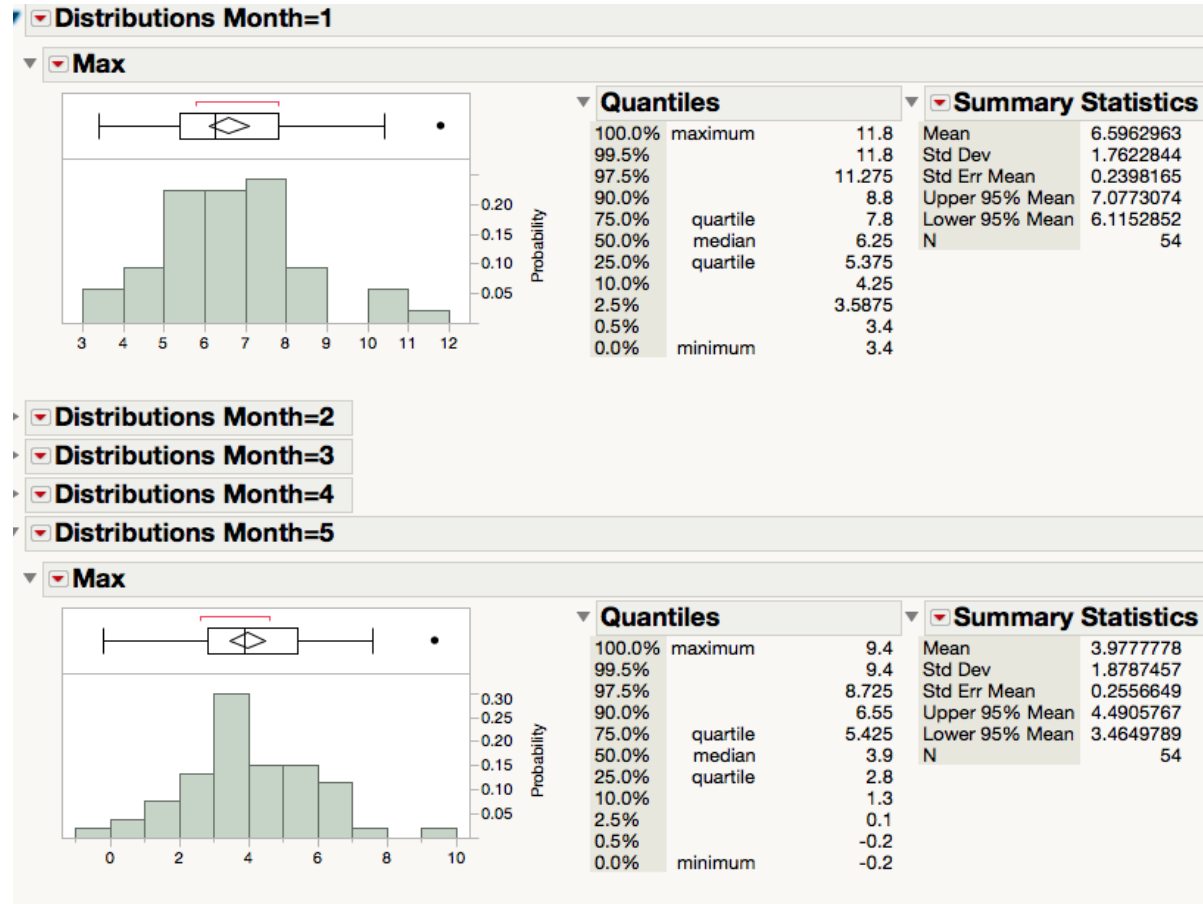
- A histogram is a very useful tool for comparing samples and seeing whether they come from the same or from different populations. We will learn more quantitative methods of comparison later in the course. What we do now is simply a visual comparison.

- **Example**

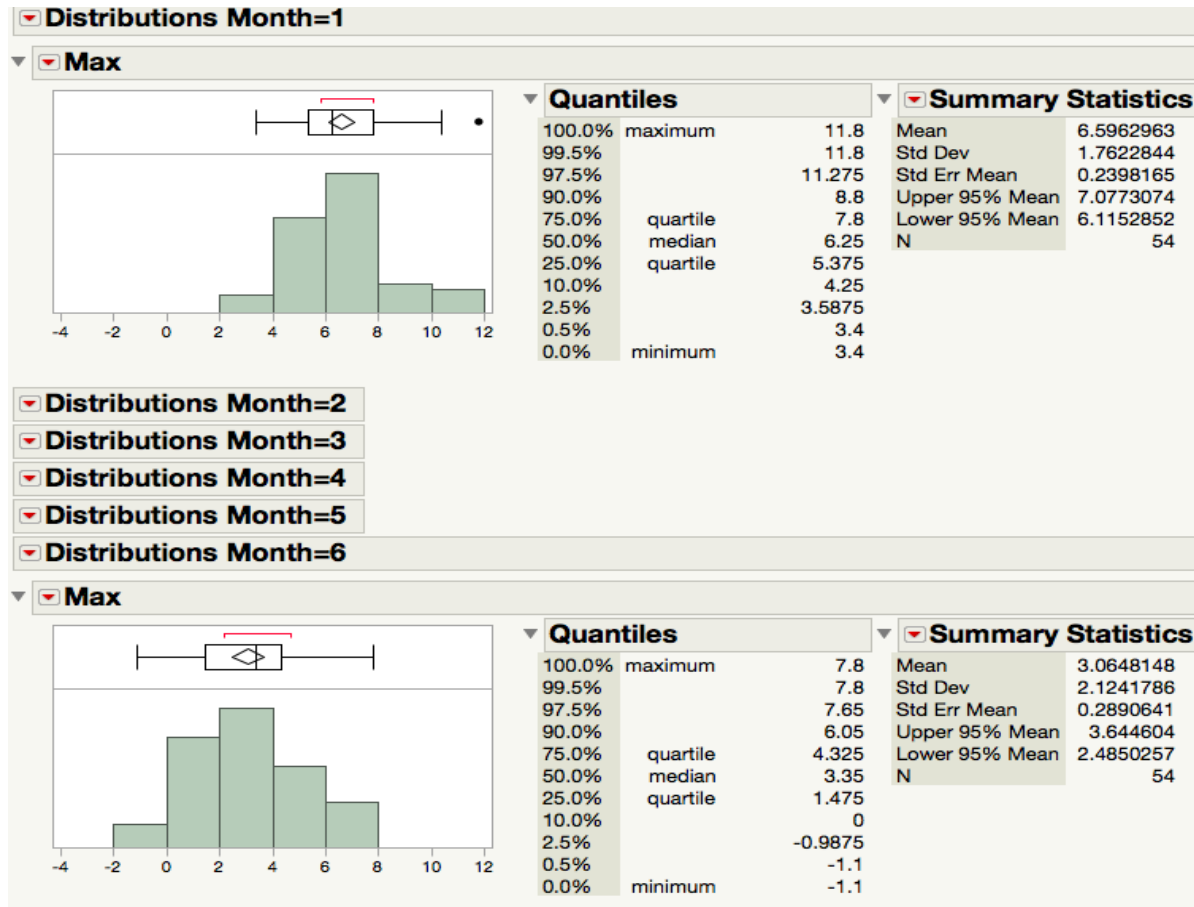
We could expect the temperatures in January in the Antarctic to be more than those in May in the Antarctic (recall that in the Antarctic, January is summer and May is winter) not that all temperatures are in Celsius.

Below are the histograms of data taken in January and a sample taken from May (maximum temp). What do you think?

Comparing temperatures in the Antarctic



Comparisons should be done using the same scaling



- The top plot the summer temperatures and the lower plot are the winter temperatures in the Antarctic between 1951-2005. What do you notice?
- We see that the histograms appear to be a shift of each other.
- How to quantify the main features and the differences?
- There are several ways to do this. One way is to consider a numerical value which describes a feature in the data, and to compare the numerical values from each sample.
 - From the point of view of statistical inference, it is much easier to compare numerical values than graphs.
 - One measure of center is the average (sample means) of the sample.

The Histogram of continuous variables

- For discrete variables the relative frequency histogram is an appropriate way to represent the frequency/distribution of a sample or population.
- However, a relatively frequency histogram cannot convey all the information in continuous variables.
- The information in the relatively frequency histogram is restricted by the selected binwidth.
- If the binwidth of the plot is two, you cannot obtain the proportions for bins less than two.
- To get over this problem (and other mathematical issues), we define a closely related cousin of the relative frequency histogram called the density plot

- The density plot is the same as the relative frequency histogram but effectively has a binwidth of zero.

The density plot

- Since the bin width is zero, the density plot is a little different to the histogram.
- It is **the area** under the graph represents the frequency of an event and **not the height**.
- To plot the distribution of the population of numerical continuous variables we will **always** use the density plots.
- The area under the curve is used to calculate probabilities. But the height of the plot will help understand which outcomes are most likely to occur.

The Shape of a distribution/density

The shape of a density gives important information about the population.

- Variables whose distributions tend to be close to symmetric:
 - Heights of a certain gender.
 - Length of bird bills and other biological lengths.
- Variables whose distributions tend to be skewed:
 - Price of houses.
 - Gestation period of a baby.
- Variables whose distributions tend to be multi-modal (have several distinct peaks).
 - The height of adult humans (both sexes),

- Number of M&Ms in a bag (all types, Peanut/Milk chocolate/Peanut butter).

Multi-modal densities suggest that it is a mix of subpopulations.

- Variables whose distribution tends to be flat (uniformly distributed):
 - The numbers in a lottery.
- Each numerical continuous variable will have its own density plot, with its own features.
- In general, the distribution will not be bell shaped. Skewed distributions very common.