Math 321 – Chapter 1 – Introduction, Data, and Basic Statistics

`(Spring 2020 version, compiled 2020/01/08-12:52:55)`

Contents

1 Introduction: Populations, Samples, Data, and Variables

Statistics. What is *statistics*? The science of statistics has two primary components: (1) to analyze data (observations of physical populations, processes, or experiments) and to understand the properties of data, and (2) to use data and our analysis of it to understand the underlying population or process that generated the data. This involves understanding variability, uncertainty, and randomness. These are loaded terms, but you probably have an intutive feel for what they mean. We'll discuss them further later.

Populations. The ultimate goal of statistics is to understand the properties of a *population*. A population is a collection of individuals and can be *concrete* or *hypothetical*. A concrete population is one where all the individuals currently exist. A hypothetical population means that some individuals do not yet exist. In some cases, a population can better be thought of as a *process* that generates individuals as in the case of a particular experimental or man-made apparatus or a natural process. Table 1 shows examples of concrete and hypothetical populations.

For example, we may be concerned with industrial machinery that produces something. Assuming the machinery has a constant condition and is maintained constantly over time, we might ask, what is this machinery capable of producing? We may have a collection of produced items already, but we are interested in all possible items (capable of being) produced, which is hypothetically an infinite population. We are really interested in the property of that machinery, say it's likelihood of producing items that meet some desired specification. We might investigate this by looking at the items produced over a given time period and test some or all of those to see if they meet the desired specification.

Concrete	Hypothetical
all US currently living citizens	all babies born in the US in the current year (some in the past and some in the future)
items produced by a factory so far this year	all items capable of being produced by a factory (Not all yet exist. Some will never exist! But they are all hypothetically capable of being produced.)
a jar of coins	an infinite sequence of coin flips from a single coin
last year's harvest of a particular crop	all possible harvests of a particular crop (with some fix set of genetic material and its variability, for example)
all voters' current opinions on a particular political issue	all votes cast on the next presidential election

Table 1: Examples of concrete and hypothetical populations.

Samples. Instead of studying an entire population, which is often unfeasible due to the size of the population or the cost of such a study, we are usually limited to studying a subset of a population. A *sample* is a subset of a population. In most cases, the goal or hope is that the sample resembles the population in the trait or quantity of interest. Statistical theory also studies ways to select samples and how to evaluate whether they are good methods.

Data. What is *data*? Data is a collection of observations of some phenomenon. There are two primary types of data: *numerical* and *categorical.* Numerical data is often referred to as *quantitative* data, and categorical data is often called *qualitative* data. Quantitative data often result from some kind of measurement, e.g. time elapsed for a given event or the physical trait of an object (length, area, volume, mass, etc.). For data to be quantitative, it must represent a quantity or amount of something, and arithmetical operations on it must be meaningful (e.g. adding two data values has a valid interpretation). Categorical data occurs when observations can be organized into a finite number of discrete classes, e.g. color, type, etc. We are primarily be concerned with quantitative data in this course.

Quantitative data can further be described as *discrete* or *continuous.* If all possible values can be counted or at least hypothetically enumerated for counting, then the data is discrete. Discrete data might have an infinite set of possible values, e.g. $\{1, 2, 3, 4, \ldots\}$, all natural numbers without any obvious maximum cut-off. If data can take any value in some interval, e.g. all real numbers in $[0, 1]$, it is continuous.

Do not get bogged down in worrying too much about classifying data as numerical/categorical or discrete/continuous. What is important to know is whether a given statistical method is valid for a particular data type. In other words, methods that we will use on continuous data do not always work well on discrete data. This is something that will be discussed for each statistical technique we encounter. If you go on to take more advanced statistics courses, then understanding more about different types of data classifications will become more important.

Variables. Quantitative data is composed of numerical values for a particular variable, we'll usually use the letter X. The variable X might represent something like water well depth (an example we'll look at shortly). Each water well has a particular depth, and different wells have different depths in most cases. Theoretically, a well can take on any depth from, say zero up to some maximum limiting value but such a maximum value is difficult to determine precisely. So we will just say that X takes on values in the continuous interval $[0, \infty)$.

We will refer to X as a *random variable.* This is a concept that we will elaborate on more later in the course, but for now just think of it as being a variable that takes on values over a population, e.g. well depth for the population of water wells. In order to find the value a random variable takes, we will need to select an individual in a population and then gather the necessary information, e.g. measure water well depth. For now, just consider the term *random* to mean that there is uncertainty in the value observed. There are many factors than can affect water well depth, for example: the detailed structure of the earth's crust, local environment, drilling company's practices, and the depth measurement process.

Variable notation. Normally a capital letter X is used for the (random) variable as an undetermined value (i.e. water well depth in general) and a lower case letter x when we are looking at a particular individual in the population, e.g. the depth of a particular well. In other words, X is the variable representing well depth, and when we look at a particular

Categorical	Quantitative
species of trees on a specific plot of land	the number of trees on that plot of land (discrete)
paint color of a vehicle	volume of paint required to coat a vehicle (continuous)
nationality of a person on a plane	the number of passengers on the plane (discrete)
variety of grain planted in a field	total harvest from that grain field (possibly discrete if number of bushels with a strict rounding rule, continuous if volume, mass, or any decimal units with reasonable precision)
responses to a multiple choice public survey on a particular political issue	proportion of a population that supports a particular political issue (generally continuous, but could be thought of as discrete if we have a fixed sufficiently small population size)
whether or not a manufactured building material failed a quality control check	force required to cause failure of a building material (continuous)
type of lightbulb (e.g. LED, incandescent, fluorescent)	time until failure of a lightbulb (time-elapsed measurements are generally treated as continuous if measuring with reasonable accuracy)
model year of a car (1995, 2018, etc. Note that the year model does not always represent the actual year that the vehicle was manufactured in)	number of years a car has been in service (discrete, if only rounding to whole years, continuous if measuring time with high decimal precision)

Table 2: Examples of categorical and quantitative data. Quantitative variables are noted as discrete or continuous.

well, X takes on an actual numerical value x. Lower case x here is your traditional variable in the algebra and calculus sense. E.g. if a particular water well is determined to have depth 127.3 feet, then 127.3 is a particular x, or we can say that $X = 127.3$ for this particular water well.

A note about continuous vs discrete. The line between discrete and continuous can sometimes be blurry. Consider money as an example. In many circumstances, money can be counted in terms of a smallest unit (e.g. the penny, \$0.01, for US physical currency coins and bills), but in many circumstances it can be counted in fractions of a cent (e.g. prices in the stock market are sometimes given out to 4 or 5 decimal places). The advent of high-value cryptocurrencies (e.g. Bitcoin) further show this.

Also consider exchange rates and that, at the present moment I am writing this, 1 Iranian Rial is worth 0.0000237876 US Dollars. In principle, money could theoretically be subdivided without limit and is often treated as continuous.

Also, measurement is limited by the accuracy of the measuring device (e.g. 1/16 of an inch resolution on a ruler), so one could similarly think of length measurements as only taking on discrete values. Or we could think of length as being the number of molecules along the edge of an object! However, in most circumstances length and most other physical measurements are though of as continuous.

Our potential measurement accuracy is usually, at least in principle, well beyond the human perceptual accuracy, and this is another justification for treating certain variables as continuous.

A note about categorical vs quantitative. What if our categories have numerical labels? The basic idea is that if doing arithmetic operations on those numerical labels does not generate numerical labels which are meaningful in the given context, then they are truly categorical and not quantitative. A quantitative variable represents a quantity or amount of something.

Consider the example of paint color. What if we give the colors numerical labels? This is still a categorical variable! If we call blue 1, red 2, and green 3, then these numbers are no longer quantitative in the sense that they are simply categorical labels of a qualitative observation. We do have that "$1 + 2 = 3$", but "red + blue = green" is not meaningful, furthermore "$2+3 = 5$" but it is not clear what color 5 should represent. We could easily have used different numerical labels for the colors without changing the underlying nature of the data.

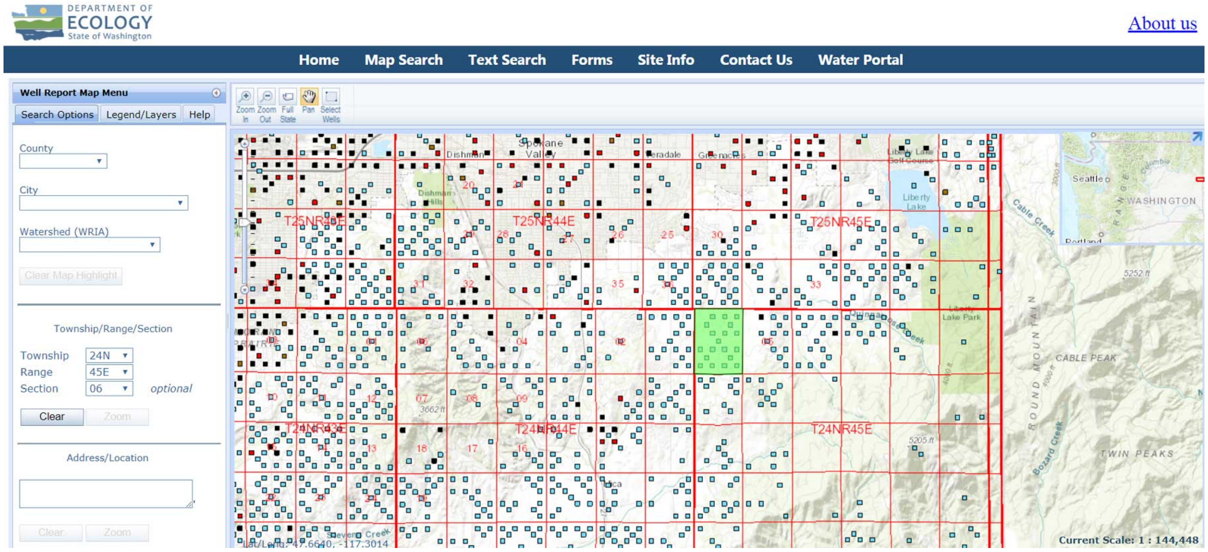
Consider calendar year as an example, e.g. year in which a historical event occurred. It is still categorical because the adding or subtracting two years generally will have no meaning within the given context. Consider two historical events, one in 900 CE (common era) and another in 800 CE. It is true that $900 + 800 = 1700$, but it is not clear how the fact of a historical event in 1700 CE is related to those in the past. If we are counting the number of siblings a person has, adding 2 siblings with 3 siblings to get 5 siblings has a very clear meaning and is thus truly quantitative.

2 Introduction to R and a preliminary dataset

Example dataset. Let's look at a dataset. Table 3 below shows all water well depths in a specific region in Spokane County, Washington (Grid location T24N-R45E-S06). Figure 1 shows the Department of Ecology website and location of the wells on a map.

115	107	114	465	250	220	690	280
520	100	450	200	80	500	120	375
575	700	80	640	640	680	260	100
300	160	440	165	660	120	560	180
110	65	120	86	360	100	300	165
500	600	600	260	300	620	480	440
240	120	420	175	585	240	460	200
200	430	482	290	290	620	660	440

Table 3: Well depths (ft) in Spokane County for grid location T24N-R45E-S06.



Township 24N
Range 45E
Section 06

Figure 1: Washington State Department of Ecology website showing the locations of water wells for a portion of Spokane County. Grid location T24N-R45E-S06 is highlighted.

Let's explore this data using R. First we need to import the data into R. There are several ways to go about doing this. I will be giving instructions for Windows. It may vary slightly if you are using another operating system.

2.1 Set Your Working Directory in R

First, we'll want to set our working directory in R. In RStudio you can do this from the main menu with:

Session → Set Working Directory → Choose Directory..

Otherwise we'll need to set it manually.

Let's say your directory is `C:\Users\myname\files\statsRwork`, then you will need to type:

```
> setwd("C:/Users/myname/files/statsRwork")
```

Notice that the slashes have reversed direction!!!

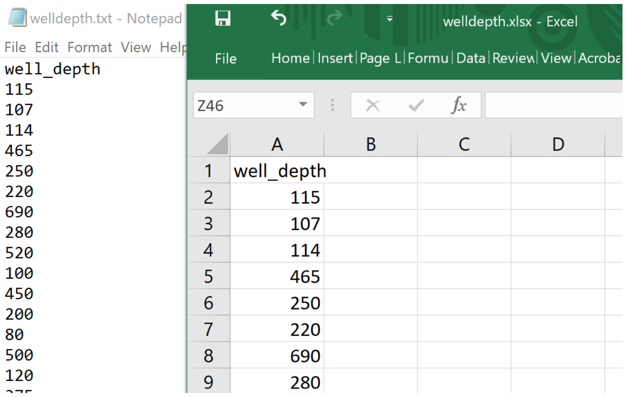
2.2 Import a Dataset into R

Import data from Excel or text. To import a single list of data into R, here is the table read method of importing data:

1. Open an Excel spreadsheet.
2. Put a name for your variable/data in the first cell (recommended text only, no spaces or special characters).
3. Type all data values in the column under the name.
4. Save as filetype "`Text (MS-Dos) (*.txt)`". Alternatively, open it in MS Notepad, and copy and paste the entire data column, including the name, into Notepad and save it as a *.txt file. Let's say we saved it as `filename.txt`.
5. Open R and execute the command:
   `> x=read.table("fileneame.txt", header=T)`

Alternatively, if the dataset is small enough that manually typing out the entire list is not a major fuss, we can just use the R command
`> x=c(x1,x2,x3,...,xn)`
where `x1,x2,x3,...,xn` is our list of actual numerical data values.

I have saved our list of water well depth data is `welldepth.txt`. Figure 2 shows how your data should look in Excel and Notepad.


welldepth.txt - Notepad
File Edit Format View Help
welldepth.xlsx - Excel
File
Home Insert Page L Formu Data Review View Acroba
Z46
fx

```
well_depth
115
107
114
465
250
220
690
280
520
100
450
200
80
500
120
```

	A	B	C	D
1	well_depth			
2	115			
3	107			
4	114			
5	465			
6	250			
7	220			
8	690			
9	280			

Figure 2: Water well depth data shown in an Excel spreadsheet column with label and in a Notepad text file.

So we can execute:
`> wd=read.table("welldepth.txt",header=T)` This will give our data a "table structure". To extract just the numerical data, we'll need to execute
`> x=wd$depthdata`

The name "depthdata" is the header name for our column of data. In R, the dollar symbol, $, is used to reference the actual data under a given name. A table of data could actually have multiple columns with each column given a unique name, e.g. a table named `y` with columns named `col1, col2, col3`. We can extract the data from the 2nd column by

```
> y$col2.
```

Import data by manually typing We can also import our well depth data directly by typing it all out separated by commas inside of a `c()` with:

```
> x=c(115,107,114,465,250,220,690,280,
      520,100,450,200,80,500,120,375,
      575,700,80,640,640,680,260,100,
      300,160,440,165,660,120,560,180,
      110,65,120,86,360,100,300,165,
      500,600,600,260,300,620,480,440,
      240,120,420,175,585,240,460,200,
      200,430,482,290,290,620,660,440)
```

> R tip: Note that in order to break lines without executing code, use a `Shift-Enter` (hold the shift key while pressing the Enter/Return key).

Import data from clipboard A very convenient way to import data is to highlight it on screen and copy it to your computer's clipboard memory. You can copy the data from a pdf, webpage, or spreadsheet. Make sure to know whether you are only selecting the data or including a header or name. If we highlight the list of data above in this pdf and copy it, we can import it with

```
> wd = read.table("clipboard")
```

This data will be imported as a "data frame" which is a specific data structure in R. If the data copied from the clipboard is simply a list of numbers for a single variable, then we will need to extract the numerical data from the data frame and turn it into a numerical list with the following command.

```
> x = as.numeric(matrix(as.matrix(wd),c(prod(dim(wd)),1)))
```

Now we have our list of well depths stored as `x`, and we are ready to analyze it with statistical techniques.

3 Graphical statistical methods

Now we will look at some graphical methods of analyzing a dataset. Graphical methods allow us to look at the entire dataset so-to-speak without actually seeing any of the numbers! They will help give us a feel for whether the data tends to cluster near a central value, if it is evenly spread out over its entire range, or if it has multiple clusters at different locations.

3.1 Stem-and-leaf plots

A stem and leaf plot consists of creating stems with the first digit of your data values, and the leaves are the remaining part of the data values, e.g. for the number 23, 2 is the stem and 3 is the leaf. This can be difficult if you have data with more than two digits or a variety of decimal places, etc.

Stem and leaf plots are useful in that they are a graphical method, but still retain some of the actual numerals in the data.

Here is the R command along with a stem plot for our water well data:

```
> stem(x)

  The decimal point is 2 digit(s) to the right of the |

  0 | 7889
  1 | 0001112222267788
  2 | 000244566899
  3 | 00068
  4 | 2344456788
  5 | 002689
  6 | 0022446689
  7 | 0
```

Notice that the `stem()` command in R rounds the data. The smallest number in the stem plot above is `0|7` with the decimal place 2 digits to the right of the "—" thus `0|7` is actually 70, but our smallest data value is actually 65. R will round all data points in order to make the leaves single digits.

If we want to make more or less leaves, we do so by setting the 'scale:'

```
> stem(x,scale=1)   (default)
> stem(x,scale=2)   (more leaves)
> stem(x,scale=0.5)   (less leaves)
```

Set the scale in powers of two: `scale` = $0.25, 0.5, 1, 2, 4$, etc.

3.2 Dotplots

A dotplot places a dot for each datapoint on a numberline, and successive dots for identical data points are stacked on top of each other. The dotplot command in R is a bit complicated. It is called a stacked stripchart, and we must specify to use the dot symbol with a "`pch=19`" option.

```
> stripchart(x,method="stack",pch=19)
```

3.3 Boxplots

A box plot is created from the five number summary. It is often called a box-and-whiskers plot as well. The box is created by the quartiles and median. The whiskers extend to the

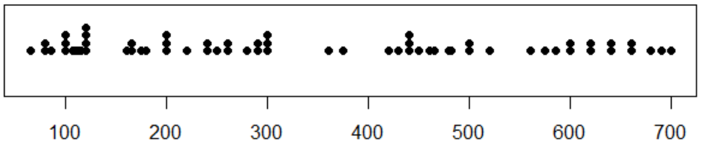


Figure 3: Dotplot of water well data.

last non-outlier data point in R. Outliers are displayed as dots beyond the whiskers.

Sometimes, depending on the source of the dotplot or the software package used to generate it, the whiskers may extend out to the minimum and maximum data values.

Here is how to generate boxplots in R.

```
> boxplot(x)    (this will generate a vertical boxplot)
> boxplot(x,horizontal=TRUE)    (this will generate a horizontal boxplot)
```

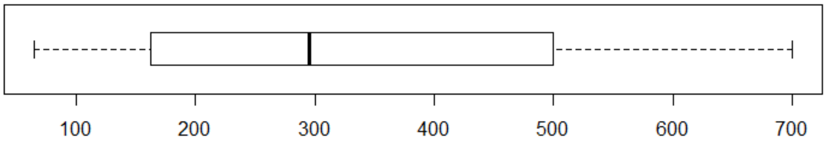


Figure 4: Boxplot of water well data.

3.4 Histograms

A histogram is probably one of the most common graphical displays of data. The data are grouped into 'bins' and the number of data points in each bin are counted. This creates a *frequency histogram*. There are three types of histograms. If we divide the frequencies by the sample size, then we create a *relative frequency histogram*. Relative frequencies sum to one. When we divide the relative frequencies by the width of the bin, then we create a *density histogram* which has total area one. The R histogram command makes frequency and density histograms only. A relative frequency histogram can be made, but it is a bit trickier.

frequency = number of data points in the bin   (histogram bar heights sum to sample size)

$$\text{relative frequency} = \frac{\text{frequency}}{\text{sample size}} \qquad \text{(histogram bar heights sum to one)}$$

$$\text{density} = \frac{\text{relative frequency}}{\text{bin width}} \qquad \text{(histogram bars have total area one)}$$

```
> hist(x)    (basic frequency histogram)
> hist(x,freq=FALSE)    (basic density histogram)
```

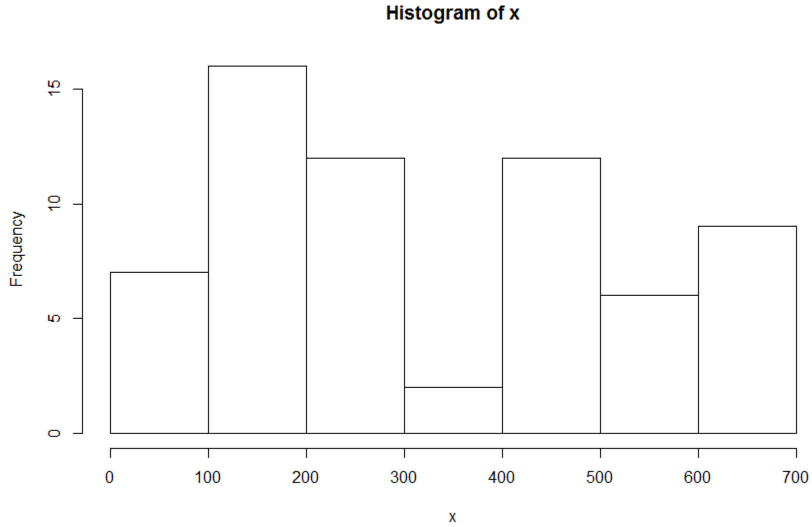


Figure 5: Frequency histogram of water well data.

It is worth knowing a bit about the structure of the R histogram object. Let's save the histogram under a name `wellhg` and then type that name and hist `Enter` to see all of the information and attributes it contains.

```
> wellhg = hist(x)
> wellhg
$‘breaks‘
[1]   0 100 200 300 400 500 600 700

$counts
[1]  7 16 12  2 12  6  9

$density
[1] 0.00109375 0.00250000 0.00187500 0.00031250 0.00187500 0.00093750 0.00140625

$mids
[1]  50 150 250 350 450 550 650

$xname
[1] "x"

$equidist
[1] TRUE

attr(,"class")
```

```
[1] "histogram"
```

Note that we can call the midpoints of our bins as

```
> wellhg$mids
[1]  50 150 250 350 450 550 650
```

We can similarly call the breaks (endpoints of bins) with `wellhg$breaks`, frequencies with `wellhg$counts`, or densities with `wellhg$density`.

Often, choosing the bins is a bit of an art. Too few bins can give a histogram that is misleading, and too many bins can give a histogram that is too jagged or doesn't sufficiently summarize the shape of the data. Figure 6 shows several different histograms for the same dataset. You can decide for yourself which is the best histogram, but I prefer the two in the middle column.

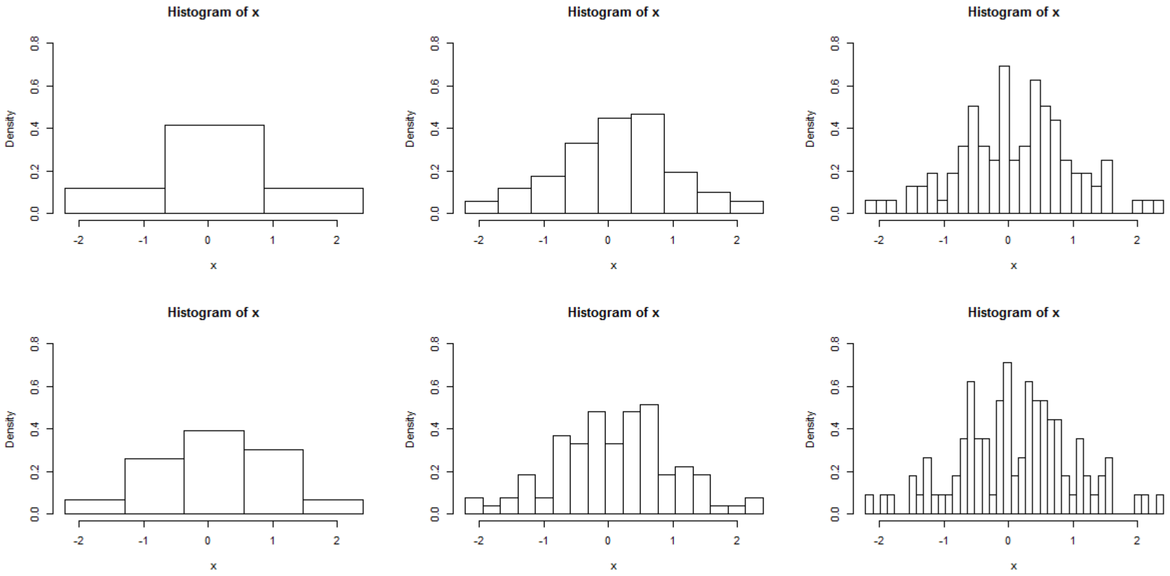


Figure 6: Histograms for the same dataset plotted with different bin sizes.

Here is the R code used to generate Figure 6:

```
> set.seed(1)
  x=rnorm(100)
  par(mfcol=c(2,3))
  nbins=3
  bins=seq(from=min(x),to=max(x),length.out=nbins+1)
  hist(x,breaks=bins,freq=FALSE,ylim=c(0,0.8))
  nbins=5
  bins=seq(from=min(x),to=max(x),length.out=nbins+1)
  hist(x,breaks=bins,freq=FALSE,ylim=c(0,0.8))
  nbins=9
  bins=seq(from=min(x),to=max(x),length.out=nbins+1)
  hist(x,breaks=bins,freq=FALSE,ylim=c(0,0.8))
  nbins=17
  bins=seq(from=min(x),to=max(x),length.out=nbins+1)
  hist(x,breaks=bins,freq=FALSE,ylim=c(0,0.8))
  nbins=29
  bins=seq(from=min(x),to=max(x),length.out=nbins+1)
  hist(x,breaks=bins,freq=FALSE,ylim=c(0,0.8))
  nbins=41
  bins=seq(from=min(x),to=max(x),length.out=nbins+1)
  hist(x,breaks=bins,freq=FALSE,ylim=c(0,0.8))
```

Here are a few different ways to create bins for histograms in R.
`> c(b0,b1,b2,...,bm)` (creates bins $[b_0, b_1], (b_1, b_2], (b_2, b_3], \ldots$)
`> seq(from=a,to=b,by=d)` (creates bins $[a, a + d], (a + d, a + 2d], \ldots$)
`> seq(from=a,to=b,length.out=m)` (creates m total bins from a to b)

3.5 Scatterplot for paired data

Often we will have a dataset with two variables that have a natural pairing. In addition to well depth, maybe we could also have the flow rate for each well. We might wish to investigate if there is a relationship between flow rate and well depth.

```
> plot(x,y)
```

4 Descriptive statistical methods

Descriptive Statistics. Now we will calculate some numerical summaries that will give us a feel for the "shape" of this dataset. Any data point or summary piece of information about a population or sample is called a *statistic*.

4.1 Measures of central tendency and location

The idea of *location* and *central tendency* entails that we want to understand where the data lies, e.g. on a numberline or within a categorical list, and what the typical data value is.

You are probably familiar with the concept of an "average" or "mean." This is often refereed to as a measure of central tendency and can be thought of as the central, typical, or expected data value. This is not the only possible way to describe the central tendency or typical data.

4.1.1 Mode

For categorical or discrete data, the *mode* is a useful measure of central tendency. The mode is the most common data point. Graphically, the mode will generally be where the highest peak on a histogram occurs.

Example. Consider the dataset: $\{1, 2, 3, 3, 3, 3, 3, 6, 7, 8, 8, 9\}$. There are five 3's, and this is the most common data value, thus 3 is the mode. If there were another data value that also occurred five times int he dataset, then the dataset would be called *bimodal* (two modes) or *multimodal* (three or more modes).

Generally, when a histogram has multiple peaks, even if they are different heights, we can call the data set multi-modal. Calling a data value a "mode" means that it has a higher frequency than nearby data values.

4.1.2 Order statistics

Consider dataset $\{x_1, x_2, \ldots, x_n\}$. This dataset may not necessarily be in increasing numerical order. We need to sort it, and thus create the *order statistics* $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ which is just the same dataset, but written from smallest to largest in numerical order, i.e. $x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$, $x_{(i)} \leq x_{(i+1)}$ for $i = 1, 2, \ldots, n-1$, and $x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$.

To sort a dataset in R and create the order statistics use the following command.

```
> sort(x)
```

R tip: If you have a dataset stored as `x`, you can overwrite `x` with its sorted list by doing:

```
> x = sort(x)
```

Generally, there will not be a reason to keep your dataset unordered thus you can replace it by its order statistics in this way.

4.1.3 Median

For quantitative data, the *median* is a useful measure of central tendency. The median can also be used for categorical data when the data has a natural order to it, e.g. letter grades or a ranked preference scale.

The median will be denoted with a "$\sim$" (referred to as a 'tilde') over a letter. The population median is denoted $\tilde{\mu}$ and the sample median is denoted $\tilde{x}$.

Once you have your dataset sorted in increasing order, to create the order statistics, you can calculate the sample median as

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even} \end{cases}$$

This gives the middle value of your sorted dataset if odd in length, or average of two middle values if even in length

```
> median(x)
```

R tip: Check your sample size with the following command.

```
> length(x)
```

R tip: Here is an alternative way to manually calculate the median:

```
> n = length(x)
> x[n/2]   (if n is even)
> x[(n+1)/2]   (if n is odd)
```

4.1.4 Mean

For quantitative data, the *mean* is another measure of central tendency. The mean is denoted by a 'bar' over a letter. The population mean is μ and the sample mean is denoted $\overline{x}$. The mean coincides with the common concept of 'average.' Note that there are multiple types of averages though! We are interested in the *arithmetic average* here.

The population mean is calculated by adding up all data values and dividing by the population size N:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The sample mean is calculated by adding up all data values and dividing by the sample size n:

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

This can be accomplished in R by:

```
> mean(x)
```

We can also calculate the mean by:

```
> sum(x)/length(x)
```

The mean is a statistic that is thought to give us some information about the sample it was calculated from, and we might even believe it gives us some information about the underlying population from which the sample was taken. In this sense, the mode, median, and mean are all "summary descriptive statistics." They are statistics that describe the central tendency of the data, or the most likely or most common observation.

4.1.5 Quartiles

Quartiles are also a measure of location. They are not generally thought of as describing central tendency. They split the data into chunks of 25% by count, e.g. a dataset of 12 points would be split into groups of 3 by the quartiles.

The median is actually the second quartile. The first quartile is the median of the first half of the data and the third quartile is the median of the last half of data. If n is odd, I prefer to exclude the median and use the first and last $(n-1)/2$ data points for calculating the quartiles. It is a valid technique to include the median as well. Don't get too hung up on which method to use, just pick one and stick with it.

```
> q1 = quantile(x,0.25)
> q3 = quantile(x,0.75)
```

Note that there are 9 different ways to calculate quartiles (or more generally, quantiles) in R, the default is `type=7`. Try `q1 = quantile(x,0.25,type=2)` for example.

Normally, the quickest way to get the above summary statistics is with the `summary()` command. It gives the minimum, maximum, mean, median, and quartiles.

```
> summary(x)
```

The minimum, 1^{st} quartile, median, 3^{rd} quartile, and maximum are referred to as the *five number summary* and can be calculated also by:

`> fivenum(x)` (similar to `summary()` but only gives a list of min, q1, median, q3, max)

4.1.6 Quantiles

These are a sort of generalization of quartiles. Instead of splitting the dataset up into 4 parts, each containing 25% of the data, we can split it up into any number of sections. Quintiles for 5 sections at 20% each (this is actually quite common in economic data such as household income), deciles for 10 sections, percentiles for 100 sections. The k-quantiles split the data up into k sections, i.e. $k = 4$ for quartiles, $k = 5$ for quintiles, $k = 10$ for deciles, and $k = 100$ for percentiles.

To calculate the j^{th} k-quantile in R:

`> quantile(x,j/k)` (e.g. $j = 3$, $k = 4$ for the third quartile)

> R tip: There are nine methods for calculating quantiles in R, by default, "quantile type 7" is used. We can try them all by specifying `quantile(x,q,type=i)` for $i = 1, 2, ..., 9$. Don't worry too much about the quantile type, just use the default when in doubt.

For a population the $q(100)^{th}$ percentile is also called the q quantile and is denoted:

$$\tilde{\mu}_q = \{\text{a value for which } q(100)\% \text{ of the population is less than or equal to and} \\ (1-q)(100)\% \text{ of the population is greater than or equal to}\}$$

Sample quantiles, v.1 To get sample quantiles, it is a bit more flexible. The simplest definition of sample quantile is that the i/n quantile is the i^{th} order statistic

$$\tilde{x}_{i/n} = x_{(i)}.$$

If we have order statistics $x_{(i-1)} = a$, $x_{(i)} = b$, and $x_{(i+1)} = c$ for a dataset of size n, then any weighted average between a, b, and c is a reasonable value for $\tilde{x}_q$ the q sample quantile when

$$\max\left(0, \frac{i-1}{n}\right) \leq q \leq \min\left(1, \frac{i+1}{n}\right)$$

with $\tilde{x}_0 = x_{(1)}$ and $\tilde{x}_1 = x_{(n)}$.

Sample quantiles, v.2 Another alternative for calculating sample quantiles is

$$\tilde{x}_{\frac{i}{n+1}} = x_{(i)}$$

or more generally $\tilde{x}_q = x_{(i)}$ for

$$\max\left(0, \frac{i-1}{n+1}\right) \leq q \leq \min\left(1, \frac{i+1}{n+1}\right)$$

with $\tilde{x}_0 = x_{(1)}$ and $\tilde{x}_1 = x_{(n)}$. Notice this is the same as the previous method but using $n+1$ instead of n in the denominator. This is because, n data points actually split the number line into $n+1$ regions.

The reason for the different methods of calculating quantiles for a sample is that we are actually trying to *estimate* the population quantiles from the sample data. This is going to be a theme in this course and is one of the main goals of the science of statistics: to estimate properties of a population from a sample dataset.

Sample quantiles, v.3 To find the q sample quantile, just multiply q by $n+1$ and then interpolate between those data points. E.g. if we have a dataset of size $n = 13$ and we want to find the 68^{th} percentile (the 0.68 quantile), then we multiply $14 \times 0.68 = 9.52$. So we can choose the 9^{th} or 10^{th} order statistics or anything between them. Note that $13 \times 0.68 = 8.84$ so that it is also reasonable to choose the 8^{th} order statistic. So any weighted average between the order statistics 8 to 10 will be a reasonable estimate.

Example. Consider the (unsorted) dataset: $\{6, 2, 3, 9, 3, 3, 1, 8, 8, 7, 3, 3\}$. There are 12 data points. Thus we will make our quantiles in increments of $1/12 = 8.\bar{3}\%$. But we need to sort this list first.

```
> x=c(6,2,3,9,3,3,1,8,8,7,3,3)
  sort(x)

  [1] 1 2 3 3 3 3 3 6 7 8 8 9
```

$\tilde{x}_{0.08\bar{3}} = 1$, $\tilde{x}_{0.16\bar{6}} = 2$, ..., and $\tilde{x}_1 = 9$. This is using `quantile(x,q,type=1)` in R.

`quantile(x,0.95)` gives $\tilde{x}_{0.95} = 8.45$.
(This is identical to `quantile(x,0.95,type=7)` by default.)

`quantile(x,0.95,type=1)` gives $\tilde{x}_{0.95} = 9$
`quantile(x,0.95,type=3)` gives $\tilde{x}_{0.95} = 8$
`quantile(x,0.95,type=4)` gives $\tilde{x}_{0.95} = 8.4$
`quantile(x,0.95,type=5)` gives $\tilde{x}_{0.95} = 8.9$

Since 95% falls between $\frac{11}{12}$ and $\frac{12}{12}$, anything between 8 and 9 is a reasonable estimate of the 95^{th} percentile also known as the 0.95 quantile.

Using the definition of population quantiles we get the following.

$$\tilde{x}_q = 1 \text{ for } 0 \leq q \leq \tfrac{1}{12}$$

$$\tilde{x}_q = 2 \text{ for } \tfrac{1}{12} \leq q \leq \tfrac{2}{12}$$
$$\tilde{x}_q = 3 \text{ for } \tfrac{2}{12} \leq q \leq \tfrac{7}{12}$$
$$\tilde{x}_q = 6 \text{ for } \tfrac{7}{12} \leq q \leq \tfrac{8}{12}$$
$$\tilde{x}_q = 7 \text{ for } \tfrac{8}{12} \leq q \leq \tfrac{9}{12}$$
$$\tilde{x}_q = 8 \text{ for } \tfrac{9}{12} \leq q \leq \tfrac{11}{12}$$
$$\tilde{x}_q = 9 \text{ for } \tfrac{11}{12} \leq q \leq 1$$

Note that some of the quantile calculation methods will vary from this slightly. If the dataset is not actually our entire population, then we don't really know what the population quantiles are, so some of the more complicated methods for calculating quantiles are trying to estimate what the population quantile is using sample data. So it gives a number between the sample data values.

4.1.7 Trimmed mean

If the median and mean of a dataset are very different, that is an indicator that there may be some outliers or a significant asymmetry in the way your data is distributed over the set of all possible values. Sometimes trimming off extreme data values on the upper and lower end, and then calculating summary statistics can be useful.

A α% trimmed mean is calculated be removing the α% smallest and α% largest data values then calculating the mean of the resulting reduced-size dataset and is denoted $\overline{x}_{tr(\alpha)}$.

Example: Consider the dataset $\{1, 1, 2, 4, 5, 8, 9, 10, 20, 30\}$ with mean $\overline{x} = 9$ and median $\tilde{x} = 6.5$. Since we have exactly 10 data points, a 10% trimmed mean would remove the largest and smallest point to get $\overline{x}_{tr(10)} = 7.375$, and a 20% trimmed mean would remove the two largest and two smallest values to get $\overline{x}_{tr(20)} = 6.333$. These are closer to the median.

An α% trimmed mean can be calculated in Ras:
`> mean(x,trim=`α`/100)` (e.g. a 10% trimmed mean is `mean(x,trim=0.1)`)

4.2 Measures of variability and spread

Calculating a mean or median does give you some information about a dataset, but in any interesting case, the data will not all be identical and will vary. Some data will fall to one side or the other any of the measures of center or location. It is also important to understand how data varies! There are several ways to characterize how data is spread out and varies over the set of all possible values.

4.2.1 Range

The difference between the maximum and minimum data value is one way to understand the spread of the data and is called the *range*:

```
> min(x)
> max(x)
> range(x)   (this give both the minimum and maximum, not their difference)
> datarange = max(x)-min(x)   (or alternatively: datarange = range(x)[2]-range(x)[1])
```

4.2.2 Interquartile range

The interquartile range is defined as the difference between the third and first quartiles $IQR = Q_3 - Q_1$ and is also a measure of the spread of the data. It is generally better than the range as a measure of spread since it will not be as affected by extremely large or small data values.

```
> IQR(x)
```

Similarly, as a measure of spread, you can take the difference between various quantiles.

4.2.3 Variance and standard deviation

Next to the mean and median, probably the most common descriptive statistics are the *variance*and *standard deviation*. These are both measures of how the data varies. The population variance is denoted σ^2 and the population standard deviation is the square root of variance σ. Sample variance is denoted s^2, and sample standard deviation is denoted $s = \sqrt{s^2}$.

Population variance is calculated by the formula

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

where N is the size of the population and $\mu$the population mean.

Sample variance is calculated by the formula

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

In R the sample variance and standard deviation can be calculated by:

```
> var(x)   (variance)
> sd(x)   (standard deviation)
```

The variance can be thought of as the *mean squared deviation (from the mean)*. If we really are interested in the average distance from the mean, then we can take the square root of the mean squared distance and thus arrive at the standard deviation.

Why the mean squared deviation? If we were to just take the average distance from the mean, we would get zero! As an exercise, you can try to show that $0 = \sum_{i=1}^{n}(x_i - \overline{x})$. We could instead take mean absolute deviation from the mean $|x_i - \overline{x}|$, but the absolute value function can be difficult to work with. These are two reasons why mean square deviation has become the standard.

Why divide by $n-1$? It turns out that the sample variance formula above is an "unbiased estimate" of the population variance. This is an advanced technical concept, but here is a way to understand it. If we really wanted to understand what the population variance was and are going to use the sample variance to estimate it, then by dividing by n in the sample variance formula instead of $n-1$ will generally give us an underestimate of the population variance.

What is so special about standard deviation? It will appear in a variety of statistical methods later on. Generally, the vast majority of your data will be within three standard deviations of the mean. This is not an absolute rule, but you will find that it is often true.

5 Summary

R commands:

Descriptive statistics:

```
mean(x)
median(x)
min(x)
max(x)
range(x)
length(x)
dim(x)
sort(x)
var(x)
sd(x)
IQR(x)
summary(x)
fivenum(x)
quantile(x)
quantile(x,p)
```

Import data:

```
x=c(x1,x2,...,xn)
x=read.table(‘‘filename.txt’’,header=T/F)
x=read.table(‘‘clipboard’’,header=T/F)
x=read.csv(‘‘filename.csv’’,header=T/F)
```

Graphical methods:

```
hist(x)
boxplot(x)
hist(x,breaks=seq(from=a,to=b,length.out=m),freq=T/F)
hist(x,breaks=seq(from=a,to=b,by=d),prob=T/F)
plot(x,y)
plot(ecdf(x))
```

Notation and formulas:

Dataset: $\{x_1, x_2, \ldots, x_n\}$ Sorted dataset, order statistics: $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$

Population mean: $\mu = \dfrac{1}{N}\sum_{i=1}^{N} x_i$ Population variance: $\sigma^2 = \dfrac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Sample mean: $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ Sample variance: $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Means and standard deviations can have subscripts on them to indicate certain information:
μ_1, σ_1^2, $\overline{x}_1$, s_1^2 (for population 1), μ_2, σ_2^2, $\overline{x}_2$, s_2^2 (for population 2), etc.
μ_X, σ_X^2, s_X^2 (for variable X), μ_Y, σ_Y^2, s_Y^2 (for variable Y), etc.
$\overline{x}_n$, s_n^2 (for sample size n), $\overline{x}_m$, s_m^2 (for sample size m), etc.

Median; population: $\tilde{\mu}$ or $\tilde{\mu}_{0.5}$ sample: $\tilde{x}$ or $\tilde{x}_{0.5}$

$q(100)^{th}$ percentile or q quantile: population: $\tilde{\mu}_q$ sample: $\tilde{x}_q$

Random variable, unknown value: X (capital letter)
Random variable, specific value: x (lower case letter)

Discrete variable on takes on specific values
Continuous variable takes any value in an interval.

Categorical variable: Generally mathematical operations are not meaningful
Quantitative variable: mathematical operations are meaningful