Analyzing Data from Distributions: Types of Data, Measures of Central Tendency, Variances

Joseph J. Luczkovich

January 28, 2015

Abstract

Populations of people, animals, or objects are normally too big to sample completely. In most cases, we are satisfied with a sample of the measured population. We choose samples at random from the population, if possible, to avoid any bias in the estimates made based on the sample. After drawing a sample, we compute some measure of central tendency (means, median, mode) and report some measure of variability (confidence intervals, variance, standard deviation, or range). Finally, we often want to know if there is a difference in the central tendency of two or more populations, given the variability observed. In this lesson, we will learn how to measure the central tendency,

1 Introduction

We rarely know the true mean (μ) and variance (σ^2) of a measurement from the whole population (this is called a census, but this is difficult to achieve). Instead, we take a sample of the population (as large as possible) and use that to estimate the true mean and variance. Samples will vary in their means and dispersion around the true mean. We compute the sample mean ($\bar{x}$), sample variance (S^2), sample standard deviation (S) or another statistic from the sample to draw conclusions about the population. The sample mean is an estimator of the central tendency of the population measurement. The sample standard deviation (S) and variance (S^2) are measures of the variability around the mean. Note that variance(S^2) is the squared standard deviation (S), and the $\sqrt{variance}$ = standard deviation.

2 Types of data

Data can mean many things to different people. It comes from the Latin word, datum (plural: data), which means a mark or reference, something given as a fact. It can mean, in current usage, facts, statistics, or pieces of information. Mostly, when scientists speak of data, they mean quantified values (numbers)

that relate to the questions being investigated, however it broadly can mean any kind of information, even photos or videos.

Types of data:

1. **Continuous variables, Interval scale**, with an arbitrary zero. Example: temperatures in Celsius or Fahrenheit.
2. **Continuous variables, Ratio scale**, with a fixed, non-arbitrary zero. Example: weight in grams. Temperature in Kelvin (absolute zero)
3. **Counts, integers only**. Example no. eggs/fish.
4. **Ordinal data**, or ranks. Example: data are ranked as 1st, 2nd, 3rd, 4th, 5th highest to lowest, or lowest to highest.
5. **Nominal data**. Example: male or female. Qualities that can be coded. Hair color. Ethnic groups. Presences or absences (0,1). Qualitative data.
6. **Text data**: strings, character data (qwert1234). DNA sequence data are an example of data that are not numbers, but letters that represent a nucleotide sequence:

Clam *Macoma* Primers	Sequence
Forward primer	GCACAGAGTTAATACATCCTGGC
Reverse primer	AGGACGCATATTAGCACCTGTAG

3 Statistical Distributions

Data can be distributed in many different ways. Some data are normally distributed, following a bell-shaped frequency or probability curve, with the mean or average value being in the center, and data falling equally above and below the mean. Formally, there are many approximations of a normal curve, which are referred to as Gaussian, Beta, Gamma, Fisher's F-distribution, Student's t-distribution. Other common distributions include the binomial (as we used coin- flipping data in HW 1), the Poisson, the Bernoulli, chi-square, log-normal and many others. In R, you can simulate each of these distributions. Below, I will generate the normal distribution with a mean of 0.0 and a SD = 1.0 (the so-called standard normal curve):

```
> #create a vector of data from -3 to 3
> q<-seq(-3,3, by = .10)
> #get the normal probability distribution (mu = 0, sd = 1) for each datum,
> # and store results in new vector
> norm<-dnorm(q)
> # Plot data by probability of drawing that number from a normal distribution
> plot(q,norm,type="b")
```

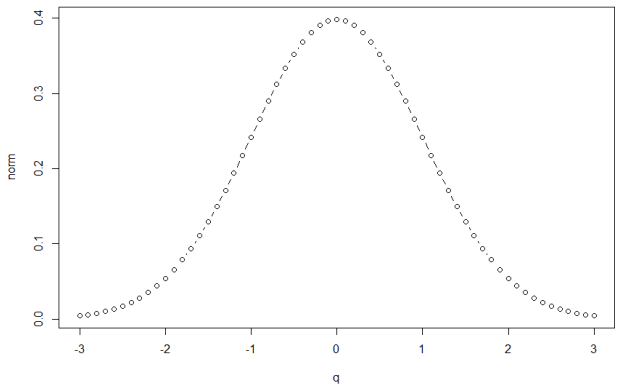

Figure 1: A plot of the normal probability distribution (with a mean of 0.0 and Standard Deviation of 1.0) versus a data set of -3.0 to 3.0

As we can easily see, the most common number drawn is 0. That is called the **mode** (the most common number in the data set). In the Normal distribution, that number is also called the **mean** and the **median**; they are all 0 in this case. The mean is the same as an average, you sum the numbers and divide by the number of values you summed. In R, this is found by typing **mean(x)**. The median is number that lies half-way between the upper tail and lower tail of the distribution (50th percentile). To find it, start dropping the highest and lowest values until you get to the middle value or two middle values (average the two middle values and you get the median). In R, the command is **median(x)**.

You can add a given mean and standard deviation of your choice to create a normal distribution curve that is similar to some data you may have (first find the mean and standard deviation by using the **mean(x)** and **sd(x)** of your data in x). The NHANES we will examine for HW 2 below has the following mean and SD for Waist Circumference (BMXWAIST) of a sample of people across the USA:

```
> library("foreign", lib.loc="C:/Program Files/R/R-3.0.2/library")
> demo<-read.xport("DEMO_G.XPT")
> bmi<-read.xport("BMX_G.XPT")
> m1<-merge(demo,bmi)
> attach(m1)
> mean(BMXWAIST,na.rm=T)

[1] 86.22398

> sd(BMXWAIST,na.rm=T)

[1] 22.36524
```

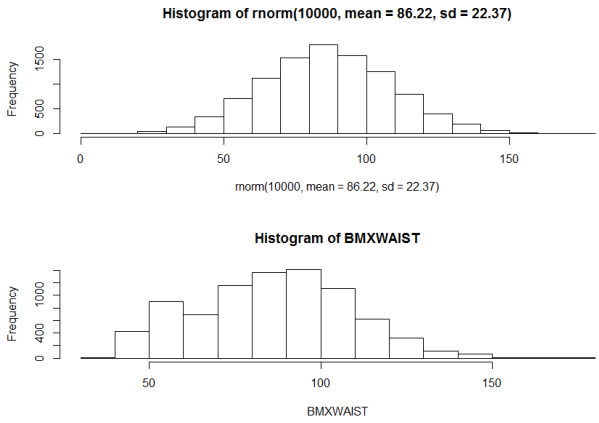

Figure 2: The histograms of a normal distribution (top) with a mean and standard deviation the same as the NHANES waist circumference data (bottom) from a sample of 9000 people in 2011-12 in the USA.

Now let's plot a histogram of the BMXWAIST data and compare it with a histogram of 10000 randomly drawn samples from a normal distribution with the same mean and SD:

```
> par(mfrow=c(2,1))
> hist(rnorm(10000,mean=86.22, sd=22.37))
> hist(BMXWAIST)
```

As we can see in Figure 2, the actual waist circumference data are similar to the normal distribution, but there are differences in these two distributions, the normal distribution at the top is equally distributed around the mean value, but the actual data are skewed to the left, means there are more people with waist circumferences below the mean, but a few large values.

Large values can make the mean very high, sometimes. Highly skewed data may make the mean a poor estimator of the central tendency of a distribution. An example of a highly skewed data set is the benthic invertebrates sampled from Chesapeake Bay in 2011. The following data was obtained from the US Army Corp of Engineers, and included all invertebrates (worms, snails, isopods, amphipods, etc.) found in core samples (VALUE=number of benthic invertebrates of each species/sample).

```
>  Ace2011Abundance <- read.csv("~/CRM7008/Ace2011Abundance.csv")
> View(Ace2011Abundance)
> attach(Ace2011Abundance)
> names(Ace2011Abundance)

 [1] "STATION"       "SAMPLE_DATE"   "SAMPLE_NUMBER" "SPEC_CODE"
 [5] "LBL"           "TSN"           "PARAMETER"     "VALUE"
```

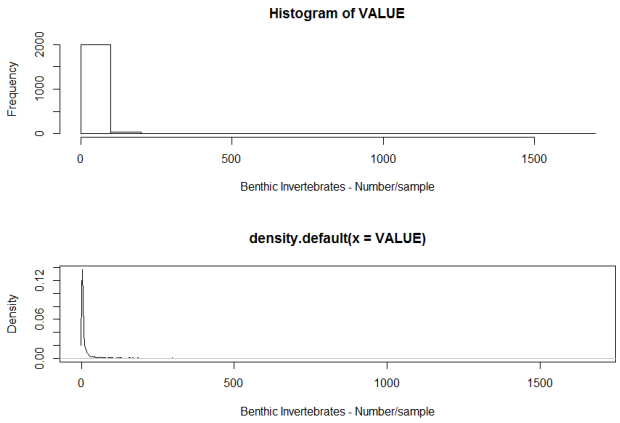

Figure 3: The histogram of the number of invertebrates of all species per sample in the Army Corps of Engineers data set (top) and a smoother **density( )** applied to the data set.

```
 [9] "UNITS"          "PARTITON"       "SOURCE"         "GMETHOD"
[13] "NET_MESH"       "CRUISENO"       "STRATUM"        "SITE"
[17] "SITE_TYPE"

> summary(VALUE)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       1       3      13       8    1678
```

The **summary()** command provides a summary of the basic stats for the data (minimum, maximum, first quartile, median, mean, and third quartile). You will see that the mean is 13 inverts/sample, but the median is 3 inverts/sample. Both are measures of central tendency, so why are they so different? It is because the underlying distribution is non-normal. Lets plot a histogram of it, then use a smoother to create the probability density curve for the data, using the **density()** command (Figure 3).

```
> hist(VALUE, xlab="Benthic Invertebrates - Number/sample")
> plot(density(VALUE), xlab="Benthic Invertebrates - Number/sample")
```

The distribution is skewed right, that is, there are large values (the tail of the distribution is right), and there are lots of zeros. In the histogram and density plot, this is reflected in the peak at 0-1 inverts/sample. This is simply a product of the benthic organisms' patterns of settlement and survival. One sample has 1681 individuals, but you can hardly see that in the distribution plots. The data need to be transformed (log transformation and others will be discussed in a later chapter) prior to data analysis for most statistics that assume a normal

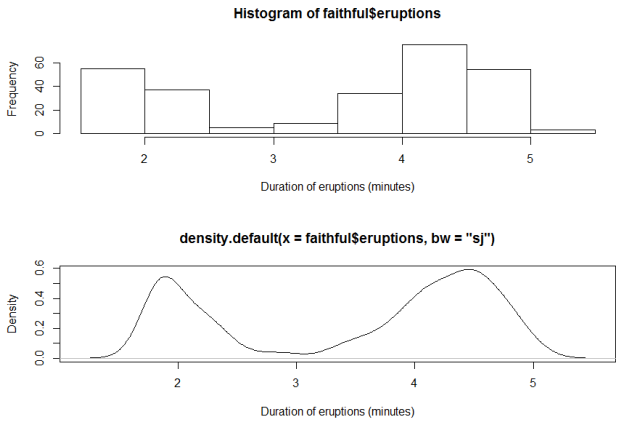

Figure 4: Bimodal distribution of the Old Faithful geyser's eruptions, showing the duration in minutes, at Yellowstone National Park. The distribution is bi-modal.

distribution. This kind of distribution is all-too-common in many sciences like ecology.

Some data are bimodal. Take the Old Faithful eruptions duration data available in R. This data set is described by typing **help(faithful)**. The histogram and plot of the data set looks like this (Figure 4):

```
> hist(faithful$eruptions, xlab="Duration of eruptions (minutes)")
> d <- density(faithful$eruptions, bw = "sj")
> plot(d, xlab="Duration of eruptions (minutes)")
```

There is no central tendency in these eruption duration data. The geyser will erupt and stay erupting either for 2 minutes or 4-5 minutes, but there is a low probability (close to zero) it will erupt for 3 minutes. The distribution for waiting times between eruptions (the other variable in the faithful data set) is also bimodal. See if you can figure out how long people should wait to see the next eruption by analyzing the waiting time data.

4 Student's t-distribution and the t-test

This distribution is closely related to the standard normal distribution. It is symmetric and bell-shaped like the standard normal, and has a mean = 0, but it has a slightly larger standard deviation. The exact shape of the t-distribution depends on a parameter called degrees of freedom (df) which is related to sample size. The parameter df, degrees of freedom, will be the minimum number of measurements needed to specify completely the behavior of the system. This in analogous to the idea that if you have 4 numbers (a,b,c,d) that when summed

Figure 5: William Sealy Gosset, known as "Student", a statisticain for the Guinness Brewery in Dublin.

must equal m, then if you choose 3 numbers, the 4th number is no longer free to vary, but must equal a single number: $d = m - \Sigma(a, b, c)$. One degree of freedom is lost every time a comparison is made between two group means, because the mean is computed. The equation $df = n - 1$ is used for the t-distribution. The df is always 1 less than the N in a sample in a group.

"Student" and the t-distribution William Sealy Gossett (born: June 13, 1876, died: October 16, 1937) was a brewer with Guinness Beer and famous statistician. He published the t-distribution under the name "Student", because Guinness would not allow him to use his real name to protect the trade secret he developed: the t-test. He worked in Dublin, Ireland on brewery quality control problems involving small samples taken from large batches of beer or barley. He needed to understand the variations in the process of making beer without sampling the whole vat. He used small samples of beer, measuring alcohol

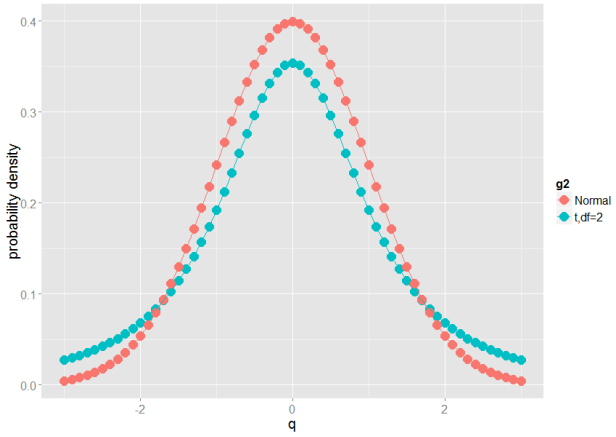

Figure 6: t-distribution (df=2) versus the normal curve

content at different stages of fermentation and a hand-cranked calculator to compute the t-distribution for Guinness, unavailable to others working at that time. While working at Guinness, he developed a new test statistic for the comparison of means based on small sample sizes and called the t-test:

$$t = (\bar{x} - \mu)\frac{\sqrt{n}}{S} \tag{1}$$

The sample mean $\bar{x}$ and the true mean μ where related to the sample standard deviation S and the square root of the size of the sample n. This was a major theoretical advance in both statistics and beer-making.

Lets look at the t-distribution and compare it with the normal. It is only slightly different, but the differences are most noticeable at low sample sizes (low df). Lets create three distributions, the $t_{df=2}$, $t_{df=10}$, and the standard normal for our vector of quantiles, q.

```
> q<-seq(-3,3, by = .10)
> t2=dt(q,df=2)
> t10=dt(q,df=10)
> norm<-dnorm(q)
```

Then, let's plot them together on the same plot (This is harder, and I did this in ggplot2, but I won't say how I did it here). Notice that the t-distribution with df=2 is pretty different, with fatter tails and lower probability density at 0, than the normal curve (Figure 6). But, when df=10, it is much more similar to the normal (Figure 7). With a larger sample (df=100) they are essentially identical. The t-distribution was developed for use with small samples in mind, and it shows. Like the normal it is symmetrical, and you can specify other means and sd's. Type **help(dt)** to see the ways to change these parameters.

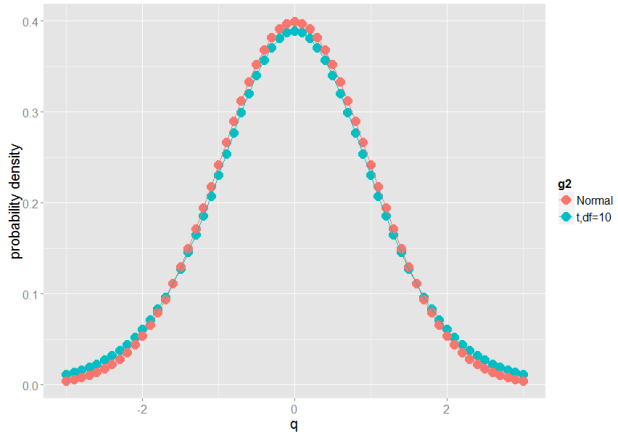

Figure 7: t-distribution (df=10) versus the normal curve

5 the t-test in R with an example

To implement a t-test in R, if you have two groups of samples, is very straight forward. Note that in the ordignal t-test equation, the test was comparing the sample mean to the true mean. We can also compare the means from two samples and test wether they differ from each other (their means should not differ significantly). Let's say we have alcohol contents for two batches of beer that were prepared using different fields of barley seeds. Does the alcohol content differ in these batches of beer?

```
> field1<-c(7.304180,  6.863088,  7.589897,  8.710749,  8.156218,
+           6.725056, 10.374051,  8.385567,  6.328037,  7.093926)
> field2<-c(10.378293,  9.550081,  8.671654, 10.518978,  9.770098,
+           8.865499,  9.518485, 10.230535, 10.342064, 11.128222)
> summary (field1)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.328   6.921   7.447   7.753   8.328  10.370

> summary (field2)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.672   9.526  10.000   9.897  10.370  11.130

> t.test(field1,field2)

        Welch Two Sample t-test

data:  field1 and field2
t = -4.771, df = 15.321, p-value = 0.0002339
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.100547 -1.188081
sample estimates:
mean of x mean of y
 7.753077  9.897391
```

The results tell us that the mean alcohol content from field 1 (7.753 % abv) is significantly lower than the alcohol content in the batch from field 2 (9.897 %abv), perhaps due to the amount of sugar or starch in the orginal barley used. The null hypothesis was that the two fields produced Guinness with the same alcohol content (difference in means = 0, $\mu = 0$). The alternative hypothesis is that the means differ in mean alcohol content. We conclude that they do differ, and with a probability of 0.0002339 that they are different due to chance alone (random factors). This is a very small probablity that they differ by chance, so we conclude the result is real, the two fields differ, field 2 is higher than field 1.

We can also test if the two fields differ from the ideal alcohol content for Guinness, which is 7.5 % abv. Specifying the the parameter μ allows us to do that.

```
> t.test(field1,mu=7.5)

        One Sample t-test

data:  field1
t = 0.6687, df = 9, p-value = 0.5205
alternative hypothesis: true mean is not equal to 7.5
95 percent confidence interval:
 6.896921 8.609233
sample estimates:
mean of x
 7.753077

> t.test(field2,mu=7.5)

        One Sample t-test

data:  field2
t = 9.8895, df = 9, p-value = 3.925e-06
alternative hypothesis: true mean is not equal to 7.5
95 percent confidence interval:
  9.349006 10.445776
sample estimates:
mean of x
 9.897391
```

Which field produced beer closest to the ideal Guinness alcohol content?

6 Homework 2

NHANES BMI data: Load the data set from the Blackboard data files folder. It is from the US Center for Disease Control NHANES (National Health and Nutritional Examination Survey) data for 2011-12, on US population Body Mass Index (BMI) measurements (BMX_E.xpt) which was obtained here:

url=http://wwwn.cdc.gov/nchs/nhanes/search/nhanes11_12.aspx

Choose menu options download the demographic and examinations data (on separate links). These data sets were exported from another common statistical program called SAS, and we will import it into R. To do this, we will need to download an R conversion package.

1. Open R, change to your default directory to whatever you use for HW 2
2. Under the menu **Packages**, choose **Load Package**
3. Select the package **foreign** from the list and click OK. (in R Studio, just click on the **foreign** box under the **Packages** tab in the lower right hand window. If **foreign** is not listed there, enter **foreign** in the search box (upper right of the window) and you will be sent to a mirror site where the package can be downloaded.
4. You have now downloaded a set of conversions for importing data from other statistical programs like SAS and SYSTAT
5. Use the command bmi<-read.xport("file=BMX_G.xpt") to read the data into an R object called **bmi** in your default directory. Make sure the NHANES data set is in the default directory before importing.
6. Use the command names(bmi) to see the names of the columns in the data. You can read the description of each one here:

   url=http://wwwn.cdc.gov/nchs/nhanes/2011-2012/BMX_G.htm)

   We wish to use the variables for weight (in kg), height (in cm) and BMI: BMXWT, BMXHT, and BMXBMI.

```
> library(foreign)

> bmi<-read.xport("BMX_G.XPT")
> names(bmi)

 [1] "SEQN"     "BMDSTATS" "BMXWT"    "BMIWT"    "BMXRECUM" "BMIRECUM"
 [7] "BMXHEAD"  "BMIHEAD"  "BMXHT"    "BMIHT"    "BMXBMI"   "BMDBMIC"
[13] "BMXLEG"   "BMILEG"   "BMXARML"  "BMIARML"  "BMXARMC"  "BMIARMC"
[19] "BMXWAIST" "BMIWAIST" "BMXSAD1"  "BMXSAD2"  "BMXSAD3"  "BMXSAD4"
[25] "BMDSADCM"
```

7. Plot weight versus height: plot(bmi$BMXHT,bmi$BMXWT). Examine the plot. How does the variation in weight of people in the US change with their height? Is there a constant variation, or does it change as people in the US grow?

8. Load the demographic data in SAS format from NHANES (DEMO_e.xpt) in the same way as the bmi data, creating a new object called"demo": demo<-read.xport(file="DEMO_G.xpt"). These demographic variables are described here: http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/vardemo_e.htm

9. Merge the data into a new R object: m1<-merge(demo,bmi). This command will use the common variable SEQN or sequence reference number to merge the data sets "demo" and "bmi". The SEQN variable keeps track of subjects by a code assigned to each person who participated in the NHANES study.

10. Using the NHANES data in the merged data structure "m1", compute the mean, median and standard deviation of the bmi for males and females in the USA using the formulas from class lecture (i.e., for mean, sum the values and divide by N; if you need help, see pages 5-6 in Dalgaard's text). To get just the males (code = 1), we first perform the command: malebmi<-m1$BMXBMI[m1$RIAGENDR==1] This get all the males' BMI data femalebmi<-m1$BMXBMI[m1$RIAGENDR==2]this gets all the females' BMI data Compare the answers you computed to those obtained using the built-in R functions mean (bmi, na.rm=TRUE), median (bmi, na.rm=TRUE), and sd(bmi, na.rm=TRUE). They should agree. Report your calculations. Use these commands to do it for the males: xbarbmi<-sum(malebmi,na.rm=T) dev<-malebmi-xbarbmi sqdev<-dev^2 $SDbmi <$ $-sqrt(sum(sqdev, na.rm = T)/(length(malebmi) - 437))$

Note that there are "NA" values in the NHANES data set. To find out how many, type summary(malebmi). You will have to account for these when computing the mean, standard deviation, and median. What does "N" mean? Why were they recorded as NA by the NHANES data collectors, rather than"0"?

11. Test the hypothesis that, over all ages, males and females have significantly different BM's in 2007-2008. One way to do that is to use the command use (for two-sided alternative, unpaired t-test):

12. test(malebmi,femalebmi) use this when the data for each group are in separate vectors or data structures or t.test(m1$BMXBMI m1$RIAGENDR). use this when the data for each group are in the same vector or column, but with a code for the groups (here gender is coded in m1$RIAGENDR) Interpret the results in light of the hypothesis above. Do you accept or reject the hypothesis? l. Now test the hypothesis that the mean BMI for all us citizens differs from the "true" population mean value of 22. Use a one-sample t-test: t.test(m1$BMXBMI,mu=22)