Math 247: Exploration 2.1

Warren Atkison

Instructions

This Exploration must be produced with an RMarkdown document. Make sure to include your name. While you are encouraged to work in groups it is important that submitted work is your own.

Sampling Words

1. Select a representative set of 10 words from the passage by bolding them with **bold**.

Four score and seven years ago, our fathers brought forth upon this continent a new nation: conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate, we cannot consecrate, we cannot hallow this ground. The brave men, living **and dead, who struggled here have consecrated it, far above** our poor power to add or detract. The world will little note, nor long remember, what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us, that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion, that we here highly resolve that these dead shall not have died in vain, that this nation, under God, shall have a new birth of freedom, and that government of the people, by the people, for the people, shall not perish from the earth.

The authorship of literary works is often a topic for debate. For example, researchers have tried to determine whether some of the works attributed to Shakespeare were actually written by Bacon or Marlow. The field of "literary computing" examines ways of numerically analyzing authors' works, looking at variables such as sentence length and rates of occurrence of specific words.

2. Complete the following data table by recording each word from your sample and then indicate the length of the word (number of letters) and whether or not the word is "short" (3 or fewer letters).

It might be easier to fill out this table in a *visual* mode. To switch into visual mode for a markdown document, use the button at the top-left of the editor toolbar.

Word	Length(# of letters)	Short (3 or fewer letters)? (Y or N)
1	3	Y
2	4	N
3	3	Y
4	9	N
5	4	N
6	4	N
7	11	N
8	2	Y
9	3	Y

Word	Length(# of letters)	Short (3 or fewer letters)? (Y or N)
10	5	N

3. Identify the observational units and the variables you have recorded on these observational units (Keep in mind that observational units do not have to be people!)

The length of the words

Definition

The ***population*** *is the entire collection of observational units of interest, and the* ***sample*** *is a subset of the population on which we record data.*

The passage is, of course, Lincoln's Gettysburg Address, given November 19, 1863, on the battlefield near Gettysburg, Pennsylvania. We are considering this passage a **population** of 268 words, and the 10 words you selected are a **sample** from this population. As before, we can use to refer to a sample proportion and to refer to the parameter, this time the proportion of short words in the population.

4. Do you think the sample proportion of short words you found will be a good estimate of the population proportion, Why or why not?

Yes, since we selected a representative set

5. Suggest a method for deciding whether the sample proportion of short words is likely to be close to the population proportion of short words. (Hint: Whereas any one sample may not produce a statistic that exactly equals the population parameter, what would we like to be true in general?)

we would want many samples to to average out to the population proportion of short words. We can take our sample and compare it to the overall sample average to see if our proportion is likely.

6. Record the sample proportion of short words in your sample of 10 words and enter your results in the *Exploration 2.1 Data Collection Sheet1*.

Reading csv file

Wait for all your classmates to enter their values into **Sheet 1**.

Then download **Sheet 1** as a .csv file (you could do it through the menu of the Google Sheets).

Next, upload this .csv file to R studio.

Then, read this .csv file into R Markdown with the commands included in the R chunk below.

Hint: to get the correct path to include between the quotation marks, locate the csv file in the *Files* tab, click on the *down arrow* to the right of *More(Blue Wheel Icon)* and choose *Copy folder path to Clipboard* from the drop-down menu.

In my case the path copied to Clipboard is *"~/Documents/Math 247 Materials/Explorations"*. Now I need to specify the name of the csv file to read at the specified location. Thus, *my full path* is :*"~/Documents/Math 247 Materials/Explorations/Exploration 2.1 Data Collection - Sheet1.csv"*.

Paste *your path copied from the Clipboard* before */Exploration 2.1 Data Collection - Sheet1.csv* within the quotation marks.

Use `head` command to make sure that the file is read correctly.

Remove **eval=FALSE** after running the code.

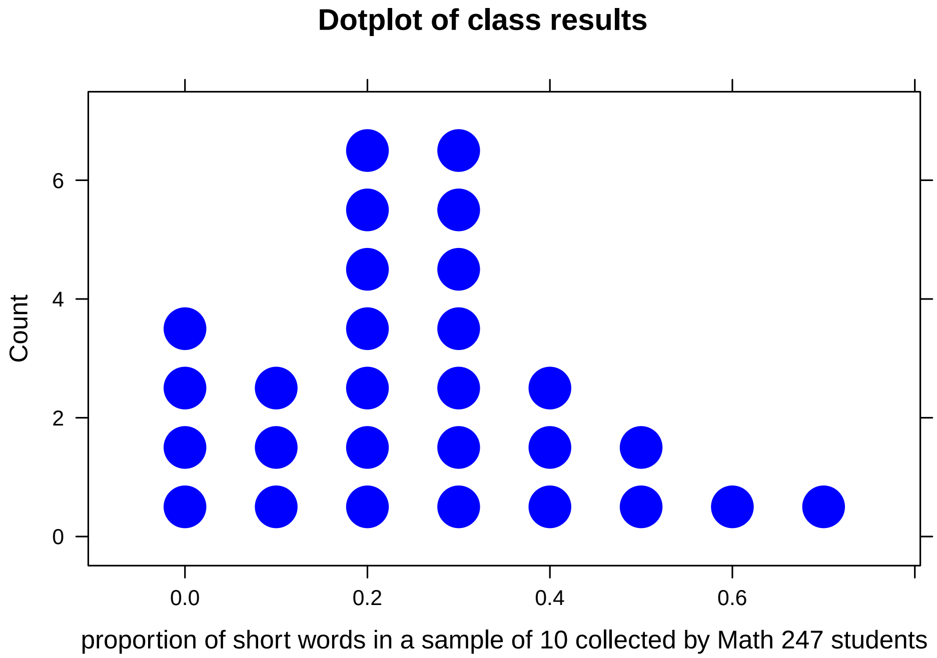
```
library(readr)
Exploration_2_1_Data_10 <-
  read_csv("~/Downloads/2.1S1.csv")
head(Exploration_2_1_Data_10, n=2)
```

```
## # A tibble: 2 x 2
##   Name  `proportion of short words`
##   <chr>                       <dbl>
## 1 Anna                            0
## 2 Ben                            NA
```

7. Produce a dotplot of the distribution of the proportion of short words in samples of 10 produced by your class. How many dots do you have? In other words, clarify what each dot represents. (Hint: If you wanted to add another dot to the graph, what would you need to do?).

Remove **eval=FALSE** after you complete and successfully run the code.

```
library(mosaic)
dotPlot(~`proportion of short words`, # tells R which variable to plot
data=Exploration_2_1_Data_10,  #tells R which data set to use
xlab = "proportion of short words in a sample of 10 collected by Math 247 students",
width=0.1, # tells R the width of a space between x values
main="Dotplot of class results", # plot title
col="blue")
```



8. Based on the class dotplot, does this sampling method tend to produce sample proportions close to the population proportion, 0.41? Justify your answer.

No, it tends to produce 0.25.

Definition

A sampling method is ***biased*** *if, when using that sampling method, statistics from different samples consistently overestimate or consistently underestimate the population parameter of interest.*

Note that bias is a property of a sampling method, not a property of an individual sample. Also note that the sampling method must consistently produce nonrepresentative results in the same direction in order to be considered biased.

9. Does asking you to quickly select 10 representative words appear to be a biased or unbiased sampling method? If biased, what is the direction of the bias (tendency to overestimate or underestimate the proportion of short words)?

Biased, it tends to underestimate the short word proportion

10. Explain why we might have expected this sampling method (asking you to quickly pick 10 representative words) to be biased for this variable.

We are more likely to pick bigger words as they stand out more to us than short words

11. Do you think that if we'd asked each of you to take 20 words instead of 10 words it would have helped with this issue? Explain.

No, because our sampling method would still be biased

12. Suggest another technique for selecting 10 words from this population in order for the sampling method to be unbiased and likely to produce samples that are representative of the population proportion of short words.

Select 10 consecutive words, or use a random number generator to pick 10 words.

Taking a Simple Random Sample

Key Idea: A **simple random sample** ensures that every sample of size n is equally likely to be the sample selected from the population. In particular, each observational unit has the same chance of being selected as every other observational unit. The key to obtaining a representative sample is using some type of random mechanism to select the observational units from the population, rather than relying on **convenience samples** or any type of human judgment.

Definition

A ***convenience sample*** *is a nonrandom sample of a population.*

Instead of having you choose "random" words using your own judgment, we will now ask you to take a simple random sample of words and evaluate your results. The first step is to obtain a sampling frame - a complete list of every member of the population where each member of the population can be assigned a number. Dataset `GettysburgAddress` within `library(ISIwithR)` contains a complete numbered list of all 268 words in the Gettysburg address. The `head` command outputs the first six observation in this dataset.

```r
library(ISIwithR)
data(GettysburgAddress)
head(GettysburgAddress)
```

```
##    word
## 1  Four
## 2 score
## 3   and
## 4 seven
## 5 years
## 6   ago
```

We can now use a `sample` function to randomly sample 5 words. Note that you would want to *use your own seed* in the `set.seed` command here, since we want every student in a class to get a different random sample of 5 words. In order to find the length (number of letters) of each word and characterize it as short or not we use a `mutate` function within `library(tidyverse)`. If using this code gives you an error you will need to install the `tidyverse` package first.

Remove **eval=FALSE** after you complete and successfully run the code.

```
set.seed(23)

library(tidyverse)
Words <- sample(GettysburgAddress, 5)
Words<-Words %>% mutate(length = nchar(word)) %>% mutate(
      type = case_when(
      length <=3 ~ "short",
      TRUE       ~  "other"
    )
  )
Words
```

```
##      word orig.id length  type
## 248 birth     248      5 other
## 121   men     121      3 short
## 171  here     171      4 other
## 103   But     103      3 short
## 149    we     149      2 short
```

14. Let's examine your sample of 5 words, as well as those of your classmates.

    a. Use the R code below to calculate the proportion of short words in your random sample (i.e. the number of times length <=3 is TRUE in a sample of 5).

Remove **eval=FALSE** after you complete and successfully run the code.

```
Words %>%
  group_by(type) %>%
  summarize(count_short = n()) %>%
  mutate(proportion_short = count_short / sum(count_short)) %>% # creates a new variable named proporti
  ungroup() %>%
  complete(type,
  fill = list(count_short = 0, proportion_short = 0)) %>%
  filter(type=="short")
```

```
## # A tibble: 1 x 3
##   type  count_short proportion_short
##   <chr>       <int>            <dbl>
## 1 short           3              0.6
```

Enter your results in the *Exploration 2.1 Data Collection Sheet2* Wait for all your classmates to enter their values into **Sheet 2**. Then download **Sheet 2** as a .csv file (you could do it through the menu of the Google Sheets). Next, upload this .csv file to R studio. Then, read this .csv file into R Markdown with the commands included in the R chunk below.

Use `head` command to make sure that the file is read correctly.

Remove **eval=FALSE** after you complete and successfully run the code.

```
library(readr)
Exploration_2_1_Data_5 <-
```

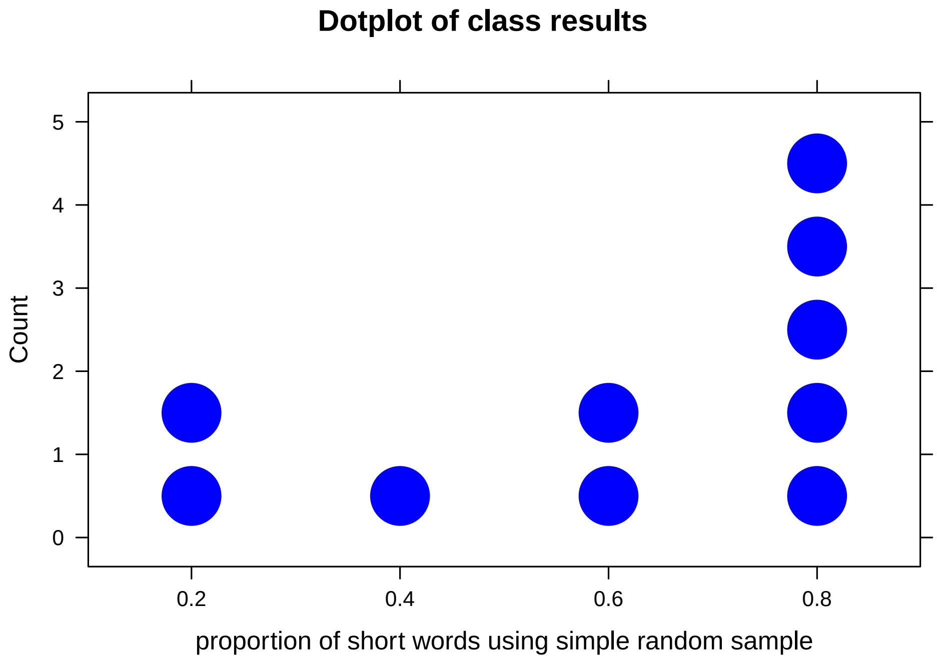
```
read_csv("~/Downloads/2.1S2.csv")
head(Exploration_2_1_Data_5, n=2)
```

```
## # A tibble: 2 x 2
##   Name  `proportion of short words using SRS`
##   <chr>                                 <dbl>
## 1 Anna                                     NA
## 2 Ben                                      NA
```

b. Again, produce a dotplot of the distribution of proportions of short words for your sample and those of your classmates.

Remove **eval=FALSE** after you complete and successfully run the code.

```
library(mosaic)
dotPlot(~`proportion of short words using SRS`,
data=Exploration_2_1_Data_5,
xlab = "proportion of short words using simple random sample",
width=0.1,
main="Dotplot of class results",
col="blue")
```



c. Comment on how this distribution compares to your class' dotplot of sample proportions from #6 which used nonrandom sampling.

People forgot to change the seed so it's mostly at 0.8…, but it should be more accurate and close to 0.4

d. Does this second method appear to be an unbiased sampling method? How are you deciding?

It appears to be biased as the seed didn't get changed in most instances

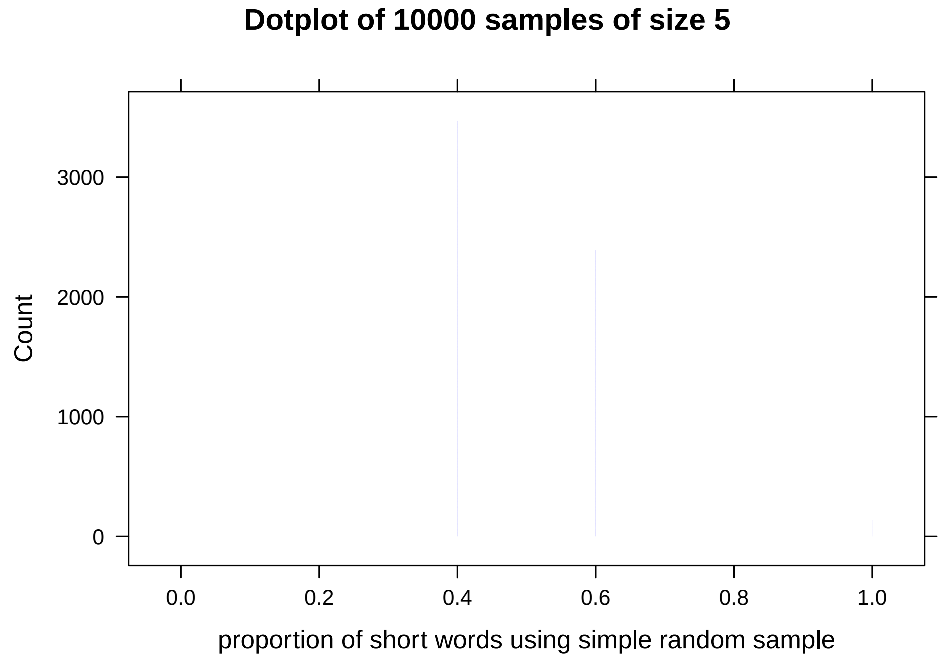
16. Suppose you take 1,000 different random samples and examined the distribution of the 1,000 sample proportions. Where do you think this distribution will be centered? What shape do you think it will have? Predict the largest and smallest values.

It wil be centered at 0.4, and the min will brobably be 0 and the max will probalby be 0.9

17. Let's create a distribution of 10000 samples . Describe the behavior of the distribution of sample proportions and contrast it with your prediction, as well as to the class distribution of nonrandom samples using 10 words.

Remove **eval=FALSE** after you complete and successfully run the code.

Keep **echo=FALSE**



Definition

A **sampling distribution** is the distribution of a statistic for all possible samples of size n randomly selected from the same population. By taking lots of samples, you can approximate this sampling distribution.

Information below outlines the simulation process you are using to approximate the sampling distribution of sample proportions of short words from the Gettysburg Address.

Parallels between population distributions and sampling distributions

Population: All 268 words in the Gettysburg Address
Parameter: Proportion of words in the population that are short π=0.41
Sample: 5 randomly selected words
Statistic: Proportion of sampled words that are short (varies from sample to sample)

One big change between the random samples and the previous convenience samples is the center of the sampling distribution should now be close to the population proportion. In other words, simple random sampling is an unbiased sampling method because there is no tendency to over- or underestimate the parameter.

Key Idea: The mean of the sampling distribution of sample proportions for all possible simple random samples from a population with success proportion will be equal to π.

As we saw, this sampling method (taking 5 **randomly** selected words) is better (**unbiased**) even though we selected fewer words.

Sampling Distribution of Sample Proportions

Based on what you learned about sample proportions in Chapter 1, you should not be too surprised that the distribution of sample proportions you produced in #17 is somewhat symmetric (because is close to 0.50), with a mean close to 0.41 (because the sampling method is unbiased). In fact, our previous formula for the standard deviation of sample proportions, $\sqrt{\pi * (1-\pi)/n}$ still works as well!

18. Verify that the standard deviation of your 10,000 sample proportions is close to $\sqrt{\pi * (1-\pi)/n}$ (Hint: What values are you using for π and n?).

Remove **eval=FALSE** after you successfully run the code.

```
## simulation-based SD
round(favstats(~proportion_short, data = Words10000)["sd"],3)
```

```
##     sd
##  0.222
```

Remove **eval=FALSE** after you complete and successfully run the code.

```
## theory-based SD
n=5
pi=0.4
SD<-sqrt(pi*(1-pi)/n)
SD
```

```
## [1] 0.219089
```

Key Idea: When sampling from a large population, the standard deviation of sample proportions is estimated by the same formula, $\sqrt{\pi * (1-\pi)/n}$ where π represents the proportion of successes in the population and n represents the sample size. *The population size is considered large enough when it is more than 20 times that sample size.*

Notice one new aspect here compared to Chapter 1: When we are sampling from a finite population rather than from an infinite random process, we want the population to be large in order to assume represents a constant probability of success. In this case, and the formula should apply. In which case, we can use the same theory-based methods we used in Chapter 1.

19. How would our prediction of the behavior (shape, mean, largest value/smallest value) of the distribution of sample proportions change if the population size had been 2,680?

Our prediction would not change

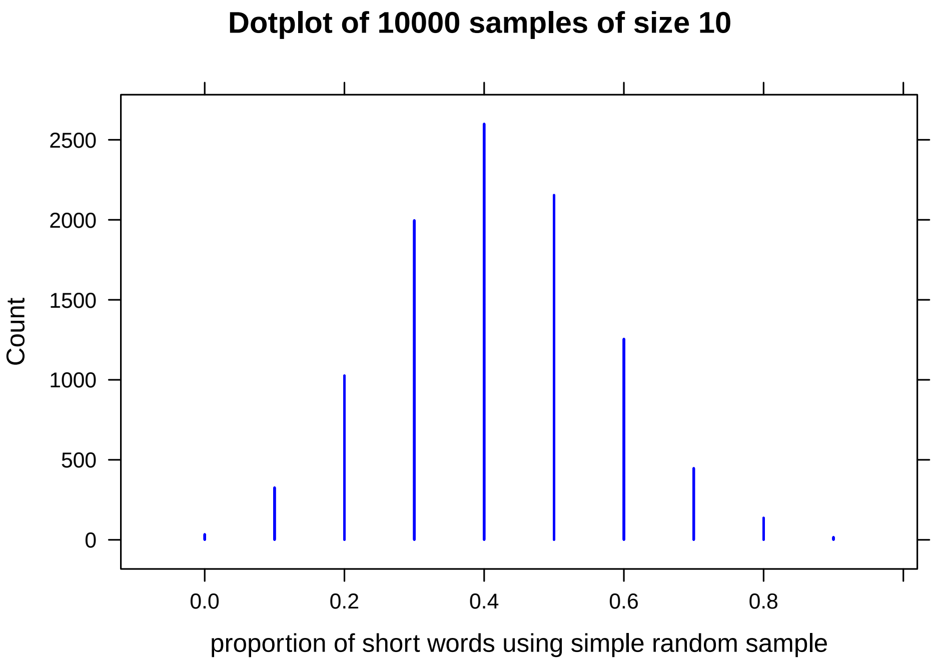
20. Name two things that do impact the standard deviation of the distribution of sample proportions.

Size of the sample, and population size

21. In R we will change the sample size from 5 to 10. Describe the behavior of the new sampling distribution. Does the standard deviation change as you predicted? Is the sampling method still unbiased? How are you deciding?

Remove **eval=FALSE** after you successfully run the code.

Keep **echo=FALSE**.

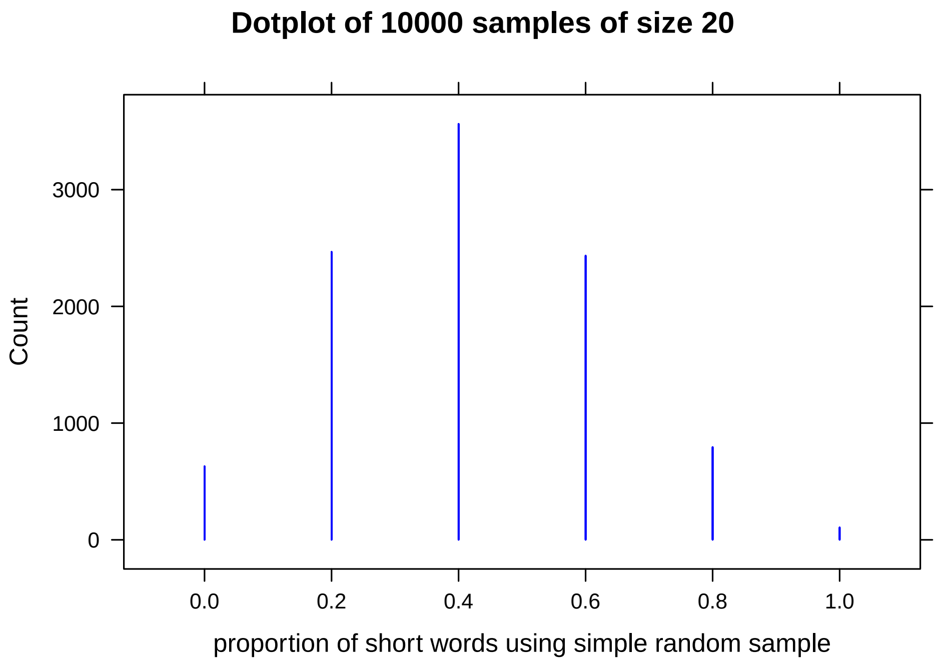


22. Repeat #21 for a sample size of 20.

Remove **eval=FALSE** after you complete and successfully run the code.

```
Words10000_5 <- do(10000) *sample(GettysburgAddress,5)
Words10000_5 <-Words10000_5 %>%
mutate(length = nchar(word)) %>%
  mutate(
      type = case_when(
      length <=3 ~ "short",
      TRUE       ~  "other"
    )
  ) %>%
  group_by(.index,type) %>%
  summarize(count_short = n()) %>%
  mutate(proportion_short = count_short / sum(count_short)) %>%
  ungroup() %>%
  complete(.index,type,
  fill = list(count_short = 0, proportion_short = 0)) %>%
  filter(type=="short")

#histogram(~proportion_short, data = Words10000_#)
dotPlot(~proportion_short,
data=Words10000_5,
xlab = "proportion of short words using simple random sample",
width=0.1,
cex=21,
main="Dotplot of 10000 samples of size 20",
col="blue")
```



Keep in mind that our population size (268) is not all that large here and you may start to see some discrepancies with the standard deviation formula with samples of 20 or more (Why?). But also notice that the shape of the distribution of sample proportions is looking more and more like a normal distribution as we increase the sample size, just as you saw in Chapter 1.

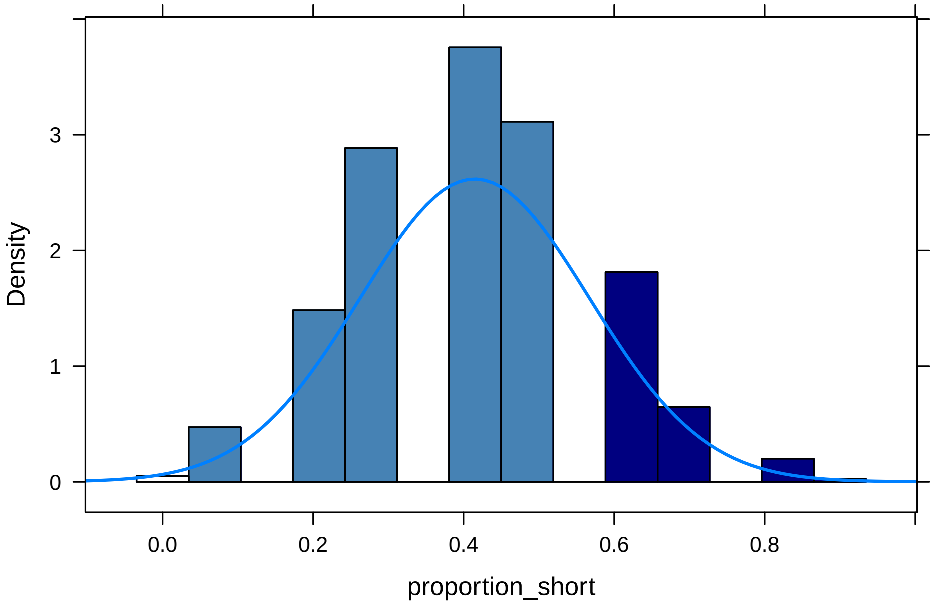
Definition

The ***Central Limit Theorem for sampling from a large finite population*** *says that the distribution of sample proportions from repeated random samples will be approximately normal if there at least 10 successes and at least 10 failures in each sample.*

In figure below, we have overlaid the normal distribution predicted by the *Central Limit Theorem* and determined the proportion of samples and the theoretical probability for a sample proportion of 0.55 and higher.

Remove **eval=FALSE** after running the code.

```
histogram(~proportion_short, data = Words10000_10, fit = "normal",
group = cut(proportion_short,
            c(0, 0.55, 1)),
      fcol = c("steelblue", "navy"))
```



23. Does the normal approximation appear to be valid here? Which p-value do you trust more? Explain your reasoning.

Remove **eval=FALSE** after running the code.

```
#Right-tailed p-value
simulation.based.p.value<-prop(~(proportion_short >= 0.55),data = Words10000)
simulation.based.p.value
```

```
## prop_TRUE
##    0.3379
```

No, since we don't have any more than 10 successes/failures

```
#Right-tail p-value
theory.based.p.value<-pnorm(0.55, mean=0.41, sd=0.22,lower.tail = FALSE)
theory.based.p.value
```

```
## [1] 0.2622697
```

Other Considerations

Another key property of our sampling method is that we had a list of every word in the population. This list is referred to as a sampling frame. An incomplete sampling frame (e.g., only the first paragraph of the speech) would also produce a biased sampling method if the observational units not in the sampling frame were systematically different from the rest of the population.

Key Idea: If a sampling frame does not include every member of the population (e.g., last paragraph missing, nouns not listed), then we cannot claim to represent those segments in the population with our sample. An incomplete sampling frame can also lead to biased sampling.

You also need to be sure there are not any other **nonsampling concerns**. These represent problems that arise even with a carefully selected sample. For example, suppose we were not able to accurately measure the lengths of the words. Or, different individuals used different definitions of "short."

Key Idea: In addition to carefully selecting the sample, you need to guard against other possible sources of bias as well. For example, when talking to people, their suspect memories and truthfulness can impact people's answers. Other examples of **nonsampling concerns** include the effects that the wording of a question can have on people's responses (e.g., how extreme the phrasing is), the effects that the interviewer can have by virtue of their demeanor, sex, race, and other characteristics, using an uncalibrated scale, and so on.

See Example 2.1 in the textbook for additional discussion of nonsampling concerns.

*Make sure to Remove all **eval=FALSE** before knitting the document.*