Data Science Visualization

The textbook for the Data Science course series is freely available online.

Learning Objectives

- Data visualization principles to better communicate data-driven findings
- How to use ggplot2 to create custom plots
- The weaknesses of several widely used plots and why you should avoid them

Course Overview

Section 1: Introduction to Data Visualization and Distributions

You will get started with data visualization and distributions in R.

Section 2: Introduction to ggplot2

You will learn how to use ggplot2 to create plots.

Section 3: Summarizing with dplyr

You will learn how to summarize data using dplyr.

Section 4: Gapminder

You will see examples of ggplot2 and dplyr in action with the Gapminder dataset.

Section 5: Data Visualization Principles

You will learn general principles to guide you in developing effective data visualizations.

Section 1 Overview

Section 1 introduces you to Data Visualization and Distributions.

After completing Section 1, you will:

- understand the importance of data visualization for communicating data-driven findings.
- be able to use distributions to summarize data.
- be able to use the average and the standard deviation to understand the normal distribution.
- be able to assess how well a normal distribution fits the data using a quantile-quantile plot.
- be able to interpret data from a boxplot.

Introduction to Data Visualization

The textbook for this section is available here

Key points

- Plots of data easily communicate information that is difficult to extract from tables of raw values.
- Data visualization is a key component of exploratory data analysis (EDA), in which the properties of data are explored through visualization and summarization techniques.
- Data visualization can help discover biases, systematic errors, mistakes and other unexpected problems in data before those data are incorporated into potentially flawed analysis.
- This course covers the basics of data visualization and EDA in R using the **ggplot2** package and motivating examples from world health, economics and infectious disease.

Code

```r
library(dslabs)
data(murders)
head(murders)
```

```
##        state abb region population total
## 1    Alabama  AL  South    4779736   135
## 2     Alaska  AK   West     710231    19
## 3    Arizona  AZ   West    6392017   232
## 4   Arkansas  AR  South    2915918    93
## 5 California  CA   West   37253956  1257
## 6   Colorado  CO   West    5029196    65
```

Introduction to Distributions

The textbook for this section is available here

Key points

- The most basic statistical summary of a list of objects is its distribution.
- We will learn ways to visualize and analyze distributions in the upcoming videos.
- In some cases, data can be summarized by a two-number summary: the average and standard deviation. We will learn to use data visualization to determine when that is appropriate.

Data Types

The textbook for this section is available here

Key points

- Categorical data are variables that are defined by a small number of groups.
  - Ordinal categorical data have an inherent order to the categories (mild/medium/hot, for example).
  - Non-ordinal categorical data have no order to the categories.
- Numerical data take a variety of numeric values.
  - Continuous variables can take any value.
  - Discrete variables are limited to sets of specific values.

Assessment - Data Types

1. The type of data we are working with will often influence the data visualization technique we use.

We will be working with two types of variables: categorical and numeric. Each can be divided into two other groups: categorical can be ordinal or not, whereas numerical variables can be discrete or continuous.

We will review data types using some of the examples provided in the `dslabs` package. For example, the `heights` dataset.

```r
library(dslabs)
data(heights)
```

```r
data(heights)
names(heights)
```

```
## [1] "sex"    "height"
```

2. We saw that `sex` is the first variable. We know what values are represented by this variable and can confirm this by looking at the first few entires:

```r
head(heights)
```

```
##      sex height
## 1   Male     75
## 2   Male     70
## 3   Male     68
## 4   Male     74
## 5   Male     61
## 6 Female     65
```

What data type is the `sex` variable?

- ☐ A. Continuous
- ☒ B. Categorical
- ☐ C. Ordinal
- ☐ D. None of the above

3. Keep in mind that discrete numeric data can be considered ordinal.

Although this is technically true, we usually reserve the term ordinal data for variables belonging to a small number of different groups, with each group having many members.

The `height` variable could be ordinal if, for example, we report a small number of values such as short, medium, and tall. Let's explore how many unique values are used by the heights variable. For this we can use the `unique` function:

```r
x <- c(3, 3, 3, 3, 4, 4, 2)
unique(x)
```

```r
x <- heights$height
length(unique(x))
```

```
## [1] 139
```

4. One of the useful outputs of data visualization is that we can learn about the distribution of variables.

For categorical data we can construct this distribution by simply computing the frequency of each unique value. This can be done with the function `table`. Here is an example:

```r
x <- c(3, 3, 3, 3, 4, 4, 2)
table(x)
```

```r
x <- heights$height
tab <- table(x)
```

5. To see why treating the reported heights as an ordinal value is not useful in practice we note how many values are reported only once.

In the previous exercise we computed the variable tab which reports the number of times each unique value appears. For values reported only once tab will be 1. Use logicals and the function sum to count the number of times this happens.

```r
tab <- table(heights$height)
sum(tab==1)
```

```
## [1] 63
```

6. Since there are a finite number of reported heights and technically the height can be considered ordinal, which of the following is true:

- [x] A. It is more effective to consider heights to be numerical given the number of unique values we observe and the fact that if we keep collecting data even more will be observed.
- [ ] B. It is actually preferable to consider heights ordinal since on a computer there are only a finite number of possibilities.
- [ ] C. This is actually a categorical variable: tall, medium or short.
- [ ] D. This is a numerical variable because numbers are used to represent it.

Describe Heights to ET

The textbook for this section is available:

- Case Study describing student heights
- Distribution Function
- CDF Intro
- Histograms

Key points

- A distribution is a function or description that shows the possible values of a variable and how often those values occur.
- For categorical variables, the distribution describes the proportions of each category.
- A *frequency table* is the simplest way to show a categorical distribution. Use `prop.table` to convert a table of counts to a frequency table. *Barplots* display the distribution of categorical variables and are a way to visualize the information in frequency tables.
- For continuous numerical data, reporting the frequency of each unique entry is not an effective summary as many or most values are unique. Instead, a distribution function is required.
- The *cumulative distribution function (CDF)* is a function that reports the proportion of data below a value a for all values of a: $F(a) = Pr(x \leq a)$.
- The proportion of observations between any two values a and b can be computed from the CDF as $F(b) - F(a)$.
- A *histogram* divides data into non-overlapping bins of the same size and plots the counts of number of values that fall in that interval.

Code

```
# load the dataset
library(dslabs)
data(heights)
```

```
# make a table of category proportions
prop.table(table(heights$sex))
```

```
##
##    Female      Male
## 0.2266667 0.7733333
```

Smooth Density Plots

The textbook for this section is available here

Key points

- *Smooth density plots* can be thought of as histograms where the bin width is extremely or infinitely small. The smoothing function makes estimates of the true continuous trend of the data given the available sample of data points.
- The degree of smoothness can be controlled by an argument in the plotting function. (We will learn functions for plotting later.)
- While the histogram is an assumption-free summary, the smooth density plot is shaped by assumptions and choices you make as a data analyst.
- The y-axis is scaled so that the area under the density curve sums to 1. This means that interpreting values on the y-axis is not straightforward. To determine the proportion of data in between two values, compute the area under the smooth density curve in the region between those values.
- An advantage of smooth densities over histograms is that densities are easier to compare visually.

A further note on histograms: note that the choice of binwidth has a determinative effect on shape. There is no "true" choice for binwidth, and you can sometimes gain insights into the data by experimenting with binwidths.

Assessment - Distributions

1. You may have noticed that numerical data is often summarized with the average value.

For example, the quality of a high school is sometimes summarized with one number: the average score on a standardized test. Occasionally, a second number is reported: the standard deviation. So, for example, you might read a report stating that scores were 680 plus or minus 50 (the standard deviation). The report has summarized an entire vector of scores with with just two numbers. Is this appropriate? Is there any important piece of information that we are missing by only looking at this summary rather than the entire list? We are going to learn when these 2 numbers are enough and when we need more elaborate summaries and plots to describe the data.

Our first data visualization building block is learning to summarize lists of factors or numeric vectors. The most basic statistical summary of a list of objects or numbers is its distribution. Once a vector has been summarized as distribution, there are several data visualization techniques to effectively relay this information. In later assessments we will practice to write code for data visualization. Here we start with some multiple choice questions to test your understanding of distributions and related basic plots.

In the murders dataset, the region is a categorical variable and on the right you can see its distribution. To the closest 5%, what proportion of the states are in the North Central region?

In the murders dataset, the region is a categorical variable and the following is its distribution.

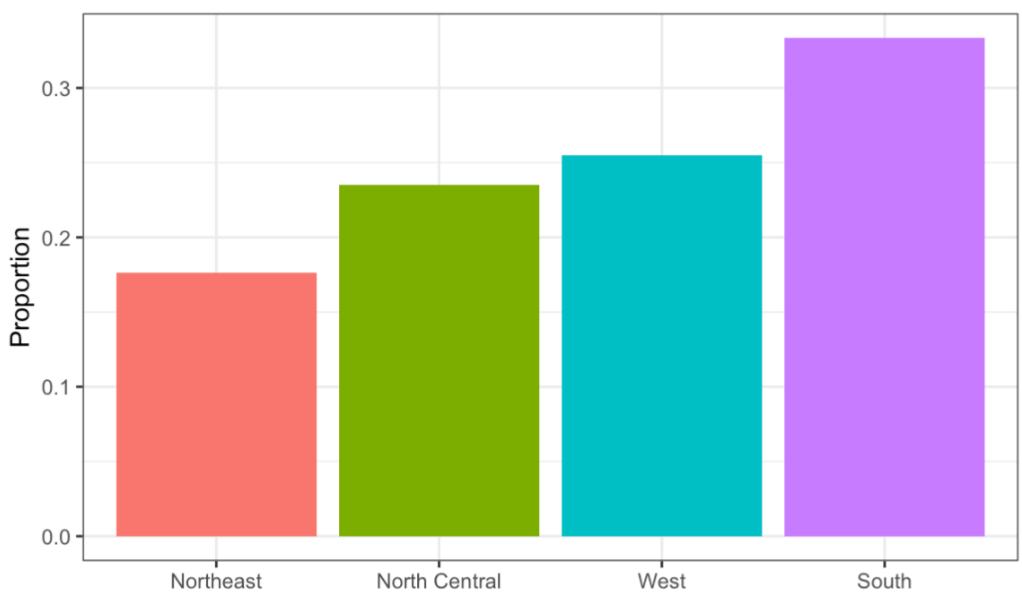

Figure 1: Region vs. Proportion

- ☐ A. 75%
- ☐ B. 50%
- ☒ C. 20%
- ☐ D. 5%

2. In the murders dataset, the region is a categorical variable and to the right is its distribution.

Which of the following is true:

☐ A. The graph above is a histogram.
☒ B. The graph above shows only four numbers with a bar plot.
☐ C. Categories are not numbers, so it does not make sense to graph the distribution.
☐ D. The colors, not the height of the bars, describe the distribution.

3. The plot shows the eCDF for male heights.

Based on the plot, what percentage of males are shorter than 75 inches?

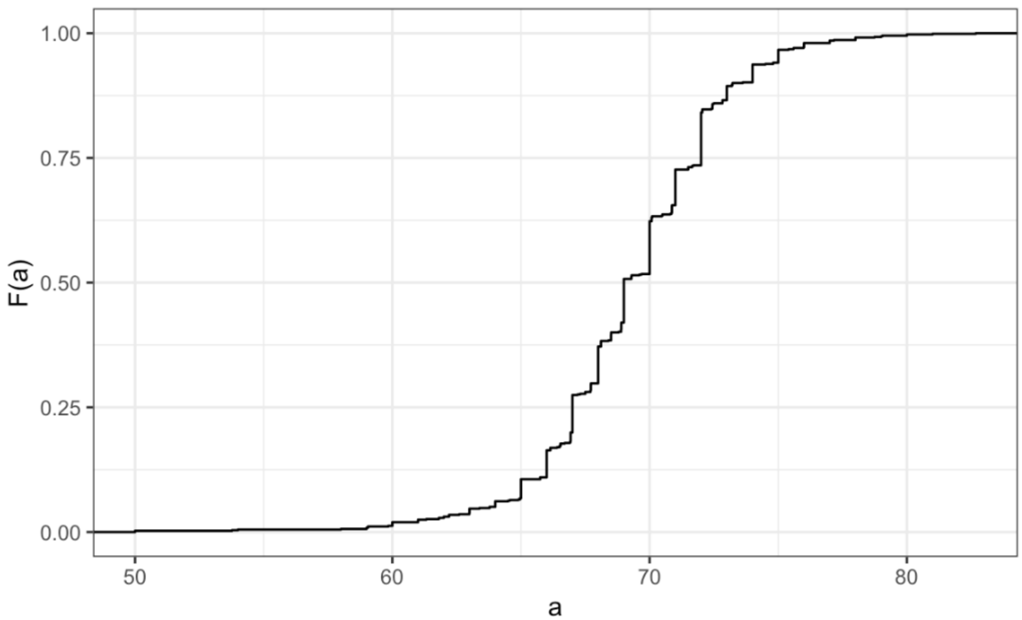

Figure 2: eCDF for male heights

☐ A. 100%
☒ B. 95%
☐ C. 80%
☐ D. 72 inches

4. To the closest inch, what height `m` has the property that 1/2 of the male students are taller than `m` and 1/2 are shorter?

☐ A. 61 inches
☐ B. 64 inches
☒ C. 69 inches
☐ D. 74 inches

5. Here is an eCDF of the murder rates across states.

Knowing that there are 51 states (counting DC) and based on this plot, how many states have murder rates larger than 10 per 100,000 people?

☒ A. 1

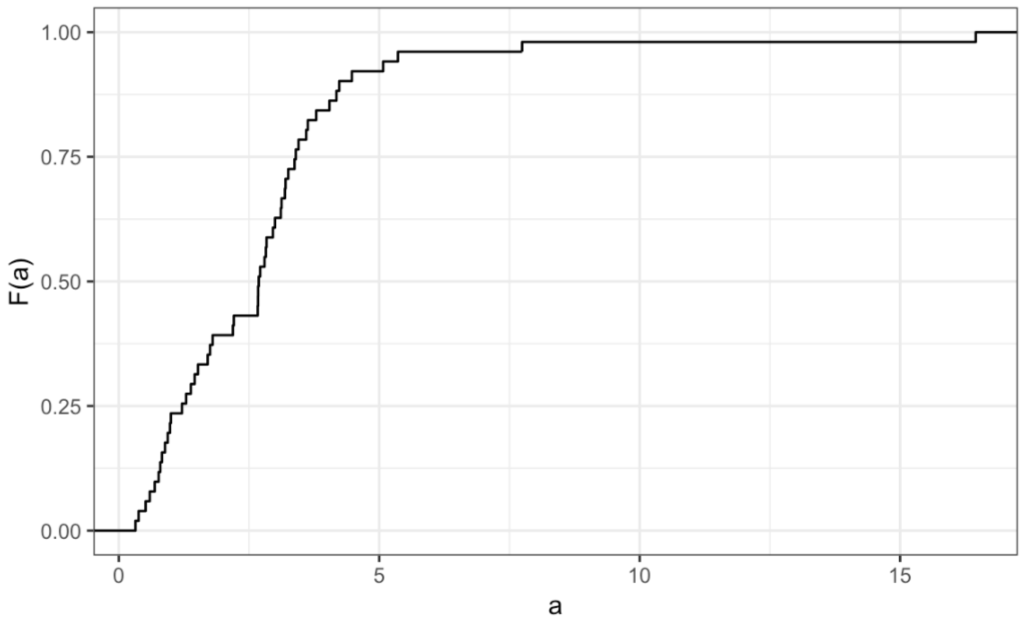

Figure 3: eCDF of the murder rates across states

☐ B. 5
☐ C. 10
☐ D. 50

6. Based on the eCDF above, which of the following statements are true.

☐ A. About half the states have murder rates above 7 per 100,000 and the other half below.
☐ B. Most states have murder rates below 2 per 100,000.
☐ C. All the states have murder rates above 2 per 100,000.
☒ D. With the exception of 4 states, the murder rates are below 5 per 100,000.

7. Here is a histogram of male heights in our `heights` dataset.

Based on this plot, how many males are between 62.5 and 65.5?

☐ A. 11
☐ B. 29
☒ C. 58
☐ D. 99

8. About what percentage are shorter than 60 inches?

☒ A. 1%
☐ B. 10%
☐ C. 25%
☐ D. 50%

9. Based on this density plot, about what proportion of US states have populations larger than 10 million?

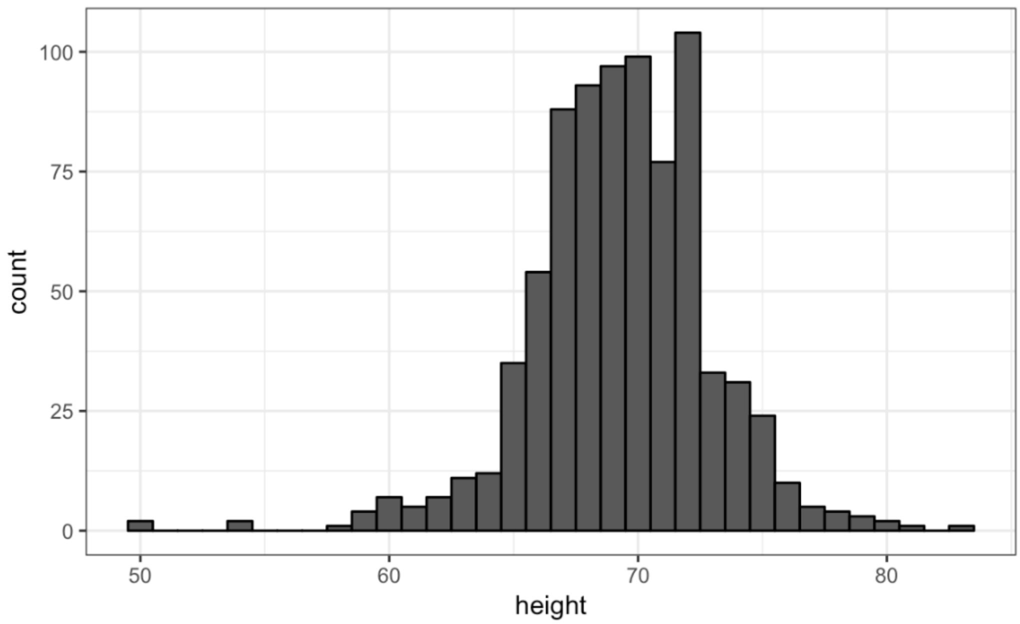

Figure 4: Histogram of male heights

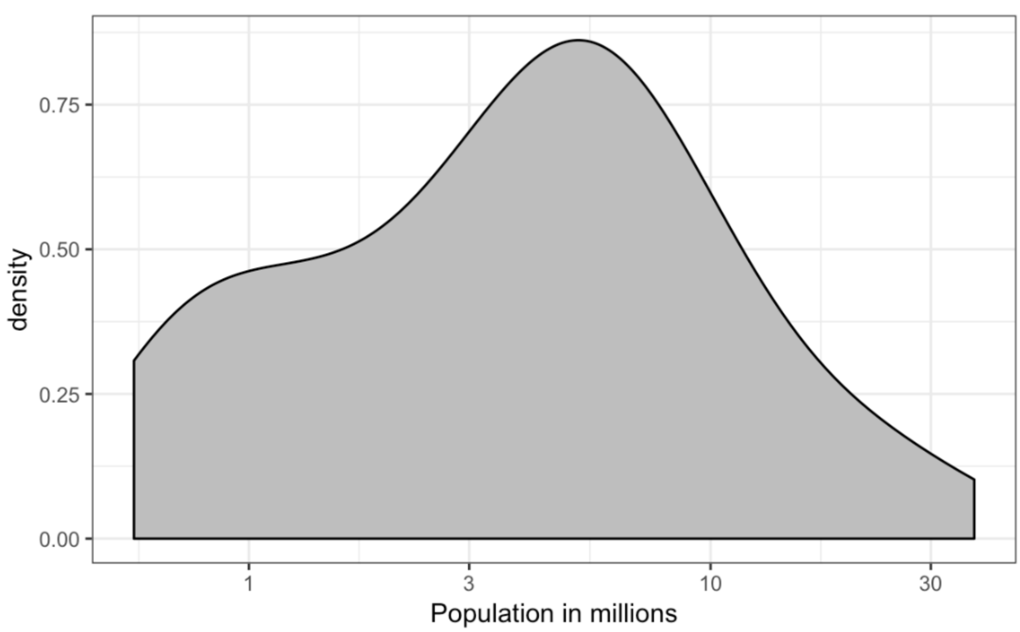

Figure 5: Density plot population

☐ A. 0.02
☒ B. 0.15
☐ C. 0.50
☐ D. 0.55

10. Here are three density plots. Is it possible that they are from the same dataset?

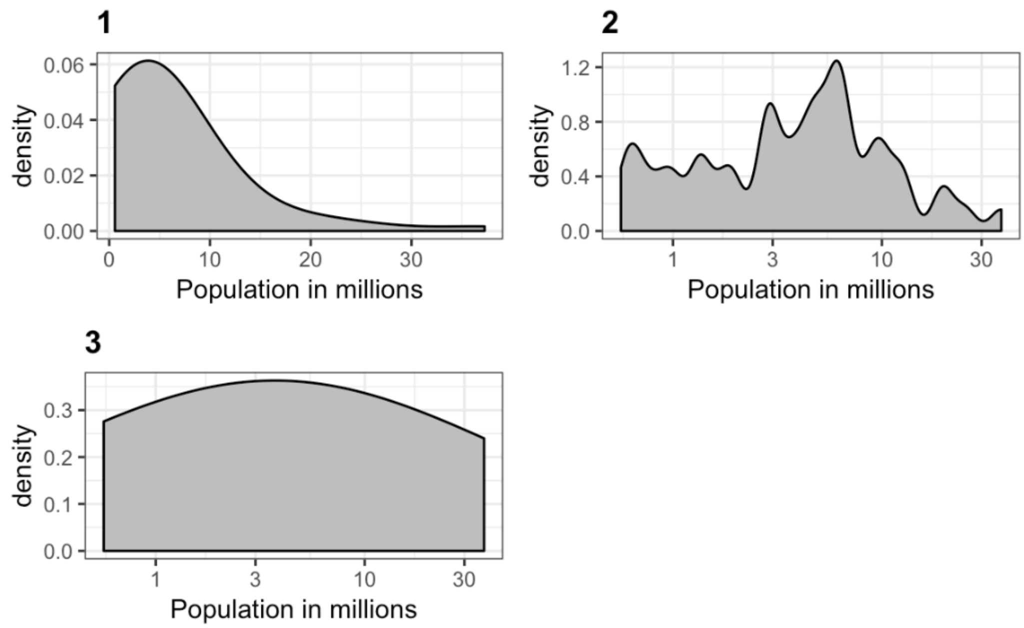

Figure 6: Three density plots

Which of the following statements is true?

☐ A. It is impossible that they are from the same dataset.
☐ B. They are from the same dataset, but the plots are different due to code errors.
☐ C. They are the same dataset, but the first and second plot undersmooth and the third oversmooths.
☒ D. They are the same dataset, but the first is not in the log scale, the second undersmooths and the third oversmooths.

Assessment 3 (Normal Distribution)

1. Proportions

Histograms and density plots provide excellent summaries of a distribution. But can we summarize even further? We often see the average and standard deviation used as summary statistics: a two number summary! To understand what these summaries are and why they are so widely used, we need to understand the normal distribution.

The normal distribution, also known as the bell curve and as the Gaussian distribution, is one of the most famous mathematical concepts in history. A reason for this is that approximately normal distributions occur in many situations. Examples include gambling winnings, heights, weights, blood pressure, standardized test scores, and experimental measurement errors. Often data visualization is needed to confirm that our data follows a normal distribution.

Here we focus on how the normal distribution helps us summarize data and can be useful in practice.

One way the normal distribution is useful is that it can be used to approximate the distribution of a list of numbers without having access to the entire list. We will demonstrate this with the heights dataset.

Load the height data set and create a vector x with just the male heights:

```
library(dslabs)
data(heights)

x <- heights$height[heights$sex == "Male"]
```

What proportion of the data is between 69 and 72 inches (taller than 69 but shorter or equal to 72)?

```
library(dslabs)
data(heights)

x <- heights$height[heights$sex == "Male"]
mean(x>69 & x<=72)
```

```
## [1] 0.3337438
```

2. Averages and Standard Deviations

Suppose all you know about the height data from the previous exercise is the average and the standard deviation and that its distribution is approximated by the normal distribution. We can compute the average and standard deviation like this:

```
library(dslabs)
data(heights)

x <- heights$height[heights$sex=="Male"]
avg <- mean(x)
stdev <- sd(x)
```

Suppose you only have avg and stdev below, but no access to x, can you approximate the proportion of the data that is between 69 and 72 inches?

Use the normal approximation to estimate the proportion the proportion of the data that is between 69 and 72 inches. Note that you can't use x in your code, only avg and stdev. Also note that R has a function that may prove very helpful here - check out the pnorm function (and remember that you can get help by using ?pnorm)

```
library(dslabs)
data(heights)

x <- heights$height[heights$sex=="Male"]
avg <- mean(x)
stdev <- sd(x)
pnorm(72, avg, stdev) - pnorm(69, avg, stdev)
```

```
## [1] 0.3061779
```

3. Approximations

Notice that the approximation calculated in the second question is very close to the exact calculation in the first question. The normal distribution was a useful approximation for this case.

However, the approximation is not always useful. An example is for the more extreme values, often called the "tails" of the distribution. Let's look at an example. We can compute the proportion of heights between 79 and 81.

```
library(dslabs)
data(heights)

x <- heights$height[heights$sex == "Male"]
mean(x > 79 & x <= 81)
```

```
## [1] 0.004926108
```

Use normal approximation to estimate the proportion of heights between 79 and 81 inches and save it in an object called approx. Report how many times bigger the actual proportion is compared to the approximation.

```
library(dslabs)
data(heights)

x <- heights$height[heights$sex == "Male"]
exact <- mean(x > 79 & x <= 81)
avg <- mean(x)
stdev <- sd(x)
approx <- pnorm(81, avg, stdev) - pnorm(79, avg, stdev)
exact/approx
```

```
## [1] 1.614261
```

4. Seven footers and the NBA

Someone asks you what percent of seven footers are in the National Basketball Association (NBA). Can you provide an estimate? Let's try using the normal approximation to answer this question.

First, we will estimate the proportion of adult men that are 7 feet tall or taller.

Assume that the distribution of adult men in the world as normally distributed with an average of 69 inches and a standard deviation of 3 inches. Using this approximation, estimate the proportion of adult men that are 7 feet tall or taller, referred to as seven footers. Print out your estimate; don't store it in an object.

```
1 - pnorm(7*12, 69, 3)
```

```
## [1] 2.866516e-07
```

5. Estimating the number seven footers

Now we have an approximation for the proportion, call it p, of men that are 7 feet tall or taller.

We know that there are about 1 billion men between the ages of 18 and 40 in the world, the age range for the NBA.

Can we use the normal distribution to estimate how many of these 1 billion men are at least seven feet tall? Use your answer to the previous exercise to estimate the proportion of men that are seven feet tall or taller in the world and store that value as p. Then round the number of 18-40 year old men who are seven feet tall or taller to the nearest integer. (Do not store this value in an object.)

```
p <- 1 - pnorm(7*12, 68, 3)

round(p * 1*10^9)

## [1] 48
```

6. How many seven footers are in the NBA?

There are about 10 National Basketball Association (NBA) players that are 7 feet tall or higher.

Use your answer to exercise 4 to estimate the proportion of men that are seven feet tall or taller in the world and store that value as p.

Use your answer to the previous exercise (exercise 5) to round the number of 18-40 year old men who are seven feet tall or taller to the nearest integer and store that value as N.

Then calculate the proportion of the world's 18 to 40 year old seven footers that are in the NBA. (Do not store this value in an object.)

```
p <- 1-pnorm(7*12,69,3)
N <- round(p*10^9)
10/N

## [1] 0.03484321
```

7. Lebron James' height

In the previous exerceise we estimated the proportion of seven footers in the NBA using this simple code:

```
p <- 1 - pnorm(7*12, 69, 3)
N <- round(p * 10^9)
10/N

## [1] 0.03484321
```

Repeat the calculations performed in the previous question for Lebron James' height: 6 feet 8 inches. There are about 150 players, instead of 10, that are at least that tall in the NBA.

Report the estimated proportion of people at least Lebron's height that are in the NBA.

```
## Change the solution to previous answer
p <- 1 - pnorm(6*12+8, 69, 3)
N <- round(p * 10^9)
150/N
```

```
## [1] 0.001220842
```

8. Interpretation

In answering the previous questions, we found that it is not at all rare for a seven footer to become an NBA player.

What would be a fair critique of our calculations? - [ ] A. Practice and talent are what make a great basketball player, not height. - [ ] B. The normal approximation is not appropriate for heights. - [X] C. As seen in exercise 3, the normal approximation tends to underestimate the extreme values. It's possible that there are more seven footers than we predicted. - [ ] D. As seen in exercise 3, the normal approximation tends to overestimate the extreme values. It's possible that there are less seven footers than we predicted.

Assessment 4 (Quantiles, percentiles, and boxplots)

1. Vector lengths

When analyzing data it's often important to know the number of measurements you have for each category. Define a variable male that contains the male heights. Define a variable female that contains the female heights. Report the length of each variable.

```
library(dslabs)
data(heights)

male <- heights$height[heights$sex=="Male"]
female <- heights$height[heights$sex=="Female"]

length(male)
```

```
## [1] 812
```

```
length(female)
```

```
## [1] 238
```

2. Percentiles

Suppose we can't make a plot and want to compare the distributions side by side. We can't just list all the numbers. Instead, we will look at the percentiles. Create a five row table showing female_percentiles and male_percentiles with the 10th, 30th, 50th, ., 90th percentiles for each sex. Then create a data frame with these two as columns.

```
library(dslabs)
data(heights)

male <- heights$height[heights$sex=="Male"]
female <- heights$height[heights$sex=="Female"]

male_percentiles <- quantile(male, seq(0.1, 0.9,0.2))
female_percentiles <- quantile(female, seq(0.1, 0.9,0.2))

df <- data.frame(female = female_percentiles, male = male_percentiles)
df
```