Week 5: Describing data and estimation

Goals

- Investigate sampling error; see that larger samples have less sampling error.
- Visualize confidence intervals.
- Calculate basic summary statistics using R.
- Calculate confidence intervals for the mean with R.

Learning the Tools

Missing data

Sometimes we do not have all variables measured on all individuals in the data set. When this happens, we need a space holder in our data files so that R knows that the data is missing. The standard way of doing this in R is to put “NA” (without the quotes) in the location that the data would have gone. NA is short for “not available”.

For example, in the Titanic data set, we do not know the age of several passengers. Let’s look at it. Load the Titanic data set:

```r
> titanicData <- read.csv("Data/titanic.csv")
```

Have R print out the list of the age variable, which you can do easily by just typing its name:

```r
> titanicData$age
```

If you look through the results, you will see that most individuals have numbers in this list, but some have NA. These NAs are the people for which we do not have age information.

By the way, the titanic.csv file simply has nothing in the places where there is missing data. When R loaded it, it replaced the empty spots with NA automatically.

Measures of location

This week we want to use R to give some basic descriptive statistics for numerical data.

```r
mean()
```

We have already seen in Week 1 how to calculate the mean of a
vector of data using `mean()`. Unfortunately, if there are missing data we need to tell R how to deal with it.

A (somewhat annoying) quirk of R is that if we try to take the mean of a list of numbers that include missing data, we get an NA for the result!

```r
> mean(titanicData$age)
[1] NA
```

**na.rm=TRUE** To get the mean of all the numbers that we do have, we have to add an option to the `mean()` function. This option is `na.rm = TRUE`.

```r
> mean(titanicData$age, na.rm = TRUE)
[1] 31.19418
```

This tells R to remove ("rm") the NAs before taking the mean. It turns out that the mean age of passengers that we have information for was about 31.2.

`na.rm = TRUE` can be added to many functions in R, including `median()`, as we shall see next.

**median()** The median of a series of numbers is the "middle" number – half of the numbers in the list are greater the median and half are below it. It can be calculated in R by using `median()`.

```r
> median(titanicData$age, na.rm = TRUE)
[1] 30
```

**summary()** A handy function that will return both the mean and median at the same time (along with other information such as the first and third quartiles) is `summary()`.

```r
> summary(titanicData$age)
   Min. 1st Qu.  Median   Mean 3rd Qu.    Max.    NA's
   0.1667 21.0000 30.0000 31.1900 41.0000 71.0000     680
```
From left to right, this output gives us the smallest (minimum) value in the list ("Min."), the first quartile ("1st Qu."), the median, the mean, the third quartile ("3rd Qu."), the largest (maximum) value ("Max."), and finally, the number of individuals with missing values ("NA’s").

The first quartile is the value in the data that is larger than a quarter of the data points. The third quartile is larger than \( \frac{3}{4} \) of the data. These are also called the 25\(^{th}\) percentile and the 75\(^{th}\) percentile, respectively. (You may remember these from boxplots, where the top and bottom of the box mark the 75\(^{th}\) and 25\(^{th}\) percentiles, respectively.)

### Measures of variability

*R* can also calculate measures of the variability of a sample. In this section we’ll learn how to calculate the variance, standard deviation, coefficient of variation and interquartile range of a set of data.

**var()** To calculate the variance of a list of numbers, use *var()*.  

```r
> var(titanicData$age, na.rm = TRUE)
[1] 217.4895
```

Note that *var()*, as well as *sd()* below, have the same need for *na.rm = TRUE* when analyzing data that include missing values.

**sd()** The standard deviation can be calculated by *sd()*.

```r
> sd(titanicData$age, na.rm = TRUE)
[1] 14.74753
```

Of course, the standard deviation is the same as the square root of the variance.

```r
> sqrt(var(titanicData$age, na.rm = TRUE))
[1] 14.74753
```

**Coefficient of variation** Surprisingly, there is no standard function in *R* to calculate the coefficient of variation. You can do this yourself, though, directly from the definition:

```r
> sd(titanicData$age, na.rm = TRUE) / mean(titanicData$age)
[1] 0.09913558
```
IQR()  
The interquartile range (or IQR) is the difference between the third quartile and the first quartile; in other words the range covered by the middle half of the data. It can be calculated easily with IQR().

```
> IQR(titanicData$age, na.rm = TRUE)
[1] 20
```

Note that this is the same as we could calculate by the results from summary() above. The third quartile is 41 and the first quartile is 21, so the difference is $41 - 21 = 20$.

Confidence intervals of the mean  
The confidence interval for an estimate tells us a range of values that is likely to contain the true value of the parameter. For example, in 95% of random samples the 95% confidence interval of the mean will contain the true value of the mean.

R does not have a simple built-in function to calculate only the confidence interval of the mean, but the function that calculates t-tests will give us this information. (See chapters 11 and 12 in the text or Week 7 and 8 of these labs for more on t-tests.) The function `t.test()` has many results in its output. By adding `$conf.int` to this function we only get back the confidence interval for the mean. By default it gives us the 95% confidence interval.

```
> t.test(titanicData$age)$conf.int
[1] 30.04312 32.34524
attr(,”conf.level”) [1] 0.95
```

As the result above shows, the 95% confidence interval of the mean of `age` in the `titanicData` data set is from about 30.0 to 32.3. R also tells us that it used a 95% confidence level for its calculation. (The confidence interval is not so useful in this case, because we actually have information for nearly all the individuals on the Titanic.)
To calculate confidence intervals with a different level of confidence, we can add the option `conf.level` to the `t.test` function. For example, for a 99% confidence interval we can use the following.

```r
> t.test(titanicData$age, conf.level = 0.99)$conf.int
[1] 29.67976 32.70861
attr(,"conf.level")
[1] 0.99
```
**R commands summary**

**Mean**

Vector of numerical variable | Ignore missing data
---|---
`mean(titanicData$Age, na.rm = TRUE)`

**Median**

Vector of numerical variable | Ignore missing data
---|---
`median(titanicData$Age, na.rm = TRUE)`

**Summary**

Vector of numerical variable | Ignore missing data
---|---
`summary(titanicData$Age)`

**Variance**

Vector of numerical variable | Ignore missing data
---|---
`var(titanicData$Age, na.rm = TRUE)`

**Standard deviation**

Vector of numerical variable | Ignore missing data
---|---
`sd(titanicData$Age, na.rm = TRUE)`

**Coefficient of variation**

Vector of numerical variable | Ignore missing data
---|---
`sd(titanicData$Age, na.rm = TRUE) / mean(titanicData$Age, na.rm = TRUE) * 100`

**Interquartile range**

Vector of numerical variable | Ignore missing data
---|---
`IQR(titanicData$Age, na.rm = TRUE)`

**Confidence interval of mean**

Vector of numerical variable | Calculate confidence interval (95% by default)
---|---
`t.test(titanicData$Age)$conf.int`
Activities

Distribution of sample means

Go to the web and open the page at http://www.zoology.ubc.ca/~whitlock/kingfisher/SamplingNormal.htm. This page contains some interactive visualizations that let you play around with sampling to see the distribution of sampling means. Click the button that says “Tutorial” near the bottom of the page and follow along with the instructions.

Confidence intervals

Go back to the web, and open the page at http://www.zoology.ubc.ca/~whitlock/kingfisher/CIMean.htm. This applet draws confidence intervals for the mean of a known population. Click “Tutorial” again, and follow along.

Questions

1. Your TA should provide a data file that includes data that your classmates collected on themselves during the first week of lab, in chapter 1. Open that file in R.

   a. Plot the distribution of heights in the class. Describe the shape of the distribution. Is it symmetric or skewed? Is it unimodal or bimodal?

   Key point: A distribution is skewed if it is asymmetric. A distribution is skewed right if there is a long tail to the right, and skewed left if there is a long tail to the left.

   b. Are there any large outliers that look as though a student used the wrong units for their height measurement. (I.e., are there any that are more plausibly a height given in inches rather than the requested centimeters?) If so, and if this is not likely to be an accurate description of an individual in your class, use filter() from the package dplyr to create a new data set without those rows.

   c. Use R to calculate the mean height of all students in the class.
d. Use \texttt{sd()} to calculate the standard deviation of height.

2. The file "caffeine.csv" contains data on the amount of caffeine in a 16 oz. cup of coffee obtained from various vendors. For context, doses of caffeine over 25 mg are enough to increase anxiety in some people, and doses over 300 to 360 mg are enough to significantly increase heart rate in most people. A can of Red Bull contains 80mg of caffeine.

   a. What is the mean amount of caffeine in 16 oz. coffees?

   b. What is the 95% confidence interval for the mean?

   c. Plot the frequency distribution of caffeine levels for these data in a histogram. Is the amount of caffeine in a cup of coffee relatively consistent from one vendor to another? What is the standard deviation of caffeine level? What is the coefficient of variation?

   d. The file "caffeineStarbucks.csv" has data on six 16 oz. cups of Breakfast Blend coffee sampled on six different days from a Starbucks location. Calculate the mean (and the 95% confidence interval for the mean) for these data. Compare these results to the data taken on the broader sample of vendors in the first file. Describe the difference.

3. A confidence interval is a range of values that are likely to contain the true value of a parameter. Consider the "caffeine.csv" data again.

   a. Calculate the 99% confidence interval for the mean caffeine level.

   b. Compare this 99% confidence interval to the 95% confidence interval you calculate in question 2b. Which confidence interval is wider (i.e., spans a broader range)? Why should this one be wider?

   c. Let’s compare the quantiles of the distribution of caffeine to this confidence interval. Approximately 95% of the data values should fall between the 2.5% and 97.5% quantiles of the distribution of caffeine. (Explain why this is true.) We can use \texttt{R} to calculate the 2.5% and 97.5% quantiles with a command like the following: (Replace "\texttt{datavector}" with the name of the vector of your caffeine data.)

   \begin{verbatim}
   > quantile(datavector, c(0.025, 0.975), na.rm =TRUE)
   \end{verbatim}

   Are these the same as the boundaries of the 95% confidence interval? If not, why not? Which should bound a smaller region, the quantile or the confidence interval of the mean?
4. Return to the class data set "students.csv." Find the mean value of "number of siblings." Add one to this to find the mean number of children per family in the class.

   a. The mean number of offspring per family twenty years ago was about 2. Is the value for this class similar, greater, or smaller? If different, think of reasons for the difference.

   b. Are the families represented in this class systematically different from the population at large? Is there a potential sampling bias?

   c. Consider the way in which the data were collected. How many families with zero children are represented? Why? What effect does this have on the estimated mean family size of all couples?

5. Return to the data on countries of the world, in "countries.csv" and the data from the class in "students.csv." Plot the distributions for ecological_footprint_2000, cell_phone_subscriptions_per_100_people_2012, and life_expectancy_at_birth_female.

   a. For each variable, plot a histogram of the distribution. Is the variable skewed? If so, in which direction?

   b. For each variable, calculate the mean and median. Are they similar? Match the difference in mean and median to the direction of skew on the histogram. Do you see a pattern?