Central Tendency (R Instructional Worksheet)

November 14, 2017

Contents

Descriptive Statistics

We will use the chimpanzee dataset that we imported earlier to explore basic descriptive statistics. But first, we need to look at the structure of the dataset.

```
str(chimp)
```

This dataset has two variables, Year and Population. Let's explore the Population variable.

```
chimp$Population  #display the data in the Population column
```

Minimum – smallest value

```
min(chimp$Population)
```

Maximum – largest value

```
max(chimp$Population)
```

Range (Minimum and Maximum)

```
range(chimp$Population)
```

Median – middle number when all of the values are organized from smallest to largest

```
median(chimp$Population)
```

Quantile

```
quantile(chimp$Population)
```

Mean is a measure of central tendency of the data, and the code to calculate it in R is as follows:

```
mean(chimp$Population)
```

The summary command calculates a group of the descriptive statistics all at once.

```
summary(chimp$Population)
```

All of the values are rounded to the nearest integer before these values are calculated, which is why they are slightly different than the values calculated above.

Missing Values

Datasets are rarely perfect, and there is often data missing. First, here is a simple example with one value missing.

```
a <- c(1, 2, 3, NA, 5)
mean(a)
```

```
## [1] NA
```

```
mean(a, na.rm = TRUE)
```

```
## [1] 2.75
```

Next, let's us see how to deal with missing data in a real dataset. Let us import the 'lion' dataset. This dataset has the population number of lions in the Serengeti Plains from 1966-1990.

First import the dataset and name it 'lion'.

```
lion <- read.csv("lion.csv", header = T)
```

If you look at the data, you'll see that there are no population sizes for the years 1969-1973.

Now, what happens if we try to find the mean of the population size?

```
mean(lion$Population)
```

```
## [1] NA
```

We can use the trick we just learned to get the mean population size.

```
mean(lion$Population, na.rm = TRUE)
```

```
## [1] 34.05
```

However, instead of having to include the 'na.rm' argument every time, we can just delete the rows that are missing data.

```
lion2 <- na.omit(lion)
```

Problems

1. Import the Lion and Black Rhino files into R that were used in previous assignments ('lionCrater.csv' and 'blackRhinoCrater.csv').

2. What is the minimum and maximum population size for lions?

3. What is the mean, median, and standard deviation of the population size of lions?

4. What is the minimum and maximum population size for black rhinos?

5. What are the quantile (0%,25%,50%,75%,100%) population sizes for black rhinos?

6. Use the *summary* command to find the descriptive statistics of population size for black rhinos. What is the mean and median?