FITTING MODELS (CENTRAL TENDENCY)

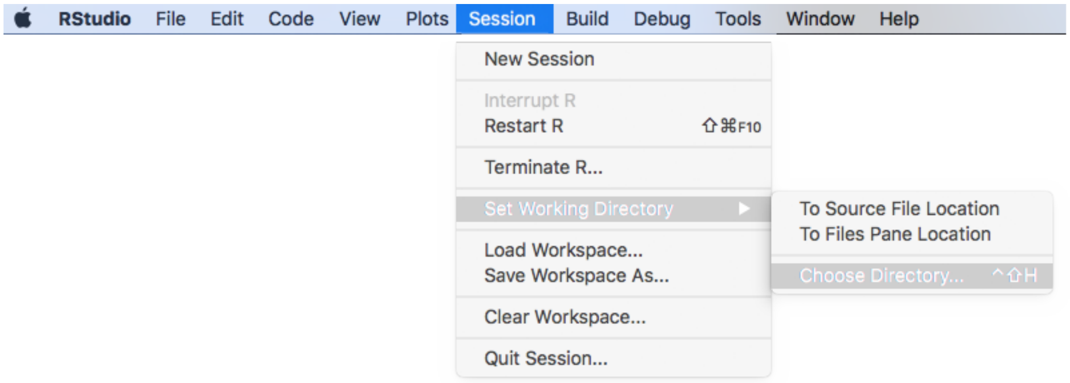
This handout is one of a series that accompanies *An Adventure in Statistics: The Reality Enigma* by me, Andy Field. These handouts are offered for free (although I hope you will buy the book).[1]

Overview

In this handout we will look at how to do the procedures explained in Chapter 4 using R an open-source free statistics software. If you are not familiar with **R** there are many good websites and books that will get you started; for example, if you like *An Adventure In Statistics* you might consider looking at my book *Discovering Statistics Using R*.

Some basic things to remember

- RStudio: I assume that you're working with RStudio because most sane people use this software instead of the native **R** interface. You should download and install both R and RStudio. A few minutes on google will find you introductions to RStudio in the event that I don't write one, but these handouts don't particularly rely on RStudio except in settingthe working directory (see below).
- Dataframe: A dataframe is a collection of columns and rows of data, a bit like a spreadsheet (but less pretty)
- Variables: variables in dataframes are referenced using the $ symbol, so *catData$fishesEaten* would refer to the variable called *fishesEaten* in the dataframe called *catData*
- Case sensitivity: **R** is case sensitive so it will think that the variable *fishesEaten* is completely different to the variable *fisheseaten*. If you get errors make sure you check for capitals or lower case letters where they shouldn't be.
- Working directory: you should place the data files that you need to use for this handout in a folder, then set it as the working directory by navigating to that folder when you execute the command accessed through the **Session>Set Working Directory>Choose Directory ...** menu



[1] This document is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, basically you can use it for teaching and non-profit activities but not meddle with it.

Packages used in this chapter

We install packages using the *install.package("package name")* function, so if you haven't already got them installed execute the following commands:

```
install.packages("modeest")
install.packages("pastecs")
```

This installs the packages **modeest**, which we use to get the mode, **pastecs** which we use to get descriptive statistics Once installed we need to initialise these packages in the current **R** session by using the *library(packagename)* function as follows.

```
library(modeest)
library(pastecs)
```

Enter data

In the book, Milton teaches Zach how to quantify the typical score in the distribution by using measures of central tendency (the mean, mode and median). They begin by looking at some scores from the relationship assessment scale (RAS). Zach has 10 scores that represent his girlfriend Alice's relationship satisfaction over the past 10 weeks. The scale ranged from 7 (not satisfied at all) to 35 (very satisfied). Alice's scores over 10 weeks are: 32, 30, 28, 30, 30, 29, 31, 29, 31, 11. We can enter these data by creating a dataframe called **AliceRAS** that contains the 10 scores (line 1). This creates a dataframe with one column that we name *Rating* in line 2. We can look at the dataframe by executing the name of the dataframe (line 3). If you execute these three lines of script you will create the dataframe, name the column, and display the results. We access these scores by typing *AliceRAS$Rating*, which means 'the variable *Rating* in the dataframe *AliceRAS*'. (Incidentally, you don't have to create a dataframe for this example but because we'll be using them in other handouts I want you to get into the habit of using them.)

```
AliceRAS<-as.data.frame(c(32, 30, 28, 30, 30, 29, 31, 29, 31, 11))
names(AliceRAS)<-c("Rating")
AliceRAS
```

```
##    Rating
## 1      32
## 2      30
## 3      28
## 4      30
## 5      30
## 6      29
## 7      31
## 8      29
## 9      31
## 10     11
```

Alternatively, you can read the data in from the csv file on the companion website for the book. To do this, execute the command below, which uses the *file.choose()* function to open a dialog box so that you can navigate to the file that you want to open, which in this case will be *AiS Ch 04 AliceRAS.csv*. The rest of the command tells R to import this file into a dataframe called AliceRAS.

```
AliceRAS<-read.csv(file.choose())
```

Central tendency

The mode

We can obtain the mode using the *mlv()* function in the **modeest** package. We place the variable that we want to know the mode for within the function (*AliceRAS$Rating*) and the *method = "mfv"* tells **R** how to estimate the mode, there are numerous ways but *mfv* gives us the most frequent value, which is the method explained in the book. Execute this command and you'll see that it returns a value of 30, the same as in the book.

```
mlv(AliceRAS$Rating, method = "mfv")
## Mode (most likely value): 30 
## Bickel's modal skewness: -0.1 
## Call: mlv.default(x = AliceRAS$Rating, method = "mfv")
```

The mean and median

We can obtain a range of descriptive statistics in **R** using the *stat.desc()* function in the **pastecs** package. There are other ways too, but I like this package. All we need to do is to reference the variable inside the function. The output tells us the median is 30 and the mean is 28.1 as computed in the book.

```
stat.desc(AliceRAS$Rating)
##      nbr.val     nbr.null       nbr.na          min          max 
##   10.0000000    0.0000000    0.0000000   11.0000000   32.0000000 
##        range          sum       median         mean      SE.mean 
##   21.0000000  281.0000000   30.0000000   28.1000000    1.9347696 
## CI.mean.0.95          var      std.dev     coef.var 
##    4.3767529   37.4333333    6.1182786    0.2177323
```

The effect of outliers

In the chapter Milton showed how the mean, mode and median are affected by outliers by removing the score of 11, which was Alice's score in the last week and was much lower than her other scores. We can ignore this outlier in several ways in **R** including creating a new data frame that doesn't include the value 11 (and there are numerous ways to achieve this). However, because we have a single variable, and the score we want to ignore is the 10th score the easiest thing to do is instead of referring to *AliceRAS$Rating* in our functions, which will use all ten scores, we instead refer to *AliceRAS$Rating[1:9]*, which will use only scores 1 to 9 (the 1:9 tells **R** to use scores 1 to 9 inclusive).

We can, therefore, recompute the mode, mean and median by using the same commands as before, but referring only to the first 9 scores. If you execute these commands, you'd see that with the outlier excluded the mode and median do not change (they are both still 30),but the mean increases to 30. Again, these values match the book.

```
mlv(AliceRAS$Rating[1:9], method = "mfv")
## Mode (most likely value): 30 
## Bickel's modal skewness: 0 
## Call: mlv.default(x = AliceRAS$Rating[1:9], method = "mfv")
stat.desc(AliceRAS$Rating[1:9])
##      nbr.val     nbr.null       nbr.na          min          max 
##   9.00000000   0.00000000   0.00000000  28.00000000  32.00000000 
##        range          sum       median         mean      SE.mean 
##   4.00000000 270.00000000  30.00000000  30.00000000   0.40824829
## CI.mean.0.95          var      std.dev     coef.var
##   0.94142225   1.50000000   1.22474487   0.04082483
```

The 'fit' of the mean: variance

In the book Milton explains to Zach how the fit of the mean (how well it represents the scores it summarizes) can be quantified with the variance and standard deviation. In fact, we already obtained these values when we obtained the descriptive statistics from the *stat.dec()* function. With the outlier excluded the variance was 1.5 and the standard deviation was 1.22, but with all scores included the variance was 37.43 (and the standard deviation was 6.12). Look for these values in the outputs above and note that they match the calculations in the book.

Dispersion

Milton also explained to Zach that the range and interquartile range could be used to look at the dispersion of scores. The range has already been computed when we obtained the descriptive statistics from the *stat.dec()* function. Look at the previous outputs and you'll see that with the outlier included the range was 21, but with it excluded the range was only 4 (as in the book).

The inter-quartile range (IQR) can be obtained using the function *IQR()* and we can obtain quantiles using the function *quantile()*. The interquartile range is based on the upper and lower quartile. The lower quartile cuts off the bottom 25% of scores, and the upper quartile cuts off the top 25%. We can use the *quantile()* function to get any quantiles, but I have specified below to get the 25% (0.25 as a proportion) quantile, the lower quartile, the 50% (0.50 as a proportion) quantile, the median, and the 75% (0.75 as a proportion) quantile, the upper quartile. These values are wrapped in the command *c()*, which binds the three values as a single object to pass into the function. To get the IQR we simple place the name of the variable for which we want the IQR into the *IQR()* function and execute. When you execute these commands you'll see that the lower quartile is 29, the median is 30, and the upper quartile is 31. The resulting interquartile range is 2 (as explained in the book).

```
quantile(AliceRAS$Rating, c(0.25, 0.5, 0.75), type = 2)
## 25% 50% 75%
##  29  30  31
IQR(AliceRAS$Rating, type = 2)
## [1] 2
```

This handout is written to be used in conjunction with: Field, A. P. (2016). An adventure in statistics; the reality enigma. London: Sage.