Robust summaries

Outliers and more robust metrics

Outliers are extreme values which affect the value of the sample mean and standard deviation. In order to handle this problem, we need more robust summaries regarding outliers. We will use one of the datasets included in R, which contains weight of chicks in grams as they grow from day 0 to day 21. This dataset also splits up the chicks by different protein diets, which are coded from 1 to 4.

Median An example of more robust summary is the median which is simply the middle point of the data.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(rafalib)
library(readr)

data(ChickWeight)
head(ChickWeight)
```

```
##   weight Time Chick Diet
## 1     42    0     1    1
## 2     51    2     1    1
## 3     59    4     1    1
## 4     64    6     1    1
## 5     76    8     1    1
## 6     93   10     1    1
```

```
chick = reshape(ChickWeight, idvar=c("Chick","Diet"), timevar="Time",direction="wide")
head(chick)
```

```
##    Chick Diet weight.0 weight.2 weight.4 weight.6 weight.8 weight.10 weight.12
## 1      1    1       42       51       59       64       76        93       106
## 13     2    1       40       49       58       72       84       103       122
## 25     3    1       43       39       55       67       84        99       115
## 37     4    1       42       49       56       67       74        87       102
## 49     5    1       41       42       48       60       79       106       141
## 61     6    1       41       49       59       74       97       124       141
##    weight.14 weight.16 weight.18 weight.20 weight.21
## 1        125       149       171       199       205
## 13       138       162       187       209       215
```

```
## 25       138       163       187       198       202
## 37       108       136       154       160       157
## 49       164       197       199       220       223
## 61       148       155       160       160       157
```

```
#We also want to remove any chicks that have missing observations at any time points (NA for "not avail
chick = na.omit(chick)
extended_chickweight4 <- c(3000,chick$weight.4)

mean_ratio <- mean(extended_chickweight4) / mean(chick$weight.4)
median_ratio <- median(extended_chickweight4) / median(chick$weight.4)
cat('Mean change ratio:',mean_ratio, ' and Median change ratio:',median_ratio)
```

```
## Mean change ratio: 2.062407  and Median change ratio: 1
```

Median Absolute Deviation (MAD) Another metric we can use is the Median Absolute Deviation (MAD) which is a robust estimate of the standard deviation. To compute the MAD, we need to find, first, the median of our sample. Afterward, we compute the distance of each point to the median. We compute the distance as the absolute value of the difference.Then we take the median of those deviations – that's where the name comes – median absolute deviation– MAD. We multiply by this factor 1.4826 to make the summary statistic unbiased. On average, it's going to be equal to the standard deviation. $MAD = 1.4826 \times median\{|X_i - median(X_i)|\}$

```
sd_ratio <- sd(extended_chickweight4) / sd(chick$weight.4)
mad_ratio <- mad(extended_chickweight4) / mad(chick$weight.4)
cat('Standard deviation change ratio:',sd_ratio,' and MAD ratio:',mad_ratio)
```

```
## Standard deviation change ratio: 101.2859  and MAD ratio: 1
```

Spearman correlation The Spearman correlation is used to measure the relationship between two ordinal variables. This metric is also used to measure relationship between two variables that are related but not linearly. To calculate the Spearman correlation, we need first to rank the scores (the data points values). The lowest score has the lowest rank and the highest point value has the highest rank. We will now use the ranks to compute

$$rs = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}\sqrt{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}}$$

Since Instead of looking at the values, we look at the ranks the Spearman correlation metric is not impacted by outliers.

```
extended_chickweight21 <- c(3000,chick$weight.21)
pearson_ratio <- cor(extended_chickweight4,extended_chickweight21) / cor(chick$weight.4,chick$weight.21
spearman_ratio <- cor(extended_chickweight4,extended_chickweight21,method="spearman") / cor(chick$weigh
cat('Pearson correlation change ratio:',pearson_ratio,' and Spearman correlation ratio:',spearman_ratio
```

```
## Pearson correlation change ratio: 2.370719  and Spearman correlation ratio: 1.084826
```