Assessment 05 - Robust Summaries with Outliers

Gabriele Mineo - Harvard Data Science Professional

Exploring the Galton Dataset - Average and Median

For this chapter, we will use height data collected by Francis Galton for his genetics studies. Here we just use height of the children in the dataset:

```
library(HistData)
data(Galton)
x <- Galton$child
```

Instructions

Compute the average and median of these data. Note: do not assign them to a variable.

```
library(HistData)
data(Galton)
x <- Galton$child
mean(x)
```

```
## [1] 68.08847
```

```
median(x)
```

```
## [1] 68.2
```

Exploring the Galton Dataset - SD and MAD

Now for the same data compute the standard deviation and the median absolute deviation (MAD).

Instructions

Compute the standard deviation and the median absolute deviation of these data.

```
library(HistData)
data(Galton)
x <- Galton$child
mad(sd(x))
```

```
## [1] 0
```

Error impact on average

In the previous exercises we saw that the mean and median are very similar and so are the standard deviation and MAD. This is expected since the data is approximated by a normal distribution which has this propoerty.

Now suppose that suppose Galton made a mistake when entering the first value, forgetting to use the decimal point. You can imitate this error by typing:

```
library(HistData)
data(Galton)
x <- Galton$child
x_with_error <- x
x_with_error[1] <- x_with_error[1]*10
```

The data now has an outlier that the normal approximation does not account for. Let's see how this affects the average.

Instructions

- Report how many inches the average grow after this mistake. Specifically, report the difference between the average of the data with the mistake `x_with_error` and the data without the mistake `x`.

```
library(HistData)
data(Galton)
x <- Galton$child
x_with_error <- x
x_with_error[1] <- x_with_error[1]*10
mean(x_with_error)- mean(x)
```

```
## [1] 0.5983836
```

Error impact on SD

In the previous exercise we saw how a simple mistake can result in the average of our data increasing more than half a foot, which is a large difference in practical terms. Now let's explore the effect this outlier has on the standard deviation.

Instructions

- Report how many inches the SD grows after this mistake. Specifically, report the difference between the SD of the data with the mistake x_with_error and the data without the mistake x.

```
x_with_error <- x
x_with_error[1] <- x_with_error[1]*10
sd(x_with_error)-sd(x)
```

```
## [1] 15.6746
```

Error impact on median

In the previous exercises we saw how one mistake can have a substantial effect on the average and the standard deviation.

Now we are going to see how the median and MAD are much more resistant to outliers. For this reason we say that they are robust summaries.

Instructions

- Report how many inches the median grows after the mistake. Specifically, report the difference between the median of the data with the mistake `x_with_error` and the data without the mistake `x`.

```
x_with_error <- x
x_with_error[1] <- x_with_error[1]*10
median(x_with_error) - median(x)
```

```
## [1] 0
```

Error impact on MAD

We saw that the median barely changes. Now let's see how the MAD is affected.

Instructions

- Report how many inches the MAD grows after the mistake. Specifically, report the difference between the MAD of the data with the mistake `x_with_error` and the data without the mistake `x`.

```
x_with_error <- x
x_with_error[1] <- x_with_error[1]*10
mad(x_with_error) - mad(x)
```

```
## [1] 0
```

Usefulness of EDA

How could you use exploratory data analysis to detect that an error was made?

Possible Answers

- Since it is only one value out of many, we will not be able to detect this.
- We would see an obvious shift in the distribution.
- A boxplot, histogram, or qq-plot would reveal a clear outlier. [X]
- A scatter plot would show high levels of measurement error.

Using EDA to explore changes

We have seen how the average can be affected by outliers. But how large can this effect get? This of course depends on the size of the outlier and the size of the dataset.

To see how outliers can affect the average of a dataset, let's write a simple function that takes the size of the outlier as input and returns the average.

Instructions

Write a function called error_avg that takes a value `k` and returns the average of the vector `x` after the first entry changed to `k`. Show the results for `k=10000` and `k=-10000`.

```
x <- Galton$child
error_avg <- function(k){
  x[1] <- k
  mean(x)
}
error_avg(10000)
```

```
## [1] 78.79784
```

```
error_avg(-10000)
```

```
## [1] 57.24612
```