11 Robust summaries

Soyoung Park

Pusan National University
Department of Statistics

Outliers

Outliers are very common in data science. Data recording can be complex and it is common to observe data points generated in error.

For example, an old monitoring device may read out nonsensical measurements before completely failing. Human error is also a source of outliers, in particular when data entry is done manually.

How do we distinguish an outlier from measurements that were too big or too small simply due to expected variability? This is not always an easy question to answer, but we try to provide some guidance.

Outliers

Suppose a colleague is charged with collecting demography data for a group of males. The data report height in feet and are stored in the object:

```
library(tidyverse)
library(dslabs)
data(outlier_example)
str(outlier_example)
#>  num [1:500] 5.59 5.8 5.54 6.15 5.83 5.54 ...
```

Outliers

Our colleague uses the fact that heights are usually well approximated by a normal distribution and summarizes the data with average and standard deviation:

```
mean(outlier_example)
#> [1] 6.1
sd(outlier_example)
#> [1] 7.8
```

and writes a report on the interesting fact that this group of males is much taller than usual. The average height is over six feet tall.

Outliers

Using your data science skills, however, you notice something else that is unexpected: the standard deviation is over 7 feet.

Adding and subtracting two standard deviations, you note that 95% of this population will have heights between -9.489, 21.697 feet, which does not make sense.

Outliers

A quick plot reveals the problem:

```
boxplot(outlier_example)
```

There appears to be at least one value that is nonsensical, since we know that a height of 180 feet is impossible. The boxplot detects this point as an outlier.

Median

When we have an outlier like this, the average can become very large. The median, defined as the value for which half the values are smaller and the other half are bigger, is robust to such outliers. No matter how large we make the largest point, the median remains the same.

Median

With this data the median is:

```
median(outlier_example)
#> [1] 5.74
```

which is about 5 feet and 9 inches.

The median is what boxplots display as a horizontal line.

The inter quartile range (IQR)

The box in boxplots is defined by the first and third quartile. These are meant to provide an idea of the variability in the data: 50% of the data is within this range.

The difference between the 3rd and 1st quartile (or 75th and 25th percentiles) is referred to as the inter quartile range (IQR). As is the case with the median, this quantity will be robust to outliers as large values do not affect it.

We can do some math to see that for normally distributed data, the IQR / 1.349 approximates the standard deviation of the data had an outlier not been present.

The inter quartile range (IQR)

We can see that this works well in our example since we get a standard deviation estimate of:

```
IQR(outlier_example) / 1.349
#> [1] 0.245
```

which is about 3 inches.

Tukey's definition of an outlier

This definition of outlier was introduced by Tukey. If we define the first and third quartiles as Q_1 and Q_3, respectively, then an outlier is anything outside the range:

$$[Q_1 - 1.5 \times (Q_3 - Q_1),\ Q_3 + 1.5 \times (Q_3 - Q_1)]$$

Tukey's definition of an outlier

When the data is normally distributed, the standard units of these values are:

```
q3 <- qnorm(0.75)
q1 <- qnorm(0.25)
iqr <- q3 - q1
r <- c(q1 - 1.5*iqr, q3 + 1.5*iqr)
r
#> [1] -2.7  2.7
```

Tukey's definition of an outlier

Using the `pnorm` function, we see that 99.3% of the data falls in this interval.

```
pnorm(r)
#> [1] 0.003488302 0.996511698
```

Tukey's definition of an outlier

Keep in mind that this is not such an extreme event: if we have 1000 data points that are normally distributed, we expect to see about 7 outside of this range. But these would not be outliers since we expect to see them under the typical variation.

If we want an outlier to be rarer, we can increase the 1.5 to a larger number. Tukey also used 3 and called these *far out outliers*. With a normal distribution, 100% of the data falls in this interval. This translates into about 2 in a million chance of being outside the range.

Tukey's definition of an outlier

In the `geom_boxplot` function, this can be controlled by the `outlier.size` argument, which defaults to 1.5.

The 180 inches measurement is well beyond the range of the height data:

```
max_height <-
  quantile(outlier_example, 0.75) + 3*IQR(outlier_example)
max_height
#>   75%
#> 6.91
```

Tukey's definition of an outlier

we take this value out, we can see that the data is in fact normally distributed as expected:

```
x <- outlier_example[outlier_example < max_height]
qqnorm(x) # Q-Q plot
qqline(x)
```

Median absolute deviation

Another way to robustly estimate the standard deviation in the presence of outliers is to use the median absolute deviation (MAD).

To compute the MAD, we first compute the median, and then for each value we compute the distance between that value and the median.

Median absolute deviation

The MAD is defined as the median of these distances. The `mad` function already incorporates this correction. For the height data, we get a MAD of:

```
mad(outlier_example)
#> [1] 0.237
```

which is about 3 inches.

Exercises

We are going to use the **HistData** package. If it is not installed you can install it like this:

```
install.packages("HistData")
```

Load the height data set and create a vector `x` with just the male heights used in Galton's data on the heights of parents and their children from his historic research on heredity.

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

Exercises

1. Compute the average and median of these data.
2. Compute the median and median absolute deviation of these data.

Exercises

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

```
x_with_error <- x
x_with_error[1] <- x_with_error[1]*10
```

How many inches does the average grow after this mistake?

Exercises

4. How many inches does the SD grow after this mistake?
5. How many inches does the median grow after this mistake?
6. How many inches does the MAD grow after this mistake?

Exercises

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

a) Since it is only one value out of many, we will not be able to detect this.

b) We would see an obvious shift in the distribution.

c) A boxplot, histogram, or qq-plot would reveal a clear outlier.

d) A scatterplot would show high levels of measurement error.

Exercises

8. How much can the average accidentally grow with mistakes like this? 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`.