# Interpretation of statistics

Zhaoqi Li

We know that in common statistic analyses, summary statistics such as mean, median, etc. are commonly used. Also, one powerful test method, hypothesis test, is also widely used in social science fields to explore the relationship between variables, and a section on statistical analysis is even required in a lot of publications. Therefore, people are paying great attention to these statistics like confidence intervals and p-values. However, they are not always good representatives of the data. In this tutorial, we will demonstrate examples when these statistics do not work.

## Mean

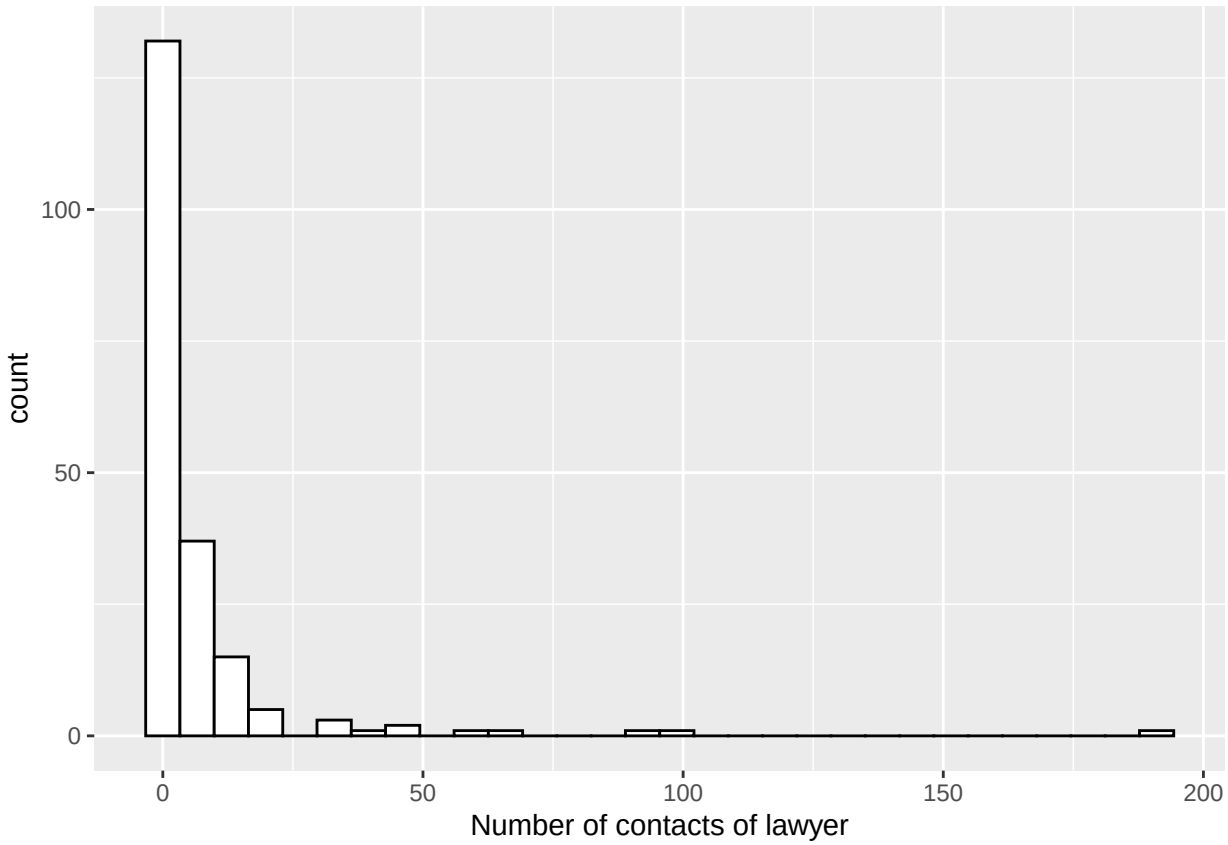
We use the example of dataset which represents lawyers' ratings of state judges in the US Superior Court. We will specifically look at the number of contacts of lawyer with judge, and we are interested in the question: how many lawyers does a judge contact with in general? To answer this question, it seems natural to compute the average.

```
mean(df$CONT)
```

```
## [1] 7.29
```

We can see that on average there are 7.29 lawyers that a judge is in contact with. However, is that really the case? We plot the histogram of all the judges.

```
ggplot(df, aes(x=CONT)) +
  geom_histogram(color="black", fill="white") +
  labs(x = "Number of contacts of lawyer")
```




We can see that in general it seems like most people have 5 or less lawyers in contact, which is very different from the average. Then, what is a good statistic we can use to capture this number? We will try calculating the median.

```
median(df$CONT)
```

```
## [1] 2
```

The median gives us 2, which is a reasonable estimate. **Question: Why is median a good representative here while mean is not?**

## Linear regression

Next, we will look at examples of linear regression. Linear regression is a widely used model for exploring relationship between different variables. Usually, to assess if a model is good or not, we will use the statistic called R-squared, a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Therefore, if the R-squared is large, then we tend to think that the model is good, as more information is explained in the model. However, this is also not always the case, as we will demonstrate in the following example of bike share dataset. We are interested in the following question: what is a good model for the number of riders registered for bikeshare?

```
bikes <- read.csv("https://www.macalester.edu/~dshuman1/data/155/bike_share.csv")
```

It seems like actual temperature is a reasonable estimate of the riders registered, so we create the model and assess its R-squared value. We get an R-squared of 0.2916, which is pretty good.

```
rides_model_1 <- lm(riders_registered ~ temp_actual, bikes)
summary(rides_model_1)$r.squared
```

```
## [1] 0.2916129
```

We would like to get a better model. To do this, we generate 300 variables with random values uniformly distributed between 1 and 100.

```
b <- paste(colnames(bikes_full[17:316]), collapse="+")
fm <- as.formula(paste("riders_registered ~ ", b, sep = ""))
```

```
rides_model_2 <- lm(fm, bikes_full)
summary(rides_model_2)$r.squared
```

```
## [1] 0.4255076
```

We get a value of 0.425, which is much better than the first one. But remember, we are using random values to fit in the model, so this model should not make sense! [Optional] **Question: why does this model have a higher R-squared value?**

## Hypothesis testing

Hypothesis testing is a statistical test for significance. In hypothesis testing, we are interested in seeing if there is a statistical significant relationship between some variables. Because of the power of significance, researchers in social science fields are trying to seek for low p-values, which indicates significant relationship between variables. In general, if the p-value of a hypothesis test is lower than 0.05, then we say that there is a statistical significance. See if you can get a p-value of 0.05 yourself through this website. https://projects.fivethirtyeight.com/p-hacking/