Correlation

Te Rutherford

February 24, 2015

So far...

- Comparing population proportion
- Investigate the effects of categorical variables on binary variables.
  - Example?
- Investigate the effects of categorical variables on continuous variables.
  - Example?
- Today
  - Investigate the effects of continuous variables on continuous variables.

Correlation is one of the coolest concepts in statistics

- The ability to predict the unknown or predict the future is one of the coolest super power you can ask for.
- Correlation is also used to describe the relationship between two quantities, which is extremely useful.
- A simple quantity called correlation can go a really long way. Let's look at an example to build an intuition.

Example

Suppose you are waiting for a bus to go to Boston and you don't know the bus schedule.

- If there are 3 people waiting, you might have to wait 20 minutes. (I might just missed the bus.)
- If there are 30 people waiting, you might have to wait 5 minutes. (People haven't been picked up for a while.)
- If you know what the waiting time is correlated to, then you can predict the future.

Pearson correlation coefficient

- Pearson correlation coefficient is a quantity that ranges from [-1, 1], corresponding to strong negative, weak negative, no correlation, weak positive, strong positive.
- The formula for it is a bit ugly and hairy, so we will use R to compute it.

Example 1 : positive correlation

```
data = read.csv('../datasets/law82.csv')
cor(data$LSAT, data$GPA)
```

```
## [1] 0.76
```

The correlation coefficient is positive. That means . . .

- If the LSAT score is high, then the college GPA is high.
- If the LSAT score is low, then the college GPA is low.

To state in general,

if X and Y are positively correlated, then X is high when Y is high and X is low when Y is low.

Example 2 : negative correlation

```
data = read.csv('../datasets/waiting_time.csv')
cor(data$num_people, data$waiting_time)
```

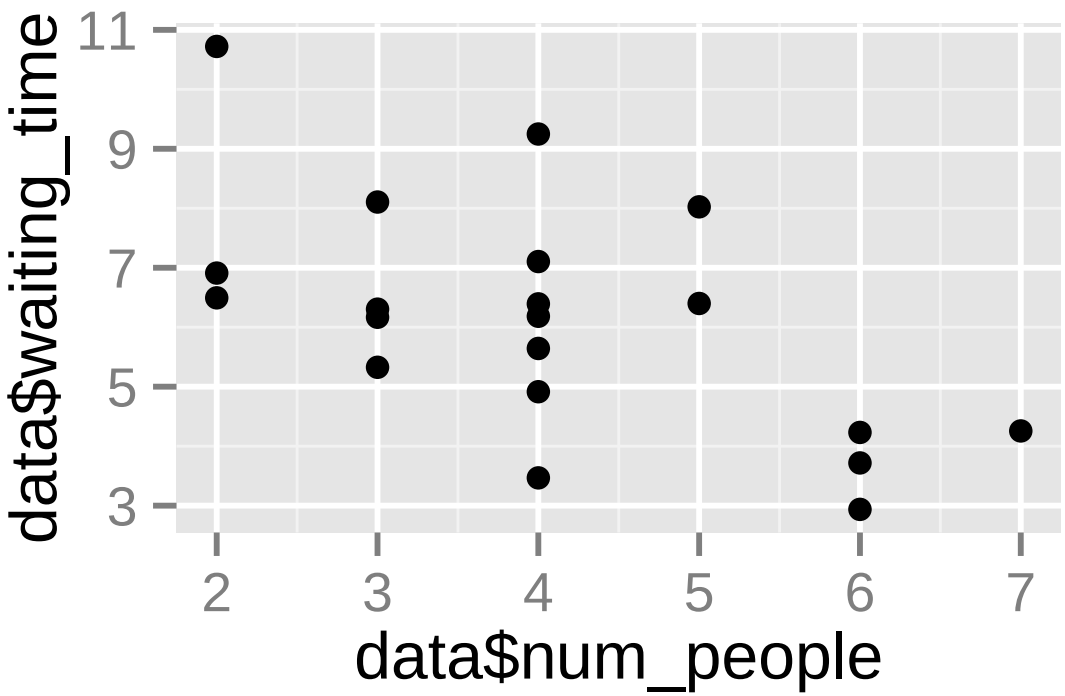
```
## [1] -0.5884
```

The correlation coefficient is negative. That means... $>$ if X and Y are positively correlated, then X is high when Y is low and X is low when Y is high.

Visualizing correlation: Scatterplot

Correlation is just a summary statistic (like mean and median) describing a relationship between two variables. Sometimes it is more helpful to show all data points (like histogram).

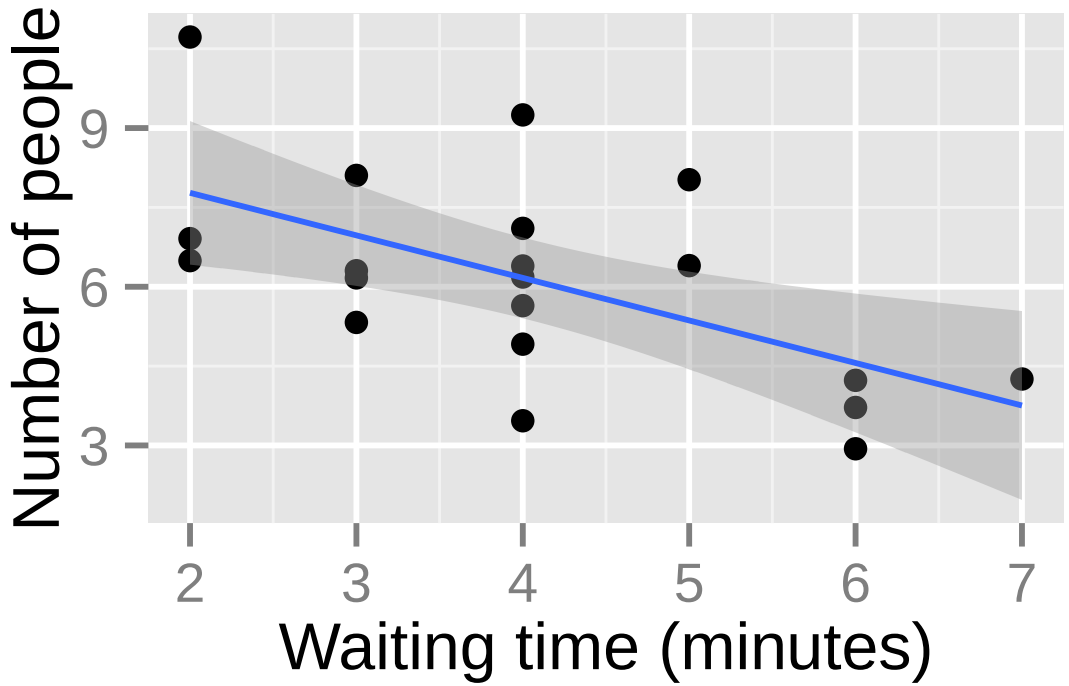
```
library(ggplot2)
data = read.csv('../datasets/waiting_time.csv')
ggplot(aes(x=data$num_people, y=data$waiting_time),
       data=data) + geom_point()
```



Linear trend line on scatterplot

Even without calculating the correlation coefficient, you can see the linear trend (a straight line). You can add a trend line.

```
ggplot(aes(x=data$num_people, y=data$waiting_time),
  data=data) + geom_point()+ geom_smooth(method='lm') +
  ylab("Number of people") + xlab("Waiting time (minutes)")
```



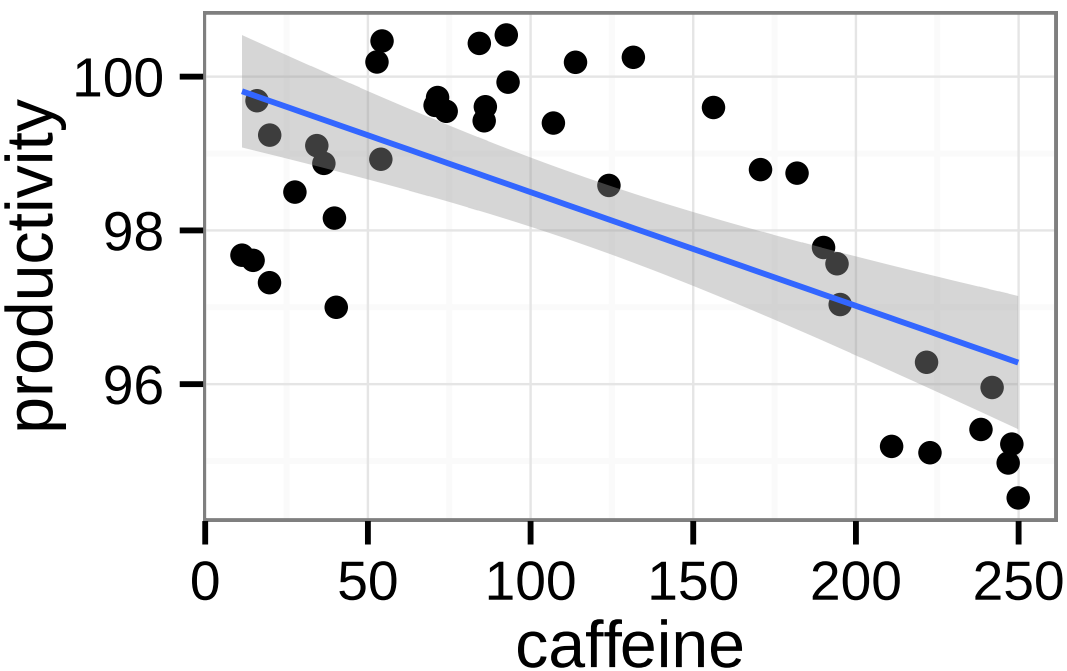
Pearson correlation coefficient only indicate linear correlation

Linear correlation means that you can draw a straight line on your scatterplot and the points will scatter not too far from it.

Example

Caffeine consumption vs productivity (e.g. number of pages one can read in three hours). We survey the amount of caffeine consumption and the number of emails the employees can send out in two hours.

```
caffeine_data = read.csv('../datasets/caffeine_data.csv')
ggplot(aes(x=caffeine, y=productivity), data= caffeine_data
  geom_point() + geom_smooth(method='lm') + theme_bw()
```



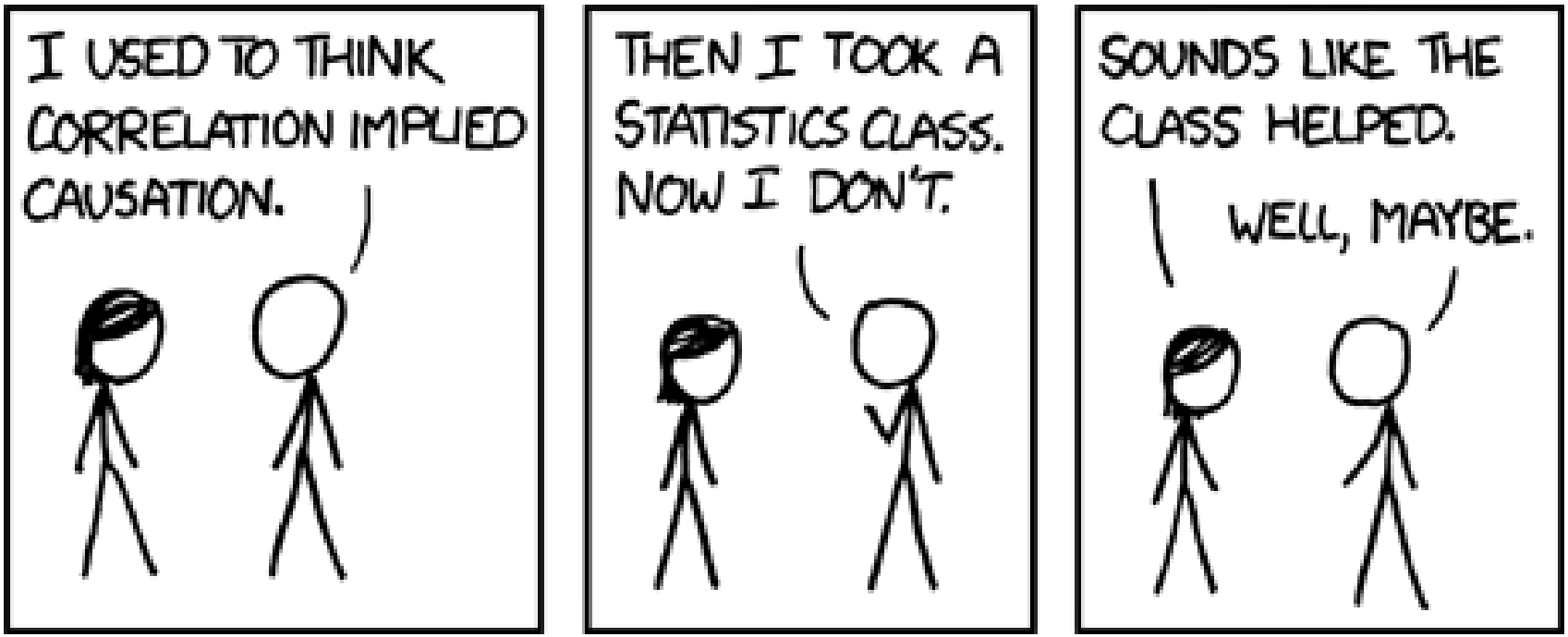
They are clearly correlated. But The points don't scatter around the line that we draw. You actually cannot draw a straight line

```
cor(caffeine_data$caffeine, caffeine_data$productivity)
```

```
## [1] -0.6629
```

We get a negative correlation, which is not the entire story for this dataset.

Correlation does not always imply causation



- It is a crime not to know this mantra. So let me say it again: correlation does not imply causation.

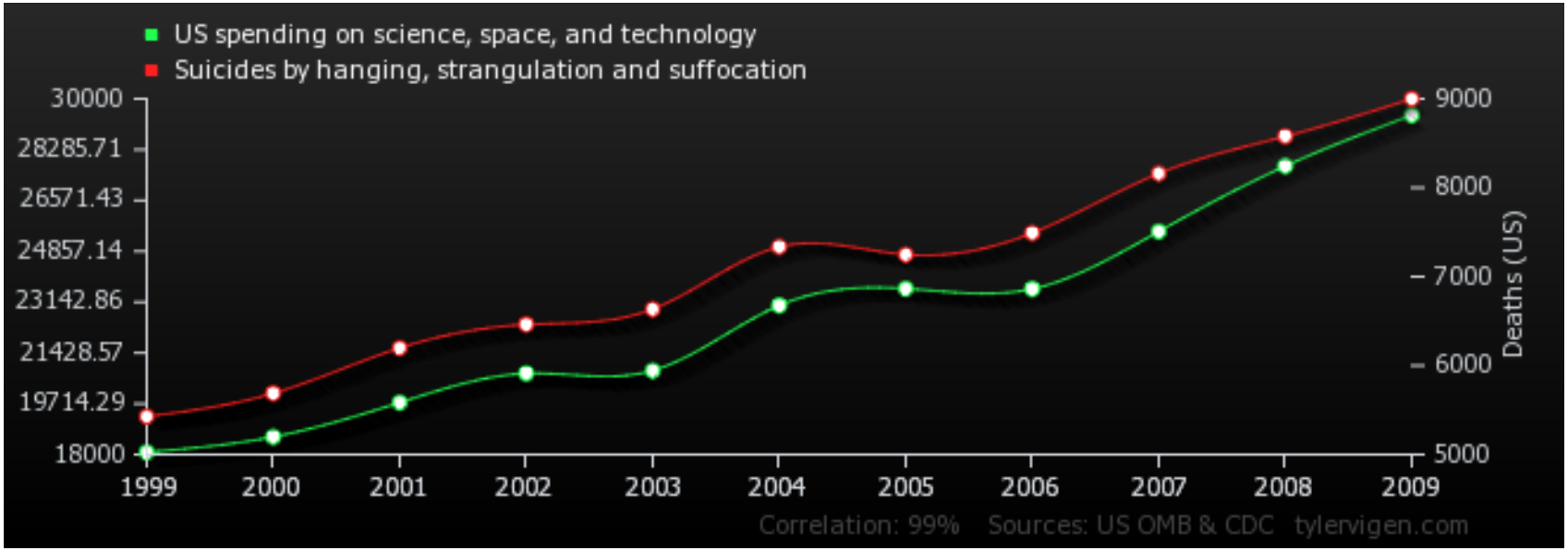
Example

- Back to our example of correlation between waiting time and number of people at the bus stop. We clearly have the negative correlation, but we cannot infer that the number of people CAUSES the waiting time to go down. Correlation does not imply causation.

Example

- If we hack into the high school database and change our gpa to be very high, the college GPA won't increase. We cannot infer that the high school GPA causes the college GPA to go up or down. Correlation does not imply causation.

To be even more absurd, look at this correlation.



We have the correlation, but we cannot infer that the US government has increased the spending on science to cause suicides. Or we cannot infer that the US spending in science makes people suicidal. The obesity rate is correlated with many things. But we can never know the cause of obesity just by looking at the data.