The Multiple Comparisons Problem

Introduction

In this lesson, you'll learn about the problems that can arise from doing multiple comparisons in a single experiment.

Objectives

You will be able to:

- Explain why multiple comparisons increases the likelihood of misleading results
- Explain the concept of spurious correlation
- Use corrections to deal with multiple comparison problems

What is the multiple comparisons problem?

Obtaining an incredibly low p-value does not guarantee that the null-hypothesis is incorrect. For example, a p-value of 0.001 states that there is still a 1 in 1000 chance that the null hypothesis is true. Yet, as you've seen, p-values alone can be misleading. For example, if you perform repeated experiments, at some point you're apt to stumble upon a small p-value, whether or not the null hypothesis is valid.

To restate this, imagine we take 100 scientific studies with a p-value of 0.03. Are all of these conclusions valid? Sadly, probably not. Remember, for any experiment with a p-value of 0.03, there is still a 3% chance that the null-hypothesis is actually true. So collectively, the probability that **all** of these null hypotheses are false is actually quite small. You can be fairly confident in each study, but there is also apt to be a false-conclusion drawn somewhere. (In fact, the p-value itself implies that, on average, 3 of these 100 conclusions will be false.)

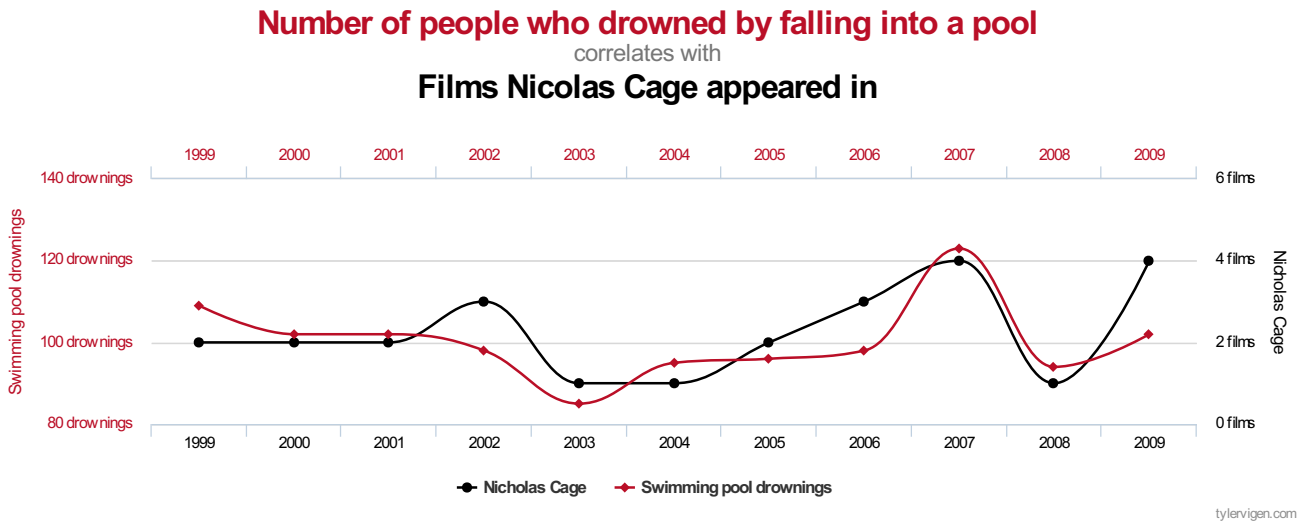
In [1]:

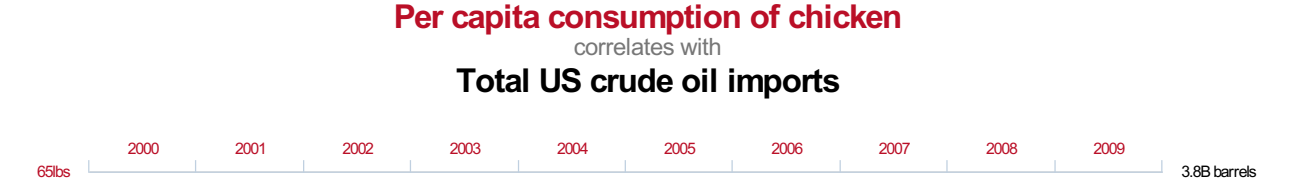
```
0.97**100 # Probability all 100 experiments with p=0.03 are all true
```

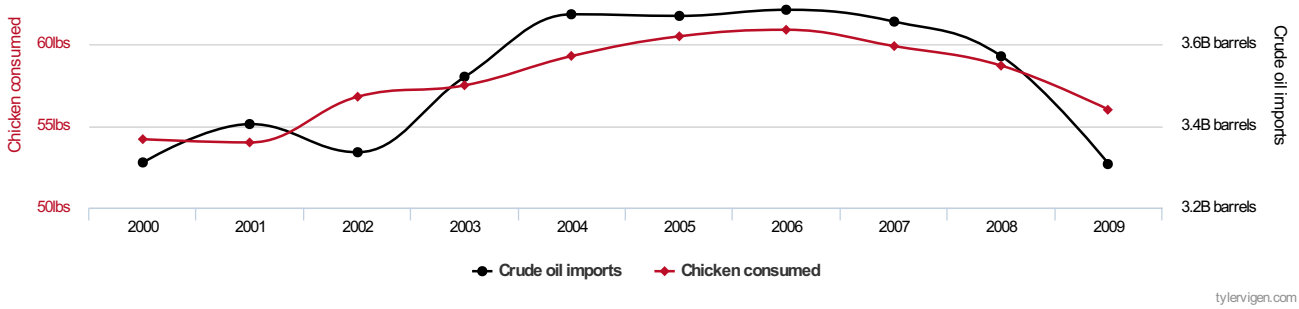
Out[1]:

```
0.04755250792540563
```

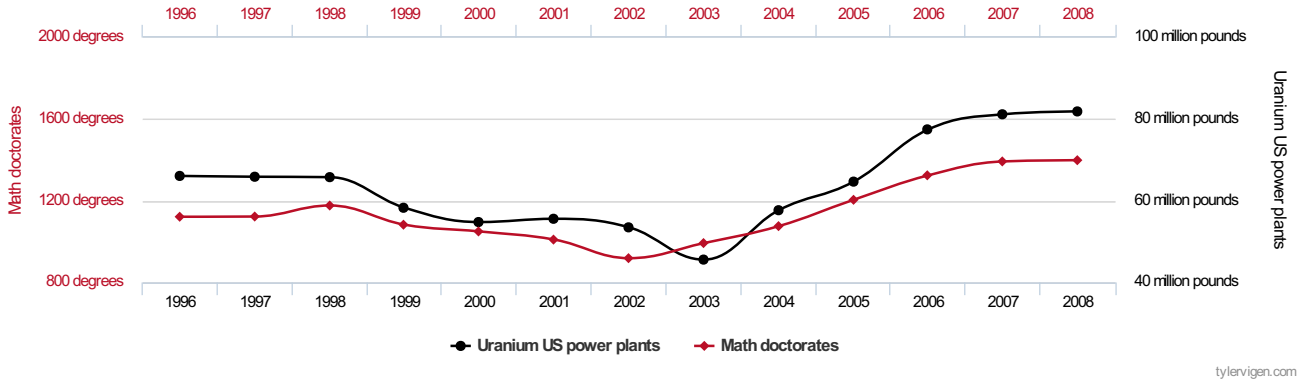
Similarly, if you are testing multiple metrics simultaneously in an experiment, the chances that one of these will satisfy your alpha threshold increases. A fun similar phenomenon is spurious correlation. If we start comparing a multitude of quantities, we are bound to find some quantities that are highly correlated, whether or not an actual relationship exists. Tyler Vigen set out to find such relationships; here are a few entertaining ones (of many):



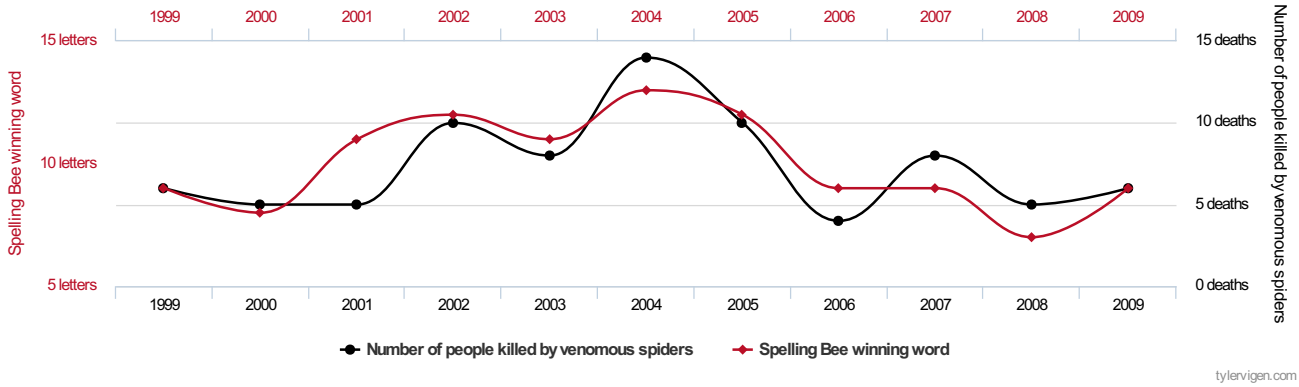




Math doctorates awarded
correlates with
Uranium stored at US nuclear power plants



Letters in Winning Word of Scripps National Spelling Bee
correlates with
Number of people killed by venomous spiders



As we can see, although these graphs show that each pair of quantities is strongly correlated, it seems unreasonable to expect that any of them have any causal relationships. Regardless of what the statistics tell us, there is no relationship through which the length of spelling bee word affects the number of people killed by venomous spiders.

How do multiple comparisons increase the chances of finding spurious correlations?

Spurious correlation is a ***Type 1 Error***, meaning that it's a type of ***False Positive***. We think we've found something important when really there isn't any. With each comparison we make in an experiment, we try to set a really low p-value to limit our exposure to type 1 errors. When we only reject the null hypothesis when $p < 0.05$, for example, we are effectively saying "I'm only going to accept these results as true if there is less than a 5% chance that I didn't actually find anything important, and my data only looks like this due to randomness". However, when we make ***multiple comparisons*** by checking for many things at once, each of the small risks of a Type 1 error becomes cumulative!

Here's another easy to way to phrase this -- a p-value threshold of less than 0.05 means that we will only make a Type 1 error 1 in every 20 times. This means that statistically, if we have 20 findings where the p-value is less than 0.05 at the same time, 1 of them is almost guaranteed to be a Type 1 error (False Positive) -- but we have no idea of which one!

The Bonferroni correction

Back to the problem of multiple comparisons. Due to the cumulative risk of drawing false conclusions when statistically testing

Back to the problem of multiple comparisons. Due to the cumulative risk of drawing false conclusions when statistically testing multiple quantities simultaneously, statisticians have devised methods to minimize the chance of type 1 errors. One of these is the ***Bonferroni correction***. With the Bonferroni correction, you divide α by the number of comparisons you are making to set a new, adjusted threshold rejecting the null hypothesis.

For example, if you desire $\alpha = 0.05$, but are making 10 comparisons simultaneously, the Bonferroni Correction would advise you set our adjusted p-value threshold to $\frac{0.05}{10} = 0.005$! The stricter p-value threshold helps control for Type 1 errors. This doesn't mean that you are immune to them -- it just helps reduce the cumulative chance that one occurs. That said, the effective power of these tests is therefore reduced (and in turn type 2 errors are more likely).

Additional Resources

- Tyle Vigen - Spurious correlations
- Nick Cage movies vs. drownings, and more strange (but spurious) correlations

Summary

In this lesson, you learned about the problems that can arise from doing multiple comparisons in a single experiment, as well as some entertaining spurious correlations that exist with real-world data.