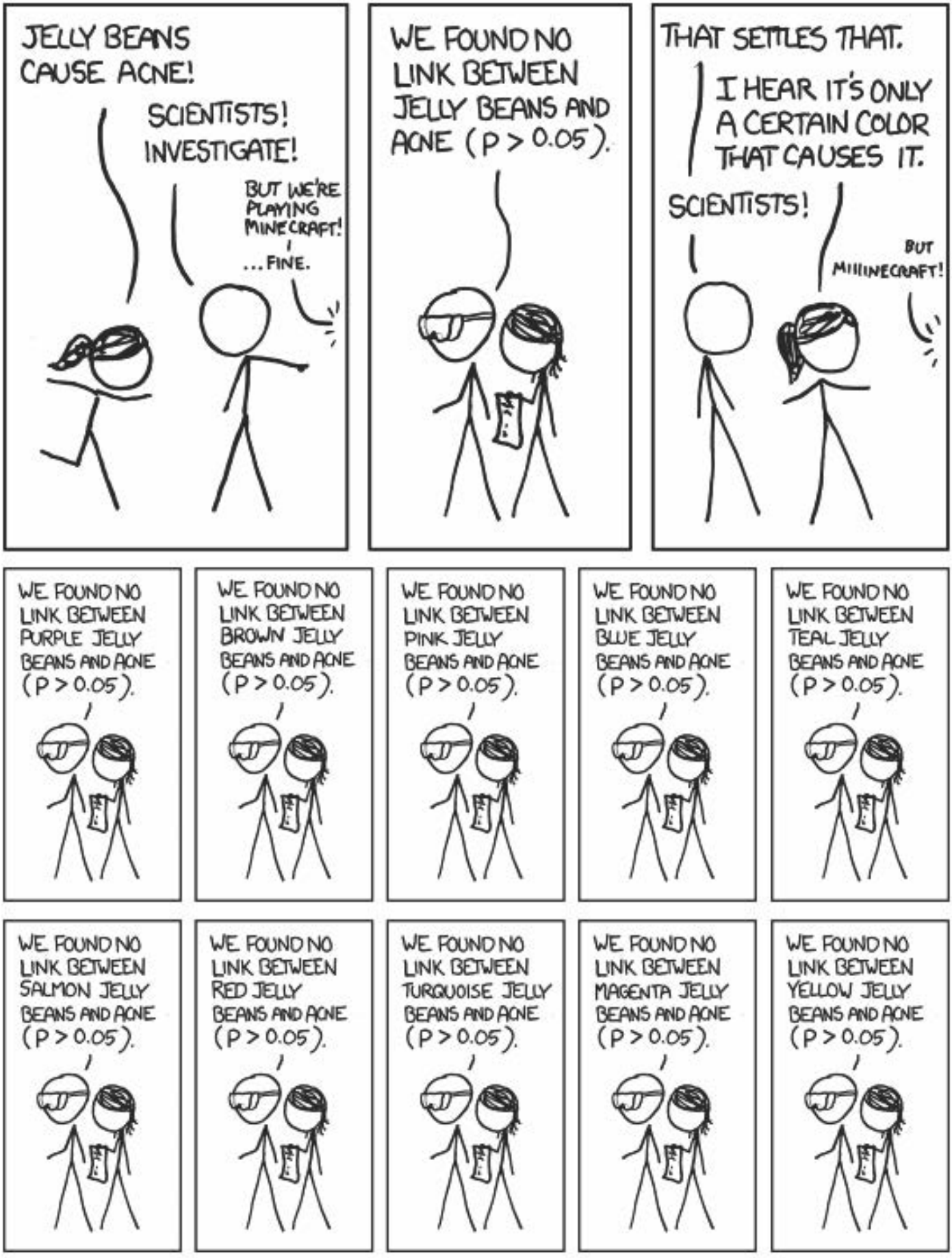
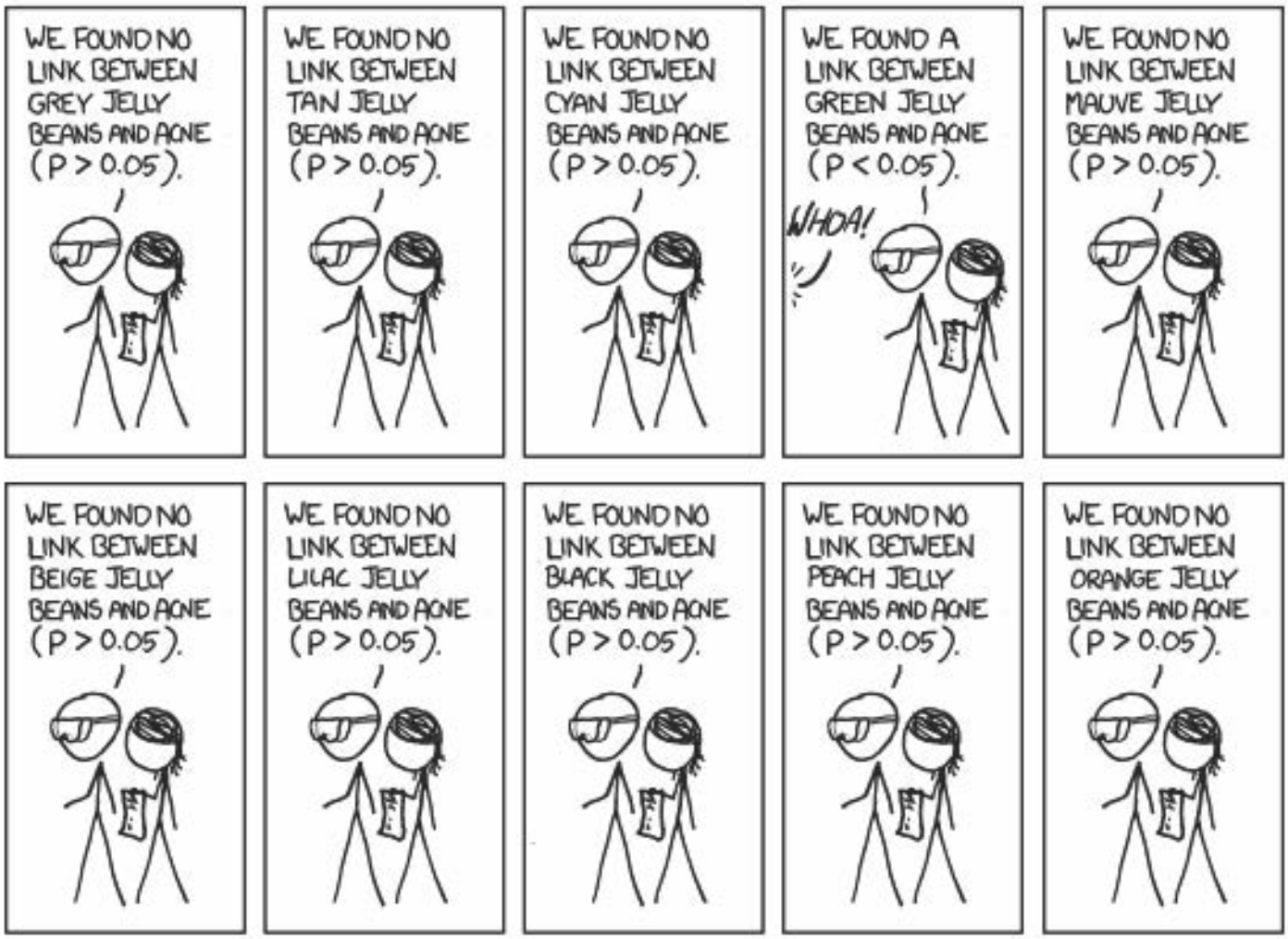
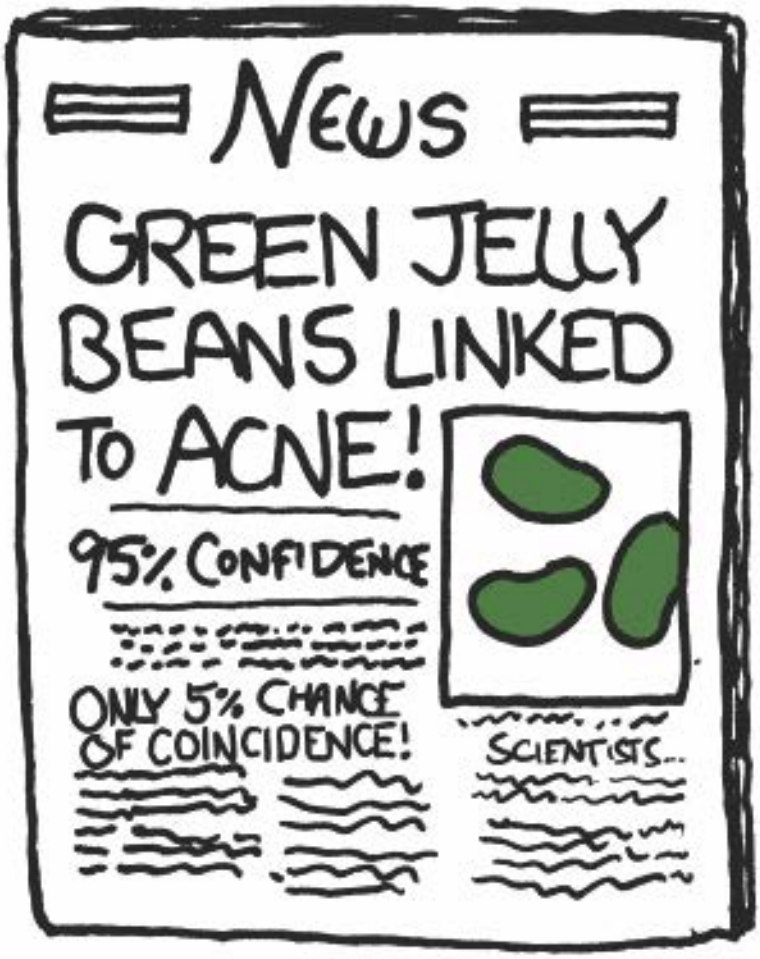
xkcd: Significant







NOTE: This document has been modified.

Explanation

This comic is about Data dredging (also known as *p*-hacking), and the misrepresentation of science and statistics in the media. A girl with a black ponytail comes to Cueball with her claim that jelly beans causes acne, and Cueball then commissions two scientists (a man with goggles and Megan) to do some research on the link between jelly beans and acne. They initially find no link, but they are egged on to keep trying until they find a link.

First some basic statistical theory. Let's imagine you are trying to find out if jelly beans cause acne. To do this you could find a group of people and randomly split them into two groups: one group who you get to eat lots of jelly beans and a second group who are banned from eating jelly beans. After some time you compare whether the group who eats jelly beans have more acne than those who don't eat jelly beans. If more people in the group who eat jelly beans have acne, then you might think that jelly beans cause acne. However, there is a problem.

Some people will have acne whether they eat jelly beans or not and some will never have acne even if they do eat jelly beans. There is an element of chance in how many people prone to acne are in each group. What if, purely by chance, all the group we selected to eat jelly beans would have had acne anyway while those who didn't eat jelly beans were the lucky sort of people who never get spots? Then, even if jelly beans did not cause acne, we would conclude that jelly beans did cause acne. Of course, it is very unlikely that all the acne prone people end up in one group by chance, especially if we have enough people in each group.

To give more confidence in the result of this type of experiment, scientists use statistics to see how probably their finding is. This is known as null hypothesis significance testing. Before they start the experiment, they choose a threshold known as the probability level. In the comic, the scientists choose a probability level of 5%. The probability level indicates the probability of obtaining test results at least as extreme as the results actually observed, assuming that the null hypothesis is correct.

The scientists (in the comic) initially find no link between jelly beans and acne, $p > .05$, meaning that there's more than a 5% probability of obtaining results as extreme, or more extreme, than their study's results. They cannot reject the null hypothesis.

But then Megan and Cueball ask the scientists to test whether another color of jelly beans is responsible. They again find a $p > .05$, meaning again that there's more than a 5% probability of obtaining results as extreme, or more extreme, than their study's results.

They continue testing more and more colors of jelly beans until they have tested 20 different colors, and only one time -- only one 1 out of 20 times (or 5% of the time) -- did they find a result that was considered significant at $p < .05$. By testing so many different colors without adjusting their *p*-value, they are likely to find a false positive. **If the probability that each trial gives a false positive result is 1 in 20, then by testing 20 different colors it is now likely that at least one jelly bean test will give a false positive.**

Moreover, by announcing only the results they found with green jelly beans, they have *p*-hacked their results.

Their *p*-hacked results lead to a big newspaper headline saying "Green Jelly Beans Cause Acne." We might later find out that the scientists tried to replicate the experiment (another key part of the scientific method), but now they no longer find any evidence for the link between acne and green jelly beans. But that's not news. The finding might not get reported in a journal, much less a newspaper. If it is reported in a newspaper, it might be reported as "Research is conflicted and more study is needed." But more study was just what the scientist already did.

This outcome is (sadly) often an issue with more serious matters than jelly beans and acne. At any time there are many news reports about possible links between substances (e.g., red wine) and illness (e.g., cancer). **By reporting only the positive results, the value of any single study is limited** - especially if the mechanism linking the two things is not known.

In 2015 some journalists demonstrated the same problem: just how gullible other news outlets are with the same sort of flawed "experimental design": How, and why, a journalist tricked news outlets into thinking chocolate makes you thin - The Washington Post.