Lecture 12

When it all goes wrong

Dr. Franco Ubaudi

The Nature of Data
Western Sydney University

Spring 2021

Power and under-power

Suppose you are testing a medication that is meant to shorten the length of colds:

- You find 20 patients with a cold
- Half get the new medication, the rest get a placebo
- You compare the average duration of colds for each group
- H_0: (the null hypothesis) is that the drug has **no** effect
- H_1: (the alternative hypothesis) is, regarding mean duration, $\mu_{drug} < \mu_{placebo}$
- Using H_1 you calculate a *p-value*

Power and under-power

Using the calculated *p-value*, say you observe a reduction in duration w.r.t the drug.

- The *p-value* is the chance of the observed reduction, if the drug has **no** effect, hence the drug is accurately described by H_0
- What if we only had 2 patients, instead of 20?
  What, if anything, would happen to your confidence in the drug?
- Understandably, our confidence in the finding is proportional to the number of patients
- As you reduce the number of patients, at what point do you consider the exercise futile?

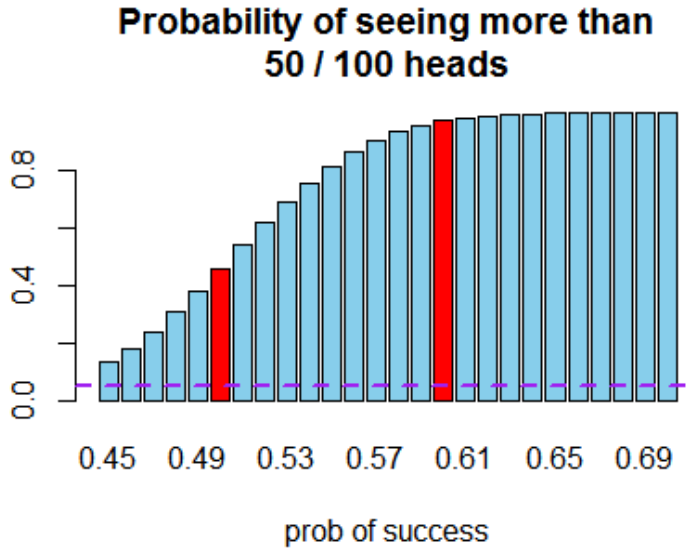
Is it fair?

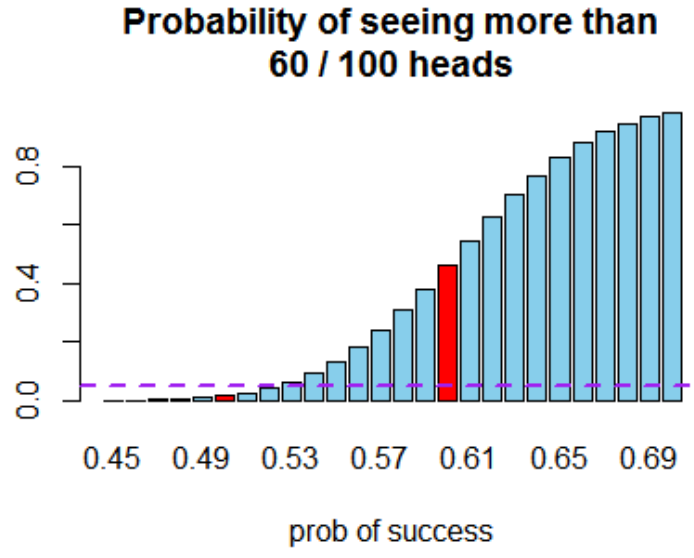
Imagine a friend has a coin and claims the coin is fair! The friend tells you the result of flipping that coin 100 times.

- You calculate the probability of the given result and suspect the coin is biased (heads 60%)
- **But** could a fair provide the observed result?
- Probably, but we need to decide some required probability threshold and if that threshold is not exceeded, we conclude that the safe **bet** is that the coin is not fair

Is it fair?

```
p <- seq(0.45, 0.7, 0.01)
d <- 1 - pbinom(heads, tosses, p)
```

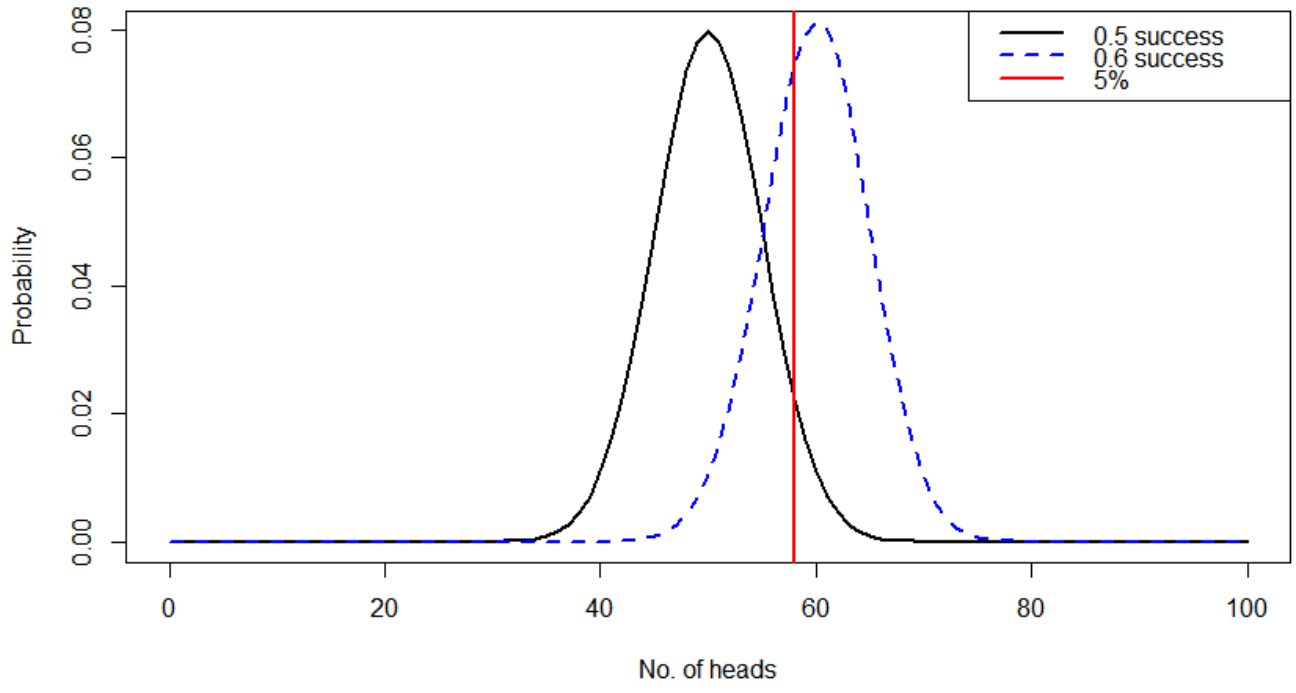




Is it fair?

Comparing a fair coin (black curve), with a 60% biased coin (blue dashed curve).

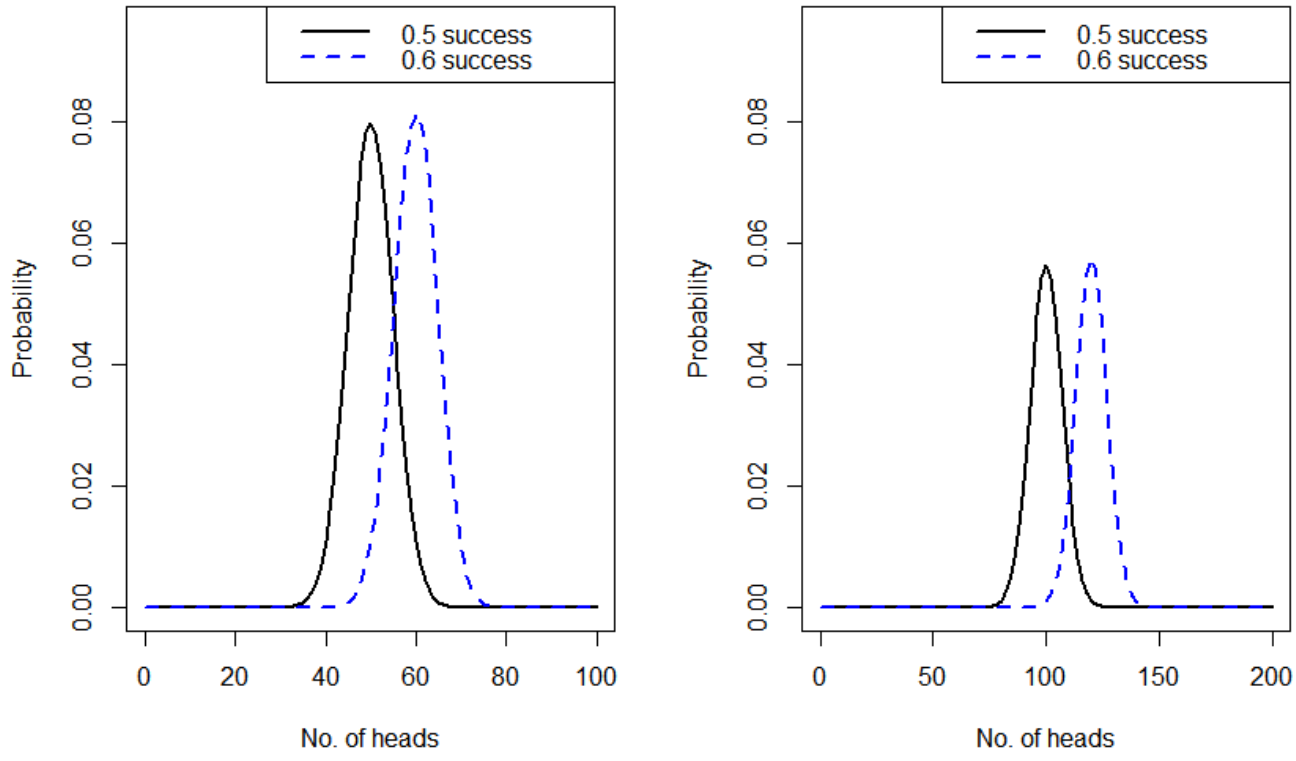
The red line shows the point where the probability of seeing more than 58 heads has fallen below 5% for a fair coin.



Is it fair?

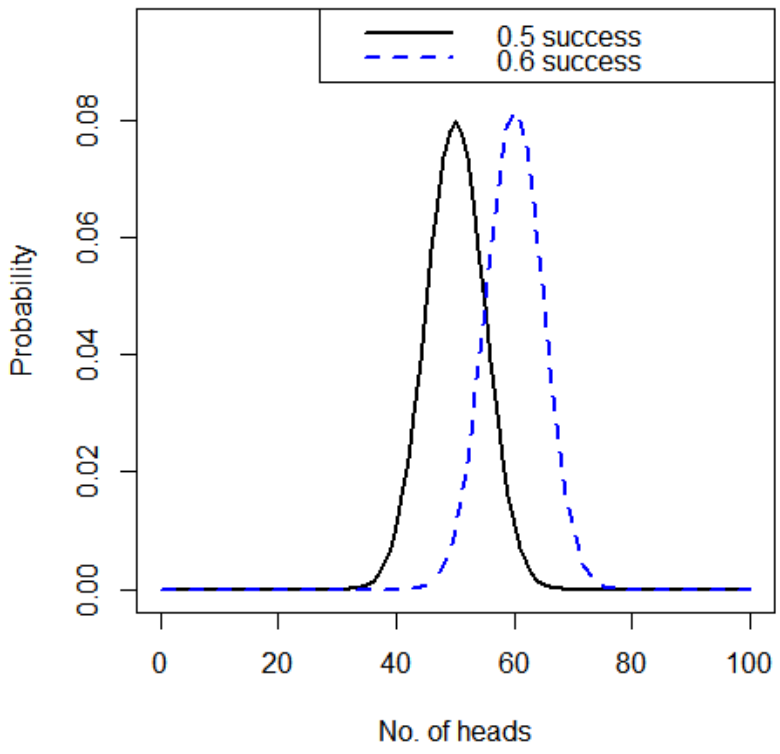
Comparing a fair coin (black curve), with a 60% biased coin (blue dashed curve).

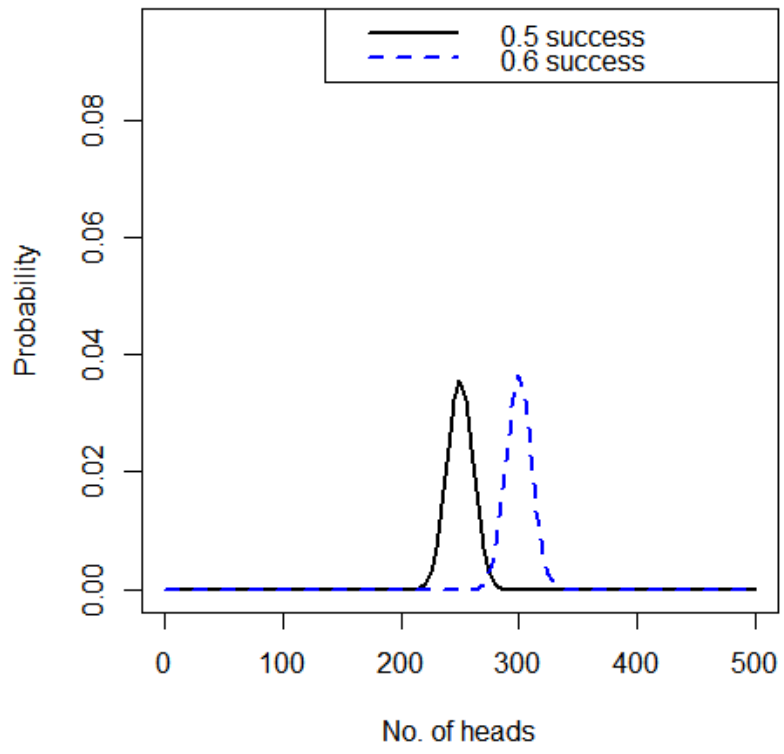
With 200 tosses, you would need to observe more than 112 heads, and the chance of doing so is 86%.



Is it fair?

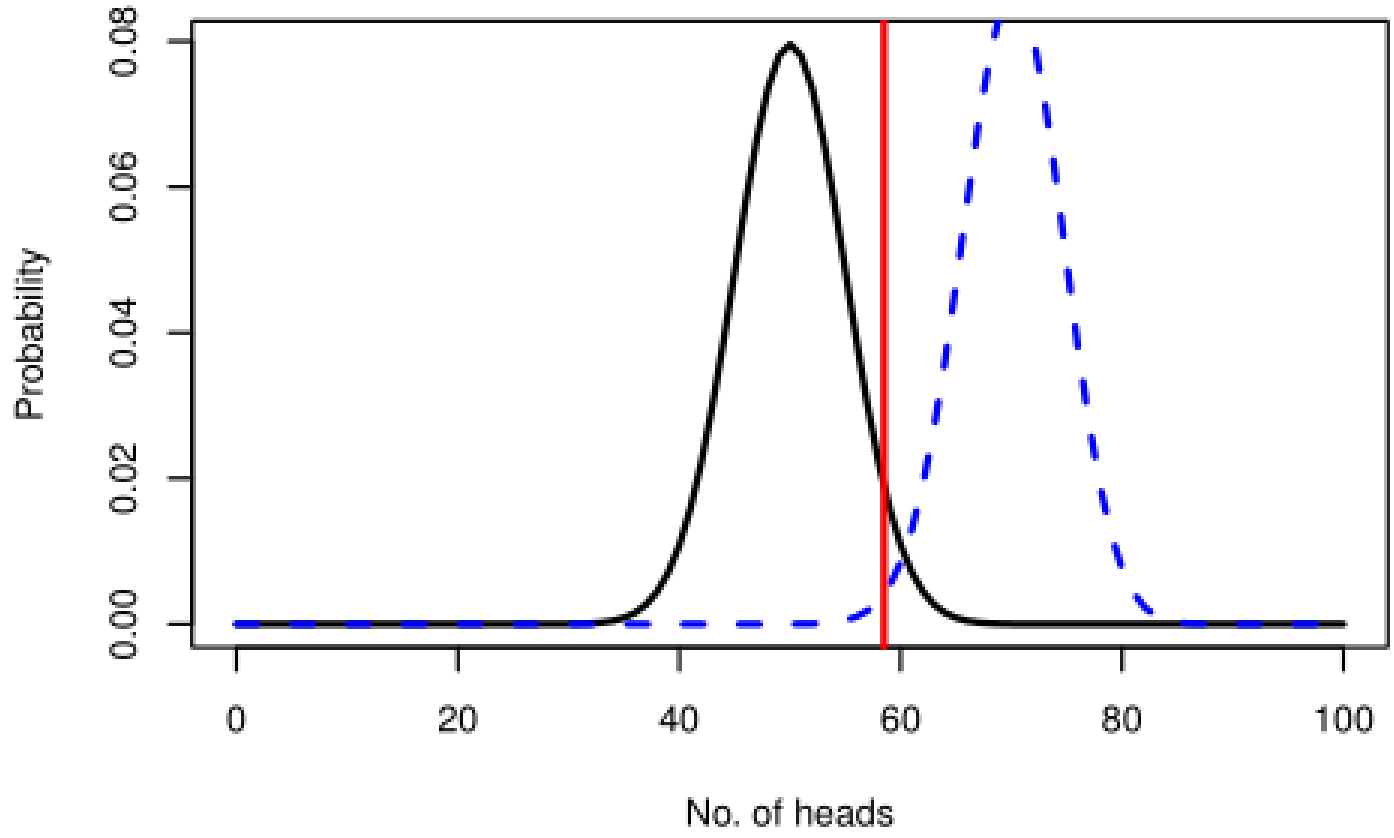
Comparing a fair coin (black curve), with a 60% biased coin (blue dashed curve).





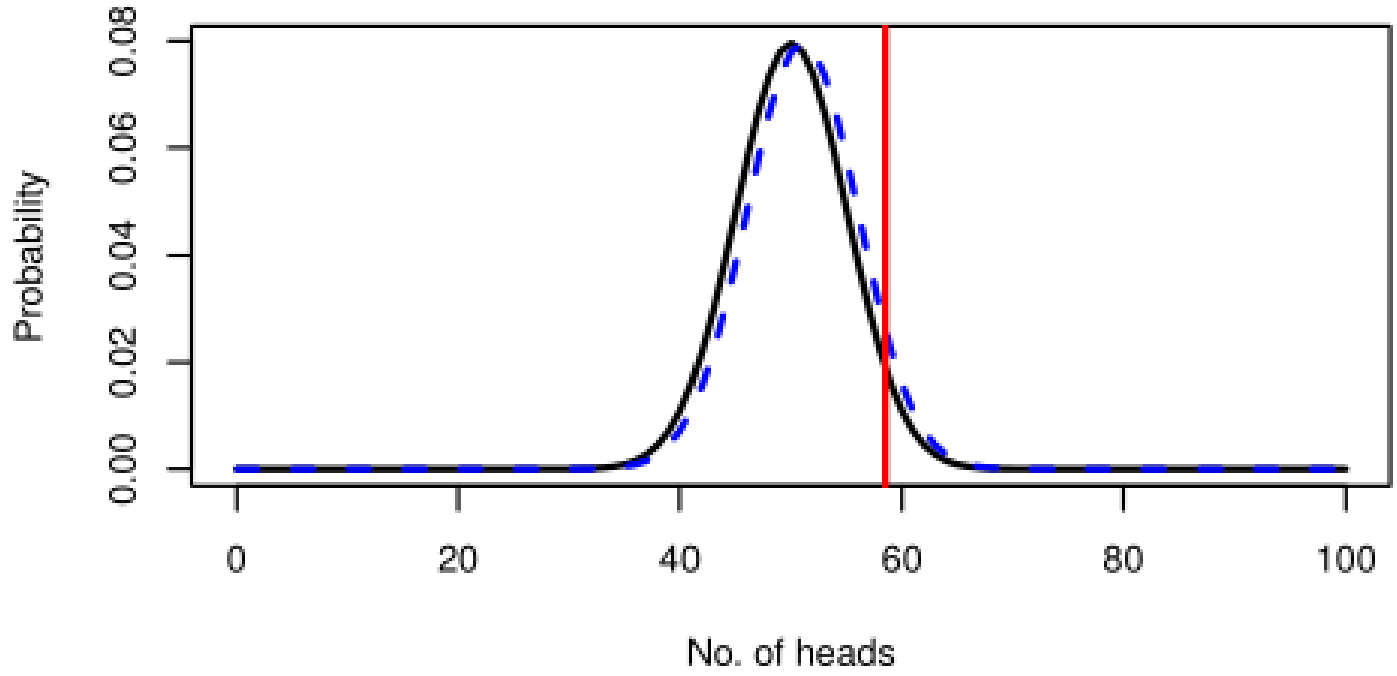
Is it fair?

Going back to 100 tosses, this time with a true (population) probability of heads 0.7



Is it fair?

Going back to 100 tosses, this time with a true (population) probability of heads 0.51



Under-powered tests

You might think it is rare for research to encounter this problem. But it occurs far more often than realised.

- JAMA. 1994 "Statistical power, sample size, and their reporting in randomized controlled trials." Moher D, Dulberg CS, Wells GA. Nearly two thirds of trials that reported no difference didn't have the power to detect a 50% difference.
- J Clin Oncol. 2007 "Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings." Bedard PL, Krzyzanowska MK, Pintilie M, Tannock IF. Only about half of the trials had enough power to detect even large differences.

Right turn on red light

In the 1970's several places in the USA introduced "right turn on red" provision.

A study of 20 intersections before and after showed that 308 incidents occurred before, and 337 afterwards (over a similar time period). No statistical significance was reported by this or by many other studies.

Finally, years later, a rather larger study showed that collisions occurred 20% more frequently, and pedestrians were being hit 60% more frequently. The previous studies lacked sufficient power to observe that.

Moral of the story: even researchers get it wrong / don't know correct statistical method. *So don't just accept what you read, have an appropriately critical perspective.*

Confidence intervals

Collapsing an entire data set to a yes / no call of statistical significance ($p < 0.05$) is wasteful.

Much more useful is to estimate the size of a difference using a confidence interval. However, this is rarely done ...

Pseudo-replication

Randomisation prevents researchers unwittingly introducing bias into their study.

Which is better, option I or II

I 2,000 patients randomised into two blood pressure medications
After the medication takes effect, the average blood pressure of the two groups is measured

II Have only ten patients per group, but measure each patient's blood pressure 100 times. Gives us 2,000 observations but! Could we say we have the same sample size?

Regarding option II $\implies$ No, since we only have 10 unique patients per group, so we just know an awful lot about each patient. This is a form of pseudo-replication.

Pseudo-replication

Another example of getting it wrong! A WSU researcher devised a study to compare two antibiotic treatments used on cows during the feed lot period just before slaughter.

Had data on 2,000 cows split across two treatments. Focus was illness / infections.

On the surface, there was significant differences.

But cows were put into the feed lots in batches. So factors like weather and location of the lot could have been confounding variables.

Pseudo-replication

Better approach for antibiotic treatments on cows.

2,000 cows in 8 batches and treatment allocated to batches.

Ideally, treatment would be randomly allocated to cows in each batch, then batches would be **blocks**. But wasn't practical.

Solution was to use *linear mixed models*[1]

Essentially involves estimating within and between batches, regarding variation and treatment differences.

This approach still showed a difference, but it was much weaker.

[1] Linear mixed models are an extension of simple linear models to allow both fixed and random effects.

Pseudo-replication

Unfortunately pseudo-replication is quite common in the published literature.

For example "Pseudo replication and the Design of Ecological Field Experiments" Stuart H. Hurlbert, Ecological Monographs 1984.

Paired t-test and pseudo-replication

The paired t-test actually allows for multiple measurements of the same thing.

On the surface it seems we have less data, **but**, by analysing only n differences, rather than $2n$ observations it looks like we have less data.

But the differences generally have much less random variation, because the individual effect is removed. Alternatively, we removed potentially confounding variables.

We demonstrated this with the New Zealand helmet example.

Base rate fallacy and multiple testing

Consider testing multiple drugs. Have 100 cancer drugs and 10 of them are really effective.

Using a threshold of 0.05 on the *p-value*, 13 drugs pass this test. But only 8 were really effective, so 5 provided *false positives*.

On average, a *p-value* of less than 0.05 will occur **by chance**, 1 in 20 times ($100/20 \approx 5$ so only 8 of the 13, or 62%, were truly effective)

Base rate fallacy and multiple testing

Some *incorrectly* state that *p-values* $< 5\%$ means less than 5% the drug was called effective by chance.

In fact we saw that 38% are of those called effective when they are not. This is called the base rate fallacy.

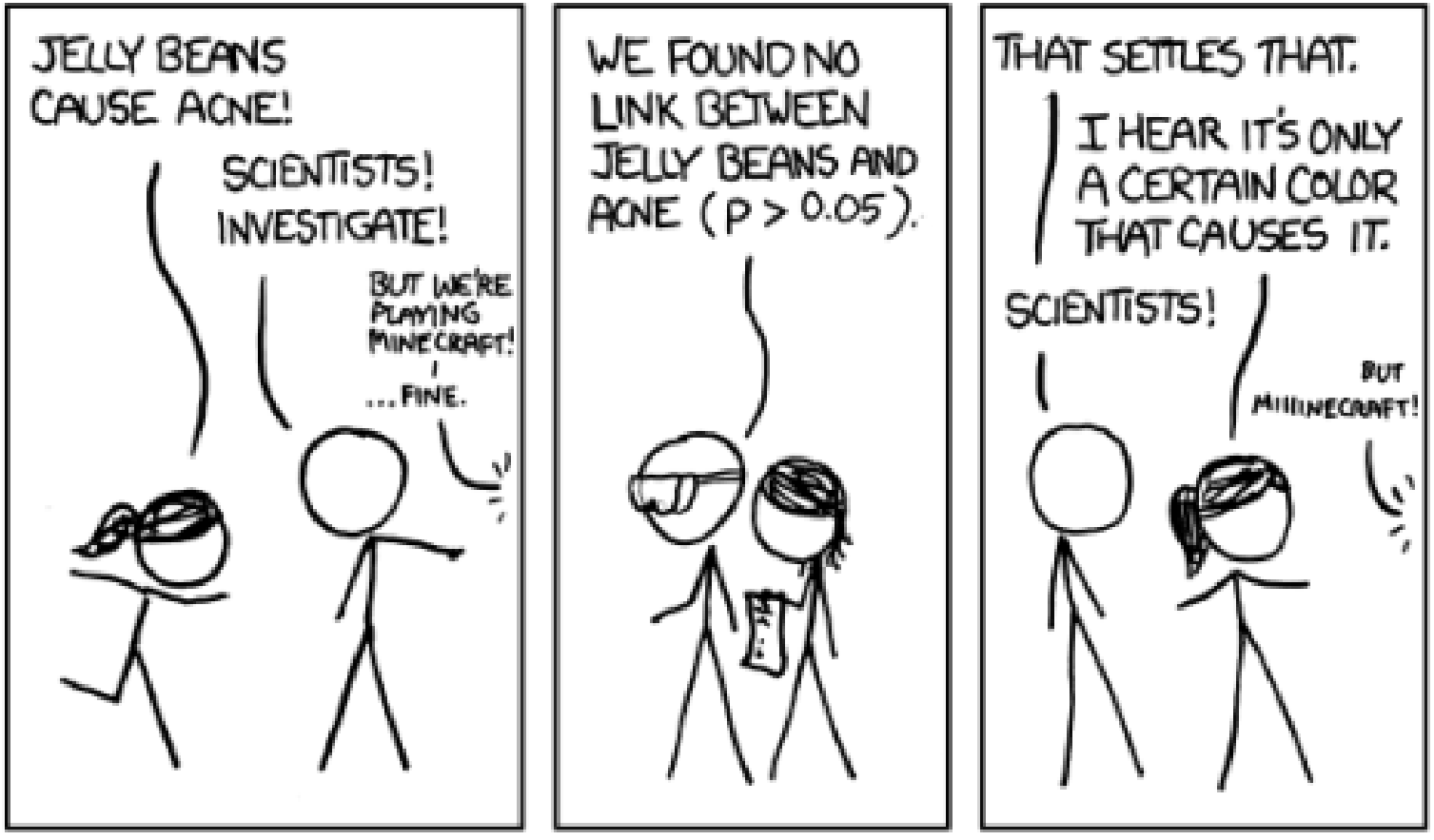
Problem

For a *p-value* of 0.01 which of the following is true?

1. You have absolutely disproved the null hypothesis.
2. There is a 1% probability the null hypothesis is true.
3. You have absolutely proved the alternative hypothesis.
4. You can deduce the probability that the alternative hypothesis is true.
5. You know, if you reject the null hypothesis, the probability that you are making the wrong decision. ✓
6. You have a reliable experimental finding, in the sense that if your experiment were repeated many times, you would obtain a significant result in 99% of trials.

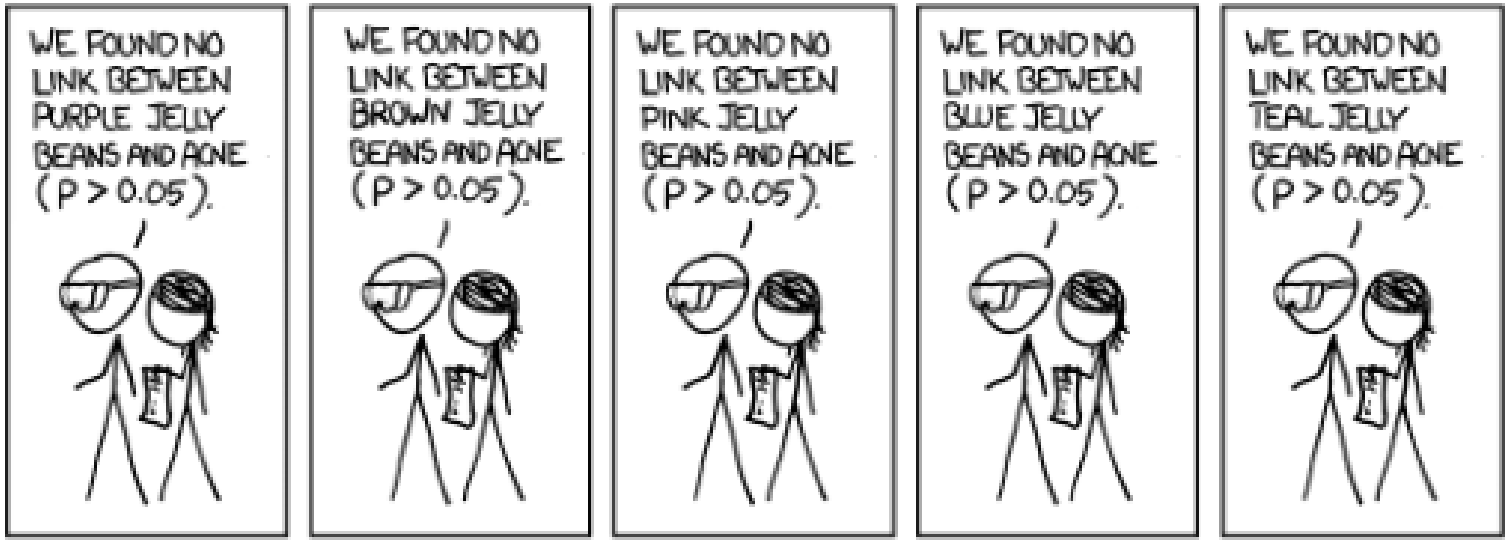
Multiple tests

Comics come from http://xkcd.com/882/



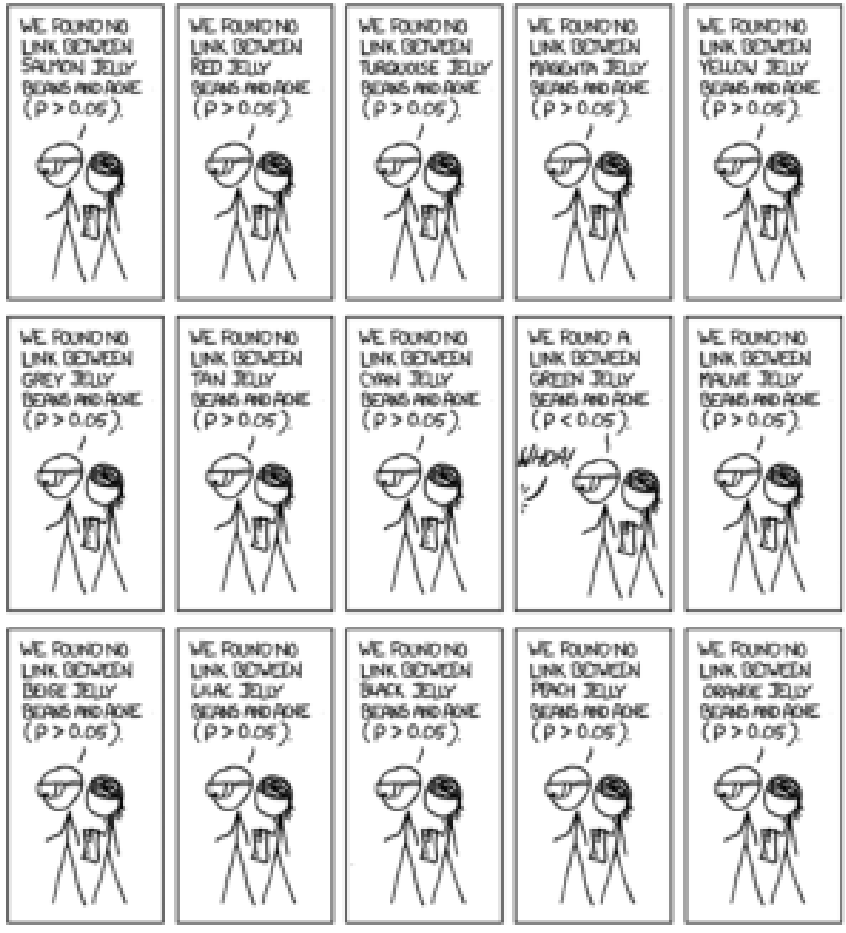
Multiple tests

What about?



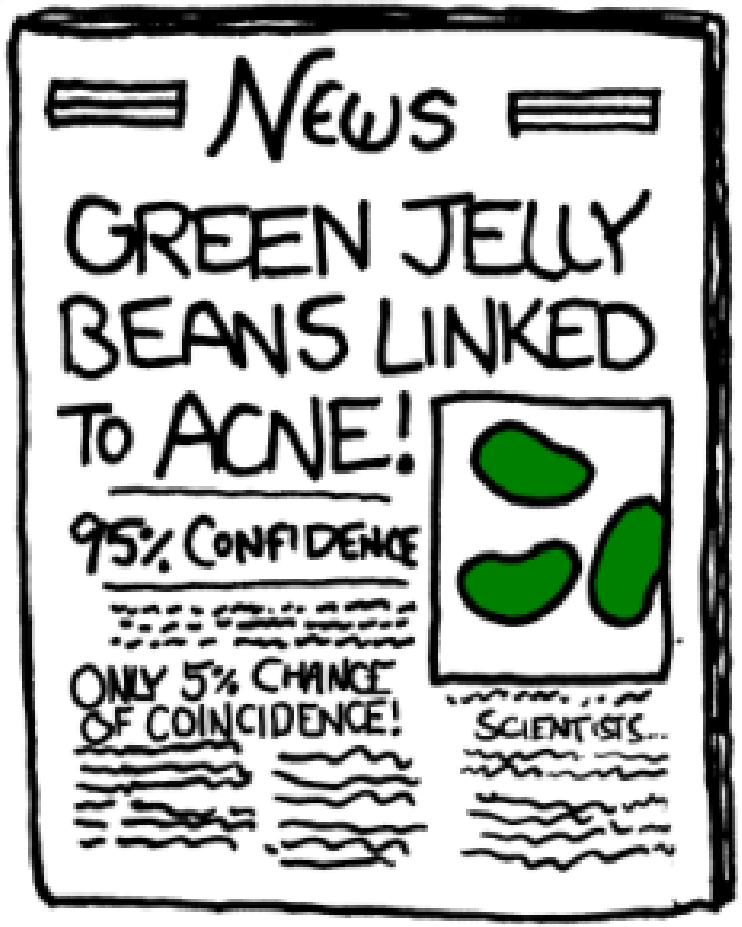
Multiple tests

What about?



Multiple tests

Eureka!!!



False discovery rate

- In genetics, its routine to test the activity of tens of thousands of genes in one go. Often this involves huge sample sizes.
- Suppose you are comparing the activity of 10,000 genes between cancer and normal tissue.
- Suppose 100 of these genes are really different in cancer and you have an 80% chance of detecting them (using a 5% *p-value* threshold).
- You detect 80 *true positive* genes, but also 5% of 9,900 genes as *false positives*. That is, you detect around 575 genes, of which only 80 (14%) are real.
- Fortunately, some of these 80 true positive genes will have a much smaller *p-value* than 5%. They will beat the threshold by a long way.
- Some of the *false positives* will have a very small *p-value* too, but this rate can be estimated. This leads to a number of methods to estimate the *False discovery rate*.

False discovery rate

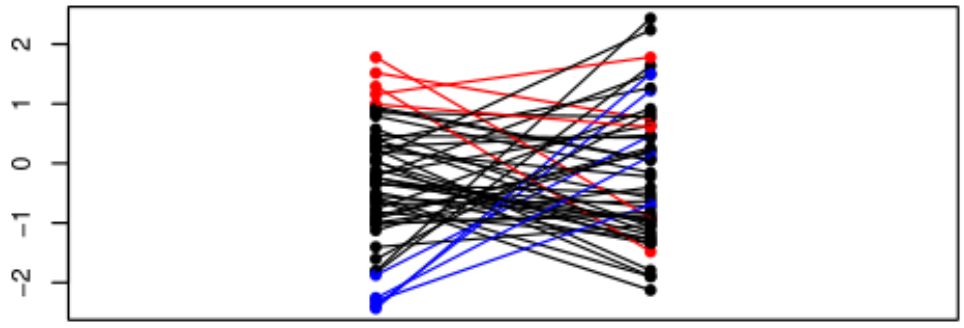
false discovery rate

For a given, *p-value* threshold, the *false discovery rate* is the fraction of tests with a *p-value* less than this threshold that are there by random chance.

Double dipping

Remember regression to the mean? Galton had noticed this effect as early as 1869. The results of many experiments are normally distributed

- Imagine trialling a new blood pressure medication
- Suppose a number of people are tested and **only** those with high blood pressure are selected
- You trial the new drug and re-test blood pressure
- You find many have reduced blood pressure. **But** some would reduce anyway due to regression to the mean.



The plot has 50 pairs of N(0,1) points

Simpson's paradox

In 1973 the University of California, Berkeley saw, of 12,763 applications for graduate study: 44% of males were accepted and 35% of females were accepted.

- Fearing a discrimination lawsuit the University investigated
- Of 101 departments, only 4 showed a statistically significant bias against women and 6 showed a bias against men
- It turns out that men and women did not apply in equal ratio to all the departments
- Two thirds of applicants in the English department were women
- Only 2% of applicants in Mechanical Engineering were women
- Some departments had higher success proportions than others
- Overall, this was sufficient to explain the perceived bias

Simpson's paradox

Consider two tennis players are to be compared across two seasons. Who is the best player?

	Wins	Games	Percent
Season 1	80	100	80%
Season 2	20	40	50%

Table: Player A statistics ✓

	Wins	Games	Percent
Season 1	78	100	78%
Season 2	2	5	40%

Table: Player B statistics

	Wins	Games	Percent
Player A	100	140	71.4%
Player B ✓	80	105	76.2%

Table: Putting everything together

Simpson's paradox

Consider two tennis players are to be compared across two seasons. Who is the best player?

	Wins	Games	Percent
Player A	100	140	71.4%
Player B ✓	80	105	76.2%

Table: Putting everything together

It is not a paradox that player A had better performance in both seasons, when considered individually, but player B had better performance overall.


WE MADE IT !!!!!