Hypothesis Testing

Owen G Ward

How do Scientist's validate theories

Scientists in almost all fields collect data which they use to validate theories. For example, chemists who develop a drug want to be able to determine if the drug is effective at treating a specific illness. Similarly, maybe engineers develop a new type of microchip which they hope can withstand higher temperatures. How do they confirm this?

To do this, generally you need to be able to compare two (or more) datasets and determine:

- If there is a difference between the two groups.
- If there is a difference, what is it? For a drug, does the difference indicate a larger proportion of people are cured, or possibly less?
- Is this difference *significant*? That is, could the difference be explained by some randomness between the two groups?

To understand these problems, we need to know the language of statistical significance.

Setting Up

Suppose we have a new drug which aims to lower blood pressure, and we wish to see if this drug is actually effective. To determine this, we obtain a *control* group, who do not receive the drug, and a *treatment* group, who are given this new drug. The question we want to answer is

> *Do the people who receive this drug have lower blood pressure on average than those who don't?*

To do this, we need to come up with a hypothesis. In statistics, this is known as the *null hypothesis*. The null hypothesis is always the default or the norm being true, i.e that there is no difference in average blood pressure between the two groups.

We also have an *alternative hypothesis*, which is the hypothesis about the data we are interested in. For the drug example, the alternative hypothesis would be that the average blood is lower in the group who receive the drug.

Informally, the idea of a hypothesis test is that we assume the null hypothesis is true and then, given that assumption, we investigate how *likely* the data we observed would occur under the null hypothesis. If there is little evidence, we will reject the null hypothesis. We will illustrate this with a famous example.

To do this, one common technique is to perform a statistical test and obtain a *p-value*. A small p-value is seen as evidence that the null hypothesis is not true. We will see a formal definition of the p-value later!

A historical example

In England in the 1700's, John Arbuthnot decided to examine whether male births were more likely than female births.

His null hypothesis, therefore, was that the probability more boys are born each year is equal to the probability more girls are born in a year. The alternative hypothesis is therefore that the probability more boys are born is greater than the probability more girls are born.

He obtained 82 years of data summarizing christenings in London. In each year, more boys were christened than girls.

Arbuthnot reasoned that if the birth rates were equal than the probability of more boys being born in a single year would be equivalent to flipping a fair coin and getting heads.

Or equivalently, the probability of having more boys born each year for 82 years would have the same probability of flipping a fair coin 82 times and getting heads each time. The probability of this happening is essentially zero, and in this scenario corresponds to a p-value.

```
birth_data %>%
  mutate(Heads = ifelse(Males - Females > 0, 1, 0)) %>%
  head() %>%
  kable()
```

Year	Males	Females	Heads
1629	5218	4683	1
1630	4858	4457	1
1631	4422	4102	1
1632	4994	4590	1
1633	5158	4839	1
1634	5035	4820	1

To see this, we can simulate flipping a fair coin 82 times and see how many heads we get. This is equivalent to saying "If the null hypothesis is true and the number of boys and girls born each year is equal, how many years would we expect there to be more boys being born?"

```
n_flips <- 82
rbinom(n = 1, size = n_flips, prob = 0.5)
```

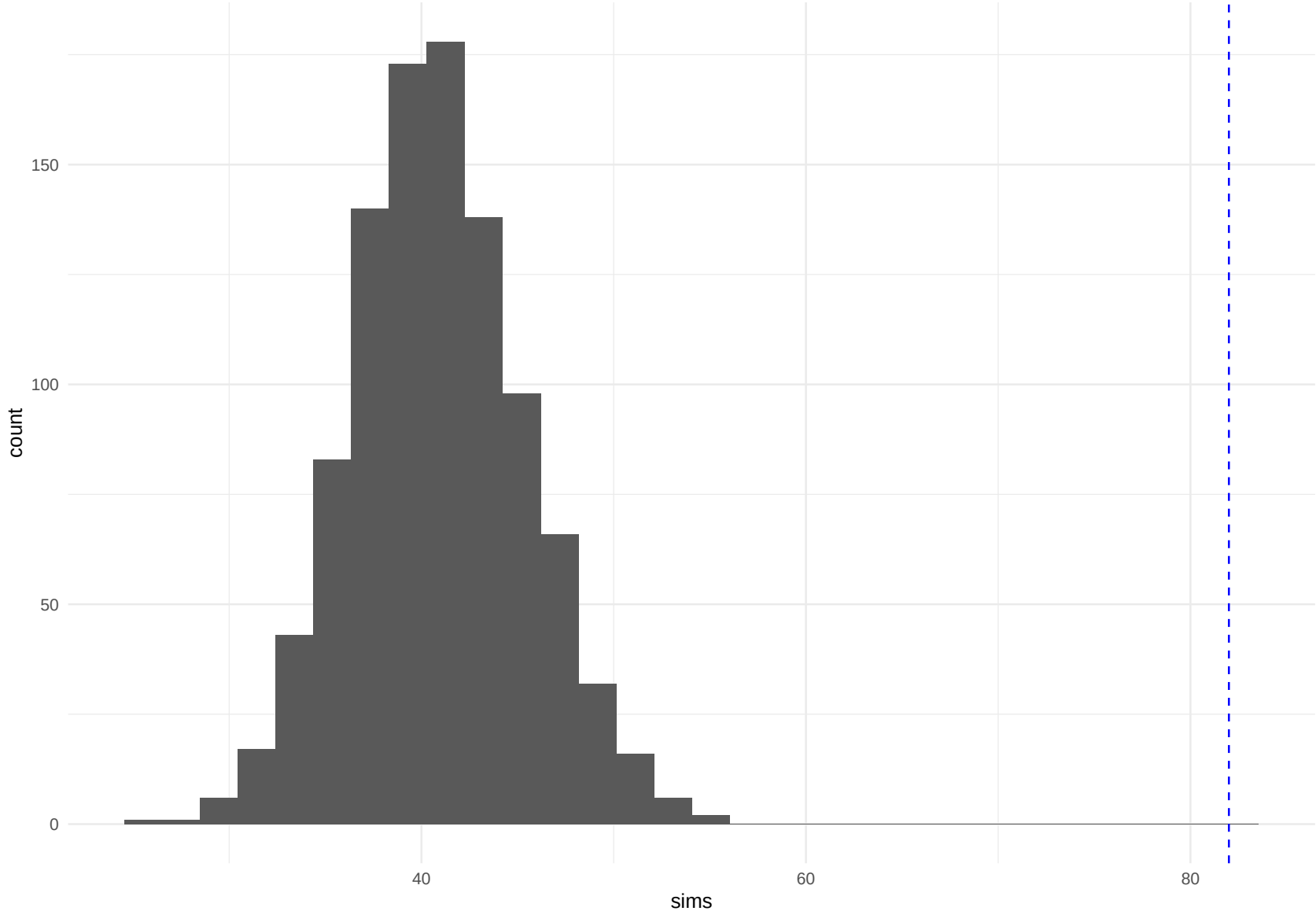
```
## [1] 39
```

What about if repeat this experiment many times and look at the distribution of the number of heads. This is the distribution of heads under the null hypothesis (that both are equally likely).

```
n_sims <- 1000
coin_sims <- tibble(sims = rbinom(n = n_sims,
                                  size = n_flips,
                                  prob = 0.5))

coin_sims %>% ggplot(aes(sims)) + geom_histogram()


# how does this compare to the data we have?
coin_sims %>% ggplot(aes(sims)) +
  geom_histogram() + geom_vline(xintercept = 82,
                                color = "blue",
                                linetype = 2)
```


150
100
50
0
count
40
60
80
sims

So the probability of getting a result, under our null hypothesis, as or more extreme than the one we observed, is essentially zero.

Arbuthnot thought this difference might be due to a "wise creator" who was accounting for the risk men faced hunting. However, his analysis does not support this.

All it shows is that the number of male christenings is more than the number of female christenings. It is possible (or even likely) that there were other reasons (financial, cultural) which meant families were less likely to christen female children.

Difficulties with statistical significance

A *p-value* is a probabilistic quantity. The formal definition is given below.

The probability of obtaining a test result as or more extreme than the one observed, if the null hypothesis was true.

As such, a p-value is far from a perfect metric for determining differences. There are also many ways it can be misused.