Hypothesis Testing

Owen G. Ward

2021-06-08

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 datasets and determine:

- If there is a difference between the the estimate from the data and what you might expect.
- 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 in the data?

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

Illustrative Example

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 pressure 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 unlikely to be 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 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.

Year	Males	Females
1629	5218	4683
1630	4858	4457
1631	4422	4102
1632	4994	4590
1633	5158	4839
1708	8239	7623
1709	7840	7380
1710	7640	7288

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.

Year	Males	Females	Heads
1629	5218	4683	1
1630	4858	4457	1
1631	4422	4102	1
1632	4994	4590	1

Year	Males	Females	Heads
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)
```

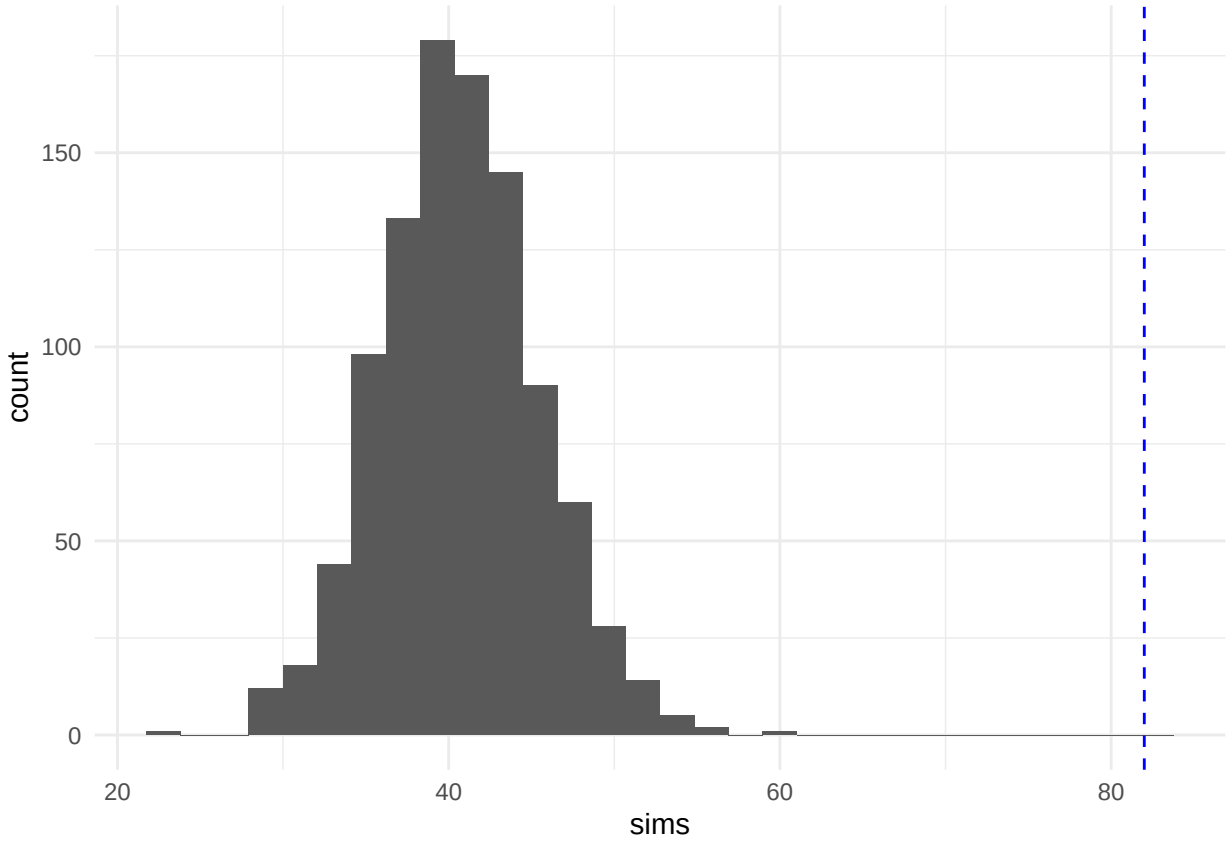
```
## [1] 40
```

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).

We are repeating sampling data from our null hypothesis, which is a distribution, and comparing the samples from this distribution to the actual data.

Assuming that each year there is a 50/50 chance more boys are born than girls. Count how many of the 82 years more boys are born than girls. Repeat this many times.

Samples from the Null



Informally, the observed data is very different to the samples from the null hypothesis.

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.

This ties into the point about data quality. The inference you can make is dependent on the data you have.

The formal framework

To do a hypothesis test we want to come up with a rule to decide between the null hypothesis and the alternative hypothesis. This rule will be based on the data we observe.

We want to be reasonably certain to reject the null hypothesis.

Don't really "accept" the alternative hypothesis. Either fail to reject the null hypothesis if there is insufficient evidence to do so, or reject the null if there is enough evidence.

Example

Suppose someone claims that 50% of New Yorkers are vaccinated.

We survey 50 people and find out if they are vaccinated or not. We want to use this data to test whether the claim is true or not.

The probability that a random New Yorker is vaccinated is some p and we want to test $p = 0.5$ vs $p > 0.5$. Let Y be the number in our sample who are vaccinated.

We have $Y \sim Binom(50, p)$ and we want to test our null hypothesis,

$$H_0 : p = 0.5$$

against the alternative

$$H_A : p > 0.5.$$

Suppose our test rejects the null hypothesis is $Y \geq 30$. What problems could there be?

We could have $Y \geq 30$ (reject H_0) even if $p = 0.5$.

Similarly, we could have $Y < 30$ (fail to reject H_0) even if $p > 0.5$.

We want our test to be able to identify $p > 0.5$ if it is, but at the same time, we don't want it to say $p > 0.5$ when it isn't true. These are opposite goals and as such there is a trade off between them.

Errors in Testing

We could make two types of errors in the previous example.

Type 1 Error where we reject H_0 when it is actually true.

Type 2 Error where we don't reject H_0 when we should.

This is similar to the disease testing example. We want the probability of both of these errors to be small.

Computing this Error

The Type 1 error is the probability $Y > 30$ given that $p = 0.5$ which we can compute using the Normal Approximation.

Similarly, for the Type 2 error, we assume $p > 0.5$ and compute the probability that $Y < 30$.

Error of a Test

In an ideal world our test would minimize both these types of error. But decreasing one leads to an increase in the other.

We construct tests which have Type 1 error α, can compute what the Type 2 error is.

Constructing a Test

To perform a test, we do a probabilistic version of proof by contradiction. We first assume the null hypothesis is true.

If the null hypothesis was true, we know the distribution of some test statistic, say the sample mean.

Then we evaluate if the observed value of the sample mean is plausible under the null distribution.

If it is implausible, we reject the null hypothesis.

Example

Suppose we have $X_1, \dots, X_n \sim Binom(1, p)$ and we want to test $H_0 : p = 0.5$ vs $H_A : p \neq 0.5$..

If the null hypothesis was true then $\frac{\bar{X}-p}{se_{p=0.5}(\bar{X})}$ is approximately Normally distributed with mean 0 and variance 1. Here is the standard error is known under the null hypothesis.

$$\frac{\bar{X} - 0.5}{\sqrt{\frac{0.5(0.5)}{n}}} \sim \mathcal{N}(0, 1)$$

We call this our **test statistic**, which is given by our point estimate minus our null value, and divided by the standard error under the null.

If the value of $\frac{\bar{X}-p}{se_{p=0.5}(\bar{X})}$ is unlikely to have come from a Normal distribution then that is evidence that the null hypothesis is incorrect, and we should maybe reject it.

How unlikely do we care about

We need to quantify how unlikely are observed value actually needs to be to reject the null hypothesis.

To do this, we go back to using a z table!

When we formally construct a test, we specify we are doing the test at some significance level α, which corresponds to the Type 1 error of the test.

This means we reject our null hypothesis if the value we observe occurs in the tail region of the distribution which has probability α.

Z Table for Testing

Example

Suppose 1000 people were asked if they support coal burning and 370 said they did. Test that the true proportion of people who support coal burning differs from 0.5 at significance level $\alpha = 0.05$.

Rejection regions

We reject our null hypothesis if we are in the tail area of our null distribution.

For significance level α and testing $H_0 : p = 0.5$ vs $H_A : p \neq 0.5$ this corresponds to

$$\frac{\bar{X} - 0.5}{\sqrt{\frac{0.5(0.5)}{n}}} < -z_{\alpha/2} \quad \textbf{or} \quad \frac{\bar{X} - 0.5}{\sqrt{\frac{0.5(0.5)}{n}}} > z_{\alpha/2}$$

So we can summarise our test by saying we reject the null hypothesis if

$$\left| \frac{\bar{X} - 0.5}{\sqrt{\frac{0.5(0.5)}{n}}} \right| > z_{\alpha/2}.$$

We can write the previous expression for any $H_0 : p = p_0$ against $H_A : p \neq p_0$ at significance level α. Under the null hypothesis we reject if

$$\left| \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right| > z_{\alpha/2}.$$

If we multiply by the positive denominator then our test is

- Reject H_0 if $|\bar{X} - p_0| \geq z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$
- Fail to reject H_0 if $|\bar{X} - p_0| < z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$

We will see that most tests look like this, which is very similar to a confidence interval.

Type 2 Error

Suppose 1000 people were asked if they support coal burning and 370 said they did. Test that the true proportion of people who support coal burning differs from 0.5 at significance level $\alpha = 0.05$.

Suppose the true probability is $p = 0.4$. What is the Type 2 error then?

p-Value of a Test

We have said that if the value is in the tail region of the z-table we reject the null hypothesis.

Equivalently we can compute a **p-value** which corresponds to the area more unlikely than what we observed.

Getting a p-value less than the significance level of your test also means you reject the null hypothesis. However it is important that you specify α before you do the test!

p-Value

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.

The p-value is the region of values under the null hypothesis which are at least as unlikely as the one we observed. This means the region even further in the tails.

As or more extreme means in terms of distance from the center of the distribution, so both very small and very large values.

Example

Suppose 1000 people were asked if they support coal burning and 370 said they did. Test that the true proportion of people who support coal burning differs from 0.5 at significance level $\alpha = 0.05$ by computing a p-value.

Confidence Intervals

There is also a direct relationship with confidence intervals.

A 95% confidence interval for p which doesn't contain 0.5 is equivalent to a hypothesis test rejecting the null hypothesis that $p = 0.5$ at significance level $\alpha = 0.05$.

This is true once both use the same α.

Examples of Statistical Testing

Normal Mean with known Variance

Suppose we have data $X_1 \dots, X_n \sim \mathcal{N}(\mu, \sigma)$ where we know σ. We want to test if $H_0 : \mu = \mu_0$ vs the alternative $H_A : \mu \neq \mu_0$ at significance level α.

Again, we have that

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1).$$

We want to reject if this value is in the region which has probability $\leq \alpha$.

So we

- reject the null if $|\bar{X} - \mu_0| \geq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
- fail to reject if $|\bar{X} - \mu_0| < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Example

A manufacturer of sprinkler systems claims that the true average system activation temperature is 130F. Suppose we test the system 9 times, with average temperature being 131.08F.

The activation time is known to be normal with standard deviation 1.5.

We want to test $H_0 : \mu = 130$ vs $H_A : \mu \neq 130$ at significance level $\alpha = 0.01$.

Our z score is $z_{0.005} = 2.33$ so we reject the null if

$$|\bar{X} - 130| \geq 2.33\frac{1.5}{3} = 1.16.$$

In this example $|\bar{X} - 130| = 1.08 < 1.16$ so we fail to reject the null at significance level $\alpha = 0.01$.

Note that if we had originally done the test at $\alpha = 0.05$ we would have rejected if

$$|\bar{X} - 130| \geq 1.96\frac{1.5}{3} = 0.98,$$

so would have came to a different conclusion.

As α gets smaller it becomes harder to reject the null, only allow very small Type 1 error.

Computing the p-Value

We can also compute the p-value for this problem, which would have given us both these results at once (but need to pick α first).

Summary

For the tests we have seen so far we have a standard procedure.

We know the standard error under our null hypothesis and what the distribution of the point estimate minus the value of the null divided by the standard error is.

If this value is unlikely to be from the null distribution we reject it. By unlikely we mean has probability $\leq \alpha$ of occuring under the null distribution.

This extends to slightly more complicated examples too.

Difference between Means

Suppose we have $X_1 \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2)$.

We know σ_1, σ_2, but we want to see if these two distributions have the same mean.

To test $H_0 : \mu_1 = \mu_2$ vs $H_A : \mu_1 \neq \mu_2$ at significance level α we do the same as before.

We want to equivalently test $\mu_1 - \mu_2 = 0$ vs $\mu_1 - \mu_2 \neq 0$. To do that, we need an estimate for $\mu_1 - \mu_2$ and a standard error for that estimate.

To get an estimate for $\mu_1 - \mu_2$ we can just get an estimate for each.

$\bar{X}$ is an estimate for μ_1 and $\bar{Y}$ is an estimate for μ_2 so $\bar{X} - \bar{Y}$ is an estimate for $\mu_1 - \mu_2$.

We need to get the standard error of $\bar{X} - \bar{Y}$.

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

So the standard error is

$$se(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}.$$

Under our null that $\mu_1 = \mu_2$ then

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{se(\bar{X} - \bar{Y})} \sim \mathcal{N}(0, 1).$$

So we reject the null at significance level α if

$$|\bar{X} - \bar{Y}| > z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}.$$

One Sided Tests

Much like with confidence intervals, we also can often consider one sided hypothesis tests.

Suppose we want to test $H_0 : p = p_0$ vs $H_A : p > p_0$ at level α.

If we have a sample our test will still be based on $\bar{X}$. We still want to reject for unlikely values, but only those which give us evidence about our hypotheses. Our null distribution is still the same

$$\frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \mathcal{N}(0, 1).$$

If our alternative is that $p > p_0$ then only large values cause us to reject the null.

We still want this test to have Type 1 error α, but unlikely values now under the null are only those which are large. Have only one rejection region, want it to have probability α under the null.

This means instead of the two sided test we will

- Reject H_0 if $\bar{X} \geq p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$.
- Fail to reject H_0 otherwise.

Example

Suppose 1000 people were asked if they support ending fracking and 570 said they did. Test $H_0 : p = 0.5$ vs $H_A : p > 0.5$ at $\alpha = 0.05$.

For this $\alpha = 0.05$ so $z_\alpha = 1.64$.

Now we reject our null if $0.57 > 0.5 + 1.64\sqrt{\frac{0.5(0.5)}{1000}} = 0.5259$. So reject the null.

One sided tests have some disadvantages and can lead to larger errors. Will generally focus on two sided tests.

T Tests (Not Directly Examinable but important)

When we have performed hypothesis tests and confidence intervals for normal means we have always assumed we know the standard deviation σ.

In reality, this is never actually the case! As such we need to estimate σ also.

The natural way to modify the test is to use the estimate for σ, which we call s, given by

$$s = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}.$$

If we had

$$\frac{\sqrt{n}(\bar{X} - \mu)}{s}$$

following a normal distribution we would be done. However, it does follow a **t-distrbution**, and we can modify the test to account for that.

A t-distribution is very similar to a normal distribution, but with more uncertainty in the tails (to account for not knowing σ).

A t-distribution has an additional parameter known as the degrees of freedom. For testing a normal mean where we have n data points this is $n - 1$.

T-Test for Normal Mean

When we knew σ our test rejected if

$$|\bar{X} - \mu| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

When we don't know σ this becomes rejecting if

$$|\bar{X} - \mu| > t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}.$$

We have

$$t_{n-1,\alpha/2} > z_{\alpha/2},$$

so harder to reject for a t-test. This makes sense, because we have more uncertainty (due to s).

Confidence Intervals

We can also construct confidence intervals when σ is unknown,

$$\left(\bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right).$$

As n gets large then $t_{n-1,\alpha/2}$ and $z_{\alpha/2}$ become closer and closer.

We can modify other normal tests in this way also. Don't need to do it for Binomial or Poisson as only parameter.