Quantitative Understanding in Biology
1.4 p-Values and Formal Statistical Tests

Jason Banfelder

September 27th, 2022

1 Introduction to p-values

We have already seen one example of formal statistical testing when we tested for the normality of a sample from a univariate distribution. To reiterate, formal statistical testing begins with a statement of the null hypothesis, H_0. The odd thing about the null hypothesis is that you will try to show that it is not plausible. By doing so, you will show that its complement, the alternative hypothesis, H_1, is likely to be true.

Once you have stated the null hypothesis, you compute a probability, called the p-value. Informally, the p-value is the probability of "getting the data that you actually got" under the assumption that the null hypothesis is true. The tricky part here is defining the colloquial "getting what you actually got" appropriately. More formally, it is the probability of getting a result as, or more, inconsistent with the null hypothesis than the data that you actually observed.

If the p-value from this computation is small (below some pre-determined cutoff value usually written as α), then you can conclude that the null hypothesis is unlikely to be true, and you reject it. You have a statistically significant result.

If the p-value is not below α, your test is inconclusive. You cannot conclude anything about the null or alternative hypotheses. Be sure to understand this point, and do not misinterpret the p-value. The p-value is **not** the probability of the null hypothesis being true.

2 The Binomial test revisited

Let's consider the contrived example of testing a coin to see if it is fair; in other words, to see if $P(H) = 0.5$. Don't confuse the p-value with P(H). We begin by stating our null

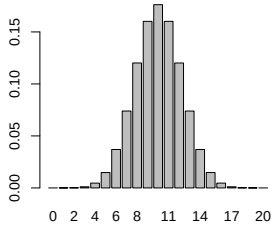
hypothesis...

$$H_0\text{: The coin being tested is fair; i.e., } P(H) = 0.5$$

...and deciding, before we collect any data, that we will work with 95% confidence ($\alpha = 0.05$). Next, we flip the coin 20 times, and observe 13 heads.

We now want to compute the probability of getting a result as or more inconsistent with the null hypothesis than the one we observed, assuming that the null hypothesis is true. To think about what it means to get a result "as or more inconsistent with the null hypothesis", let's plot the distribution under the null hypothesis:

```
barplot(dbinom(0:20, size = 20, prob = 0.5), names.arg = 0:20)
# The null hypothesis is "prob = 0.5"
```



Looking at the distribution we can reason that getting a result as or more inconsistent with the null hypothesis is the sum of the probabilities of observing 0, 1, 2, 3, 4, 5, 6, 7, 13, 14, 15, 16, 17, 18, 19, or 20 heads, using the assumption that $P(H) = 0.5$. You can now see why we formulate a null hypothesis; it gives us additional information that we need to complete a mathematical model and calculate a probability.

We already know several ways to compute this probability using R. It is equal to $1 - (P(8) + P(9) + P(10) + P(11) + P(12))$, which is...

```
1 - (pbinom(12, 20, p = 0.5 ) - pbinom(7, 20, p = 0.5))
```

```
## [1] 0.263176
```

It is comforting that this is the p-value that R reports when we ask it to perform a binomial test...

```
binom.test(13, 20, p = 0.5)
```

```
##
##  Exact binomial test
##
```



```
## data:  13 and 20
## number of successes = 13, number of trials = 20, p-value =
## 0.2632
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4078115 0.8460908
## sample estimates:
## probability of success
##                   0.65
```

Since the p-value we computed is not less than our cutoff of 0.05, our test is inconclusive. Also, note that the 95% confidence interval reported by the binomial test is consistent with our result; since it includes 0.5, it is plausible that the coin is perfectly fair.

If we tested a coin with the same success rate, but had planned to conduct additional trials, we'd gain statistical power and might be able to detect unfairness. Suppose we flip the coin 200 times, and observe 130 heads.

```
binom.test(130, 200)
```

```
##
##  Exact binomial test
##
## data:  130 and 200
## number of successes = 130, number of trials = 200, p-value =
## 2.653e-05
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5795494 0.7159293
## sample estimates:
## probability of success
##                   0.65
```

Since our p-value is less than 0.05, we are 95% sure that the coin is unfair. As we expect, this result is consistent with the reported 95% CI, which now excludes $P(H) = 0.5$.

To make sure you understand the cut-off, α, let's conduct a simulation. What if we repeated the same experiment – 20 coin tosses with a fair coin – 10,000 times. If we did a binomial test with a null hypothesis of a fair coin, in how many of those cases would you expect to get a p-value less than 0.05? We would be rejecting the null hypothesis even though it is true. This is known as a Type I error; we'll come back to this again at the end of the class.


```r
x <- rbinom(10000, 20, p = 0.5)
pval <- vector(length = 10000)
for (n in 1:10000) {
  pval[n] <- binom.test(x[n], 20, p = 0.5)$p.value
}
print(sum(pval <= 0.05) / length(x)) # sum of booleans is number of TRUEs
```

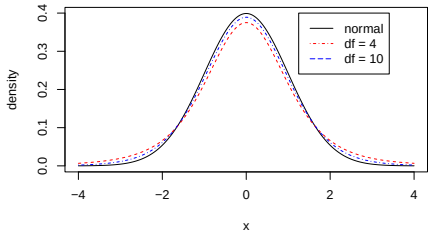
```
## [1] 0.0445
```

3 Comparing Means with the t-test

A very common statistical test is comparing two means. We saw the equations to compute the CI of the difference between two sample means in an earlier session; now we'll see how to do it with p-values. This test is called a t-test. The null hypothesis for a t-test is that the means of the two sampled populations are the same. Using our notation from our previous session, we have:

$$H_0\text{: } \overline{x_2} - \overline{x_1} = 0$$

Now, the means of your two samples have some observed difference, Δ, which is presumably not zero. We can compute the probability of taking two samples from the hypothesized distributions and obtaining sample means as far apart, or further, from each other as Δ. This probability is given by the t distribution, which we alluded to earlier. It happens to look a lot like a Gaussian for large n, but is heavier tailed for small n.



The basic t-test assumes that the data come from normal distributions with the same SD. The first assumption is not that important when n is large (central limit theorem), and the second can be corrected for.

A few examples of a t-test:


```
x1 <- rnorm(100); x2 <- rnorm(100); t.test(x2, x1)
```

```
##
##  Welch Two Sample t-test
##
## data:  x2 and x1
## t = -0.48822, df = 195.31, p-value = 0.6259
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3630220  0.2189514
## sample estimates:
##  mean of x  mean of y
## 0.02717270 0.09920797
```

```
x1 <- rnorm(100); x2 <- rnorm(100, mean = 1); t.test(x2, x1)
```

```
##
##  Welch Two Sample t-test
##
## data:  x2 and x1
## t = 6.942, df = 197.98, p-value = 5.404e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7466337 1.3391444
## sample estimates:
##   mean of x   mean of y
##  0.94489625 -0.09799281
```

```
x1 <- rnorm(100); x2 <- rnorm(100, mean = 0.2); t.test(x2, x1)
```

```
##
##  Welch Two Sample t-test
##
## data:  x2 and x1
## t = 1.553, df = 197.29, p-value = 0.122
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05638169  0.47430699
## sample estimates:
##  mean of x  mean of y
## 0.25776951 0.04880686
```

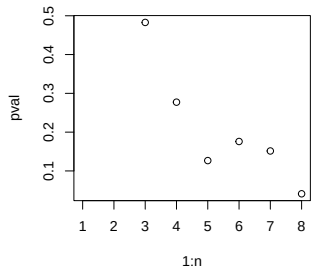
As demonstrated, the t-test (in this case), can distinguish between two distributions with

an SD of one that are separated by a mean of one, but not when they are as close as 0.2.

You can compute a p-value (and a CI) with just a mean, SD, and n from each group. Often, a paper will report a p-value, and you'll want to compute a CI; it can be done (see Motulsky, *Intuitive Biostatistics*).

One more simulation, so that you will hopefully avoid a dangerous but all too common mistake. What if you do a small experiment, get a p-value greater than your α, and decide to collect more data until you have big enough sample sizes to obtain a significant p-value? We'll set the random seed for this simulation, because the simulation can occasionally be very long.

```
set.seed(2)
n <- 3
x <- rnorm(n)
y <- rnorm(n)
pval <- c(NA, NA, t.test(x, y)$p.value)
while(pval[n] > 0.05) {
  x[n + 1] <- rnorm(1)
  y[n + 1] <- rnorm(1)
  test <- t.test(x, y)
  pval[n + 1] <- test$p.value
  n <- n + 1
}
plot(1:n, pval)
```



Does this convince you that you should decide on your sample size before conducting any tests? We'll learn how to do this based on your expected results in a later lecture on power calculations.

4 One-Tailed and Two-Tailed p-values

The p-values we have been computing so far are two-tailed p-values, because they compute the probability of seeing a difference as or more inconsistent with the null hypothesis in either direction. You will occasionally see one-tailed p-values reported. As you might have guessed, a one tailed p-value is usually just half of a two tailed p-value. This is the case for the t-test (because the t-distribution is symmetric), but isn't the case, for example, with Fisher's Exact test.

Whether you report a one- or two-tailed p-value is sometimes debatable (although one-tailed values are falling out of favor). If you are in a position to not care about or are willing to chalk up deviations in a direction other than that which you expect to random sampling only, it may be appropriate to use a one-tailed p-value. In our coin flipping example, if a casino were running a game where it wins when heads appears, and you were a gaming inspector looking for cheating casinos, you might use one-tailed p-values since you don't care about coins weighted towards tails. Doing this is tantamount to cutting off one side of a CI on the basis that you know it is impossible for the difference between means to be, say, negative. If you are not comfortable with this, then you should not use a one-tailed p-value.

In general, it is safer to stick with two-tailed p-values. If you choose to use a one-tailed value, you should have good reasons for doing so, and should state them before you collect any data. Computing a two-tailed p-value of 0.07 and then using a one-tailed value so you report a significant result is certainly not proper. Additional arguments in favor of two tailed p-values are that they are more conservative, and that they are consistent with CIs.

5 p-Value Cutoffs

Throughout the bulk of our work so far, we have been computing 95% CIs and using a p-value cutoff of $\alpha = 0.05$. While this is very common, it is important to realize that there is nothing sacred about this arbitrary number. It is often appropriate to choose a different cutoff, and we will spend a little time exploring the motivations and implications of doing so. Most importantly, it is important to choose your cutoff before you collect any data.

Some like to joke that statistics is never having to say you are wrong. When we "reject" a null hypothesis, we are not saying it is wrong, just that it is unlikely to be true; we are always trying to hedge. One never rounds a p-value to zero. You will see low p-values reported as "$< 10^{-9}$", or something similar, indicating that the null hypothesis is 'very, very unlikely to be true', but not 'impossible'.

Unfortunately for the cautious, when we perform a statistical test, compute a p-value, and compare it to a pre-established α, we are usually planning to make some decision and take some action (or not) based on the result. The decision will have consequences, sometimes quite profound ones. It is therefore important to understand what errors we might be inclined to make, how likely they are, and what their consequences are.

When we perform a test and we find a statistically significant result when, in fact, the null hypothesis should not have been rejected, we have committed a Type I error. Whenever you test a hypothesis that is not, in fact, deserving of a statistically significant conclusion, the probability of making a Type I error is α. The overall rate of Type I errors across many tests is dependent on how many hypotheses you test.

When we fail to reject the null hypothesis when it is in fact false, we are committing a Type II error. As above, determining the Type II error rate requires additional information on the proportion of tests that should lead to a significant conclusion.

As α decreases, the probability of making a Type I error goes down, and the probability of making a Type II error goes up. You can compensate for the latter by increasing the sample size. In order to quantify the Type II error rate, you need to know something about the rate of occurrence of the outcomes in the population(s) you are studying.

Consider the following statistical analyses:

Screening compounds for a drug that might have a desired biological effect:

Type I error: An ineffective compound is viewed as putatively effective. The consequence of making this error is the additional time and money invested in further testing an unhelpful compound.

Type II error: An effective compound is abandoned. The consequence of making this error is a missed opportunity to improve human health (or make big money for your employer).

In this case, you might want to increase α; even a value of 0.2 might be appropriate.

Phase III clinical trial of a drug to treat a disease that currently has no effective treatment:

Type I error: You determine that the drug is effective when it is not. The consequence is that patients are paying for a placebo, and unnecessarily enduring any side effects that might be present.

Type II error: You abandon an effective drug, and leave patients with no viable treatment. The consequence is a missed opportunity to improve human health.

In this case, you may consider increasing α, weighing the costs and side effects of the drug.

Phase III clinical trial of a drug to treat a disease that already has an effective treatment:

Type I error: You determine that your drug is more effective than the existing one when it is not. You have actively deprived patients of better care.

Type II error: You deprive patients of a better or less expensive treatment, but they are still adequately treated.

In this case, you may consider lowering α; a value as low as 0.01 might be appropriate. You will probably need a larger study to prove that your drug is better than the existing therapy; this seems appropriate when changing something that is recognized to work well.

6 Exercises

1. In the first code example of section 3, we performed a t-test on two univariate distributions where we knew that the difference between the true means was zero.

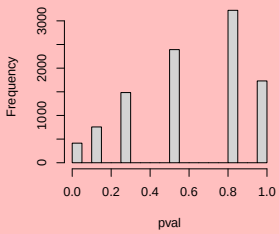
   ```
   x1 <- rnorm(100); x2 <- rnorm(100); t.test(x2, x1)
   ```

   Since the null hypothesis in this case is true, we expect that, most of the time, we should not reject it (i.e., that the p-value will be above 0.05). How often do you think we will be wrong and reject the null hypothesis even though it is correct? Confirm and demonstrate your answer by writing a simulation in R that performs this numerical experiment 10,000 times, and then summarizes the results.

   Augment your simulation so that it stores all 10,000 p-values. Plot a histogram of the p-values. What distribution do the p-values look like they come from?

   It looks like the uniform distribution, and in fact it is.

2. Challenge: In section 2, we performed an experiment similar to the one above, except that we simulated coin flips (i.e., sampling from the binomial distribution). Prepare a histogram of the p-values observed from that experiment. What do you think of the result?



We are seeing the effects of the fact that the distribution is discrete. Since there are only 21 possible outcomes of each experiment, and the distribution is symmetric, there are a limited number of p-values that we can see. An alternative way to visualize this may be with a boxplot and a stripchart.

3. In this problem, we'll consider a case where there is a very small difference between two groups, and where there is much data available. We'll simulate this as follows:

```
x <- rnorm(150000, mean = 5.00)
y <- rnorm(150000, mean = 5.01)
```

How big is the difference between the two groups relative to the mean? Relative to the standard deviation? Can you think of a case where such a small effect could be biologically significant?

We know that the effect size is 0.2% of the mean (0.01/5), and 1% of the SD (0.01/1). This is a pretty small effect, and wouldn't be considered biologically significant in most cases.

How often will a t-test call datasets generated like this as statistically significantly different. (You can answer this question by running a simulation not unlike the others in this assignment.)

About 79% of the time

