Introduction to Hypothesis Testing
or: “Everything you think you know about p-values is wrong”

BMR 617

James Denvir, Ph.D.

March 16th 2021

Example

- Using our usual data set, let's just look at body weight of the two groups of mice fed the standard Chow diet

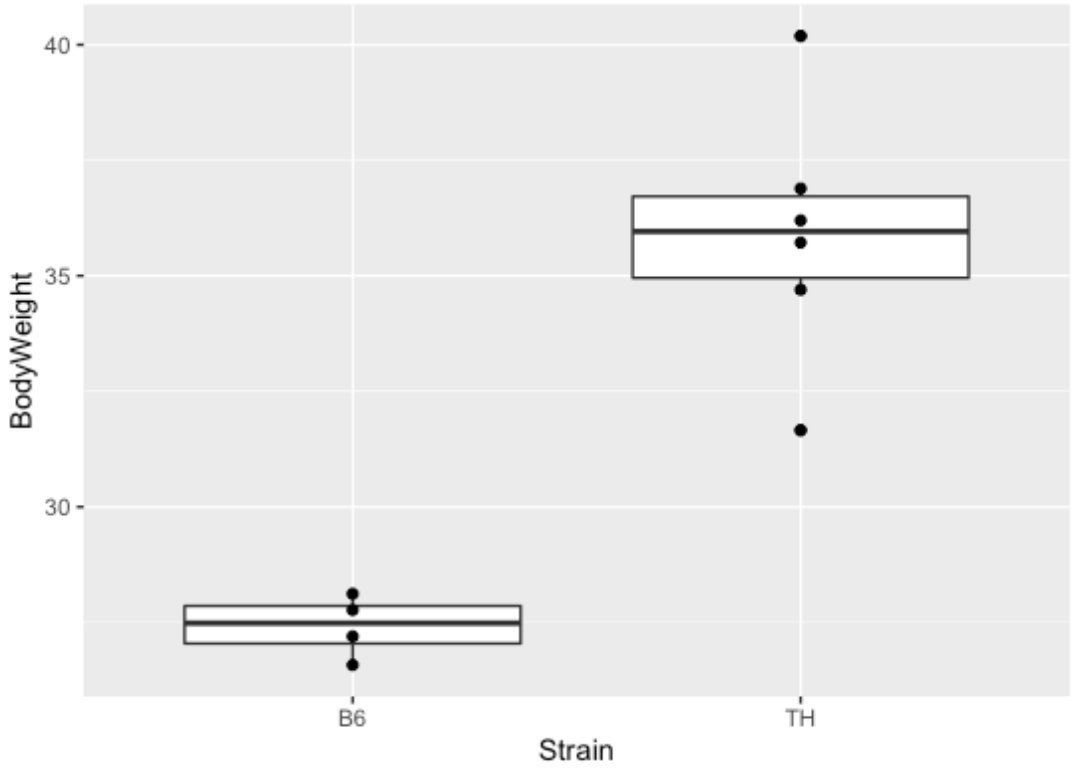
```
library(tidyverse)

met <-
read_csv("https://denvirlab.marshall.edu/BMR617-2021/data/TH-B6-metabolic.csv")

met <- separate(met, MouseID, sep="-", into=c("Strain", "Diet", "ID"))

chow_only <- filter(met, Diet=="Chow")

ggplot(chow_only, aes(x=Strain, y=BodyWeight)) + geom_boxplot() + geom_point()
```



Hypothesis testing

- Statistical hypothesis testing is
  Assessing evidence provided by the data in favor of or against some claim about the population
- In this example, we are going to test the hypothesis that the strain affects body weight, i.e. that the body weight is different in different strains
- Hypothesis testing is always formulated as testing two competing hypotheses. Generally speaking:
  - The *null hypothesis* is the "default"; i.e. the hypothesis we would tend to conservatively believe without any evidence to the contrary
  - The *alternative hypothesis* is (typically) the hypothesis we want to prove
- The null and alternative hypotheses should be mutually exclusive and exhaustive (i.e. one and only one of them is true)

Hypothesis testing: example

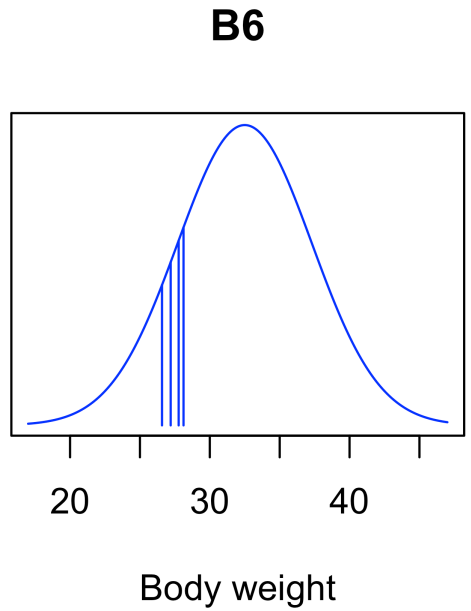
- In our example, the null hypothesis is
  - There is no difference in body weight between the B6 and TH strains on the chow diet
- The alternative hypothesis is
  - The body weights of TH mice fed the chow diet are different to the body weights of B6 mice fed the chow diet
- Note that one of these two must be true

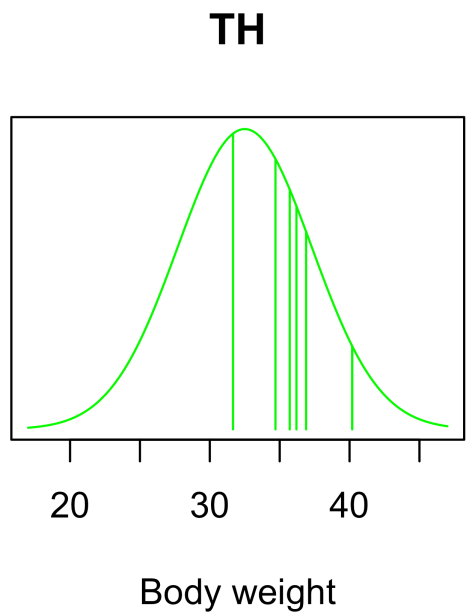
Hypothesis testing: theoretical and philosophical considerations

- Note that our data cannot *prove beyond any doubt* that the alternative hypothesis is true
- The hypothesis concerns the entire population (all possible TH and B6 mice)
- We only have data from a small sample
- Since the data vary, even if the null hypothesis is true, there is always some chance that we just happened to sample one group of data predominantly from one side of the distribution, and one group of data from the other

Null hypothesis is possible

- Even with our data, the null hypothesis is possible
- In the graphs, which illustrate the null hypothesis, the mean (population) body weight of both strains is around 32.5g
- In this scenario, it just happened that our B6 samples came from the left of the distribution, and our TH samples came from the right of the distribution
- Essentially, we are interested in how likely this scenario is (or isn't)
  - Unfortunately, **we can't actually measure this!**





Aside: mathematical proof

- In math, we prove things with complete certainty.
- On the right is a proof that no rational number (i.e. no fraction) is equal to the square root of 2
- We cannot possibly test all fractions
- So our strategy is to assume the statement is false, and show this leads to a contradiction
- This is called a “proof by contradiction”

Assume $\left(\frac{m}{n}\right)^2 = 2$, where m and n are integers and m/n is in its lowest terms

Then $m^2 = 2n^2$, so m^2 is even, and m is even. Write $m = 2k$.

So $m^2 = (2k)^2 = 4k^2 = 2n^2$, and $n^2 = 2k^2$, so n^2 is even, and n is even.

Now both m and n are even, so $^m/_n$ is not in its lowest terms.

Consequently, no fraction written in lowest terms can be equal to $\sqrt{2}$.

Hypothesis testing: process

- In statistical hypothesis testing, we cannot prove anything completely
  - Our data are subject to randomness
- We can only talk about probabilities
- The approach is similar to a “proof” by contradiction:
  - We collect data to provide evidence for our hypothesis of interest (the alternative hypothesis)
  - We then assume that the null hypothesis is true (i.e. that the hypothesis of interest is false)
  - Under this assumption, we calculate how likely it is we would observe data at least as “extreme” as the data we actually observed
  - If this probability is small, we reject the null hypothesis and conclude that the alternative hypothesis is true

The p-value

- In hypothesis testing, we calculate the probability that, assuming the null hypothesis is true, we would obtain data at least as extreme as the data we observe
- For example, in the B6 vs TH body weights, under the Chow diet, the mean body weight for B6 mice is 27.4g and for TH mice is 35.9g. The difference is 8.5g.
- So when we calculate the p-value, we are asking the question: “If the body weights of B6 and TH mice came from distributions with the same mean, what is the probability we would see a difference of 8.5g in our samples?”

Typical strategy

- Let's make the additional assumption that the body weights are normally distributed: suppose the *population* standard deviation for body weights of B6 mice is σ_{B6} and the population standard deviation for body weights of TH mice is σ_{TH}
- Our random variable, X, is the difference in means between B6 and TH. From our rules for normally-distributed random variables, X is also normally-distributed, and
  - The mean of X is the difference of means of B6 and TH
    - Under the null hypothesis this is zero
  - The standard deviation of X is $\sqrt{\frac{\sigma_{B6}^2+\sigma_{TH}^2}{2}}$
- There is a problem: we don't know the *population* standard deviations
  - We only have the standard deviations of our *samples*, which are an approximation

The t-distribution

- If we knew the population standard deviations, we would know that the difference in means was normally distributed, and we could calculate the probability it was at least 8.5g
- Typically, as in this case, we only know the sample standard deviation
- Fortunately, it is known that if m_{B6} and m_{TH} are the sample means, and if s_{B6} and s_{TH} are the *sample* standard deviations, then

$$\mathrm{t} = \frac{m_{TH} - m_{B6}}{\sqrt{\frac{s_{B6}^2 + s_{TH}^2}{2n}}}$$

follows a distribution called the t-distribution

The t-distribution

- The t-distribution is really a family of distributions
  - The actual distribution depends of a *degree of freedom*
  - For a single variable this is one less than the sample size
- Knowing the distribution, we can calculate the p-value
  - Remember, this is the probability that we would get a difference of at least 8.5g between the two groups of sample, if the groups were drawn from populations of equal means
  - Another way to think about it is this: even if the body weights of B6 and TH mice were the same, our means would be different just by sampling. How likely is it we would see the difference we actually saw
- The calculation of this p-value is called a t-test (sometimes a "Student's t-test") after the distribution

T-test in R

- R has a `t.test` function
- Try the following:
  ```
  t.test(BodyWeight ~ Strain, data=chow_only)
  ```
- What is the p-value?
- What are the estimated means? What's the difference between them?

Interpreting a hypothesis test

- Care is needed in interpreting a hypothesis test
- Very commonly misinterpreted, even by advanced professional scientists
- In our test, the p-value was 0.00043
- So, if there were no difference in body weight between B6 and TH mice, there would be only a 0.043% chance of obtaining a difference in means this big with these sample sizes
  - Since this is so unlikely, we would reject the null hypothesis and conclude that the two strains had different body weights

P-value thresholds

- The natural question here is “how small is a small p-value?”
- The standard approach is to choose a threshold, below which we will reject the null hypothesis
  - Do this *before* collecting and analyzing data
- Hypothesis testing was invented in the late 1800s by Ronald Fisher
- In his original paper, as an example, he used a threshold of 0.05
- This has become, for no particularly good reason, a “standard” threshold

Important lesson #1: The p-value is *not* the probability the null hypothesis is false

- The p-value is the probability of obtaining data as extreme as that observed, under the assumption that the null hypothesis holds
- This is not the same as the probability that the null hypothesis is true
- Commonly mistaken as such
- We can almost never compute the probability the null hypothesis is true

An imaginary scenario

- Prof. Cavendish O'Leary ("Cav") runs a lab in the Institute for Unlikely Discoveries. Over the course of his career, he conducts 1000 studies. His ambition is to discover something that will make him famous and chooses studies that, if successful, will be groundbreaking and paradigm shifting in the field.
- Prof. Prudence Dent ("Pru") runs a lab in the Institute for the Establishment of Known Facts. Over the course of her career, she also conducts 1000 studies. Her aim in life is to build a steady body of solid, reproducible publications, and as such she only studies things with an abundance of solid evidence.

The a-priori probability

- Because Cav O'Leary and Pru Dent exist only in a fictional universe, of which we are the creators and omnipotent beings, we know that 2% of the hypotheses that Prof. O'Leary tests are really true, and that 90% of the hypotheses that Prof. Dent tests are really true
  - These values are called the "a-priori" probabilities of the hypotheses, and these are never known in the real world
- Both Prof. O'Leary and Prof. Dent use a p-value threshold of 0.05 to determine "statistical significance" (i.e. decide whether or not to reject the null hypothesis)
- Furthermore, both design their experiments and choose their sample sizes to give a *statistical power* of 80%: if their hypothesis is true, there is an 80% chance they will get a p-value less than the threshold
- Suppose Pru Dent performs an experiment and rejects the null hypothesis. What is the probability that the null hypothesis is false? What about Cav O'Leary?

Pru Dent computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes			
	No			
	Total			1000

Dr. Dent performs 1000 experiments in her career

Pru Dent computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes			100
	No			900
	Total			1000

In 90% of the experiments, her hypothesis is true (a-priori probability, unknown in reality)

Pru Dent computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes			100
	No	180	720	900
	Total			1000

When the null hypothesis is false, there is an 80% chance of correctly rejecting it (statistical power)

Pru Dent computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes	95	5	100
	No	180	720	900
	Total			1000

If the null hypothesis is true, there is a 0.05 chance of incorrectly rejecting it (definition of p-value and choice of threshold)

Pru Dent computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes	95	5	100
	No	180	720	900
	Total	275	725	1000

Complete column totals: Dr. Dent rejects the null hypothesis 725 times out of her 1000 experiments, and fails to reject it 275 times

Pru Dent computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes	95	5	100
	No	180	720	900
	Total	275	725	1000

Out of the 725 times Dr. Dent rejects the null hypothesis, that is a correct decision 720 times. So if she rejects the null hypothesis, the probability her hypothesis is true is 720/725 = 0.993

Cav O'Leary computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes			
	No			
	Total			1000

Dr. O'Leary performs 1000 experiments in his career

Cav O'Leary computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes			980
	No			20
	Total			1000

Only 2% of Dr. O'Leary's hypotheses are actually true

Cav O'Leary computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes			980
	No	4	16	20
	Total			1000

When the null hypothesis is false, there is an 80% chance of correctly rejecting it (statistical power)

Cav O'Leary computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes	931	49	980
	No	4	16	20
	Total	935	65	1000

Complete column totals: Dr. O'Leary rejects the null hypothesis 65 times out of his 1000 experiments, and fails to reject it 935 times

Cav O'Leary computation

		Reject null hypothesis		
		No	Yes	Total
Null hypothesis really true	Yes	931	49	980
	No	4	16	20
	Total	935	65	1000

Out of the 65 times Dr. O'Leary rejects the null hypothesis, that is a correct decision 16 times. So if he rejects the null hypothesis, the probability his hypothesis is true is 16/65 = 0.246

Conclusions

- Many people assume that if we reject the null hypothesis (with a p-value threshold of 0.05), it means there is a 95% chance the null hypothesis is false (i.e. that the hypothesis of interest is true)
- This fictional scenario demonstrates this is not true
- The probability the null hypothesis is false, given we have a p-value less than 0.05, depends greatly on the *a-priori* probability
  - The prior probability that the null hypothesis was false
  - The problem is, again, that we never actually know this
- In our (somewhat extreme) examples, the probabilities we calculated were 0.993, and 0.246,respectively.

Important lesson #2: A “non-significant” p-value doesn’t lead to any conclusion

- In the hypothesis-testing scenario, we choose a p-value threshold (below which we reject the null hypothesis)
- We then collect data and perform our analysis, computing the p-value
- If the p-value is below the threshold, we reject the null hypothesis and conclude the hypothesis of interest is correct
- If the p-value is above the threshold, we simply fail to reject the null hypothesis
  - Notice here that we don’t draw any conclusion
  - We certainly cannot conclude the null hypothesis is true

P > 0.05 is often misinterpreted

- This last point is very commonly misunderstood
- It is very common to see statements, even in published papers, like "There was no difference in the blood pressures of patients on drug A or drug B (p>0.05)"
  - This is not a valid conclusion from a "non-small" p-value!
  - At best, you can conclude that the experiment or study failed to demonstrate a difference

The Frequentist versus Bayesian argument

- The hypothesis-testing approach is known as the “Frequentist” approach to statistics
  - How frequently would we get data like this if the null hypothesis were true
- Advantages include the fact that it is unbiased
  - Computation of the p-value doesn’t rely on any guess work or subjective evaluation of the veracity of the hypothesis
- Disadvantage is that it doesn’t really tell us what we want to know:
  - We want to know the probability the null hypothesis is true
- The approach we demonstrated in the tables is known as the “Bayesian” approach
  - Essentially uses Bayes theorem to “update” the probability the null hypothesis is true in light of the data we observe
- Advantage is that it’s closer to the question we really want to answer
- Disadvantage is that we need an initial, subjective, estimate of that probability

Frequentists are still winning

- The frequentist approach is still, by far, the most prominent in scientific literature
  - Particularly in biomedical sciences
- Both approaches have their merits and sometimes it is useful to use tools from each
- We will spend some time examining hypothesis testing, but bear in mind the limitations
  - We will see more scenarios where we have to be careful how to interpret p-values later

Further reading

- In March 2019, the American Statistical Association published a special edition of their journal, *The American Statistician*, addressing misunderstandings of hypothesis testing and p-values
- https://www.tandfonline.com/toc/utas20/73/sup1
  - Read (as a minimum) the editorial from this issue
- Nature published an accompanying article, signed by 800 academic and professional statisticians
  - https://www.nature.com/articles/d41586-019-00857-9
- The key point these articles make is to stop categorizing and dichotomizing results based solely on p-values
  - Instead, *thoughtfully* assess all the data and analyses