Introduction to Probability and Statistics Eleventh Edition

More About Hypothesis Testing

p-values

- Remember from last week that the p-value tells us the probability that the null hypothesis is true?

- NO! This is a common misunderstanding of the p-value. Don't make the mistake of thinking a p-value tells you this.

p-values

- If a test statistic has a p-value of 0.05 we are saying that if the null hypothesis is true then we have a 5% chance of observing an effect this size or larger every time we test a random sample.
- So, if H_0 is true then we have a 5% chance of choosing a sample at random that happens to be in the rejection region for p = 0.05.
- This means that, even if the null hypothesis is true, 1 in 20 times we do an experiment we will reject it anyway!

p-values

- On the other hand, if H_0 is false then we may choose a sample that happens to have a p-value greater than our significance level.
- This would mean we would keep believing the null hypothesis even though it is false!

Errors

Four Possible Outcomes in a Hypothesis Test

Decision	Null Hypothesis, H_0, is True	Null Hypothesis, H_0, is False
Reject H_0, and conclude H_a	Type I error **False positive!**	Correct decision
Do not Reject H_0 (Accept H_0)	Correct decision	Type II error **False negative!**

Four Possible Outcomes in a Hypothesis Test

Decision	Null Hypothesis, H_0, is True	Null Hypothesis, H_0, is False
Reject H_0, and conclude H_a	Type I error **False positive! (α)**	Correct decision
Do not Reject H_0 (Accept H_0)	Correct decision	Type II error **False negative! (β)**

Types of Error

Definition 1

A **Type I error** for a statistical test is the error of rejecting the null hypothesis when it is true. The probability of making a Type I error is denoted by the symbol α.

Definition 2

A **Type II error** for a statistical test is the error of accepting the null hypothesis when it is false. The probability of making a Type II error is denoted by the symbol β.

Note that 'accepting' the null hypothesis just means we don't reject it. We don't necessarily have evidence that it is true. For example, suppose the null hypothesis is that a ball is blue. I test it by checking whether it is red. If the test succeeds then I know the ball is red, and so not blue. But if the test fails then all I know is that it is not red, but that doesn't prove it is blue (e.g. it could be green).

Type I errors

- We can measure the likelihood of making a type I error. It's just the level of significance we set for the test (e.g. .01, .05 etc.)
- When we reject H_0 we can quantify our risk of making a mistake.
- This risk is the significance level α.

Type II errors

- The probability of making a type II error isn't usually known or controlled by the experimenter.
- The alternative hypothesis usually gives a range of values for, e.g., the mean, so we can't build a distribution around it like we do for the null hypothesis.
- So when we reject H_a it's harder quantify the risk that we've made a mistake.
- We won't say much about it on this course.

Four Possible Outcomes in a Hypothesis Test

Decision	Null Hypothesis, H_0, is True	Null Hypothesis, H_0, is False
Reject H_0, and conclude H_a	Type I error	Correct decision **Sensitivity**
Do not Reject H_0 (Accept H_0)	Correct decision **Specificity**	Type II error

Statistical Power

- When we have a binary hypothesis test, we say the **power** or **sensitivity** of the test is how likely it is to reject the null hypothesis when it is false.
- That is, how likely it is to correctly reject H_0.
- For example, if we are screening for an illness, how likely is our test to correctly identify positive cases.
- The power of a test is 1 minus the chance of making a type II error $(1 - \beta)$.

$$\text{sensitivity} = P(\text{reject } H_0 | H_0 \text{ is false}) = P(\text{accept } H_a | H_a \text{ is true})$$

Specificity

- The **specificity** of a binary hypothesis test is how likely the test is to keep the null hypothesis when it is true.
- For example, if we are screening for an illness how likely are negative cases to be correctly identified.
- This is 1 minus the probability of making a type I error $(1 - \alpha)$.

$$\text{specificity} = P(\text{keep } H_0 | H_0 \text{ is true}) = P(\text{reject } H_a | H_a \text{ is false})$$

Four Possible Outcomes in a Hypothesis Test

Decision	Null Hypothesis, H_0, is True	Null Hypothesis, H_0, is False
Reject H_0, and conclude H_a	Type I error	Correct decision **Sensitivity**
Do not Reject H_0 (Accept H_0)	Correct decision **Specificity**	Type II error

Four Possible Outcomes in a Hypothesis Test

Decision	Null Hypothesis, H_0, is True	Null Hypothesis, H_0, is False
Reject H_0, and conclude H_a	Type I error (α)	Correct decision ($1-\beta$)
Do not Reject H_0 (Accept H_0)	Correct decision ($1-\alpha$)	Type II error (β)

Four Possible Outcomes in a Hypothesis Test

Decision	Null Hypothesis, H_0, is True	Null Hypothesis, H_0, is False
Reject H_0, and conclude H_a	Type I error **(α)** **False positive**	Correct decision **Sensitivity (1-β)**
Do not Reject H_0 (Accept H_0)	Correct decision **Specificity (1-α)**	Type II error **(β)** **False negative**

A Warning – base rate fallacy

- Suppose we are running a screening program for a new deadly disease called boneitis.
- We have a test with a sensitivity of .95 and a specificity of .99.
- So 95% of the people who have the disease will be correctly diagnosed.
- And 99% of people who don't have the disease will be correctly diagnosed.
- This is a pretty good test right?
- Maybe NOT.

A Warning – base rate fallacy

- Suppose we take the test and it comes back positive, saying we have boneitis.
- This is very worrying. The **specificity** of the test is .99, so the chance of us having boneitis is 99% right?
- WRONG! There's one important piece of information we need to know to calculate how likely it is we actually have boneitis.
- We need to know the percentage of the population that has boneitis (the **base rate**).

A Warning – base rate fallacy

- Suppose boneitis is **very rare**, and only affects **1** in every **million** people.
- Then if we randomly test 1,000,000, only 1 of them should have boneitis, and the other **999,999** will **not have it**.
- However, our test only **correctly identifies negative cases 99%** of the time.
- So from the 999,999 people who do not have boneitis, **1** in **every hundred** will be diagnosed as having it **incorrectly**.
- This is roughly **10,000** people. So from our sample of 1 million we get one true positive (if it is detected) and **10,000 false positives**, and the likelihood that we have boneitis is actually only around **1/10,000=0.0001.**
- This is a real problem for medical screening procedures.

Class Activity 12

1. **Hotel Costs.** We explored the average cost of lodging at two different hotel chains. We randomly select 50 billing statements from the computer database of Marriott and Westin, and record the nightly room rates. A portion of the sample data is shown in the table.

	Marriott	Westin
Sample size	50	50
Sample mean ($)	150	165
Sample standard deviation	17.2	22.5

Use the critical value approach to determine if there is a significant difference in the average room rates for the Marriott and the Westin hotel chains. Use $\alpha = 0.01$.

Class Activity 12

2. **Treatment versus Control.** An experiment was conducted to test the effect of a new drug on a viral inflection. After the infection split into 2 groups of 50. The first group, the control group, received no treatment for the infection, and the second group received the drug. After a 30-day period, the proportions of survivors, $\widehat{p_1}$ and $\widehat{p_2}$, in the two groups were found to be 0.36 and 0.60, respectively.

Is there sufficient evidence to indicate that the drug is effective in treating the viral infection? Use $\alpha = 0.05$.

Class Activity 12

3. **Taste Testing.** A food store is doing a taste test of a local brand vs a national brand. They believe that the national brand should be better. In other words, they believe that most people will prefer the national brand to the local brand.

   a) State the null and alternative hypothesis to be tested. Is this a one- or a two-tailed test?

   b) Suppose the store tests the hypothesis by asking 35 people chosen at random. Suppose that 8 out of the 35 people prefer the national brand. Use this information to test the hypothesis in part a). Use $\alpha = 0.01$. What practical conclusions can you draw from the results?

Videos

- Type errors and the power of the test https://www.youtube.com/watch?v=7mE-K_w1v90
- Base rate fallacy https://www.youtube.com/watch?v=VeQXXzEJQrg