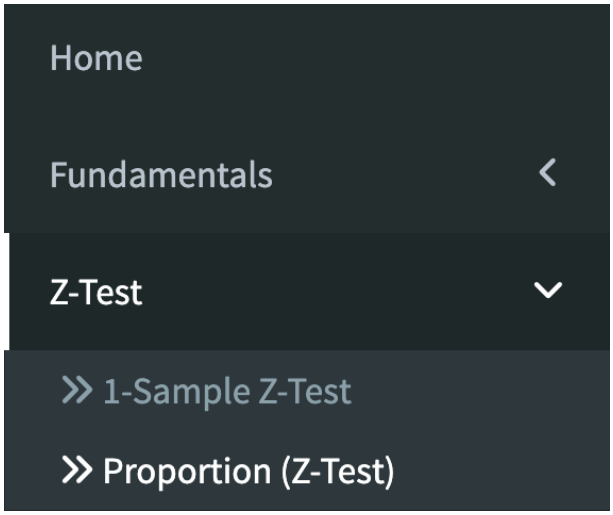
Week 10

Some of the questions this week were very tricky, so I wanted to provide additional help and pointers to help with your revision and approach to these questions.

Regarding question 3, which asked you to create a confidence interval and then to do a sanity check on this interval, please review the document here. I spent a lot of time working on this to hopefully help explain this content! I intend to turn it into a blog post at some point.

Regarding question 4, please play around with the 'Hypothesis Tests Visualised' website I made. In particular, go to the 'Proportion (Z-Test)' page:



- This website page should help you understand the process of doing a proportion hypothesis test. You should try to implement the steps using code.
- For question 4, in some labs, I said that we were doing a 2-sided t-test. This is actually wrong. We are doing a 1-sided test as the question specifically asks for **greater than 50%**. Hence, your hypotheses should look like this:

Null Hypothesis

$$H_0 : p = 0.5$$

Alternate Hypothesis

$$H_1 : p > 0.5$$

Regarding question 5, we didn't go through this in the lab, but please look at the answers when they are released at the end of the week.

Week 13

Checking the assumptions behind the first question (2-sample t-test).

When it came to checking the assumptions of the 2-sample t-test, I found it hard to tell whether the assumptions were satisfied or not. The main reason it was hard was that we had such a small sample size (one group had 10 responses, and the other had 8). On thinking about this more, I want to place some more concrete thoughts down.

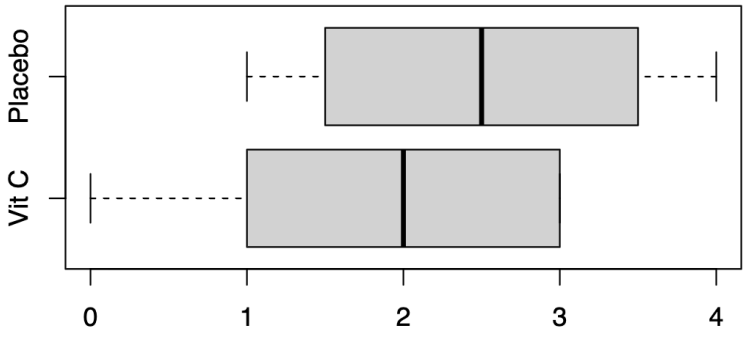
As a reminder, these are the assumptions for a 2-sample t-test:

- Independence and random samples: The two samples must be independent, and the observations within each sample should also be independent. The samples should also be random.
- Equal variances (for the pooled or standard 2-sample t-test): The two groups should have approximately equal population variances. This assumption is not required if you use the Welch's 2-sample t-test.
- Normality: The sampling distribution of the difference in means should be approximately normal.

Let's consider the assumptions for the question in the lab!

Firstly, regarding the independence and random samples assumption, we assume this to be true, but we cannot really assess this as we do not know enough about the data collection methodology. In the real world, you could determine this by reading the scientific paper and reading the methods they used.

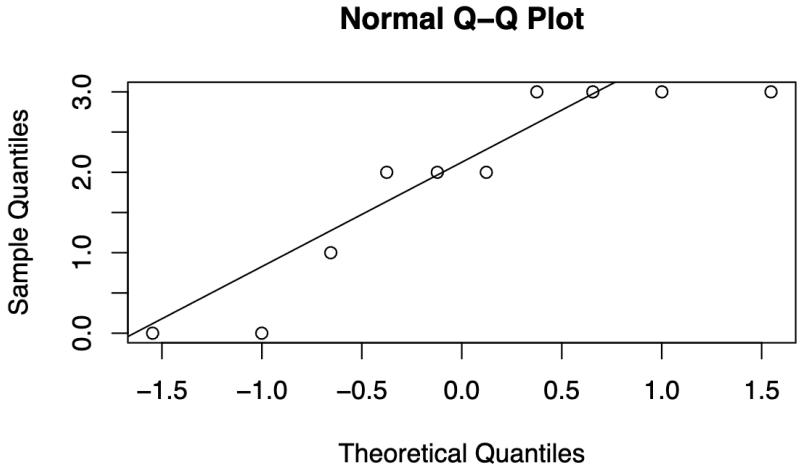
Let's consider the equal variance assumption. From the boxplots below, we can see that the spread of the boxplots is similar:



Also, we confirmed that the standard deviation of each sample is also very similar (1.197 and 1.195). Overall, we can say that these assumptions are satisfied. Even though the Vitamin C group does not have an upper whisker, this is likely a result of the small sample size.

Now, in order to **check the normality assumption,** there are two things that we can do. Firstly, we can look at whether the boxes are reasonably symmetric or not. We can say that they are reasonably symmetric (the median appears right in the middle of the box). However, we can also see that the Vitamin C group is actually a bit asymmetric. We can tell this by observing how there is no upper whisker for the Vitamin C group, and also when we expect the values in the Vitamin C group, there are many more “3” values than other values. Overall, though, we do claim that this is reasonably symmetric, yet there is some doubt.

The second thing we can do is look at the qq plots for each of the samples. The qq-plot for the placebo sample looks good, and we can claim that this data is normally distributed. However, when looking at the qq-plot for the Vitamin C sample (as shown below), we can see substantial deviation from the line in the right tail.



This is likely because of the integer nature of the data, and because we have a very small sample size. In this case, I would be leaning towards claiming that the normality assumption does not hold. When considering a qq-plot, small deviations at the tails are expected, but I think it is concerning to have so much deviation here with such a small sample.

This next paragraph is extension material! If I had this case in the real world, where I think there is serious doubt whether the normality assumption holds or not, one thing I would if I could is go and collect a larger sample, and see what happens to the trend. I worry that this plot could be a result of such a small sample size (a single value could distort things). You could also do a Shapiro-Wilk test on each sample, which is a hypothesis test to test whether the data follows a normal distribution. If you do conclude that the data is not normally distributed, you could do another type of hypothesis test that does not require the normality assumption. In this case, it would be a Mann–Whitney U test.

I hope this clears things up :)

Credit: Generative AI (ChatGPT) was used in writing this.