Wk5 exam qn discussion:

The student-t test assumes normality in the data being analyzed, so once the data was partitioned by the gene of interest and the outcome variable, I performed a Shapiro-Wilks test to determine whether the data follows a normal distribution, and can see it has a p-value of < 2.2e-16 and equal to 1.746e-06 for tumor and normal tissues, respectively. With this test, the hypothesis is that the data is normally distributed, as the values are far less that 0.05, the assumption of normality is rejected which indicates the data is significantly different from a normal distribution. This could suggest the data follows a different distribution, is skewed or has outliers.

Due to this result I can't use the t-test on the raw data. Due to the central limit theorem however, I can use the t-test if I have enough independent identically distributed random variables, as the random variables converge to a normal distribution as n increases. At this point, the original paired structure is lost, as the focus is now on comparing independent sample means derived from tumor and normal tissues. Consequently, the appropriate test to use in this scenario is a two-sample t-test rather than a paired t-test. The null hypothesis for this test is that there is no difference in the underlying means of ERBB2 expression between the tumor and normal samples.

The p-value from the two-sample t-test is <2.2e-16. This extremely low p-value means that if there were no true difference between the groups, the probability of observing the current (or a more extreme) difference would be nearly zero. Therefore, we reject the null hypothesis and conclude that there is a statistically significant difference in ERBB2 expression between the two outcomes. Additionally, the test provides a 95% confidence interval of approximately 391 to 580, which indicates that we can be 95% confident that the true mean difference in expression levels lies within this range. This confidence interval being far from zero, also reinforces the significance of the difference observed. Lastly, the sample estimate shows that, on average expression of tumor samples is 11143.5, and 10653.33 for normal samples, giving an average difference of 490 between the two. We can see that the paired t-test results are very similar, also with p-value of <2.2e-16 and mean difference of 490. Using a two-sample t-test on the original paired data would have been problematic since it ignores the dependency between paired observations. But as we transformed this data with the simulations of independent samples, the two-sample test is the best method.