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.

qn2.

Pearson's Correlation

* Use When: You have continuous data and want to measure the linear relationship between two variables.
* Example: Testing if there is a linear relationship between height and weight.

Spearman's Correlation

* Use When: You have continuous or ordinal data and want to measure the monotonic relationship between two variables.
* Example: Testing if there is a consistent increase or decrease in one variable as the other variable increases, even if the relationship is not linear.

Both the coefficient estimates and the p-values are very similar from both tests. The Pearson coefficient estimate represents the strength and direction of the relationship between the two variables, while the spearmans coefficient estimate represents...

We can see that the p-values for each are very similar at 0.002 and 0.003, both of which are less than 0.05. Therefore we can reject the null hypothesis, which states there is no correlation between the two variables, and accept the alternate hypothesis that there is a significant relationship between the two variables.

While I would expect there to be a predictive relationship between mRNA levels of a gene and protein levels of the same gene, it doesn’t necessarily mean it would be completely linear, there are many factors that can interfere with mRNA being translated and then post translational modifications resulting in an active protein. While both methods work well for continuous data, Pearson’s correlation assumes normality, where Spearman’s correlation makes no assumptions on the data and tests whether there is a consistent increase or decrease based on the other variable, even if it’s not particularly linear. Even though the results were very similar between methods, to test which test is perferred I checked the normailty of each variable, and found that Shapiro-Wilks p-value for the mRNA data was p=~0.001, and the RPPA data p=0.74. Now we’re in a situation where one set of data follows a normal distribution and the other doesnt. Since one set is not normal, I would still lean towards using the Spearman method as the Pearson method is assuming both sets of data are normal.

\_\_

Both the coefficient estimates and the p-values are very similar from both tests. The Pearson coefficient estimate represents the strength and direction of a linear relationship between the two variables, while the Spearman coefficient estimate represents the strength and direction of a consistent, unidirectional relationship between the two variables, regardless of whether the relationship is linear.

We can see that the p-values for each are very similar at 0.0029 and 0.0032, both of which are less than 0.05. Therefore, we can reject the null hypothesis, which states there is no correlation between the two variables, and accept the alternate hypothesis that there is a significant relationship between the two variables.

While I would expect there to be a predictive relationship between mRNA levels of a gene and protein levels of the same gene, it doesn’t necessarily mean it would be completely linear. There are many factors that can interfere with mRNA being translated and then post-translational modifications resulting in an active protein. While both methods work well for continuous data, Pearson’s correlation assumes normality, whereas Spearman’s correlation makes no assumptions about the data and tests whether there is a consistent increase or decrease based on the other variable, even if it’s not particularly linear.

Even though the results were very similar between methods, to determine which test is preferred, I checked the normality of each variable and found that the Shapiro-Wilk p-value for the mRNA data was approximately 0.001, and the RPPA data p-value was 0.74. Now we’re in a situation where one set of data follows a normal distribution and the other does not. Since one set is not normal, I would still lean towards using the Spearman method, as the Pearson method assumes both sets of data are normal.

qn3.

After segregating the data and constructing a contingency table of results, I was ready to proceed with selecting a signifcance test for two categorical variables. Based on the lecture notes, to use Pearson’s chi-square test the counts should be relatively even across cells of the contingency table, and the count of observations should be at least 5-10 in each cell. According to another article I found, expected counts (computed based on the fact that there is no association between groups – so P(HER2-ICH, ER-ICH) = P(HER2-ICH)\*P(ER-ICH)) should also be at least 5 in a minimum of 80% of cells [1]. Even though the counts did not look particulary uniform, I decided to see what the expected table would look like. One cell (HER2-ICH positive, ER-ICH negative) had an expected value of 4.57, so 25% of the table was <5. Therefore, I continued on to use Fishers exact two-sample test which looks for a correlation in either direction.

The p-value is greater than 0.05, which means we can accept the null hypothesis that states that the ICH status (presence) of the HER2 protein and ER protein are not correlated in breast cancer patients (or should I say in our sample of breast cancer patients?). Looking at the confidence interval returned from the Fishers test we can see that the range includes 1, and that our estimated odds ratio is 0.85. The odds ratio represents a measure of association between our groups, where a ratio of 1 means they are independent of each other. Having our confidence interval span across 1 suggests that the true underlying odds ratio of the population could be 1 (not that it would change, the population ratio is fixed). This further suggests that there is no significant association between the variables. Even though our estimated odds ratio is slightly less than 1, since the p-value is >0.05 it is not a significant difference. In contrast if the test returned an odds ratio of 0.85 and a p-value of 1.2e-7, this could signify ...

\_\_

After segregating the data and constructing a contingency table of results, I was ready to proceed with selecting a significance test for two categorical variables. Based on the lecture notes, to use Pearson’s chi-square test the counts should be relatively even across cells of the contingency table, and the count of observations should be at least 5-10 in each cell. According to another article I found, expected counts (computed based on the assumption that there is no association between groups – so P(HER2-IHC, ER-IHC) = P(HER2-IHC) \* P(ER-IHC)) should also be at least 5 in a minimum of 80% of cells [1]. Even though the counts did not look particularly uniform, I decided to see what the expected table would look like. One cell (HER2-IHC positive, ER-IHC negative) had an expected value of 4.57, so 25% of the table was <5. Therefore, I continued on to use Fisher's exact two-sample test which looks for a correlation in either direction.

The p-value is greater than 0.05, which means we can accept the null hypothesis that states the IHC status (presence) of the HER2 protein and ER protein are not correlated in the sample of breast cancer patients. Looking at the confidence interval returned from Fisher's test we can see that the range includes 1, and that our estimated odds ratio is 0.85. The odds ratio represents a measure of association between our groups, where a ratio of 1 means they are independent of each other. Having the confidence interval span across 1 suggests that the true underlying odds ratio of the population could be 1 (not that it would change, the population ratio is fixed). This further suggests that there is no significant association between the variables. Even though our estimated odds ratio is slightly less than 1, since the p-value is >0.05 it is not a significant difference.

reference:

McHugh M. L. (2013). The chi-square test of independence. Biochemia medica, 23(2), 143–149. <https://doi.org/10.11613/bm.2013.018>

qn4.

If our output is tumor mutation burden, it means that the regression model aims to calculate, as our input changes, what happens to our output. Therefore, tumor mutation burden (TMB) is the dependent variable in this analysis, and fraction genome altered (FGA) is the independent variable.

summary(FgaTmb.lm)$coeff...

The regression coefficients section of the results are really important to interpreting our model. The intercept estimate represents the expected value of our output (dependent variable) when the independent variable is zero. In other words, it represents the value of the tumor mutation burden where there is no association with fraction genome altered. The standard error, t value, and Pr(>|t|) measure the average deviations of observed points from the calculated regression line, the absolute value of the coefficient divided by its standard error, and the probability of obtaining a t-value the same or more extreme than our result, respectively. The higher the t-value, the greater the indication that the intercept is significantly greater than zero, and the lower the p-value the stronger the evidence against the null hypothesis (for this row it states the intercept is zero). Because the p-value  1.09e-5 is less that our cutoff, we can conclude that the intercept estimate is not due to chance and that when there is no alteration in the genome, the tumor mutation burden has a baseline of 1.2363.

Looking at the “FGA” row of the coefficients section, the estimate is equal to 1.8378 with a p-value of 0.0154. Since we are now looking at the coefficient of the independent variable, the estimate relates to the slope or rate of change per one measurement of FGA. This means that for each one-unit increase in the fraction of the genome that is altered, the Tumor Mutation Burden is expected to increase by approximately 1.8378 units.

The p-value indicates that this relationship is statistically significant at the 5% significance level, meaning that there is strong evidence to suggest that FGA has a predictive relationship with TMB, which is our alternate hypothesis for this row.

The 95% confidence interval for the coefficient of FGA ranges from 0.3590 to 3.3165. This interval suggests that we can be 95% confident of our model in predicting the true effect of FGA on TMB lying within this range. Since the confidence interval does not include zero, it also supports our conclusion that there is a significant positive relationship between the fraction of the genome that is altered and tumor mutation burden.

Lastly the R-squared value represents the “explained” variance, or the proportion of the variance between the means of each variable. In this case R-squared is 0.05844, which means that only 5.84% of the variation seen is due to the relationship between the two variables. This is very small and indicates that there is a high level of variance in TMB that is unaccounted for.