GENOME 560
Problem Set #2

(Due April 26th 8:59am)

---

1. **[20 points] The t-test**

   You are interested in the transcriptional changes during early stages of the innate immune response. You obtain lymphoblast cell lines from 10 individuals and for each one measure expression levels at baseline (untreated) and following treatment with the drug immiquimod (which is a TLR8 agonist). The following table shows gene expression levels for a particular transcript.

Individual	Baseline	Stimulated
1	-0.24	1.74
2	0.25	2.1
3	1.12	1.65
4	-0.06	2.65
5	0.46	3.11
6	0.17	2.31
7	0.02	1.87
8	1.10	3.21
9	0.55	2.19
10	0.98	1.75

   (a) Perform a one sample t-test to test the hypothesis that baseline expression levels are significantly different than zero. Clearly state the null and alternative hypotheses and submit R code, test statistic value and p-value.

   (b) Use a paired t-test to test the hypothesis that gene expression levels are significantly different between baseline and stimulated conditions. Again, clearly state the null and alternative hypotheses and submit R code, test statistic value and p-value.

   (c) An alternative way of analyzing the data as opposed to a paired two sample t-test (part b) is to create a new phenotype for each individual defined as the difference between stimulated and baseline expression. Formally, let x_i and y_i denote the expression level for the $i-th$ individual in baseline and stimulated conditions, respectively. Then define $z_i = y_i - x_i$. Perform a one sample t-test on the vector of z_i values. Clearly state the null and alternative hypotheses and submit R code, test statistic value, and p-value. How does your result compare to that obtained from part b?

2. **[40 points] A Simulation Study to Investigate the Power of the t-test**

   (a) Generate two groups of n=5 observations, each from a standard normal distribution. Perform a t-test to test $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$ and record the resulting p-value. Now repeat these steps 2000 times and calculate the number of times you would have rejected the null hypothesis with $\alpha = 0.05$. Explain this result and why it is/is not expected. How would the result change if we used $\alpha = 0.01$?

   (b) Repeat problem a, but now change the sample size to 50 and then to 500. Do the results change? Why or why not?

   (c) Repeat problem a, but now change the mean of one of the groups to be $\mu = 1$. What did you expect would happen? Summarize your results.

   (d) Repeat problem b, but now change the mean of one of the groups to be $\mu = 1$. What did you expect would happen? Summarize your results.

   (e) Plot your results for problems 1-4. Try to produce one plot that summarizes the results from all four problems, showing how your power (probability of rejecting the null) changes based on sample size and the difference in means. The plot should be self-explanatory.

3. **[40 points] Chi-Square Test**

   An expensive private school also asks for donations. Here are (actual) data (we have anonymized the school) on how many of the parents in each of its graduating classes have donated to the school's fund drive. The school is K-12. These are from the same year (2009) so the 2009 class means current 12-th graders, 2010 means current 11th graders, and so on.

   (a) Do a Chi-square analysis (be careful to set up the table correctly) to find out whether there is any sign of parent burnout – are donations equally likely in all grades?

   (b) Is this to be done as one-tailed or two-tailed? Why? (Does a low Chi-square mean a departure from the expected proportions?)

   (c) Think of some way to lump parts of the table to make the test focus more on the question at issue, and not waste effort on detecting whether there are differences that do not represent a long term trend. Carry it out and describe the results. Note that you can sum column 2 of rows 1 to 8 of a table by the R command sum(a[1:8,2])

Year	Donated	Total Parents
2009	35	51
2010	42	56
2011	39	70
2012	37	60
2013	38	53
2014	35	54
2015	32	53
2016	19	27
2017	22	31
2018	20	30
2019	17	32
2020	34	34
2021	28	32

(d) What is the effect on the Chi-square test, on average, if some of the parents have two (or more) children in more than one grade, and thus are listed as donating (or not) in both of those grades on the basis of the same donation or the same non-donation?

4. [**40 points**] **Wilcoxon Rank Sum Test**

(a) Generate two groups of observations, each with $n = 5$ form a standard normal distribution. Perform a t-test for a difference in location for these two groups, and record the p-value. Now perform a Wilcoxon Rank-Sum test for a difference in location for these two groups and record the p-value.

(b) Repeat this process 1,000 times. At a level of $\alpha = 0.05$, how often do you reject the null hypothesis for each test? What can you conclude about the type 1 error rate of each test, given these results? Which test is more conservative and why?

(c) Repeat parts a and b, but now let one group have a mean of 1 instead of 0. How often do you reject the null hypothesis for each test? What can you conclude about the power of each test for this sample and effect size? Try using different sample and effect sizes and summarize your results either in words or with one plot.

(d) Repeat parts a and b, but now let one group have a variance of 3^2. How often do you reject the null hypothesis for each test? Which test seems more robust to unequal variances? How do your results chance if you use var.equal = TRUE and var.equal = FALSE in your t-test?

5. **[30 points] P-values**

   Imagine you have collected two groups of observations, one with $n = 3$ and one with $n = 4$ from some populations with unknown distributions. You want to test for a difference in location of these two distributions, so you perform a Wilcoxon Rank Sum test. Assume that there are no ties in your observations.

   (a) What is the minimum p-value you could possibly obtain with this test? What is the maximum p-value you could possibly obtain with this test?

   (b) We learned that ties are given midrank values when the WRS test statistic is calculated. Ties in the WRST are handled the same way. What happens to the test statistic distribution when a tie is present? We learned that we can approximate the WRS test statistic with a normal distribution. If we used this approximation for a WRST with ties present, how should we adjust the normal approximation of the test statistic distribution to account for the ties?

6. **[30 points] Wilcoxon Rank Sum Test Function**

   (a) Write a function in R to calculate T_{WRS} the Wilcoxon Rank Sum test statistic. Your function should take as input two groups of observations and return the test statistic. Your function should NOT make use of or contain code from the built-in wilcox.test() function.

   (b) Add a parameter to your function that enables the function to return either the value of T_{WRS} or the p-value for a two-sided test for equality of medians (using the normal approximation is fine).