GENOME 560, Spring 2017
Problem Set #3

(Due May 8th 8:59 am)

Please include both your R code and your answers.

1. **[50 points] Bootstrap**

   In homework 2, you were investigating the *de novo* mutation rate of *R. Waterstonii.* In a given generation, you observed the following mutation counts in 12 different isolates:

Isolate	# of mutations	Isolate	# of mutations
1	12	7	18
2	15	8	12
3	10	9	7
4	4	10	11
5	6	11	5
6	15	12	14

   You previously assumed that these data came from a Poisson distribution, and estimated λ from the sample. Now, your collaborator asks you for the variance of your estimate of λ - i.e. how precise is your estimate?

   (a) Obtain 1,000 bootstrap samples from your data and recompute $\hat{\lambda}$ for these bootstrap examples.

   (b) Compute the variance of λ.

   (c) Generate a 95% confidence interval for $\hat{\lambda}$ using the normal approximation method.

   (d) Generate a 95% confidence interval for $\hat{\lambda}$ using the percentile method.

   (e) Compare the results from c and d. Which do you prefer, and why?

2. **[50 points] Choose Your Own Adventure!**

   You are studying the effect of a new enzyme on the number of PCR cycles required for a reaction. You split each of your 16 samples into two and perform the reactions with the old and new enzymes. The following table summarizes your data on the number of cycles required for each pair of reactions using the old and new enzymes:

Old	New
7	6
5	8
7	8
7	2
10	9
12	9
5	7
13	7
11	4
9	7
9	8
8	7
8	7
13	11
6	8
10	6

You want to make a conclusion about whether the new enzyme requires fewer cycles than the old enzyme. Analyze this data using two different techniques or tests (we have covered many). For each approach, make sure you answer the following questions:

(a) Briefly describe (in words) your strategy. Why did you choose this approach?

(b) What are the null and alterative hypotheses for this test?

(c) What assumptions does your approach make? Are they all satisfied? How do you know?

(d) What is the p-value from your analysis?

(e) What do you conclude (be precise about the conclusion)?

3. **[50 points] Probability Theory Review**

In class, we discussed conditional probability. The conditional probability distribution over A given B can be computed as:

$$P(A|B) = \frac{P(A,B)}{P(B)},$$

and the probability conditional distribution over B given A can be computed as:

$$P(B|A) = \frac{P(A,B)}{P(A)}.$$

Bayes' rule defines a relationship between P(A|B) and P(B|A) based on the two conditional probabilities above:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Based on the definition of Bayes' rule given above, answer the following questions:

(a) [**20 points**] After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e. the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

(b) [**30 points**] It is quite often useful to consider the effect of some specific propositions in the context of some general background evidence that remains fixed, rather than in the complete absence of information.

- Prove the conditionalized version of the general product rule:

$$P(A, B|E) = P(A|B, E)P(B|E)$$

- Prove the conditionalized version of Bayes' rule:

$$P(A|B, E) = \frac{P(B|A, E)P(A|E)}{P(B|E)}$$

4. [**50 points**] **Maximum Likelihood Estimation**

In lecture last Thursday, we learned a general method to compute the maximum likelihood estimate (MLE). Now, we are interested in applying this method to obtaining the MLE of the number of spontaneous deleterious mutations in humans. Say that you developed a potential Nobel prize worthy method of counting spontaneous deleterious mutations. You applied it to ten individuals and obtained the following data:

Our goal is to estimate the parameter for the number of spontaneous deleterious mutations based on these data from 10 samples.

(a) [**15 points**] Write out the likelihood function for this data. (Hint: we discussed in lecture #3 the probability distribution that this type of data follows.)

Individual	Number of deleterious mutations
1	4
2	3
3	0
4	1
5	9
6	2
7	3
8	8
9	0
10	6

(b) [**20 points**] Calculate the MLE of the average number of spontaneous deleterious mutations per individual from the data above. You can do this in any of the three ways that we discussed in class (graphically, numerically or calculus).

(c) [**15 points**] Plot the log-likelihood of the data as a function of the parameter.

5. [**50 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?