Project 3 Theory Addendum

Group 2: Kasen Teoh, Chung En Pan, Nathan Fallahi, Parsa Ganjooi, Eamon Jarrett-Mann

Theory

Hypothesis Testing

Hypothesis tests typically have two hypotheses: a null and an alternative hypothesis. The null hypothesis answers yes to the topic in question while the alternative hypothesis says the opposite. Usually, with hypothesis tests, we test at the 95% confidence level and an α of 5%, although it is not required to use these levels.

In the event that the calculated p-value for the test statistic is less than the α or if the test statistic is greater than the critical value of the χ^2 distribution with a specified degree of freedom, then we are able to reject the null hypothesis in favor of the alternative hypothesis because we have sufficient evidence that the alternative hypothesis is true. On the other hand, when the p-value is greater than the α level or the test statistic is less than the critical value, we fail to reject the null hypothesis because we do not have sufficient evidence to prove that the alternative hypothesis is correct.

When rejecting or failing to reject the null hypothesis, there are errors involved with the decision process. A type I error is described as when we have rejected the null hypothesis when we did not have sufficient evidence to reject it. The probability of a type I error occurring is equal to that of the α level. A type II error is described as when we have failed to reject the null hypothesis when the alternative hypothesis was true. The probability of a type II error occurring is equal to β. Lastly, power describes the ability to correctly reject the null hypothesis when the alternative hypothesis is actually true and it is calculated as 1 - β (the probability of a type II error occurring). Typically, as α grows larger, β grows smaller and power increases as well, and vice versa.

Poisson process

The Poisson process is a model for random occurrences such as waiting time for a bus, arrival time for customers waiting in the queue. This process is distributed with no regularity and follow with the characteristic that:
- rate (λ) at each hit does not change with each location
- the hit in each region are independent
- one hit can only be in exactly one location

Detecting the unusual scatter of the palindromes, the poisson process becomes very useful as the process is a model for uniform random scatter, which represents a good reference model for the DNA strand, as DNA strands are composed of A,C,T, and G. DNA strand can be treated a line that the palindromes are randomly scatter across the line, the location of each palindromes are unique with no overlapped region and the hits of the palindromes in certain interval are independent to other interval. These all align with characteristics for the poisson process. With the properties of the Poisson process, we can check whether Poisson process is a reference model for the DNA strand data.

Chi-Square Test

A chi-squared test is used to determine whether a data comes from a specific distribution in categorical variables. This is used to compare the observed frequency versus the expected frequency to check whether the data is statistically significant. The null hypothesis is that the data does follow the specific distribution while the alternative hypothesis is that the data does not follow the specific distribution. We are able reject the null hypothesis if the observed value of the test statistic is larger than the 95% quantile or is the p-value of the test-statistic is greater than the specified alpha level. The data distributions that were tested against the data were uniform distributions, poisson distributions, exponential, and gamma distributions. The formula for the test statistic is

$$\sum_{i=1}^{m} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

We used the chi square test in question two to compare the observed location and spacing to the theoretical or expected location. When comparing the test statistic, we found the 95 percentile of the chi-square distribution using a degree of freedom of the number of bins minus the number of parameters estimated minus one $\mathrm{df} = m - k - 1$ where m is the number of bins and k is the number of estimated parameters.

Choice of number of Bins

Typically we want the choice of r to be such that $n \cdot p_j \geq 5$ for all j's where p_j is the probability of a specific data point ending up in a particular bin. Another way that we are able estimate r is by setting it equal to $2 \cdot n^{\frac{2}{5}}$. However, when using this result, it may lead to a slightly large choice of r; hence, when using this estimation of r, it depends on a tuning parameter called bandwidth choice. To show that our choice of r is sufficient, performing a sensitivity analysis on the r should not result in significant changes in our results.

Poisson Distribution

$$H_0 : \text{Counts of the palindromes follow a Poisson Distribution}$$

$$H_A : \text{Counts of the palindromes do not follow a Poisson Distribution}$$

The poisson distribution is used because it closely resembles the data that we are given to work with. The poisson distribution quantifies the counts of hits in fixed intervals of the poisson point process. That is, it is a way of estimating what the occurrence count of an event will be in some given interval. The poisson distribution's probability mass function is

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

We used chi-square testing to determine how well the observed palindrome counts within arbitrarily chosen intervals can be modeled by using a Poisson distribution. These are used to determine whether there are unusual clusters of palindromes, relative to what one would expect if it was determined by random chance alone. We chose to use a wide variety of intervals to ensure the cluster findings are not some strange result of the intervals chosen.

Typically the parameter λ is unknown and utilize the maximum likelihood estimator method, where we essentially take the average.

$$\hat{\lambda} = \frac{\#\text{ of palindromes}}{\#\text{ of intervals of equal length}}$$

The test statistic is:

$$T = \sum_{j=1}^{r} \frac{(\tilde{V_j} - n\tilde{p_j})^2}{n\tilde{p_j}}$$

Where $\tilde{V_j}$ is the number of intervals with j palindromes, and $n \cdot \tilde{p_j}$ is the total number of intervals times the probability of that a particular intervals has x palindromes.

We reject the null hypothesis if the test statistic T is greater than 95% percentile of the χ^2 distribution with the degrees of freedom.

The p-value:

$$P\left[\text{Theoretical distribution of T under H}_0 > \text{the observed value of the test statistic}\right] = \text{p-value}$$

For the Chi-Square goodness of fit test against a Poisson distribution, we had tried different interval lengths to see whether our results would change. When using a smaller interval length of around 80, we calculated an interval length of around 13.8 and a p-value of 0.008. Hence, we concluded that with a smaller interval length, we would reject the null hypothesis, of the counts of palindromes following a Poisson Distribution.

When we had tested against a larger interval of around 5500, we calculated a large test statistic of 290.637 with an extremely small p-value. Hence, we a larger interval length, we fail to reject the null hypothesis that the counts of palindromes does follow a Poisson Distribution.

By testing different interval lengths, this allows us to see that with a larger interval length, more palindromes are captured in the interval and is more similar to a Poisson distribution. With a smaller interval length, less palindromes are captured in each interval and and the distribution of the counts of palindromes is possibly skewed. Hence, when testing against a Poisson Distribution, we used use a larger interval length.

Uniform Distribution

$$H_0 : \text{Locations of the palindromes follow a Uniform Distribution}$$

$$H_A : \text{Locations of the palindromes do not follow a Uniform Distribution}$$

Analyzing the locations of the palindromes, we had split the data up into equal length intervals. Then, because the number of palindromes in each interval is known, we are able to test the data against a uniform distribution. Hence, the expected counts of each interval is the total number of palindromes divided by the number of intervals

$$E\left[V_j\right] = \frac{\text{total number of palindromes}}{\text{the number of intervals}}$$

and because we did not need to estimate any parameters, the degree of freedom is the number of intervals minus one, so

$$H_0 : T \sim \chi^2_{\text{number of equal length intervals}-1}$$

We had chosen the number of bins to be 20 when testing for a uniform distribution because more than 20 would have resulted in the data being skewed, i.e many of the bins would have been empty or contain very little palindromes. Using 20 bins, we expected to get around 14.8 palindromes in each interval.

Once again, we compare both the test statistic with the 95th percentile of the chi-square distribution and the p-value with the alpha level of 0.05 to determine whether we reject or fail to reject the null hypothesis.

We had calculated a test statistic of around 22.1 with a degree of freedom of 19 because we did not estimate any parameters and a p-value of 0.279. Hence, we had failed to reject the null hypothesis, i.e the location of palindromes does follow a uniform distribution.

Exponential and Gamma Distribution

When analyzing the spacings between consecutive, consecutive pairs, and consecutive triple palindromes, we use the chi-square test to test again an exponential and gamma distribution because we are analyzing the distance until the next palindrome. In other words, we are measuring the time until the next hit, where the time is the distance and the hit is the location of a palindrome.

$$P(\text{distiance between the first and second hits} > \text{t}) = P(\text{no palindromes in an interval of length t}) = e^{-\lambda \cdot t}$$

When testing against exponential and gamma distribution, we estimated the parameter of λ, hence, the degree of freedom is the number of bins minus two.

When we are analyzing the distance between consecutive pairs and triplets of palindromes, we use a goodness of fit test against a Gamma distribution with parameters of 2 or 3 (depending on whether it is pairs or triplets) and lambda as the second parameter.

Consecutive Palindromes:
H_0 : Distance between consecutive palindromes $\sim$ Exponential Distribution(λ)

Pairs of Palindromes:
H_0 : Distance between consecutive pairs of palindromes $\sim$ Gamma Distribution$(2, \lambda)$

Triplets of Palindromes:
H_0 : Distance between consecutive triplets of palindromes $\sim$ Gamma Distribution$(3, \lambda)$

When we had carried out the chi-square test for exponential and gamma distributions, we had used a bin length of 50 and counted the number of consecutive pairs and triplets in each bin. For these three tests, we had estimated the parameter of $\lambda = \frac{1}{\bar{x}}$, and hence our degree of freedom was 48. In all three tests, we had calculated a test statistic that was consistently greater than that of the 95% percentile of the χ^2 distribution with degrees of freedom 48. Hence, the locations of consecutive, consecutive pairs, and consecutive triplet palindromes do not follow an exponential, gamma, and gamma distribution, respectively.

Scan Statistic

The scan statistic is used to find the unusual cluster. An example of an unusual cluster of palindromes is when a large number of palindromes are really close together. Before finding the unusual cluster we have to decide what an usual cluster will look like first. To do this we have to decide on the maximum number of palindromes per different intervals. It is important to choose the correct length so we are able to identify the palindromes and the number of them in each cluster. The first step in this process is to see what the distribution of T_m is under the null hypothesis. In our case, we want to calculate the probability that the maximum number of palindromes in an interval, given m intervals, is greater than or equal to the max value we observe in our data a.

T_m = the maximum number of palindromes over m intervals
$H_0 : T_m \sim$ Poisson Distribution(λ)

Where under the null hypothesis, T_m has a distribution of a maximum of independent Poisson(λ) random variables.

$$P(T_m > a) = P\left[\text{maximum of independent Poisson}(\lambda) \le a\right]$$

$$= P\left[\text{all of independent Poisson}(\lambda) \le a\right]$$

$$= P[\text{first interval has Poisson}(\lambda) \le a]^m$$

$$= P([\text{Poisson}(\lambda) = 0] + [\text{Poisson}(\lambda) = 1] + \cdots + [\text{Poisson}(\lambda) = a])^m$$

$$= e^{-\lambda \cdot m}\left[1+\hat{\lambda}+\frac{\hat{\lambda}^{2}}{2!}+\cdots+\frac{\hat{\lambda}^{a}}{a!}\right]$$

$$\text{p-value} = \mathrm{P}(T_m > \text{observed test statistic}) = 1 - \left[P\left(T_m < \text{observed test statistic}\right)\right]^m$$

$$= 1 - \left(e^{-\widehat{\lambda}}\left[1+\widehat{\lambda}+\frac{\widehat{\lambda}^{2}}{2!}+\ldots+\frac{\widehat{\lambda}^{\text{observed test statistic}}}{(\text{observed test statistic})!}\right]\right)^m$$

Where $\hat{\lambda}$ is the Maximum likelihood estimator of parameter λ of Poisson Distribution

Histograms

- Histograms are estimators of the probability density function for a given distribution. Histograms consist of bins that count the number of observations within each interval. The number of bins must be predetermined.

  $$\hat{f}_x(a) = \frac{1}{n}(\# \text{ of observations within} B_l) \times \left(\frac{1}{\text{length of bin} B_l}\right)$$

  Where a is a specific location in the distribution and B_l is a bin that contains a
  - Histograms have lower bias the more number of bins are used.
  - Histograms provide useful information about the spread of the data and can be used to identify potential outliers, whether the underlying distribution is unimodal/bimodal or symmetric/asymmetric, and what proportion of the distribution lies above or below a certain value.