Patterns in DNA

Joshua Castro, Kevin Elkin, Zunlin Lin, Sneheil Saxena, Yuka Sukamaki, Tony Wu

I.INTRODUCTION

Human Cytomegalovirus (CMV) is a common virus that infects people of all ages. More than half of adults by the age of 40 have been infected with CMV. (cdc.gov) a potentially life-threatening disease - especially for people who have a deficient immune system. In order to suppress the harmful effects of this disease, scientists and researchers are examining the way the virus replicates; they are specifically concerned with areas in the DNA that contain the genetic information encoding its replication. From this, scientists hope to be able to make better advancements in medicine to help counter and potentially eradicate the disease entirely.

Although scientists are unsure of how the virus is spread, it is suspected to be transmitted through bodily fluids and sexual contact. After the virus is contracted, it lies dormant in the lymphocytes permanently until it is reactivated. Once active, CMV produces symptoms consisting of fevers, night sweats, tiredness and uneasiness, sore throats, swollen glands, joint and muscle pain, low appetite and weight loss.

Herpes Simplex and Epstein-Barr are two viruses from the same family as CMV (the herpes family) and have the origin of their replication marked by a complementary palindrome. The Herpes Simplex virus is marked by a palindrome of 144 letters while the Epstein-Barr virus is marked by a series of several shorter palindromes clustered near the origin of replication.

DNA has a double helical structure composed of two long chains of nucleotides. A single nucleotide has three parts: a sugar, a phosphate and a base. Any virus or cell like structure contains DNA with base pairing rules consisting of the following: A with T - the purine adenine (A) always pairs with the pyrimidine thymine (T), C with G - the pyrimidine cytosine (C) always pairs with the purine guanine (G). The human genome has nearly 3 billion base pairs with 99% of the base pairs being the same for all individuals. Using this knowledge, we can determine how we can find clusters of palindromes and can thus conclude with some degree of certainty whether a cluster is by chance or is a replication site for a virus such as CMV.

II.DATA

A. Source

Our dataset, from Chee et al. (1990), contains 229,354 nucleotides, which were analyzed by Leung M et al. (1991) to extract complementary palindrome locations.

B. Data Summary

We have N = 296 palindromes with their reported location. On average, we see that the average interval between each palindrome is 775.5119 nucleotides, so we see that we can sufficiently analyze clusters of palindromes within a given interval. The locations are based on the starting index of the palindrome. Palindromes shorter than 10 nucleotides were ignored and not reported. No matter the length of the intervals, there always seem to be an outlier of intervals containing a higher number of palindromes. Also, we can observe that the intervals of the random hits do not display such outliers thus, strengthening the suggestion that a deeper investigation into the strand should expose peculiarities within the CMV strand.

III.BACKGROUND

CMV, Cytomegalovirus, is a member of the herpes virus family, a virus commonly known to infect those of all ages. Once a person is infected with CMV, the virus is carried dormantly until reactivated or until the carrier is reinfected with a different strain. Common symptoms of CMV include fever, sore throat, fatigue, and swollen glands. It is worth noting that those who are infected with CMV and have a weakened immune system may carry more urgent symptoms affecting the eyes, lungs, liver and more. Also, infants born with CMV may experience brain, liver, spleen, or growth problems, but most commonly hearing loss. [CDC]

Occurrence of CMV geographically fluctuates from 30% to 80%, though it is normal for around 10-15% of children to contract this virus before 5 years of age. As stated, upon contracting this virus, the infection tends to remain restricted until becoming a young adult. As a young adult, CMV carriers may present symptoms similar to mononucleosis, as the infection levels increase. Should this virus remain within developing children, this may pose as a serious problem for those whose immune system may be in a depressed state.

The CMV DNA molecule contains 229,354 complementary pairs of letters or base pairs. In comparison, human DNA has more than 3 billions base pairs. By locating CMV, we may assist virologists develop a vaccine to effectively combat the virus. It has been found that entropic perimetry may be an effective and inexpensive screening test or CMV retinitis in hospitals [CDC]. However, we would like to investigate whether or not we can distinguish any

sort of clusters of palindromes in the DNA strands in order to find some sort of potential replication site.

We have also been given a dataset regarding human cytomegalovirus epidemiology and its relationship to diseases such as tuberculosis and cardiovascular disease risk factors in a rural Ugandan cohort. We want to use this data to find whether or not we can find some relationship between a presence of CMV and a presence of HIV. We would like to investigate how the presence of these two infections may reflect in the health of the person, as it has been previously found that infants with CMV infection within the first on top of their HIV-1 infection at birth have higher rates of disease progression and central nervous system disease than those with HIV-1 infections alone, from a previous study [Kovac]. With that being founded, we would like to build on top of these findings, to see how the distribution of the BMI from m the given dataset is affected by the presence of CMV and HIV, in hopes to find some relationship between the general well being of a person and the presence of these diseases.

IV.HYPOTHESIS

1. How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?

2. Additional hypothesis: Given the extra dataset, can we find some correlation between the presence of CMV and HIV testing results? [Stockdale]

V. INVESTIGATION

1) Random Scatter

Fig. 1 Distribution of the actual number of palindromes contained in the data at each location (bin size 4000)

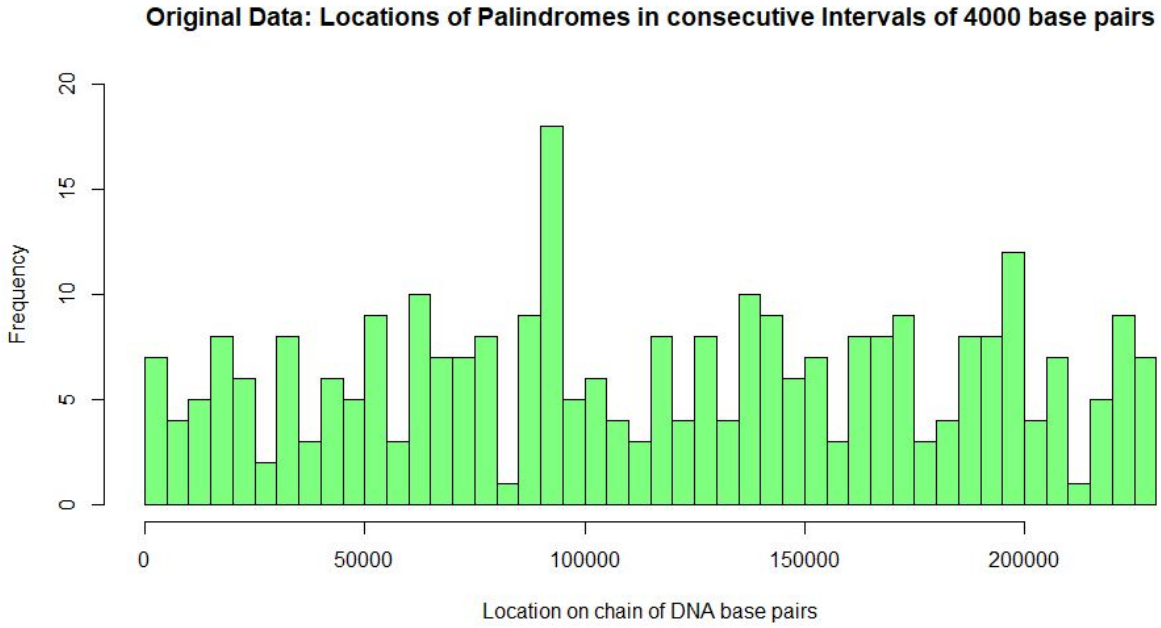
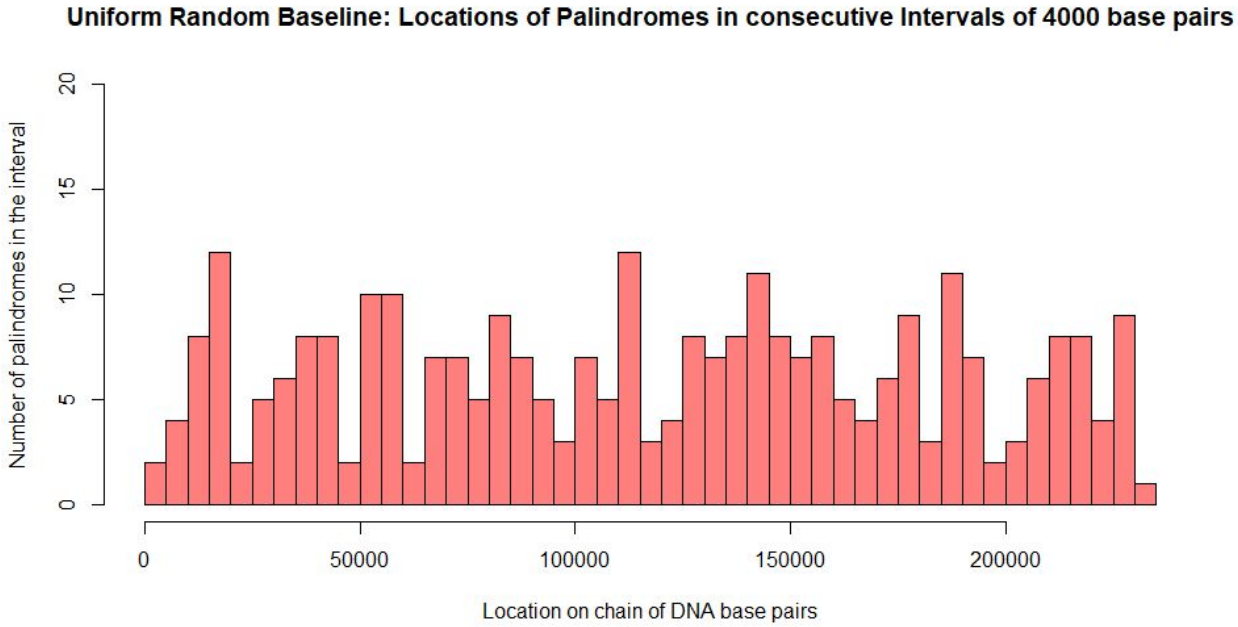


Fig. 2 Distribution of the uniform random baseline number of palindromes contained in the data at each location (bin size 4000)



We begin by examining the complementary palindrome clusters by investigating how they compare to the baseline distribution (Fig 2 - a uniform random distribution) which will serve as our null hypothesis. The primary focus of this is to see if there is any significant difference between the original data (Fig 1) and the baseline distribution; if such a difference exists, then more conclusive tests and investigations will be conducted to get a more definitive reason as to why and whether we should reject our null hypothesis.

296 palindromes were randomly scattered along a DNA sequence of 229,354 base pairs and a uniform random distribution was created in Fig 2 which shows the number of hits from a uniform random distribution. Locations are grouped into bins with intervals of 4,000 nucleotides. Fig 1 shows the original 296 locations of the palindromes. Upon examination of Fig 1 and Fig 2, it is difficult to notice any striking differences between the original and simulated data. We notice in Fig 1 there is one interval around the nucleotide 90,000 with a large number of palindromes compared to the rest of the data in the distribution for both histograms. This could be a potential outlier but more conclusive testing is required to determine whether there is any significance in this bin.

Fig. 3 A scatter plot mapping the distance between consecutive base pair palindromes and their distance/location from one another (actual data)

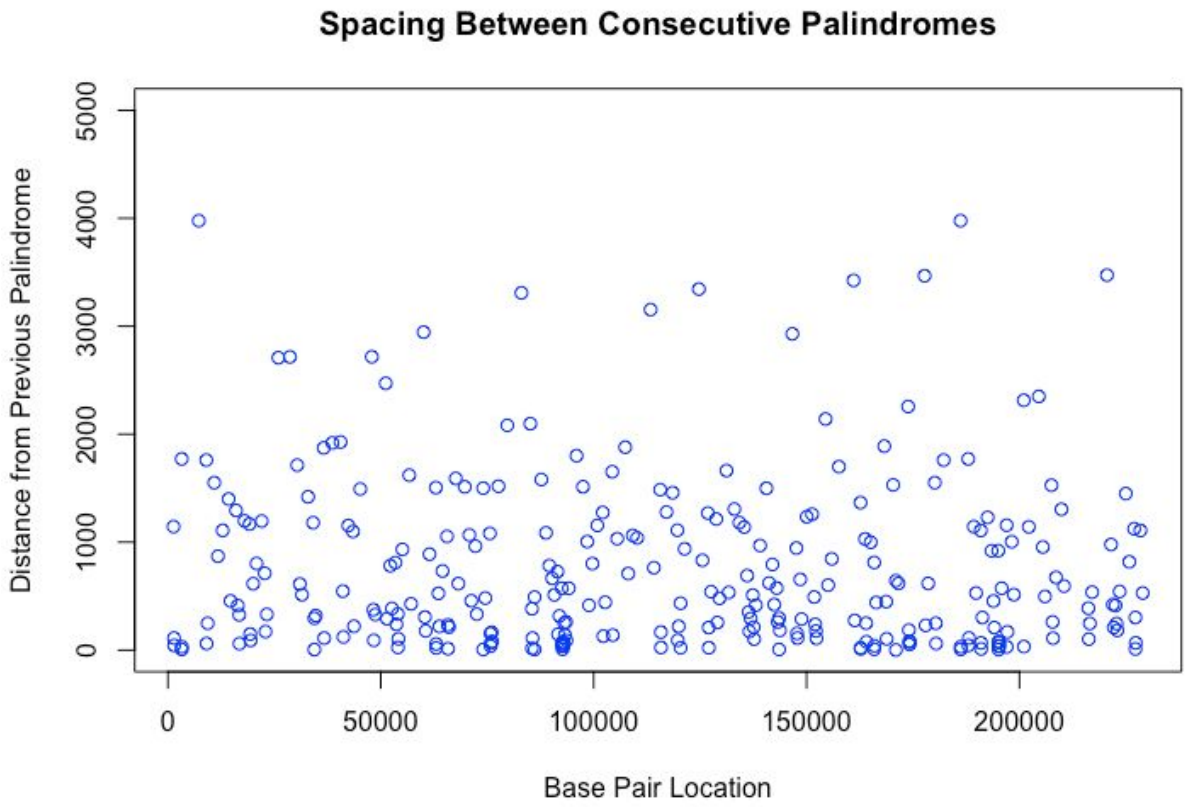
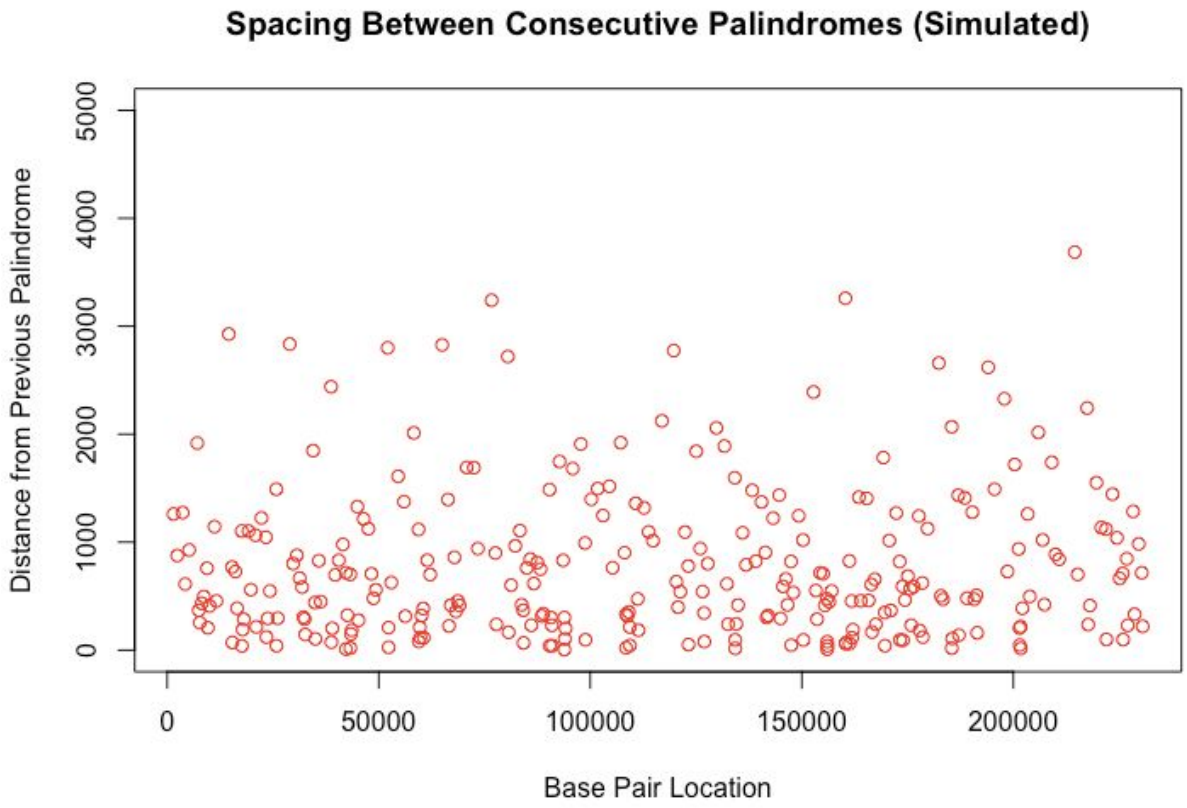


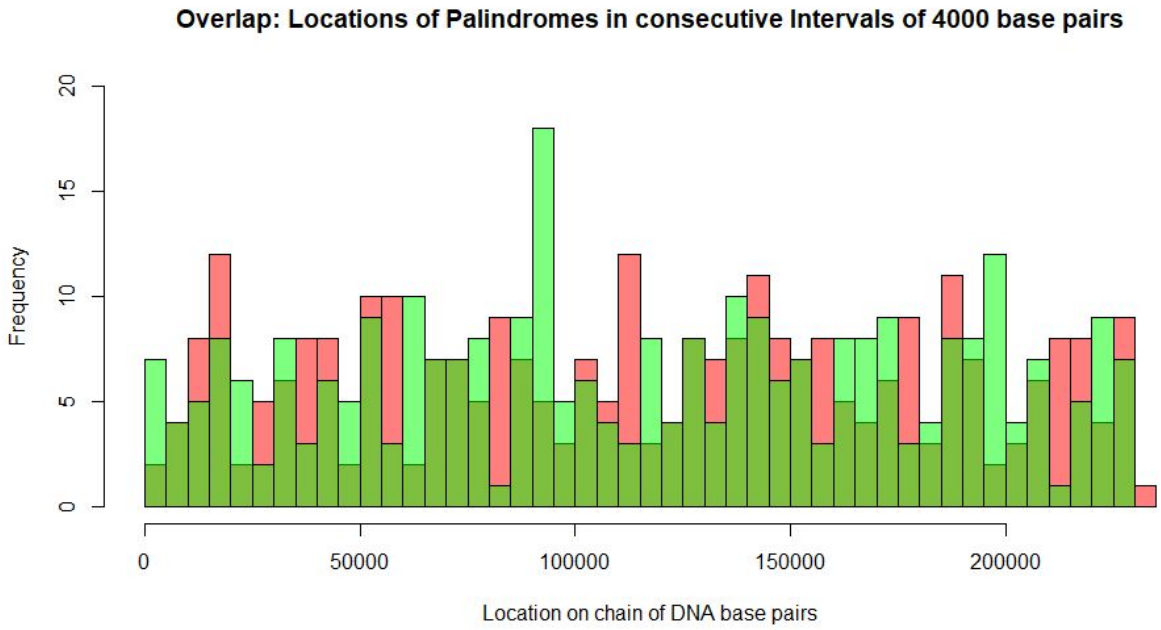
Fig. 4 A scatter plot mapping the distance between consecutive base pair palindromes and their distance/location from one another (simulated)



We can see from Fig. 3 and Fig. 4 that we are unable to identify any sort of pattern or outlier high density areas. This supports the findings stated previously, as we are generally cannot find any patterns from the original data or the random scatter data. Hence, in our following analyses, we will investigate any possible patterns regarding the spacing between consecutive palindromes to answer our hypothesis.

2) Locations and Spacings

Fig. 5 Distribution of the overlapping histograms of the actual number of palindromes contained in the data at each location (Fig 1) and of the uniform random baseline number of palindromes contained in the data at each location (Fig 2) (bin size 4000)



The CMV DNA strand is segmented and partitioned into intervals of equal length (4,000) to determine the location of palindromes and palindrome clusterings. As mentioned before, the location of palindromes are expected to follow a uniform distribution; conducting an χ^2 Goodness of Fit test can help in determining whether or not the observed frequency of palindromes differs from the theoretical frequency. In Fig 3 we can see the overlap between the expected and observed locations of palindromes (Fig 1 and Fig 2) and a residual plot in Fig 4. Fig 3 and Fig 4 both show that there is an unusual clustering occuring around the nucleotide at 90,000 and 194,000 base pairs of the DNA sequence.

The CMV DNA strand is segmented and split up into 58 equal intervals of 4,000 base pairs; the number of palindromes is then counted and added in each interval on the histogram plot(s). As of now, the unusual clustering occurs around interval 90,000 and 94,000 where 15+ palindromes. Due to this following a uniform distribution, we would expect the frequency for

each bin of interval 4,000 to have 296/58 or 5.10 palindromes per interval which means χ^2 = 38.78716, thus for $\alpha = 0.05$, the p-value is 0.1902511.

Fig. 6 QQ Plot depicting the comparison of Original Data and Baseline Uniform

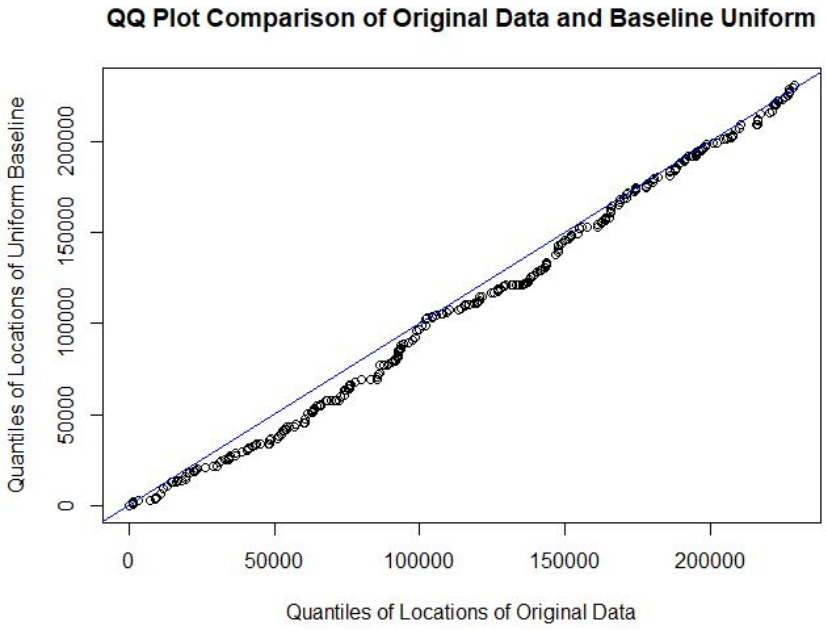


Fig 6 depicts a QQ plot showing the comparison of original data and the random baseline uniform. We can see that the distributions tends to be below the normal line, suggesting that the quantiles of locations of the original data occur before the quantiles of locations of uniform data. This suggests that the distribution of the location of the original data may follow the uniform distribution, and this backs the graphical analysis provided by Fig. 5, as we can see that the two sets of data both follow a somewhat uniform distribution, neglecting certain outliers.

Also, we see in Fig. 7 that the range of values within [-3,3], thus we do not have reason to believe there is a statistically significant difference of locations of nucleotides. We expect to see that the original data set follows a uniform distribution from this comparison.

Fig. 7 Standard normal residuals from the observed and expected location of palindromes with interval length 4000.

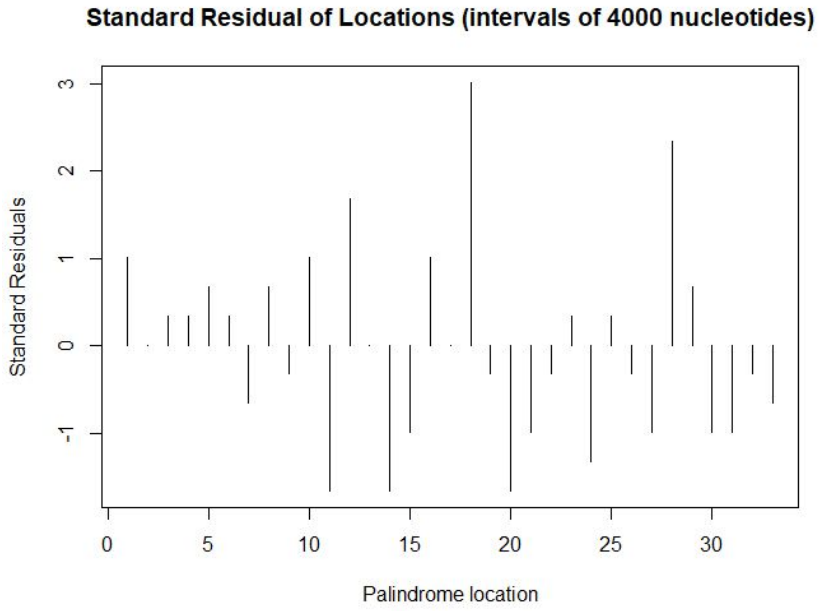
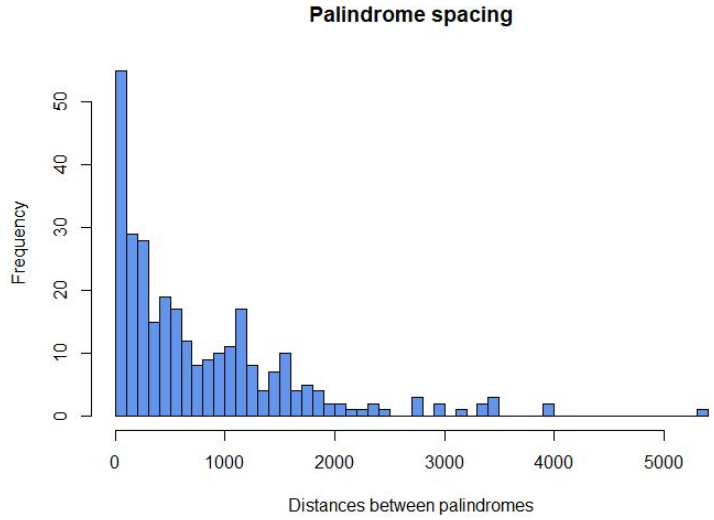


Fig. 8 The distribution of the difference in spacing between adjacent palindromes.

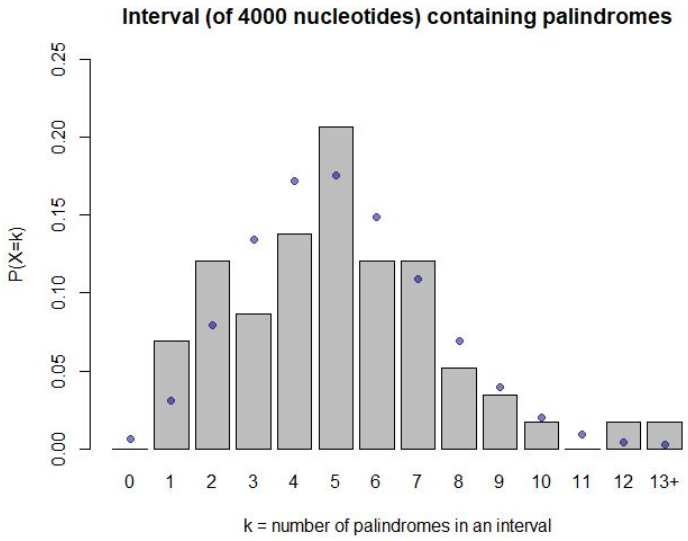


3) Counts

We partitioned the CMV DNA strand into nucleotide intervals of 4,000 and observed the number of palindromes in each interval. We expect the number of palindromes to be uniformly scattered

throughout the DNA strand, therefore we expect the probability of encountering k palindromes P(X = k) to be modeled as a Poisson distribution.

Fig. 9 The probability distribution of an interval of 4000 nucleotides containing k palindromes, overlaid with the estimated Poisson Distribution, where $\widehat{\lambda} = 5.1$.



Using a Maximum Likelihood Estimate, we see that parameter λ of Poisson distribution is best estimated by $\widehat{\lambda}$ which is equivalent to the mean of the observations. We see in Fig. 9 the observed probabilities of the number of palindromes in an interval, compared to the expected probability, looks like it fits the model well.

To assess the fit of our distribution, we will visualize the standard residuals of the counts, modeled by $\frac{N_i - \mu_i}{\sqrt{\mu_i}}$, where N_i is the observed count and μ_i is the expected count for each interval i. Residuals that are outside of the range [-3, 3] indicate a poor model fit. From Fig. 10, we see that there are no noticeable outliers and are within the range.

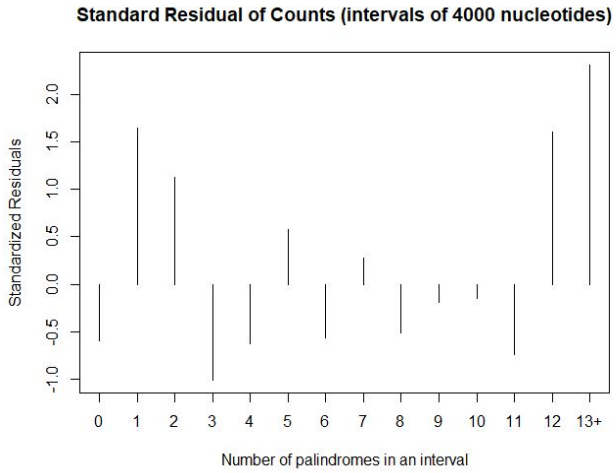
Table I. Observed number of palindromes in each interval vs expected number of palindrome in each interval (interval length of 4000)

Palindrome count	Observed Counts	Expected Counts
0	0	0.3524
1	4	1.7984
2	7	4.5891

3	5	7.8067
4	8	9.9603
5	12	10.1664
6	7	8.6473
7	7	6.3044
8	3	4.0218
9	2	2.2805
10	1	1.1639
11	0	0.5400
12	1	0.2296
13+	1	0.1392

Therefore, we use a χ^2 Goodness of Fit test to determine if the Poisson Distribution well models the locations model. Our null hypothesis H_0 states that there is no difference in the number of palindromes in any interval. We see that our test statistic $\chi^2 = 15.19$, with degrees of freedom = 12, compared to the an $\alpha = 0.05$ with the same degrees of freedom $\chi^2 = 21.03$. So, we see our p-value = 0.231, thus we fail to reject the null hypothesis, despite any bias in the intervals in Fig. 9. So, we can assume the observed data is modeled by the Poisson Distribution.

Fig. 10 Standard normal residuals from the observed and expected count of palindromes in each interval.



4) The biggest cluster

Table II: A table of resulting p-values and cluster sizes based on different interval lengths

Interval length of nucleotides	Subintervals	$\widehat{\lambda}$ - estimator	Max cluster size	P-value
2500	92	3.22	13	8.490572e-86
4000	58	5.10	16	3.077013e-53
5500	42	7.05	15	0.7553284
7000	33	8.97	18	0.1460012

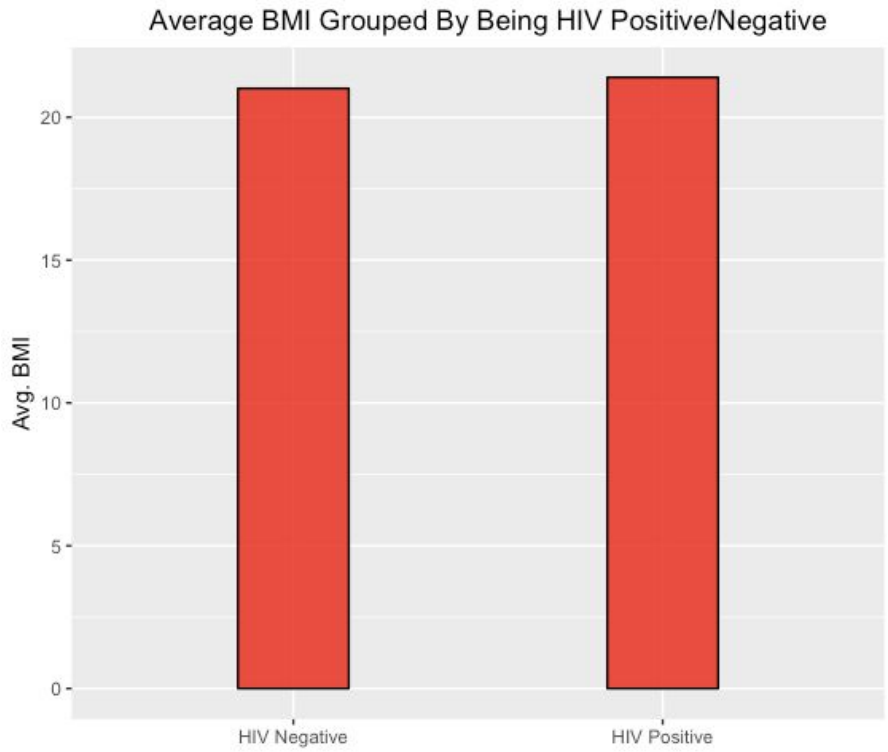
We saw that the number of palindromes in an interval are independent observations from a Poisson distribution under the Poisson process model. Therefore we see that the interval with the most palindromes is the maximum of these independent random variables. The probability that *k* is the largest cluster size can be approximated with a sample rate $\widehat{\lambda}$. Therefore, we will use hypothesis testing to determine if the cluster size is unusual, thus providing evidence that the interval with the most palindromes is a potential site of replication.

We see from Table II that some strands that are split into different interval lengths have different amount of subintervals and estimators. Thus, the cluster sizes will vary and we calculate the p-value that the largest cluster size is expected from the Poisson process. We see that with $\alpha = 0.05$, there is statistically significant evidence that the largest clusters of intervals 2500 and 4000 are larger than expected from the Poisson process.

Additional Hypothesis

We try to see if there is any correlation between being HIV Positive and the person's BMI.
The HIV virus attacks the body's immune system, and towards the later stages it can result in loss of appetite and consequently, severe weight loss. Based on that, we expected a person with the HIV virus to have a lesser BMI than someone without it on average.

Fig. 11 Barplot of average BMI of HIV positive and negative people



Average BMI for someone with HIV: 21.3986
Average BMI for someone without HIV: 21.0123

Based on the bar plot above, we can see that the average BMI for someone with HIV is actually higher than someone without HIV, contrary to our expectations.

VI.DISCUSSION & CONCLUSION

In Scenario 1, we found that we are unable to identify any sort of pattern from the given dataset, using the original dataset or a uniform baselines. Recall from Fig. 1 and 2 that the distributions given from the original data and the uniform base line do not indicate any sort of pattern, and this is reinforced by Fig. 3 and 4, displaying a random scatter given from both the original data and the uniform baseline. This compels us to use other forms of statistical analysis and graphical analysis to investigate our situation.

Then in Scenario 2, we wish to compare the locations of palindromes in the uniform baseline and the original dataset. We find that using the chi-squared test, there does not exist a statistical difference between observed and expected locations of palindromes. Further analysis with a Q-Q plot, we are able to see that the quantiles of the original data remain less than the quantiles of the uniform baseline, allowing us to continue with our analysis.

From Scenario 3, we found the number of palindromes within each interval. Using the maximum likelihood estimate, we get the expected counts of palindromes of each interval. We then utilized the chi-squared goodness of fit test to find that the number of palindromes within each interval follows a Poisson distribution. Despite biases indicating the data does not follow a Poisson distribution, we can find that is in fact false, as given in our analysis.

As for Scenario 4, we look for the max cluster size for different interval sizes. We found that the different subintervals result in different p-values, and based on the sub-interval, the p-value will change. Hence, we conclude that there is statistically significant evidence that the largest clusters of certain intervals are larger than expected than the Poisson process. This suggests the interval with the most palindromes is a potential site of replication within the DNA.

Therefore, we find that we are able to use the Poisson process to locate potential sites of replication within the DNA.

It is worth noting that we are limited by the data, as the length of the palindromes remain unknown. The dataset also contains a few instances of overlap in palindromes, as we see palindromes separated by less than 10 base pairs. This affects our analysis of Scenarios 2 and 3. This study would have benefited from information regarding the characteristics of the palindromes.

VII.METHODS AND THEORIES

1. Poisson Distribution

In statistics and probability theory Poisson distribution is a discrete event with a specific random variable X that counts discrete events occurring in a given time interval, announced by mathematician Simeon Doni Poisson with probability theory in 1838 It is a probability distribution. For a discrete event, the Poisson distribution shows the probability of the number of occurrences within a given time, and the exponential distribution shows the probability of the occurrence period.

For the constant $\lambda > 0$, a random variable X that takes an integer equal to or larger than 0 as the value is;

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

When random variable X satisfy with above, the random variable X follows Poisson distribution of parameter λ.

P (X = k) corresponds to "the probability that an event occurring λ times on average on a given time occurs exactly k times (k is a nonnegative integer)". For example, if an event occurs on average every 2 minutes, the number of times an event occurs within 10 minutes is obtained using the Poisson distribution model with $\lambda = 5$.

Notation:

$E[X] = \lambda$

$V[X] = \lambda$

Characteristic: Poisson distribution has reproducibility. That is, when X and Y are independent random variables and follow the Poisson distribution with parameters λ and μ respectively, the sum X + Y of the random variables follows the Poisson distribution of the parameter $\lambda + \mu$.

Approximation: If λ is sufficiently large (eg $\lambda > 1000$), the normal distribution of mean λ, standard deviation $\sqrt{\lambda}$ is a very good approximation of this Poisson distribution. Approximately $\lambda > 10$, the normal distribution is an excellent approximation of this Poisson distribution only if

corrective correction to the continuous distribution has been made. For example, for P ($X \leq x$), if x is a nonnegative integer it can be replaced with P ($X \leq x + 0.5$).
Limit theorem: In the binomial distribution where the parameters are n and $p = n/\lambda$ when n is made close to infinity while keeping λ constant, the distribution approaches the Poisson distribution of the average λ. Therefore,

$$\lim_{\lambda=np,n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

is consisted. This is called Poisson's limit theorem. The name of this theorem derives from mathematician Simeon Doni Poisson who gave results in 1837 in the book "Recherches sur la probabilite des jugements". Note that the Poisson distribution has been derived for the first time as the limit of the binomial distribution among them.

1a. Poisson Process
We find that the counting process is a Poisson Process with rate λ with $\lambda>0$ if

1. N(0) = 0, meaning no events occur at time $t = 0$
2. The increments of the process are independent
    a. Say we have range of time t to $t+s$. Then with the increments of the process being independent, we can find that the number of events that occurred at time t is independent of the events that occurred within the specified range. This characteristic means that N(t) is independent of the increment N(t+s)-N(t).
3. The number of events within the interval of length t follows a Poisson distribution with an average of λt.
    a. It is equivalent to state the following:

$$P(N(t+s) - N(s) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \text{ for } n = 0, 1, 2, \ldots$$

    b. From this, we also find that:
        i. The Poisson process uses stationary increments
            1. This means the distribution of the number of events which occur within an interval of time within the Poisson process depends solely upon the interval length of the time
        ii. $E[N(t)] = \lambda t$

2. Goodness of Fit test
Goodness of Fit test
Goodness of Fit test is the test that if the sample data fits a certain distribution.
We first introduce discrete models: n observations are grouped into t classes.
Then, we use hypothesis test:

$$H_0: p_1 = p_{1_o},\ p_2 = p_{2_o},\ \ldots,\ p_t = p_{t_o}$$

$$H_1: p_i \neq p_{i_o} \text{ for at least one } i$$

which is equivalent to test if $(X_1, X_2, ..., X_t)$ comes from a multinomial population with probabilities $p_1 = p_{1_o}, p_2 = p_{2_o}, ..., p_t = p_{t_o}$

Example: we test $H_0 : p_1 = 0.3, p_2 = 0.5, p_3 = 0.2.$

So, in this case, $p_{1_o} = 0.3,\ p_{2_o} = 0.5,\ p_{3_o} = 0.2.$

Chi-Squared Test Statistics (known parameter)

Theorem: Let $r_1, r_2, ..., r_t$ be the set of possible outcomes (or ranges of outcomes) associated with each of n independent trials, where $P(r_i) = p_i, i = 1, 2, ..., t.$ Then, let

$X_i = number\ of\ times\ r_i\ occurs,\ i = 1, 2, ..., t.$

Then, The random variable

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i}$$

which is $X^2 test\ statistic$ has approximately a $X^2 distribution$ with t-1 degrees of freedom. For the approximation to be adequate, the t classes should be defined so that $np_i \geq 5$, for all i.

For t=2, $X^2 test$ using D is equivalent to perform a test using Z test statistic.

We can derive

$$\begin{aligned} D &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{[n - X_1 - n(1 - p_1)]^2}{n(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2(1 - p_1) + (-X_1 + np_1)^2 p_1}{np_1(1 - p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} \quad = \left[\frac{X_1 - E(X_1)}{\sqrt{\mathrm{Var}(X_1)}}\right]^2 \end{aligned}$$

Notation : If X^2 is large, then it is a poor fit.

If X^2 is small, then it is a good fit.

Decision rule: Let $k_1, k_2, ..., k_t$ be the observed frequencies for the outcomes $r_1, r_2, ..., r_t$, respectively, and let $np_{1_o}, np_{2_o}, ..., np_{t_o}$ be the corresponding expected frequencies based on the null hypothesis. At the α level of significance, H_0 is rejected if

$$d = \sum_{i=1}^{t} \frac{(k_i - np_{i_o})^2}{np_{i_o}} \geq \chi^2_{1-\alpha, t-1}$$

where $np_{i_o} \geq 5$ for all i.

Here, we summarize four steps of $X^2 test$ with known parameters.

Step1: State a null hypothesis H_0 and an alternative hypothesis H_1.

$$H_0: p_1 = p_{1_o}, p_2 = p_{2_o}, \ldots, p_t = p_{t_o}$$

$$H_1: p_i \neq p_{i_o} \text{ for at least one } i$$

Step 2: The data are divided into several classes. The test statistic

$$D = \sum_{i=1}^{t} \frac{(X_i - np_i)^2}{np_i}$$

Has approximately a X^2_{t-1} distribution if H_0 is true.

Step 3: Compare the observed test statistic with $X^2_{1-\alpha,t-1}$ or calculate the p-value.

Step 4: Conclusion.

Chi-Squared Test Statistics (unknown parameter)

Suppose that a random sample of n observations is taken from $f_Y(y)$ *or* $p_x(k)$, a pdf having s with unknown parameters. Let $r_1, r_2, ..., r_t$ be a set of mutually exclusive ranges or outcomes associated with each of the n observations. Let $\widehat{p}_i$ =estimated probability of r_i, i=1,2,...,t as calculated from $f_Y(y)$ *or* $p_x(k)$ after the pdfs, s with unknown parameters have been replaced by their maximum likelihood estimates. Let X_i denote the number of times that r_i occurs, i=1,2,...,r. Then, the random variable

$$D_1 = \sum_{i=1}^{t} \frac{(X_i - n\widehat{p}_i)^2}{n\widehat{p}_i}$$

Has approximately a X^2 distribution with t-1-s degrees of freedom. For the approximation to be Fully adequate, the r_i's should be defined so that $n\widehat{p}_i \geq 5$ for all i.

Decision rule: $k_1, k_2, ..., k_t$ are the observed frequencies of $r_1, r_2, ..., r_t$, respectively, and $n\widehat{p}_{1o}, n\widehat{p}_{2_o}, ..., n\widehat{p}_{t_o}$ are the corresponding estimated expected frequencies based on the null hypothesis. If,

$$d_1 = \sum_{i=1}^{t} \frac{(k_i - n\widehat{p}_{io})^2}{n\widehat{p}_{i_o}} \geq X^2_{1-\alpha,t-1-s}$$

H_0 should be rejected. (The r_i's should be defined so that $n\widehat{p}_{i_o} \geq 5$ for all i.)

Here we summarize five steps of $X^2 test$ with unknown parameters.

Step 1: Find the maximum likelihood estimator for s unknown parameters.

Step 2: : Plug the estimated parameters into the distribution to be tested, calculate the estimated probabilities $\widehat{p}_{1o}, \widehat{p}_{2_o}, ..., \widehat{p}_{t_o}$,then state the null hypothesis H_0 as

$$H_0 : p_1 = \widehat{p}_{1o}, p_2 = \widehat{p}_{2_o}, ..., p_t = \widehat{p}_{t_o}$$

Step 3:The test statistic

$$d_1 = \sum_{i=1}^{t} \frac{(k_i - n\hat{p}_{io})^2}{n\hat{p}_{io}} \geq X^2_{1-\alpha,t-1-s}$$

Has approximately a X^2_{t-1-s} distribution if H_0 is true.

Step 4: Compare the observed test statistic with $X^2_{1-\alpha,t-1-s}$.

Step 5: Write a conclusion.

P-value

If we let the significance level as 5%, then the result of the hypothesis test is rejecting the null hypothesis at a significance level of 5%. However, it would arise questions that "what happens when the significance level is 1% or 0.1%?" and "what is the limit?" So, P-value solves that question.

If P-value is smaller than a significance level, then we reject the null hypothesis H_0. In other words, the limit value of the significance level such that the null hypothesis is rejected is P-value.

It P-value is small, which means

① the probability of taking an extreme value is smaller than the statistic calculated from the data.

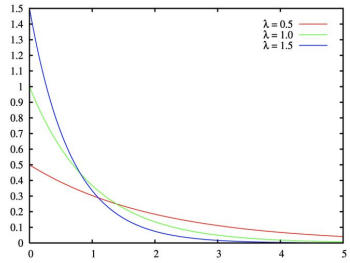
② we can reject null hypothesis.

Therefore, the smaller the P-value is, the more reliable to reject the null hypothesis.

3. Exponential Distribution

Exponential distribution is a continuous probability distribution in probability theory and statistics. This describes, for example, the Poisson process, the time interval of events following events in which events occur continuously and independently at a constant incidence.

The exponential distribution has a probability density function for parameter $\lambda > 0$ such that;



$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

Then, CDF is;

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

Notation:

$E(x) = \frac{1}{\lambda} = \theta$

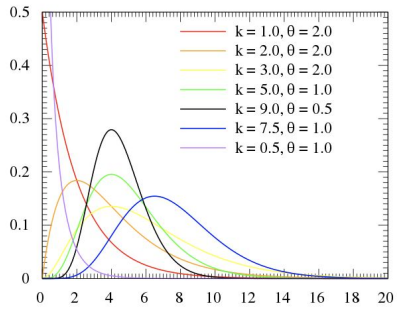
$V(x) = \frac{1}{\lambda^2} = \theta^2$

Characteristic: The exponential distribution is a special case where the parameter of gamma distribution shape is 1. Also, the Chi-squared distribution with 2 degrees of freedom coincides with the exponential distribution with θ = 2. It is also a special case where the coefficient m = 1 in the Weibull distribution.

4. Gamma Distributions

In probability theory and statistics, the gamma distribution is a continuous probability distribution. Its character is characterized by two parameters of shape parameter k and scale parameter θ. It is mainly applied to lifetime distribution of electronic parts in reliability engineering and distribution of waiting time of traffic in communication engineering. It also applies to income distribution.

For the gamma distribution, the probability density function uses the shape parameter k> 0 and the scale parameter θ> 0 such that;



$$f(x) = \frac{1}{\Gamma(k)\,\theta^k} x^{k-1} e^{-x/\theta} \quad \text{for } x > 0$$

Then, CDF is that;

$$F(x) = \int_0^x f(u)\,du = \frac{\gamma(k, x/\theta)}{\Gamma(k)} = \frac{\gamma(k, \lambda x)}{\Gamma(k)}$$

Notation:

$E(X) = k\theta = \frac{k}{\lambda}$

$V(X) = k\theta^2 = \frac{k}{\lambda^2}$

Characteristic: The gamma distribution has reproducibility. In other words, let X1 be the random variable of the gamma distribution with the shape parameter k1 and the scale parameter θ as the parameter, and the random parameter of the gamma distribution with the shape parameter k2 and the scale parameter θ as the parameter be X2, the probability variable The sum X 1 + X 2 follows the gamma distribution of the shape parameter k 1 + k 2 and the scale parameter θ.

5. Method of moments

Method of moments: univariate parameter (d=1)

Let g : R → R be a function such that

$$m(\theta) = E_\theta g(X) = \int g(x) f(x;\theta) dx$$

is continuous and monotonic. Then θ can be uniquely identified by the value m(θ), i.e., there exists an inverse function m^{-1} such that $\theta = m^{-1}(m(\theta))$. Although is m(θ) unknown, replacing it with its empirical counterpart leads to

$$\tilde{\theta} = m^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right)$$

Usually we choose $g(x) = x$ or $g(x) = x^2$, which explains the name of the method. This method was proposed by Karl Pearson in 1902 and is historically the first regular method of constructing a statistical estimate.

Method of moments: multivariate parameter (d ≥ 2)

MOM can be extended to the multivariate case. Let $g(y) = (g_1(y), ..., g_d(y))^T$ be a function $R \to R^d$. Define the moment $m(\theta) = (m_1(\theta), ..., m_d(\theta))^T$ by

$$m_j(\theta) = E_\theta g_j(X) = \int g_j(x) f(x;\theta) dx$$

The vector-valued function g is chosen such that m is invertible, i.e. for any $t = (t_1, ..., t_d)^T \in R^d$, the system of equations $m_j(\theta) = t_j$, $j = 1, ..., d$ has a unique solution. The empirical counterpart M_n of unknown moments $m(\theta)$ is

$$\mathbf{M}_n \stackrel{\text{def}}{=} \left(\frac{1}{n}\sum_i g_1(X_i), \ldots, \frac{1}{n}\sum_i g_d(X_i)\right)^{\mathsf{T}}.$$

Then the MOM estimator of is defined as

$$\tilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \mathbf{m}^{-1}(\mathbf{M}_n)$$

6. Maximum Likelihood

Maximum likelihood estimation: univariate parameter

The maximum likelihood method was first used by Sir Ronald Fisher in 1922. In fact, the method originated in the works of Gauss and Bernoulli.

Likelihood function: Let $X_1, X_2, ..., X_n$ be an iid sample with PMF/PDF $f(x;\theta)$. The likelihood function is the joint density of the data, defined by

$$L_n(\theta) = \prod_{i=1}^{n} f(x;\theta)$$

Likelihood function is not a density function, but a function of the parameter θ.

Maximum Likelihood estimator (MLE): The maximum likelihood estimator (MLE), denoted by $\widehat{\theta}$ or $\widehat{\theta}_n$, is the value of θ that maximized $L_n(\theta)$, i.e.

$$\hat{\theta} \in \arg\max_{\theta \in \Theta_0} L_n(\theta)$$

where Θ_0 is the parameter space.

Remarks:

① $L_n : \Theta_0 \rightarrow [\,0,\ \infty\,)$

② For any $\theta \in \Theta_0, L_n(\theta)$ is a random variable.

③ Multiplying $L_n(\theta)$ by any positive constant not depending on θ will not change the MLE. Hence, we often drop constants in the likelihood function. Note that $L_n(\theta)$ is the product of n functions. It would be easier to work with its logarithm, which is the sum of n functions.

Maximum likelihood estimation: multivariate parameter

Most of the time the model is complex enough so that there are multiple unknown parameters, denoted as a vector $\theta = (\theta_1, ..., \theta_d)^T$. in this case, the likelihood L_n is a function $R^d \rightarrow [0, \infty)$, the parameter space Θ_0 is a subset of R^d, and the first order condition becomes $\nabla_\theta L_n(\theta) = 0$,i.e.

$$\frac{\partial}{\partial\theta_1}\ell_n(\theta_1, \ldots, \theta_d) = 0,\ \ldots,\ \frac{\partial}{\partial\theta_d}\ell_n(\theta_1, \ldots, \theta_d) = 0.$$

A remarkable feature of the MLE is the invariance property. Briefly stated, $\widehat{\theta}$ is a MLE for θ, then $g(\widehat{\theta})$ is a MLE for $g(\theta)$, where g is a function from R^d to R^m. This result was proved by Zehna.

7. Mean Squared Error

The mean squared error of an estimator $\widehat{\theta}$ is defined as $MSE(\widehat{\theta}) = E(\widehat{\theta} - \theta)^2$.

If $\widehat{\theta}$ is unbiased, then $MSE(\widehat{\theta}) = var(\widehat{\theta})$.

To see this,

$$\begin{aligned} MSE(\widehat{\theta}) &= E(\widehat{\theta} - E\widehat{\theta} + E\widehat{\theta} - \theta)^2 \\ &= E(\widehat{\theta} - E\widehat{\theta})^2 + E\{(\widehat{\theta} - E\widehat{\theta})(E\widehat{\theta} - \theta)\} + E(E\widehat{\theta} - \theta)^2 \\ &= E(\widehat{\theta} - E\widehat{\theta})^2 + (E\widehat{\theta} - \theta)E(\widehat{\theta} - E\widehat{\theta}) + E(E\widehat{\theta} - \theta)^2 \\ &= E(\widehat{\theta} - E\widehat{\theta})^2 + E(E\widehat{\theta} - \theta)^2 \\ &= var(\widehat{\theta}) + \{bias(\widehat{\theta})\}^2 . \end{aligned}$$

Note:

When choosing among several different estimators, first we would like to select one that is unbiased. However, in the previous case, even though the MLE $\widehat{\theta}$ is biased, the bias is very small and is negligible in the long run, i.e. when n is large.

8. Asymptotic Distribution

Fisher Information

The Fisher information amount is an amount that appears in statistics or information theory, and represents the amount of "information" that the random variable X has with respect to the parameter θ. It was named after statistician Ronald Fisher.

Let θ be the parameter and let X be the random variable whose probability density function is expressed by $f(x\,|\theta)$, then Likelihood function of θ is defined

$$L(\theta|x) = f(x|\theta)$$

The score function is a derivative of log-likelihood function;

$$V(x;\theta) = \frac{\partial}{\partial\theta}\ln L(\theta|x)$$

The Fisher information is defined by the second moment of the score function;

$$\begin{aligned}\mathcal{I}_X(\theta) &= \mathrm{E}[V(x;\theta)^2|\theta] \\ &= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln L(\theta|x)\right)^2 \middle| \theta\right]\end{aligned}$$

Since the expected value is taken for X, the amount of Fisher information depends only on the probability density function $f(x|\theta)$ followed by X. Therefore, if X and Y have the same probability density function, their Fisher information amount is the same.

The Fisher information of $X \sim f(x;\theta)$ is

$$\mathcal{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right]$$

The Fisher information is a quadratic derivative with respect to θ of the logarithm of f plus minus. The Fisher information amount can also be regarded as the "sharpness" of the support curve near the maximum likelihood estimator for θ.

Asymptotically unbiased estimator

A point estimator $\widehat{\theta_n}$ is said to be asymptotically unbiased if its bias $E(\widehat{\theta_n}) - \theta$, as a function of sample size n, tends to 0 as $n \to \infty$. In other words, $\lim_{n\to\infty} E(\widehat{\theta_n}) = \theta$.

Given an iid sample $X_1, X_2, ..., X_n$ from a normal population $N(\mu, \sigma^2)$ with unknown μ and σ^2. The MLe of σ^2 is given by $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$, which satisfies $E\widehat{\sigma^2} = (1 - n^{-1})\sigma^2$. Cleary, $E\widehat{\sigma^2} \to \sigma^2$ as $n \to \infty$, so it is asymptotically unbiased.

The most commonly used asymptotic property is consistency, which refers to the shape of the CDF of $\widehat{\theta_n}$ and how that shape changes as a function of n.

Asymptotic properties of MLE

Under certain conditions on the model $\{f(x;\theta) : \theta \in \Theta\}$, the MLE $\widehat{\theta_n}$ possesses many properties that make it an appealing choice of estimator. The main properties of MLE are:

1. The MLE is consistent: $\widehat{\theta_n} \to \theta^*$, where θ^* is the true value in Θ;
2. The MLE is invariant: if $\widehat{\theta_n}$ is the MLE of θ, then $g(\widehat{\theta_n})$ is the MLE of $g(\theta)$;
3. The MLE is asymptotically normal: $(\widehat{\theta_n} \to \theta^*)/se \to N(0, 1)$, where se is the standard error that can often be computed analytically;
4. The MLE is asymptotically efficient: among all well-behaved estimators, the MLE has the smallest variance, at least for large samples.

Asymptotic Normality of MLE

Let $se = var(\widehat{\theta}_n)^{1/2}$. Under appropriate regularity conditions, the following hold;

1. $(\widehat{\theta}_n \to \theta^*)/se \to N(0,1)$ and $se/\{nI(\theta^*)^{-1/2} \to 1$ as $n \to \infty$;

2. $(\widehat{\theta}_n \to \theta^*)/se \to N(0,1)$, where $\widehat{se} = \{nI(\theta^*)^{-1/2}$.

These theorems indicate that;

$$\sqrt{n}(\widehat{\theta}_n \to \theta^*) \to N(0, \tfrac{1}{I(\theta^*)})$$

This statement says that this is still true if we replace the Fisher information $I(\theta^*)$ by its estimated version $I(\widehat{\theta}_n)$. From this fact we can construct an asymptotic confidence interval.

9. Hypothesis Tests

In the statistical hypothesis test, assuming that the hypothesis is correct, the probability of extracting the specimen actually observed is obtained from the population according to it, and judgment is made based on the value. If the probability is sufficiently small, we reject the hypothesis.

The statistical hypothesis test (for single proportions)is carried out in the following four procedures.

Step 1: State a null hypothesis $H_0 : p = p_0$ and an alternative hypothesis H_1 .

Two sided test H_0 : $H_0 = p_0$

One sided to the left H_0 : $H_0 < p_0$

One sided to the right H_0 : $H_0 > p_0$

Step 2: Construct a test statistic and compute its value from data (a number that reflects to what extent the data agree with H_0 and to what extent they favor H_1)

	Not Reject (Retain) Null	Reject Null
H_0 ture	✓	Type I error
H_1 ture	Type II error	✓

Type I error: rejecting H_0 when H_0 is true.

Type Ⅱ error: not rejecting H_0 when H_1 is true.

Step 3: Design a test based on the test statistic.

We use the test statistics;

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Where the point estimator is;

$$\hat{p} = \frac{x}{n}$$

Step 4: Make a conclusion.
If the result from step 3 is out of below interval, we reject H_0.
Confidential Intervals;

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

The statistical hypothesis test (for mean)is carried out in the following four procedures.
Assume that we observe an iid sample $X_1, X_2, ..., X_n$ from a normal population with $X \sim N(\mu, \sigma^2)$ unknown mean μ and known variance σ^2. For a given value μ_0, we wish to test;
Two sided test H_0: $H_0 = p_0$
One sided to the left H_0: $H_0 < p_0$
One sided to the right H_0: $H_0 > p_0$
Step 1: State a null hypothesis and an alternative versus .
Step 2: Construct a test statistic from the data;

$$Z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

Under the null hypothesis H_0, it holds $Z_n \sim N(0, 1)$.
Step 3: Pick a significance level (e.g. 5% or 10%), and compute the critical value $c_\alpha = z_{\alpha/2}$.
Step 4: make a decision.
If $H_1 : \mu > \mu_0$, we reject H_0 if $z \geq z_\alpha$, where z is the value of Z_n .
If $H_1 : \mu < \mu_0$, we reject H_0 if $z \leq -z_\alpha$, where z is the value of Z_n .
If $H_1 : \mu \neq \mu_0$, we reject H_0 if $|z| \geq z_{\alpha/2}$, where z is the value of Z_n .
P-value
There are two general ways to quantify the amount of evidence against that is contained in a given data set. The first involves the level of significance concept. Using this format, the user selects a value for α (typically 5%) before any data are collected. Once α is specified, a corresponding critical region can be identified. If the test statistic falls in the critical region, we reject at the level of significance. The second strategy is to calculate the test statistic's -value. Reporting "reject H_0 " or "fail to H_0 " is not very informative. Instead, we could ask, for every , whether the test rejects at that level. Generally, if the test rejects at level α , it will also reject at level $\alpha' > \alpha$. Hence, there is a smallest at which the test rejects. We call this number the value, which is a random number constructed from the data such that;
• For any -value, we reject at level
• For any -value, we fail to reject at level .
The P-value associated with an observed test statistic is the probability of getting a value for that test statistic as extreme as or more extreme than what was actually observed given that H_0 is true.
Remarks:

1. In short, the P-value is the probability (under H_0) of observing a value of the test statistic the same as or more extreme than what was actually observed.
2. Test statistics that yield small P-values should be interpreted as evidence against H_0 .
3. If the P-value calculated for a test statistic $\leq \alpha$, the null hypothesis can be rejected at the significance level α.
4. The P-value is the smallest α at which we can reject H_0.

VIII.ACKNOWLEDGEMENTS

Andrea, K., M.D. (1999). Cytomegalovirus Infection and HIV-1 Disease Progression in Infants Born to HIV-1–Infected Women. *The New England Journal of Medicine*.

Stockdale L, Nash S, Nalwoga A, Painter H, Asiki G, Fletcher H, Newton R (2018) Human cytomegalovirus epidemiology and relationship to tuberculosis and cardiovascular disease risk factors in a rural Ugandan cohort. PLOS ONE 13(2): e0192086. https://doi.org/10.1371/journal.pone.0192086

Stockdale L, Nash S, Nalwoga A, Painter H, Asiki G, Fletcher H, Newton R (2018) Data from: Human cytomegalovirus epidemiology and relationship to tuberculosis and cardiovascular disease risk factors in a rural Ugandan cohort. Dryad Digital Repository. https://doi.org/10.5061/dryad.d1k17

"Exponential distribution." *Wikipedia*, Wikimedia Foundation, 3 Mar. 2019, en.wikipedia.org/wiki/Exponential_distribution

"Poisson distribution." *Wikipedia*, Wikimedia Foundation, 4 Mar. 2019, en.wikipedia.org/wiki/Poisson_distribution

"Mean squared error." *Wikipedia*, Wikimedia Foundation, 28 Jan. 2019, en.wikipedia.org/wiki/Mean_squared_error

"Gamma distribution." *Wikipedia*, Wikimedia Foundation,15 Feb. 2018, en.wikipedia.org/wiki/Gamma_distribution

"Method of moments (statistics)." *Wikipedia*, Wikimedia Foundation, 6 Feb. 2019, en.wikipedia.org/wiki/Method_of_moments_(statistics)

"Goodness of fit." *Wikipedia*, Wikimedia Foundation, 21 Feb. 2019, en.wikipedia.org/wiki/Goodness_of_fit

"Chi-squared test." *Wikipedia*, Wikimedia Foundation, 14 Jan. 2019, en.wikipedia.org/wiki/Chi-squared_test

"Maximum likelihood estimation." *Wikipedia*, Wikimedia Foundation, 19 Feb. 2019, en.wikipedia.org/wiki/Maximum_likelihood_estimation

"Fisher information." *Wikipedia*, Wikimedia Foundation, 20 Feb. 2019, en.wikipedia.org/wiki/Fisher_information

"Statistical hypothesis testing." *Wikipedia*, Wikimedia Foundation, 20 Feb. 2019, en.wikipedia.org/wiki/Statistical_hypothesis_testing

"Bias of an estimator." *Wikipedia*, Wikimedia Foundation, 16 Jan. 2019, en.wikipedia.org/wiki/Bias_of_an_estimator