# CHAPTER 4: PATTERNS IN DNA

math 189 : data analysis and inference : winter 2019

Jelena Bradic
http://www.jelenabradic.net/
Associate Professor, Department of Mathematics and Halicioglu Data Science Institute
University of California, San Diego
jbradic@ucsd.edu

# Introduction

## The data

## Background

## Investigations

## Theory


Jelena Branc http://www.jelenabranic.net/ Associate Professor, Department of Mathematics and Halicioglu Data Science Institute University of Califo

# Introduction

* The human cytomegalovirus (CMV) is a potentially life-threatening disease for people with suppressed or deficient immune system..
* To develop strategies for combating the virus, scientists study the way in which the virus replicates.
* In particular, they are in search of a special place on the virus' DNA that contains instructions for its reproduction: origin of replication.

# DNA

* A virus' DNA contains all of the information necessary for it to grow, survive and replicate.
* DNA can be though of as a long, coded message made from a four-letter alphabet: A, C, G, T.
* DNA sequences contain many patterns, as the alphabet is small.
* Some of these patterns may flag important sites on the DNA, such as the origin of replication.
* Complementary palindrome is one type of pattern. In DNA, the letter A is complementary to T, and G is complementary to C, and complementary palindrome is a sequence
of letters that reads in reverse as the complement of the forward sequence :

```
GGGCATGCCC
```

.

# Search

* The origin of replication for two viruses from the same family as CMV, the herpes family, are marked by complimentary palindromes. One of them, Herpes simplex, is marked by a long palindrome of 144 letters.The other, the Epstein-Barr virus, has several short palindromes and close repeats clustered at the origin of replication.
* For the CMV, the longest palindrome is 18 basepairs, and altogether, contains 296 palindromes between 10 and 18 base pairs long. Biologist conjectured that clusters of palindromes in CMV may serve the same role as the single long palindrome in Herpes simplex, or the cluster of palindromes and short repeats in the Epstein-Barr virus' DNA.
* To find the origin of replication, DNA is cut into segments and each segment is tested to determine whether it can replicate. If it does not replicate, then the origin of replication must not be contained in the segment.

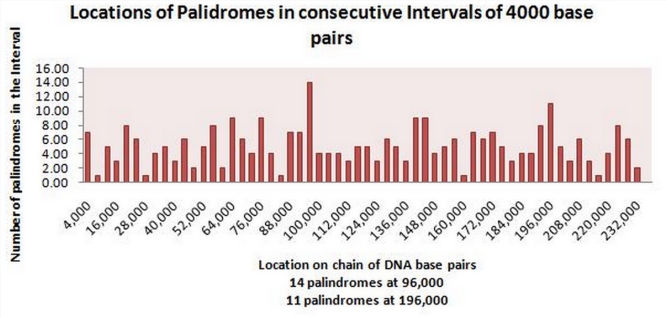
# Search

* This process can be very time consuming and expensive without leads on where to begin the search. A statistical investigation of the DNA to identify unusually dense clusters of palindromes can help narrow the search and potentially reduce the amount of testing needed to find the origin of replication.
* In this lab we will search for unusual clusters of complementary palindromes.

# Description

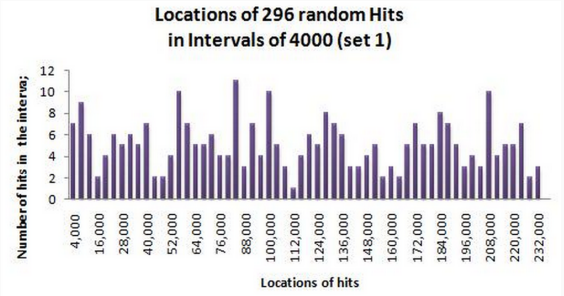
* DNA sequence of CMV was published in 1990 (Chee et al.)
* Leung et al. (1991) implemented search algorithms to screen the sequence for many types of patterns
* Altogether, 296 palindromes were found that were at least 10 letters long.
* The longest ones found were 18 letters long and occurred in locations 14719, 75812, 90763 and 173893 along the sequence.
* Palindromes shorter than 10 letters were ignored.
* The CMV DNA is 229,354 letters long.

* One way to begin to group the data of the 296 palindromes found is to segment the DNA chain into intervals of base pairs and count the number of palindromes found in each interval.

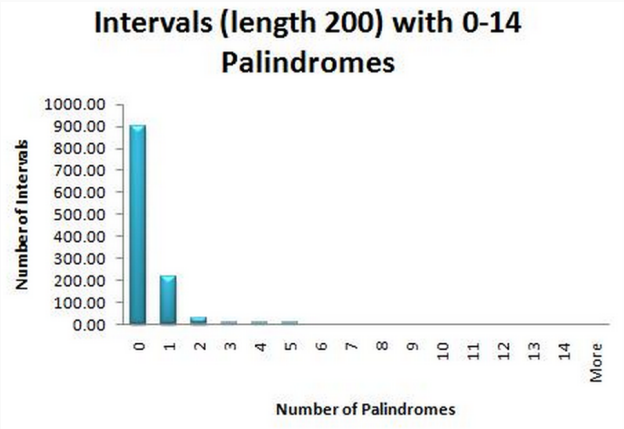



** From these histograms, it is fairly easy to see that no matter the length of the interval, there appear to be clusters of palindromes in at least two locations: around the 93,000th and 195,000th pairs of DNA. This is enough to formulate a hypothesis which claims that the clusters at theses two locations are exceptions within the typical structure of the DNA chain, i.e. that the clusters are not due to chance.

** By comparing histograms of the actual palindromes to histograms based on randomly generated numbers we can see that the random sets of numbers present no pattern of clusters at any given point, no matter what size intervals we use.

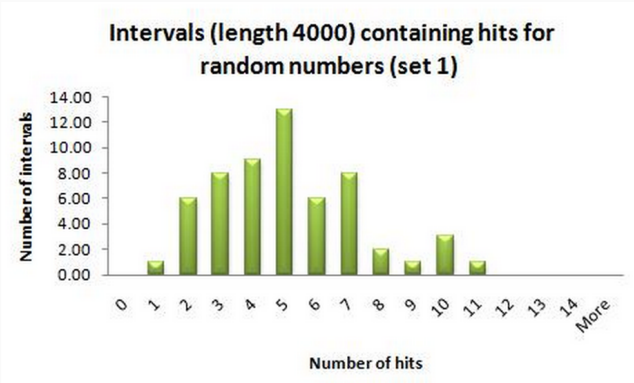


* We can also see that the observed palindromes present higher spikes of number of palindromes per intervals. In addition, there does not appear to be any consistent pattern of clusters of hits with the random numbers.

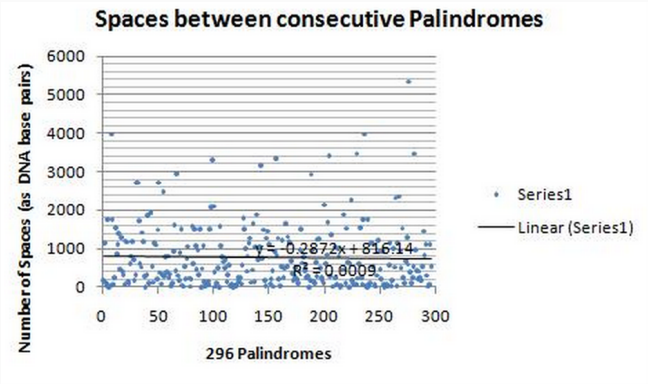



** We can see that no matter the length of the intervals, there always seem to be one or two outliers of intervals containing a higher number of palindromes.

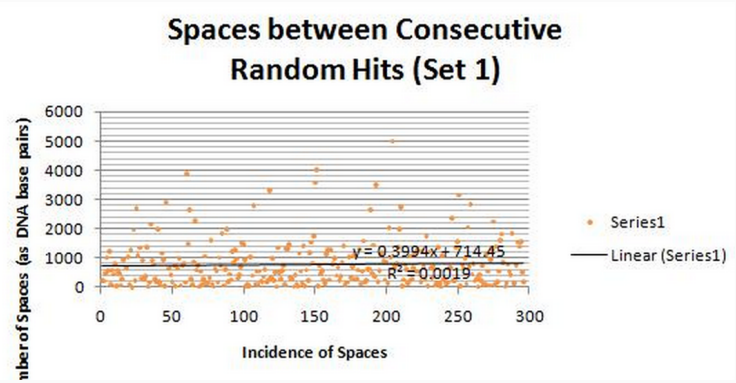
** We can observe that the intervals of the random hits do not display such outliers. Therefore it would seem logical to deduce that the outliers on the DNA are atypical and worth examining for the replication code.



* The Investigations in the chapter suggest looking at the spaces between the palindromes.




** A scatterplot of the spaces between the palindromes doesn't seem to show any patterns that may be useful.




** Perhaps there is another way of analyzing the spaces that is more useful.

# DNA

1944 Avery, MacLeod and McCarty showed that DNA was the carrier of hereditary information.

- 1953, Franklin, Watson and Crick found that 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.

- All the sugars in the DNA are deoxyribose.

- The basis come in four types: adenine, cytosine , guanine and thymine, or A,C,G,T for short.

- As the basis vary from one nucleotide to another, they give the appearance of a long, coded message.

# DNA(cont.)

The two strange of the nucleotides are connected at the bases, forming complementary pairs. The bases on one strand are paired to the other strand: A to T, C to G, G to C and T to A

.

- 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.

# Viruses

Viruses are very simple structures with two main parts: a DNA molecule wrapped within a protein shell called a capsid.

- The DNA stores all the necessary information for controlling life processes, including its own replication

- The DNA for viruses typically ranges up to several hundred thousand base pairs in length.

- For example, E coli. replication begins when a ”snipping” enzyme cuts the DNA strand apart at a small region called the origin. In the neighborhood are plenty of free nucleotides. When a free nucleotide meets its complementary base on the DNA, it sticks, while the ”wrong” nucleotides bounce away.

- As the snipping enzyme opens the DNA further, more nucleotides are added, and a clipping enzyme puts them together.

# Human Cytomegalovirus

CMV is a member of the herpes virus family.

Incidence of CMV varies geographically from 30% to 80%. Typically 10%-15% of children are infected with CMV before the age of 5. The infection then levels off until young adulthood, when it again increases and presents symptoms often similar to mononucleosis.

Once infected, CMV lays dormant. It only become harmful when the virus enters a productive cycle in which it quickly replicated tens of thousands of copies.

In this cycle it poses a major risk for people in immune-depressed states: transplant patients, AIDS patients, etc.

Locating the origin of replication for CMV may help virologist find an effective vaccine against the virus.

Jelena Branc http://www.jelenabranic.net/ Associate Professor Department of Mathematics and Halicioglu Data Science Institute University of Califo

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

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

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

* [**Random scatter**] To begin, pursue the point of view that structure in the data is indicated by departures from a uniform scatter of palindromes across the DNA.
  ** Of course, a random uniform scatter, does not mean that the palindromes will be equally spaced as milestones on a freeway. There will be some gaps on the DNA where no palindromes occur, and there will be some clumping together of palindromes.

  To look for structure examine the locations of the palindromes, the spacing between palindromes, and the counts of palindromes in non overlapping regions of the DNA. One starting place might be to see first how random scatter looks by using a computer to simulate it.

  ** A computer can simulate 296 palindrome sites chosen at random along a DNA sequence of 229,354 bases using a pseudo random number generator. When this is done several times,by making seller sets of simulated palindrome locations, then the real data can be compared to the simulated data.

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

* [**Random scatter**] To begin, pursue the point of view that structure in the data is indicated by departures from a uniform scatter of palindromes across the DNA.
  ** Of course, a random uniform scatter, does not mean that the palindromes will be equally spaced as milestones on a freeway. There will be some gaps on the DNA where no palindromes occur, and there will be some clumping together of palindromes.

  To look for structure examine the locations of the palindromes, the spacing between palindromes, and the counts of palindromes in non overlapping regions of the DNA. One starting place might be to see first how random scatter looks by using a computer to simulate it.

  ** A computer can simulate 296 palindrome sites chosen at random along a DNA sequence of 229,354 bases using a pseudo random number generator. When this is done several times,by making seller sets of simulated palindrome locations, then the real data can be compared to the simulated data.
* [**Locations and spacings**] Use graphical methods to examine the spacings between consecutive palindromes and sum of consecutive pairs, triplets, etc, spacings. Compare what you find to what would you expect to find in a random scatter. Also, use graphical methods to compare locations of the palindromes.

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

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

* [**Counts**] Use graphical methods and more formal statistical tests to examine the counts of palindromes in various regions of the DNA. Split the DNA into nonoverlaping regions of equal length to compare the number of palindromes in an interval to the number of that would you expect from uniform random scatter. The counts for shorter regions will be more variable than those for longer regions. Also consider classifying the regions according to their number of counts.

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

* [**Counts**] Use graphical methods and more formal statistical tests to examine the counts of palindromes in various regions of the DNA. Split the DNA into nonoverlaping regions of equal length to compare the number of palindromes in an interval to the number of that would you expect from uniform random scatter. The counts for shorter regions will be more variable than those for longer regions. Also consider classifying the regions according to their number of counts.
* [**The biggest cluster**] Does the interval with the greatest number of palindromes indicate a potential origin of replication?Be careful in making your intervals, for any small, but significant ,deviations from random scatter, such as a tight cluster of a few palindromes, could easily go undetected if the regions examined are too large. Also, if the regions are too small, a cluster of palindromes may be split between adjacent intervals and not appear as a high-count interval.

How would you advise biologist who is a bout to start experimentally searching for the origin of replication? Write your recommendations in the form of a report that a team members including biologist will read.

Jelena Bradic http://www.jelenabradic.net/ Associate Professor Department of Mathematics and Halicioglu Data Science Institute University of Califo

# Goals

Understand a random model that describes the behavior of ”counts” of the number of palindromes and for a ”uniform” aka random scatter of palindromes.

* To determine the estimation procedure in such a model.
* To understand how to find statistical discrepancies between a model with clusters and model without clusters.
  * Is a model a good model
  * Can we formulate spacings as well as counts in the model
  * What is a hypothesis tests
  * How is uniform distribution related to the problem

The Homogeneous Poisson Process is a model for random phenomena such as arrival times of telephone calls at an exchange, the decay times of radioactive particles, and the position of stars in parts of the sky.

The Homogeneous Poisson Process is a model for random phenomena such as arrival times of telephone calls at an exchange, the decay times of radioactive particles, and the position of stars in parts of the sky.

The process arises naturally from the notion of points haphazardly distributed on a line with no obvious regularity.

The characteristic features of the process are

- The underlying rate $\lambda$ at which points, called hits, occur and is such that id doesn't change with location (homogeneity).
- The number of points falling in separate regions are independent.
- No two points can land in exactly the same place.

These three properties are enough to derive the formal probability model for The Homogeneous Poisson Process.

# The Homogeneous Poisson Process

The poison process is a good reference model for making comparisons because it is a natural model for uniform random scatter.

* The strand of the DNA can be thought of as a line, and the location of a palindrome can be thought of as a point on the line
* The uniform random scatter model says: palindromes are scattered randomly and uniformly across the DNA
* The number of palindromes in any small piece of DNA is independent of the number of palindromes in another, non overlapping piece
* The chance that one tiny piece of DNA ha palindrome in it is the same for all tiny pieces of the DNA.

# Counts and the Poisson Distribution

Counts of the number of points in different regions follow Poisson distribution with rate $\lambda$.

$$P(\text{k points in a unit interval }) = \frac{\lambda^k}{k!} e^{-\lambda}, \text{ for } k = 0, 1, \cdots \ldots$$

* Te rate $\lambda$ is the rate of hits per unit.
* E of Poisson random variable is $\lambda$, hence it stands for the expected number of hits per unit interval
* In most examples rate $\lambda$ is unknown. A good estimate is the empirical average number of hits per unit interval
* This method of estimation is called

  method of moments.

* Another method of estimation is called

  maximum likelihood method.

  For poisson distribution they result in the same estimator.

# Goodness of Fit for Probability Distributions

We often hypothesize that the observations are realizations of independent random variables from a specified distribution such is Poisson distribution. We do not believe that data follow this distribution exactly, but rather that this distribution isa good proxy for the randomness we observe in the data.

* If Poisson distribution fits the data well then it could be useful in searching for the unusual clusters.

We would want to use the The Homogeneous Poisson Process as a reference model against which to seek an excess of palindromes. This only makes sense if the model more or less fits the data.

A technique foe accessing how well the reference model fits to the data is to apply the chi-square goodness of fit test

* Divide the CMV DNA into 57 non overlapping regions of length 4000 bases, and tally the number of complementary palindromes in each segment

| Palindrome counts | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 7 | 1 | 5 | 3 | 8 | 6 | 1 | 4 | 5 | 3 |
| 6 | 2 | 5 | 8 | 2 | 9 | 6 | 4 | 9 | 4 |
| 1 | 7 | 7 | 14 | 4 | 4 | 4 | 3 | 5 | 5 |
| 3 | 6 | 5 | 3 | 9 | 9 | 4 | 5 | 6 | 1 |
| 7 | 6 | 7 | 5 | 3 | 4 | 4 | 8 | 11 | 5 |
| 3 | 6 | 3 | 1 | 4 | 8 | 6 | | | |

*
* There is nothing special about the number 4000. It is chosen to make the number of observations in the table reasonable.
* The distribution of these counts appears in the PICS
* The last column in the table above contains the expected number of segments continuing the specified number of palindromes as computed from the hypothesized Poisson distribution.

# Basic principle (cont.)

* | Palindrome count | Number of intervals Observed | Expected |
|---|---|---|
| 0–2 | 7 | 6.4 |
| 3 | 8 | 7.5 |
| 4 | 10 | 9.7 |
| 5 | 9 | 10.0 |
| 6 | 8 | 8.6 |
| 7 | 5 | 6.3 |
| 8 | 4 | 4.1 |
| 9+ | 6 | 4.5 |
| Total | 57 | 57 |

  $57P(0, 1 \text{ or } 2 \text{ palindromes in an interval of length } 4000) = 57e^{-\lambda}[1+\lambda+\lambda^2/2]$

* The rate $\lambda$ is not known. There are 294 palindromes in the 57 intervals of length 4000, so the sample rate is 5.16 per 4000 base pairs.
* Plugging this estimate into calculations above yields 0.112 for the change that an interval of 4000 base pairs has 0, 1, or 2 palindromes. Hence the approximate expected number is

$$57 \times 0.112 = 6.4.$$

  This is approximate as we are using an estimated value of $\lambda$.

# Chi-Squared Test Statistics

To compare the observed data to the expected, we compute the following statistic:

$$\frac{(7-6.4)^2}{6.4}+\frac{(8-7.5)^2}{7.5}+\frac{(10-9.7)^2}{9.7}+\frac{(9-10)^2}{10}$$
$$+\frac{(8-8.6)^2}{8.6}+\frac{(5-6.3)^2}{6.3}+\frac{(4-4.1)^2}{4.1}+\frac{(6-4.5)^2}{4.5}=1.0$$

* If the random scatter model is true, then the test statistic computed here has an approximate chi-square distribution (also written $\chi^2$) with six degrees of freedom.
* The size of the actual test statistic is a measure of the fit of the distribution.
* Large values indicate that the observed data were quite.

# Chi-Squared Test

We use the $\chi^2$ distribution to compute the chance of observing a test statistic at least as large as ours under the random scatter model:

$$\mathrm{P}\left(\chi^2_6 \text{ random variable } \geq 1.0\right) = 0.98.$$

From this computation, we see that deviations as large as ours (or larger) are very likely. Hence, we conclude that it appears that the Poisson is a reasonable initial model.

The hypothesis test performed here is called a chi-square goodness of fit test.

# Chi-Squared Test

In general, to construct a hypothesis test for a discrete distribution, a distribution table is constructed from the data, where m represents the number of categories or values for the response and $N_j$ stands for the number of observations that appear in category j, $j = 1, \ldots, m$. These counts are then compared to what would be expected under the null hypothesis, i.e. under the assumption that the data does follow poisson distribution:

$$\mu_j = np_j, \qquad p_j = P(\text{ an observation is in category j }).$$

Note that $\sum p_j = 1$ so $\sum_j \mu_j = n$.

# Chi-Squared Test

Sometimes a parameter of the distribution needs to be estimated in order to compute the probabilities. In this case, data are used to estimate the unknown parameter(s). The measure of discrepancy between the sample counts and the expected counts is

$$\sum_{j=1}^{m} \frac{(\text{jth sample count} - \text{jth Expected count})^2}{\text{jth Expected count}} = \sum_{j=1}^{m} \frac{(N_j - \mu_j)^2}{\mu_j}.$$

When the statistic computed in the hypothesis test (called test statistic) is large it indicated a lack of fit of the distribution

Assuming that the data are generated from the hypothesized distribution, we can compute the chance that the test statics would be as large, or large, than that observed. This chance is called the observed significance level, or p-value.

To compute p-value we use $\chi^2$ distribution. If the probability model is correct, then the test statistic has na pproximate chi-squared distribution with $m - k - 1$ degrees of freedom, where m is the number of categories and k is the number of parameters estimated to obtain the expected counts.

$\chi^2_{m-k-1}$ is a continuous distribution on the positive real line and the density has a long right tail. As the degrees of freedom increase it starts to look symmetric and a lot like normal.

# p-value (cont.)

If the p-value is small, then there is a reason to doubt the fit of the distribution.

When this is the case, a residual plot can help determine where the lack of fit occurs. For each category, plot the standardized residuals

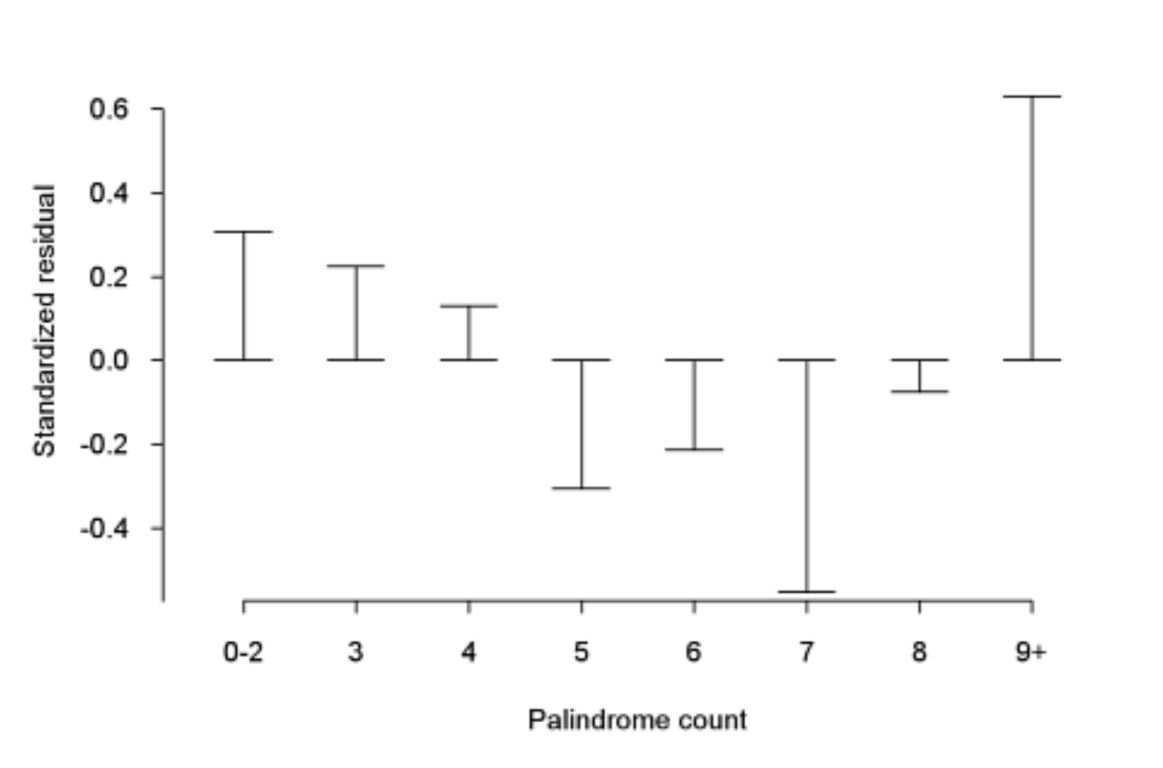
$$\frac{\text{sample count} - \text{Expected count}}{\sqrt{\text{Expectedcount}}} = \frac{N_j - \mu_j}{\sqrt{\mu_j}}.$$

The denominator transforms residuals in order to give them approximately equal variance. Square root make sense for meaningful comparisons across categories.

Note: Sum of residuals is always zero but the sum of standardized residuals is not.

# p-value (cont.)

Values of standardized residual larger than 3 indicate a lack of fit.



Under the Poisson process model for random scatter, if the total number of hits in an interval is known, then the positions of the hits are uniformly scattered across the interval.

In other words, the Poisson process on a region can be viewed as a process that first generates a random number, which is the number of hits, and then generated locations for the hits according to the uniform distribution.

Hence, for the CMV DNA, under the uniform random scatter, the positions of these palindromes are like 296 independent observations from a uniform distribution. Hence, these locations can be compared to the expected locations from the uniform distribution.

* If the DNA is split into 10 equal subintervals, the according to the uniform distribution, we would expect each interval to contain 1/10 of the palindromes.
* Hence, we perform another $\chi^2$ test.

| Segment | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Observed | 29 | 21 | 32 | 30 | 32 | |
| Expected | 29.6 | 29.6 | 29.6 | 29.6 | 29.6 | |
| Segment | 6 | 7 | 8 | 9 | 10 | Total |
| Observed | 31 | 28 | 32 | 34 | 27 | 296 |
| Expected | 29.6 | 29.6 | 29.6 | 29.6 | 29.6 | 296 |

# Which test?

Why did we use 57 intervals over 4000 base pairs in our goodness of fit test?

If we based the test on much shorter interval lengths, we would get many more intervals but a larger proportion of them would contain zero palindromes.

For example with an interval length of 400 base pairs, we would get 522 of the 573 intervals have 0 or 1 palindromes. The distribution of the counts is now highly skewed and the test is uninformative because a large proportion of the counts are in two categories (0 or 1 palindromes).

Alternatively, why not use large intervals? Suppose we divide the DNA into 10 large, equal-sized intervals. If we do this, we have hardly enough data to compare observed and expected numbers of intervals for a particular palindrome count. Our sample size is 10 but the 10 intervals have 8 different palindrome counts.

Distances between successive hits follows an Exponential distribution.

$$P(\text{the distance between the first and second hits } > t) \tag{1}$$
$$= P(\text{ no hits in an interval of length } t) = e^{-\lambda t} \tag{2}$$

Distances between the hits that are two apparatus, follows a Gamma distribution with parameters $2, \lambda$.

---

Note: $\mathcal{E}(\lambda) - \Gamma(1, \lambda)$; $\chi^2_k - \Gamma(k/2, 1/2)$

# Maximum Number of Hits

Under the Poisson process model, the numbers of hits in a set of non-overlapping intervals of the same length are independent observations from a Poisson distribution. This implies that the greatest number of hits in a collection of intervals behaves as the maximum of independent Poisson random variables. If we suppose that there are m such intervals then

$$
\begin{aligned}
&\mathrm{P}(\text{ maximum count over m intervals } \geq k) && (3)\\
&= 1 - \mathrm{P}(\text{ maximum count over m intervals } < k) && (4)\\
&= 1 - \mathrm{P}(\text{ all interval counts} < k) && (5)\\
&= 1 - \mathrm{P}(\text{ first interval counts} < k)^m && (6)\\
&= 1 - \left[\lambda^0 e^{-\lambda} + \cdots + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}\right]^m && (7)
\end{aligned}
$$

For a given estimate of $\lambda$, from the above expression, we ca find the approximate chance that the greatest number of hits is at least k. If this chance is unusually small, then it provides evidence for a cluster that is larger than the expected from the Poisson process. We can use the maximum palindrome counts as a test statistic, and the computation above provides the p-value for the test statistic.

# Method of moments

Suppose we have an independent sample

$$x_1, \ldots, x_n$$

from a Poisson distribution with unknown rate parameter $\lambda$.

Method of moments is one estimation technique that proceeds as follows:

1. Find $E(X)$ where $X$ has Poisson distribution with rate $\lambda$
2. Express $\lambda$ in terms of $E(X)$
3. Replace $E(X)$ with $\bar{x}$ to produce an estimate of $\lambda$, called $\hat{\lambda}$.

For Poisson distribution

$$E(X) = \lambda \Longrightarrow \bar{x} = \hat{\lambda}.$$

If higher moments need to be computed then $E(X^2)$ is replaced with $\sum_i x_i^2/n$.

Suppose we have an independent sample

$$x_1, \ldots, x_n$$

from a Poisson distribution with unknown rate parameter $\lambda$.

Maximum Likelihood method searches among all Poisson distributions to find the one that places the highest chance on the observed data.

For Poisson distribution, the chance of observing $x_1, , x_n$ is

$$\frac{\lambda^{x_1}}{x_1!}e^{-\lambda} \times \cdots \frac{\lambda^{x_n}}{x_n!}e^{-\lambda} = \frac{\lambda^{\sum_i x_i}}{\prod_i x_i!}e^{-\lambda} := L(\lambda)$$

For given data, this is a function of $\lambda$ that is called the likelihood function. Maximum likelihood estimates the unknown parameter by the $\lambda$-value that maximized the likelihood function.

# Maximum Likelihood (cont.)

Since the function is monotonically increasing, the log likelihood function, denoted with l, is maximized at the same value as L. To find the maximum we consider solving the first-order equation

$$\frac{\partial}{\partial \lambda} l(\lambda) = \frac{\partial}{\partial \lambda} \left[ \sum_i x_i \log(\lambda) - n\lambda - \sum_i \log(x_i!) \right] = \sum_i /\lambda - n = 0.$$

By solving the last equation for $\lambda$ we obtain:

$$\hat{\lambda} = \bar{x}.$$

# Maximum Likelihood (cont.)

Maximum-likelihood for continuous distributions is the same. Suppose we have an independent sample

$$x_1, \ldots, x_n$$

from an Exponential distribution with the unknown parameter $\theta$. Now, the Likelihood function, given the data is

$$L(\lambda) = \theta^n e^{-\theta \sum_i x_i},$$

and the log-likelihood function

$$l(\theta) = n \log(\theta) - \theta \sum_i x_i.$$

By solving the last equation for $\theta$ we obtain:

$$\hat{\theta} = \frac{1}{\bar{x}}.$$

# Mean Squared Error

To compare and evaluate parameter estimates, we use mean squared error, defined as

$$\text{MSE}(\hat{\lambda}) = \mathbb{E}(\hat{\lambda} - \lambda)^2 = \underbrace{\text{Var}(\hat{\lambda})}_{\text{variance}} + \underbrace{\left[\mathbb{E}(\hat{\lambda}) - \lambda\right]^2}_{\text{squared BIAS}}$$

Many of the estimators we use are UNBIASED, but sometimes an estimator with a small bias will have a small MSE.

### Theorem

Under certain regularity conditions, as the sample size increases, the Maximum-likelihood estimator, $\hat{\lambda}$ satisfies

$$\hat{\lambda} \to \lambda$$

$$\hat{\lambda} \sim \mathcal{N}\left(\lambda, \frac{1}{n\text{I}(\lambda)}\right)$$

where $\text{I}(\lambda)$ is called the Fisher's Information Matrix.

Fisher's Information matrix is defined as

$$I(\lambda) = \mathbb{E}\left(\frac{\partial}{\partial\lambda}\log f_\lambda(X)\right)^2 = -\mathbb{E}\left(\frac{\partial^2}{\partial\lambda^2}\log f_\lambda(X)\right).$$

Hence, as n increases

$$\sqrt{nI(\lambda)}\left(\hat{\lambda} - \lambda\right) \sim \mathcal{N}(0, 1).$$

The approximate normal distribution can be used to build the 95% confidence interval for the unknown $\lambda$ as

$$\hat{\lambda} \pm 1.96\sqrt{nI(\lambda)}.$$

---

Note: An asymptotic variance of the MLE is a lower bound for any other unbiased parameter estimate.

# Hypothesis Tests

The $\chi^2$ goodness-of-fit test and the test for the maximum number of palindromes in an interval, are two examples of hypothesis tests.

Here we provide another example of a hypothesis test, one for parameter values. We use it to introduce the statistical terms in testing.

# An Example (cont.)

In Hennepin County, a simple random sample of 119 households found an average radon level of 4.6 pCi/l with a standard deviation as 3.4pCi/l. In neighboring Ramsey County, a simple random sample of 42 households has an average radon level of 4.5 pCi/l with a standard deviation as 4.9pCi/l. It is claimed that the households in these two counties have the same average radon level and that the difference observed in the sample averages is due to chance variation in the sampling procedure.

To investigate this claim we introduce a hypothesis test.

# An Example (cont.)

We begin with a Probability Model.

Let $X_1, \ldots, X_{119}$ denote the radon levels for the Hennepin County and let $Y_1, \ldots, Y_{42}$ denote the radon levels for the Ramsey County. Also set $\mu_H$ and $\mu_R$, and $\sigma_H$ and $\sigma_R$, denote the average, and standard deviation of the radon levels in these two counties respectively.

The Null hypothesis is that the average radon levels are the same

$$H_0 : \mu_H = \mu_R$$

and the Alternative hypothesis is

$$H_R : \mu_H \neq \mu_R.$$

In hypothesis testing we assume that the $H_0$ is true and find out how likely our data are under this model.

# An Example (cont.)

We continue by finding estimators for the unknown parameters.

$\bar{X}$ and $\bar{Y}$ are good estimators of $\mu_H$ and $\mu_R$. They are independent and asymptotically normally distributed :

$$\bar{X} \sim \mathcal{N}\left(\mu_H, \frac{\sigma_H^2}{119}\right), \qquad \bar{Y} \sim \mathcal{N}\left(\mu_R, \frac{\sigma_R^2}{42}\right)$$

This implies that

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_H - \mu_R, \frac{\sigma_H^2}{119} + \frac{\sigma_R^2}{42}\right)$$

and that under the null hypothesis (that is if the null hypothesis is true)

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(0, \frac{\sigma_H^2}{119} + \frac{\sigma_R^2}{42}\right).$$

# An Example (cont.)

Next step is to find a Test Statistics.

Since $\bar{X} - \bar{Y}$ has approximately normal distribution a good candidate for the test statistic is its rescaled version, that under the null satisfies

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_H^2}{119} + \frac{\sigma_R^2}{42}}} \sim \mathcal{N}(0, 1)$$

We call this test statistic a Z-test, as it is based on normal approximations.

# An Example (cont.)

Next step involves finding the unusual values of the Test Statistic.

Values of Z such that $\{|Z| > 0.12\}$ are unusual for this setup.

Why ? The magic number 0.12 comes from the observations, i.e. the observed value of the Test statistics is 0.12.

Then the p-value is computed as :

$$\mathbb{P}\left(|Z| > Z_{\text{observed}}\right) = \mathbb{P}\left(|Z| > 0.12\right) = 0.90$$

We are ready to conclude or make a decision:

As p-value > 5%, we conclude that the observations support the Null hypothesis.

If the p-value was $< 5\%$ we would have concluded that the observations do not support the null, and we would reject the null in favour of the alternative.

The cutoff of 5% is called significance level of a test.

# An Example (cont.)

Next step involves discovering if we have made erroneous decision!

Note that the p-value is not the chance that the null hypothesis is true: the hypothesis is either true or not.

When we reject the null hypothesis we don't know if we have been unlucky with our sampling and observed a rare event or if we are making the correct decision.

| | | **Decision** | |
|---|---|---|---|
| | | fail to reject $H_0$ | reject $H_0$ |
| **Truth** | $H_0$ true | | |
| | $H_A$ true | | |

# An Example (cont.)

Next step involves discovering if we have made erroneous decision!

Note that the p-value is not the chance that the null hypothesis is true: the hypothesis is either true or not.

When we reject the null hypothesis we don't know if we have been unlucky with our sampling and observed a rare event or if we are making the correct decision.

| | | **Decision** | |
|---|---|---|---|
| | | fail to reject $H_0$ | reject $H_0$ |
| **Truth** | $H_0$ true | | |
| | $H_A$ true | | |

# An Example (cont.)

Next step involves discovering if we have made erroneous decision!

Note that the p-value is not the chance that the null hypothesis is true: the hypothesis is either true or not.

When we reject the null hypothesis we don't know if we have been unlucky with our sampling and observed a rare event or if we are making the correct decision.

| | | **Decision** | |
|---|---|---|---|
| | | fail to reject $H_0$ | reject $H_0$ |
| **Truth** | $H_0$ true | ✓ | |
| | $H_A$ true | | |

# An Example (cont.)

Next step involves discovering if we have made erroneous decision!

Note that the p-value is not the chance that the null hypothesis is true: the hypothesis is either true or not.

When we reject the null hypothesis we don't know if we have been unlucky with our sampling and observed a rare event or if we are making the correct decision.

| | | **Decision** | |
|---|---|---|---|
| | | fail to reject $H_0$ | reject $H_0$ |
| **Truth** | $H_0$ true | ✓ | |
| | $H_A$ true | | ✓ |

# An Example (cont.)

Next step involves discovering if we have made erroneous decision!

Note that the p-value is not the chance that the null hypothesis is true: the hypothesis is either true or not.

When we reject the null hypothesis we don't know if we have been unlucky with our sampling and observed a rare event or if we are making the correct decision.

| | | **Decision** | |
|---|---|---|---|
| | | fail to reject $H_0$ | reject $H_0$ |
| **Truth** | $H_0$ true | ✓ | Type 1 Error |
| | $H_A$ true | | ✓ |

# An Example (cont.)

Next step involves discovering if we have made erroneous decision!

Note that the p-value is not the chance that the null hypothesis is true: the hypothesis is either true or not.

When we reject the null hypothesis we don't know if we have been unlucky with our sampling and observed a rare event or if we are making the correct decision.

| | | **Decision** | |
|---|---|---|---|
| | | fail to reject $H_0$ | reject $H_0$ |
| **Truth** | $H_0$ true | ✓ | Type 1 Error |
| | $H_A$ true | Type 2 Error | ✓ |

# An Example (cont.)

Next step involves computing the errors: Type I and Type II errors!

Luckily,

Type I error $= \alpha$ or the significance of the test!

Unluckily, Type II error is a bit complicated !

Type II error $:= \beta$. Power of a test $:= 1 - \beta$.

Typically $\alpha$ is set in advance and $\beta$ is computed for various values of the alternative hypothesis. High power is a sign of a good test.

# An Example (cont.)

For example, the power of the test in this example for $\alpha = 0.05$ and $\mu_H - \mu_R = 0.5$ (which is one specific value in the alternative) is

$$\mathbb{P}\left(\frac{|\bar{X} - \bar{Y}|}{0.81} > 1.96\right) \tag{8}$$

$$= \mathbb{P}\left(|\bar{X} - \bar{Y}| > 1.96 * 0.81\right) \tag{9}$$

$$= \mathbb{P}\left(\bar{X} - \bar{Y} > 1.58\right) + \mathbb{P}\left(\bar{X} - \bar{Y} < -1.58\right) \tag{10}$$

$$= \mathbb{P}\left(\frac{\bar{X} - \bar{Y} - 0.5}{0.81} > 1.34\right) + \mathbb{P}\left(\frac{\bar{X} - \bar{Y} - 0.5}{0.81} < -2.58\right) = 0.09 \tag{11}$$

That is, the chance that we would reject the null of no difference, given an actual difference of 0.5 is about 1 in 10. This test is not very powerful in detecting difference of 0.5 in the population means. A larger sample size would have given a more powerful test.