Statistical Analysis of Project Tycho's Measles cases in the United States from 1928 to 2002

Rachel Hussmann

Stockton University

CSCI 3327: Probability and Applied Statistics

Professor Byron Hoy

May 1, 2025

Abstract

The purpose of this research was to use statistical analysis methods on instances of the Measles virus contained within Project Tycho's dataset of disease case totals within the United States from 1916 to 2011. The dataset included the total cases for multiple diseases including Hepatitis A, Measles, Mumps, Pertussis, Polio, Rubella, Smallpox and Diphtheria. The data also contained the year and epidemiological week of the count and the state the disease took place in. The methodology of the research included general analysis of the data using standard statistics such as mean and median, statistical distributions and visualization methods for the data. The software used for calculations and visualization of the data included Excel, Desmos, and Python. The research then goes extensively into the results of the statistical analysis. While most of the data was not statistically significant, the main takeaway from the data was that Measles vaccine is effective and continued use of it can prevent the number of cases pre-vaccine.

Introduction

According to the United States Centers for Disease Control and Prevention (CDC), Measles, also known under its scientific name of Rubeola, is an extremely contagious virus that can cause symptoms such as a widespread rash and fever (Centers for Disease Control and Prevention, 2024b). The CDC began mandating the tracking and reporting of Measles cases in 1912, creating a large source of data in the infection rate and spread of the Measles virus (Centers for Disease Control and Prevention, 2024a). After extensive research and testing, in 1963, a vaccine was created and approved for use (Centers for Disease Control and Prevention, 2024a). This vaccine has been used since its creation, although it is typically given in combination with vaccines for other diseases (Centers for Disease Control and Prevention, 2024a). The Measles, Mumps, Rubella (MMR) vaccine is used to reduce the number of needles needed for full immunity (Centers for Disease Control and Prevention, 2024a). Project Tycho, a research database created by researchers at the University of Pittsburg, was created to make case data of many different diseases available to the general public (Tannery, 2014). The goal of this research is to analyze the Measles data contained within Project Tycho's database of disease cases in the United States from 1916 to 2011 to report any statistically significant findings.

Description of Dataset

The dataset that was used for this research is the Project Tycho Level 1 dataset of counts of multiple diseases in the United States from 1916 to 2011 (Van Panhuis et al., 2018). The diseases included in the dataset were Hepatitis A, Measles, Mumps, Pertussis, Polio, Rubella, Smallpox and Diphtheria. The dataset had a total of seven columns, with the respective titles: epi_week, state, loc, loc_type, disease, cases, and incidence_per_100000. After receiving the data, some preprocessing, using a Python script and the Python pandas library, was completed to assist with organization of data and the statistical analysis of the data. The only preprocessing that was completed was to omit the diseases that were not being analyzed. So, after using the Python script, only the data associated with Measles cases was left behind in a .CSV file. The actual reported data that was left behind after processing was unaltered. With the data left behind, there was a total of 75 years' worth of Measles data from most, if not all, of the states in the United States.

Methodology

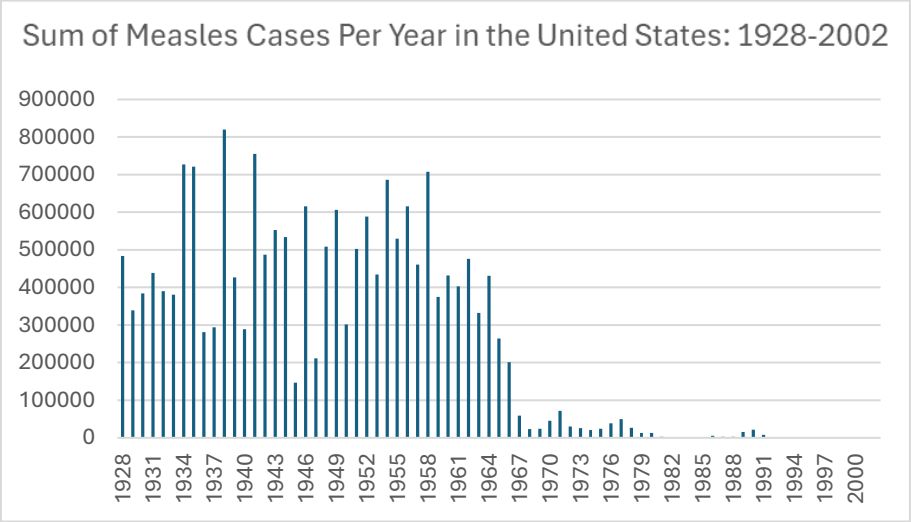
This analysis was mainly for an exploratory purpose. Its goal was to analyze Measles data from the Project Tycho database and report any interesting or statistically significant findings. The main parts of the methodology include:

- General Analysis: The main statistics, such as the mean, median, variance and standard deviation were calculated per year and per state to get a basic understanding of the data.
- Statistical Distributions: Distributions for discrete random variables, continuous random variables, and multivariate analysis were used to get the most statistical data and the clearest picture of how the Measles virus affected the United States between 1928 to 2002.
- Visual Representation: Graphs using Excel and Desmos were created to create clear visualizations of trends and patterns for ease of analysis.

The software used for visualization and analysis were Desmos and Excel. Software used for calculations with large numbers was completed using Python.

Results

To begin with the analysis, the total number of cases per year had to be calculated. After completing this calculation, the following histogram was created.



Based on the histogram, it can be seen that the data is skewed right, based on the considerable number of cases at the beginning of the graph that eventually taper off toward zero. This is as expected for the number of Measles cases within the United States due to vaccine creation in 1963 and continued efforts in vaccinating the United States population after that point.

Synthesizing the data given in the dataset, the mean (μ), median (M), variance (σ^2) and standard deviation (σ) were calculated.

$$\mu = 248{,}948.8133 \text{ cases}$$

$$M = 201{,}064 \text{ cases}$$

$$\sigma^2 = 64{,}322{,}641{,}725.592 \text{ cases}^2$$

$$\sigma = 253{,}619.0879 \text{ cases}$$

Analyzing the statistics calculated shows that the data has an exceptionally large variance, which means that the data has a large spread of values. This is expected due to various outbreaks before vaccine creation and the eventual adoption of the vaccine, leading total cases to plummet significantly. The original statement that the histogram is skewed right is confirmed by the data due to the mean being greater than the median.

To further complete more statistical analysis, the sample space for the dataset was created. It was found that within the dataset there are 3,272 data points for each state and each year from 1928 to 2002. This is lower than expected because, due to the mn rule, the total number of events in the sample space should be 3,750, as there are 50 states that need to be combined with every year between 1928 to 2002, which is 75 years. This means that the dataset does not have all the years and states accounted for within the context of Measles data. Considering the age of this data, this is as expected.

The median number of cases was found to be 201,064 cases. This means that half of the years are below the median and half of the years are above the median. Supposing that four years were randomly chosen from the dataset, the following events were defined:

A: At least two of the years chosen had more than the median

B: Exactly two years chosen had more than the median

C: Exactly one year chosen had less than the median

The sample space was found for this problem (H is a year selected that was above the median, L is a year selected that was below the median):

$$S = \begin{Bmatrix} (HHHH), (HHHL), (HHLH), (HLHH), \\ (LHHH), (HHLL), (HLHL), (LHHL), \\ (HLLH), LHLH), (LLHH), (HLLL), \\ (LHLL), (LLHL), (LLLH), (LLLL) \end{Bmatrix}$$

Using this sample space, the number of simple events within the compound events A, B and C were calculated, which were then used to calculate the probability of the compound event occurring.

$$A = 11 \text{ outcomes} \qquad P(A) = \frac{11}{16} = 0.6875 = 68.75\%$$

$$B = 6 \text{ outcomes} \qquad P(B) = \frac{6}{16} = 0.375 = 37.5\%$$

$$C = 4 \text{ outcomes} \qquad P(C) = \frac{4}{16} = 0.25 = 25\%$$

The probability of choosing at least two years with Measles case counts greater than the median is about 69%. The chance of choosing exactly two years with Measles case counts greater than

the median is about 38% and the probability of choosing exactly one year with Measles case counts lower than the median is 25%.

As referenced above, the Measles vaccine was created in 1963. The purpose of this section was to calculate how many different combinations of five years can be made using a subset of ten years, where five were before the creation of the vaccine and five were after the creation of the vaccine.

$$C(10,5) = \frac{10!}{5!\,(10-5)!} = 252$$

By using the combination formula, it is calculated that there are 252 ways to pick five years from a subset of ten years.

For the next section of analysis, probabilities associated with Measles cases per state needed to be found. To begin this process, the mean and median of total cases per state were calculated.

$$\mu = 372{,}158.32 \text{ cases}$$

$$M = 213{,}414.5 \text{ cases}$$

Of the 50 states in the United States, 25 of them had a total number of Measles cases above the median, while 25 of the states had a total number of Measles cases below the median. Out of the 50 states, nine of them are in what is considered the Northeast. Of the states in the Northeast, five had a total number of cases above the median. The goals for this section were to find the probability that a randomly selected Northeastern state had a total number of Measles cases above the median and the probability that a randomly selected state came from the Northeast.

$$P(Above\ median|Northeast) = \frac{Number\ of\ states\ in\ the\ Northeast\ AND\ cases\ above\ median}{Total\ number\ of\ states\ in\ the\ Northeast} = \frac{5}{9}$$

$$P(Northeast) = \frac{Northeast\ states}{All\ states} = \frac{9}{50}$$

Therefore, the probability that a randomly selected state had a total number of cases above the median was a conditional probability problem, resulting in a probability of about 55.56%. The probability that a randomly selected state was from the Northeast was a standard probability problem, resulting in a chance of about 18%.

Continuing the analysis of cases by state, the events A and B are defined as:

A: A randomly selected state has a total number of cases greater than the mean

B: A randomly selected state has a total number of cases greater than the median

The probability of the events are as follows:

$$P(A) = 0.28$$

$$P(B) = 0.5$$

28% of the states had a total number of cases greater than the mean and 50% of the states had a total number of cases greater than the median. The goal was to find the probability of A or B occurring at the same time. Using Excel and the dataset, the intersection of A and B was found.

$$P(A \cap B) = 0.28$$

28% of the states had a total number of Measles cases above the mean and median. To find the union of these probabilities, the Additive Rule for Probabilities was used:

$$P(A \cup B) = 0.28 + 0.5 - 0.28 = 0.5$$

Therefore, there is a 50% chance that any randomly selected state would have a total number of cases above the mean or above the median.

Based on the dataset, 45.3% of the years are before 1963 and 54.6% of the years are from 1963 and onwards. A year is considered to have a high number of cases if the case count was greater than the mean of 248,948.8133 cases. It was found that of the years before 1963, 94% had a high number of cases, while in the years after 1963, only about 8% had a high number of cases. The aim of this section was to find the probability that the year was before 1963 given that the year had a high number of cases. To find this, Bayes' theorem was applied. Event A was defined as the probability that the year had a high number of cases and event B as the probability of the year being before 1963.

$$P(B) = 0.453$$

$$P(\neg B) = 0.546$$

$$P(A|B) = 0.94$$

$$P(A|\neg B) = 0.08$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) \cdot P(A|\neg B)P(\neg B)} = \frac{(0.94)(0.453)}{(0.94)(0.453) + (0.08)(0.546)} = \frac{0.42582}{0.4695}$$

$$P(B|A) = 0.90696 = 90.696\%$$

Therefore, there is a 90.696% chance that given a chosen year had a high number of cases, it was a year before 1963.

Using discrete random variables, such as Y, introduces more information to the statistical analysis. Now, the dataset can be evaluated using specific distributions such as binomial, geometric, negative binomial, hypergeometric and Poisson. From a binomial standpoint, information that has only two outcomes can be calculated. For the purposes of this research, high and low case years are defined as the two outcomes. The dataset has a total of 75 years. A year is defined as having a high number of cases when the total number of cases for the year is greater

than the mean of 248,948.8133 cases. Using Excel, 36 of the 75 years are considered to have a high number of cases. Let Y denote the number of high case years that are randomly chosen. If three of the years were chosen at random, replacing them after they have been chosen, what would the probability distribution look like for Y? Using a binomial distribution, with the following variables results in the following probabilities for Y.

$$n = 3, p = \frac{36}{75} = 0.48, Y = 0,1,2,3$$

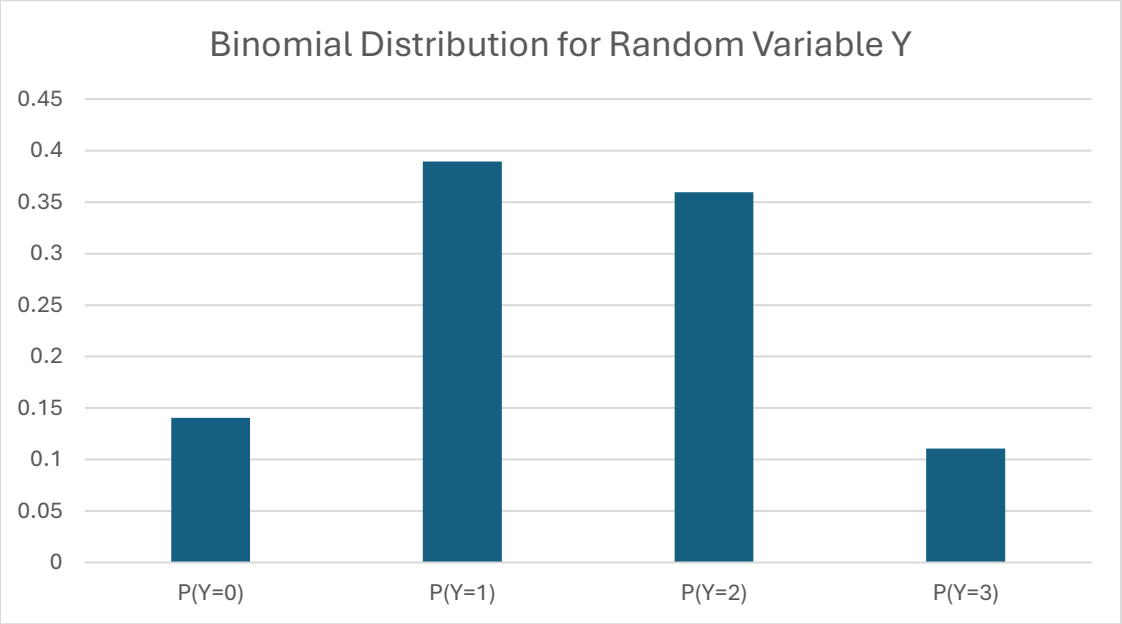
$$p(0) = \binom{3}{0} 0.48^0 0.52^3 = 0.1406$$

$$p(1) = \binom{3}{1} 0.48^1 0.52^2 = 0.3894$$

$$p(2) = \binom{3}{2} 0.48^2 0.52^1 = 0.3594$$

$$p(3) = \binom{3}{3} 0.48^3 0.52^0 = 0.1106$$

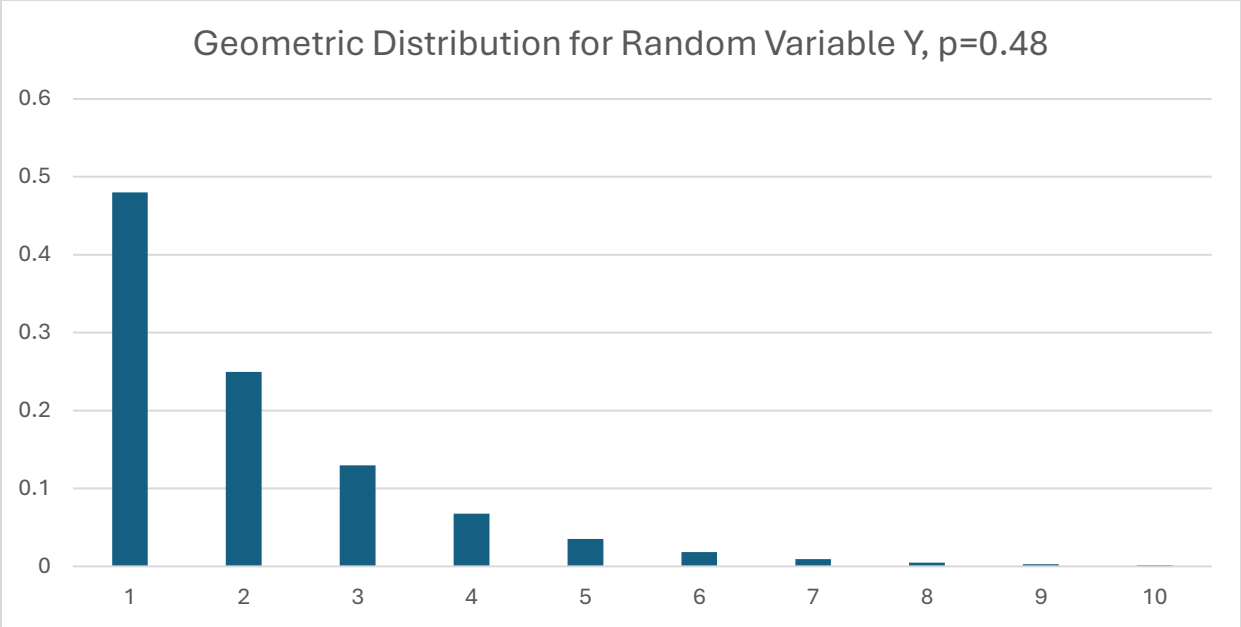
With these probabilities, the histogram for the binomial distribution of random variable Y can be created.



The calculated mean for this distribution is $\mu = 1.44$ years. This means that on average, when choosing three years from the dataset at random, at least one year will be considered a high case year.

Using a geometric distribution gives additional information about the dataset. Let Y be the number of high case years when randomly choosing a year. What would the probability of Y

be if the high case year appeared on the 1st, 2nd, 3rd, . . ., nth trial? Using a geometric distribution with the variables $p = 0.48, Y = 0, 1, 2, 3, \ldots, n$ creates graphable information, displayed in the histogram referenced below.



The mean calculated for this distribution is $\mu = 2.08$ years. This means that on average, the first high case year will appear on the second trial.

Using the negative binomial distribution allows for the calculation of probabilities for a specific number of trials and successes. As previously explained, the probability of picking a high case year is $p = 0.48$. Due to the length of calculating a negative binomial distribution, a specific probability will be evaluated instead. Calculating the probability that the third high year is chosen after eight total years is completed as follows:

$$p(8) = \binom{7}{2} 0.48^3 0.52^5 = 0.0883 = 8.83\%$$

The calculated mean for this distribution is $\mu = 6.25$ years. This means that if we wanted three high years, on average, we would have to pick more than six years from the dataset to get those years.

Using the hypergeometric distribution allows for the probability of Y to be calculated when the years are chosen without replacement. Let $r = 3$ years chosen from the entire dataset of 75 years. Let the random variable Y be the number of high case years. The probability distribution for Y is as follows:

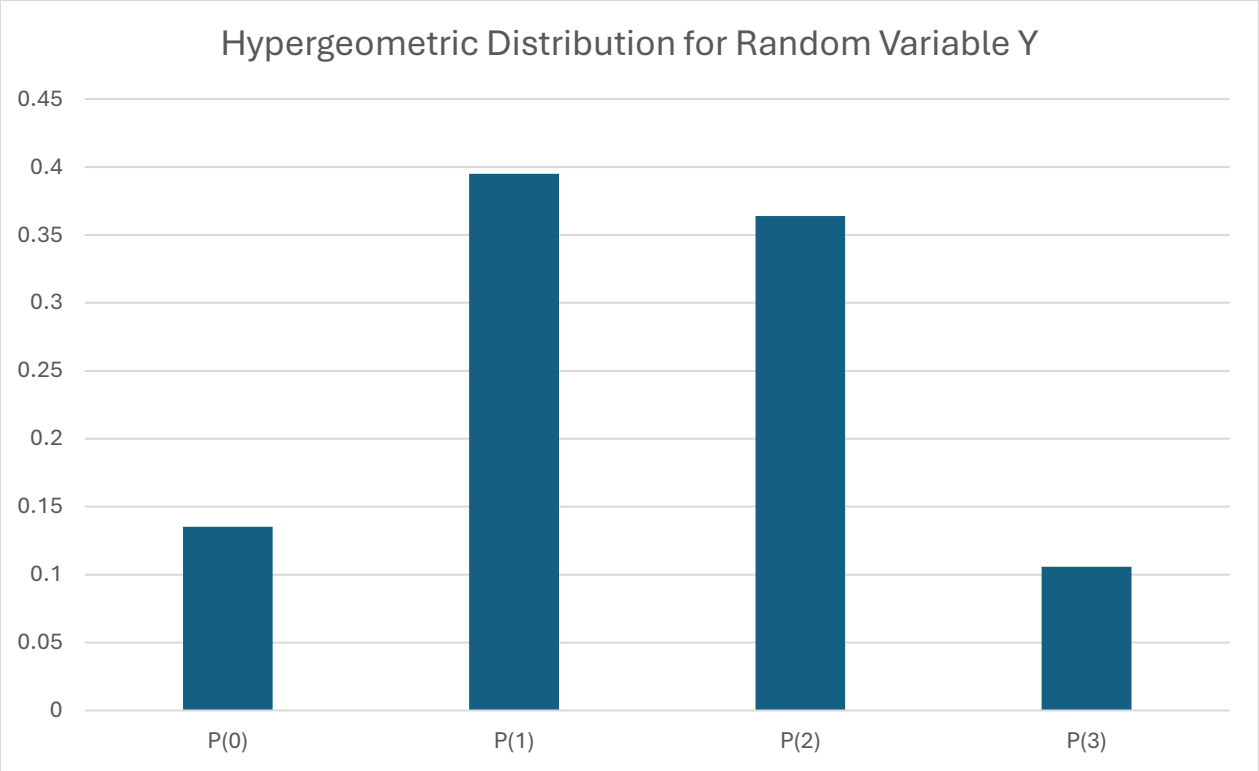
$$p(0) = \frac{\binom{36}{0}\binom{39}{3}}{\binom{75}{3}} = \frac{1 \cdot 9139}{67525} = 0.1353 = 13.53\%$$

$$p(1) = \frac{\binom{36}{1}\binom{39}{2}}{\binom{75}{3}} = \frac{36 \cdot 741}{67525} = 0.3951 = 39.51\%$$

$$p(2) = \frac{\binom{36}{2}\binom{39}{1}}{\binom{75}{3}} = \frac{630 \cdot 39}{67525} = 0.3639 = 36.39\%$$

$$p(3) = \frac{\binom{36}{3}\binom{39}{0}}{\binom{75}{3}} = \frac{7140 \cdot 1}{67525} = 0.1057 = 10.57\%$$

Using these values to create a histogram creates this visual.



The mean of this distribution is calculated to be $\mu = 1.44$ years. This means that on average, when choosing three years from the entire dataset, at least one year will be a high case year.

To use a Poisson distribution for this dataset, it is assumed that the mean and the variance are the same. For this case, the mean will be used as the lambda (λ). Once again, due to the complexity of a Poisson distribution, this research will only cover a specific value of Y. Let $\lambda = 248{,}948.8133$ cases. What is the probability of Y being greater than 250,000? These calculations involve exceptionally large numbers that make calculating by hand difficult. To prevent errors in calculations, a Python script was used to calculate the following answer:

$$P(Y > 250{,}000) = 1 - P(Y \leq 250{,}000) = 0.0176 = 1.76\%$$

This means that picking a random year from the dataset, there is 1.76% chance that the year will have a total number of cases greater than 250,000. This probability is extremely small compared to the size of the numbers that the dataset has given. This means that small fluctuations in large numbers can make a significant difference in the probability of an event.

Using Tchebysheff's theorem, the interval for an amount of data can be found. Let 75% be the percentage of data to find the interval for. Knowing that two standard deviations worth of data contains at least 75% of the data creates the following interval:

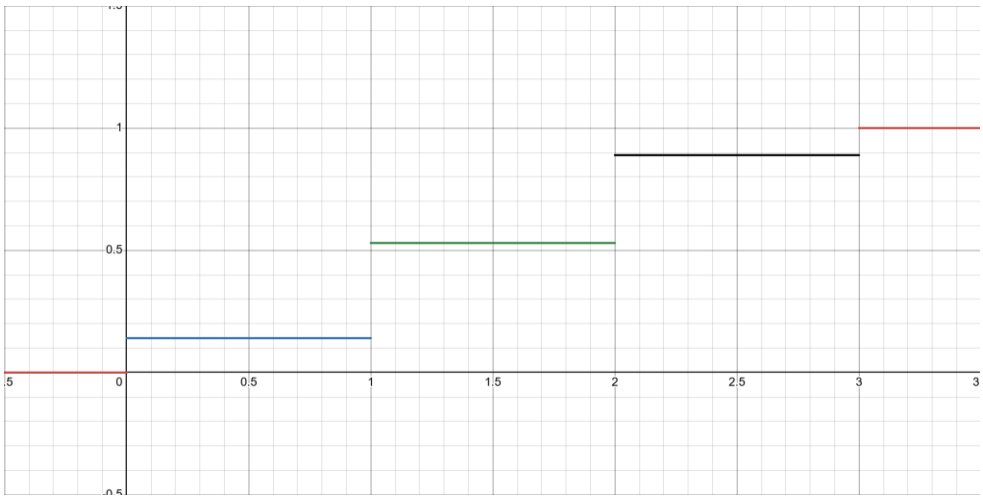
$$\mu \pm 2\sigma = 248{,}948.8133 \pm 507{,}238.1758 = (-258{,}289.3625, 756{,}186.9891)$$

Considering the bounds of the dataset, this interval can be adjusted to $(0, 756{,}186.9891)$. Therefore, at least 75% of the data is contained between the intervals of 0 and 756,186.9891 cases.

Cumulative measures, such as cumulative probability functions, uniform distributions, and the exponential distribution can display information about how the data accumulates over multiple points. A discrete probability function can be converted into a cumulative distribution function by adding up the values as they grow. This can be shown more efficiently using a piecewise function. The data from the binomial distribution will be used to create the cumulative probability function.

$$F(y) = P(Y \leq y) = \begin{cases} 0, & for\ y < 0 \\ 0.1406, for\ 0 \leq y < 1 \\ 0.53, for\ 1 \leq y < 2 \\ 0.8894, for\ 2 \leq y < 3 \\ 1, for\ y > 3 \end{cases}$$

Using Desmos, the cumulative probability function can be visualized.



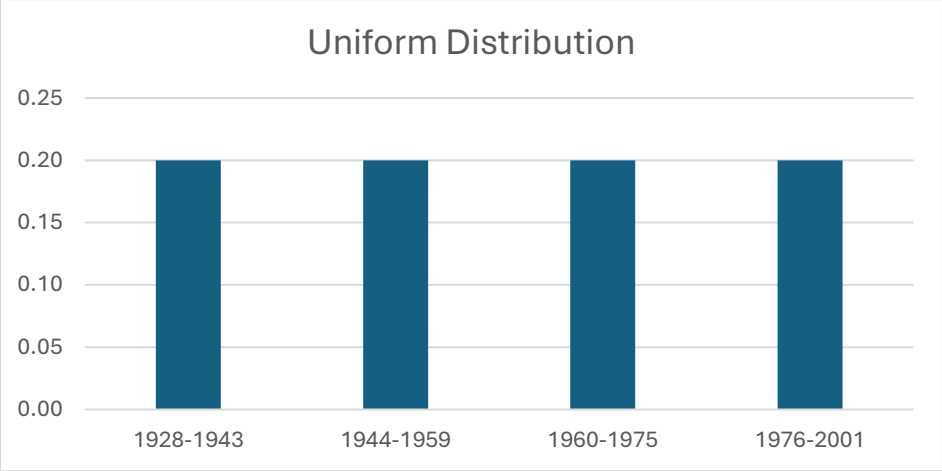
The bounds of the function can clearly be seen, going from 0 to 1, showing how calculus could be used to find the probability under the curve. Because this is a discrete cumulative function,

using calculus would be excessive, but it can be done, and calculus is helpful when variables are added to the probability functions.

One of the distributions gained from this new knowledge is the uniform distribution. The uniform distribution is where the area under the curve is the same across all values of Y. An example from the dataset is the probability of each year. There are 75 years in the dataset, each with a probability of $p = \frac{1}{75}$. The goal of this section is to find the probability of choosing a year between 1928 and 1950. The complete formula would be as follows:

$$P(1928 < Y < 1942) = \frac{1943 - 1928}{2003 - 1928} = \frac{15}{75} = \frac{1}{5}$$

Because this is based on a uniform distribution, this means that for every possible set of 15 years that could be chosen, the probability would be the same.

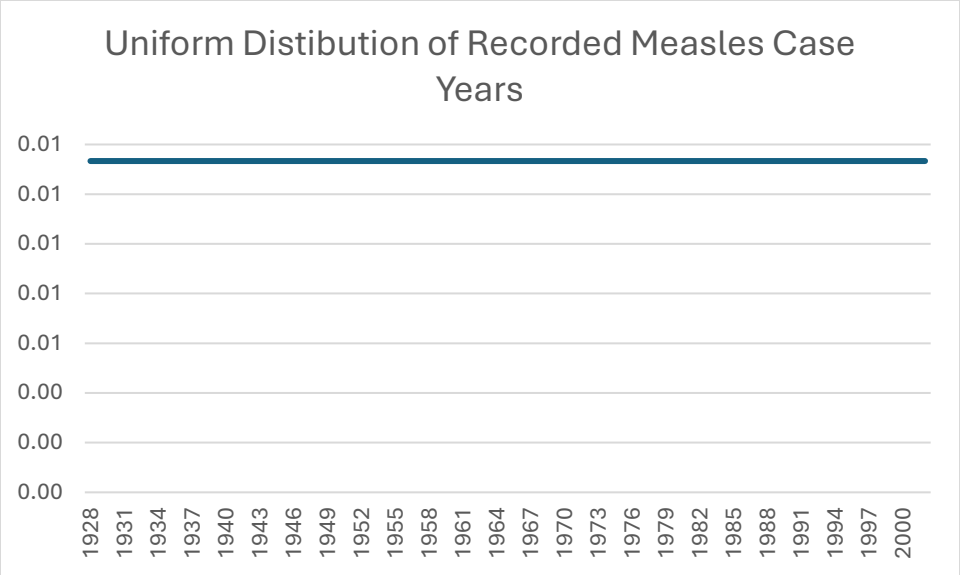


The mean and variance associated with this uniform distribution is as follows:

$$\mu = \frac{1928 + 2003}{2} = 1965.5$$

$$\sigma^2 = \frac{(2003 - 1928)^2}{12} = 469.75$$

This means that the average year for the uniform distribution is between 1965 and 1966, which given the distribution, should be directly in the middle of the graph. This is shown better in the graph where each year is given its respective probability.



Now it can be seen that the value right in the middle of the graph is 1965.5. The variance does not mean much here, but once it is turned into the standard deviation for the distribution, it is more meaningful.

$$\sigma = \sqrt{469.75} = 21.67 \text{ years}$$

This number makes sense in the context of the graph. So, the average deviation from the mean is 21.67 years, which makes sense given the substantial number of years given for the dataset.

Another meaningful distribution for this dataset is the exponential distribution. Certain parts of the data do follow an approximate exponential curve, so valuable information can be derived from them. Specifically, between the years of 1932 and 1934, the total number of Measles cases follow an approximate exponential distribution with a mean of 499,255.67 cases. To find the probability of Y being greater than 500,000, the formula for an exponential distribution would be used.

$$P(Y > 500{,}000) = \int_{500000}^{\infty} \frac{1}{499255.67} e^{-\frac{y}{499255.67}} = e^{-\frac{500000}{499255.67}} = .3673 = 36.73\%$$

This means that for the actual data given by Project Tycho, the probability of the year having more than 500,000 Measles cases is 36.73%. In terms of virus spread, that is a high percentage. This means that during that time, Measles was continually active and hard to control.

Statistics also has the ability to calculate functions with more than one variable. This is called multivariate analysis. In the case of this dataset, the two variables that are given are the years of data and the states. This allows for calculations of the probability of outbreaks per state to be performed. Three of the states, New Jersey, Ohio and Massachusetts, have an approximately equal chance of an outbreak from 1928 to 2002. This is defined by the fact that each state's average number of cases were similar or the same as the other state's average

number of cases. New Jersey's average was $\mu = 13{,}044.70833$ cases, Ohio's average was $\mu = 11{,}972.0411$ cases and Massachusetts' average was $\mu = 11{,}037.44595$ cases. The states' similar averages demonstrate that each have similar values, which in turn concludes that they have a similar probability for an outbreak. For the calculations, let each state have a chance to possess either zero, one, two or three outbreaks. Let X be the number of outbreaks in New Jersey and Y be the number of outbreaks in Ohio. The joint probability function for X and Y is referenced below.

		X			
		0	1	2	3
Y	0	$\frac{1}{27}$	$\frac{3}{27}$	$\frac{3}{27}$	$\frac{1}{27}$
	1	$\frac{3}{27}$	$\frac{6}{27}$	$\frac{3}{27}$	0
	2	$\frac{3}{27}$	$\frac{3}{27}$	0	0
	3	$\frac{1}{25}$	0	0	0

From the table referenced above, the marginal probability functions for X and Y can also be found by adding up the columns for X and the rows for Y.

X

$P(0)$	$P(1)$	$P(2)$	$P(3)$
$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$

Y

$P(0)$	$P(1)$	$P(2)$	$P(3)$
$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$

Based on these marginal distributions, it can also be concluded that X and Y are dependent variables. Based on the formula for independent variables, $p(x, y) = p(x)p(y)$, when values from the joint probability and marginal distributions are substituted, the values fail.

$$p(0,0) = p(0)p(0)$$

$$\frac{1}{27} = \frac{8}{27} \cdot \frac{8}{27}$$

$$\frac{1}{27} \neq \frac{64}{729}$$

This means that X and Y are dependent upon each other.

Conclusion

Most of the data contained within this statistical analysis was either expected based on the history of the Measles virus or was not statistically significant. Regardless of this, the analysis completed the purpose of the research, which was to explore Project Tycho's dataset of case totals for the Measles virus in the United States from 1928 to 2002 and report if there was any significant data. The only significant discovery of this statistical analysis was that the Measles vaccine does help to prevent people from becoming infected with the Measles virus and prevents its spread. The vaccine should be continued to be used to prevent infection rates from pre-vaccine United States.

References

Centers for Disease Control and Prevention. (2024a, May 9). *Measles history*. https://www.cdc.gov/measles/about/history.html

Centers for Disease Control and Prevention. (2024b, May 9). *Measles symptoms and complications*. https://www.cdc.gov/measles/signs-symptoms/index.html

Tannery, N. (2014, January). *Project Tycho™: Public health data to help fight deadly contagious diseases – HSLS Update*. University of Pittsburgh Health Sciences Library System. https://info.hsls.pitt.edu/updatereport/2014/january-2014/project-tycho%E2%84%A2-public-health-data-to-help-fight-deadly-contagious-diseases/

Van Panhuis, W., Cross, A., & Burke, D. (2018). *Project Tycho Level 1 data: Counts of multiple diseases reported in United States of America, 1916-2011* (Version 1.0.0) [Data set]. Project Tycho. https://doi.org/10.5281/zenodo.12608992