2013

Actuarial Mathematics with Applications

Ryan Varnals
University of Redlands

ACTUARIAL MATHEMATICS CAPSTONE

By Ryan Varnals

INTRODUCTION

What is the probability of a person dying? How can these calculations be made? These are all questions that Actuaries seek to answer. Actuaries are mathematicians that calculate the finance behind risk and uncertainty. They calculate the probabilities of undesirable events in order to minimize the financial cost of these undesirable events. Examples of undesirable events are car accidents, the death of a person, earthquakes, and tornadoes. Most of these events destroy either buildings or cars, which is why actuaries are usually employed by insurance companies. Actuaries are also employed by health care companies, investing companies, and even banking companies. An undesirable event for an investing company could be a big drop in stock prices, and an actuary would use probability to minimize the financial cost of this. In order to understand how an actuary calculates these probabilities, a few terms must be defined.

DEFINITIONS

The age of a person is important to determine the probability of surviving or dying. The age of any give person is referred to as x. T_x is the random variable for calculating the age-at-death for a person. For example, the probability that a sixty year old person dies within a year of turning sixty is expressed as the following probability.

$$P(T_{60} \leq 1)$$

The lifetime distribution $F_x(t)$ is the probability density distribution of the mortality of a person. The t is the length of time after the person's current age x. The lifetime distribution is defined as the below equation.

$$F_x(t) = P(T_x \leq t)$$

This equation is stating the probability that the age-at-death of a person at the age of x is less than or equal to t. The actuarial notation for $F_x(t)$ is ${}_tq_x$, so I will use the actuarial notation from this point on. The survival distribution $S_x(t)$ is the probability density function of the survivability of a person. The survival distribution is defined as the below equation.

$$S_x(t) = P(T_x > t)$$

This equation is stating the probability that the age-at-death of a person at the age of x is greater than t. Since the age-at-death is either equal to t, less than t, or greater than t, then the following must hold true for probabilities.

$$P(T_x \leq t) + P(T_x > t) = 1$$

$$P(T_x \leq t) = 1 - P(T_x > t)$$

$$F_x(t) = 1 - S_x(t)$$

The actuarial notation for $S_x(t)$ is ${}_tp_x$, so this will be used from this point on for the survival distribution.[1] The force of mortality μ_x is the instantaneous rate of change of the fatality probability function, or the rate of failure. The force of mortality is defined as:

$$\mu_x = \lim_{dx \to 0^+} \frac{1}{dx} [{}_{dx}q_x]$$

The limit is only in the positive direction. This is because time cannot be negative. As dx goes to zero, the μ_x becomes the instantaneous rate of change. This dx is thought to be extremely small – for example, $dx = 0.00274$ years is very small, and 0.00274 years is equivalent to 1 day.

There are two aspects to actuarial mathematics. There is the statistic and stochastic side, and there's the financial mathematics side. The financial mathematics aspect is related to economics, while the stochastic side related to mathematics and statistics. For this reason, the stochastic processes involved with actuarial mathematics will be the only focus of this paper.

ASSUMPTIONS

In order to do most of the calculations, some assumptions must be made:

1. If time has not progressed, a person cannot die. Then, we assume ${}_0p_x = 1 = P(T_x > 0)$.
2. Since everyone will die some day, we assume $\lim_{t\to\infty} {}_tp_x = 0$.
3. If $t_1 < t_2$, then ${}_{t1}p_x \geq {}_{t2}p_x$. The survival probability function is non-increasing. The probability of a person at the age of x surviving 10 years cannot be higher than the probability of that same person surviving 5 years because in order for that person to survive 10 years, they must survive 5 years.
4. $\frac{d}{dt}[\,{}_tp_x]$ exists $\forall t > 0$. This ensures the survival distribution function exists.
5. To ensure the expected value exists, then $E(T_x)$ is defined as $E(T_x) = \int_0^\infty {}_tp_x \cdot dt$. This is shown using the definition of an expected value.

$$E(Y) = \int_0^\infty y \cdot f(y)dy$$

By definition of an integral, $y = \int_0^y 1 \cdot dt$. So using substitution and changing the order of integration:

$$E(Y) = \int_0^\infty \int_0^y 1 \cdot f(y)\, dt\, dy = \int_0^\infty \int_t^\infty f(y)\, dy\, dt = \int_0^\infty P(T \geq t)\, dt$$

Since ${}_tp_x = P(T_x \geq t)$, then:

$$E(Y) = \int_0^\infty {}_tp_x \cdot dt$$

SURVIVAL MODELS

An example of three survival functions on the same graph are shown below.[2]

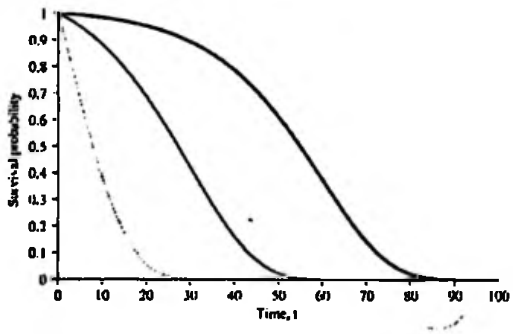
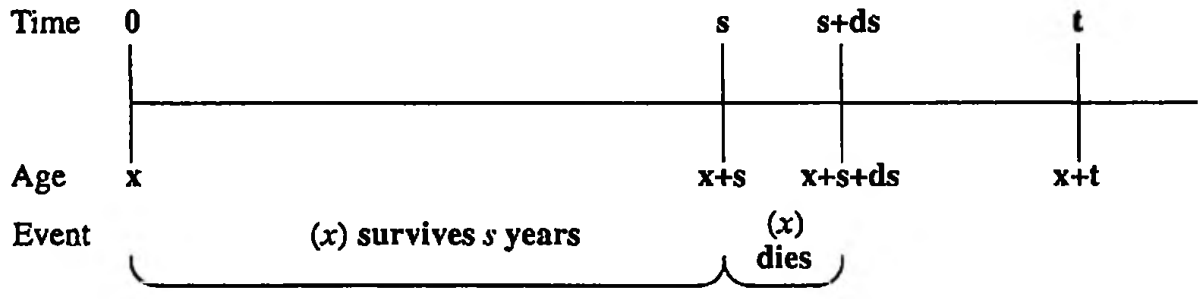


Figure 2.1 $S_x(t)$ for x = 20 (bold), 50 (solid) and 80 (dotted).

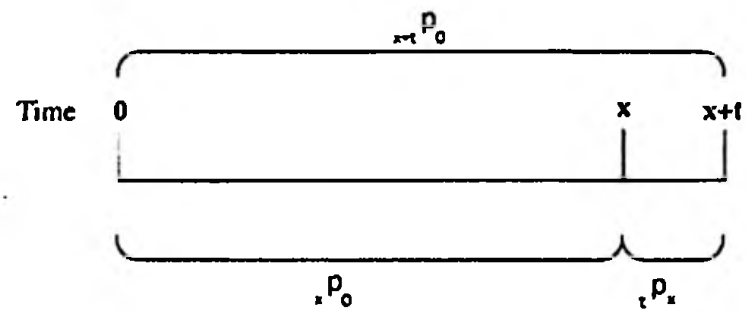
The graph shows the probability function for three different ages. The worst line on the graph is when x = 80, and the probabilities of survival quickly diminish. In the graph, x = 80 is the furthest line to the left, whereas x = 20 is furthest to the right. The furthest left means that x = 80 has the highest probability to survive the least amount of time. The x = 20 graph has the highest probability to survive the longest amount of time. The below graph is a visual representation of times and ages of people using different variables and illustrates what x and t represent in the survival probability function. (Dickson, David, Mary Hardy, and Howard Waters: Page 28)



Some important theorems for survival models are:

- Theorem 3: $${}_tp_x = \frac{{}_{x+t}p_0}{{}_xp_0}$$
- Theorem 5: $${}_{t+u}p_x = {}_tp_x \cdot {}_up_{x+t}$$

Theorem 3 states that the probability of a person aged x years old surviving t years is equal to the probability of that person at birth surviving $x + t$ years divided by the probability of the person at birth surviving x years. This can be seen below.

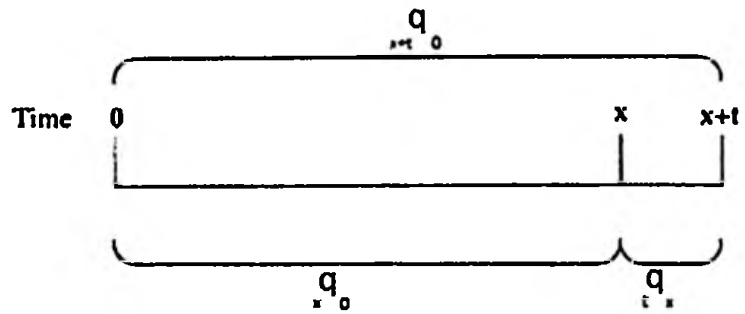


Looking at $\frac{{}_{x+t}p_0}{{}_xp_0}$, the ${}_{x+t}p_0$ is the top probability while ${}_xp_0$ is the bottom left probability. Since most of these two intervals overlap each other, meaning they cancel out. The only portion that does not overlap is ${}_tp_x$. This means ${}_tp_x \cdot {}_xp_0 = {}_{x+t}p_0$ since the two intervals are the same length. This is the same equation as Theorem 3. Theorem 5 is similar to Theorem 3, except the ages for the given person does not necessarily have to be from birth.

An alternate definition given for the lifetime distribution is:

$$ {}_tq_x = \frac{{}_{(x+t)}q_0 - {}_xq_0}{{}_xp_0} $$

This equation is easier to view using the chart below.2



In order for a person to die between x and x+t, the person must have lived past x. So in order to calculate ${}_tq_x$, the ${}_tp_0$ must be calculated also since the person must live past x. Looking at the overlap in the chart, ${}_{(x+t)}q_0 - {}_xq_0$ will be the top interval minus the bottom interval, which leaves the smaller interval on the right. But since ${}_tq_x$ must be calculated with ${}_tp_0$, then:

$$ {}_xp_0 \cdot {}_tq_x = {}_{(x+t)}q_0 - {}_xq_0 $$

This can be rewritten by dividing ${}_xp_0$ on both sides.

$$ {}_tq_x = \frac{{}_{(x+t)}q_0 - {}_xq_0}{{}_xp_0} $$

From this equation and substituting ${}_tq_x = 1 - {}_tp_x$ into every lifetime probability distribution (the 'q' probabilities), then:

$$ 1 - {}_tp_x = \frac{(1 - {}_{(x+t)}p_0) - (1 - {}_xp_0)}{{}_xp_0} = \frac{(-{}_{(x+t)}p_0) + {}_xp_0}{{}_xp_0} $$

$$ 1 - {}_tp_x = \frac{(-{}_{(x+t)}p_0)}{{}_xp_0} + \frac{{}_xp_0}{{}_xp_0} = 1 - \frac{{}_{(x+t)}p_0}{{}_xp_0} $$

$$ - {}_tp_x = -\frac{{}_{(x+t)}p_0}{{}_xp_0} $$

$$ {}_tp_x = \frac{{}_{(x+t)}p_0}{{}_xp_0} $$

This illustrates that the probability of survival at any age can be found out as long as the probability function at birth is known.

The force of mortality is important because it shows the rate of failure. Having a high force of mortality means the chances of survival are decreasing dramatically. Since the force of mortality is calculated using only the lifetime distribution, then we can solve for ${}_tq_x$ to find another way to find the lifetime distribution if given the force of mortality. Using the definition:

$$\mu_x = \lim_{dx\to 0^+} \frac{1}{dx}[{}_{dx}q_x]$$

Using $F_x(t) = 1 - S_x(t)$, then the equation becomes:

$$\mu_x = \lim_{dx\to 0^+} \frac{1}{dx}[1 - {}_{dx}p_x]$$

Substituting Theorem 3, ${}_tp_x = \frac{x+tp_0}{xp_0}$, in for ${}_tp_x$:

$$\mu_x = \lim_{dx\to 0^+} \frac{1}{dx}\left[1 - \frac{x+dxp_0}{xp_0}\right]$$

$$\mu_x = \lim_{dx\to 0^+} \frac{1}{dx}\left[\frac{xp_0 - x+dxp_0}{xp_0}\right]$$

Since the limit is with respect to dx, then $\frac{-1}{xp_0}$ can be factored out of the fraction.

$$\mu_x = \frac{-1}{xp_0} \lim_{dx\to 0^+} \left[\frac{x+dxp_0 - xp_0}{dx}\right]$$

The definition of a derivative is $f'(x) = \lim_{dx\to 0}\left[\frac{f(x+dx)-f(x)}{dx}\right]$, so substituting the derivative for $\lim_{dx\to 0^+}\left[\frac{x+dxp_0 - xp_0}{dx}\right]$ gives the following.

$$\mu_x = \frac{-1}{{}_xp_0} \cdot \frac{d}{dx}[{}_xp_0]$$

$\frac{d}{dx}[\log h(x)] = \frac{1}{h(x)}\frac{d}{dx}[h(x)]$ is true for all functions $h(x) > 0$, so $\log {}_xp_0$ can be substituted in for $\frac{1}{{}_xp_0} \cdot \frac{d}{dx}[{}_xp_0]$.

$$\mu_x = -\frac{d}{dx}[\log {}_xp_0]$$

Since μ_x is probability over time, then integration can be used to find the probability over the time.

Taking the integral yields:

$$\int_0^t \mu_x \, dx = -[\log {}_xp_0] \,|_0^t = -[\log {}_xp_0] \,|_0^t$$

$$\int_0^t \mu_x \, dx = -[\log {}_tp_0 - \log {}_0p_0]$$

Using the assumption that ${}_0p_0 = 1$, then the $\log 1 = 0$,

$$\int_0^t \mu_x \, dx = -\log {}_tp_0$$

$$-\int_0^t \mu_x \, dx = \log {}_tp_0$$

Now raise both sides with base e:

$${}_tp_0 = \exp\{-\int_0^t \mu_x \, dx\}$$

Given the rate of failure, or the force of mortality, the probability of the survival distribution can be calculated. Once the survival distribution is found, the lifetime distribution can be found also.

LIFE TABLES

A life table is a table of values containing life expectancies and mortality probabilities for the different age groups for some population. The life table is expressed as a function, where we must be given ${}_tp_x$ to create the life table. The life table shows the expected number of survivors for some population based on ages. This expected number, denoted as l_x, is defined over $x_0 \leq x \leq \omega$, where x_0 is the initial age and ω is the maximum age. We let l_{x_0} be some arbitrary positive number, referred to as the Radix of the table, that represents the number of survivors at the initial age. For $0 \leq s \leq \omega - x_0$:

Definition 1.1: $l_{x_0+s} = l_{x_0} \cdot {}_sp_{x_0}$

This equation gives a way to find out the expected number of people in an age group. Since $0 \leq t$, then $x_0 \leq x \leq x + t \leq \omega$ must be true. x_0 doesn't have to be 0 since it's the starting age. We want to find l_x for any age, not just for l_{x_0}, so substitute s from the top equation with $x + t - x_0$ to get:

$$l_{x_0+x+t-x_0} = l_{x_0} \cdot {}_{(x+t-x_0)}p_{x_0}$$

$$l_{x+t} = l_{x_0} \cdot {}_{(x+t-x_0)}p_{x_0}$$

Then the equation below can be applied to the above equation.

$${}_{v+u}p_w = {}_vp_w \cdot {}_up_{w+v}$$

In the life table equation, we want to use this equation, where $= x - x_0$, $u = t$, and $w = x_0$. So substituting into the definition produces

$$(x-x_0)+t p_{x_0} = (x-x_0)p_{x_0} \cdot \ {}_t p_{x_0+(x-x_0)}$$

$$x+t-x_0 p_{x_0} = x-x_0 p_{x_0} \cdot \ {}_t p_x$$

Now substituting ${}_{x-x_0}p_{x_0} \cdot \ {}_t p_x$ into the life table equation for ${}_{x+t-x_0}p_{x_0}$ yields:

Equation 1.0: $l_{x+t} = l_{x_0} \cdot {}_{x-x_0}p_{x_0} \cdot \ {}_t p_x$

Using the definition of life tables but substituting $= x - x_0$, the below equation is produced

$$l_{x_0+(x-x_0)} = l_{x_0} \cdot {}_{(x-x_0)}p_{x_0}$$

Equation 1.1: $l_x = l_{x_0} \cdot {}_{x-x_0}p_{x_0}$

Substituting l_x from equation 1.1 for $l_{x_0} \cdot {}_{x-x_0}p_{x_0}$ in equation 1.0 constructs the following equation.

$$l_{x+t} = l_x \cdot \ {}_t p_x$$

$${}_t p_x = \frac{l_{x+t}}{l_x}$$

As shown, the probability ${}_t p_x$ is a ratio of the expected number of survivors at age x+t over the expected number of survivors at age x.

The life table is a collection of data for different age groups. This information is the expected number of survivors and the difference between survivors from one age group to the next group. Below is the calculation for this difference.

$$d_x = l_x - l_{x+1}$$

This difference calculates the number of people that didn't survive, so we can rewrite this as:

$$d_x = l_x\left(1 - \frac{l_{x+1}}{l_x}\right) = l_x(1 - {}_1p_x) = l_x \cdot {}_1q_x$$

This equation provides a simple way to find the mortality probability if given a life table that does not have ${}_1q_x$. Once ${}_1q_x$ is known, then ${}_1p_x$ can be calculated along with the force of mortality. Life tables give an enormous amount of information, so creating life tables are crucial for actuaries. Since making life tables yield so much information, I will develop my own life table.

LESLIE MATRICES

The Leslie Matrix is important for calculating populations because it includes the fertility rates, or the births of a population. This is an important part of populations, and should not be left out. The way to compute a population can be written in $m + 1$ equations, where the final age group is from age $m - 1$ to m. Some definitions that Leslie provided are listed below. [3]

- $n_{x,t}$ is the number of people expected to be alive in the age group x to x+1 at time t for some population.
- P_x is the probability of a person of age x to x+1 at time t being alive in the age group from x+1 to x+2 at time t+1.
- F_x is the number of people born from t to t+1 per person aged x to x+1 at time t. These newborn's must be alive in the group 0 to 1 for time t+1.

The P_x is the same probability used for life tables and survival models, which is ${}_1p_x$. The $n_{x,t}$ is similar to l_x, except there is another variable 't' that is included. The age distribution from the beginning of time is given by

$$\sum_{x=0}^{m} F_x n_{x,0} = n_{0,1}$$

$$P_0 n_{0,0} = n_{1,1}$$

$$P_1 n_{1,0} = n_{2,1}$$

$$P_2 n_{2,0} = n_{3,1}$$

$$\vdots$$

$$P_{m-1} n_{m-1,0} = n_{m1}$$

The top equation is the number of people born per person in each age group multiplied by the number of people. This yields the number of people in age group 0 after 1 year [n_{01}]. The rest of the equations are the probability for a person at an age of x surviving, multiplied by the number of people at age x for time t. This yields the number of people at age x+1 for time t+1.

These equations can be rewritten by grouping all of the $n_{x,0}$'s into a single vector of length m+1, denoted as $\overrightarrow{n_0}$. This leaves the rest of the values to be put into a matrix. The matrix is labeled M, and there are m+1 rows and m+1 columns,

$$M = \begin{bmatrix} F_0 & F_1 & F_2 & & F_{m-2} & F_{m-1} & F_m \\ P_0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & P_1 & 0 & & 0 & 0 & 0 \\ & \vdots & & \ddots & & \vdots & \\ 0 & 0 & 0 & & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & P_{m-2} & 0 & 0 \\ 0 & 0 & 0 & & 0 & P_{m-1} & 0 \end{bmatrix}$$

This is the Leslie Matrix. Notice that if $F_m = 0$, then M is singular because the det(M) = 0. This means that M has no inverse. Using this vector form, the age distribution using the Leslie Matrix is:

$$M\overrightarrow{n_0} = \overrightarrow{n_1}$$

The $\overrightarrow{n_1}$ is generated from combining the $n_{x,1}$ values together into a vector. If the number of people expected to be alive after two years was to be calculated, then this is done using the following equation ($\overrightarrow{n_2}$ is the vector for all of the $n_{x,2}$ values).

$$M^2\overrightarrow{n_0} = M\overrightarrow{n_1} = \overrightarrow{n_2}$$

The main differences between the Life Table and the Leslie Matrix are:

1. The Leslie Matrix includes birth rates while the Life Table does not.
2. The Life Table has one initial age group x_0 where the Leslie Matrix has several initial age groups, every element of the vector $\overrightarrow{n_0}$ is an initial age group.

Every step in the Life Table from one line to the next is the same as going from $n_{x,t}$ to $n_{x,t+1}$. Since the Leslie Matrix has several starting age groups, it contains much more information about a population. However, the Leslie Matrix takes up much more space than a Life Table. Also, the Leslie Matrix has many elements that are filled with zeros, making the Leslie Matrix waste space. Given several life tables for different initial age groups and the birth rates for the population, the Leslie Matrix can be created.

DATA

One major problem I had was getting large amounts of data about peoples deaths. I did not want to have to manually find out the ages of each person. This led me to decide to not focus my research on people since the data would be hard to gather in large quantities. This did not prove to be a problem though. On the west coast, buildings fall to major earthquakes (major earthquakes are magnitude six or higher). On the east coast, buildings fall to hurricanes and tornadoes. Vehicles are at risk every day of being totaled by other drivers. Insurance can also be applied to any insurable

object, such as home insurance being applied to houses, so I decided to focus on buildings in California and how earthquakes affect California buildings. I acquired data for earthquakes of magnitude six and higher, noting the time of the earthquake, the exact magnitude, and the location of the epicenter of the earthquake.[4] I entered this data into Excel, then looked at ways to measure times between earthquakes so I can apply earthquake damage toward the life span of buildings. Calculating the survivability of buildings will include the use of what I have researched.

Analysis of the Data

Since I am applying earthquakes to building damage, I want to calculate the time between earthquakes. Knowing the length of time between earthquakes will give me an understanding of how long the buildings will stand before getting struck by an earthquake. To find the time between earthquakes, the difference for two consecutive earthquakes must be calculated. The data I found had 61 earthquakes with their times that they struck, meaning there will be 60 differences because every difference is calculated from two times. The first step when the data is collected is to find the mean and standard deviation. These measure the center and spread of the data. These calculations are shown below. The mean is calculated by using the following equation.

$$\overline{Y} = \frac{1}{n}\sum_{i=0}^{n} Y_i$$

The variance is calculated using the equation below (Wackerly, Mendenhall, and Scheaffer : Pages 9-10).

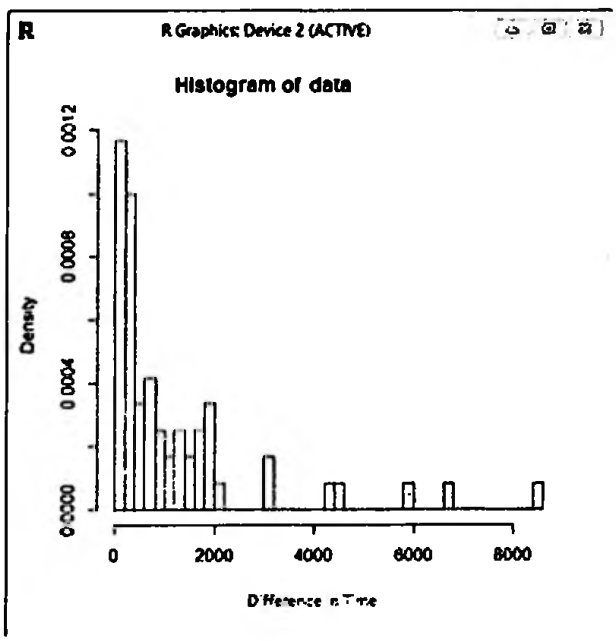
$$s^2 = \frac{1}{n-1}\sum_{i=0}^{n} (Y_i - \overline{Y})^2$$

The standard deviation is simply the square root of the variance. Using these equations, the center and spread is shown below.

Mean	1199.767
Variance	2912958
Std Dev	1706.739

The units for mean and standard deviation are in seconds since the data

A histogram of the data is the next step to get an idea of how the data acts. The histogram will find the number of observations over a bin size, but since this is the density probability histogram, these observations are divided by the total number of observations to give the probability of each bin. There are 50 bins set over about 8500 in this histogram, so each bin has length 170.



The data is skewed right, meaning that most of the data is grouped to the left and there is a tail to the right. This means that most of the data has small values, or small differences in times between earthquakes. Since the data is skewed right, matching the data to a distribution becomes simple.

Distributions for the data

Since the data is skewed right and most of the observations are small numbers, then two distributions seem to fit the data. These are the exponential distribution and the gamma

distribution. Both of these distributions are skewed right like the density of the distribution, which is why I chose them. I have never used the weibull distribution, so this distribution was not considered from lack of experience.

The two distributions that were chosen have the following probability density functions. [5]

Distribution	Probability Density Function	Mean	Variance
Gamma	$f(x) = \left[\frac{1}{\Gamma(\alpha)\cdot\beta^{\alpha}}\right] x^{(\alpha-1)}\cdot e^{-\frac{x}{\beta}}$ $x \in \mathfrak{X}$	$\alpha\cdot\beta$	$\alpha\cdot\beta^2$
Exponential	$f(x) = \frac{1}{\beta}\cdot e^{-\frac{x}{\beta}}$ $x \in \mathfrak{X}$	β	β^2

Notice that the exponential function is a special case of the exponential distribution, where $\alpha = 1$. To plot the gamma distribution, I first need to find the estimators for α and β. This was done using the method of moments. The second moment was calculated using the variance instead of using the second moment, or $E(X^2)$. Since the mean and variance of the data can be calculated, then we can find α and β since we have two equations and two unknowns. β is the quickest to find since

$$\frac{\alpha\beta^2}{\alpha\beta} = \beta = \frac{Variance}{Mean}$$

Then once β is known, α can be found using

$$\alpha = \frac{Mean}{\beta}$$

For the exponential distribution, there is only one unknown. The one unknown, which is β, is used to calculate both the mean and the variance. Since the data has two different values for the mean and standard deviation, then having two different exponential distributions can be used to approximate the data. The two different exponential distributions are found below.

Exponential Distribution 1:$\beta = Mean$

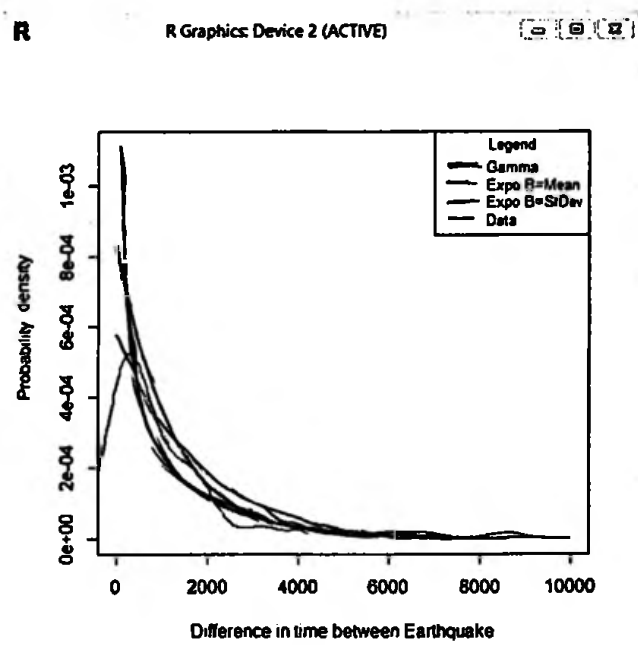
Exponential Distribution 2:$\beta^2 = Variance$

$\beta = Standard\ Deviaton$

Applying the values to their distributions gives the below table.

Distribution	Probability Density Function	Parameters
Gamma	$f(y) = \left[\frac{1}{\Gamma(\frac{1}{2}) \cdot (2400)^{\frac{1}{2}}}\right] y^{(\frac{1}{2})} \cdot e^{-\frac{y}{2400}}$	$\alpha \approx \frac{1}{2}$ $\beta \approx 2400$
Exponential 1	$f(y) = \frac{1}{1200} \cdot e^{-\frac{y}{1200}}$	$\beta \approx 1200$
Exponential 2	$f(y) = \frac{1}{1707} \cdot e^{-\frac{y}{1707}}$	$\beta \approx 1707$

Applying all of these distributions on one graph is shown below.



To analyze this graph, look at how close the lines are to the distribution. The green line is above the black density function, which means it will overestimate if chosen as the distribution. The red line looks as if it will overestimate the black line as well, since the line is slightly above at most of the points. The blue line starts above the black density line, but decreases to go under the density line. The blue one seems to be more balanced than the other two; however, concrete evidence will be needed to prove or disprove this.

Chi Square Goodness of Fit Test

To start, the null hypothesis and alternate hypothesis needs to be defined. The null hypothesis is that the distribution fits the data, while the alternate hypothesis is that the distribution does not fit the data. If the probability is below 0.05, then we reject the null hypothesis and come to the alternative hypothesis, which is the distribution does not fit the data. If the probability is above 0.05, then we fail to reject the null hypothesis. Next, the bins must be calculated to calculate the Chi-Square value. Having at least five observations in each bin is recommended, so selecting the bins based off of 10% intervals will yield bin sizes with six observations in them. The table below shows the computed bins. The top row contains the percentiles, the middle has the value in time for the percentile, and the last row is the number of observations from the last percentile to the current percentile. Each bin has six observed values.

Perc	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Value	11.9	104.8	257.9	362.8	587.5	836.2	1292.1	1723.8	3172.4	8572.0
Obs	6	6	6	6	6	6	6	6	6	6

The Chi-Square value is defined as: (Wackerly, Mendenhall, and Scheaffer : Page 715)

$$X^2 = \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)}$$

The n_i is the number of observations at i, where $i = 1,2,3, \ldots, k$. The $E(n_i)$ is the expected number of observations at i. Since the exponential distribution and gamma distribution seem to fit the data, then the different distributions must be compared. To use the Chi-Squared Test, I must first find the expected number of observations for each of time. Since there are sixty observations in the data, multiplying the probabilities by sixty will give the expected number of earthquakes, of magnitude six or higher, in a sample of size sixty for each of the distributions.

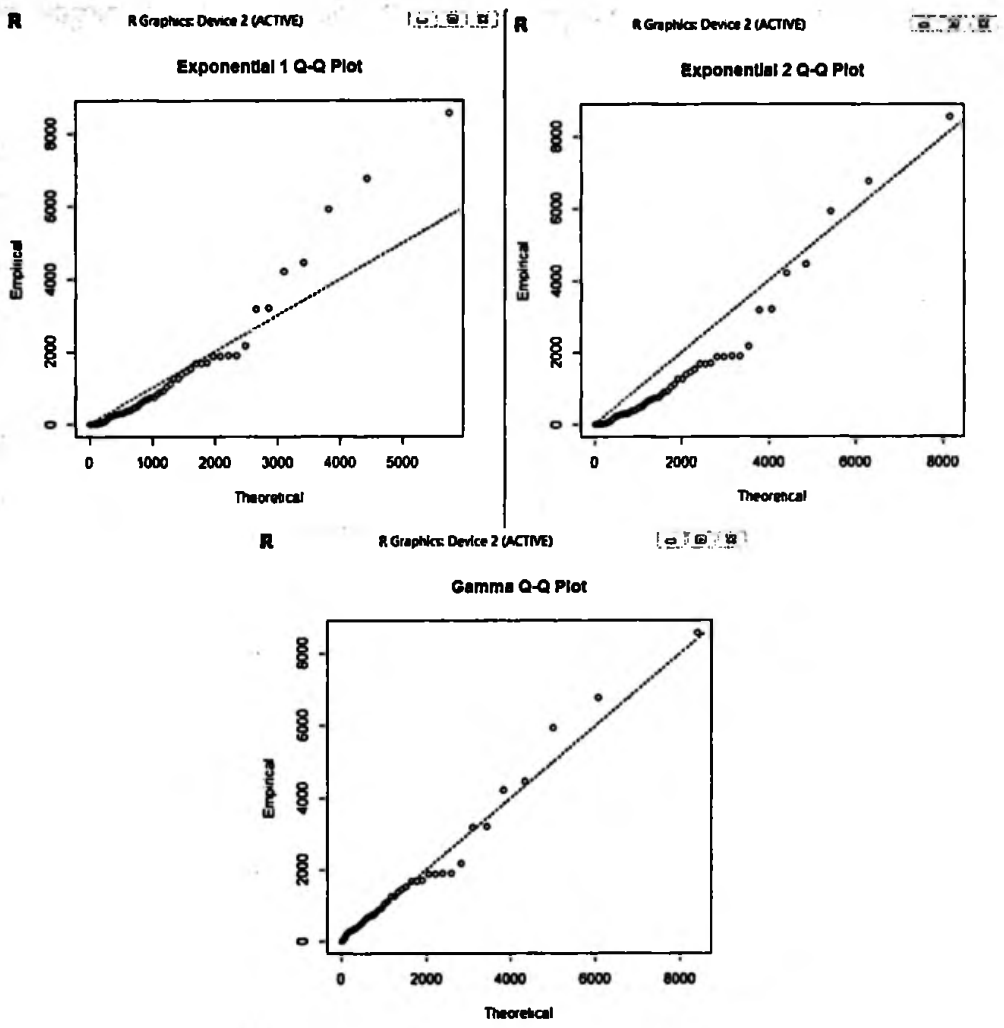
Chi-Square Analysis

The Chi-Square test sums the difference of expected values and observed values squared, so the more accurate distribution will have the smallest difference between the expected values and observed values. To calculate the following Chi-square values were calculated for these five bins, and are given below. These Chi Square values produced the following probability values for each of the distributions.

Distribution	Chi-Square Value	Deg. of Freedom	P-value
Gamma	5.923260693	8	0.6558
Exponential 1	59.91860959	9	0.0000
Exponential 2	92.58583755	9	0.0000

Both of the exponential distribution have probability less than 0.05, so we reject these distributions. Since the gamma distribution has a probability greater than 0.05, then I fail to reject this null hypothesis. Out of the three distributions, the gamma distribution is the only one I fail to reject, so this is the best distribution of the three. However, since the two parameters for the gamma distribution were found using an approximate version of the method of moments, so is this estimator good enough? To answer this, I looked into Q-Q plots.

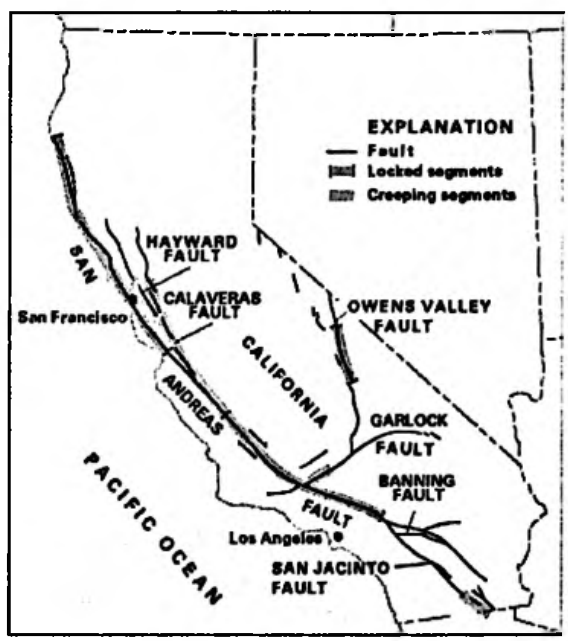
A Q-Q plot is a Quantile-Quantile plot, and it compares the quantiles for the distribution with the quantiles of the data. If they are related, then the quantiles should be linear. Each of the Q-Q plots are given below, and the dashed line is the linear line x=y.



The two exponential distributions both are not on the line, although exponential 1 seems to be pretty close. However, the gamma plot seems to have most of the points on the line. This leads to believe that the estimator that was selected seems to fit the distribution well, so the estimator is good enough.

LOCATIONS IN DATA

The data collected is for earthquakes throughout California, and adding location to the data analysis is essential. An earthquake that strikes San Francisco will not affect buildings in San Diego. For this reason, location needs to be included to accurately effect buildings within an area of an earthquake. California has earthquakes every day, some smaller than others. The reason why is because California lies directly on top of fault lines, which is the cause of earthquakes. The fault lines are shown below.



The biggest fault line is the San Andreas Fault. Since these faults are where earthquakes occur, I want to determine the probability that a house at a given location will be hit by a magnitude six or higher earthquake. I divided the fault line into equal zones so that an earthquake that hits within a zone will affect the buildings in that zone. Each zone is about 25 miles long, and if an earthquake of magnitude six or higher strikes in that zone, all houses would be shaken in that 25

mile area zone, causing building damage since the magnitude of the earthquake is extremely high. Since a building that needs to be repaired due to an earthquake will cost money for an insurance company, then buildings that have a higher chance of being hit by an earthquake will cost more money to be insured. The separation of the zones is shown below.

After the fault line is divided, I wanted to find the number of earthquakes that struck each zone. The locations of the earthquakes were not directly on the fault line. The epicenter of the earthquakes occurred around the fault line, so to capture the earthquakes in a zone, I transformed each divided line into an area. Most of the earthquakes fell into one of these areas. These areas are shown below.

There are a total of 23 zones that were created. There are sixteen zones that are along the San Andreas Fault along the west coast. The rest of the zones are inland. This is illustrated below.

Now getting the location of each earthquake in the data, I added the number of earthquakes that occurred in each zone. This is known as the frequency of the number of earthquakes that struck in each zone. This is represented below.

As shown above, the most earthquakes are at the very top zone with twelve of the sixty one earthquakes striking within the zone. A few other zones had many earthquakes in them also. I considered the ones with five or more to be the most dangerous zones. The most dangerous zones will have buildings that need the most repair, and these buildings will be the highest cost to be insured. Most of the earthquakes occurred toward the coast and rarely struck inland.

DATA ANALYSIS

Life Tables

Below I have constructed life tables for buildings in California. The zones where no earthquakes occurred are not important because I want to find out which zones are the most dangerous. From

what I learned about life tables, I need l_{x_0} and ${}_tq_x$ in order to create a life table. The l_{x_0} is the initial number of survivors, but for the data, it is the initial number of standing buildings. The initial number is set to some arbitrary number, so I used 50,000 houses for each zone. The important part of the life table isn't the initial number. The important part is to see what is expected to change in the number of survivors. For ${}_tq_x$, I will use the best fitting distribution to accurately find the probability of an earthquake damaging the buildings in each zone. Using the frequency of earthquakes that occurred in each zone, I divided the frequency by the total number of earthquakes to get the probability of an earthquake of magnitude six or higher striking in each zone. I multiplied this probability by the gamma probability density function, which is the below density function. The higher the frequency for a given zone, the higher the value is when doing the above step, giving a greater mortality probability for the zone.

$$f(y) = \left[\frac{1}{\Gamma(\alpha) \cdot \beta^{\alpha}}\right] y^{(\alpha-1)} \cdot e^{-\frac{y}{\beta}}$$

The difference is denoted as d_x, which is included in the life tables and is defined as:

$$d_x = l_x - l_{x+1}$$

These three together make up the life tables, which are shown below.

Zone 1				
Days	Years	qx	lx	dx
1	0	0.002368286	50000	118.4143
365	1	0.000102899	49881.59	5.132749
730	2	6.23944E-05	49876.45	3.112009
1095	3	4.37402E-05	49873.34	2.181471
1460	4	3.25418E-05	49871.16	1.622899
1825	5	2.50127E-05	49869.54	1.247373
3650	10	8.31265E-06	49868.29	

Zone 2				
Days	Years	qx	lx	dx
1	0	0.000197361	50000	9.868055
365	1	8.57506E-06	49990.13	0.428668
730	2	5.19963E-06	49989.7	0.259928
1095	3	3.64509E-06	49989.44	0.182216
1460	4	2.71187E-06	49989.26	0.135565
1825	5	2.08444E-06	49989.13	0.104199
3650	10	6.92735E-07	49989.02	

Zone 4				
Days	Years	qx	lx	dx
1	0	0.00039471	50000	19.73552
365	1	1.71496E-05	49980.26	0.857142
730	2	1.0399E-05	49979.41	0.519734
1095	3	7.28996E-06	49978.89	0.364344
1460	4	5.42358E-06	49978.52	0.271063
1825	5	4.16875E-06	49978.25	0.208347
3650	10	1.38543E-06	49978.04	

Zone 5				
Days	Years	qx	lx	dx
1	0	0.000986782	50000	49.33909
365	1	4.28743E-05	49950.66	2.141599
730	2	2.59975E-05	49948.52	1.298539
1095	3	1.8225E-05	49947.22	0.910289
1460	4	1.3559E-05	49946.31	0.677224
1825	5	1.04219E-05	49945.63	0.52053
3650	10	3.46359E-06	49945.11	

Zone 6				
Days	Years	qx	lx	dx
1	0	0.001381504	50000	69.0752
365	1	6.00244E-05	49930.92	2.997074
730	2	3.63968E-05	49927.93	1.817218
1095	3	2.55152E-05	49926.11	1.273875
1460	4	1.89828E-05	49924.84	0.947712
1825	5	1.45908E-05	49923.89	0.72843
3650	10	4.84906E-06	49923.16	

Zone 7				
Days	Years	qx	lx	dx
1	0	0.00039471	50000	19.73552
365	1	1.71496E-05	49980.26	0.857142
730	2	1.0399E-05	49979.41	0.519734
1095	3	7.28996E-06	49978.89	0.364344
1460	4	5.42358E-06	49978.52	0.271063
1825	5	4.16875E-06	49978.25	0.208347
3650	10	1.38543E-06	49978.04	

Zone 8				
Days	Years	qx	lx	dx
1	0	0.00039471	50000	19.73552
365	1	1.71496E-05	49980.26	0.857142
730	2	1.0399E-05	49979.41	0.519734
1095	3	7.28996E-06	49978.89	0.364344
1460	4	5.42358E-06	49978.52	0.271063
1825	5	4.16875E-06	49978.25	0.208347
3650	10	1.38543E-06	49978.04	

Zone 9				
Days	Years	qx	lx	dx
1	0	0.000986782	50000	49.33909
365	1	4.28743E-05	49950.66	2.141599
730	2	2.59975E-05	49948.52	1.298539
1095	3	1.8225E-05	49947.22	0.910289
1460	4	1.3559E-05	49946.31	0.677224
1825	5	1.04219E-05	49945.63	0.52053
3650	10	3.46359E-06	49945.11	

Zone 10				
Days	Years	qx	lx	dx
1	0	0.000197361	50000	9.868055
365	1	8.57506E-06	49990.13	0.428668
730	2	5.19963E-06	49989.7	0.259928
1095	3	3.64509E-06	49989.44	0.182216
1460	4	2.71187E-06	49989.26	0.135565
1825	5	2.08444E-06	49989.13	0.104199
3650	10	6.92735E-07	49989.02	

Zone 11				
Days	Years	qx	lx	dx
1	0	0.000789433	50000	39.47163
365	1	3.42997E-05	49960.53	1.713633
730	2	2.07982E-05	49958.81	1.039055
1095	3	1.45801E-05	49957.78	0.728392
1460	4	1.08473E-05	49957.05	0.541901
1825	5	8.33762E-06	49956.51	0.416518
3650	10	2.7709E-06	49956.09	

Zone 12				
Days	Years	qx	lx	dx
1	0	0.00039471	50000	19.73552
365	1	1.71496E-05	49980.26	0.857142
730	2	1.0399E-05	49979.41	0.519734
1095	3	7.28996E-06	49978.89	0.364344
1460	4	5.42358E-06	49978.52	0.271063
1825	5	4.16875E-06	49978.25	0.208347
3650	10	1.38543E-06	49978.04	

Zone 13				
Days	Years	qx	lx	dx
1	0	0.000592071	50000	29.60357
365	1	2.57247E-05	49970.4	1.285472
730	2	1.55986E-05	49969.11	0.779448
1095	3	1.09351E-05	49968.33	0.546406
1460	4	8.13546E-06	49967.79	0.406511
1825	5	6.25318E-06	49967.38	0.312455
3650	10	2.07816E-06	49967.07	

Zone 14				
Days	Years	qx	lx	dx
1	0	0.000789433	50000	39.47163
365	1	3.42997E-05	49960.53	1.713633
730	2	2.07982E-05	49958.81	1.039055
1095	3	1.45801E-05	49957.78	0.728392
1460	4	1.08473E-05	49957.05	0.541901
1825	5	8.33762E-06	49956.51	0.416518
3650	10	2.7709E-06	49956.09	

Zone 16				
Days	Years	qx	lx	dx
1	0	0.001184143	50000	59.20714
365	1	5.14493E-05	49940.79	2.569421
730	2	3.11972E-05	49938.22	1.557932
1095	3	2.18701E-05	49936.67	1.09212
1460	4	1.62709E-05	49935.57	0.812497
1825	5	1.25064E-05	49934.76	0.624502
3650	10	4.15633E-06	49934.14	

Zone 17				
Days	Years	qx	lx	dx
1	0	0.00039471	50000	19.73552
365	1	1.71496E-05	49980.26	0.857142
730	2	1.0399E-05	49979.41	0.519734
1095	3	7.28996E-06	49978.89	0.364344
1460	4	5.42358E-06	49978.52	0.271063
1825	5	4.16875E-06	49978.25	0.208347
3650	10	1.38543E-06	49978.04	

Zone 21				
Days	Years	qx	lx	dx
1	0	0.000197361	50000	9.868055
365	1	8.57506E-06	49990.13	0.428668
730	2	5.19963E-06	49989.7	0.259928
1095	3	3.64509E-06	49989.44	0.182216
1460	4	2.71187E-06	49989.26	0.135565
1825	5	2.08444E-06	49989.13	0.104199
3650	10	6.92735E-07	49989.02	

Zone 23				
Days	Years	qx	lx	dx
1	0	0.000197361	50000	9.868055
365	1	8.57506E-06	49990.13	0.428668
730	2	5.19963E-06	49989.7	0.259928
1095	3	3.64509E-06	49989.44	0.182216
1460	4	2.71187E-06	49989.26	0.135565
1825	5	2.08444E-06	49989.13	0.104199
3650	10	6.92735E-07	49989.02	

CONCLUSION

The most dangerous zone is zone one since it had the highest frequency out of all the zones. Many of the zones in the long run lose less than 100 houses due to earthquakes. However, zone 1 loses over 100 houses in the first year. Zone 1 is the deadliest zone, and houses in that area have the highest probability of being struck by a giant earthquake. Since zone 1 is the deadliest zone, the cost to insure these buildings are the greatest since the chance of needing to repair the buildings are the greatest. In order for insurance companies to make money from insuring these building with high probabilities of earthquakes occurring, the cost of the insurance policy will cost the most. Having these life tables helps determine which group is more risky than the other groups. The reason for making a life table is to calculate the amount of risk involved with different groups, which for my data is different zones.

The process of calculating actuarial based mathematics consists of analyzing survival models of the population, which includes the Leslie Matrix, and creating Life Tables. These are important because reducing the financial costs of undesirable events leads to maximum profit. These undesirable events include calculating mortalities and finding the probabilities of survivability. The time between tremendous earthquakes closely illustrates the mortality and survivability for buildings in California. When the next earthquake strikes, will the buildings around you stay standing, or fall to the ground?

APPENDIX

R Statistics Software

R is a program for statistical computations and statistical graphics. R is used in graduate schools and in many companies, so learning how to use this program is essential for most mathematicians. Ican use R for my data, but I must import the data into R. In order to import my data, I had to change the working directory to the directory that has my data using

setwd("C:/Users/varna_000/documents/R")

The "setwd" stands for set working directory. Once I have R set to the proper directory, I need to store the observed data into R. The data was entered in Excel, so the most efficient way to get the data in R is to import the data. R cannot read Excel files that end in the xlsx extension – which is for 2007 Excel of newer. The file had to be saved as a "Comma-Separated Variables " extension. R can import csv files easily by setting a variable equal to the csv file. I called this variable "data" and imported the data by using:

data <- read.csv("Time Data.csv", header = TRUE)

The "<-" indicates to the program to set the value of the left side to the value of the right side. The read.csv function reads in the file that is indicated in quotes. The csv file must be in the working directory described above in order for this to work. The "header = TRUE" is saying that the file has headers for the columns. If I would have put "header = FALSE", then the first row would have been input as numbers.

Analysis and Graphics in R

After the data is in R, analysis begins. The first step is to find the mean and variance of the data. The following shows this valuable information. To calculate the mean, the function 'mean' is

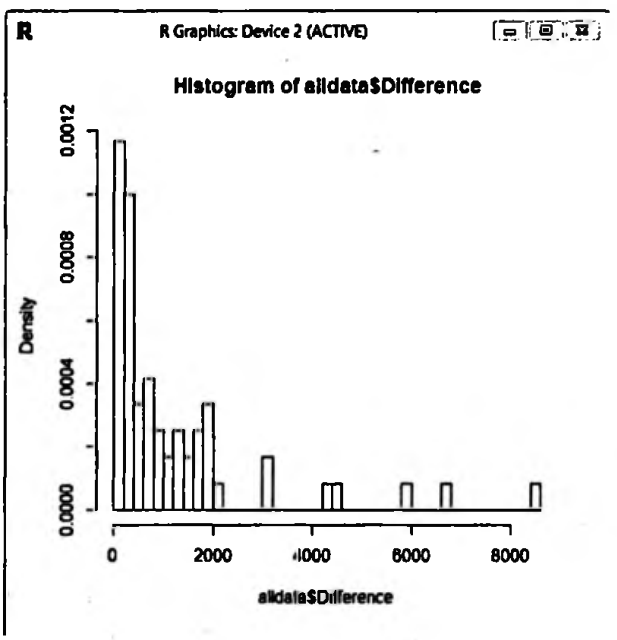
input into R. To calculate the variance, the input is 'var'. To calculate the standard deviation, the input is 'std'.

Mean	1199.767
Variance	2912958
Std Dev	1706.739

To get a better understanding of the data, the next step is to create a histogram. Histograms show the shape of the observations, and to show which observations occur more frequently than other observations. This is done in R by using

hist(data$Difference, breaks=50, prob=TRUE)

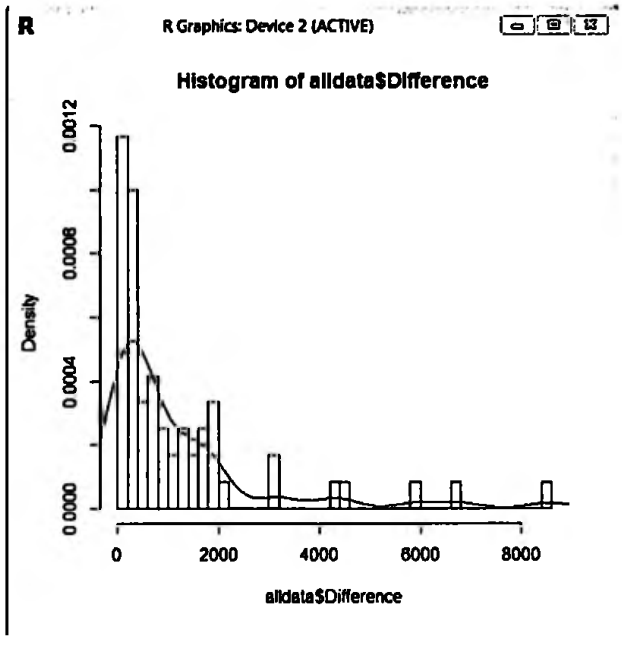
Since I imported multiple observations for each earthquake, the variable 'data' contains a table of three columns – one being difference in time between earthquakes, another being magnitude of the earthquake, and the last column of the table being the location of the earthquake. To only look at the differences column in the variable 'data', I used 'data$Difference'b. The 'breaks=50' part means to create 50 equally spaced bins for the histogram. Since I want to look at the probabilities and not the frequencies, then I put 'prob=TRUE'. These together create the following histogram.



Most of the differences between earthquakes are close to zero. This histogram shows that the data is skewed right, which means that most of the data is on the left side and little data is on the right side. R can calculate the density, so in order to add a line of the density into the histogram, I entered

```
lines(density(data$Difference))
```

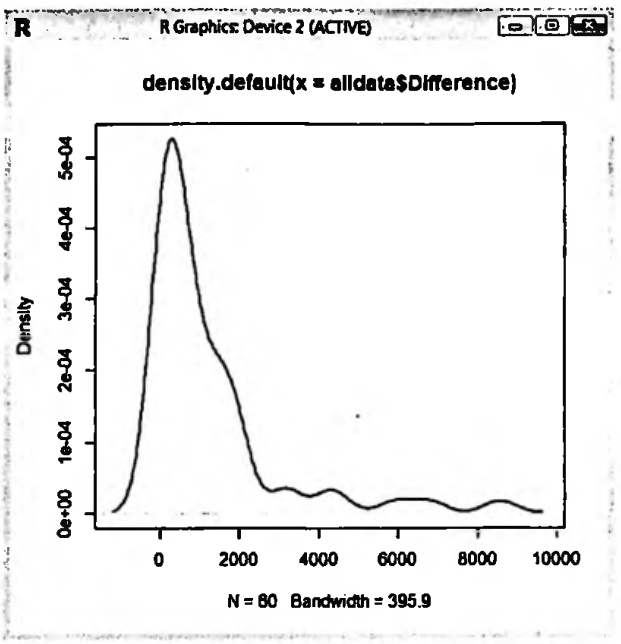
This produced the following graph.



Comparing the density of the data with the density of a distribution will give valuable information about what distribution will fit the data most accurately. To get just the density by itself, the input is:

```
plot(density(data$Difference))
```

The 'plot' function creates a new graph, while the 'lines' function adds to the current graph. The above input produces the graph of only the density.



To calculate the parameters for the gamma distribution in R, variables were created to store α and β, called 'alphadata' and 'betadata'. The calculation for 'betadata' is the same as β for the gamma distribution.

```
betadata <- var(data$Difference)/mean(data$Difference)
```

To calculate 'alphadata', I use the equation for α for the gamma distribution using the new variable 'betadata' where β is used.

```
alphadata <- mean(data$Difference)/betadata
```

The same thing needs to be done for both exponential distributions. For exponential 1, the β is calculated by setting β equal to the mean.

```
expo1beta <- mean(data$Difference)
```

For exponential 2, the β is found by setting the β equal to the standard deviation.

```
expo2beta <- sd(data$Difference)
```

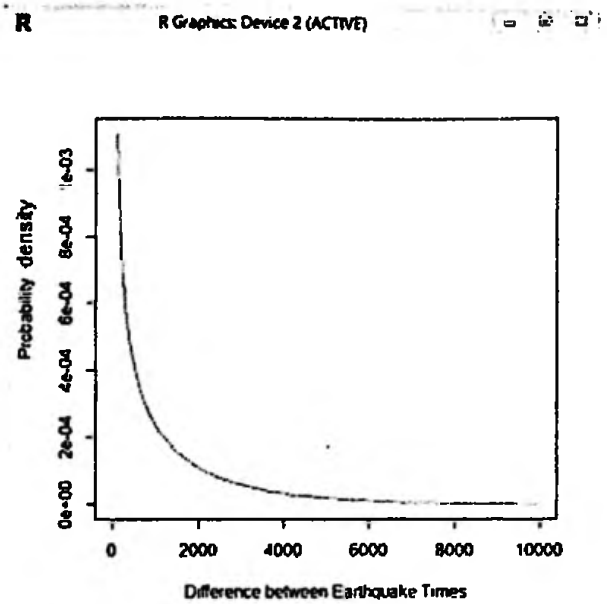
All of these calculation gave the following parameters for each of the distributions.

Distribution	Probability Function	Parameters
Gamma	$f(y) = \left[\frac{1}{\Gamma(\frac{1}{2}) \cdot (2400)^{\frac{1}{2}}}\right] y^{(-\frac{1}{2})} \cdot e^{-\frac{y}{2400}}$	$\alpha \approx \frac{1}{2}$ $\beta \approx 2400$
Exponential 1	$f(y) = \frac{1}{1200} \cdot e^{-\frac{y}{1200}}$	$\beta \approx 1200$
Exponential 2	$f(y) = \frac{1}{1707} \cdot e^{-\frac{y}{1707}}$	$\beta \approx 1707$

Then plotting the gamma function becomes a matter of using β for the scale and α for the shape since R has the probability function built into it.

```
curve(dgamma(x, scale=betadata, shape=alphadata), from=0, to=10000, col="red",
    xlab="Difference between Earthquakes", ylab="Probability")
```

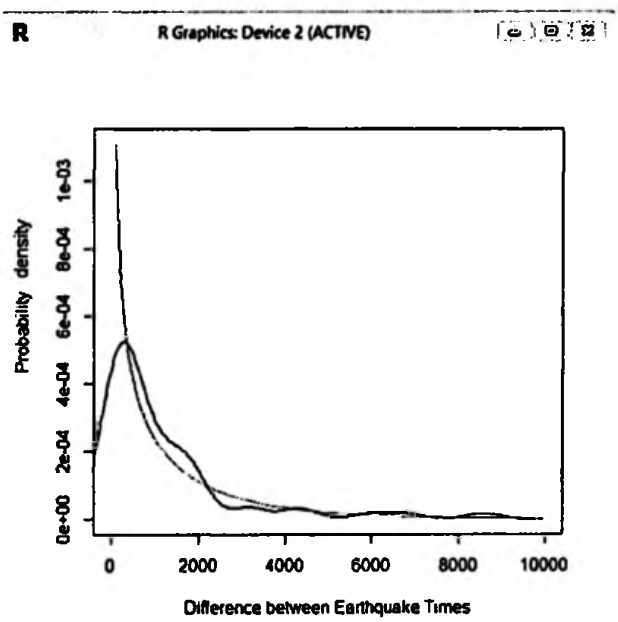
dgamma stands for the density function for gamma. The 'x' is the variable of the probability function that is changing, which goes up to 10000. The 'col' stands for color, and makes the line the color that is typed. The 'xlab' stands for x label, which is to label the x-axis. Similarly, the 'ylab' is for labeling the y-axis. This input produces the graph for the gamma distribution, which is below.



However, further analysis needs to be done by comparing this gamma distribution to the density function of my data. To add the density function of the data to this plot, the input into R is:

lines(density(data$Difference))

This yields the plot of both functions on the same graph below.

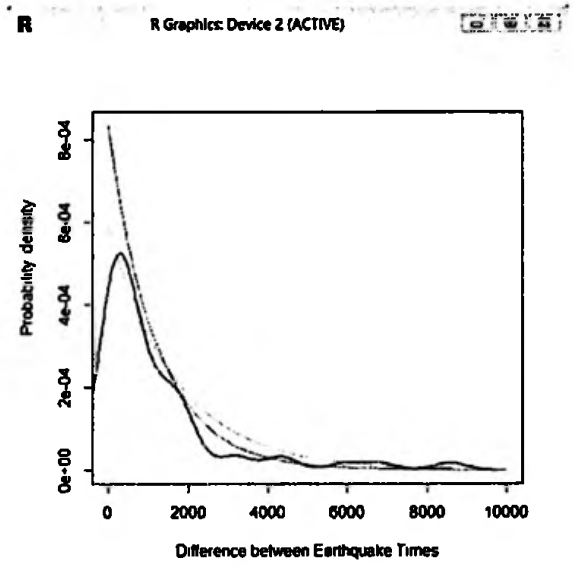


However, this is for only the gamma distribution. For further analysis, the exponential distribution needs to be plotted to see which has the better fit.

To show what the exponential distributions look like compared to the density of the data, I input the following.

```
curve(dexp(x, rate=1/expobeta), from=0, to=10000, col="red", xlab="Difference between
    Earthquake Times", ylab="Probability")
lines(x, expo2data, col="green")
lines(density(data$Difference))
```

This produces the following graph.

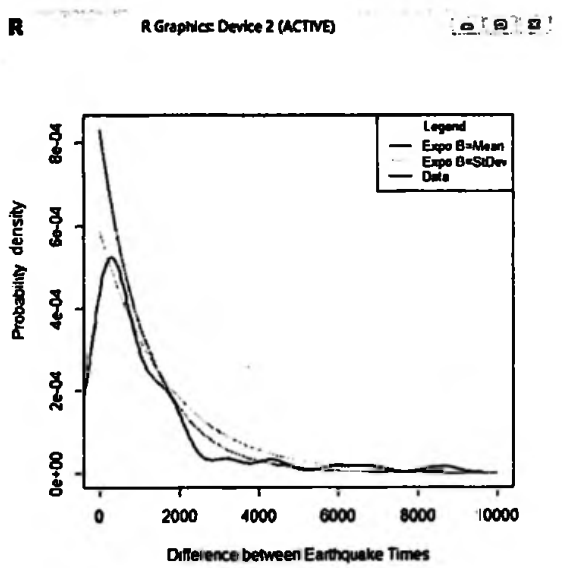


This graph has one major flaw in it. The colors are not represented on it. This graph needs a legend. The legend is created by inputting

```
legend("topright",c("Expo B=Mean","Expo B=StDev","Data"), title="Legend", cex=0.8,
    col=c("red","green","black"), lty=1)
```

The 'topright' is for the location of the legend, so the top right corner should suffice. The next part is declaring a vector, which contains the words for the legend. The 'cex' is the font size used for the

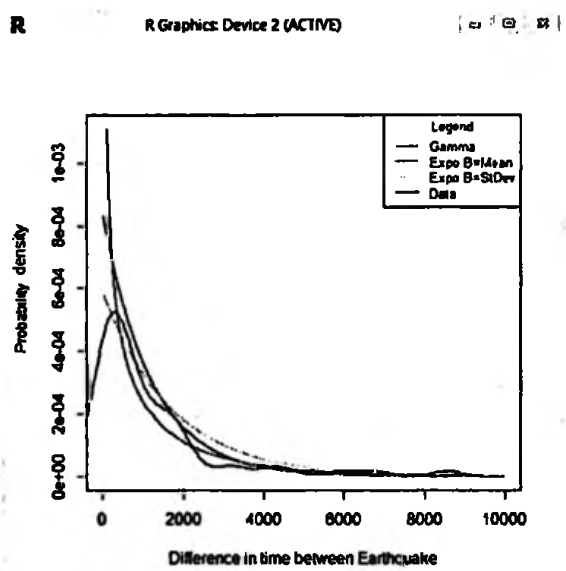
legend. The color (declared as 'col') in this case is a vector since there is more than 1 color being used. The 'lty' stands for line type, and setting it equal to one means it's a simple, solid line. This produces the graph below



This figure only compares the two different exponential distributions. To add the gamma distribution, the input below is used.

lines(x, dgamma (x, scale=betadata, shape=alphadata), col="blue")

This produces the following graph.



Chi Square Test

To use the Chi-Squared Test in R, I must first find the expected number of observations for each of time. Since there are 60 observations in the data, multiplying the probabilities by 60 will give the expected number of earthquakes, of magnitude six or higher, in a sample of size 60 for each of the distributions. I began with the gamma distribution by entering the below lines into R (where gammaobs is the expected observations for the gamma distribution and the dataobs is the actual number of observations for the data).

```
gammadiff <- gammaobs - dataobs
gammadiffsq <- (gammadiff)^2
gammadiffsq <- (gammadiffsq/gammaobs)
gammachisq <- sum(gammadiffsq)
```

These calculations are the same as the definition for the Chi-Square Test from above where each line is a single step in the definition. Doing the same calculations for the exponential distributions was done by using the variable 'expo1' substituted for 'gamma' and 'expo2' substituted for 'gamma'.

BIBLIOGRAPHY

1. Crowther, David. "Apptuary Free!" iOS App. N.p, 8 Mar. 2013. <http://www.apptuary.com/>
2. Dickson, David, Mary Hardy, and Howard Waters.Actuarial Mathematics for Life Contingent Risks. New York: Cambridge University Press, 2009. Print. Pages 25-26
3. On the Use of Matrices in Certain Population Mathematics, 1945. Biometrika; vol 33 no 3. Oxford, England: Bureau of Animal Population. Leslie, Paul. Feb 20th 2013. Pages 183.
4. U.S. Department of the Interior, . "Historic Earthquakes in the United States and Its Territories." *United States Geological Survey*. Web. 7 Mar. 2013. <http://earthquake.usgs.gov/earthquakes/states/historical_state.php>
5. Wackerly, Dennis, William Mendenhall, and Richard Scheaffer. Mathematical Statistics with Applications. 7th ed. Belmont, CA: Thomson Higher Education, 2008. Print. Page 839