(1) First, we will examine the probabilities of large earthquakes on each fault. That is, we will examine the probability, given an earthquake has happened, that the earthquake is large instead of small (18 points total):

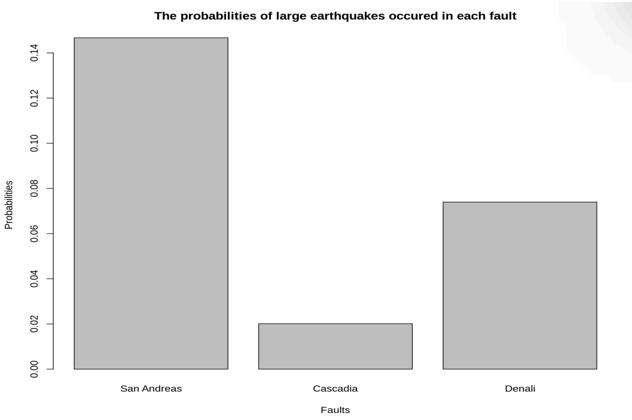
a. Choose a distribution for this variable and explain in 1-3 sentences whether the assumptions of your chosen distribution are met. (Note: it is ok if some of the assumptions are not 100% certain. Choose the distribution that is the best match, and call into question any assumptions that might not be true; 6 points)

The variables met the assumption of Bernoulli distribution. First, there are only two outcomes: large or small. Secondly, the number of trails is fixed: how many earthquakes have happened are shown in the dataset. Third, each trail is independent from one another. Whether the earthquake is small or large won't affect the next earthquakes to be small or large.

b. Estimate the probabilities that an earthquake on each of the three listed faults is of magnitude higher than 6.0 (this is our definition of "large"; 4 points).

For San Andreas, the probability that an earthquake has a magnitude higher than 6.0 is 14.67%.
202/1377=0.1466957
For Cascadia, the probability that an earthquake has a magnitude higher than 6.0 is 2.01%.
79/3927=0.02011714
For Denali, the probability that an earthquake has a magnitude higher than 6.0 is 7.40%.
168/2272=0.07394366

c. Make a bar plot summarizing these probabilities (4 points).



d. Calculate for the Cascadia Fault the probability that, out of 30 earthquakes in a given year, more than one exceeds magnitude 6.0 (4 points).

The probability that in Cascadia Fault, out of 30 earthquakes in a given year, more than one exceeds magnitude 6.0 is 12.17%.

(2) Now, we will examine the average number of earthquakes to happen on each fault (22
points total):

a. Choose a distribution for this variable and explain in 1-3 sentences whether the assumptions of your chosen distribution are met (6 points).
Those variables met the assumptions of the Poisson distribution. First, there is no binary variables. Secondly, there is no clear max for the outcomes. Third, it is counting something — the average number of earthquakes happened on each fault, which is the parameter of Poisson distribution.

b. Estimate the average number of earthquakes on each fault per year (4 points).
The average number of earthquakes on San Andreas per year is 44.41935.
The average number of earthquakes on Cascadia per year is 126.6774.
The average number of earthquakes on Denali per year is 73.29032.

c. Calculate the empirical variance in the number of earthquakes on each fault per year. Compare these variances to the means. Do your results match what you would expect based on your chosen distribution? (4 points)
The empirical variance in the number of earthquakes on San Andreas per year is 72.98495.
The empirical variance in the number of earthquakes on Cascadia per year is 2853.826.
the empirical variance in the number of earthquakes on Denali per year is 96.6129.

My results don't match what I would expect. For Poisson distribution, the variance should be the same with the average. However, the results turned out that they are not the same. I think it is because of sampling issue, and this caused a difference between the real values and the empirical values. Especially when it comes to Cascadia, the huge difference between the average and the variance implies that the Poisson distribution may not be the best distribution for this dataset.

d. Your variable is currently calibrated to an annual time scale, since this is how the data is presented. Rescale to a decadal time scale. Use this new variable to estimate the probability that the Denali Fault will have more than 700 earthquakes over a 10-year period (4 points)
The probability that the Denali Fault will have more than 700 earthquakes over a 10-year period is 88.49%.

e. Your variable is currently calibrated to account for the whole fault line. Assume that earthquakes are uniformly distributed across the length of the fault line. Suppose that King County takes up 38 km of the Cascadia fault line. Estimate the probability that King County will be hit by at least 10 earthquakes in a given year. (4 points)
The probability that King County will be hit by at least 10 earthquakes in a given year is 1.486%.

Applied Independent Learning Problem:
In class this week we discussed applied ILPs. In most future weeks, the ILP will require you to run some basic calculations and interpret the results. This week the ILP covers the geometric distribution, which we started working on in class. Remember the steps of an applied ILP:
(1) Read about the technique
(2) Break down into steps how you will determine when the technique should be used
(3) Break down into steps how the technique is used in practice
(4) Attempt the given calculations.

For this week, this process will be formalized in the questions themselves. In future weeks, we will skip straight to the calculations, but you should still break down the when and how for your own learning.

(1)
a. When should the geometric distribution be used? Break down into steps how you can tell. (4 points)
There are a few steps that we used to make sure whether we can apply geometric distribution.

The geometric distribution is similar to the Bernoulli distribution. First, the binary variables are presented in a dataset. Secondly, each trail is independent. Finally, the probability of success is the same for each trail. However, there are a few assumptions that the geometric distribution is different from the Bernoulli distribution: geometric distribution doesn't have a fixed number of trails — in geometric distribution, we continue to run trails until a success occurs. Also, the geometric distribution only has one parameter: p, the probability of success.
Break down into steps:
-whether the binary variables are presented in a dataset?
-whether each trail is independent?
-whether the probability of success is the same for each trail?
-whether the number of trails are fixed?
-How many parameters are presented/what is the parameter?

b. Give an example of a situation when the geometric distribution can be used (2 points)
Question: how many times that I need to roll a dice to get a 6?
Steps:
-The binary variables are presented: success(get 6), and failure(doesn't get 6) in a roll
-Trails are independent: each trail is not related to one another.
-The probability of success are all 1/6.
-The number of trails are not fixed: how many times for us to get 6 is not determined.
-There is only one parameter: p=1/6

(2) How can you use the geometric distribution to estimate probabilities? Break down into steps. (4 points)

1. Set the probability of success as "p"
2. Set the probability of failure as "1-p"
3. The probability of finding the first success in the n trial is given by: 「[(1-p)^n-1]*p」 *100%

Example: What is the probability that we would get 6 within the third roll?
「[(1-1/6)^(3-1)]*1/6」 *100%=11.57%
The probability that we would get 6 within the third roll is 11.57%.

a. Using the example from part (1) make up some fake data (1 point) and estimate a probability using the PMF (2 point) and the CDF (2 point)

Using PMF to calculate the probability that we get 6 in the nth roll:
PMF: calculating probability for one single trail
$P(n=1) = (1/6)$
$P(n=2) = (5/6)*(1/6)$
$P(n=3) = (5/6)^2 * (1/6)$

Using CDF to calculate the probability that we get 6 within the first n rolls, n=3:
CDF: calculating the probability for a range of trails.
$P(3)= P(1) + P(2) + P(3) = (1/6)+ (5/6)*(1/6)+ (5/6)^2 * (1/6)$