

INDIAN STATISTICAL INSTITUTE

Roll No – BS1613

SUBHRAJYOTY ROY

Assignment of STAT Methods II

Fitting Known Distributions to Real Datasets

Introduction

We often come in touch of several real life data which resembles our known distribution’s figures in a way. That often leads to fit a known distribution to the real life dataset and calculating several sample features, with the help of the features of the fitted distribution that has already been developed in theory.

Here, I have presented some real-life datasets in which some well-known distributions can be fitted. I have estimated the parameters of fitted distribution and then tested the goodness of fit measure.

Analysis of Data

# Data 1:

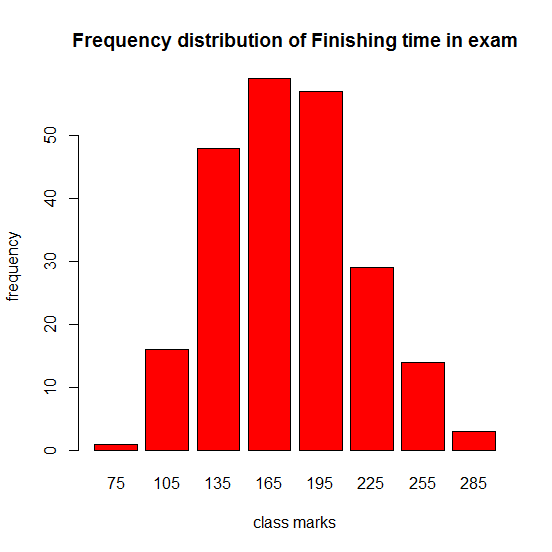
The dataset contains marks and finishing time of 227 students in an exam. The exam was designed in a way to see whether the finishing time of exam has any impact on the marks of the student. I have chosen only the time as my variable. It has been taken from [www.openmv.net](http://www.openmv.net).

**Observed Dataset**

Here is the observed frequencies of the finishing time of 227 students. The time is regarded as continuous variable (measurement is in minutes), and is broken into classes of interval of 30 minutes.

|  |  |
| --- | --- |
| **Class (in minutes)** | **Frequency (No. of students)** |
| 60-90 | 1 |
| 90-120 | 16 |
| 120-150 | 48 |
| 150-180 | 59 |
| 180-210 | 57 |
| 210-240 | 29 |
| 240-270 | 14 |
| 270-300 | 3 |
| **Total** | 227 |

Below we have the barplot of the frequency distribution of the dataset. As the classes are of equal length, the histogram will merely be a change of scale, but the shape will remain same. It can be seen that the plot can be approximated by a normal curve. So, I will try to fit a normal curve in the dataset. Also, by the frequency distribution, we see that the mean should be somewhere between 165 and 180.

**Fitting a distribution to the Dataset**

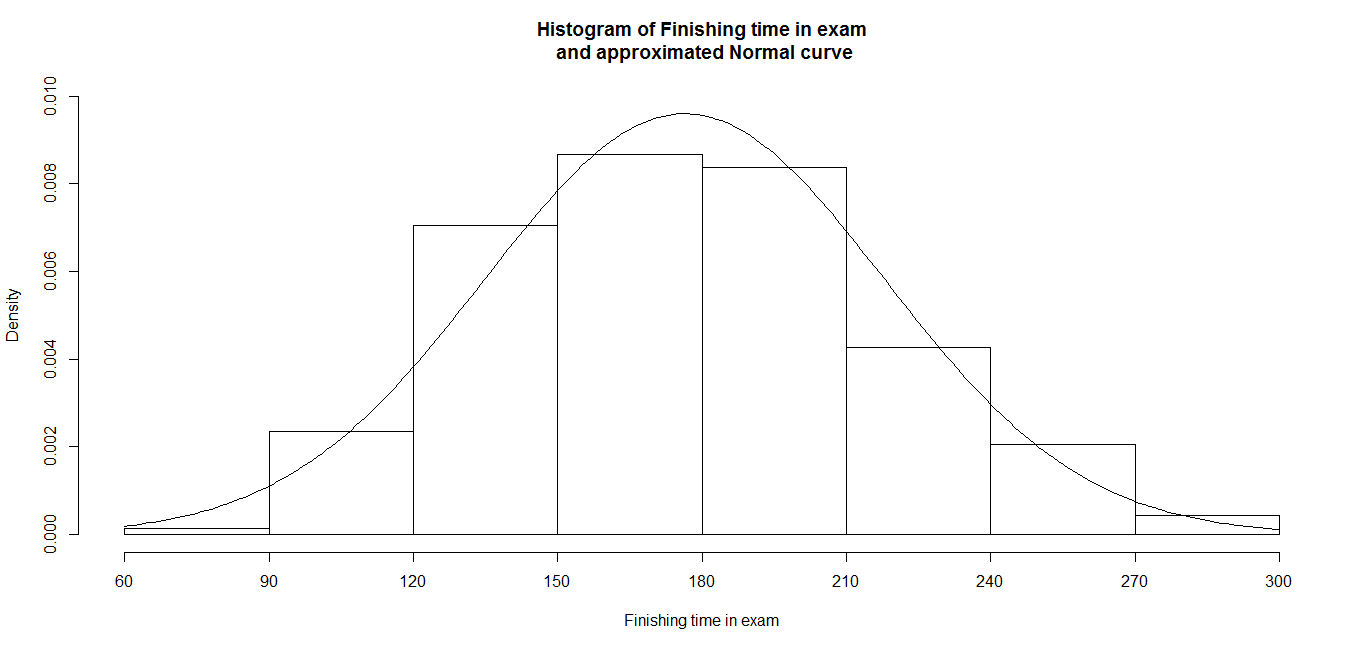
As suggested in the previous section, I will use normal distribution to fit the dataset. Normal distribution has the following density function;

By method of moments, the estimated parameters of the normal distribution comes out to be and . Based on these distribution, I get the expected frequencies and see that there are two classes with less than 5 expected frequency. Hence, I coalesced them with adjacent classes and finally get the following frequency distribution;

|  |  |  |
| --- | --- | --- |
| **Class (in minutes)** | **Observed Frequency** | **Expected Frequency** |
| <120 | 17 | 20 |
| 120-150 | 48 | 40 |
| 150-180 | 59 | 62 |
| 180-210 | 57 | 58 |
| 210-240 | 29 | 33 |
| >240 | 17 | 14 |
| **Total** | 227 | 227 |

We get the Pearsonian chi-square test statistic as, on (6- 1- 2) = 3 degrees of freedom (since, I have estimated 2 parameters of the normal distribution). Clearly, expected value of should be 3 and we have observed the test statistic as 3.34, close enough to 3. Hence, we may conclude that the fit is good. We see that p-value of chi-square, i.e. , indicating the fit is indeed good.

The approximated normal curve is given in the following diagram.



**Critical Observation**

As our initial guess suggests that **we should have a long left hand tail in our dataset,** since it contains the finishing time of students in the exam. It is our common intuition that we should have more students giving their answer scripts at the very last moment in the exam. **But, we find a symmetric distribution like Normal distribution to fit the dataset!** This leads to unintuitive and interesting thing in dataset.

It struck to me and I went back to the source to find why this anomaly happened. It may be a very easy exam where most of the students could complete it before the destined time. This is what exactly happened. They designed the exam in a way where the students were given a very long time to submit their answers.

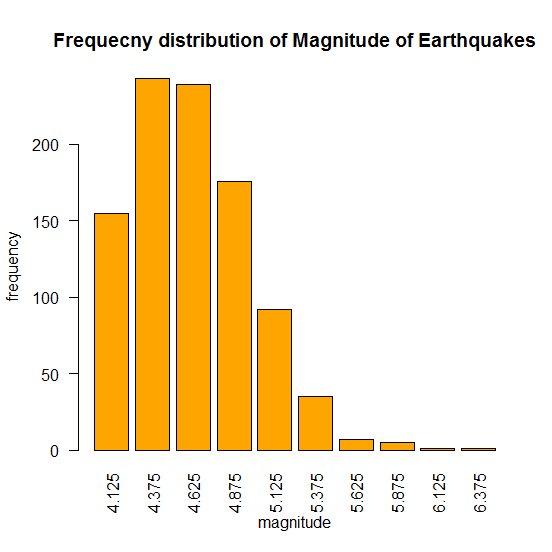
# Data 2:

This data contains earthquakes data in the island of Fiji. Fiji is an archipelago with more than 300 islands on South Pacific Ocean. The dataset 954 earthquakes’ descriptions along with their locations and magnitudes in Richter scale. I have taken the magnitudes as my variable. The data is taken from R’s built-in datasets package.

**Observed Dataset**

The magnitude can be measured upto 2-3 decimal point precision, and as it ranges over 4.0 to 6.5, I have regarded the variable as a continuous one. The observed frequency table is as follows:

|  |  |
| --- | --- |
| **Classes (in Richter unit)** | **Observed Frequency** |
| 4-4.25 | 155 |
| 4.25-4.5 | 243 |
| 4.5-4.75 | 239 |
| 4.75-5 | 176 |
| 5-5.25 | 92 |
| 5.25-5.5 | 35 |
| 5.5-5.75 | 7 |
| 5.75-6 | 5 |
| 6-6.25 | 1 |
| 6.25-6.5 | 1 |
| **Total** | 954 |

 I have presented below the barplot of the frequency distribution of observed dataset. Again, the classes having equal width, histogram will look exactly similar with merely a change of scale.

As suggested by the barplot, (or the similar shaped histogram), the distribution follows an exponentially downtrend. But, the rapid growth at the beginning of the distribution, leads us to think of Gamma distribution should fit the data. So, I will try with that.

**Fitting of the distribution to the Dataset**

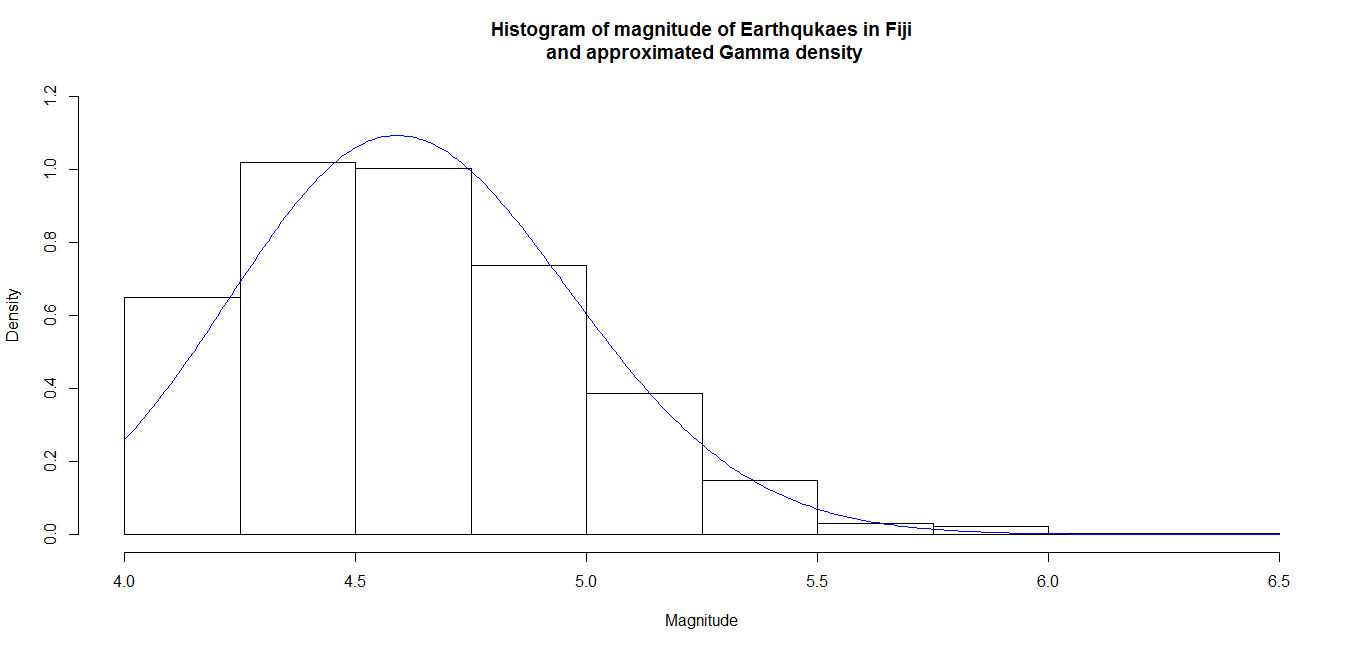
As suggested in the previous section, I will try ‘Gamma’ distribution to fit the data. Gamma distribution has the following density,

Where or the shape and or the rate are two parameters. By the method of moments, the estimated parameters of the distribution comes out to be; and .

Now, based on this density, I have calculated expected frequencies of each class and whenever the expected frequency is lower than 5, I have coalesced it with adjacent classes. Finally, we get the following frequency distribution;

|  |  |  |
| --- | --- | --- |
| **Classes (in Richter unit)** | **Observed Frequency** | **Expected Frequency** |
| <4.25 | 155 | 149 |
| 4.25- 4.5 | 243 | 215 |
| 4.5-4.75 | 239 | 255 |
| 4.75-5 | 176 | 193 |
| 5-5.25 | 92 | 98 |
| 5.25-5.5 | 35 | 33 |
| >5.5 | 14 | 11 |
| **Total** | 954 | 954 |

By the above table, we see that expected frequency is somewhat closer to the observed ones. Pearsonian chi-square test statistic comes out to be with degrees of freedom as (7- 1- 2) = 5 (since we are estimating two parameters of a distribution). Hence, the expected value of chi-square should be 5 and we are off by a margin of 2.696, which is tolerable as lies within sd of the chi-square distribution with df 5. P-value of chi-square i.e. , not so small to ignore. This indicates that the fit is good.

Here I have presented the approximated gamma curve with the histogram; 

**Critical Observation**

One common man would have expect most of the earthquakes have magnitude closer to 4-5, and the frequency will decrease to both the sides assuming a normal or symmetric trend. But, there are almost no earthquakes with magnitude lower than 4, and the very definition of earthquake suggests it is called an earthquake only if its magnitude is over 3. So, rather than having a symmetric curve, we have a frequency distribution with long right hand tail and could fit gamma distribution to magnitudes.

Also, one can get how much probable it is to come another earthquake of greater magnitude from the fitting, and can take necessary precaution to prevent/ minimize the damages later on. I found that, in Japan, as it being one of the most vulnerable country to earthquake, they build up architectures such that it can withstand the most probable magnitude of earthquakes.

# Data 3

The source of this dataset is “The World Almanac, Book of Facts”, where the number of “great” discoveries and scientific developments has been recorded for each year between 1860 and 1959. The dataset contains all the descriptions of those ‘great’ discoveries.

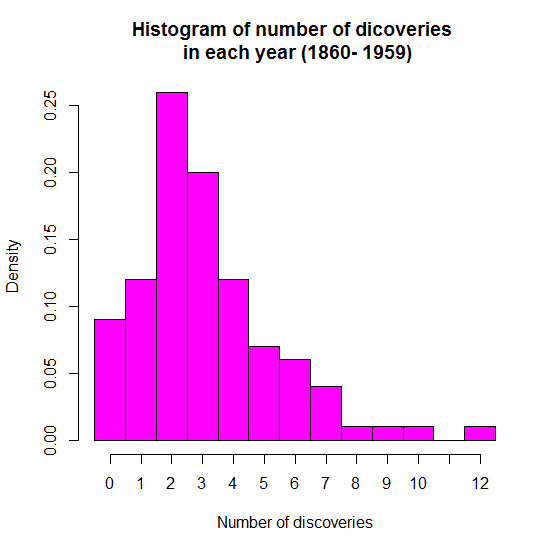
I have considered the number of discoveries as my variable, and the frequency as the number of years with that many discoveries. I have taken it from built-in datasets package of R.

**Observed Dataset**

The observed dataset contains the frequency distribution as follows:

|  |  |
| --- | --- |
| **Classes / Values (in no. of discoveries)** | **Observed frequencies** |
| 0 | 9 |
| 1 | 12 |
| 2 | 26 |
| 3 | 20 |
| 4 | 12 |
| 5 | 7 |
| 6 | 6 |
| 7 | 4 |
| 8 | 1 |
| 9 | 1 |
| 10 | 1 |
| 11 | 0 |
| 12 | 1 |
| **Total** | 100 |

The histogram of the observed dataset is given below in the diagram. As suggested by the histogram, the frequency density drops down rapidly to 0. In this situation, we expect that Poisson distribution should give a good fit to the dataset. So, we will try that.

The backdrop against using Poisson distribution that in any given year, there is no limit to the number of discoveries that can be made (suggesting ) and the probability of success in discovering something “great”, is small (suggesting ). But, i.e. expected number of discoveries should be finite and moderate, suggesting the growth of scientific advancement of humankind.

**Fitting distribution to the Data**

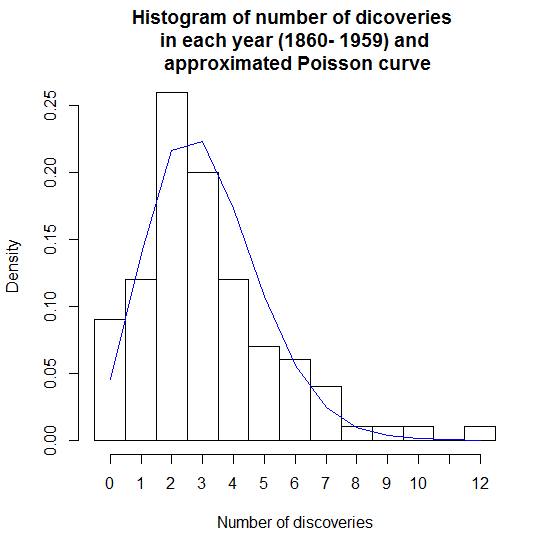
I have tried to fit Poisson density first. The Poisson density is given by the following formula:

By the method of moments, estimated parameter comes out to be . Using that, expected frequencies are found and finally the classes with lesser than 5 as expected frequencies are coalesced together. Finally, we get the following table;

|  |  |  |
| --- | --- | --- |
| **Classes/ No of discoveries in a year** | **Observed frequency** | **Expected frequency** |
| 0 | 9 | 5 |
| 1 | 12 | 14 |
| 2 | 26 | 22 |
| 3 | 20 | 22 |
| 4 | 12 | 17 |
| 5 | 7 | 11 |
|  | 14 | 9 |
| **Total** | 100 | 100 |

Based on these observed values and expected values, the calculated value of Pearsonian chi-square test statistic comes out to be, on ( 7- 1- 1) = 5 degrees of freedom. The expected value of should be 5, but we have something around 10 here, a bit off. But, essentially standard deviation of being (2\*5) = 10, we have the observed value between ‘mean sd’ range, essentially suggesting that the fit is not so bad. (We may say there is not enough evidence to reject that the distribution is not fitting the data.)

The p-value of the observed chi-square test statistic comes out to be 0.0727, not extremely small probability (or so called negligible probability).

 The estimated curve along with the histogram is given below;

Now, as the fit is not so good, I have tried to fit another distribution to this data, with some distribution following the same trend, but with more parameters to adjust. One thing to notice here is that, for Poisson to fit well, one good approach could be checking whether . But, in this case, **mean number of discoveries is 3.1, while the variance is much larger 5.08, implying that Poisson would not be a good fit to the dataset.**

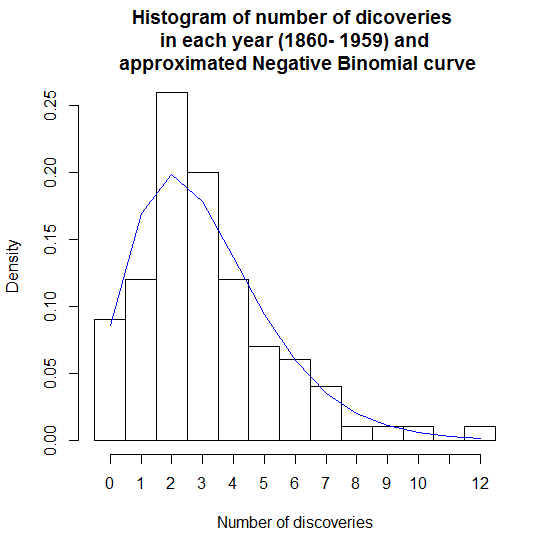
Hence, I tried with Negative binomial distribution, having pmf as;

with parameters p and r.

Again based on the method of moments, the estimated parameters comes out to be as follows:

Size or and probability of success, . Based on these, I have calculated the expected frequencies and finally coalesced the classes containing less than 5 expected frequencies. It gives the following table:

|  |  |  |
| --- | --- | --- |
| **Classes/ No of discoveries in a year** | **Observed frequency** | **Expected frequency** |
| 0 | 9 | 9 |
| 1 | 12 | 17 |
| 2 | 26 | 20 |
| 3 | 20 | 18 |
| 4 | 12 | 14 |
| 5 | 7 | 9 |
| 6 | 6 | 6 |
|  | 8 | 7 |
| **Total** | 100 | 100 |

 As seems from the above table, it should be a very good fit to the dataset. Indeed, the chi-square test statistic comes out to be on (8 - 1 - 2) = 5 degrees of freedom (Number of estimated parameters is 2). The expected value of is 5, and the observed one is pretty close to that. So, we might conclude that the fit seems good enough. The p-value of the chi-square is 0.501992, as it lies on the left hand side of the mean. (). The fitting distribution’s curve is given below:

One important thing to notice is that, there is no good interpretation of why should this dataset follows Negative Binomial Distribution. Clearly, it is not that case that whenever we have 4 or 5 “great” discoveries in a year, scientists stops thinking or does not try to discover anything else (as suggested by the estimated size of negative binomial distribution to be around 5.) Also, though we have a good interpretation to fit Poisson into the dataset, we find that it does not go so well.

However, Negative binomial distribution overfits the given dataset slightly, but as it is off by a slight margin, one may disregard the problems with overfitting and conclude that the fit is good.

**Critical Observation**

One critical observation could be that only the method of moments is not enough to know whether a distribution will fit the data or not. We should check whether the dataset has properties similar to the distribution or not.

* For instance, to fit Poisson to any data, we may consider checking whether the dataset has mean and variance almost same. Since, for Poisson distribution, mean and variance are same.
* To fit Binomial distribution to any dataset, we may check whether mean is more than variance or not. Since, binomial mean binomial variance .
* To fit Negative Binomial, we check the opposite that whether mean is lower than the variance in the dataset.
* To fit normal distribution, we may check skewness and kurtosis measures to see whether the frequency distribution in the dataset is approximately symmetric and mesokurtic.

Eventually, we need to look at the histogram of the dataset and think about all the known distributions that has the similar shape approximating the frequencies. We expect that it would give us a good chi-square test statistic of goodness of fit.