Part 1:

Discrete random variables:

1- Number of times I wash my hands per day. Domain is time (which day) and the range is the number of times.

2- Value of a dice roll. The domain is the time at which the dice was rolled, and its range is the value (from 1 to 6).

3- Number of males in a random sample of 100 students. The domain is the sample, and the range is the number of males (from 0 to 100).

4- Number of leaves in a tree picked at random. The domain is the tree, and the range is its number of leaves ([0, ∞[)

5- Number of words in the first page of a book picked at random. The domain is the book, and the range is the number of words.

Continuous random variables:

1- Weight of an apple picked at random. The domain is which apple we are measuring, and the range is weight (in kilograms).

2- Distance between two people picked at random. The domain is two-dimensional; the first dimension is who the first person is, and the second dimension is who the second person is. The range is the distance (in kilometers).

3- Diameter of a leaf picked at random. The domain is which leaf we are measuring, and the range is the diameter (in centimeters).

4- The time I wake up at on every morning. The domain is which day we are at, and the range is the time I wake up at.

5- The exact amount of a water in a water bottle picked at random. The domain is the water bottle in question, and the range is the quantity of water in it (in milliliters).

Part 2:

The data I used is the eruption times (in minutes) of the Old Faithful Geyser in Yellowstone National Park, Wyoming, USA. The data has a total of 272 measurements and was obtained from the R programming package: <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/faithful.html>. The raw data can be found in figure 1 and its estimated density function in figure 2.

Here is the raw data:

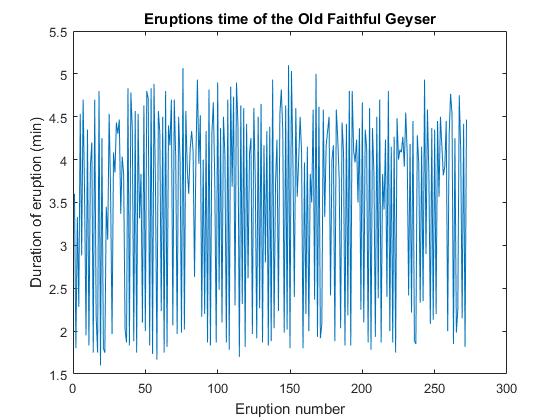


Figure 1: Eruption times of the Old Faithful Geyser across eruptions.

Here is the estimated probability density function:

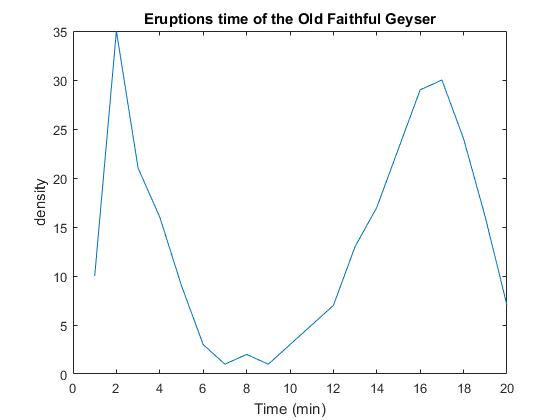
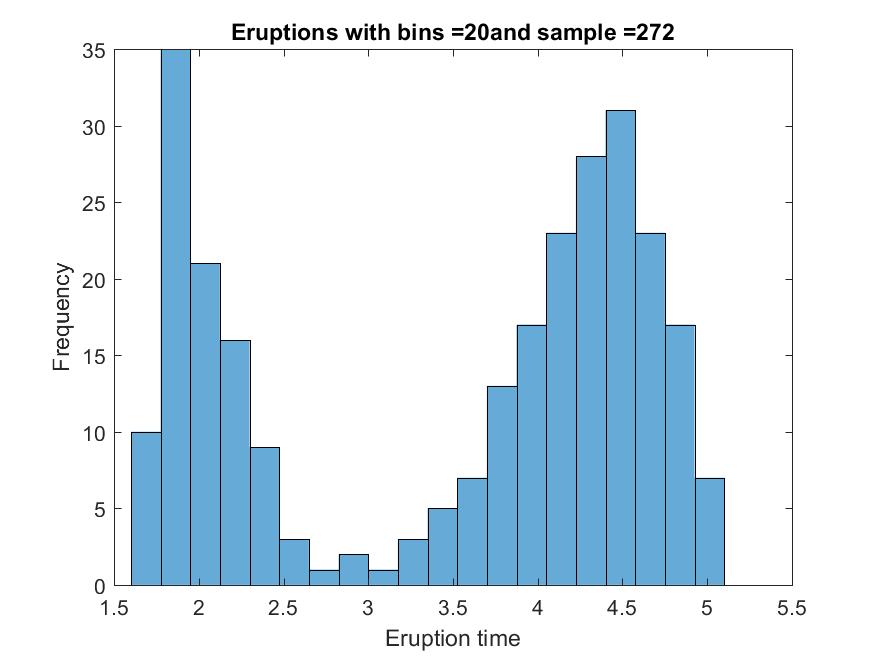
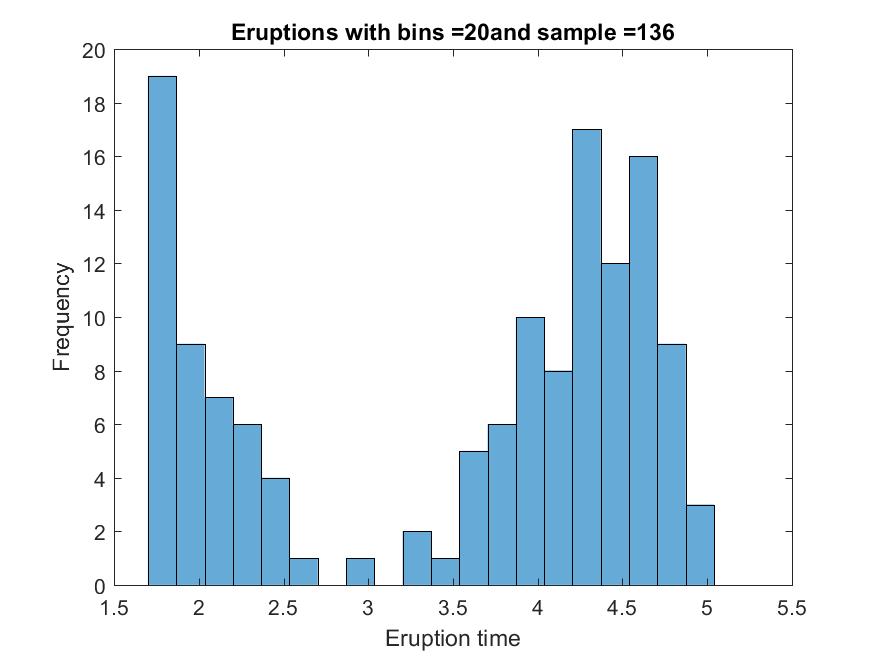
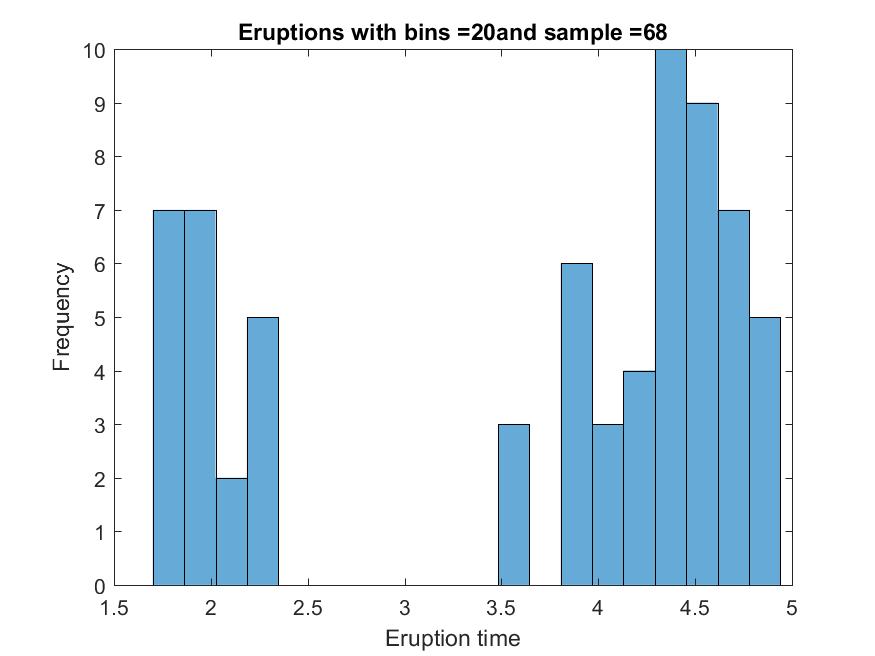
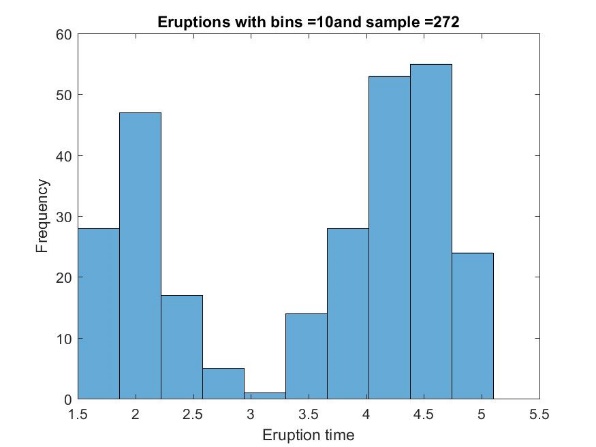
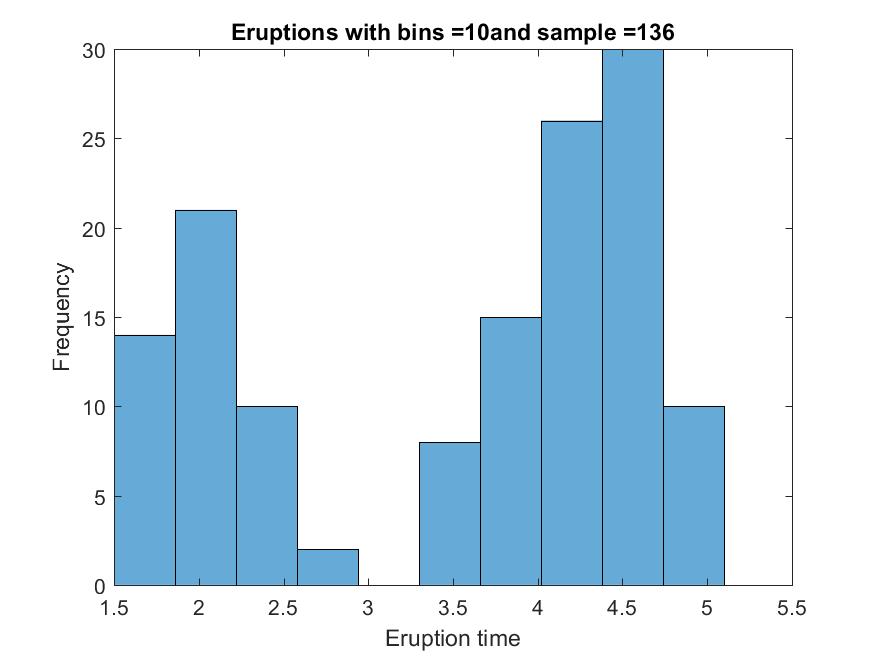
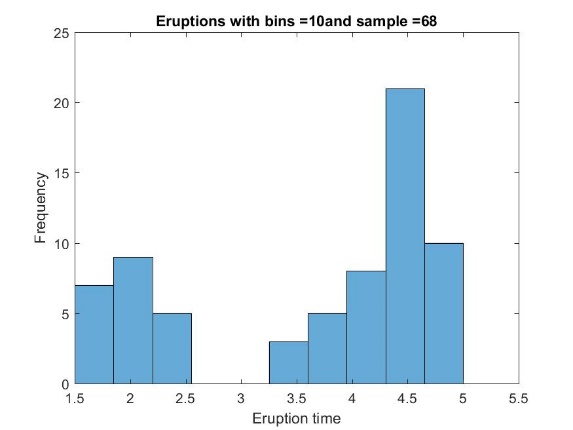


Figure 2: Estimated pdf of the eruptions times of the Old Faithful Geyser.

I investigated histograms with diverse number of bins and samples. I tried every combination with either 10, 20 or 40 bins, and with a sample of 272, 136 (half) or 68 (a quarter).



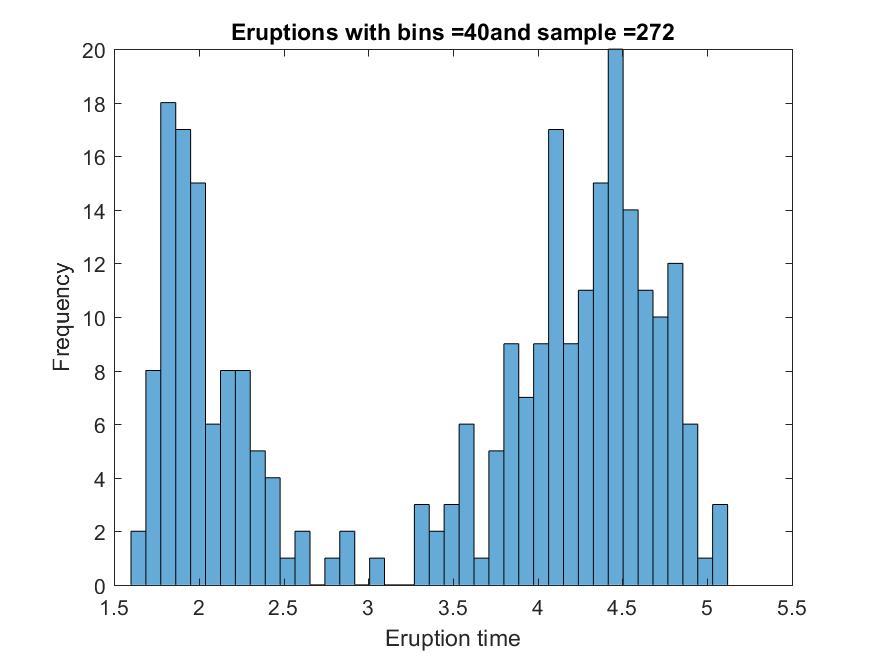
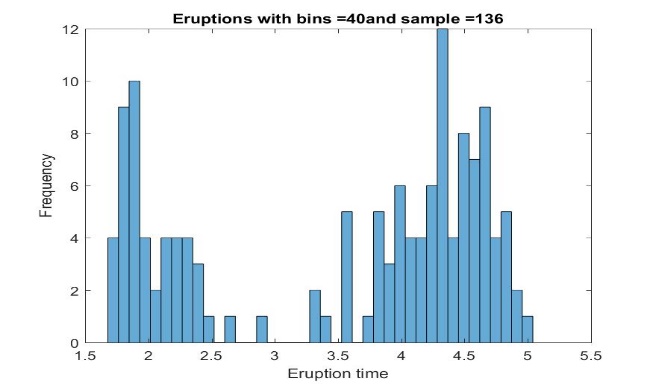
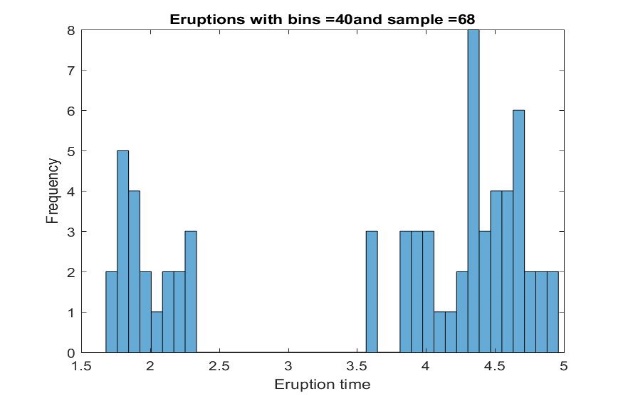


Figure 3: Frequency distribution of the eruption times estimated using different numbers of bins (10, 20 and 40) and different sample sizes (272, 136 and 68 data points sampled randomly).

As we can see in figure 3, the more bins we have, the more noise there is in the frequencies. Smaller samples also increase the noise in the histogram. The frequencies seem to follow a consistent pattern with 20 bins and the full sample; this suggests that our sample is large enough to estimate frequencies with 20 bins. 20 bins also seem to give frequencies that are precise enough.

Part 3:

The mean, median and mode of every variable are listed in table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Mode | Std | Type |
| X1 | 5.005947 | 4.991819 | -13.5101 | 5.214861 | Continuous |
| X2 | 5.021712 | 3.48154 | 0.000713 | 5.065176 | Continuous |
| X3 | -1.6E-17 | -4.9E-16 | 0.999921 | 0.707142 | Continuous |
| X4 | 3.502786 | 3.502741 | 2.001011 | 0.86825 | Continuous |
| X5 | 2.22E-20 | -4.9E-16 | 0.999921 | 0.707142 | Continuous |
| X6 | 5.0311 | 5 | 4 | 2.245359 | Discrete |
| X7 | 14.0079 | 14 | 14 | 2.048672 | Discrete |
| X8 | 0.4949 | 0 | 0 | 1.117541 | Discrete |
| X9 | 26.9999 | 26.87757 | -3.97136 | 14.57826 | Continuous |
| X10 | -0.00946 | -0.03595 | -11.1221 | 4.458542 | Continuous |

Table 1: Mean, median, mode, standard deviation and type (continuous or discrete) of each variable from assignment 2.

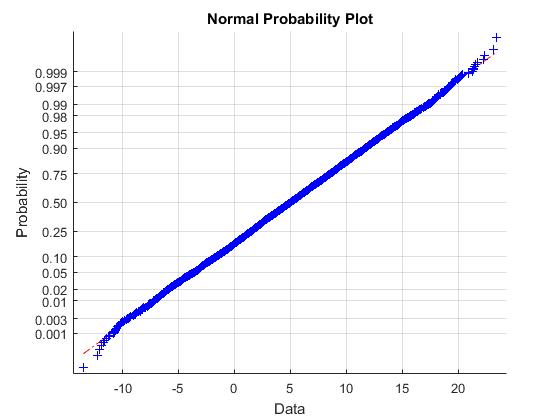


Figure 4: Normal probability plot for x1. The x-axis are frequency values obtained from the data, and the y axis is the corresponding frequency assuming the density function is normal.

As is shown in Figure 4, X1 is a normal distribution. We can estimate the mean and standard deviation of the distribution using the mean (5.006) and standard deviation (5.21) of the sample.

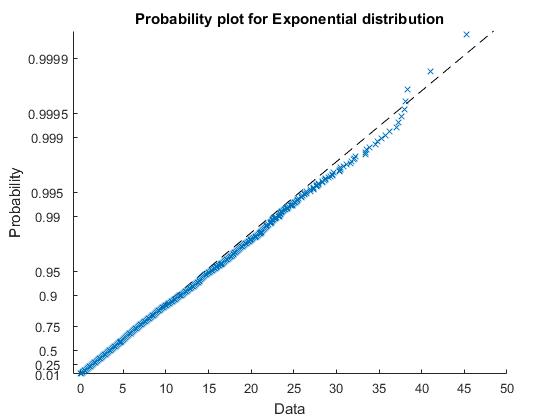


Figure 5: Exponential probability plot for x2. The x-axis are frequency values obtained from the data, and the y axis is the corresponding frequency assuming the density function is normal.

As shown in Figure 5, X2 can be very well approximated by an exponential distribution. We need to know mean of the distribution, which we can estimate by the mean of the sample (µ = 5.022).

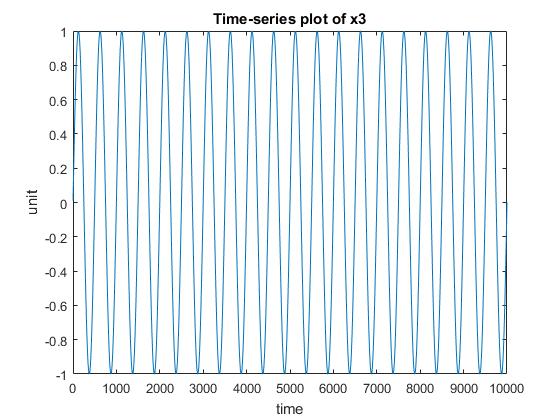


Figure 6: Time plot of x3. The x-axis is time and the y-axis represent the values of the signal.

X3 is a deterministic sinusoid function. It follows the function sin(x\*pi/250).

X4 is a uniform distribution. We need to know the minimum (min = 2.001) and maximum (max = 4.9996) of the sample.

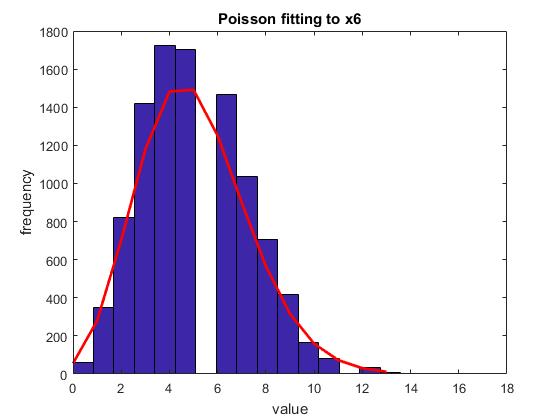


Figure 7: Fitting of a Poisson distribution to x6.

X6 is a Poisson distribution. We can estimate µ (lambda) using the mean of the sample (5.0311).

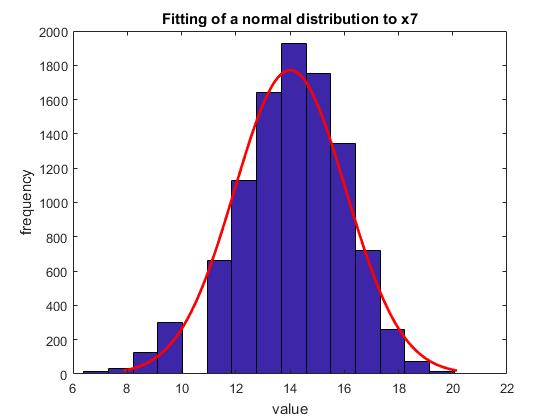


Figure 8: Fitting of a normal distribution to x7.

X7 is a binomial distribution, which is supported by the fact it can be approximated properly using a normal distribution. To estimate the distribution, we need to know the number of independent events (N) and the probability of a success occurring (p). Since there are 10 000 data points, it is safe to assume the number of independent events per data point is equivalent to the maximum of the distribution (N = 20). Knowing this, we can estimate p by dividing the mean of the distribution by N (p 14/20 = 0.7).

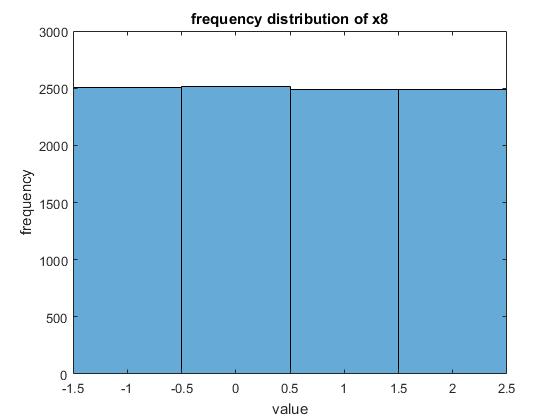


Figure 9: Frequency distribution of x8.

X8 is a discrete rectangular distribution. We need to know the total number of possible outcomes in the sample (N = 4).

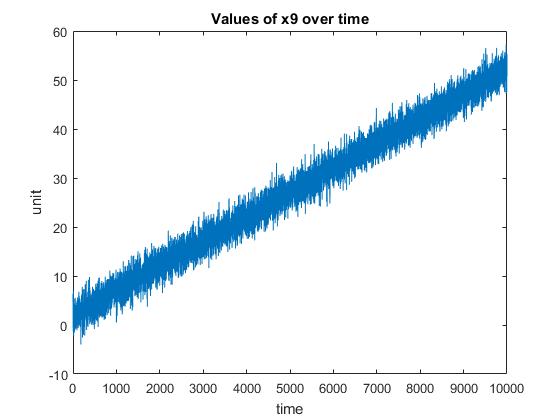


Figure 10: Time-series data of x9. The x-axis represents time and the y-axis the values.

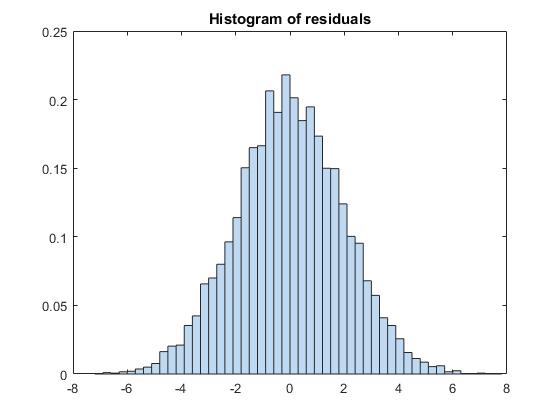


Figure 11: Probability density function of the residuals of x9 after fitting a linear model.

X9 is a linear model with some gaussian noise, as can be seen in figures 10 and 11. The linear model is y = 0.05x + 2. The mean of the residual is 0 with a standard deviation of 1.97.

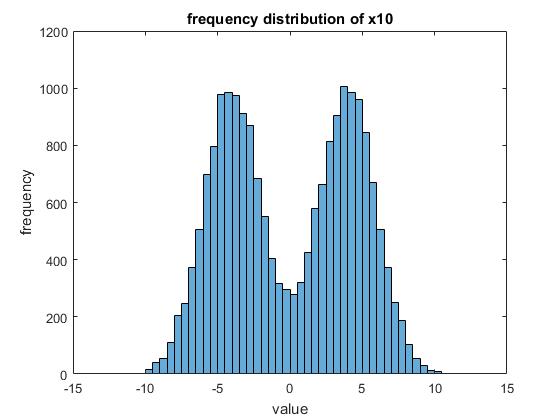


Figure 12: Frequency distribution of x10.

X10 is a normal bimodal distribution. We need to know the mean and standard deviation of both normal distributions. The means are -4 and 4, respectively. Using a homemade method, the standard deviation of both normal distributions was estimated to be approximately 2.3.

Part 4:

Here are the answers to questions a) through g) for all of the 10 signals.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | a) | b) | c) | d) | e) | f) | g) |
| X1 | 0.4986 | 0.159 | 0.1537 | 0.6873 | 0.0233 | 0.0253 | 0.9514 |
| X2 | 0.3638 | 0.1341 | 0 | 0.8659 | 0.0507 | 0 | 0.9493 |
| X3 | 0.4998 | 0.25 | 0.25 | 0.5 | 0 | 0 | 1 |
| X4 | 0.5 | 0.2111 | 0.214 | 0.5749 | 0 | 0 | 1 |
| X5 | 0.4997 | 0.25 | 0.25 | 0.5 | 0 | 0 | 1 |
| X6 | 0.3925 | 0.1422 | 0.1226 | 0.7352 | 0.0296 | 0.0058 | 0.9646 |
| X7 | 0.4165 | 0.1066 | 0.1134 | 0.78 | 0.0086 | 0.0173 | 0.9741 |
| X8 | 0.4973 | 0.2485 | 0.2509 | 0.5006 | 0 | 0 | 1 |
| X9 | 0.4983 | 0.2059 | 0.205 | 0.5891 | 0.0003 | 0.0004 | 0.9993 |
| X10 | 0.4991 | 0.2058 | 0.20405 | 0.59015 | 0.0033 | 0.0034 | 0.9933 |