Applied Statistics

Exam

Andras Csepreghy
xgj708@alumni.ku.dk

University of Copenhagen — January 17, 2020

I – Distributions and probabilities:

1.1 Question: Assuming the "El Clasico" football match is an even game ($p = 0.5$), what is the probability, that the score after 144 non-draw league games is exactly even?

1.1 Answer: The probability of a football team winning n times with probability p follows a **binomial distribution** shown in Figure 1 with $p = 0.5$ and number of trials $n = 144$. The distribution for this problem is binomial because we have a discrete number of events each with an unchanging probability p. Since we need an exactly even score the question can be rephrased as such: What is the probability of the team winning exactly 72 times.

The probability of winning exactly 72 times can be retrieved from the PDF by plugging in numbers for $x = 72$ or by this equation:

$$\binom{144}{72} = \frac{144!}{72!(144-72)!}\left(\frac{1}{2}\right)^{72}\left(\frac{1}{2}\right)^{72} = 0.06637$$

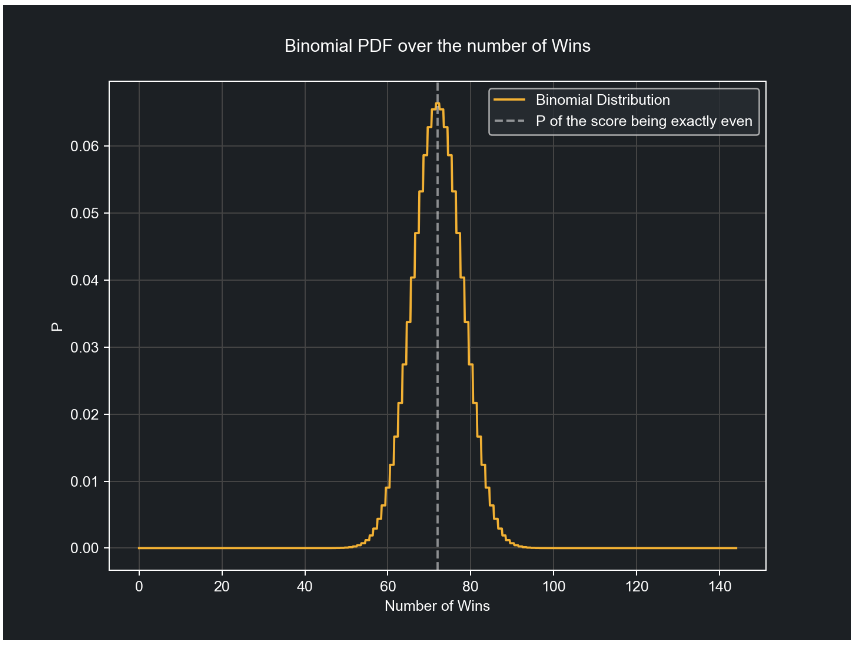


Figure 1: Binomial Distribution with $p = 0.5$ and number of trials $n = 144$

1.2 Question: Brad Pitt and Edward Norton are shooting golf balls at a window with $p_{\text{hit}} = 0.054$ chance of hitting. How many golf balls do they need to be 90% sure of hitting the window?

1.2 Answer: Since the probability of hitting the window is $p = 0.054$, the probability of hitting the window in n tries is $1-(1-p)^n$, which nicely enough gives 0.054 for $n = 1$. To calculate the solution I used a loop increasing n by 1 each time since we are considering a discrete number of tries (balls). At $n = 42$ we obtain a $p = 0.9028$ for hitting the window (I have a sense that 42 was intentional here). Figure 2 shows the probabilities for hitting the window for each n between 0 and 80. To state it again **Brad Pitt and Edward Norton need 42 golf balls to be 90% sure of hitting the window**. I bet Tiger Woods could do it in 10.

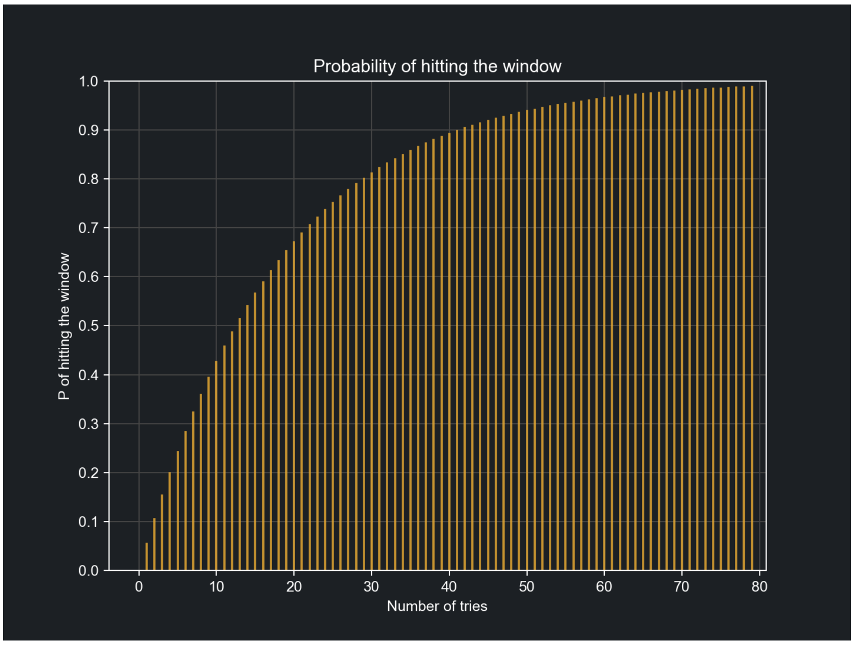


Figure 2: Probability of hitting the window for each n tries

II – Error propagation:

2.1 The Hubble constant h has been measured by seven independent experiments: 73.5 ± 1.4, 74.0 ± 1.4, 73.3 ± 1.8, 75.0 ± 2.0, 67.6 ± 0.7, 70.4 ± 1.4, and 67.66 ± 0.42 in (km/s)/Mpc.

Question: What is the weighted average of h? Do the values agree with each other?

Answer: To calculate the weighted mean of h I used the formula:

$$\hat{\mu} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}$$

This and the formula for calculating the uncertainty on the weighted mean gave the weighted mean for h to be $h = 70.14 \pm 0.38$. (I rounded 70.139 and 0.377) See Figure 3 for the plot. To check whether these values agree with each other I used chi2:

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \mu_i)^2}{\sigma_i^2}$$

The chi2 for these 7 values is:

$$Prob(\chi^2 = 70.39, N_{dof} = 6) = 3.389 \times 10^-13$$

I define the null hypothesis to be that these 7 values agree with each other. From the p value obtained from the chi2 test, I can safely reject the null hypothesis choosing a 1% threshold.

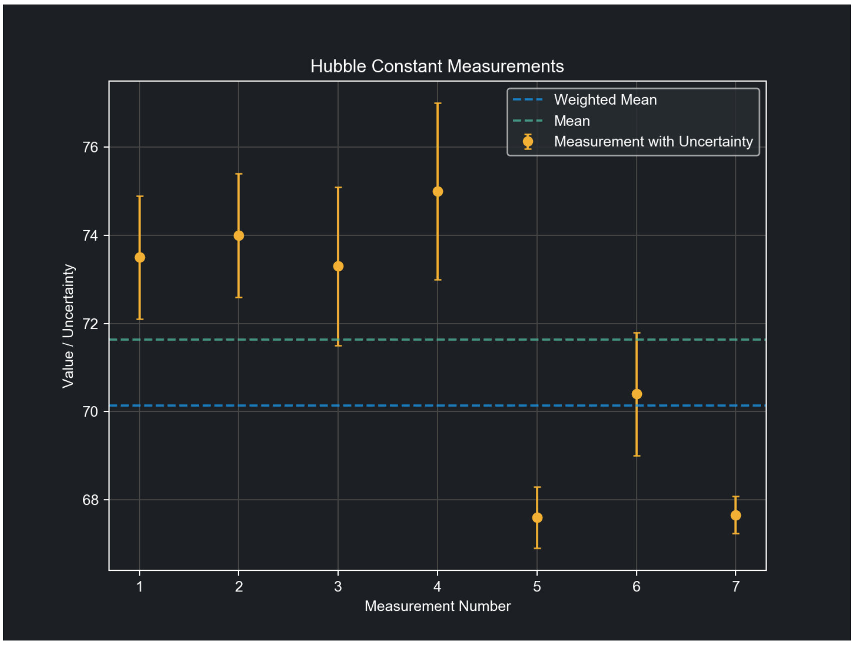


Figure 3: The 7 Hubble constant measurements with their uncertainties

Question: The first four measurements are based on a different method than the last three. Do the values from the same method agree with each other?

Answer: Using the same formulas on the first 4 measurements I obtained these values:

$$Prob(\chi^2 = 0.50, N_{dof} = 3) = 0.92$$

Using the same threshold of 1% I can safely conclude that these 4 values agree with each other. As for the last 3 measurements the chi2 test shows

$$Prob(\chi^2 = 4.18, N_{dof} = 2) = 0.12$$

Using the same threshold of 1% I can safely conclude that these 3 values also agree with each other. They don't agree as much as the first 4 did, but this is because those 4 have significantly larger uncertainties, which makes their chi2 value lower, hence resulting in a better p.

2.2 Using Coulomb's law you want to measure a charge, $q_0 = Fd^2/k_eQ$. Assume that Coulomb's constant $k_e = 8.99 \times 10^9$ Nm^2/C^2 and the instrument charge $Q = 10^{-9}$ C are known.

Question: Given force $F = 0.87 \pm 0.08$ N and distance $d = 0.0045 \pm 0.0003$ m, what is q_0?

Answer: Because the text doesn't mention correlation I will assume that F and d are uncorrelated. Here is the general formula with which one can calculate the error propagation in $q0$:

$$\sigma_y^2 = \sum_i^n \left[\frac{\delta y}{\delta x_i}\right]_{\bar{x}=\bar{y}}^2 \sigma_i^2$$

I prefer simulation, so I will continue explaining how I used simulation to do error propagation. I first took $F = 0.87 {\pm} 0.08$ and generated 10,000 normally distributed numbers having a mean of 0.87 and a standard deviation of 0.08. I repeated the same for $d = 0.0045 \pm 0.0003$ with normally distributed numbers having their mean and standard deviation corresponding to the value and uncertainty of d accordingly. This gave the result for q_0:

$$q_0 = 1.95 \pm 0.37 \times 10^{-08}$$

In order to calculate the error propagation without simulation one would have to calculate the derivative of Fd^2 with respect to F and d and then substitute in the formula written out above.

Question: Where does the largest contribution to the uncertainty on q_0 come from? F or d?

Answer: As d appears squared while F appears linearly, d has to be determined with twice the relative precision compared to F.

$$2 \times \frac{\sigma(d)}{d} = \frac{\sigma(F)}{F}$$

Twice the relative error for d:

$$2 \times \frac{0.0003}{0.0045} = 0.13$$

Relative error for F:

$$\sigma_d \frac{0.08}{0.87} = 0.092$$

Which means d contributes more to the error than F

Question: If you could measure F and d with uncertainties $\pm\, 0.01$ N and $\pm\, 0.0001$ m respectively, at what distance should you expect to measure the charge in question q_0 most precisely?

Answer: (Ran out of time) For this problem I would first rewrite the uncertainties in F and d, then create a function in which I leave d as a free parameter. Then I would re-do the simulation and keep track of what the uncertainty of the final result is for each iteration. Then pick the value of d that minimized the uncertainty of q_0. The uncertainties should fall on a parabola.

2.3 Sub-saharan humans tend not to have any Neanderthal DNA, while all others have a few percent. The file: http://www.nbi.dk/ petersen/data_NeanderthalDNA.txt contains the fraction of Neanderthal DNA for 2318 Danish high school students.

Question: Plot the distribution of Neanderthal DNA fraction, and calculate the mean and RMS.

Answer:

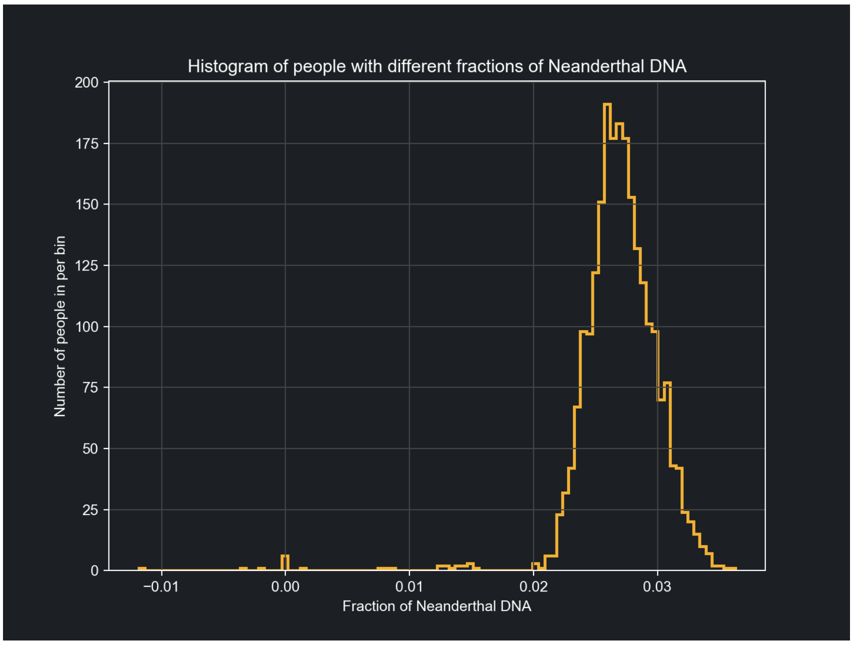
- The plot can be seen in **Figure 4**.
- Mean: **0.0269**
- RMS: **0.00335**

Question: Do you find any mismeasurements or outliers from the main population in the data?

Answer: I can see two kinds of discrepancies in the data, one of which can be easily explained without invoking errors in the data.

- There is a significant bump at the number 0.0, which is not surprising since we can assume some fraction of the high school students are descendants of African families, and therefore have no Neanderthal DNA.
- There are 3 negative fractions (-0.00208, -0.01185, -0.00326) for which I found no good explanation except human error. In this dataset a negative fraction isn't valid. After some consideration I would either remove them or make them equal to 0.

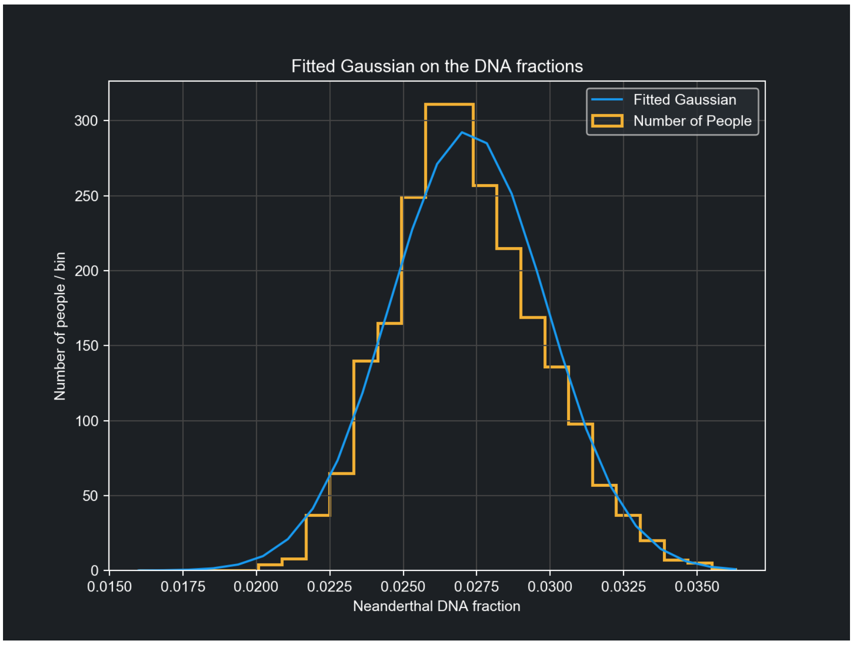
Figure 4: The distribution of fractions of Neanderthal DNAs in people



Question: Fit the main population data with distributions of your choice, and comment on the fits.

Answer: Choosing the minimum value of x to be 0.15 still doesn't give a good Gaussian fit, however limiting the population to have Neanderthal DNA from 0.16 to 0.0363 gave better results. I chose the upper bound to be 0.0363 since there appears to be no suspicious values at the upper end. The resulting Gaussian fit seems promising but the p value the chi2 gives is very low. Chi2 = 52.617, p = 0.000649.

Figure 5: Gaussian fit with the distribution of fractions of Neanderthal DNAs in people



III – Monte Carlo:

3.1 Assume that the outcome of an experiment can be described by first drawing a random number x from the distribution $f(x) = C(c_1 + x^{c_2})$ for $x \in [1, 3]$, where $c_1 = 1$ and $c_2 = 2$ and then using this x value to calculate $y = x\exp(-x)$.

Question: What is the value of C? And what is the mean and RMS of $f(x)$?

Answer: To figure out the value of C I computed the original PDF's integral which is 10.666. To get the value of C it is simply $C = \frac{1}{10.666}$. I recomputed its integral with this C and it conveniently gives 1.

- Mean = **0.500**
- RMS = **0.218**

Question: What method(s) can be used to produce random numbers according to $f(x)$? Why?

Answer: The two methods that are most convenient to use for this PDF are the *Transformation Method or Inverse Transform Sampling* and *Accept-Reject Method or Rejection Sampling* also known as *Von Neumann Method*.
Inverse Transform Sampling:
For the transformation method it is necessary that the $f(x)$ can be integrated, and then its integral can be inverted. The integral of a PDF is known as a Cumulative Distribution Function (CDF). So we are looking for the inverse of the PDF's corresponding CDF. Then we can sample u from a uniform distribution $u_i \sim U(0,1)$ then $x_i = CDF^-1(u_i)$, so x will be sampled according to our original $f(x)$. In our case it looks like $f(x)$ can be integrated and its integral can be inversed. When it is possible to use this method it is more computationally effective since we don't throw away any sampled numbers where as in Von Neumann method we reject some portion of our sampled numbers.
Von Neumann Method:
Rejection sampling requires a the function to be finite in x and y, which works well in our case. However even with a function that is infinite in x and y can be approximated by choosing large enough upper bounds. Using this method we can generate random numbers according to $f(x)$ by first $x_i \sim U(1,3)$ and then $y_i \sim U(0.1875, 0.9375)$. Reason for choosing 0.9375 as the upper bound is because it is the largest value the function takes while 0.1875 is the smallest one which it has in $x = 1$. Then we decide to accept or reject the value for x if the sampled y value falls under the curve of $f(x)$ meaning its value is smaller, otherwise we reject it and sample again. As the number of samplings go to infinity the sampled random numbers approximate the distribution more and more.

Question: Produce 5000 random numbers distributed according to $f(x)$ and y and plot these.

Answer:

To produce these numbers I used the Accept-Reject method. Please see Figure 6

Question: What is the linear (Pearson) correlation between the produced x and y values?

Answer: The Pearson correlation between x and y is: **0.992**

Question: Fit the distribution of the produced x values to $f(x)$, with c_1 and c_2 as parameters.

Answer: To fit the probability distribution to the sampled values I minimized its chi2 leaving c1 and c2 floating. This resulting in having c1 = 0.97223 and c2 = 2.0174. Rerunning gives similar results, within about 3% range if these above stated values. Please see Figure 7 for results.

Question: How many measurements of x would you need, in order to determine c_1 and c_2, respectively, with a precision better than 1% of their values?

Answer: This may vary depending how many bins one chooses. For this solution I chose a 100 bins and used brute force in a for loop to check different values. I started getting good results around 80,000, however interestingly enough at a 100,000 c1 stayed around 2-3% off. From 200,000 and up they both stayed very close to being in the 1% range

IV – Statistical tests:

Figure 6: Randomly sampled values (orange) according to f (blue)

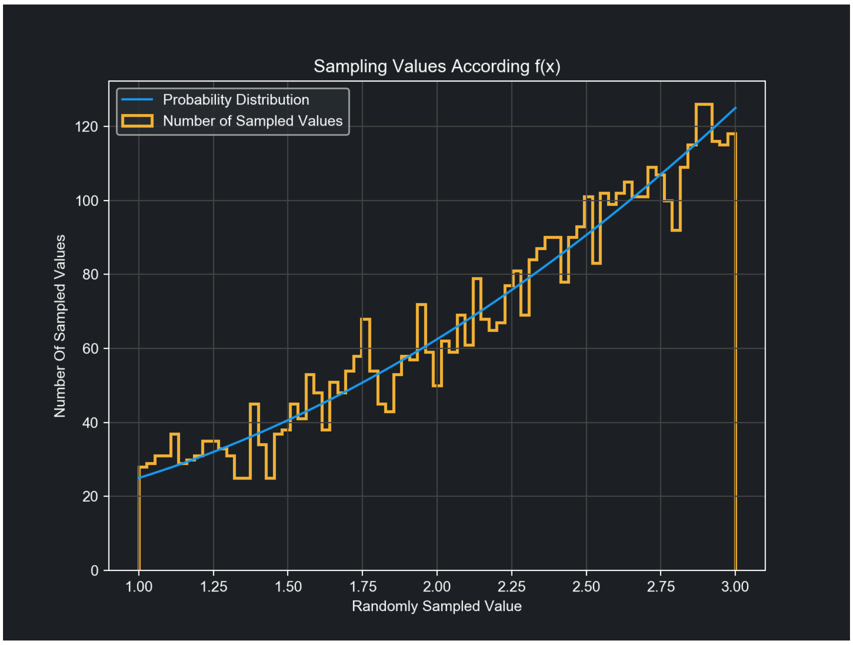
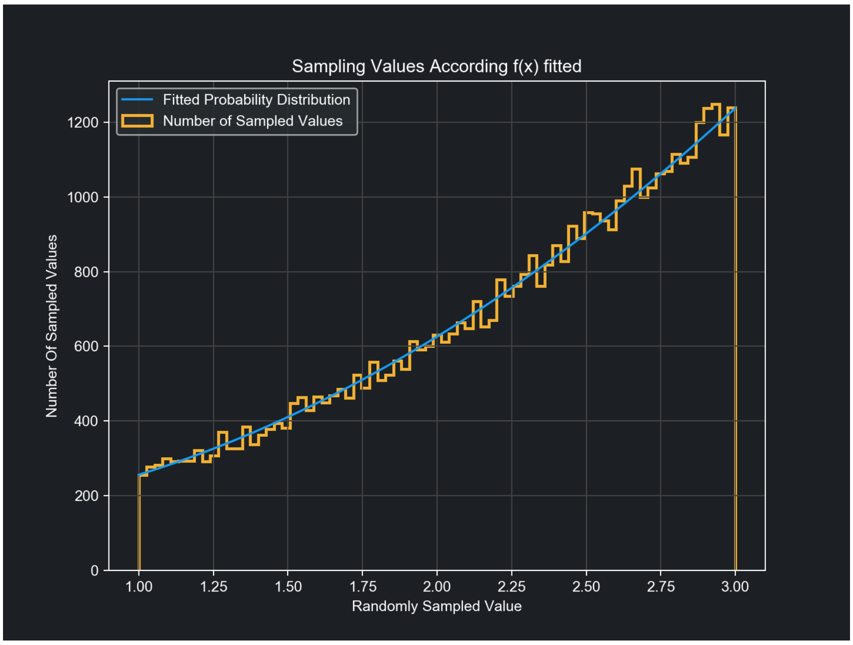


Figure 7: f(x) fitted to the sampled values by minimizing its chi2



4.1 The length (l in μm) and transparency (T) of two types of cells (P and E) can be found for 4690 cells in the file: http://www.nbi.dk/ petersen/data_Cells.txt**www.nbi.dk/**$\sim$**petersen/data_Cells.txt**.

Question: Selecting P-cells by requiring $l < 9\,\mu$m what is the rate of type I and type II errors?

Answer: Type I error is when we reject a true null hypothesis also called a false positive. Type II error is accepting a false null hypothesis, also called false negative. In this example by making the cut at $l < 9\,\mu$m we get:

- Type I Error: 72
- Type I Error Rate: 0.0375
- Type II Error: 72

- Type II Error Rate: 0.0970

Question: Which of the two variables l and T is best at distinguishing between P and E cells?

Answer: I may have found an unusual way of solving this, but I believe it is a valid solution nonetheless. I chose a simple machine learning classifier called Support Vector Machine (SVM), then took all the data and divided it to either only have the size parameter or the transparency parameter. Then I trained these two classifiers on the full dataset gaining to accuracy scores.

Having only the size resulted in an predicting accuracy of 0.9341 while having only the transparency resulted in an accuracy of 0.9179. To further test this method one could run a few different classifiers. However, from this expoeriment I conlcude that the size parameter has a better predictive power or in other words separates the data beter.

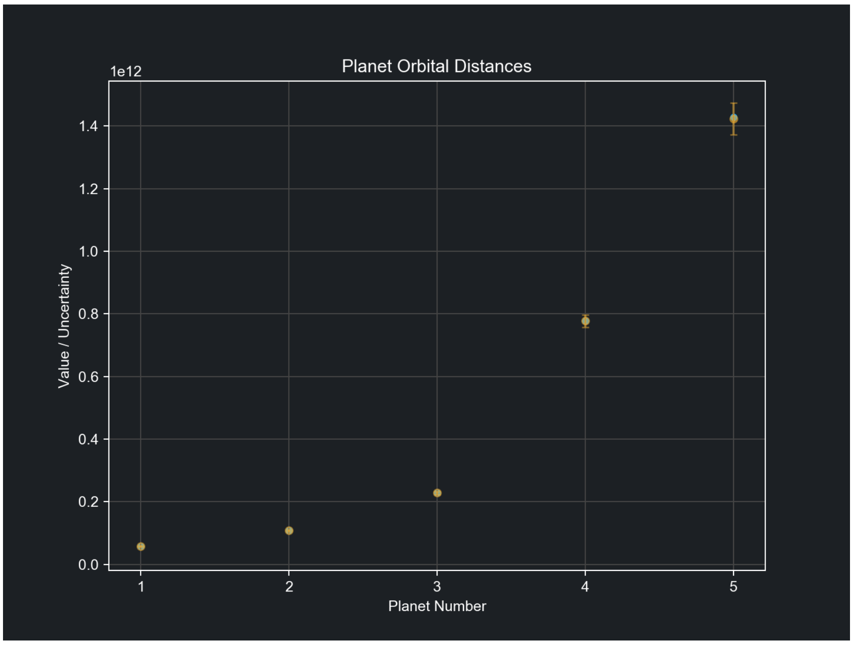
Question: Separate P and E cells using l and/or T, and draw a ROC curve of your result.

V – Fitting data:

5.1 Question: Plot the five non-Earth values and fit these to Kepler's third Law: $a = C \times T^{2/3}$.

On Figure 8 the slightly bigger orange dots are the values with their uncertainties that are given in the dataset and in the middle of each dot there is a faint blue one that shows the fitted values. It's hard to see, because they fit perfectly, couldn't make them nicer. Having a log scale is no help either. I obtained the fit by minimizing the chi2. There is less than 1% difference between the measured values and the ones I obtain from the fit. From the fit I get a chi2 of 0.0519 and a p of 0.99.

Figure 8: a values fitted with Kepler's third law



Question: In this fit, which planet seems to follow this relation least well? Is it critical?

Answer: The biggest difference between observed and predicted by the model comes from Saturn, but looking at the chi2 it does not appear significant.

Question: From the value you obtain for C and G_{1778} estimate the solar mass $M = 4\pi^2 C^3/G$ in kg.

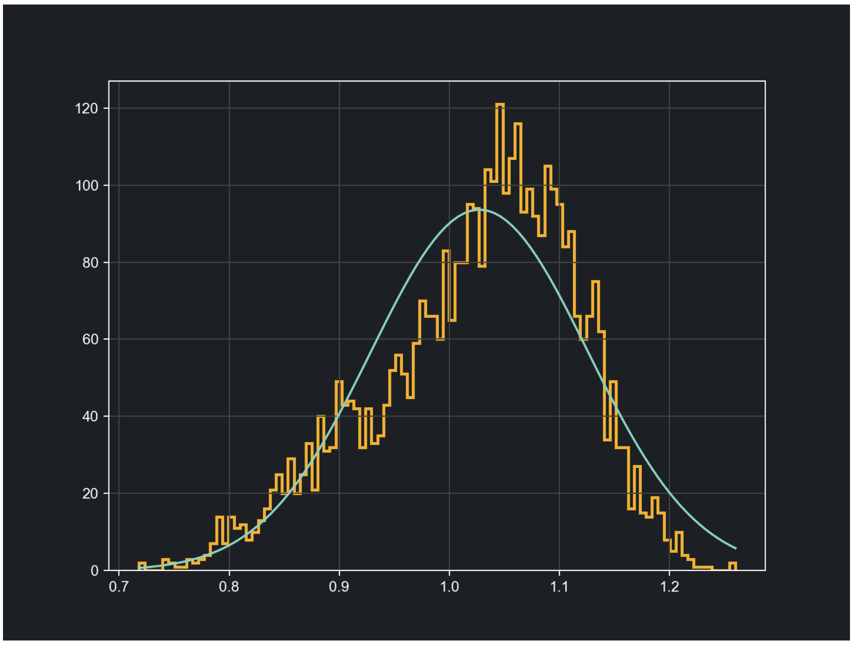
Answer: I have not had time to do this, however for this solution I would do the fitting again by minimizing the chi2 with the large values and see if the chi2 obtain from this fit is better than the one without the extra parameters. My gut feeling is, that the fit would improve.

Question: Expand the fit to Kepler's third law by further adding two parameters: $T = C \times (a^{c_1} + c_2)$. Does this formula match the data well? Are the two additional parameters necessary?

5.2 Question: What is the resolution of β_{init}? And is it consistent with a Gaussian distribution?

Answer: To fit a Gaussian distribution I have minimized the chi2 as in previous exercises, see Figure 9. I have done a Kolmogorov-Smirnov test between the fitted Gaussian and the histogram of beta values. The Kolmogorov-Smirnov test between resulted in a p value of 0.177 when using 200 bins and 0.47 when using 100. From these p values I cannot reject the null hypothesis which is tha these two distributions are the same. However I must mention the Kolmogorov-Smirnov test is meant for continuous values although it is sometimes a good approximation for binned values too. Even though I cannot reject the null hypothesis based on the KS test, I looking at the p value that the chi2 gave, 4.2925e-33, I can conclude that a gaussian is not a good approximation. It seems that the data is producing an assymetrical distribution.

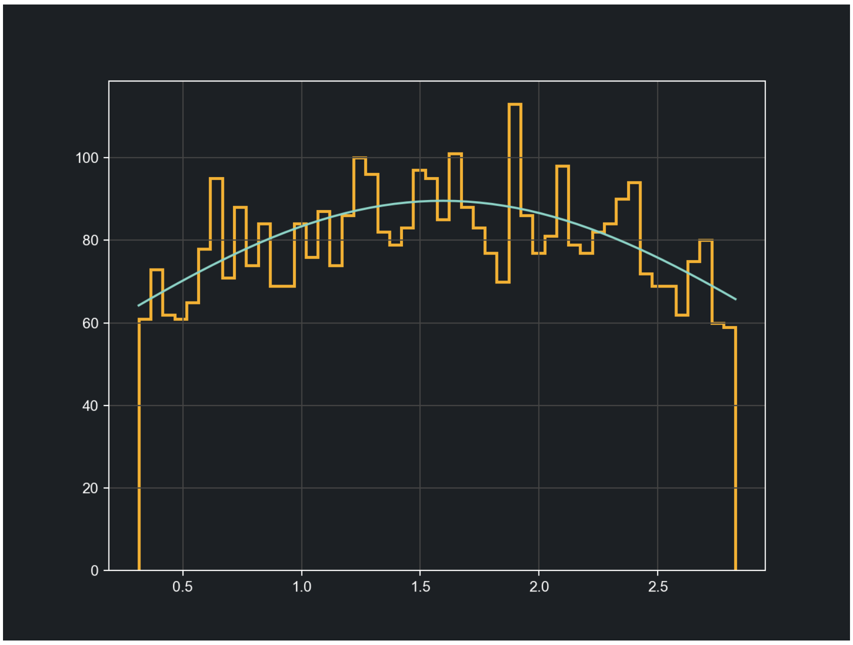
Figure 9: Gaussian fit on beta



Question: Is the distribution in θ consistent with being symmetric around $\pi/2$?

Answer: The mean of theta is 1.57551 which is within 0.3% of $\pi/2$, and to check its symmetry I fitted a Gaussian to it by minimizing the chi2. The chi2 after the fit is 53.101 and the corresponding p value is 0.3191 from which I would conclude that the theta is consistent with being symmetric around $\pi/2$

Figure 10: Gaussian fit on theta


100
80
60
40
20
0
0.5
1.0
1.5
2.0
2.5