Applied Statistics
Problem set

Simon C. Debes, rlq306

January 4, 2021

Problem 1 - Distributions and probabilities

1.1

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?

Because there only are two possible outcomes, this can be described binomially, with $n = 144, r = n/2$.

$$P(r;p,n) = p^r(1-p)^{n-r}\frac{n!}{r!(n-r)!} \tag{1}$$

$$= 0.5^{72}(1-0.5)^{72}\frac{144!}{72!(72)!} = 0.0664$$

The probability that the score is exactly even after 144 non-draw league games is $0.0664 = 6.64\%$.

1.2

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

We'll reword it to consider the how many balls they would need to hit to have 10% chance of not hitting, and then subtract that from 1. Now $P = 0.1, r = 0$.

$$P(r;p,n) = p^r(1-p)^{n-r}\frac{n!}{r!(n-r)!}$$

$$0.1 = (0.946)^n \Rightarrow n = 41.478$$

So if they hit 42 golf balls, they can be 10% sure of not hitting i.e. 90% sure of hitting.

Problem 2 - Error propagation

2.1

The Hubble constant h has been measured by seven independent experiments:

$73.5 \pm 1.4, 74.0 \pm 1.4, 73.3 \pm 1.8, 75.0 \pm 2.0, 67.6 \pm 0.7, 70.4 \pm 1.4,$ **and** 67.66 ± 0.42 **in** $(km/s)/Mpc$.

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

The weighted mean is

$$\hat{\mu} = \frac{\sum(x_i/\sigma_i)^2}{\sum(1/\sigma_i)^2}, \tag{2}$$

and the weighted error is

$$\hat{\sigma} = \sqrt{\frac{1}{\sum(1/\sigma_i)^2}}. \tag{3}$$

So the weighted mean is

$$\hat{\mu} = (68.8 \pm 0.3)(km/s)/Mpc$$

Taking the χ^2, we get a probability of $1.45 \cdot 10^{-9}$, which is very low, so the values do not agree with each other.

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?

The χ^2 probability of the first 4 measurements is 0.92, which is very nice, and the measurements agree. The χ^2 probability of the last 3 measurements is 0.16, which is alright too, and the measurements agree.

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

First, to find the value of q_0 we plug in the values and get:

$$q_0 = 0.87N \cdot (0.0045m)^2/(8.99 \cdot 10^9 Nm^2/C^2 \times 10^{-9}C) = 1.963 \cdot 10^{-6}C$$

Then to find the uncertainties, we'll use

$$\frac{\sigma_{x^y}}{x^y} = y\left(\frac{\sigma_x}{x}\right) \implies \frac{\sigma_{d^2}}{d^2} = 2\left(\frac{\sigma_d}{d}\right)$$

to write

$$\frac{\sigma_{q_0}}{q_0} = \sqrt{\left(\frac{\sigma_F}{F}\right)^2 + 2\left(\frac{\sigma_d}{d}\right)^2}$$

$$\frac{\sigma_{q_0}}{q_0} = \sqrt{\left(\frac{0.08N}{0.87N}\right)^2 + 2\left(\frac{0.0003m}{0.0045m}\right)^2} = 0.1317$$

$$\sigma_{q_0} = 0.1317 \cdot q_0 = 3 \cdot 10^{-7}$$

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

The contribution of d is larger because the F-part:

$$\left(\frac{\sigma_F}{F}\right)^2 = \left(\frac{0.08N}{0.87N}\right)^2 = 0.085$$

is smaller than the d-part:

$$2\left(\frac{\sigma_d}{d}\right)^2 = 2\left(\frac{0.0003m}{0.0045m}\right)^2 = 0.089$$

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

Now we can write the uncertainty of q_0 as a function of d. We then differentiate it, set it equal to 0, and solve for d.

$$\sigma_{q_0}(d) = q_0\sqrt{\left(\frac{\sigma_F}{F}\right)^2 + 2\left(\frac{\sigma_d}{d}\right)^2} = q_0\sqrt{\left(\frac{0.01N}{0.87N}\right)^2 + 2\left(\frac{0.0001m}{d}\right)^2}$$

$$\sigma'_{q_0}(d) = q_0\sqrt{\left(\frac{\sigma_F}{F}\right)^2 + 2\left(\frac{\sigma_d}{d}\right)^2} = -\frac{2\sigma_d^2}{d^3\sqrt{\left(\frac{\sigma_F}{F}\right)^2 + \left(\frac{\sigma_d}{d}\right)^2}}$$

This function doesn't have a minimum. Seems like the larger the d, the smaller the σ_{q_0}.
FORGET IT
From the equation given in the problem text, I isolated F, such that I could substitute F for q_0kQ/d^2, and then express σ_q as a function of d, only:

$$\sigma_q = \sqrt{\left(\frac{0.01}{q_0kQ/d^2}\right)^2 + 2\left(\frac{0.0001}{d}\right)^2}\, q_0$$

I plotted σ_q as a function of d, and found that the minimum for σ_q is at $d = 0.0056$.

2.3

Sub-saharan humans tend not to have any Neanderthal DNA, while all others have a few percent. The file: www.nbi.dk/ petersen/data_DNAfraction.txt contains the fraction of Neanderthal DNA for 2318 Danish high school students. Plot the distribution of Neanderthal DNA fraction, and calculate the mean and RMS.

Figure 1: Histogram of the students' fractions of Neanderthal DNA.

The mean is the sum of the data points divided by the number of data points. The root mean square is...

$$\text{Mean} = 0.027$$

$$\text{RMS} = 0.003$$

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

There were negative data points which I discarded, before analysis, since you can't be negative part Neanderthal. I also calculated that with our sample size of 2309, we would expect 0.2 data points to lie outside 4 sigma, so I discarded any data point outside 4 sigma.

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

Figure 2: Histogram of the students' fractions of Neanderthal DNA.

After cleaning up the data, the probability rose many orders of magnitude, but is still quite low.

Problem 3 - Monte Carlo

3.1

Assume that the outcome of an experiment can be described by first drawing a random number x from the distribution $f(x) = C(c_1 + c_2)$ for $x \in [1, 10]$, where $c_1 = 5$ and $c_2 = 2$ and then using this x value to calculate $y = xexp(-x)$.

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

We'll want our distribution f to be normalized, so we'll demand that its integral over the interval on which it is defined is equal to one:

$$\int_1^{10} C(5 + x^2)dx = \left[C\left(5x + \frac{1}{3}x^3\right)\right]_1^{10}$$

$$= C\left(50 + \frac{1000}{3}\right) - C\left(5 + \frac{10}{3}\right) = 1$$

$$\implies C = \frac{1}{378}$$

I'll find the mean and RMS like it is done for a Gaussian in eq. (3.19) and (3.20) in Barlow.

$$\mu = \int_1^{10} xf(x)dx = 7.268$$

$$\sigma^2 = \int_1^{10} (x - \mu)^2 f(x)dx = 4.493$$

$$\sigma = 2.120$$

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

Accept/reject method. Define min and max values for x and y. Generate random numbers x_r, y_r from a uniform distribution within these limits. Accept if $y < f(x)$, reject if $y > f(x)$.

• Produce 5000 random pairs (x, y) and calculate the correlation(s) between the (x, y) values.

I calculate the correlation of the x-values found using the accept/reject method, and the y-values found by plugging those x-values into $y = x \cdot exp(-x)$, using eq. (2.20) in Barlow.

$$\rho = \frac{\bar{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} = -0.774.$$

The correlation assumes a value between 1 and -1, where ±1 means a total positive/negative correlation respectively, and a correlation of zero, means no correlation. The function $y(x)$ looks like a line drawn from the top left corner, drawn down to the bottom right corner, so high y-values corresponds well to low x-values, and low y-values, corresponds well to high x-values. Thus we would also expect a large negative correlation, like -0.774.

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

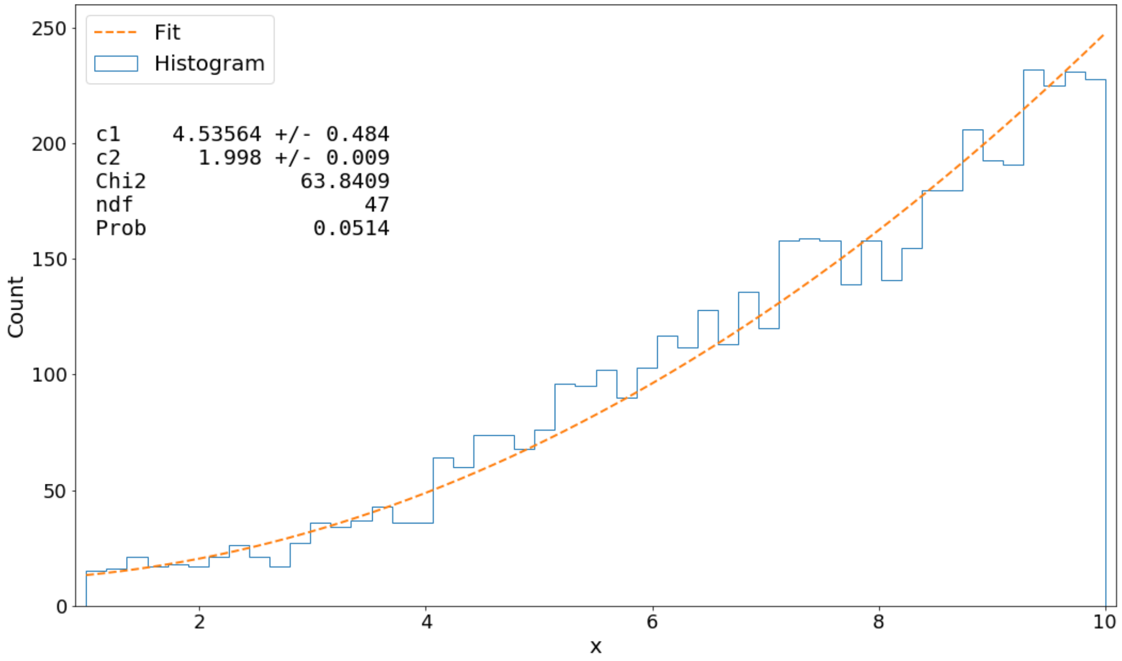

Figure 3: A histogram of all the accepted values, and a fit of f.

As you can see in fig 3, $c_{1,fit} = 4.5 \pm 0.5$ and $c_{2,fit} = 1.998 \pm 0.009$.

• 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?
The relative precision of c_1 is

$$\frac{c_1}{\sigma_{c_1}} = 0.106,$$

and the relative precision of c_2 is

$$\frac{c_2}{\sigma_{c_2}} = 0.004.$$

c_2 is already determined with a precision of 0.4% < 1%. For c_1 to fulfill this requirement, we need to improve the precision with a factor of $\frac{\sigma_{c_1}}{0.01c_1} = 10.675$. And as the error goes as $\sqrt{N}$, we need to up the N with a factor of $10.675^2 = 113.96$, such that $N = 113.96 \cdot 5000 = 569804.0$.

$$N = 5000 \left(\frac{\sigma}{0.01c_1}\right)^2 = 569804.0$$

After changing N to 569804, we find that the relative uncertainty σ_{c_1}/c_1 has fallen just below 1% to 0.009578 = 0.9578%. We calculated the value for N pretty precisely, so the reason σ_{c_1}/c_1 is a good bit below 0.01, instead of something like 0.00999 is that as N increases, the fit value for c_1 also changes with around 0.35. So the actual value is somewhere around $N = 550000$.
Doing the same calculation for c_2, to find the lowest N needed for a sub 1% precision, gave me the result $N = 1043$. For $N = 1043$, I got a precision of 1.02%. This result is a bit off for the same reason as I explained earlier, but because the value for c_2 decreased as N decreased, we landed a bit over this time.

Problem 4 - Statistical tests

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: www.nbi.dk/etersen/data_Cells.txt.

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

We say that the ones in the first column of the data represent the E cells, and the zeros represent the P-cells.
There are 2771 cells with $l < 9m$, and 2574 of them are P-cells, and 197 are E-cells.
Type I errors refer to false positives, and type II errors refer to false negatives. Thus the rate of type I errors is the ratio of E-cells classified as P-cells, to E-cells. The rate of type II errors is the ratio of P-cells classified as E-cells, to P-cells.

Table	E-cells	P-cells
Accepted	9.3%	100%
Rejected	90.7%	0%

Table 1: A table showing how many of the cells were accepted or rejected, based on their type.

As we see from table 1, the rate of type I errors (i.e. false positives), is 0.093, and the rate of type II errors (i.e. false negatives) is 0.

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

The separation of two variables can be calculated as such:

$$\Delta = \frac{|\mu_A - \mu_B|}{\sqrt{\sigma_A^2 + \sigma_B^2}}$$

From this equation, I found $\Delta_l = 1.49$ and $\Delta_T = 2.22$.
Since T has a higher separation, it is also better at distinguishing between P and E cells.

• Separate P and E cells using l and/or T, and draw a ROC curve of your result.
I found the Fisher discriminant, plotted the histogram shown in fig. 4, and calculated the separation to be 2.61, which is higher than for both the length and the transparency.

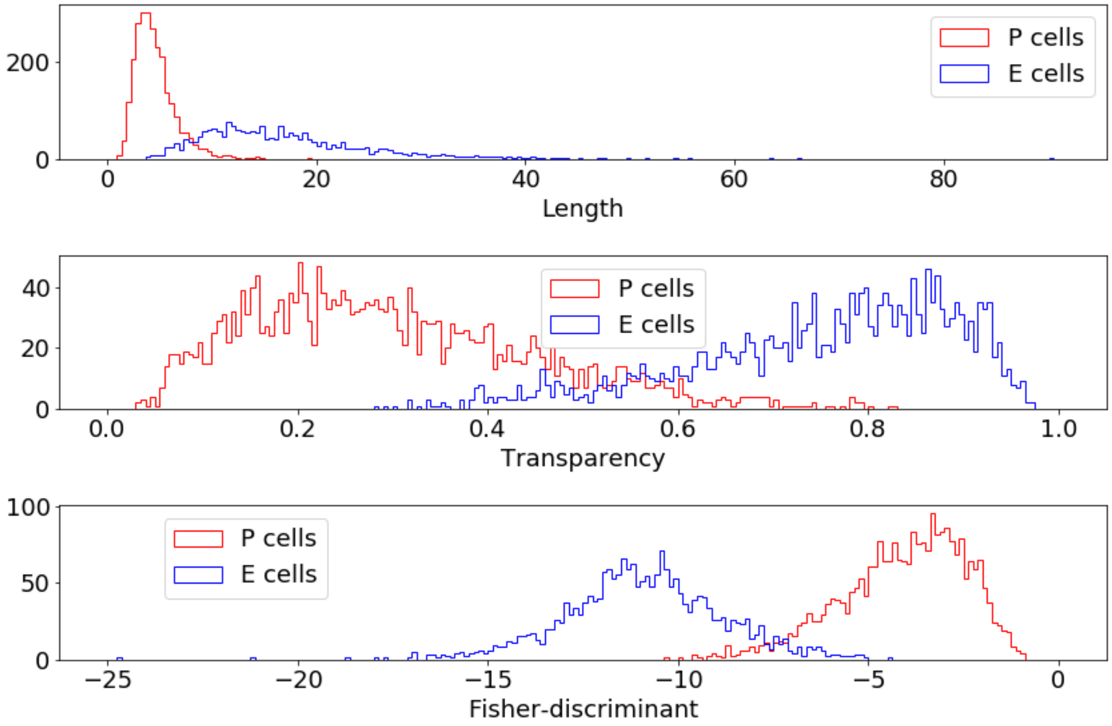

Figure 4: Three histograms of the P- and E-cells, of length, transparency, and the Fisher discriminant respectively.

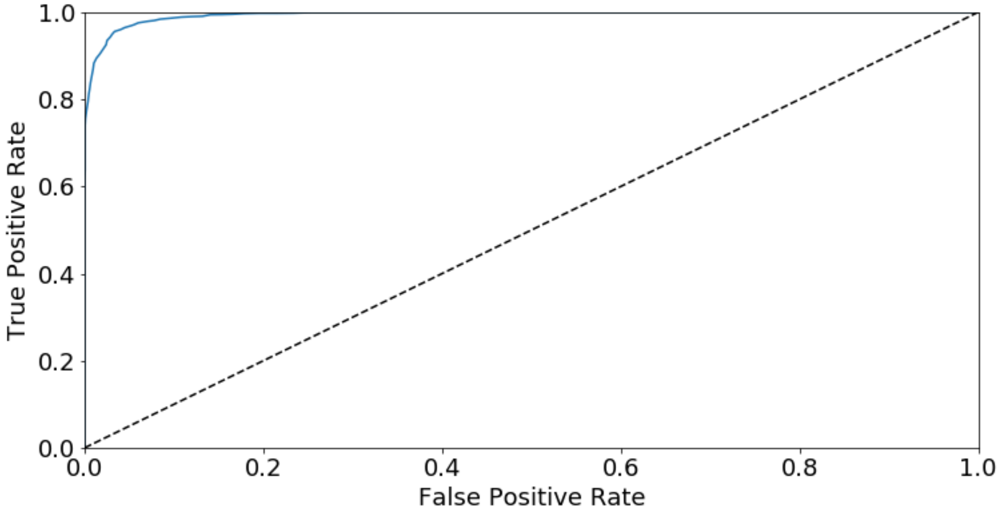

Figure 5: ROC curve of the Fisher histograms.

Problem 5 - Fitting data

5.1

Kepler's third law states that "the square of the orbital period (T) of a planet is directly proportional to the cube of the semi-major axis (a) of its orbit". The table lists values for T in days (known very precisely) and a in AU (= 149597870700m) at the time of the first measurement (in 1778) of the gravitational constant $G_{1778} = (7.5 \pm 1.0)10^{-11} m^3 kg^{-1} s^{-2}$.

Planet	T (days)	a (AU)
Mercury	87.77	0.389 ± 0.011
Venus	224.70	0.724 ± 0.020
Earth	365.25	1 (definition)
Mars	686.95	1.524 ± 0.037
Jupiter	4332.62	5.20 ± 0.13
Saturn	10759.2	9.51 ± 0.34

Table 2: The table given in the problem text.

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

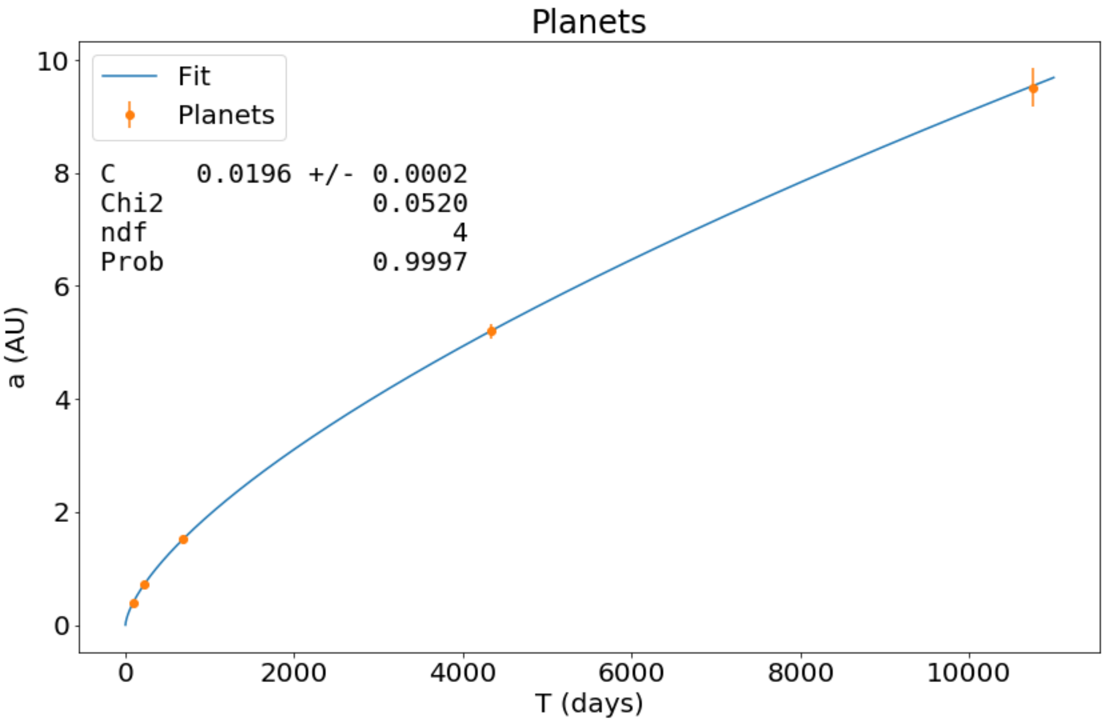

Figure 6: The semi-major axes of five non-Earth planets as a function of the orbital period, fitted to Kepler's third law, $a = CT^{2/3}$.

• In this fit, which planet seems to follow this relation least well? Is it critical?
In this fit, Mercury seems to follow the relation least well, being 0.1895 sigma away from the fit value. This was calculated by simply taking the a for a planet, and subtracting the a predicted by the fit, and dividing by the associated sigma. i.e.

$$\Delta = \frac{a - f(T, c)}{\sigma}$$

$$\Delta_{Mercury} = \frac{0.389 - 9.55}{0.34} = 0.1895$$

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

Using the equation

$$M = 4\pi^2 C^3/G,$$

where $C = 0.0196 \pm 0.0002$, and $G = G_{1778} = (7.5 \pm 1.0) \times 10^{-11} m^3 kg^{-1} s^{-2}$, we find that

$$M = (1.8 + 0.2) \times 10^{30} kg$$

• Expand the fit to Kepler's third law by further adding two parameters: $a = C \times (Tc_1 + c_2)$. Does this formula match the data well? Are the two additional parameters necessary?

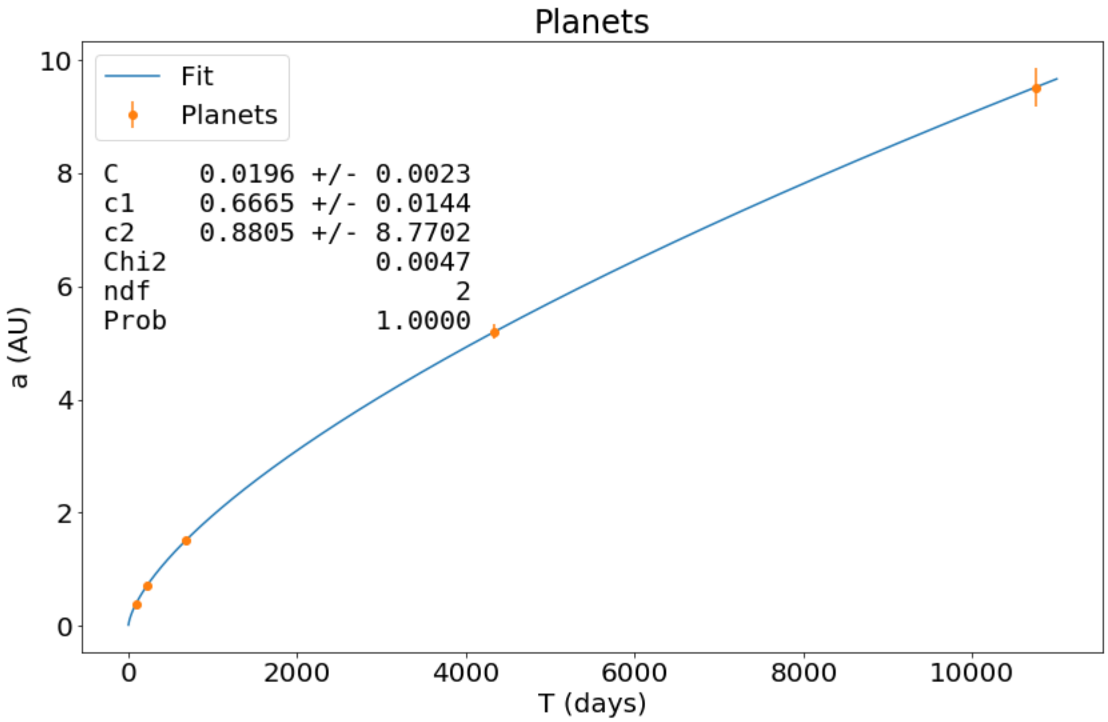

Figure 7: The semi-major axes of five non-Earth planets as a function of the orbital period, fitted to a modification of Kepler's third law, $a = C(T^{c_1} + c_2)$.

From the figure, you can see that C keeps that same value, 0.0196, but with a larger uncertainty, and the exponent ends up having 2/3 within the uncertainty anyway. The new term c_2 is also close to zero, with an uncertainty ten times larger than its value. The χ^2 probability is increased a little bit, but was very high to begin with. So the addition of the two new parameters doesn't tell us very much, and can be deemed unnecessary.

5.2

Searching for slow moving (compared to speed of light) particles at CERN's LHC accelerator, you are calibrating the speed measurement $\beta = v/c$ of the candidate particles, using a control sample of particles known to (effectively) travel at the speed of light, i.e. $\beta = 1$. The file www.nbi.dk/ petersen/data BetaCalibration.txt contains 4000 control sample measurements of initial speed estimate (β_{init}), energy (E) in GeV, angle with respect to the beam axis (θ) in radians, and time since start of experiment (T) in seconds, respectively.

- **What is the resolution of β_{init}? And is it consistent with a Gaussian distribution?**

The resolution is just the standard deviation,

$$\sigma = \sum_{i}^{N} (x_i - \mu)^2,$$

which in this case is 0.091.

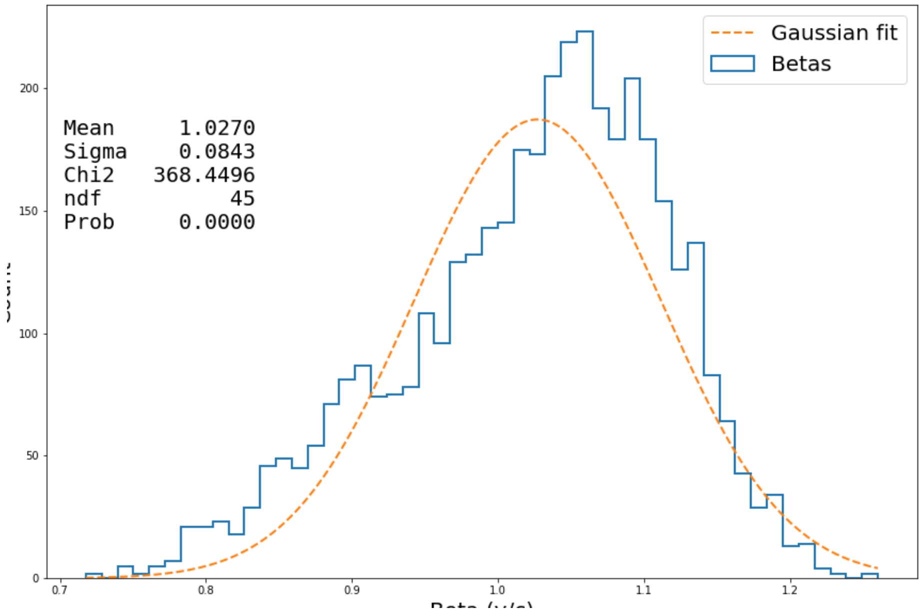

Figure 8: A Gaussian fit of the initial β values.

I made a Gaussian fit, and got a χ^2 probability of $2 \cdot 10^{-52}$, so no, it is not consistent with a Gaussian distribution.

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

Firstly, we would want the mean to be at least close to $\pi/2$, and it was calculated to be $\pi/2 - 0.00478$, a deviation of 0.27 degrees. We expect a symmetric distribution to have zero skew, as skew is a measure of asymmetry. The skew was calculated to be -0.016, meaning there was little asymmetry. So far we cannot say that the distribution is not consistent with being symmetric around $\pi/2$, as the arguments against are lacklustre.

I made a histogram of the data, spanning 20 bins. Then, treating the first 10 bins as one histogram, and the last 10 bins as another, the χ^2 for the two histograms was calculated using the following equation,

$$\chi^2 = \sum_{i \in bins}^{N} \frac{(O1_i - O2_i)^2}{(O1_i + O2_i)}$$

Here the nominator is the difference between the two subsets, and the denominator is the squared error of the two subsets. We assume the errors are Poissonian, and are thus the square root of the number of data points. The result was $\chi^2 = 0.091$, with a probability of 0.652. In conclusion, yes, the distribution in θ is consistent with being symmetric around pi.

- **Test if the mean of β_{init} is constant as a function of energy.**

I made a 2d histogram of the relative betas $(\beta - 1)$ and the energies, plotted the means, and then fitted a line to the means. The result, seen in fig x, is that there is some correlation. If the parameters of the fit had been zero, the mean of β_{init} would have been constant as a function of the energy.

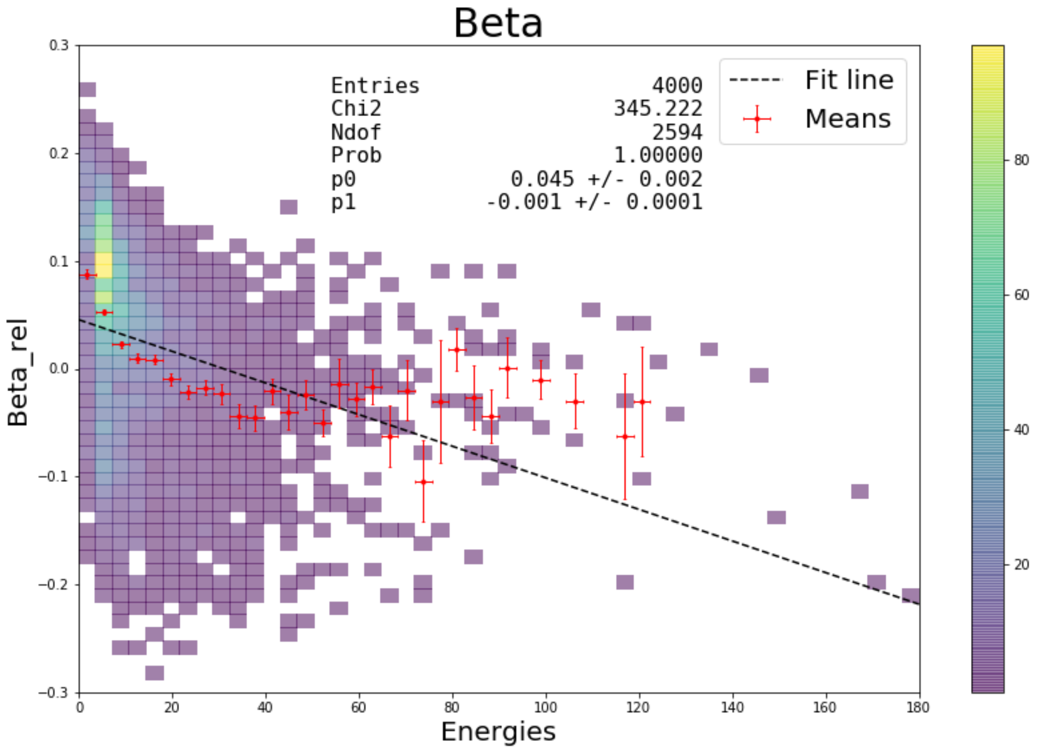

Figure 9: Energy calibration

• Due to shifts in timing, the central value of init shifted with time T, smearing the resolution. Calibrate β_{init} with respect to T and determine the obtained resolution on $\beta_{T-calib}$.

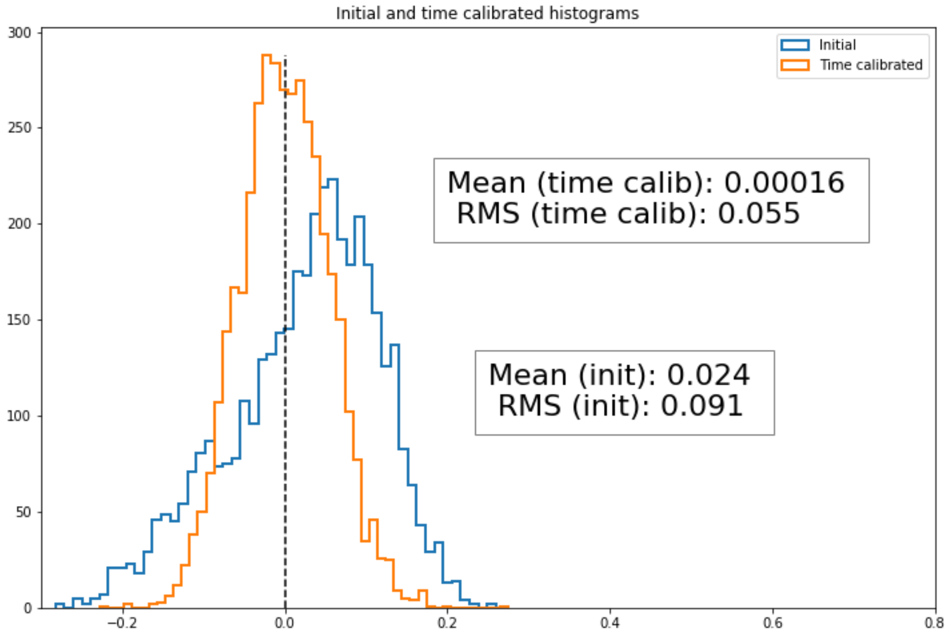

Figure 10: β_{init} calibrated with respect to time. Note that the uncalibrated histogram has had all its elements subtracted by one, in order to show it on top of the calibrated histogram, so that the difference is clearer. Thus the actual mean of the uncalibrated histogram is 1.024.

- **Using all information available, what is the best calibration of β you can produce?**