

Problem set in Applied Statistics 2020

Kian Gao SHP593

The philosophy of some problems is discussed with my group members Alicia, Timo and Jonas, meanwhile the code is completely written on **my own**.

I – Distributions and probabilities:

1.1 (4 points) 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?

To make it even, the number of both team's victories should be the same. Thus, we can use binomial distribution to actually calculate the probability.

$$P_e = C_{144}^n p^n (1-p)^{n-n}$$

The probability of even in Barcelona verses RM is: 0.06637504645119732
If using the scipy library, the answer is: 0.06637504645119337

1.2 (4 points) 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?

To hit the window with $p_{hit} = 90\%$, we need to make sure that the former shoots don't hit, means that $p_{nohit} = 10\%$.

The number of golf balls they use to make sure the probability over 90% is: 40

II – Error propagation: 2.1 (10 points) 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.

- What is the weighted average of h ? Do the values 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?

Answer:

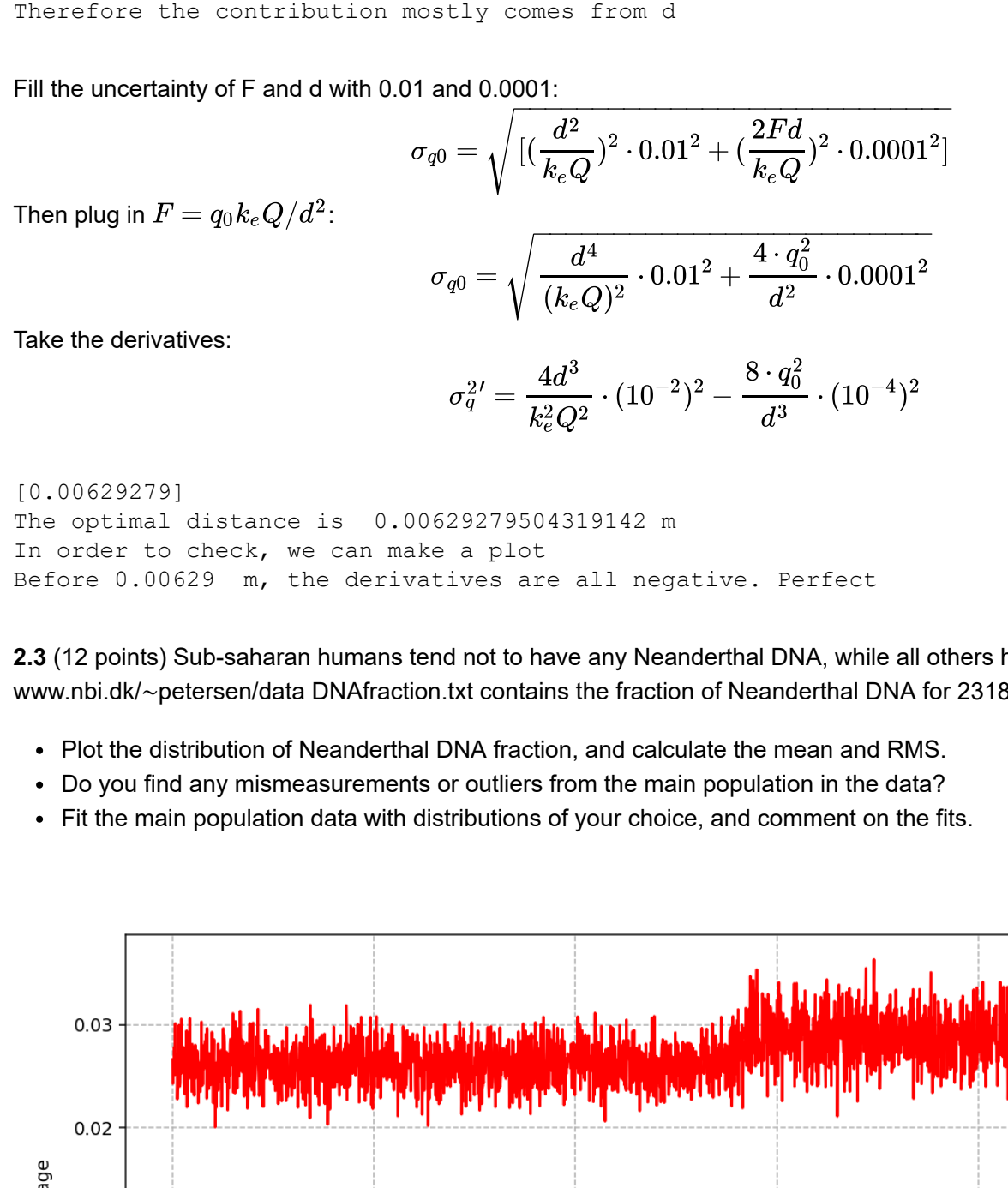
1. The weighted mean is defined as

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

68.78925107187163

And the values don't agree with each other. The weighted mean is roughly 68.79, but we can easily see that the second and the fourth measurements need more than 3 σ distance to touch the weighted average.

Meanwhile, we can do a chi2 test for the constant.



From the Chi2 regression test and numerical estimation, we can safely say that the values don't agree with each other

The Chi2 probability of the first four values are: 0.9852357656317927

The Chi2 probability of the last three values are: 0.2584838114720961

Therefore, through the Chi2 test we could safely say that the first four values and the last three values agree each other within their own methods.

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

- Given force $F = 0.87 \pm 0.08$ N and distance $d = 0.0045 \pm 0.0003$ m, what is q_0 ?
- Where does the largest contribution to the uncertainty on q_0 come from? F or d ?
- If you could measure F and d with uncertainties ± 0.01 N and ± 0.0001 m, respectively, at what distance should you expect to measure the charge in question q_0 most precisely?

q_0 is $1.96e-06 \pm 3.17e-07$ C

The uncertainty contribution on F is $1.802002224694104e-07$

The uncertainty contribution on d is $2.6129032258064514e-07$

Therefore the contribution mostly comes from d

Fill the uncertainty of F and d with 0.01 and 0.0001:

$$\sigma_{q_0} = \sqrt{\left(\frac{d^2}{k_e Q} \cdot 0.01^2 + \left(\frac{2Fd}{k_e Q}\right)^2 \cdot 0.0001^2\right)}$$

Then plug in $F = q_0 k_e Q / d^2$:

$$\sigma_{q_0} = \sqrt{\frac{d^4}{(k_e Q)^2} \cdot 0.01^2 + \frac{4 \cdot q_0^2}{d^2} \cdot 0.0001^2}$$

Take the derivatives:

$$\sigma_{q_0}' = \frac{4d^3}{k_e^2 Q^2} \cdot (10^{-2})^2 - \frac{8 \cdot q_0^2}{d^3} \cdot (10^{-4})^2$$

[0.00629279]

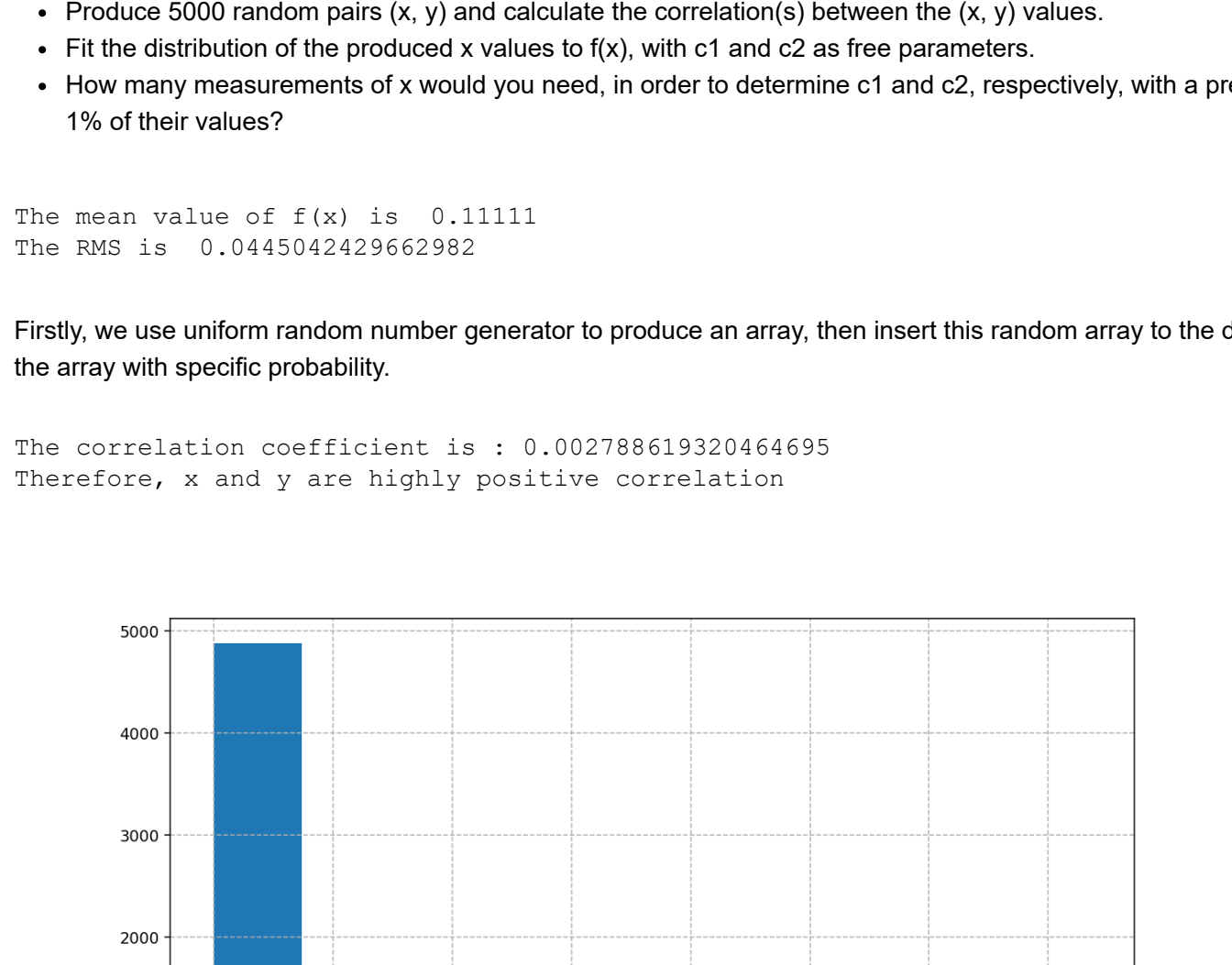
The optimal distance is 0.00629279504319142 m

In order to check, we can make a plot

Before 0.00629 m, the derivatives are all negative. Perfect

2.3 (12 points) 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.
- Do you find any mismeasurements or outliers from the main population in the data?
- Fit the main population data with distributions of your choice, and comment on the fits.



The outlier numbers are: [2291 2292 2293 2294 2295 2296 2297 2298 2299 2300 2301 2302 2303 2304]

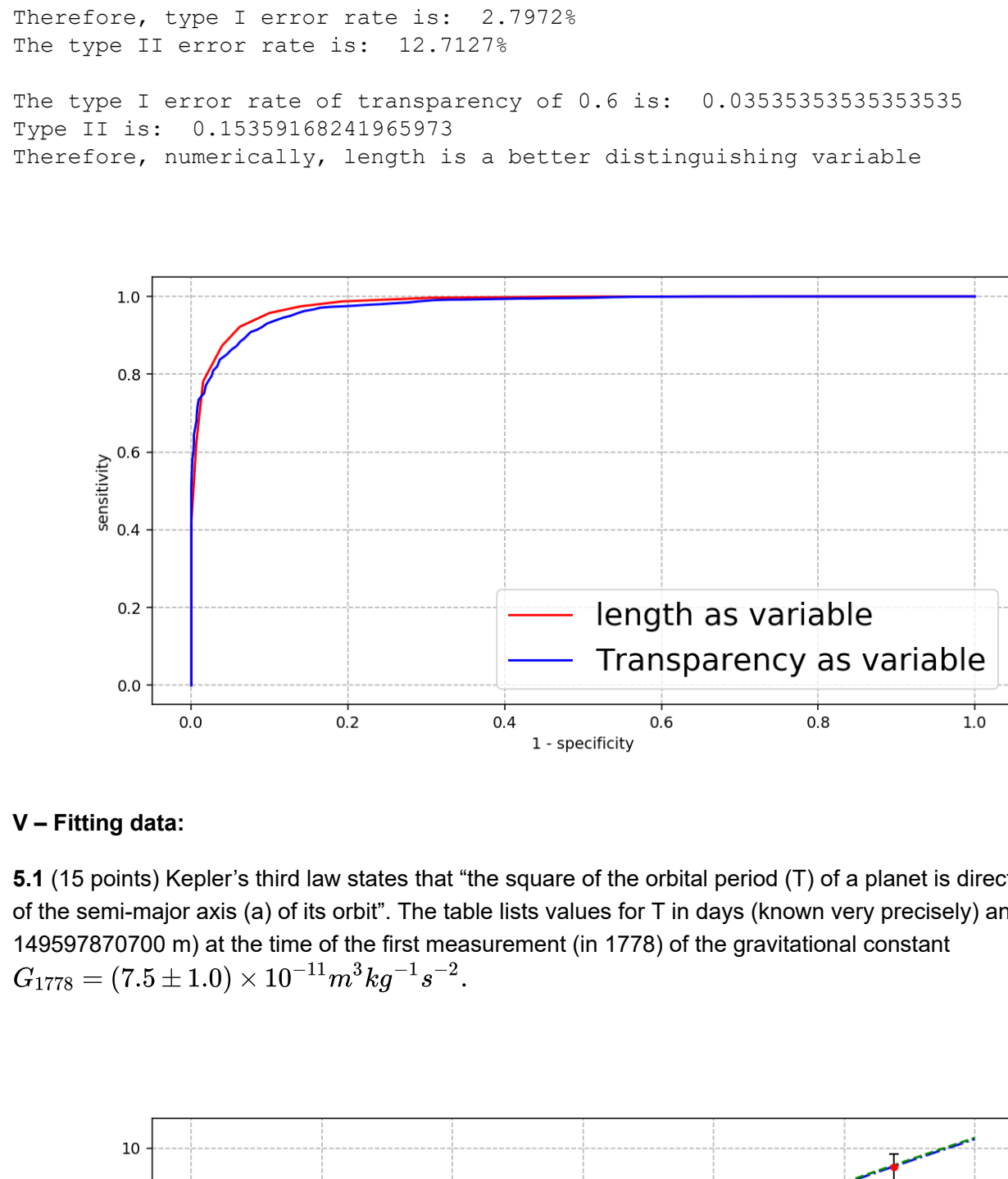
2305 2306 2307 2309 2310 2311 2312 2313 2314 2315 2316 2317]

Where they get the percentage deviate away from 3 sigma(std)

Meanwhile, the mismeasurements (also includes in outliers) are: [2303 2304 2307]

Where the percentage is smaller than 0, physically not acceptable

First we move the outlier out, where they exist in 3 sigma out from average



The distribution chose is Gaussian

We create the histogram of dna_data, where array[0] is bin data's number, [1] is bins' position

The original chi2 is: 3185.835458738805

The dof is: 61

92.7224680401674 0.026955491659177957

The gaussian fit with Chi2 probability is ~56%, pretty neat.

III – Monte Carlo: 3.1 (15 points) 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 $x \in [1, 10]$, where $c_1 = 5$ and $c_2 = 2$ and then using this x value to calculate $y = x \exp(-x)$.

- What is the value of C ? And what is the mean and RMS of $f(x)$?
- What method(s) can be used to produce random numbers according to $f(x)$? Why?
- Produce 5000 random pairs (x, y) and calculate the correlation(s) between the (x, y) values.
- Fit the distribution of the produced x values to $f(x)$, with c_1 and c_2 as free parameters.
- 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?

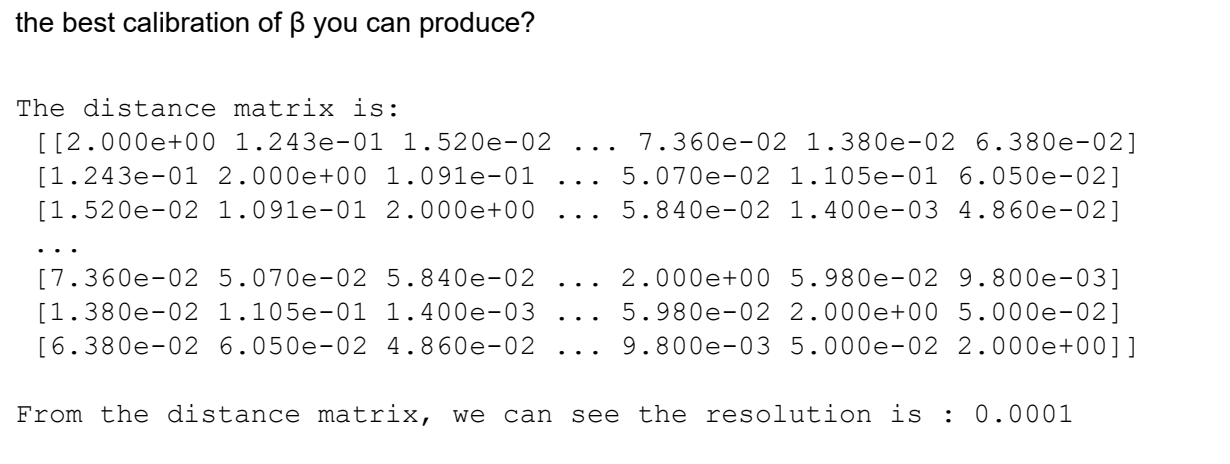
The mean value of $f(x)$ is 0.11111

The RMS is 0.044504243662982

Firstly, we use uniform random number generator to produce an array, then insert this random array to the distribution to get the array with specific probability.

The correlation coefficient is : 0.002788619320464695

Therefore, x and y are highly positive correlation



c1 fit is: 5.001322181076822 c2 fit is: 2.0000276739729297

P_chi2 is 1.0

IV – Statistical tests: 4.1 (15 points) 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/~petersen/data/Cells.txt](#).

- Selecting P -cells by requiring $l < 9 \mu m$ what is the rate of type I and type II errors?
- Which of the two variables l and T is best 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.

The total num of P cells smaller than $9 \mu m$ is : 2502

The E cells num is : 269

The total num of E cells greater than $9 \mu m$ is : 1847

The P cells num is : 72

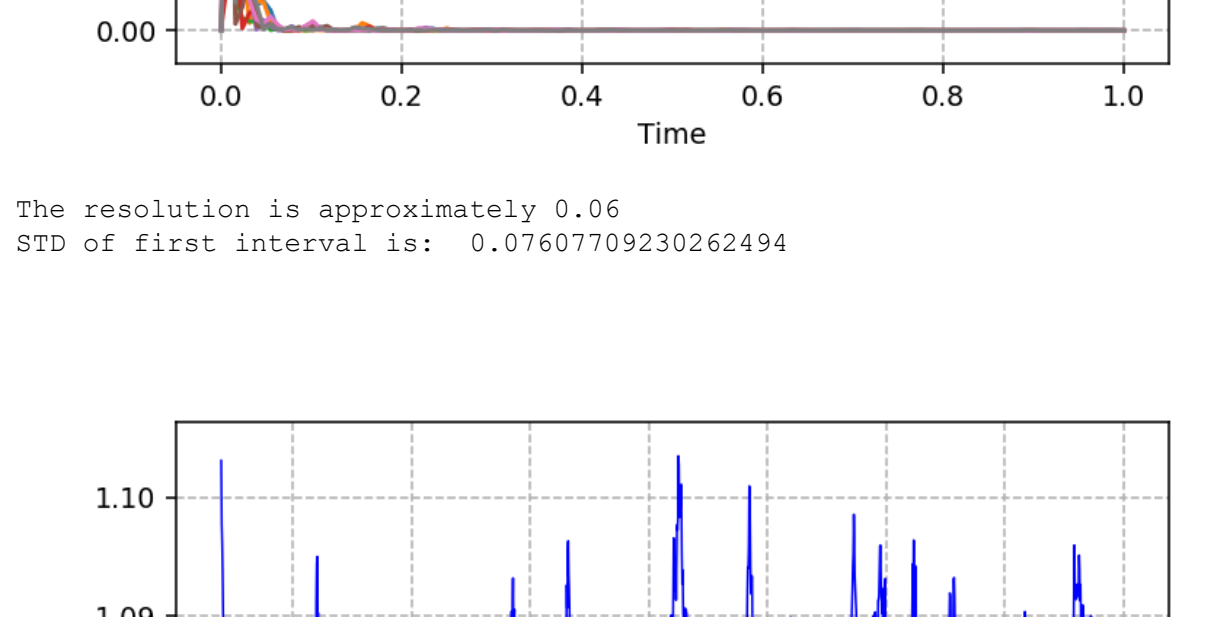
Therefore, type I error rate is: 2.7972%

The type II error rate is: 12.7127%

The type I error rate of transparency of 0.6 is: 0.03535353535353535

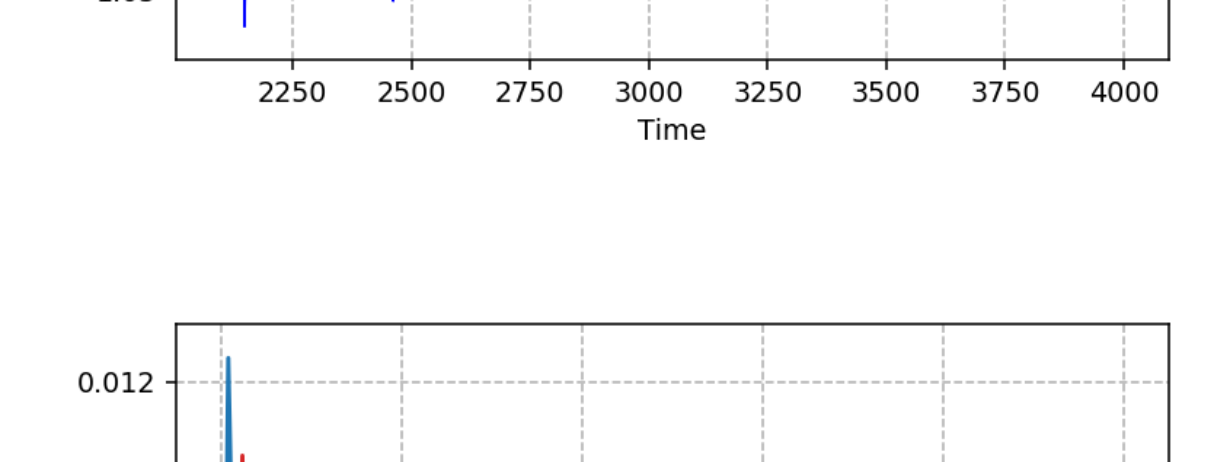
Type II is: 0.15359168241965973

Therefore, numerically, length is a better distinguishing variable



V – Fitting data:

5.1 (15 points) 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 in AU (= 149597870700 m) at the time of the first measurement (in 1778) of the gravitational constant $G_{1778} = (7.5 \pm 1.0) \times 10^{-11} m^3 kg^{-1} s^{-2}$.

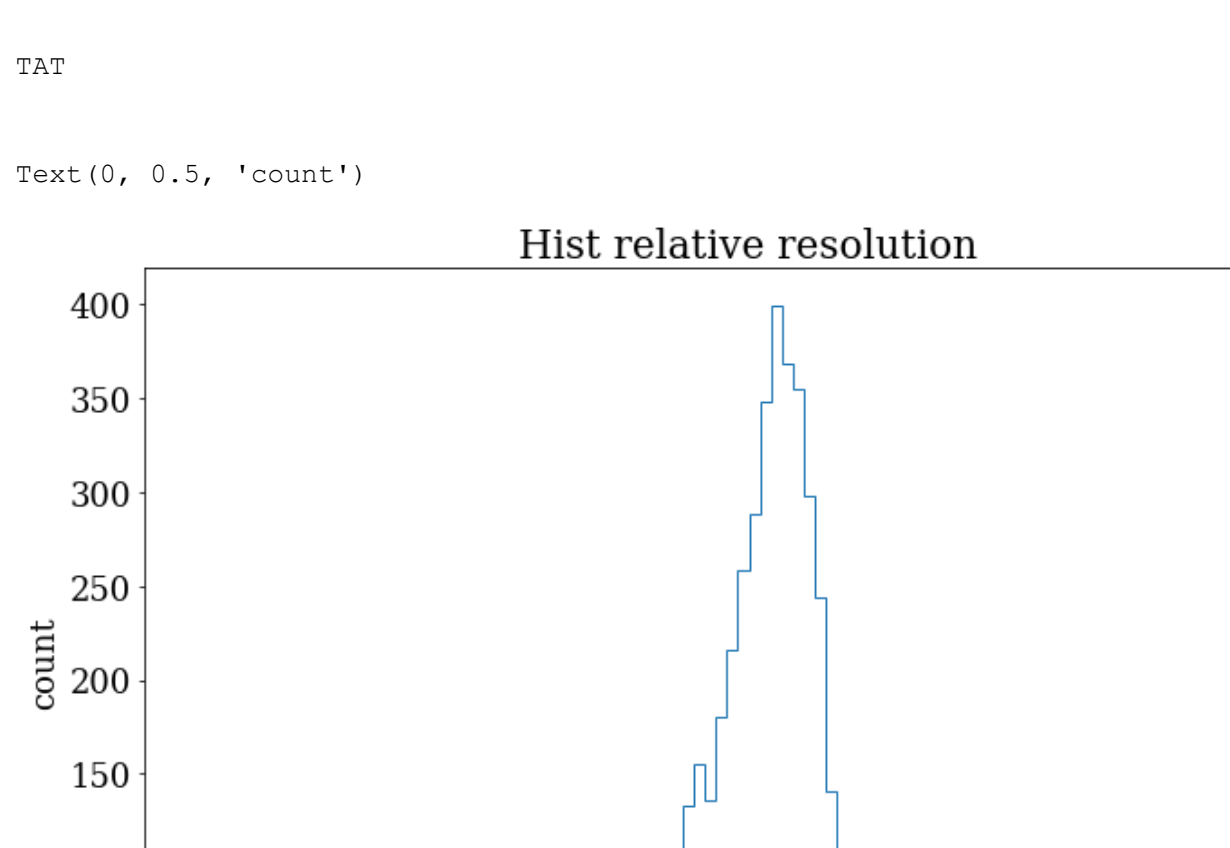


<ErrorbarContainer object of 3 artists>

The deviation in the unit of our own uncertainty are individually:
[166441.17488035 171314.95758697 195061.54302034 189516.51018511 132881.54039788]

Therefore, the first planet, Mercury, deviates the most but not critical

solar mass is: 1.75766e+30 +/- 2.344e+29 kg



The deviation in the unit of our own uncertainty are individually:
[0.01678535 0.04205357 0.01402375 0.04075394 0.04502363]

This formula match slightly better except for the last planet. However, from my perspective, the addition of this two parameters is not necessary since the chi2 probability doesn't really change much (say the simple formula's chi2 probability is high enough). Therefore, no need to add these two parameters before we get more datapoint.

5.2 (15 points) 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_{ini}), 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_{ini} ? And is it consistent with a Gaussian distribution? • Is the distribution in θ consistent with being symmetric around $\pi/2$? • Test if the mean of β_{ini} is constant as a function of energy. • Due to shifts in timing, the central value of β_{ini} shifted with time T , smearing the resolution. Calibrate β_{ini} with respect to T and determine the obtained resolution on $\beta_{T-calib}$. • Using all information available, what is the best calibration of β you can produce?

The distance matrix is:

[[0.00e+00 1.243e-01 1.520e-02 ... 7.360e-02 1.380e-02 6.380e-02]
[1.243e-01 2.000e+00 1.091e-01 ... 5.070e-02 1.105e-01 6.050e-02]
[1.520e-02 1.091e-01 2.000e+00 ... 5.840e-02 1.400e-03 4.860e-02]
...
[7.360e-02 5.070e-02 5.840e-02 ... 2.000e+00 5.980e-02 9.800e-03]
[1.380e-02 1.105e-01 1.400e-03 ... 5.980e-02 2.000e+00 5.000e-02]
[6.380e-02 6.050e-02 4.860e-02 ... 9.800e-03 5.000e-02 2.000e+00]]

From the distance matrix, we can see the resolution is : 0.0001

The original chi2 is: 12156.402657186032

The dof is: 97

P_chi2 is: 0.0

98.0022831923988 1.043759626573496

Not consistent with Gaussian distribution

The skewness of theta is : -0.016267892096180927

The average value is: 1.5795733749999998

Where diff between pi/2 and avg is: 0.004777048205103274

Therefore, we can safely say that the angle is symmetric along pi/2 axis

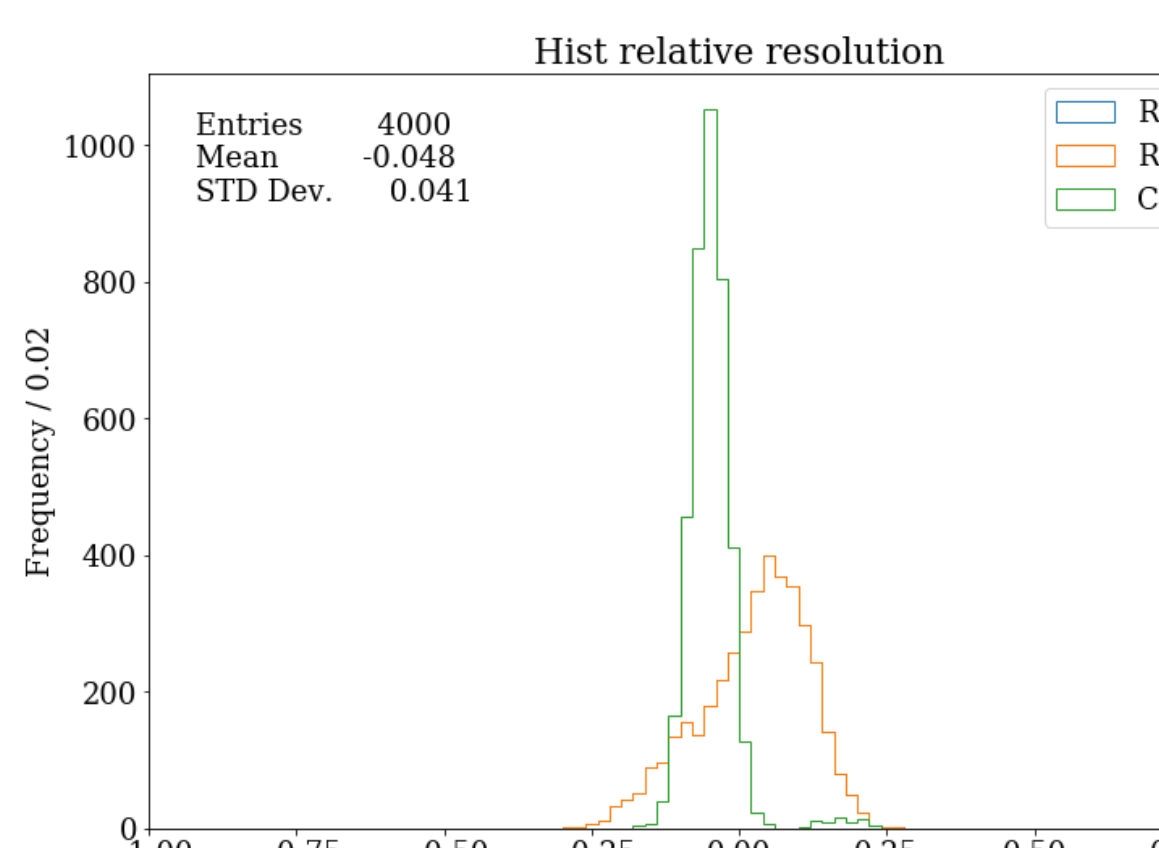
Use covariance matrix to do the test

[[2.47393707e-02 -4.33266321e-01]
[-4.33266321e-01 6.30866790e-03]]

The $C_{[01]}$ of cov matrix shows that the correlation between beta ini and energy is negative. Therefore, we can say that the two variables are to some extend, negative correlation

Hence, beta is not constant as function of energy

Use convolution to test the variance of beta under the evolution of T
Since there is a huge gap at around 2000s, we treat the two intervals separately



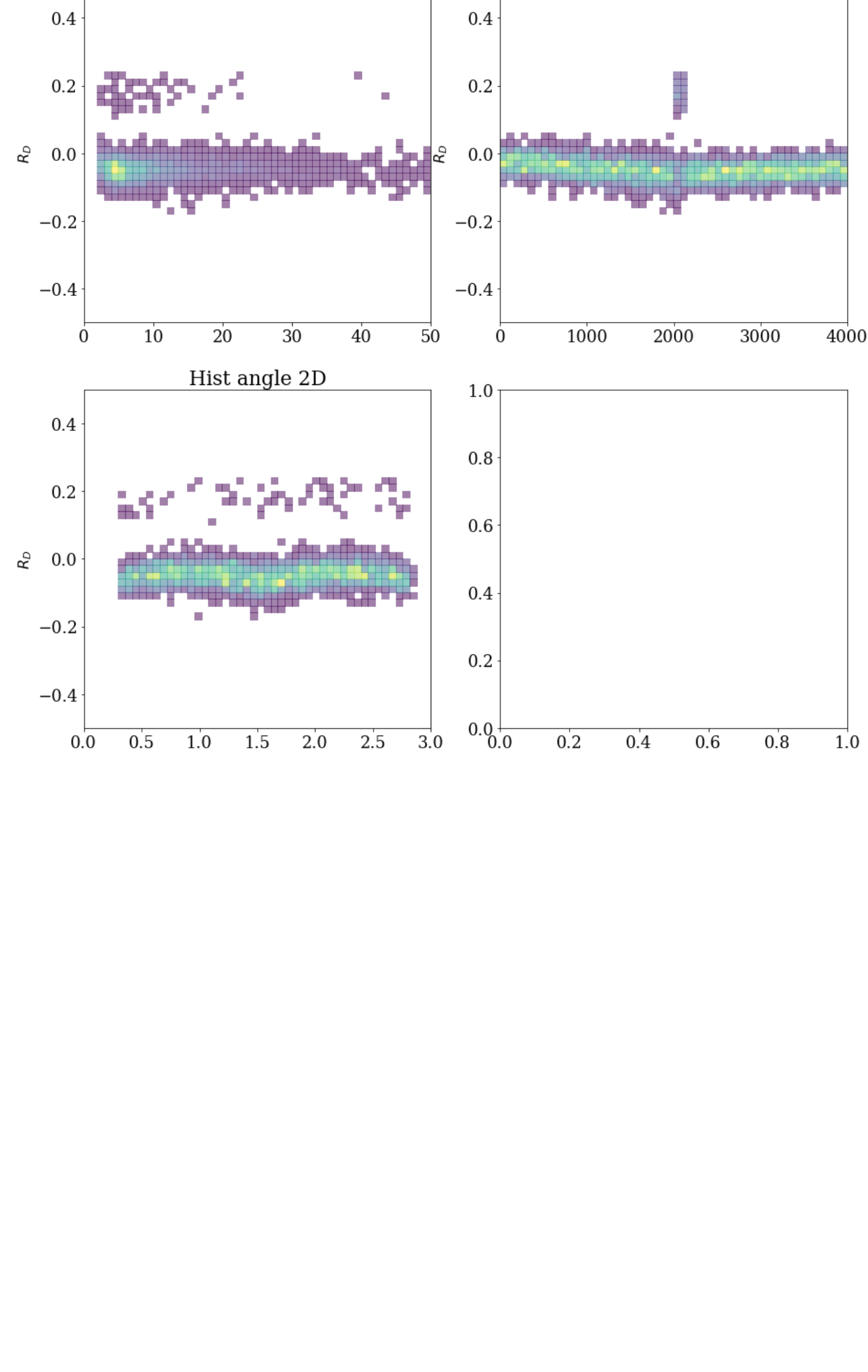
The resolution is approximately 0.06

STD of first interval is: 0.07607709230262494



[-0.03268303 -0.04664392 -0.00309181 ... -0.01656032 -0.0408306 -0.06008171]

Text(0, 0.5, 'SR_(\$')')



[NbConvertApp] Converting notebook ProblemSet2020_KianGao.ipynb to PDF via HTML
[NbConvertApp] Writing 3525975 bytes to ProblemSet2020_KianGao.pdf