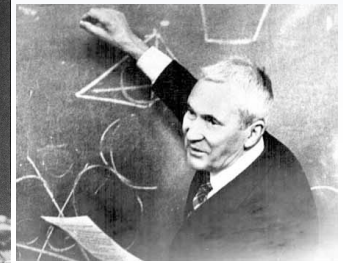
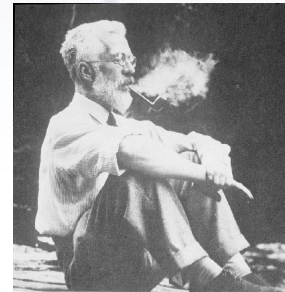
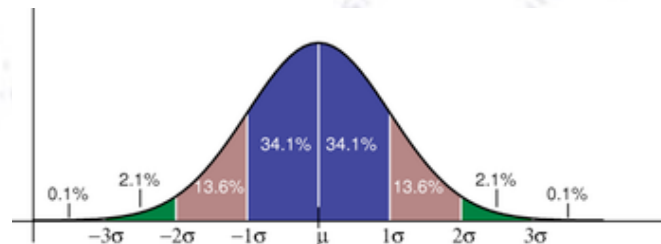


Applied Statistics

Problem Set Solution and Discussion



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

A faded nautical chart background. It features magnetic isogonic lines (lines of equal magnetic variation) labeled with values like 0, 30, 60, 90, 120, 150, 180, 210, 240, and 270. A specific magnetic variation is noted as "VAR 10° 15' W" with a small cross symbol. The word "MAGNETIC" is also visible. In the upper right, there is a label "ICE BITTER END TACHT/ALUB".

Overall comments

Make solutions CLEAR

The solution is not a style contest, but making your results CLEAR to the reader is important, and please do not put in any code! For the exam, we get the code anyway.

1.5 IV - Statistical Tests

```
[51]: # Calculate ROC curve from two histograms (hist1 is signal, hist2 is
      ↪background):
def calc_ROC(hist1, hist2) :

    # First we extract the entries (y values) and the edges of the histograms
    y_sig, x_sig_edges, _ = hist1
    y_bkg, x_bkg_edges, _ = hist2

    # Check that the two histograms have the same x edges:
    if np.array_equal(x_sig_edges, x_bkg_edges) :

        # Extract the center positions (x values) of the bins (both signal or
        ↪background works - equal binning)
        x_centers = 0.5*(x_sig_edges[1:] + x_sig_edges[:-1])

        # Calculate the integral (sum) of the signal and background:
        integral_sig = y_sig.sum()
        integral_bkg = y_bkg.sum()

        # Initialize empty arrays for the True Positive Rate (TPR) and the
        ↪False Positive Rate (FPR):
        TPR = np.zeros_like(y_sig) # True positive rate (sensitivity)
        FPR = np.zeros_like(y_sig) # False positive rate ()

        # Loop over all bins (x_centers) of the histograms and calculate TN,
        ↪FP, FN, TP, FPR, and TPR for each bin:
        for i, x in enumerate(x_centers):

            # The cut mask
            cut = (x_centers < x)

            # True positive
            TP = np.sum(y_sig[-cut]) / integral_sig # True positives
            FN = np.sum(y_sig[cut]) / integral_sig # False negatives
            TPR[i] = TP / (TP + FN) # True positive rate

            # True negative
            TN = np.sum(y_bkg[cut]) / integral_bkg # True negatives
            ↪(background)
            FP = np.sum(y_bkg[-cut]) / integral_bkg # False positives
            FPR[i] = FP / (FP + TN) # False positive rate
            ↪

        return FPR, TPR
```

2 Error propagation

2.1 The Hubble constant

2.1.1

The weighted mean of h is 68.8 ± 0.3 (km/s)/Mpc. The χ^2 value of this is $\chi^2 = 52.54$ with $p = 1.454 \cdot 10^{-9}$, so the values do not agree with each other.

2.1.2

The first method has: $h = 73.9 \pm 0.8$ with $\chi^2 = 0.4978$ and $p = 0.9194$. The second method has: $h = 67.8 \pm 0.3$ with $\chi^2 = 3.640$ and $p = 0.1620$. Since both p -values are in the trusted range, the values from the same method agree with each other in both cases.

2.2 Coulomb's law

2.2.1

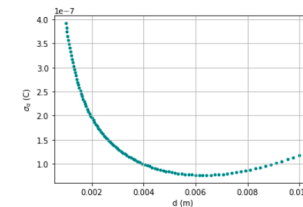


Figure 3: The uncertainty on q_0 with respect to the value of d .

The charge q_0 is

$$q_0 = \frac{Fd^2}{k_e Q} = 1.959 \cdot 10^{-6} C \quad (4)$$

with an error of

$$\sigma_{q_0} = \sqrt{\left(\frac{d^2}{k_e Q}\right)^2 \sigma_F^2 + \left(\frac{2dF}{k_e Q}\right)^2 \sigma_d^2} = 3.174 \cdot 10^{-7} C \quad (5)$$

so the resulting q_0 is $q_0 = 2.0 \pm 0.3 \cdot 10^{-6} C$

2.2.2

The contribution from F is:

$$\sigma_{q_0, F} = \sqrt{\left(\frac{d^2}{k_e Q}\right)^2 \sigma_F^2} = 1.8 \cdot 10^{-7} C \quad (6)$$

The contribution from d is:

$$\sigma_{q_0, d} = \sqrt{\left(\frac{2dF}{k_e Q}\right)^2 \sigma_d^2} = 2.6 \cdot 10^{-7} C \quad (7)$$

So the largest contribution comes from d .

Give rationale, not story

You should for every solution give your **rationale/explanation**, and make sure that you **quantify your answers**.

Do NOT derive or write basic equations, e.g. it is enough to state that “I use a weighted average”.

Do NOT give a (long) story. To some extent, much of science is often to give the short and concise answer.

Do NOT include any code, and if you do, it better be short and very well reasoned.

Do NOT repeat the problem, but rather start on your answer (“3.1.1 Using x...”).

where $N = 4000$ is the number of angle measurements and the standard deviation is found to be $\sigma_\theta = 0.69$. Comparing this to the expected value of $\pi/2$ using a z-test gives $z_{\theta} = (\theta_{\text{mean}} - \pi/2)/(\sigma_{\text{mean}}) \cong 0.4$. Considering the uncertainty on the mean this is 0.4σ away from the expected value and thus in agreement with being symmetric around $\pi/2$ considering only the mean. Furthermore I looked at the number of points above and below $\pi/2$ finding that $N_{\text{above}}/N_{\text{below}} = 2011/1989 \cong 1.011$, which shows that the amount of measurements above $\pi/2$ is only 1.1% larger than the amount of measurements below, confirming quantitative symmetry around $\pi/2$. Combining these insights it is reasonable to conclude that the data is symmetric around $\pi/2$.

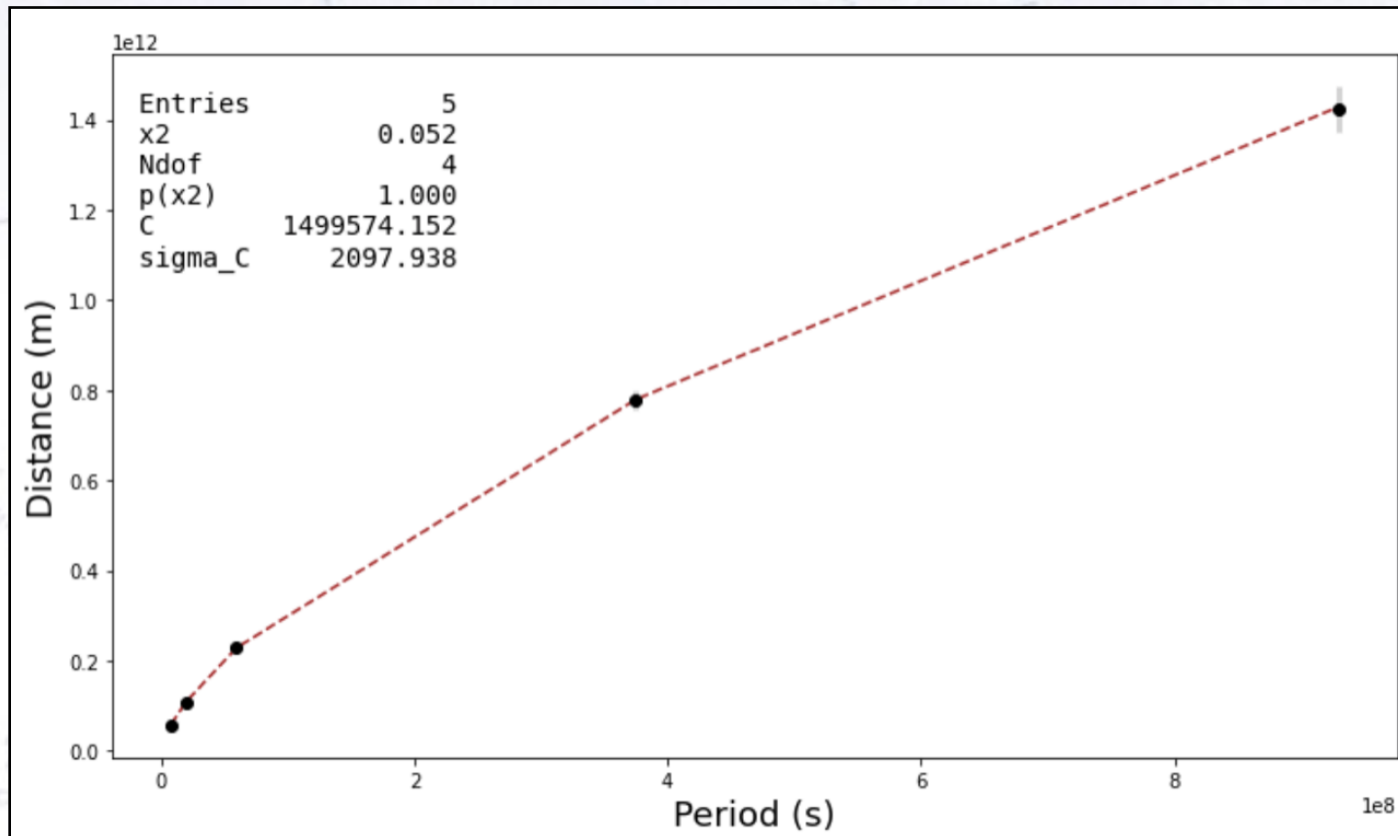
To test if β_{init} is constant as a function of the energy I have binned the energies in equidistant intervals ranging from E_{min} to E_{max} using a number of $n_{\text{bins}} = 50$ resulting in a binwidth of $\Delta E = 3.545 \text{ GeV}$. Then I took the corresponding values of β_{init} and found the mean of these within the area of energies. The result is shown in figure 15 where a number of decisions must be addressed. I did the binning for a range of number of bins and decided on this one, as it demonstrates the challenge of a low number of data points for high values of E . As is seen in the illustration the error on the values of β increases with the energy, because of this. Some points only contain one value therefore resulting in an uncomputable error as sketched in the plot. To avoid this the binning could be changed to a lower number, however this would come at a cost of the resolution for the lower values of E , and even bin numbers as low as 20 would still contain data points with only one measurement of β . And exactly the resolution of the lower values of E is important, as the plot indicates that there is a skew going from high values of β for small E to lower values of β for larger E . This however can only be concluded for the relatively low energies, i.e. in the order below 50 GeV, as the uncertainties become too large for larger values of E .

Given the information that there is a smearing due to a shift in timing I inspect the plot of β as a function of T as shown in figure 16. To me this looks like the time is shifted by a negative, linear relation until a point around $T = 2000$. I found one of the last points before the time got readjusted, by determining the last point in time between $T = 2000$ and $T = 2200$ where $\beta < 0.9$. This value does not appear after the readjusted time in this interval, however it appears frequently for the non-adjusted time. Using this value, $T_{\text{shift, end}} = 2047$ I then fitted the values of β from $T = 0$ to $T_{\text{shift, end}}$ to a straight line using a χ^2 fit, which is also seen in figure 13. The probability of fit being equal to 1 is because of the lack of errors on the measurements. In order to correct for this systematic shift of the values of β , I subtracted this linear relation from the given values until the time of $T = T_{\text{shift, end}} = 2047$ and adding the constant from the linear relation. A

Continuous models

The planet case is NOT “small statistics”. This only goes for counting statistics, e.g. in histograms, when bins with small statistics do not have Gaussian errors.

Careful when drawing your models / functions... they should be continuous.



But admittedly, I had a hard time finding bad figures... you should be proud!

Various remarks

In problem 3 (MC simulation), some solutions fail to calculate mean and STD!

If you do a Fisher transformation, please include either a histogram of the distributions projected on the new Fisher axes, or a value for the separation you achieved (and preferably both). Better than just a ROC curve.

Everytime when you calculate a weighted mean: include both the value of the mean and its uncertainty, and also include the chi2 value and probability to check if you're actually allowed to combine the data in a weighted mean!

For readability and happy TAs (and censors!) it would be great if they could mark (for example with bold font or colour) what their final answer is.

Various remarks

An example of a (favourite) solution to question 2.1 (about the Hubble tension), is the following solution. Not because it has the most pretty figures (it has none), but just because this took 2 seconds to look and to realise it was correct. Efficiently transferring information is great.

2.1

	All measurements	First four measurements	Last three measurements
weighted average (mean)	68.8	73.9	67.8
Error on mean	0.32	0.80	0.35
ChiSquare	52.54	0.50	2.64
Ndof	6	3	2
Probability	$1.45 \cdot 10^{-9}$	0.92	0.16
Do the values agree with each other ?	No, because $p < 0.01$	Yes, because $0.01 < p < 0.99$	Yes, because $0.01 < p < 0.99$

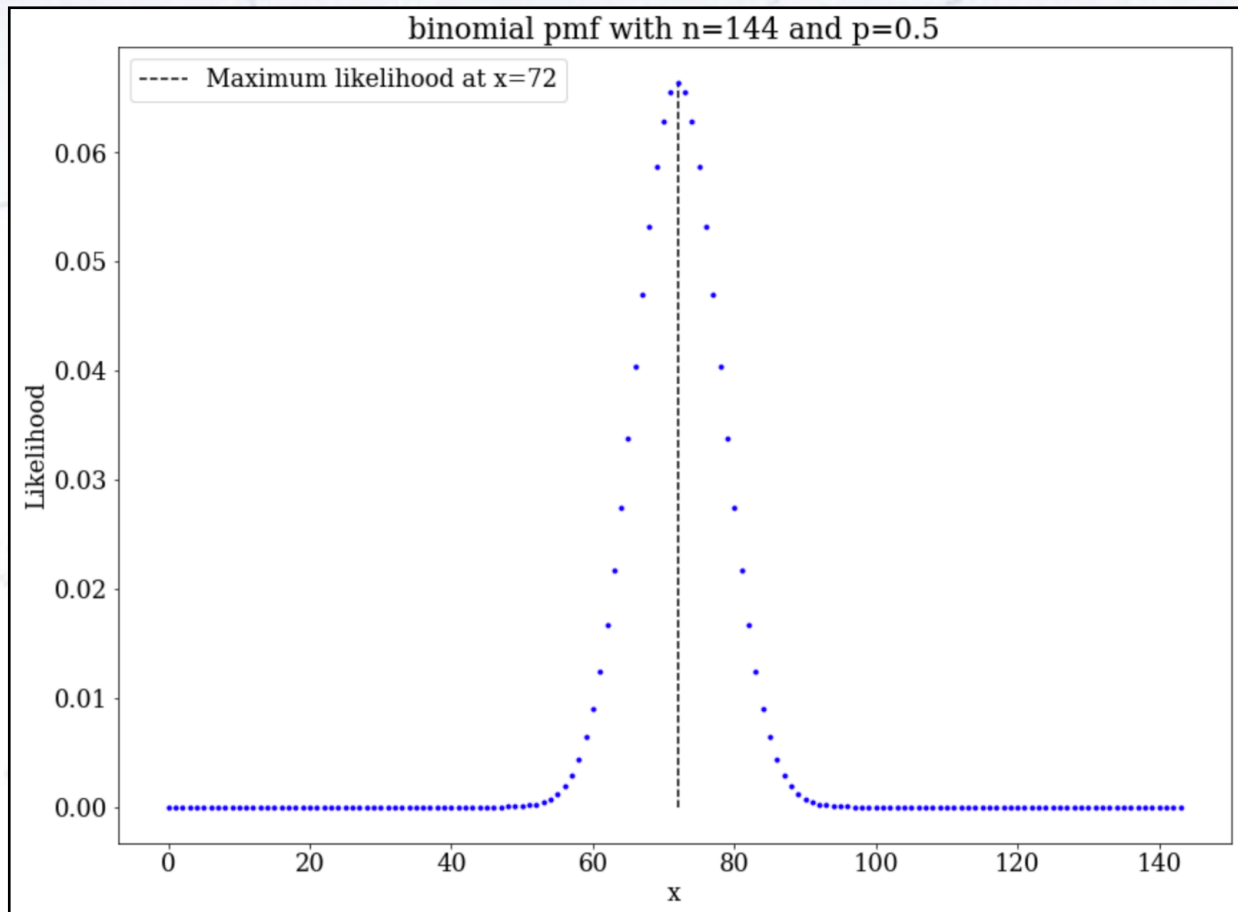
A faded nautical chart showing magnetic isogonic lines. The chart includes a grid of latitude and longitude lines. A prominent line is labeled "MAGNETIC" and "VAR 10° 15' W". Other lines are labeled with values like 30, 60, 90, 120, 150, 180, 210, 240, and 270. The text "THE BITTER END YACHT CLUB" is visible in the upper right corner. The overall image has a light blue and white color scheme with a subtle grid pattern.

The solutions

Problem 1.1

1.1 The probability, that the score after 144 non-draw league games is exactly even can be calculated using the binomial distribution with $N_{\text{trials}} = 144$, $p_{\text{win}} = 0.5$ and the number of wins $n_{\text{wins}} = N_{\text{trials}}/2 = 72$. Using `stats.binom.pmf(nwins, Ntrials, pwin)` the probability that the score is even is

$$P_{\text{even}} = 0.0664$$



Problem 1.2

The local probability of hitting the window is $p_{\text{hit}} = 0.054$, so for the global probability to be at least 90%, the trial factor can be found from

$$\begin{aligned} p &= 1 - (1 - p_{\text{hit}})^{N_{\text{trials}}} \Leftrightarrow \log(1 - p) = \log((1 - p_{\text{hit}})^{N_{\text{trials}}}) \\ \Rightarrow N_{\text{trials}} &= \frac{\log(1 - p)}{\log(1 - p_{\text{hit}})} = \frac{\log(1 - 0.9)}{\log(1 - 0.054)} = 41.5 \end{aligned} \quad (2)$$

Hence to be at least 90 % sure of hitting the window, they need **42** golf balls. This probability should also follow the infinite sum of binomial distributions

$$0.90 \leq \sum_{r=1}^{\infty} (0.054)^r (1 - 0.054)^{N_{\text{trials}} - r} \frac{N_{\text{trials}}!}{r!(N_{\text{trials}} - r)!} \quad (3)$$

where the first value of N_{trials} that solves this is **42**.

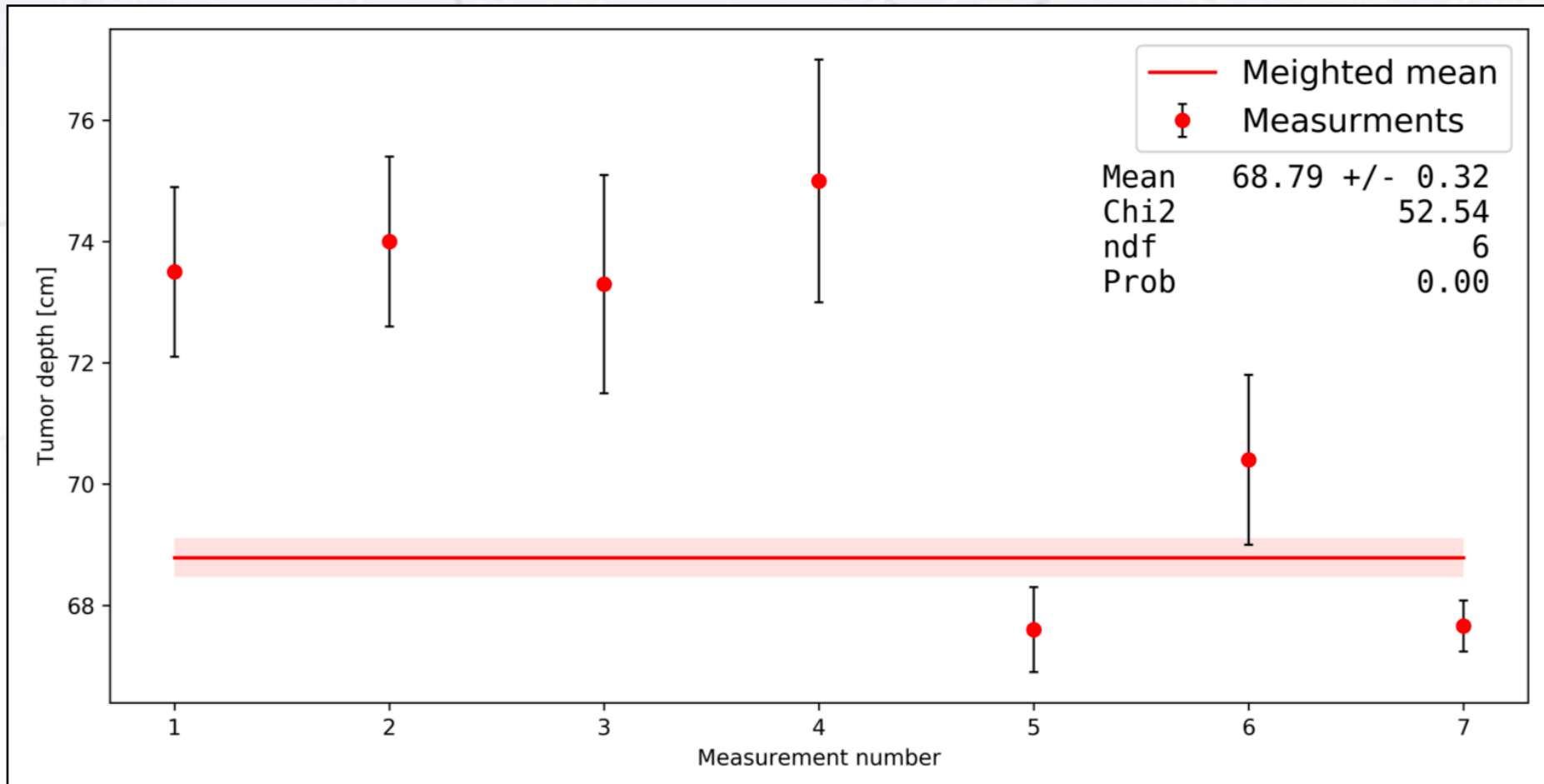
Problem 1.2



Problem inspired by the movie "Fight Club"

Problem 2.1

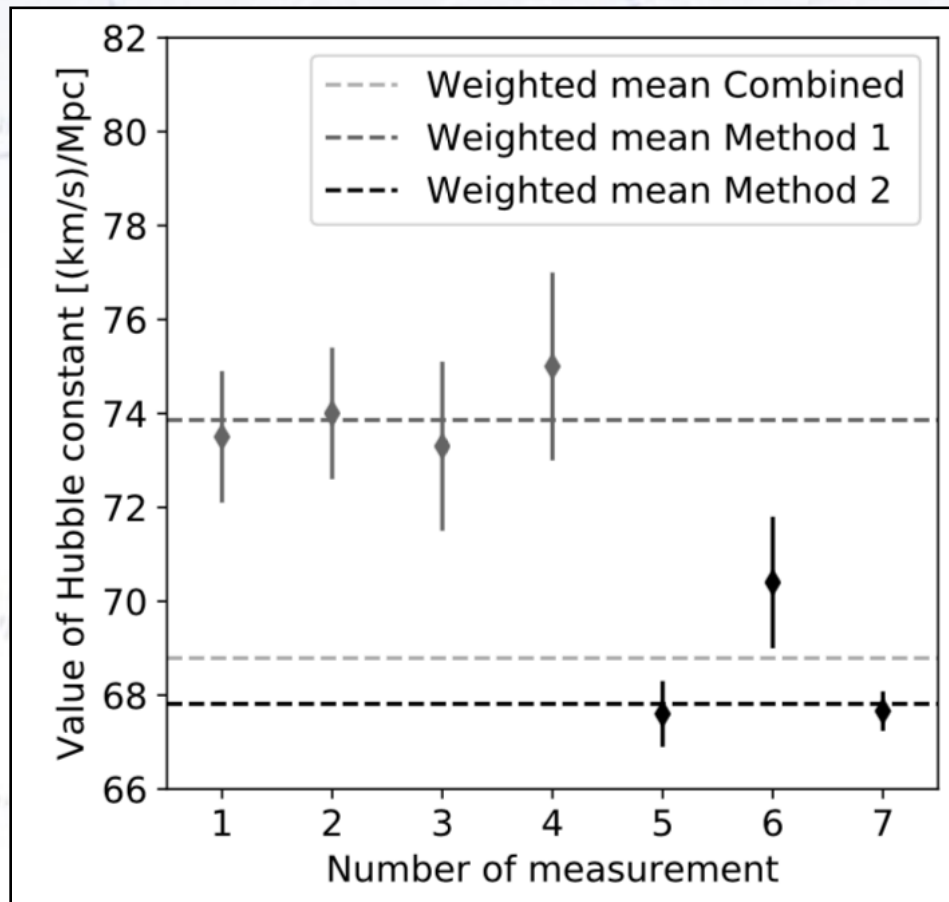
The problem originates from the “Hubble tension”, which have been a classic problem in many fields of science... two methods yield different results! Typically, the outcome is either improved understanding or discovery of something completely new.



Problem 2.1

$$\bar{h}_1 = (73.9 \pm 0.8) \text{ (km/s)/Mpc} \quad \chi^2 = 0.50 \quad p_{\chi^2,3} = 0.92$$

$$\bar{h}_2 = (67.8 \pm 0.3) \text{ (km/s)/Mpc} \quad \chi^2 = 3.6 \quad p_{\chi^2,2} = 0.16$$



Problem 2.2

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

The end result is: $q_0 = 2.0 \pm 0.3 \mu\text{C}$

The contribution from F is:

$$\sigma_{q_0}^F = \sqrt{\frac{d^4}{Q^2 k_e^2} \sigma_F^2} = \underline{\underline{0.18 \mu\text{C}}}$$

The contribution from d is:

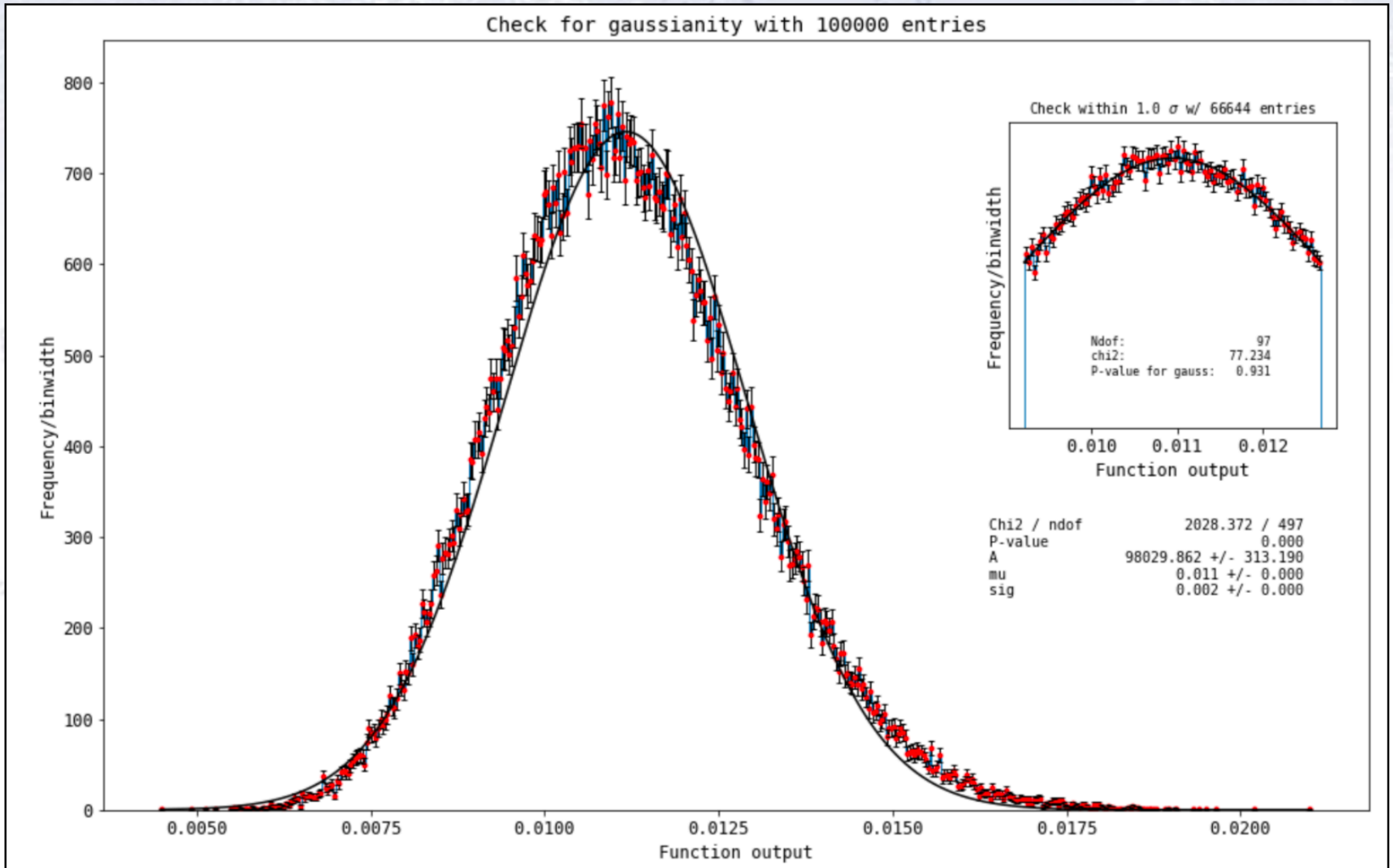
$$\sigma_{q_0}^d = \sqrt{\frac{4F^2 d^2}{Q^2 k_e^2} \sigma_d^2} = \underline{\underline{0.26 \mu\text{C}}}$$

Hence, the largest contribution comes from the distance, d .

Notice, that this is what I was asking for, when I wanted you to state the influence of each input measurement in the Project...

Problem 2.2

Checking if error propagation result is Gaussian...



Problem 2.2

Most of you got the below expression for the uncertainty on q_0 .

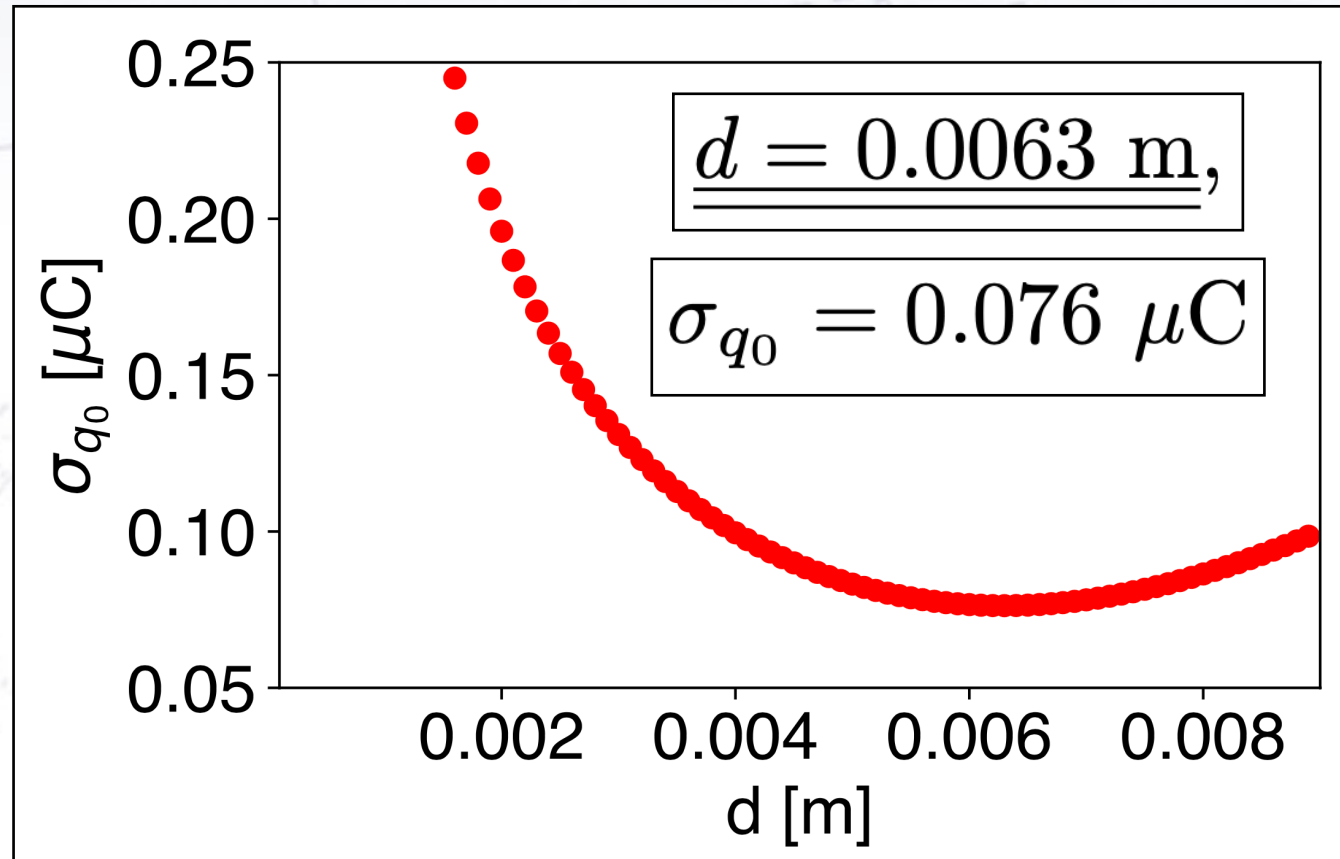
$$\sigma_{q_0} = \sqrt{\frac{4F^2 d^2}{Q^2 k_e^2} \sigma_d^2 + \frac{d^4}{Q^2 k_e^2} \sigma_F^2}$$

Looking at it at first, it seems minimal when $d=0$.

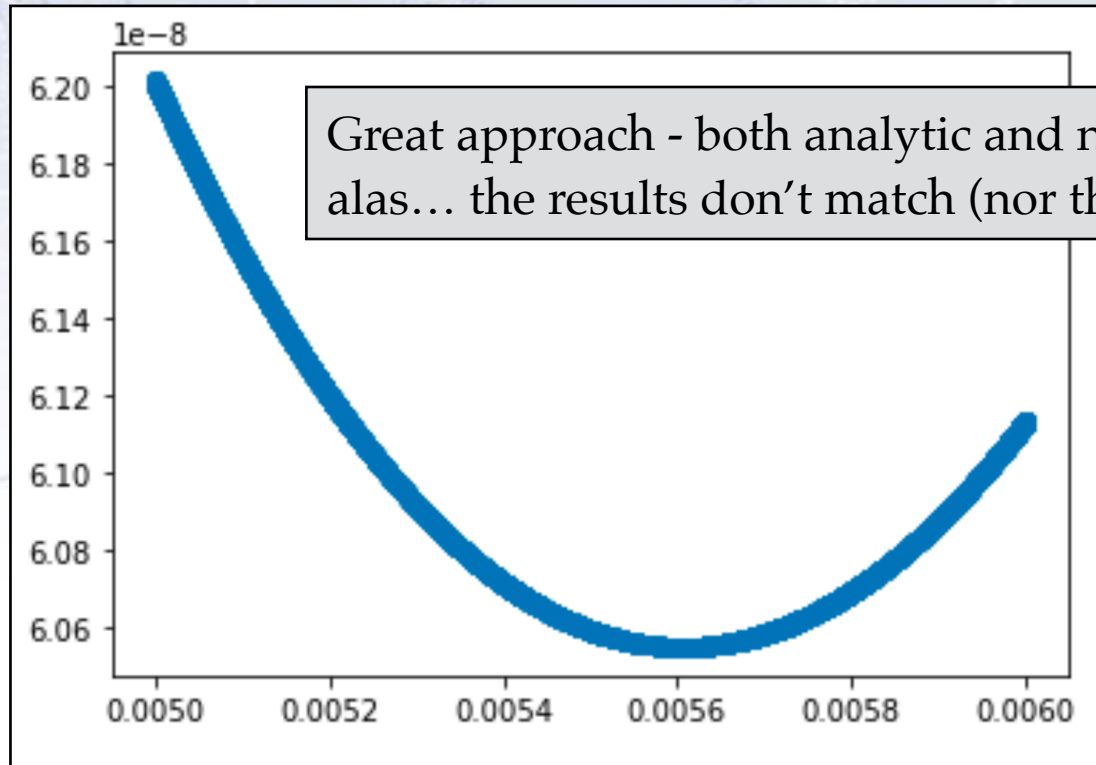
The thing to realise is, that **F changes with distance**, so this needs to be substituted into the equation!

Problem 2.2

$$\sigma_{q_0} = \sqrt{\frac{4 \left(\frac{q_0 Q k_e}{d^2} \right)^2 d^2}{Q^2 k_e^2} \sigma_d^2 + \frac{d^4}{Q^2 k_e^2} \sigma_F^2} = \sqrt{4 \left(\frac{q_0}{d} \right)^2 \sigma_d^2 + \frac{d^4}{Q^2 k_e^2} \sigma_F^2}$$



Problem 2.2



0.5.4 Answer 2.2.3

```
print(f'Distance at Minimum Uncertainty: Calculus - {d_min:.2e} m, Numerical -  
↪{ds_min:.2e} m')
```

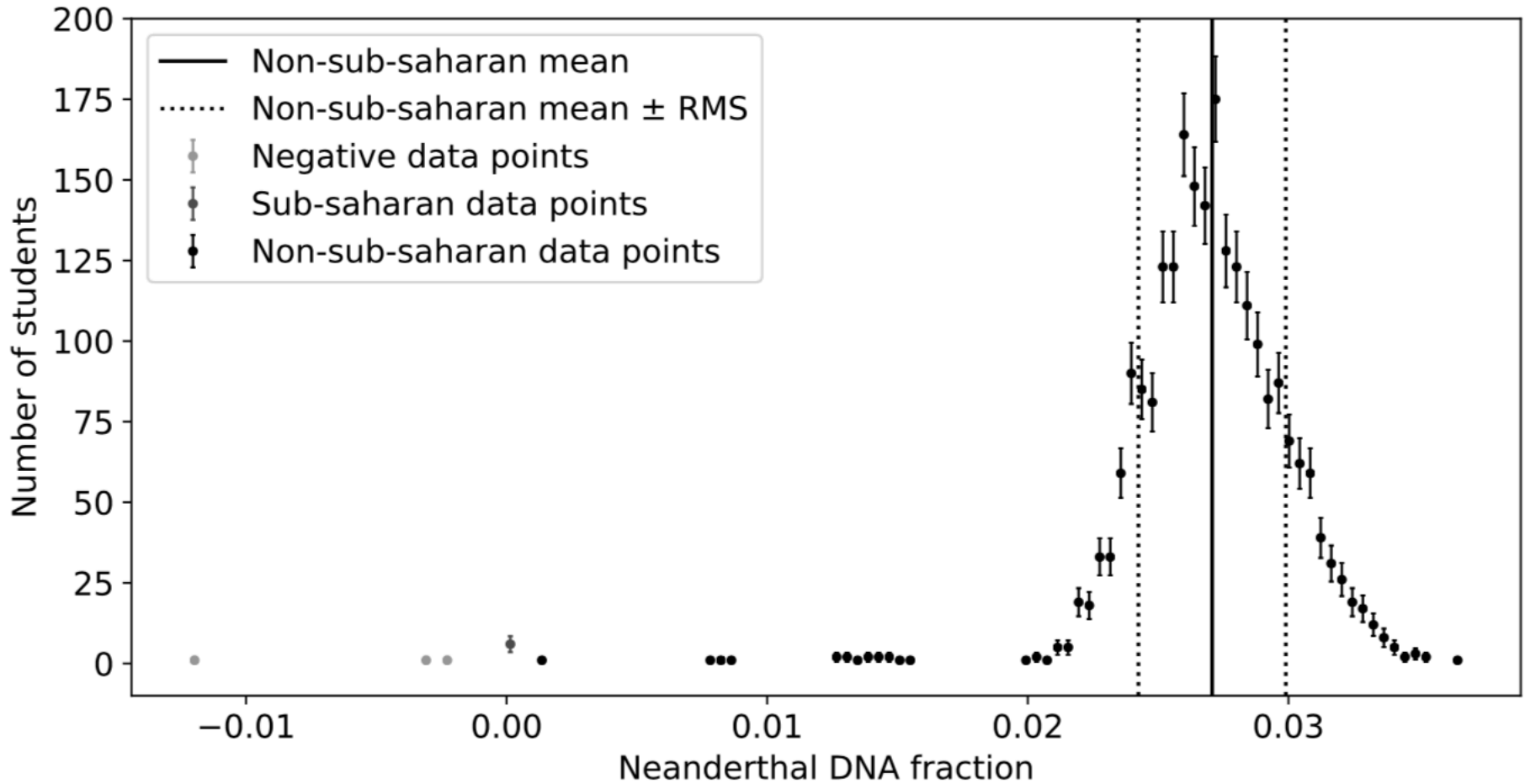
Distance at Minimum Uncertainty: Calculus - 5.88e-03 m, Numerical - 5.61e-03 m

Not sure why the result from calculus and numerical don't agree. The uncertainty on q_0 could be propagated through also to provide an uncertainty on the optimum distance.

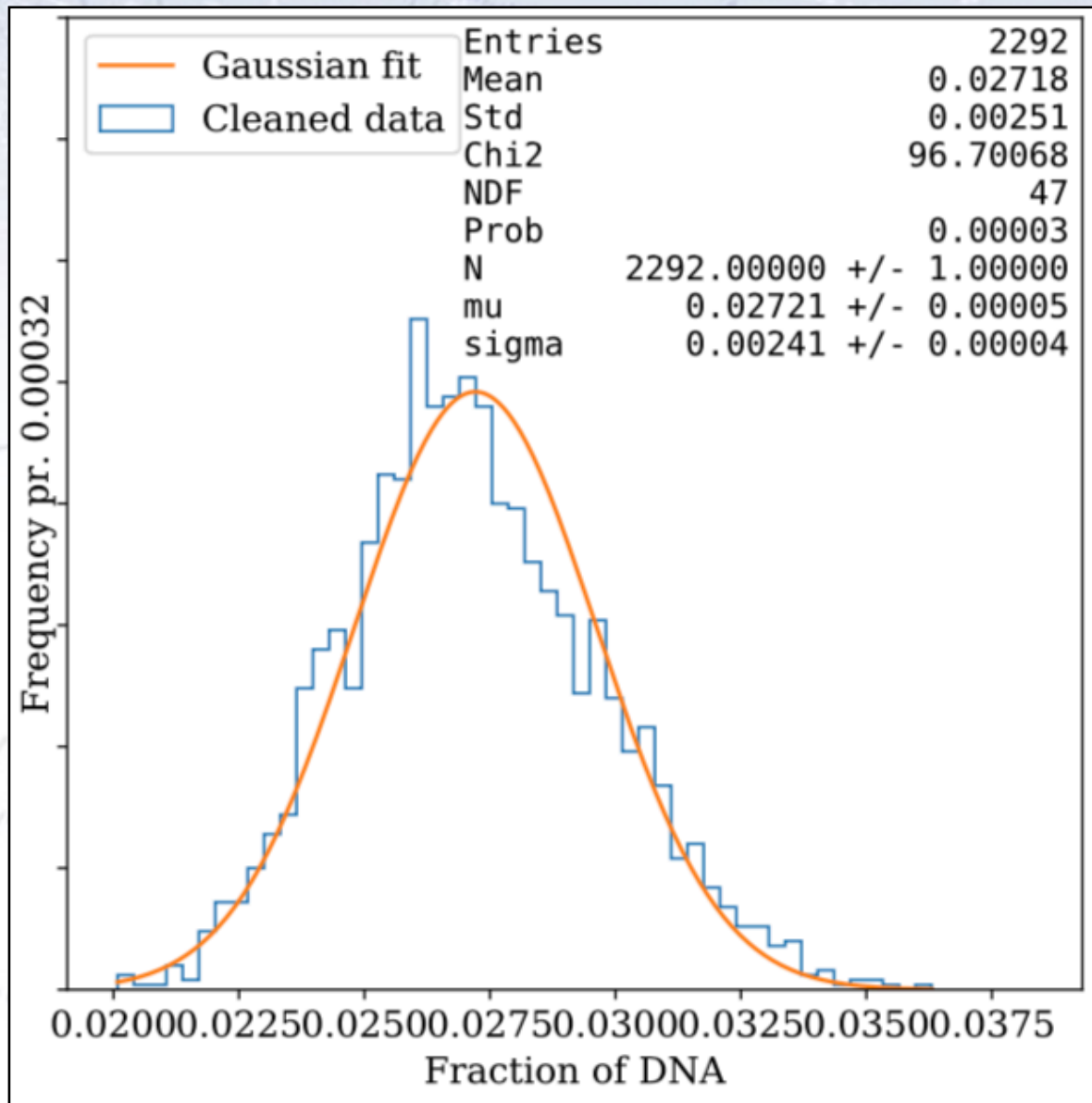
Problem 2.3

$$\mu_{\text{DNA}} = 0.02696 \pm 0.00007$$

$$\text{RMS}_{\text{DNA}} = 0.0034$$

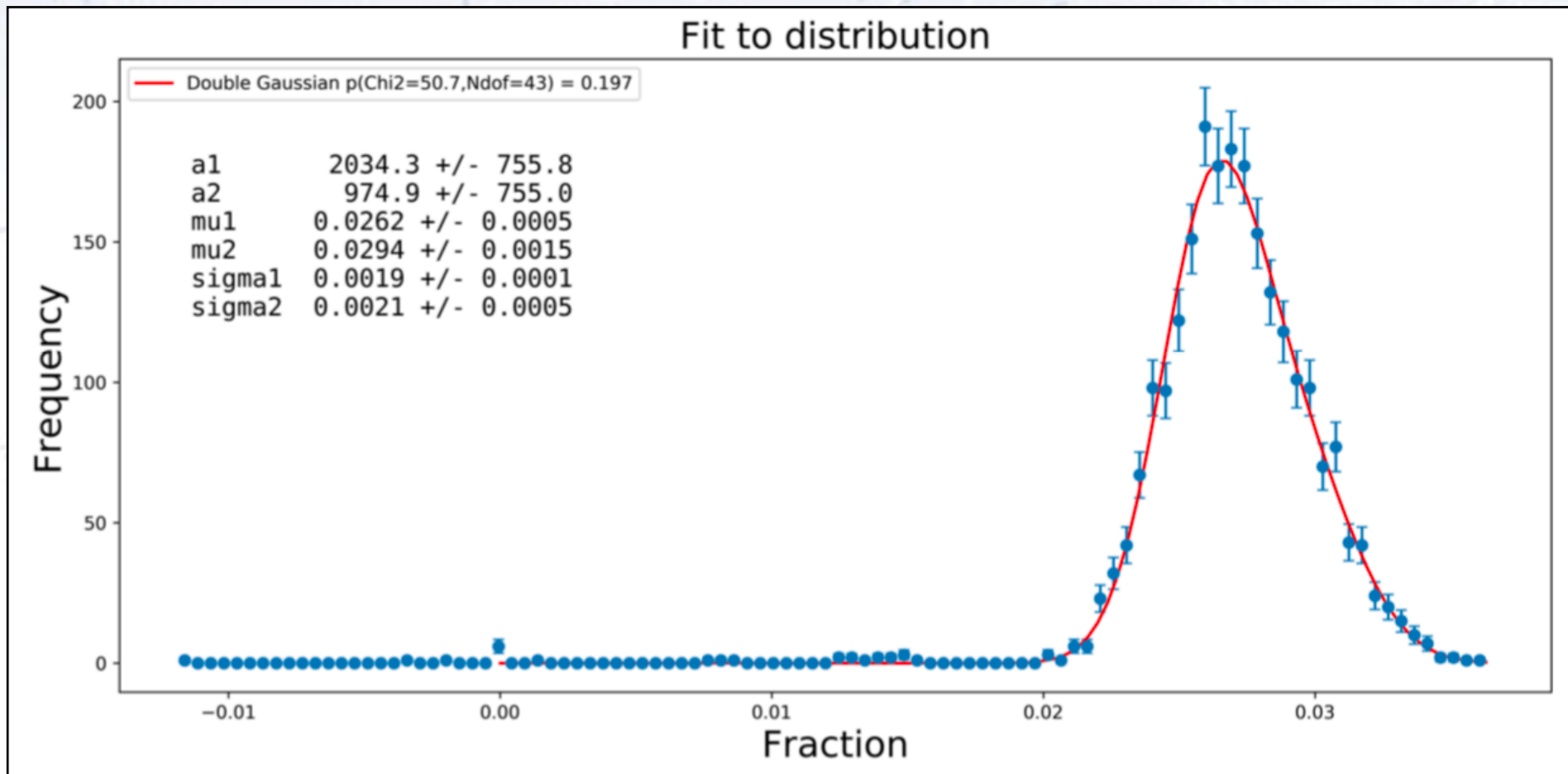


Problem 2.3



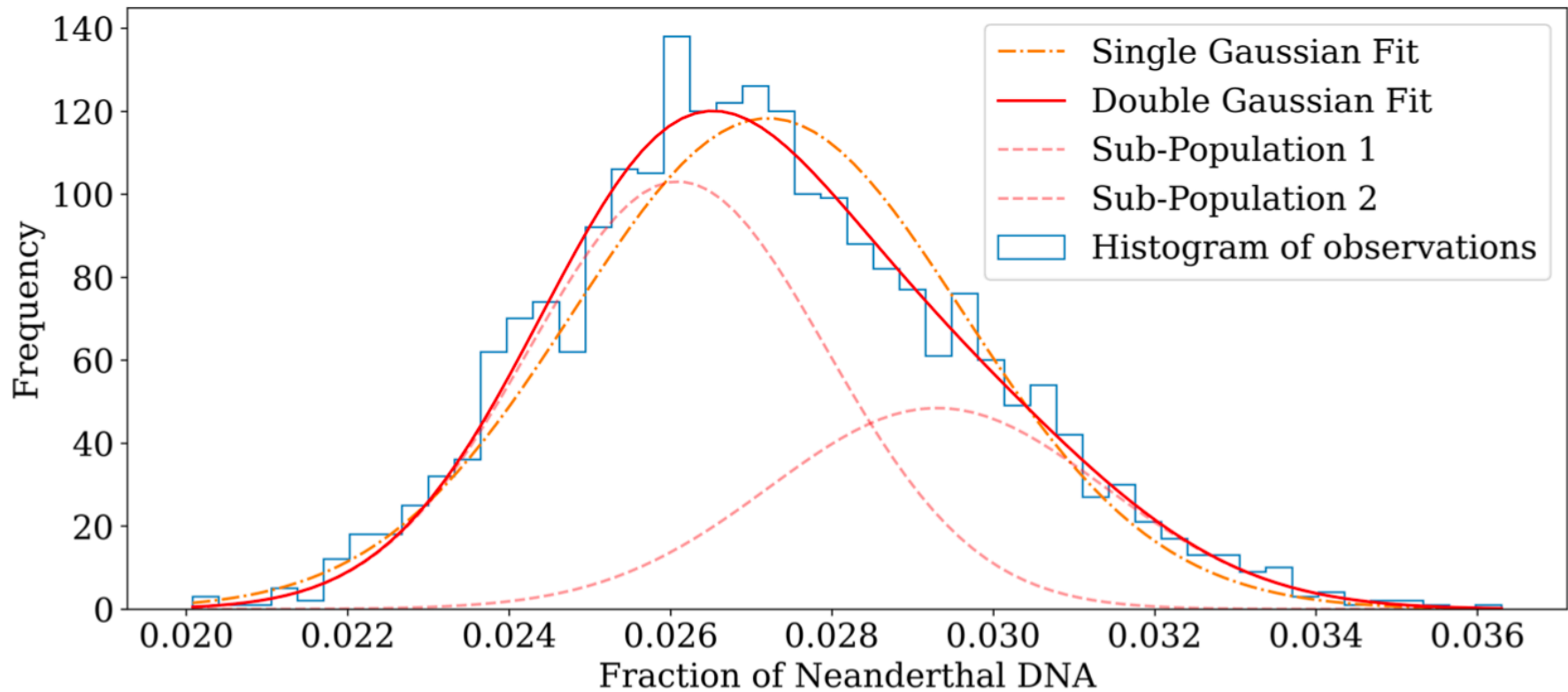
Problem 2.3

A **double Gaussian** does a good job... and this is actually, what I produced the data with! The original data looks very much like this data, but does not have any negative values, and fit the single Gaussian well (not what I wanted!).



Problem 2.3

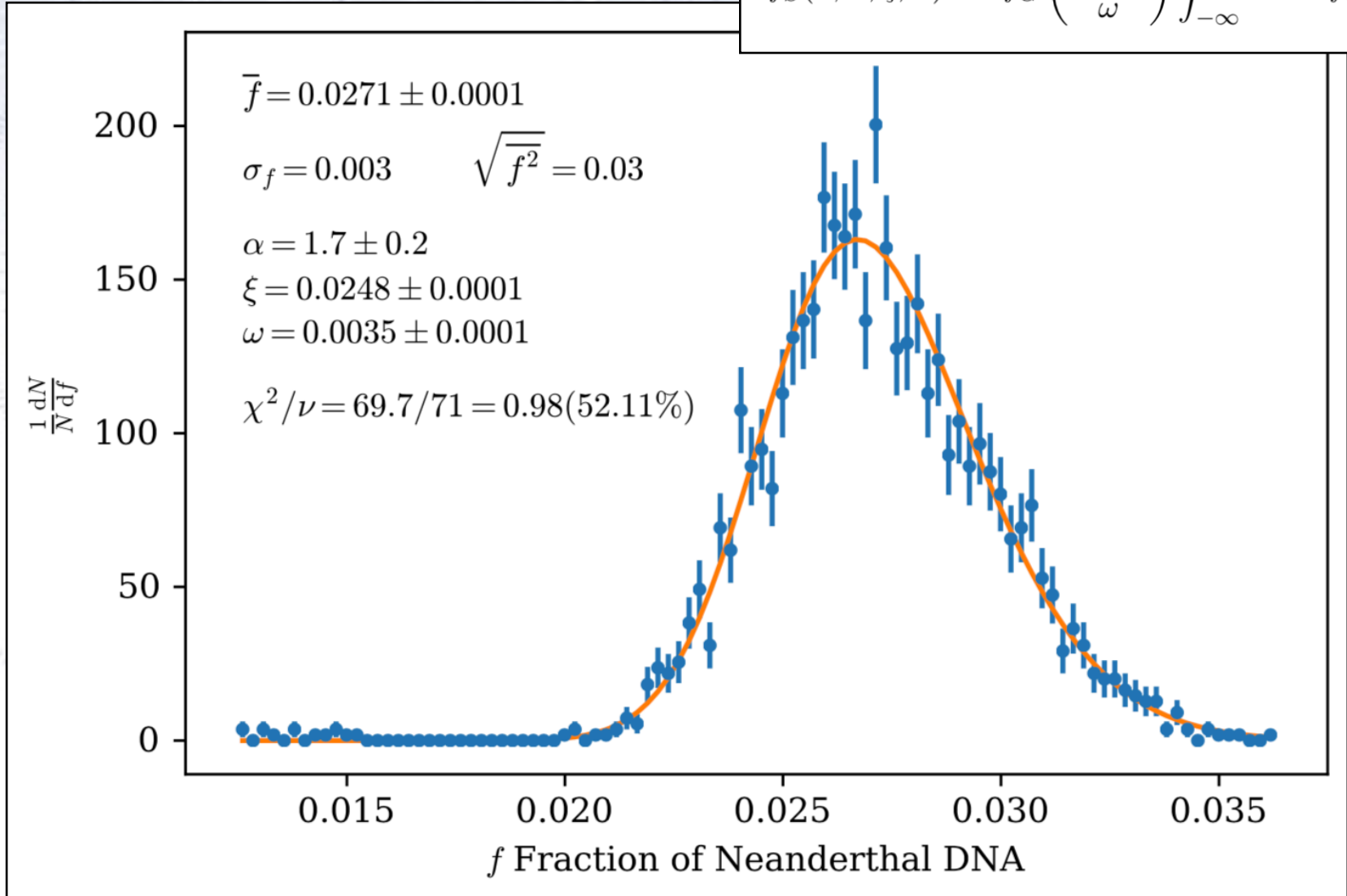
A double Gaussian does a good job... and this is actually, what I produced the data with! The original data looks very much like this data, but does not have any negative values, and fit the single Gaussian well (not what I wanted!).



Problem 2.3

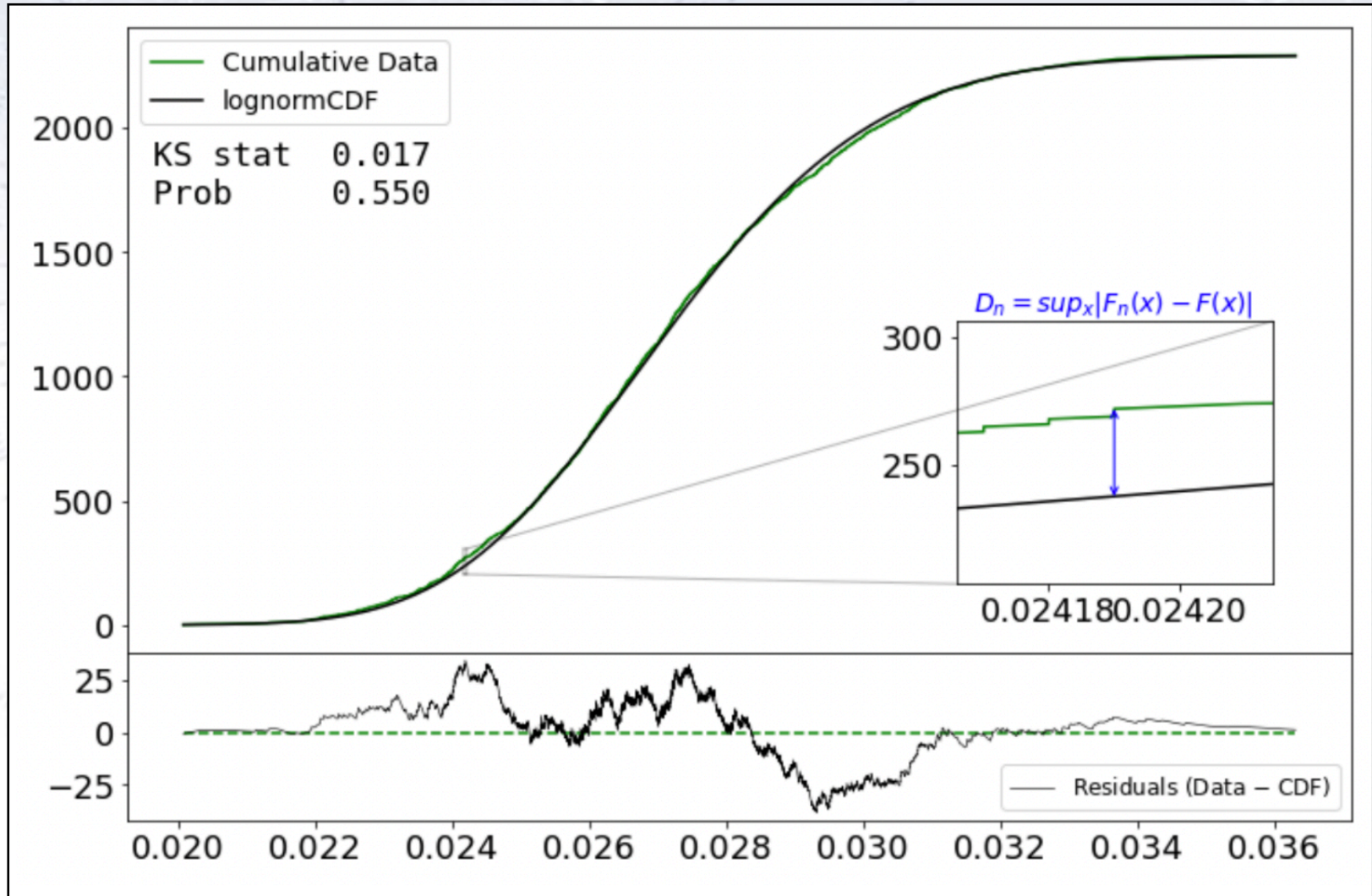
Skewed normal distribution:

$$f_S(x; \alpha, \xi, \omega) = 2f_G\left(\frac{x - \xi}{\omega}\right) \int_{-\infty}^{\alpha \frac{x - \xi}{\omega}} dx' f_G(x')$$



Problem 2.3

KS-test to what degree the distribution fitted the data:



Problem 3.1

3.1 The value of C is obtained by normalizing $f(x)$:

$$\int_1^3 f(x)dx = 1 \Leftrightarrow \int_1^3 C(1+x^2)dx = C \left[x + \frac{1}{3}x^3 \right]_1^3 = C \left(3 + \frac{1}{3}3^3 - 1 - \frac{1}{3}1^3 \right) = \frac{32}{3}C = 1 \Leftrightarrow \boxed{C = \frac{3}{32}}$$

The mean is found by calculating $\int_1^3 x \cdot C(1+x^2)$ using `scipy.integrate.quad()`. The RMS is calculated in two steps. First, the variance is obtained by calculating $\int_1^3 (x - \mu)^2 \cdot C(1+x^2)$ using `scipy.integrate.quad()`, next the RMS is calculated by taking the square root of the variance. Then,

$$\boxed{\mu = 2.25 \quad \text{RMS} = 0.536}$$

$f(x)$ is shown in fig. 4, left, along with μ and the area within $\mu \pm \text{RMS}$.

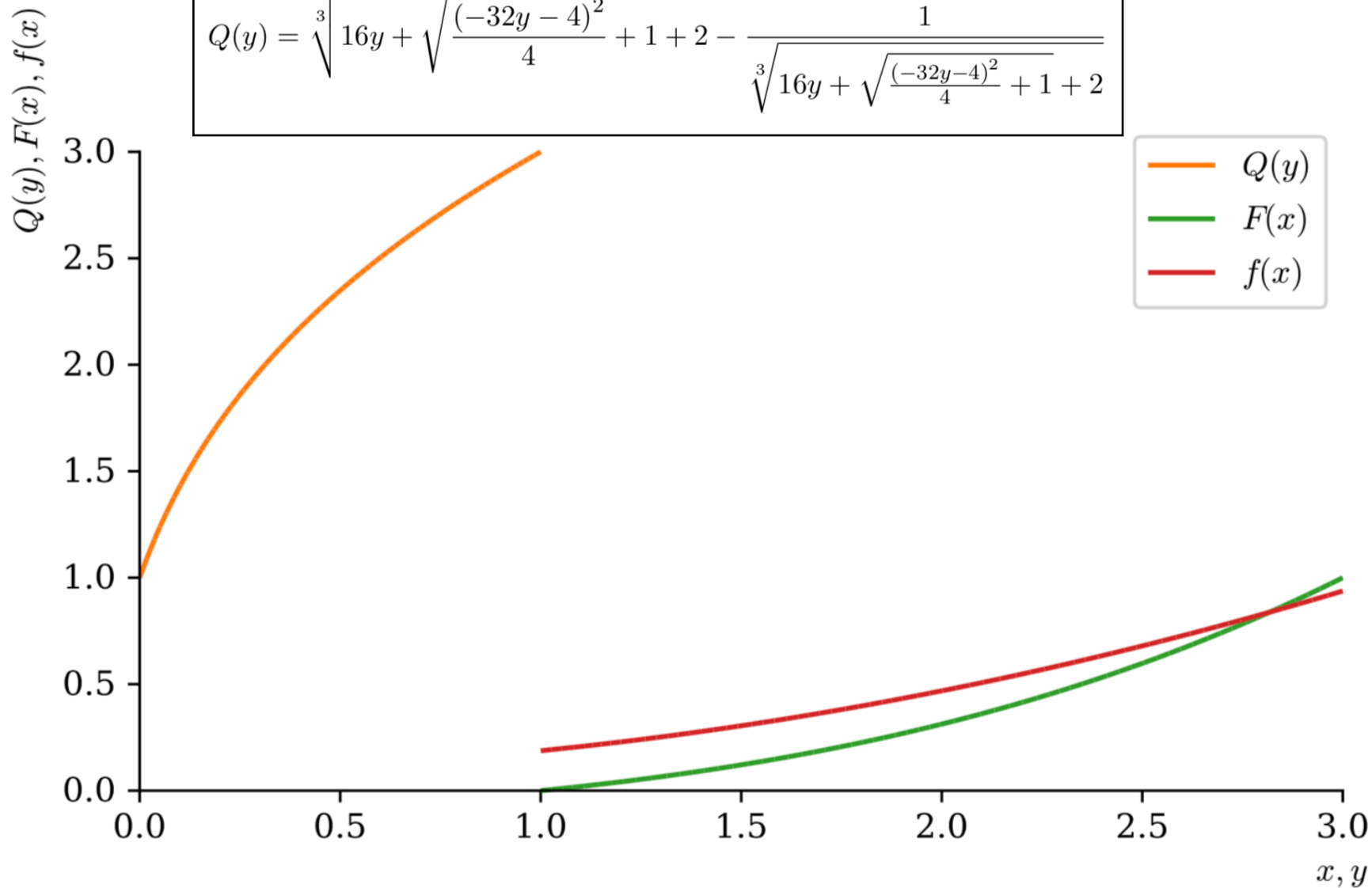
There are generally two ways to generate random numbers: the transformation method and the accept/reject method.

Both the accept/reject and the transformation method can be used

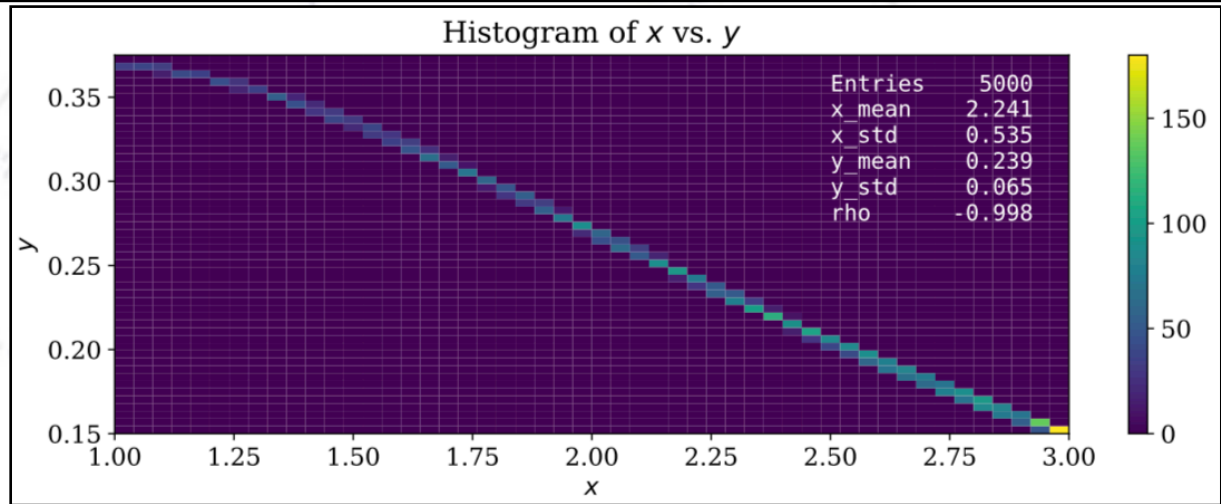
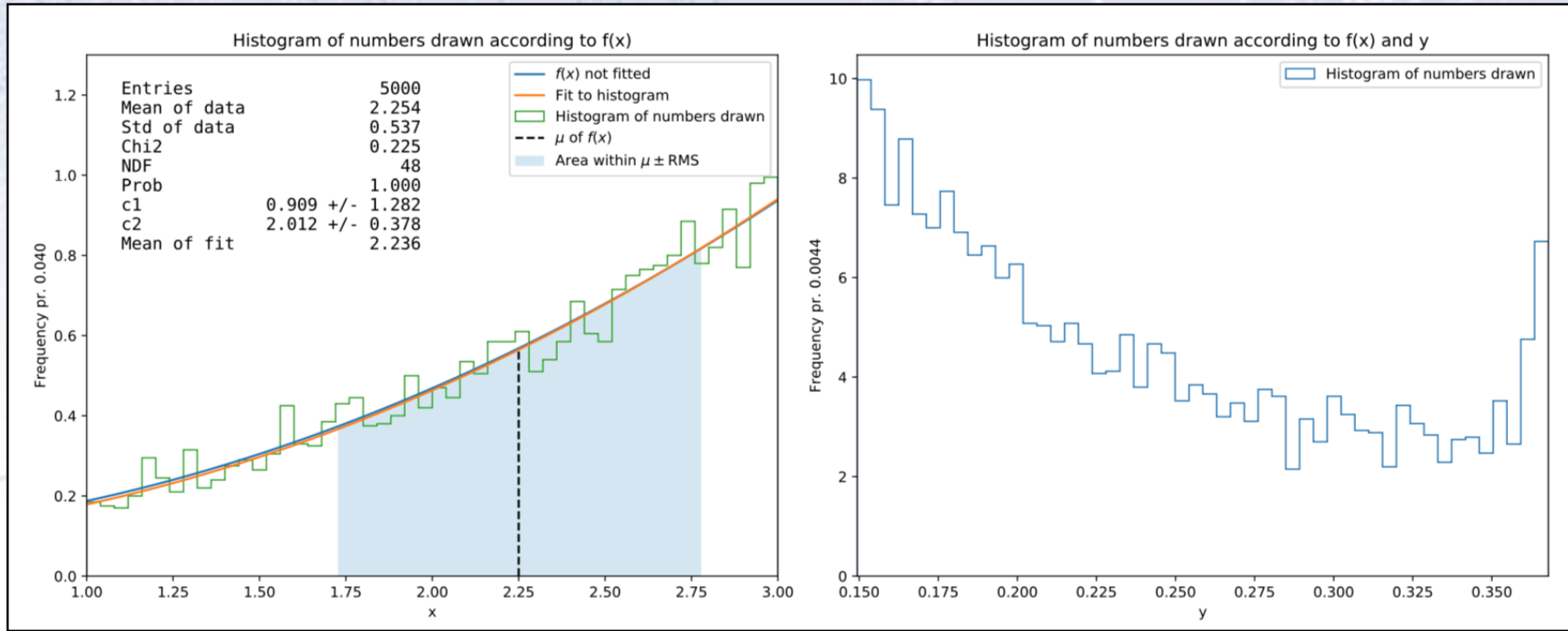
The transformation method works, you may think???

Problem 3.1

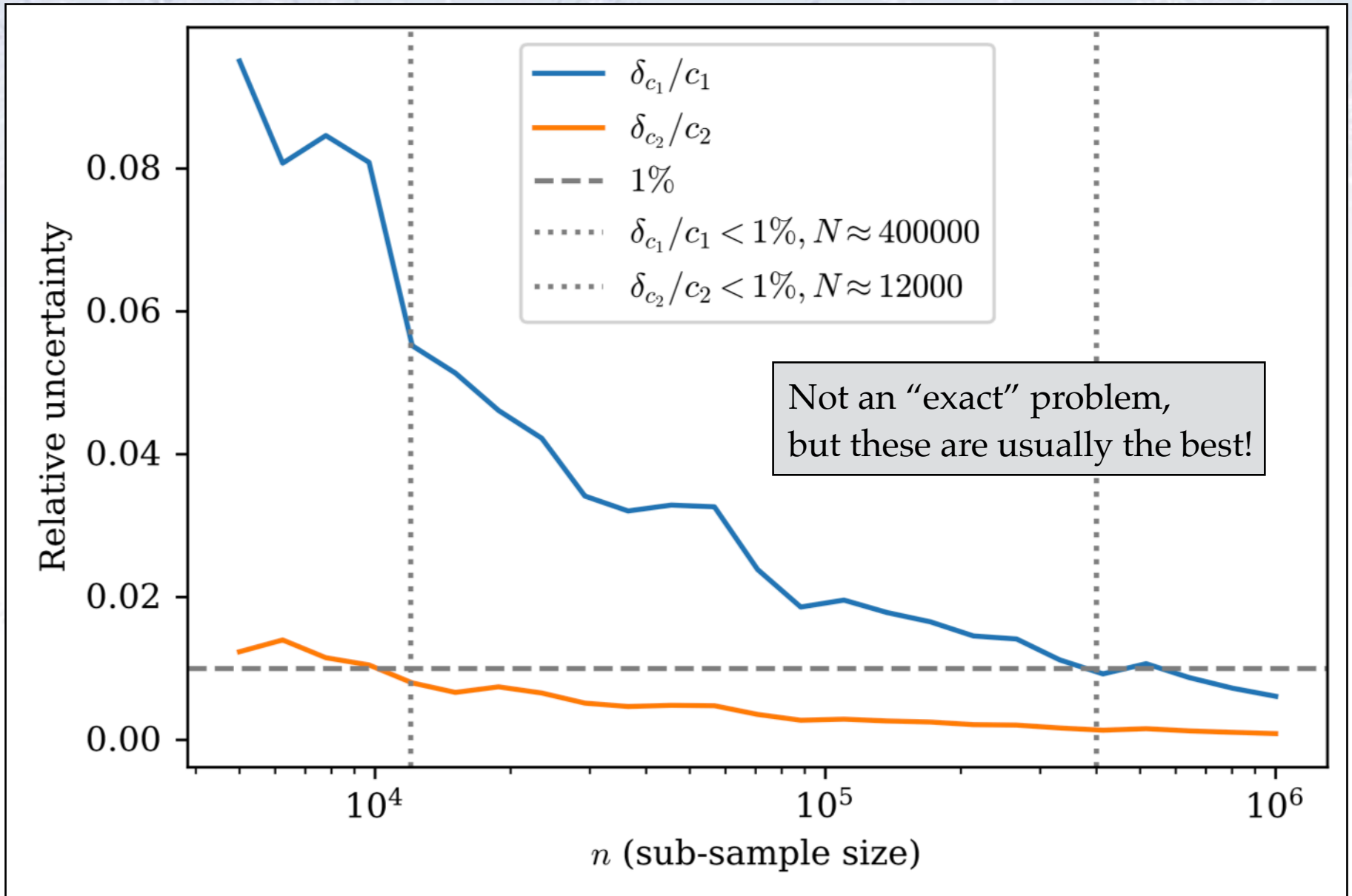
$$Q(y) = \sqrt[3]{16y + \sqrt{\frac{(-32y - 4)^2}{4} + 1} + 1 + 2} - \frac{1}{\sqrt[3]{16y + \sqrt{\frac{(-32y - 4)^2}{4} + 1} + 1 + 2}}$$



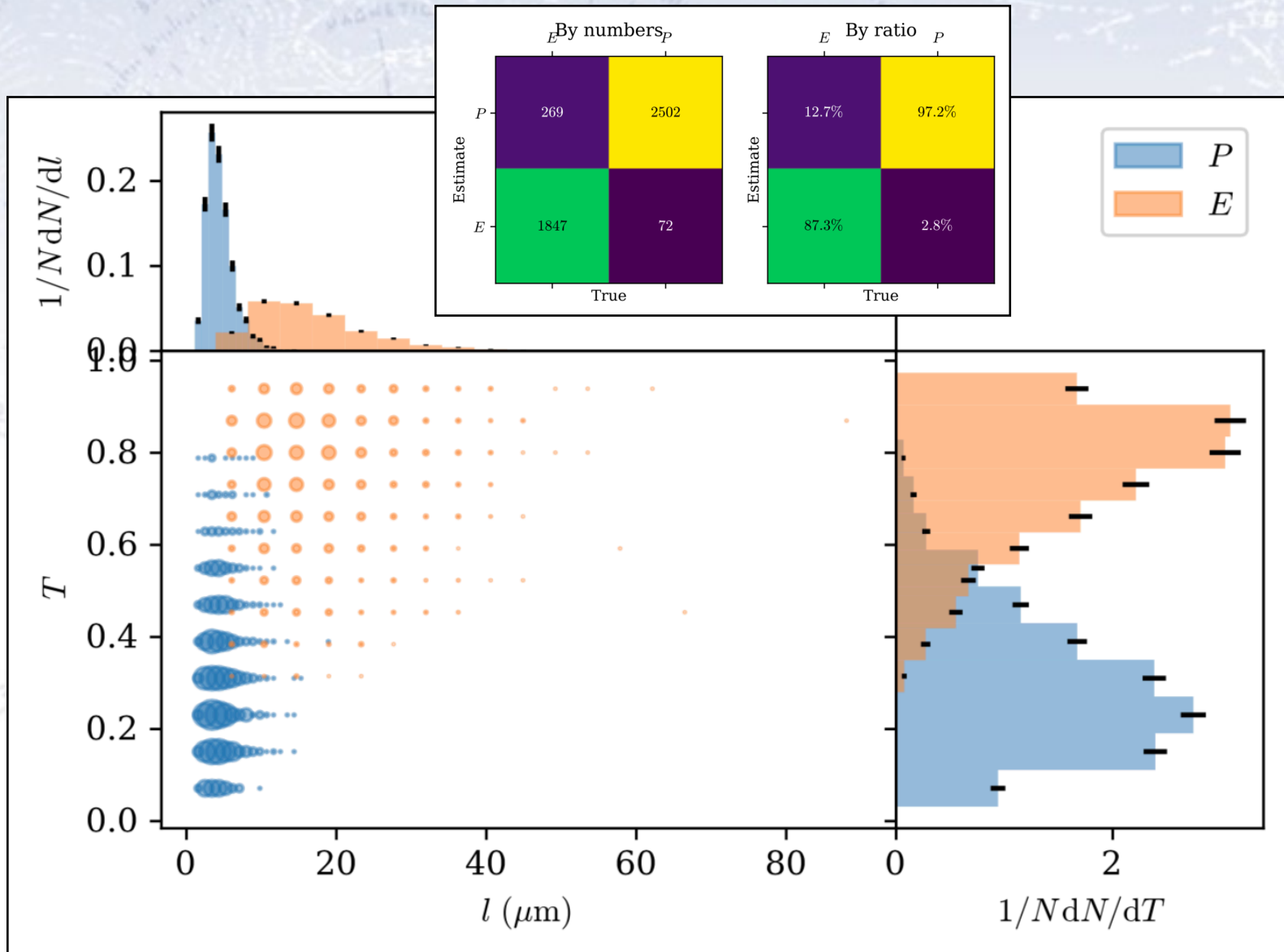
Problem 3.1



Problem 3.1

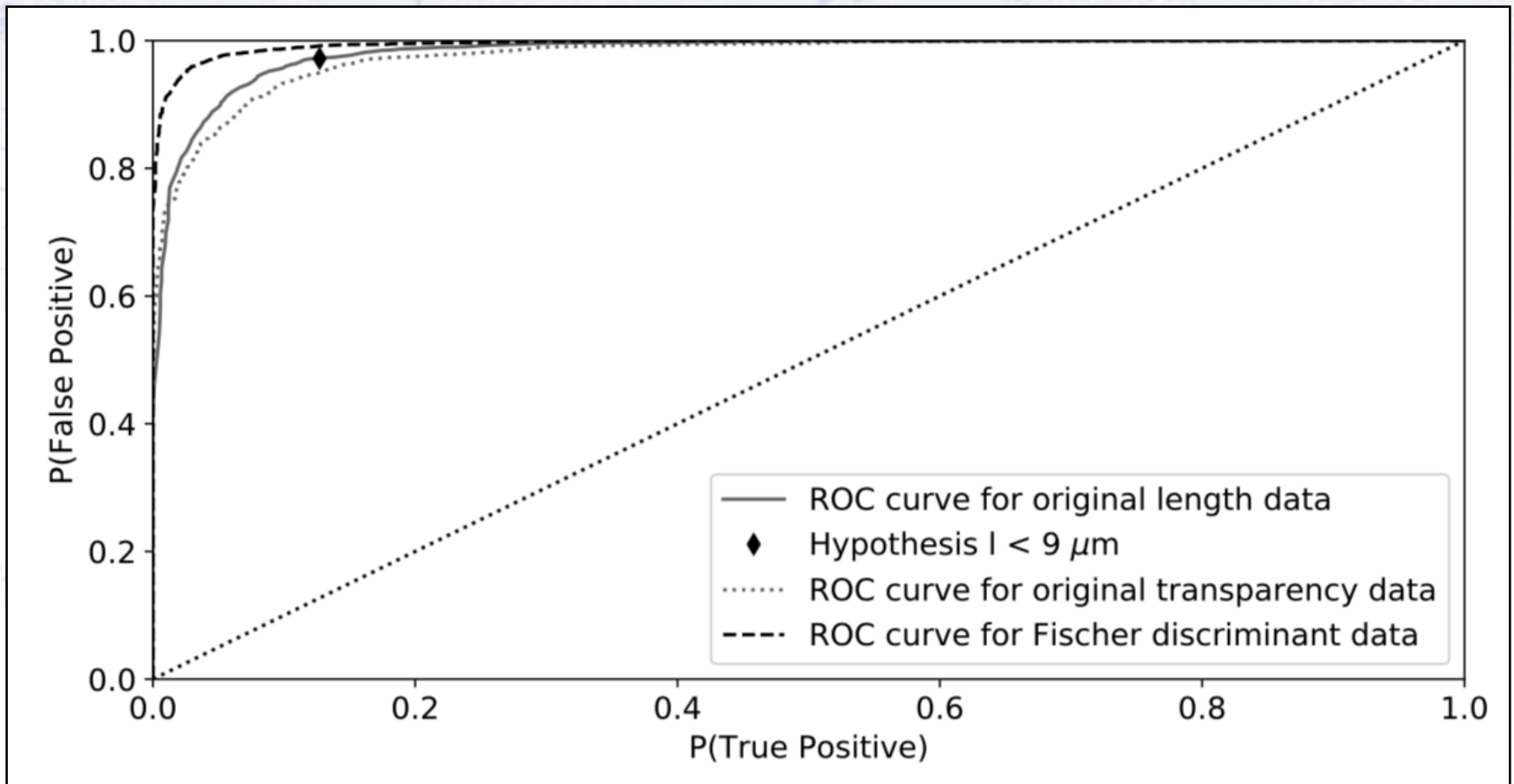


Problem 4.1



Problem 4.1

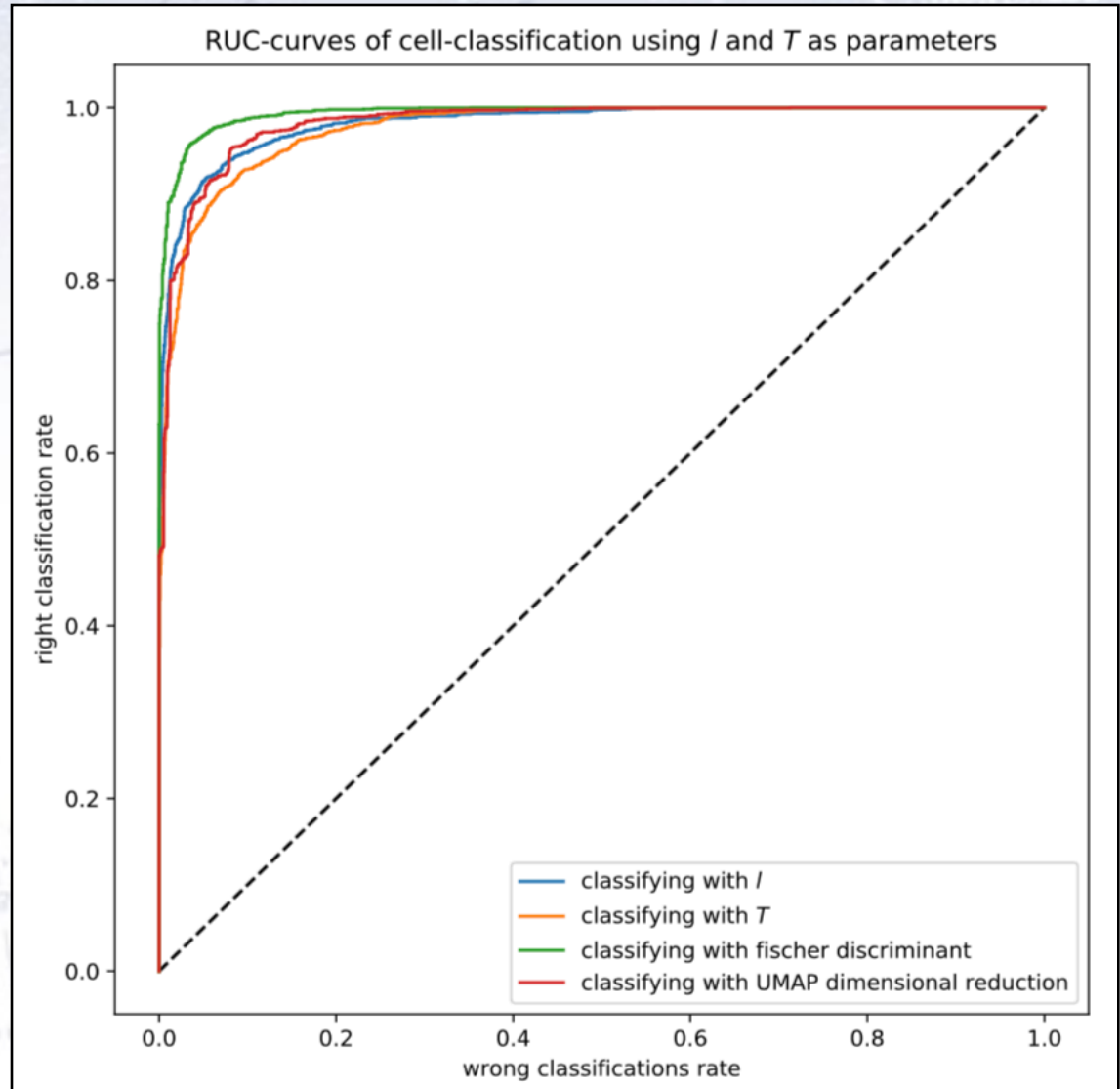
The ROC curve shows, that I is better than T, even if the “sigma distance” is better for T. Due to large tail...



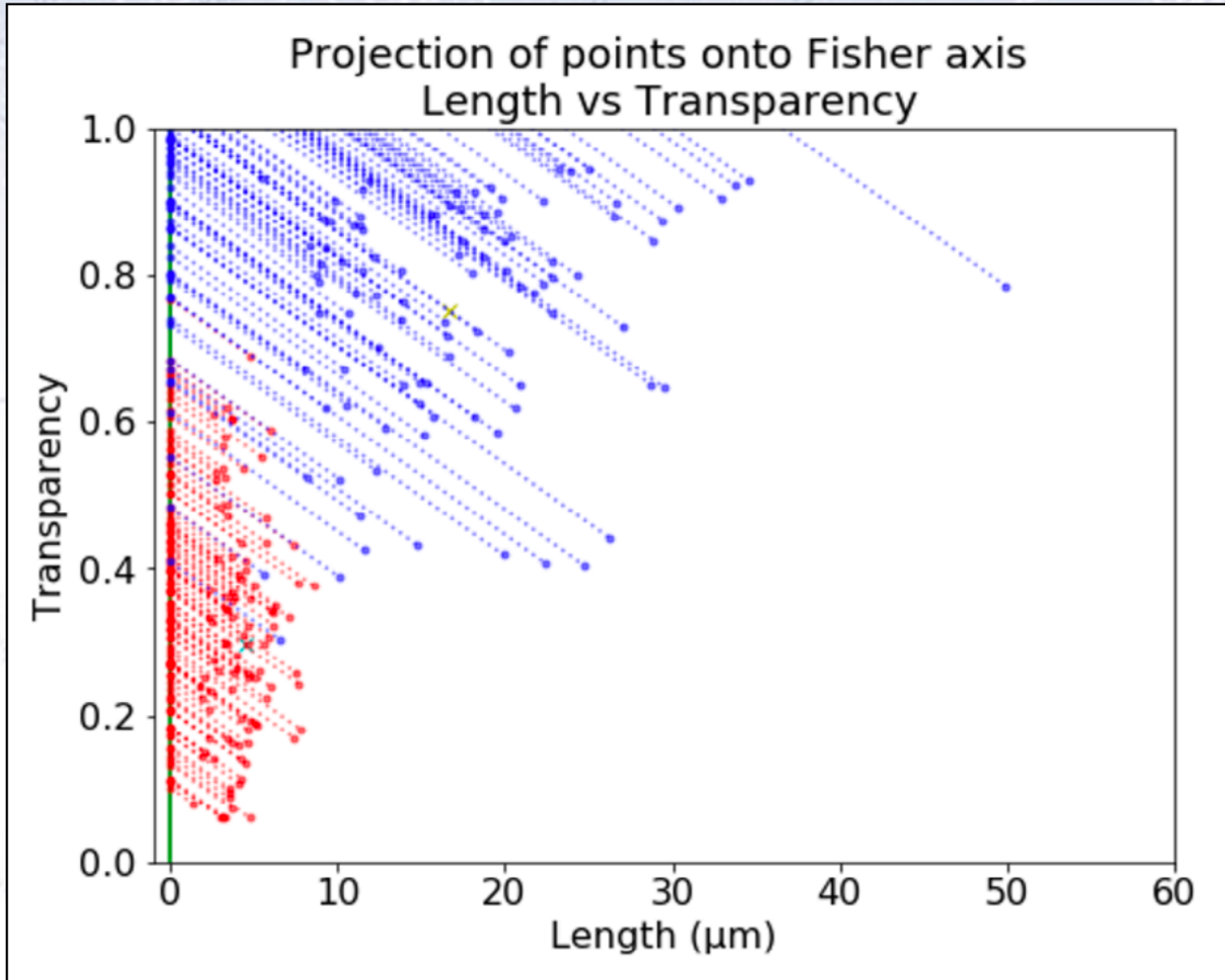
Problem 4.1

With some surplus time,
ML was tried.

Fun to see, that the good
old Fisher was still the
better method.



Fisher Illustrated...



Problem 5.1

$$\sigma_M^2 = \left(\frac{dM}{dC}\right)^2 \sigma_C^2 + \left(\frac{dM}{dG}\right)^2 \sigma_G^2, \quad (26)$$

where σ_C and σ_G are the standard deviation on C and G , respectively. To calculate the solar mass M with uncertainty σ_M , the derivatives are needed

$$\frac{dM}{dC} = \frac{12\pi^2 C^2}{G} \quad \text{and} \quad \frac{dM}{dG} = -\frac{4\pi^2 C^3}{G^2}. \quad (27)$$

Using these equations, the solar mass can be calculated to be $M = (1.775 \pm 0.246) \cdot 10^{30}$ kg.

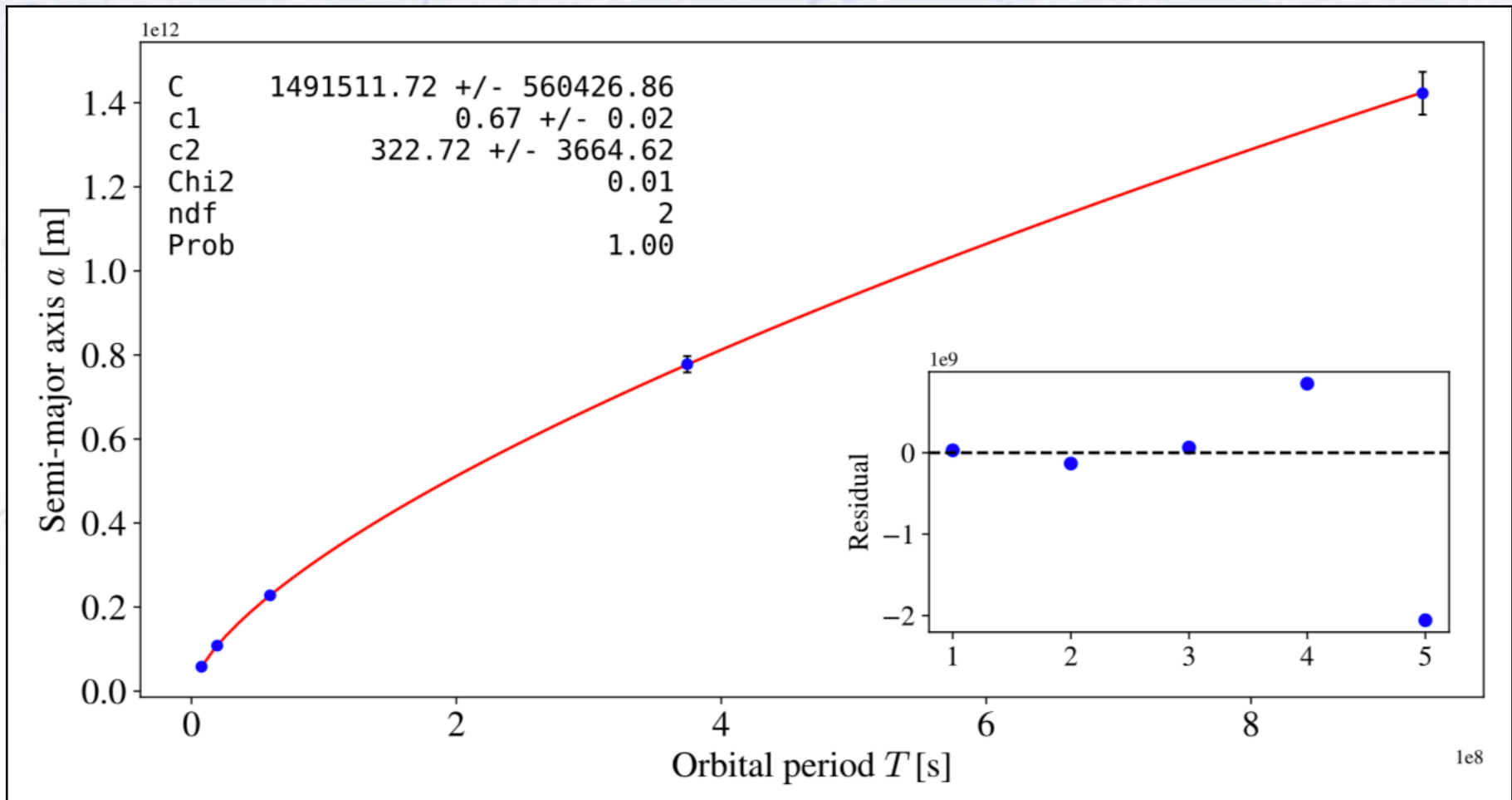
In order to calculate the solar mass in kg I need C in the right units. From the fit C is in the units $AU \cdot days^{3/2}$, so by using that $1AU = 149597870700m$ and the number of seconds in a day is $s/day = 60s/min \cdot 60min/hour \cdot 24hour/day$ is converted to be $C = 5727779217714.285 \pm 70292052150.28448$. Inserting this into the equation for M gives $M = (1.8 \pm 0.3) \cdot 10^{30} kg$. Wikipedia tells me that the real value is $1,989 \cdot 10^{30}[1]$ (with no uncertainties, so technically this information is useless!), meaning that my value is roughly 0.12 percentage off, which is fairly okay seeing as I am also using G_{1778} , which is also 14% off by the real value of $6.67 \cdot 10^{-11} \frac{m^3}{kg s^2}$.

The number of decimals given here is “a little excessive” :-)

The Wikipedia answer for $M(\text{sun})$ has no error (probably because it is very small).

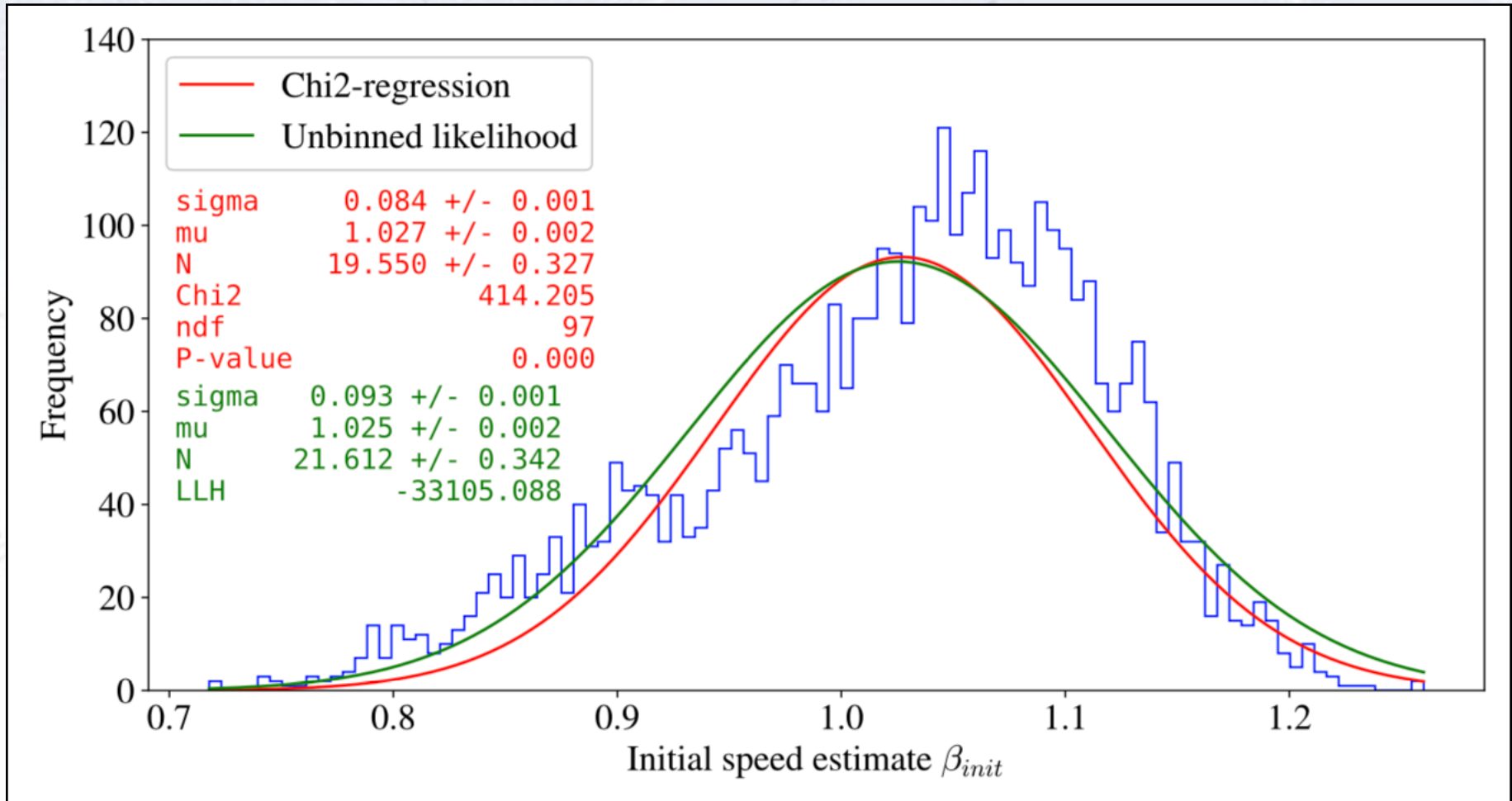
Problem 5.1

The additional freedoms do not improve the fit, but simply shows to what degree Kepler's formula is correct.



Problem 5.2

However you try to fit, it doesn't help you. Also, the LLH fit is hard to evaluate!



Problem 5.2

This was a surprise for you... and me!

I would have thought, that you would take those above/below, and compare them with a Kolmogorov-Smirnof test, but hardly anybody did this!

$$\text{datapoint} = \frac{\text{below}}{\text{above}} = \frac{1989}{2011} = 0.989 \approx 1 \quad (49)$$

This shows there is close to equal amount of data on each side, with only $1 - 0.989/1 = 0.011 = 1.1\%$ difference, indicating symmetry around the axis. We found the mean= 1.576 ± 0.011 , median= 1.577 and std= 0.694 , which again indicates symmetry around $\pi/2 = 1.5707$, since the values lie really close. I have also checked how far away the median is from the symmetry line, because there is a small difference:

$$\text{median} : \frac{\pi/2 - 1.577}{1} = -0.0062 \approx 0.62\% \quad (50)$$

To check how far away the mean is from the symmetry line, a one sample z-test have been made , because we found an uncertainty on the mean as $\sigma_{\text{mean}} = \text{std}/\sqrt{N}$, where N is total number of data points.

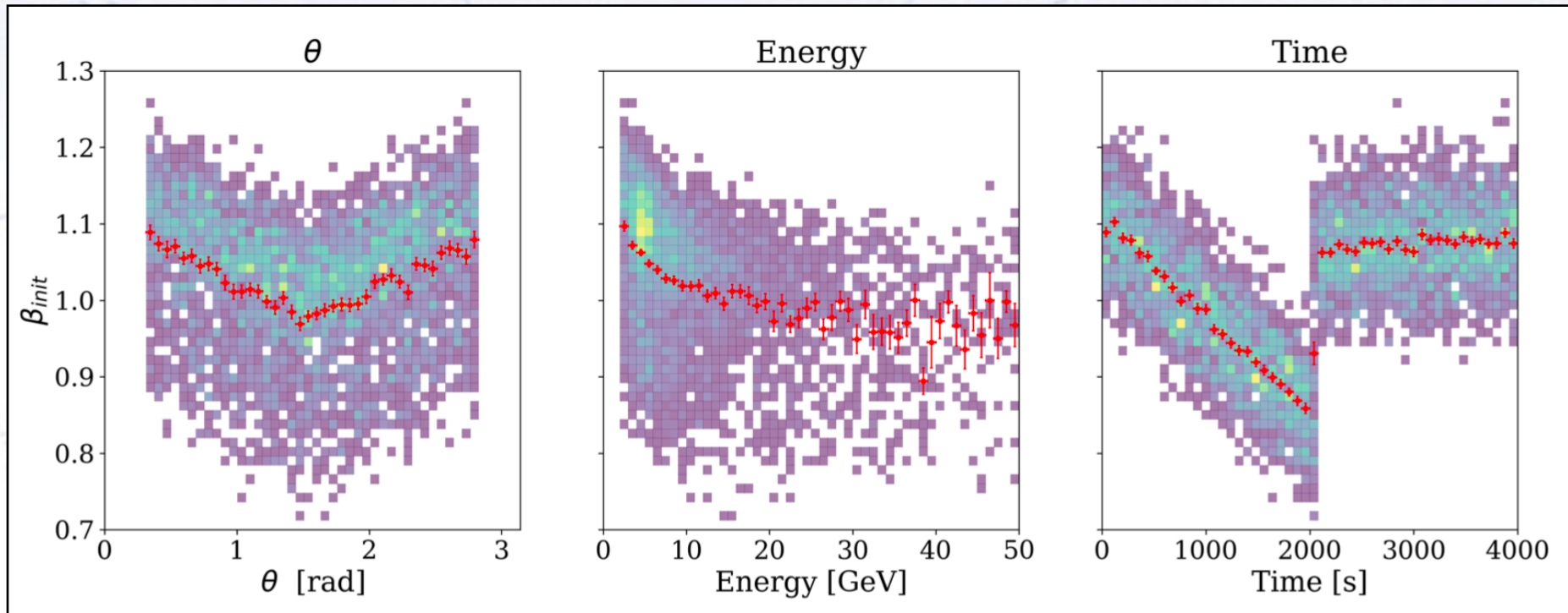
$$z_{\text{mean}} = \frac{\text{mean} - \pi/2}{\sigma_{\text{mean}}} = 0.436 \approx \mathbf{0.4\sigma} \quad (51)$$

Problem 5.2

Several things to comment on:

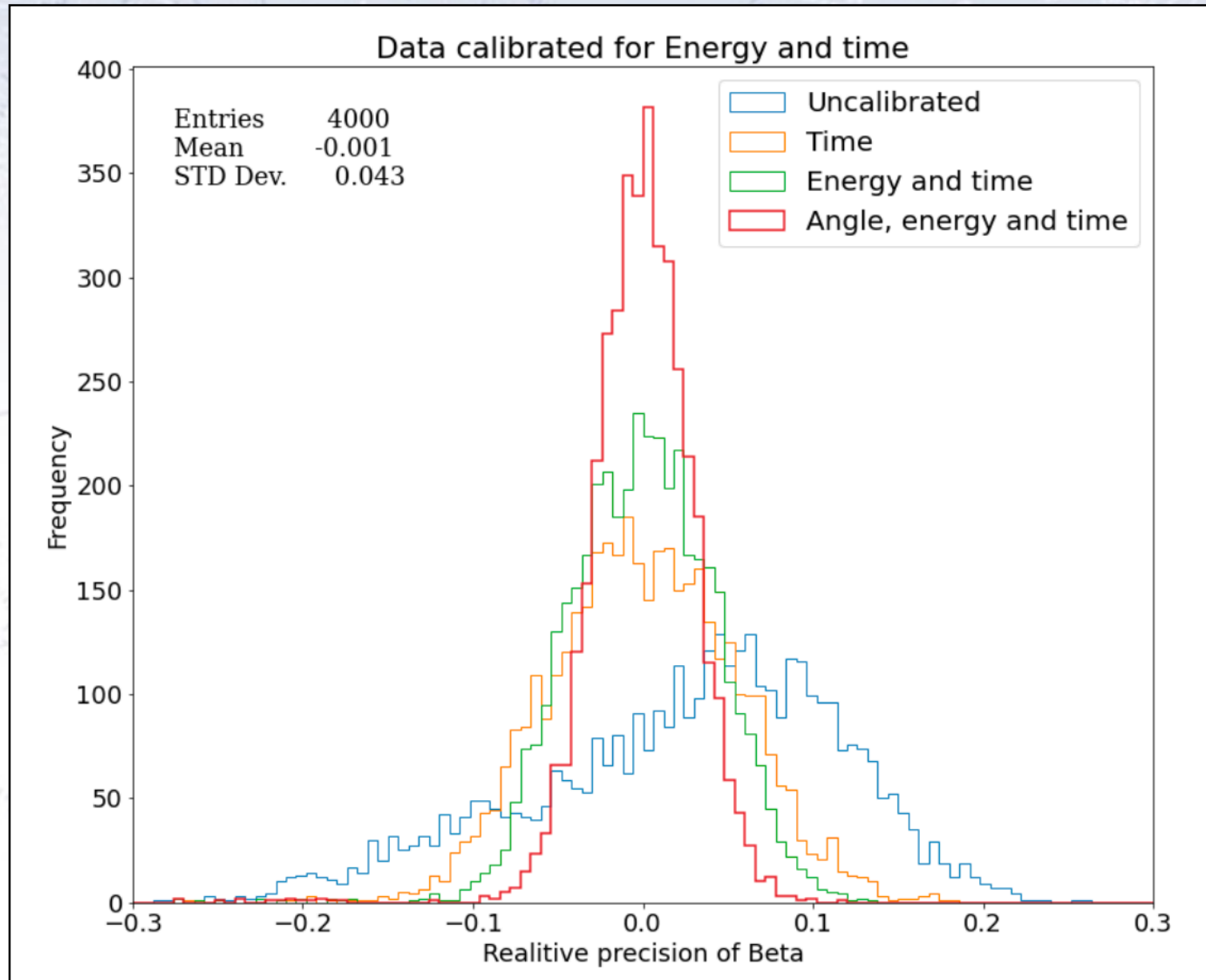
Do you want to use RMS or error on mean, when testing if there is a dependency?

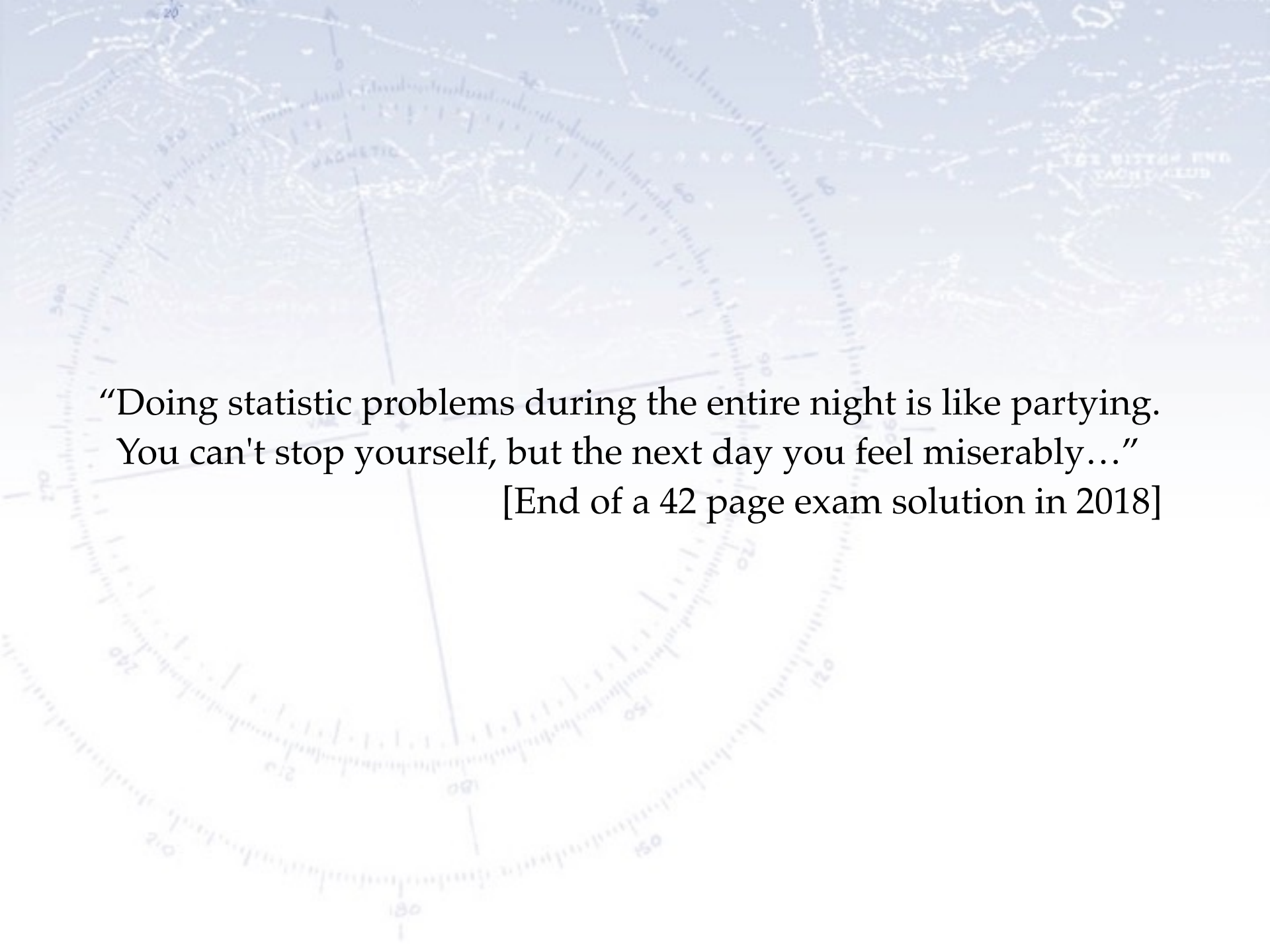
And how about the low stat. bins at high E?



But clearly all three variables have an impact, and things should improve with each correction.

Problem 5.2





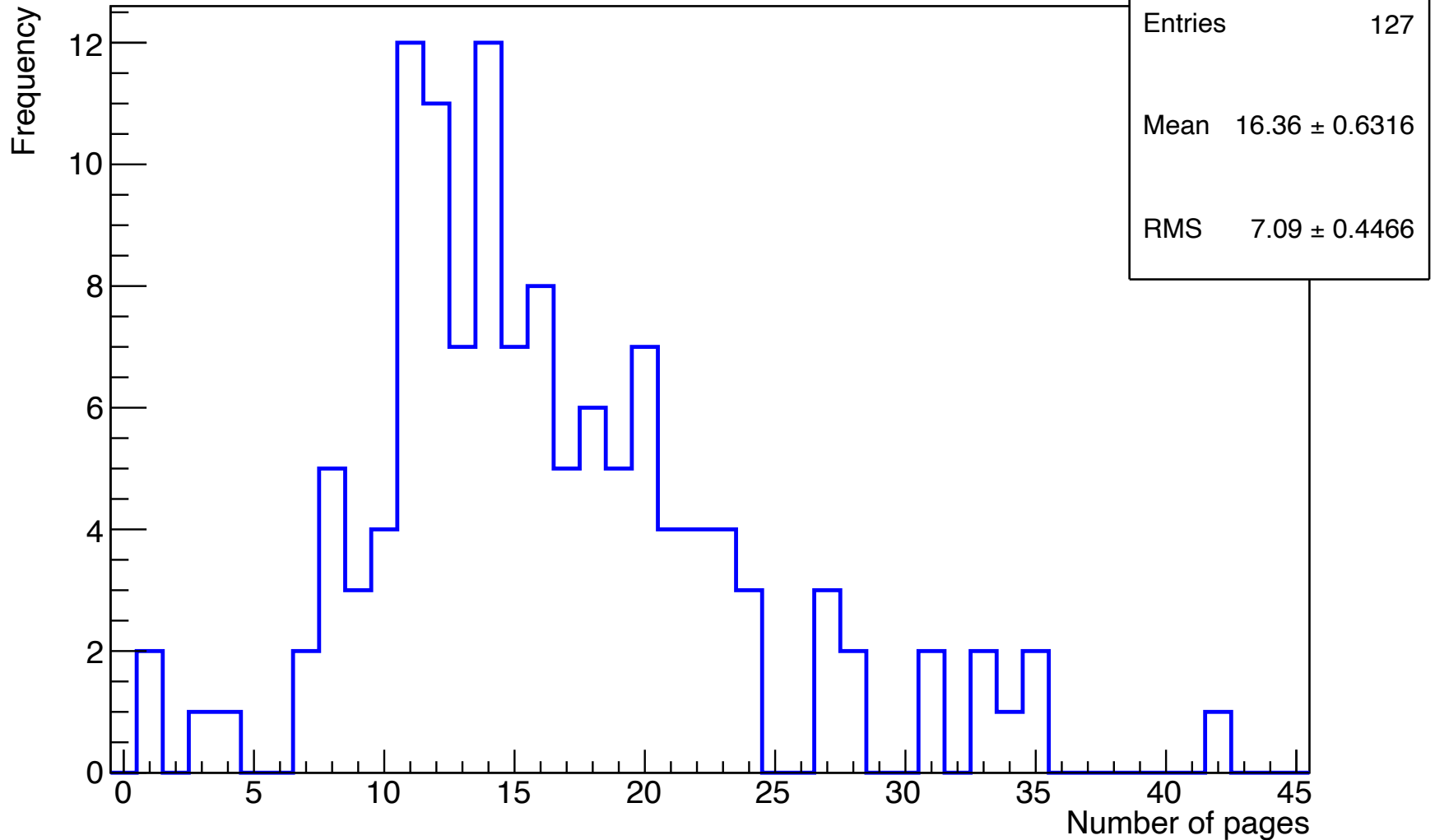
“Doing statistic problems during the entire night is like partying.
You can't stop yourself, but the next day you feel miserably...”
[End of a 42 page exam solution in 2018]



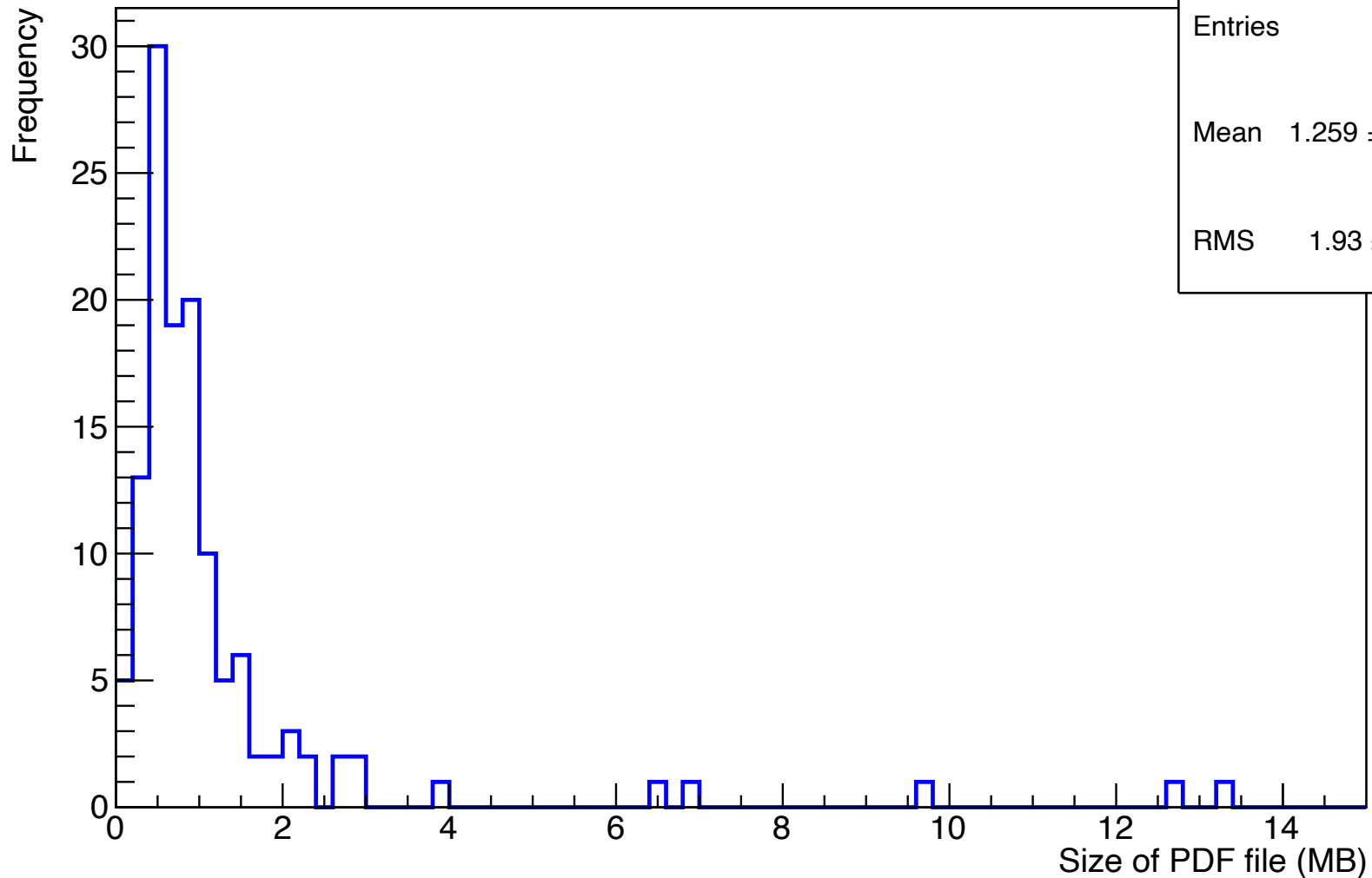
Some statistics

From last year!

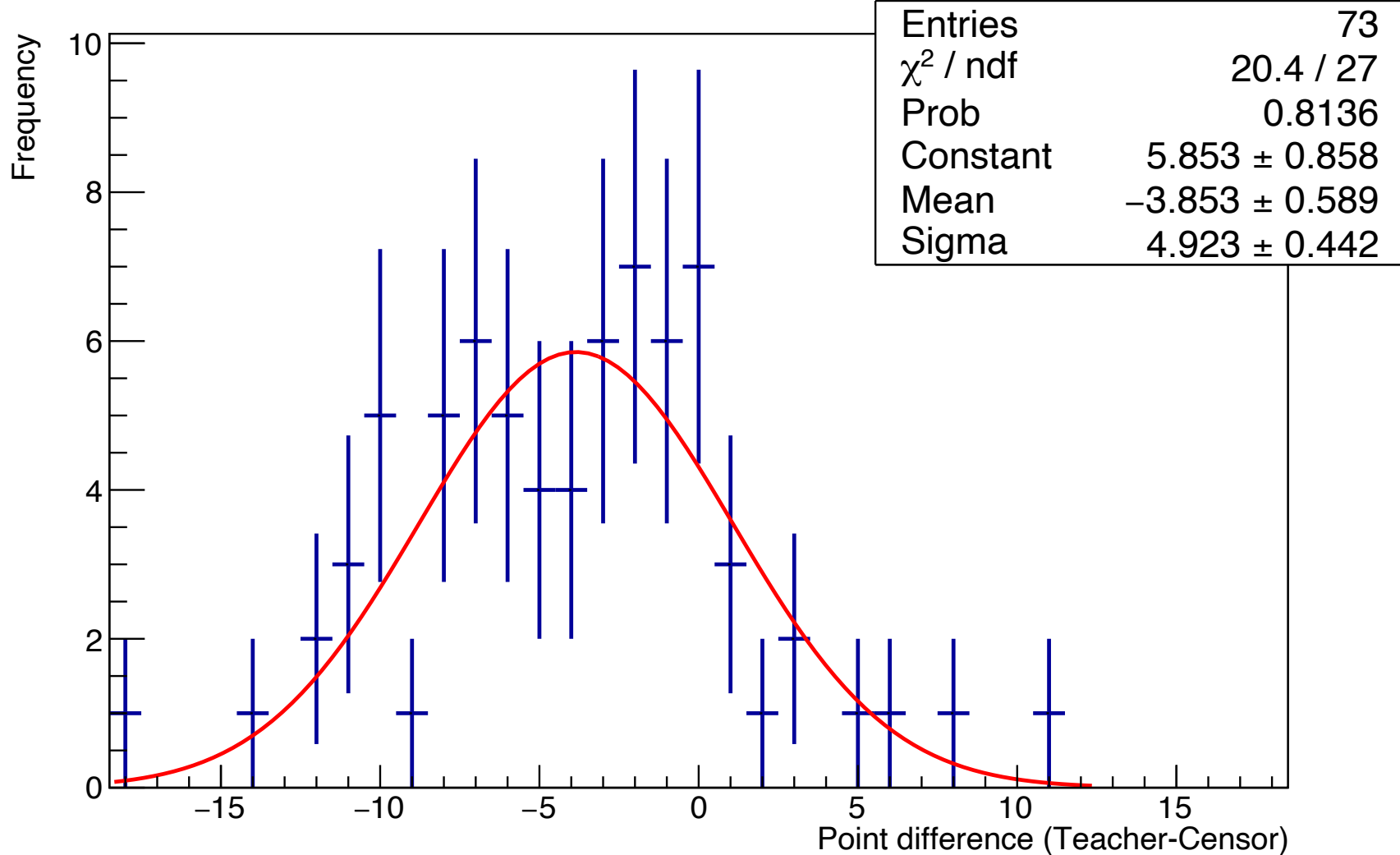
Number of pages in solution



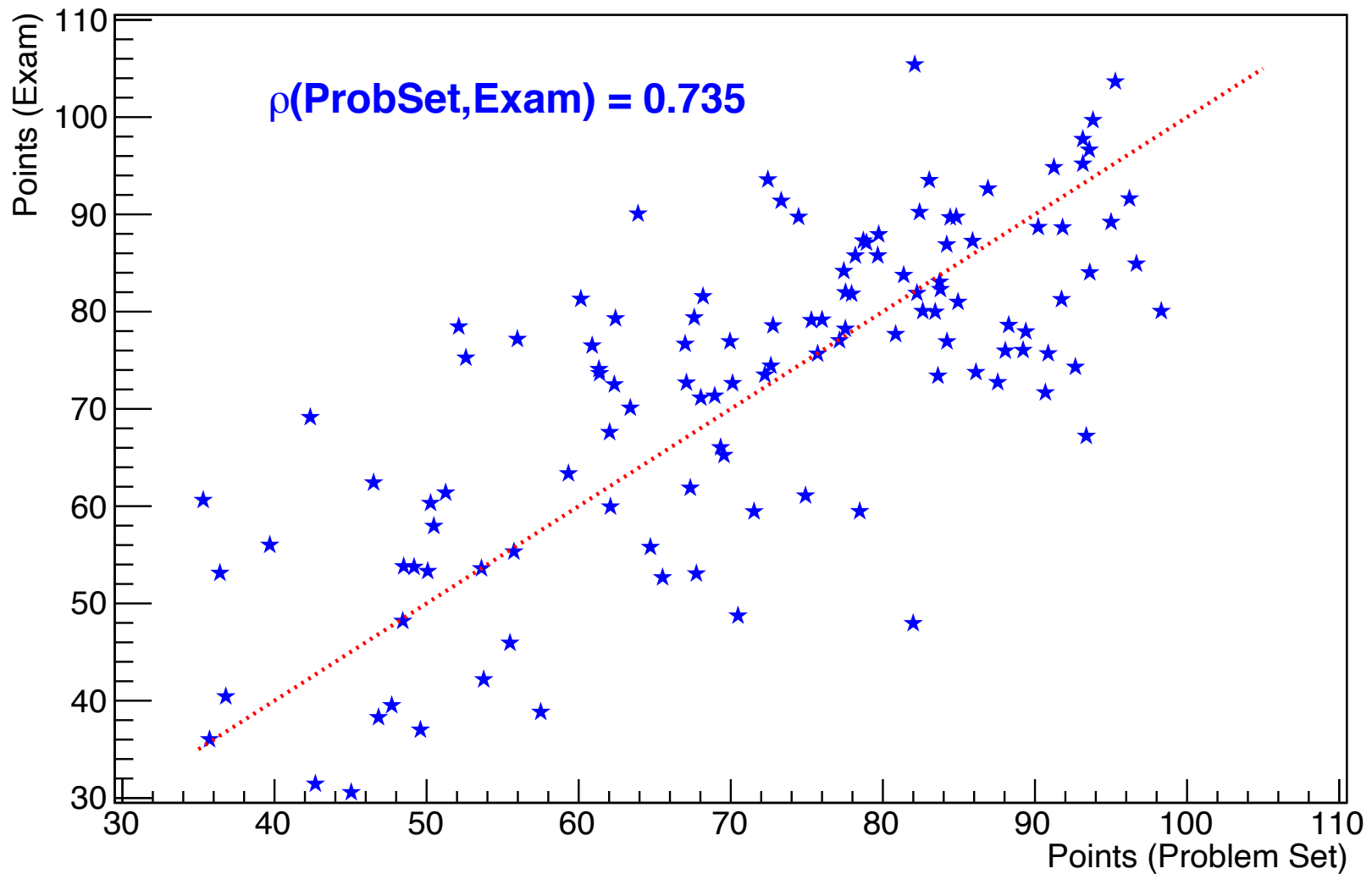
Size of PDF file



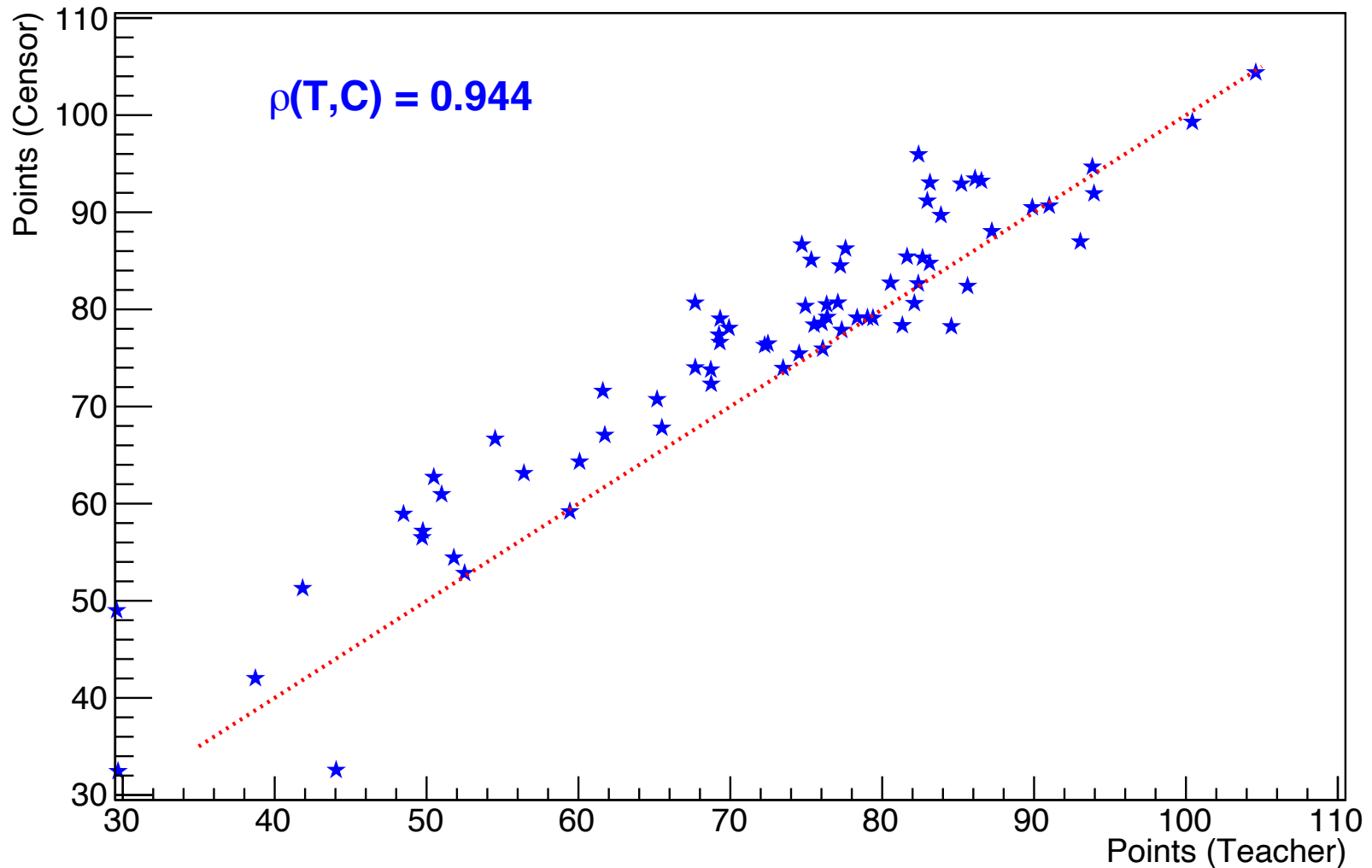
Teacher-Censor difference



ProblemSet-Exam correlation



Teacher-Censor correlation



Comment on code sharing!

To cross check, we run Moss (Measure Of Software Similarity) on your code, which is an automatic system for determining the similarity of programs (e.g. detecting plagiarism in programs).

Don't worry - nothing suspicious was found. Thank you!

Moss Results

Tue Jan 22 04:54:30 PST 2019

Options -l python -d -m 4

Moss Results

Tue Jan 22 05:11:04 PST 2019

Options -l python -d -m 1000000

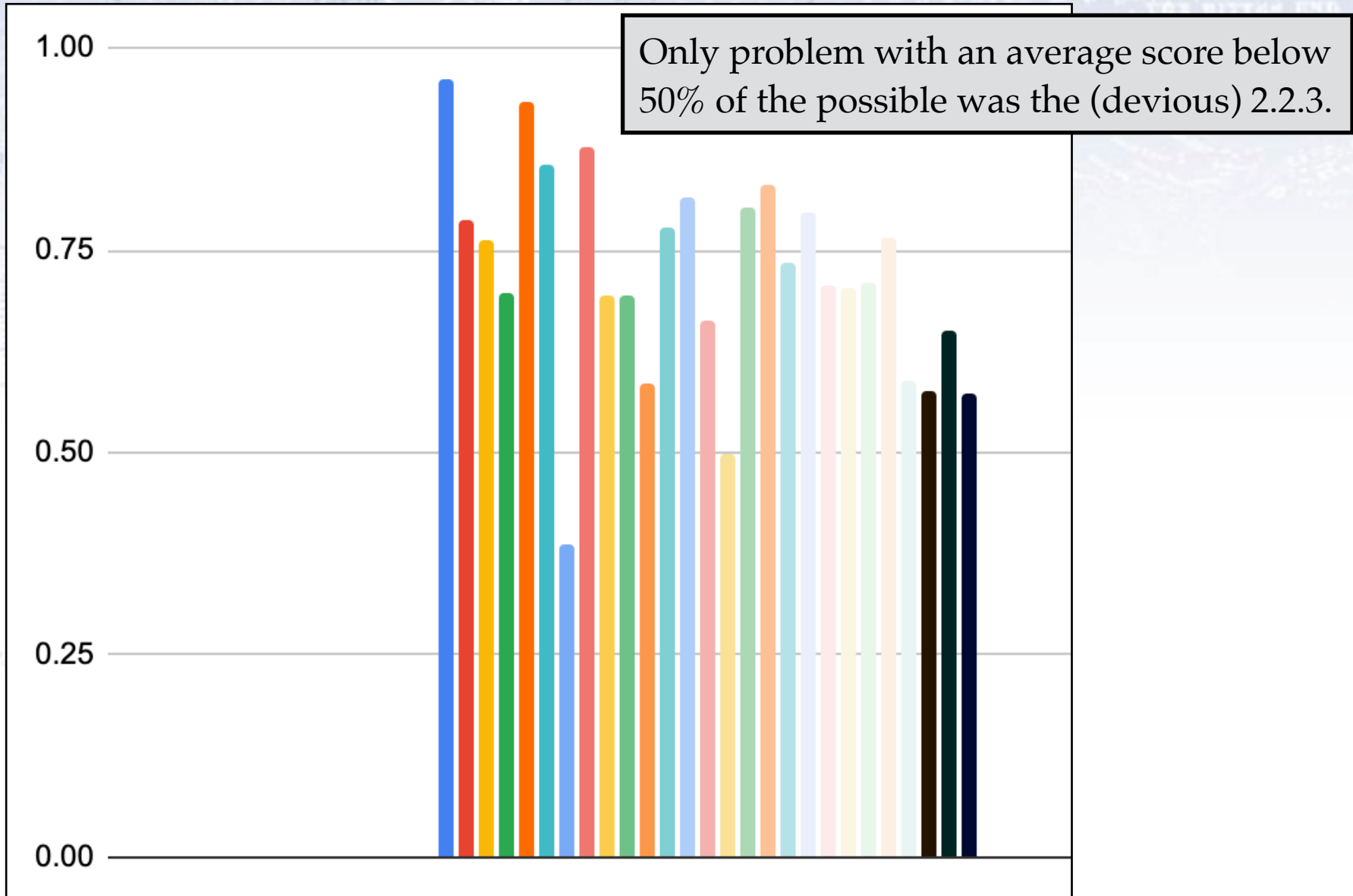
[[How to Read the Results](#) | [Tips](#) | [FAQ](#) | [Contact](#) | [Submission Scripts](#) | [Credits](#)]

File 1	File 2	Lines Matched
py/ [REDACTED] (85%)	py/ [REDACTED] 30/ (23%)	936
py/ [REDACTED] (36%)	py/ [REDACTED] (48%)	458
py/ [REDACTED] (31%)	py/ [REDACTED] (%)	424
py/ [REDACTED] (6%)	py/ [REDACTED] (6%)	38
py/ [REDACTED] / (7%)	py/ [REDACTED] / (7%)	54
py/ [REDACTED] / (5%)	py/ [REDACTED] / (5%)	69
py/ [REDACTED] (11%)	py/ [REDACTED] (12%)	62
py/ [REDACTED] (3%)	py/ [REDACTED] (2%)	51
py/ [REDACTED] (4%)	py/ [REDACTED] / (4%)	43
py/ [REDACTED] (3%)	py/ [REDACTED] (4%)	30
py/ [REDACTED] 3/ (2%)	py/ [REDACTED] / (2%)	55
py/ [REDACTED] / (2%)	py/ [REDACTED] / (2%)	77

A faded background image of a nautical chart. The chart features a compass rose with a 'V' indicating magnetic variation. Concentric lines represent magnetic isogons, with values such as 0, 30, 60, 90, 120, 150, 180, 210, 240, 270, and 300. The word 'MAGNETIC' is printed on the chart. In the upper right, the text 'THE BITTER END YACHT CLUB' is visible. The overall image is light blue and serves as a background for the text.

Your results....

Individual problem scores



Score distribution

The score distribution is very much as expected. The peak at 0 is from the 24 not following the course for credit. The average of the 140 students with a non-zero score is 72.0.

Be reminded, that all scores are relative. Final scale is yet to be set (by the censors!).

