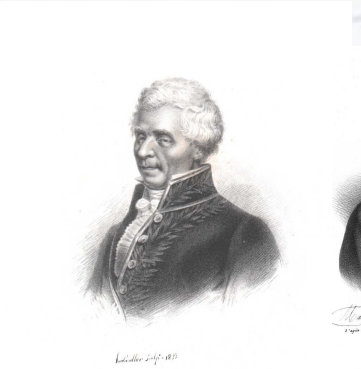
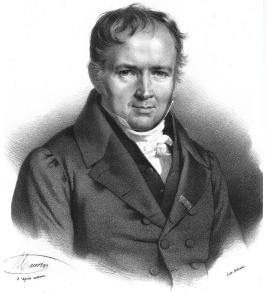
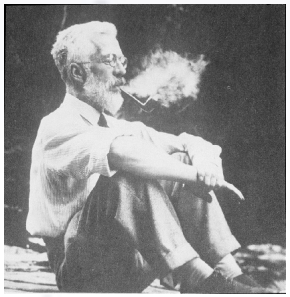
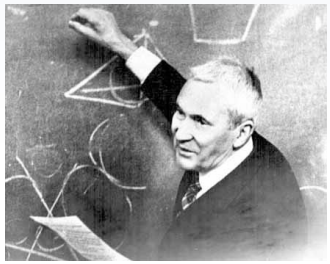
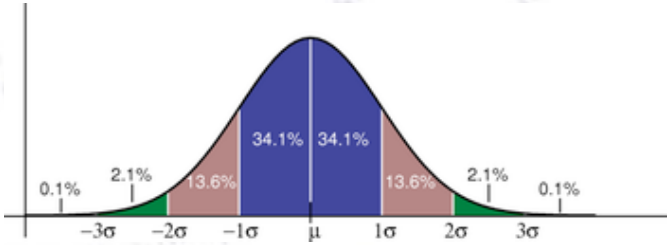
Applied Statistics

Problem Set Example Solutions



Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

The solutions

Problem 1.1

Hypergeometric formula (or simple logic) solves this problem…

1.1.1

When taking a marble and not putting it back, we are changing the probabilities for the next marble. However these changes are dependent of what exactly we marble we draw at the start. This setup follows the hypergeometric probability mass distribution:

$$p(k, M, n, N) = \binom{n}{k} \cdot \frac{\binom{M-n}{N-k}}{\binom{M}{N}} \tag{1}$$

Here k is the number of observed successes, M is the total amount of marbles, n is the number of success marbles in the bag and N is the number of marbles that we draw. This then takes care of all the different micro states that satisfy our conditions.

Now we take two marbles, and want at least one of them to be white, meaning that we have 3 success marbles. This gives us:

$$p(k \geq 1, 15, 3, 2) = p(1, 15, 3, 2) + p(2, 15, 3, 2) \tag{2}$$

$$p(k \geq 1, 15, 3, 2) = 0.3714 \tag{3}$$

This means that there is a probability of 37.14% chance of getting 1 or more white marble.

Problem 1.1

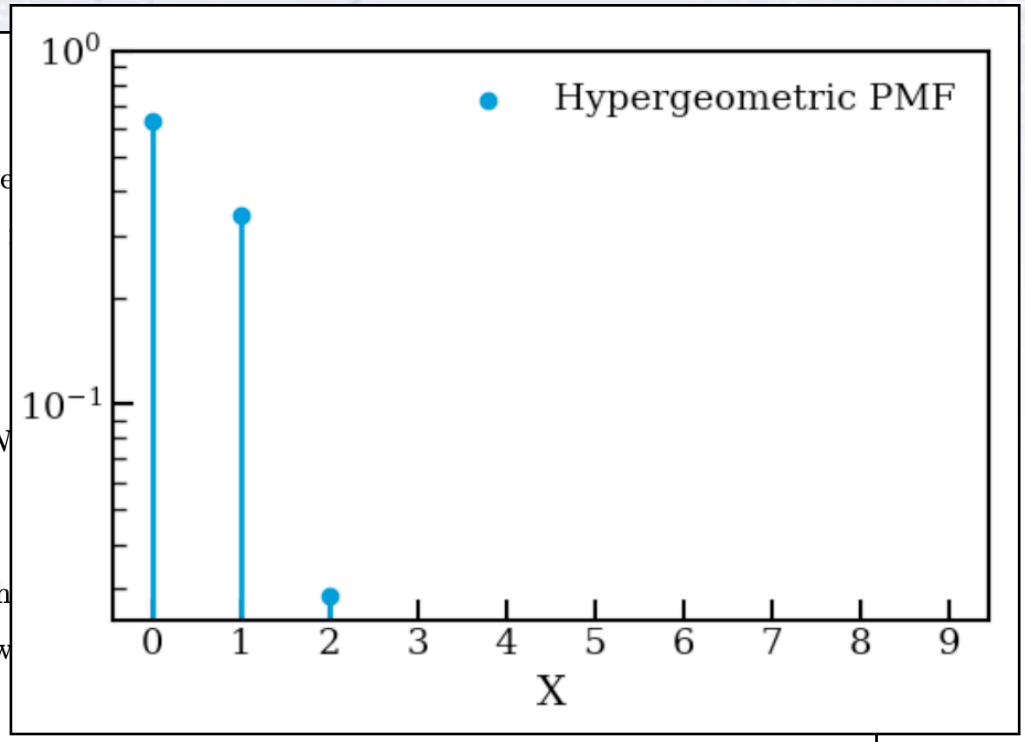
Hypergeometric formula (or simple logic) solves this problem…

1.1.1

When taking a marble and not putting it back, we
these changes are dependent of what exactly we
geometric probability mass distribution:

$$p(k, M, n, N$$

Here k is the number of observed successes, M is th
in the bag and N is the number of marbles that w
that satisfy our conditions.



Now we take two marbles, and want at least one of them to be white, meaning that we have 3 success marbles. This gives us:

$$p(k \geq 1, 15, 3, 2) = p(1, 15, 3, 2) + p(2, 15, 3, 2) \tag{2}$$

$$p(k \geq 1, 15, 3, 2) = 0.3714 \tag{3}$$

This means that there is a probability of 37.14% chance of getting 1 or more white marble.

Problem 1.1

Binomial solutions…

1.1.2

This probability follows the binomial distribution and we can white it up as:

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

For 18 marples out of 25 this gives:

$$P(18;\frac{7}{15},25) = \frac{7}{15}^{18}(1-\frac{7}{15})^{25-18}\frac{25!}{18!(25-18)!} = 0.0065$$

There is a 0.65% chance of drawing exactly 18 grey balls out of 25 tries.

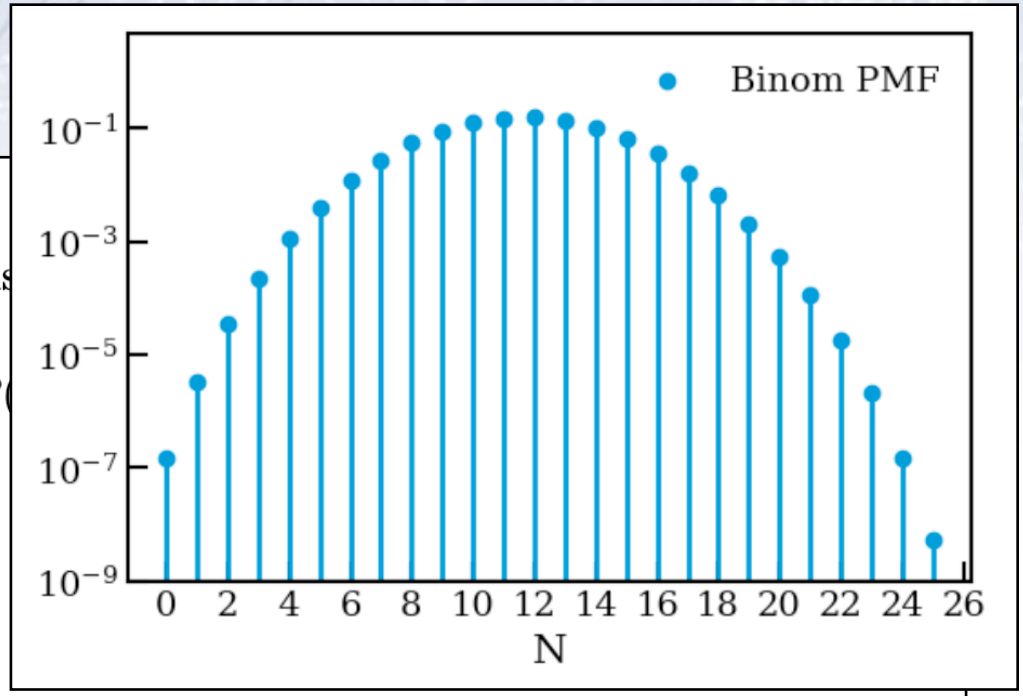
If it was instead 18 or more, the probability is now the sum:

$$P(\geq 18;\frac{7}{15},25) = \sum_{r=18}^{25}\frac{7}{15}^{r}(1-\frac{7}{15})^{25-r}\frac{25!}{r!(25-r)!} = 0.93\%$$

There is a 0.93% chance of drawing 18 or more grey balls out of 25 tries.

Problem 1.1

Binomial solutions…



1.1.2

This probability follows the binomial dis

$$P($$

For 18 marples out of 25 this gives:

$$P(18; \frac{7}{15}, 25)$$

There is a 0.65% chance of drawing exactly 18 grey balls out of 25 tries.

If it was instead 18 or more, the probability is now the sum:

$$P(\geq 18; \frac{7}{15}, 25) = \sum_{r=18}^{25} \frac{7}{15}^{r} (1 - \frac{7}{15})^{25-r} \frac{25!}{r!(25-r)!} = 0.93\%$$

There is a 0.93% chance of drawing 18 or more grey balls out of 25 tries.

Problem 1.1

The p-value of the test, p_{test}, is the "integral" under the two tails of the binomial distribution including only values that are further away (or equally far) from μ (because the absence of grey marbles would be equally suspicious), i.e:

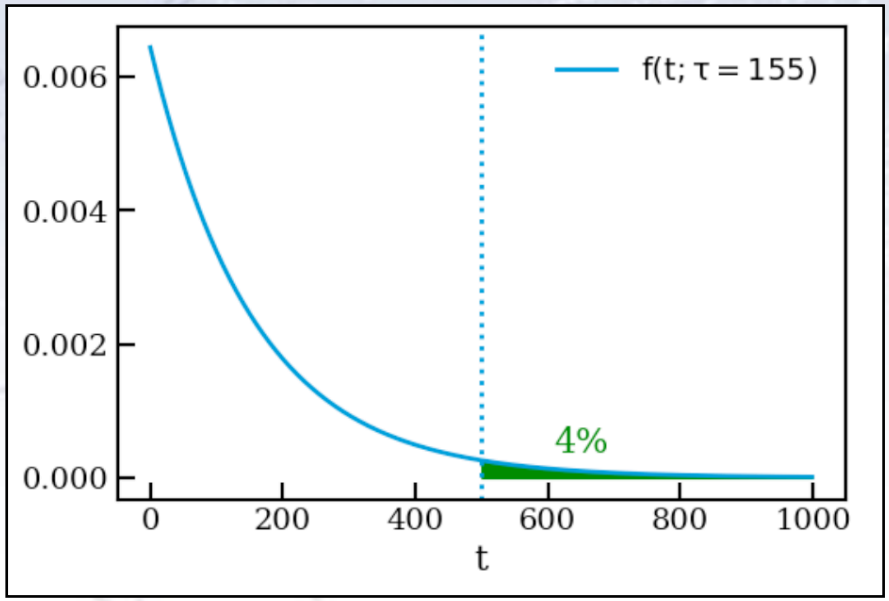
$$p_{test} = P(k \geq 18) + P(k \leq 5) = 1.47 \cdot 10^{-2} \tag{2}$$

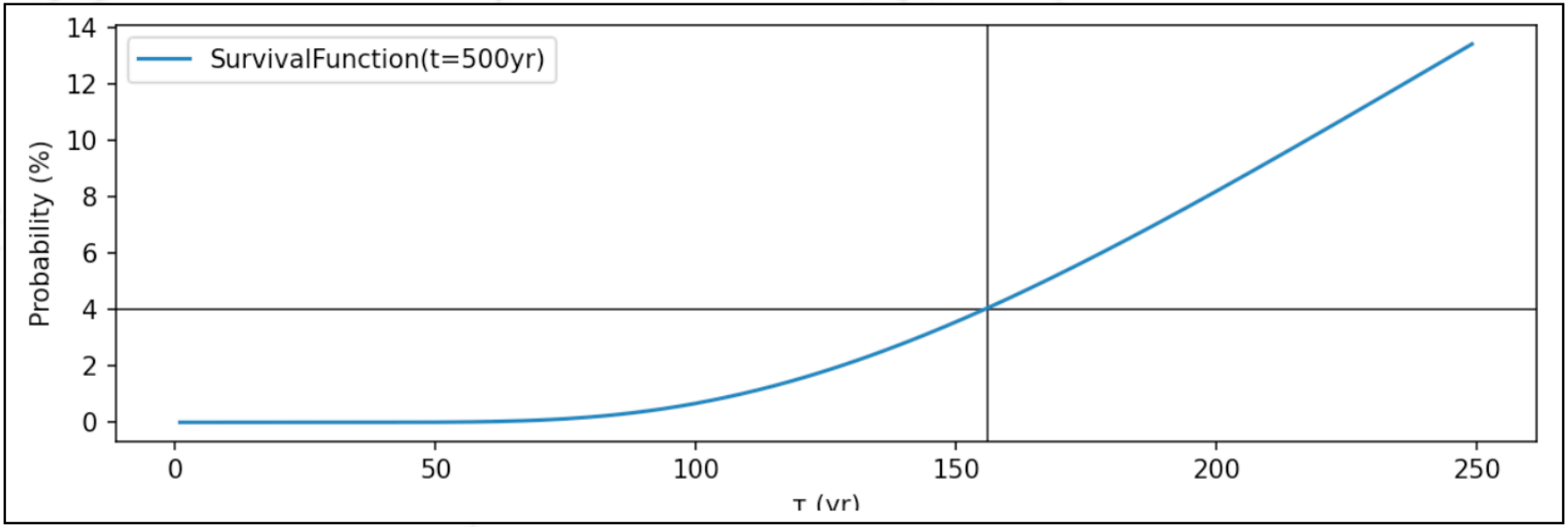
where the two probabilities have been calculated with scipy's built-in binomial cdf and survival function.
Thus I cannot reject that my friend is telling the truth on a 1% confidence level.
Note that if this was not my friend but instead a more suspicious person I might have chosen another more restrictive confidence level.

Problem 1.2

Almost all got this one right.

A few “inverted” it, and got 1/155!





Problem 1.3

This was a problem to illustrate why one considers “this or more extreme…”

1.3.1

I assume that the "signal" is some kind of background process such that they arrive to the telescope at a constant rate and that the telescope measures continuously in time. Under these assumptions, the number of signals measured in a period of time is a poisson process. Since the experiment has been going on for many hours, the expected number of signals during an hour is $\lambda = 241089/24 \approx 10045$ (the average number of signals per day divided by 24).
The probability of observing exactly 9487 signals in an hour, $P(k = 9487)$ can then be calculated with scipy's built-in poisson distribution. **This gives** $P(k = 9487) = 5.55 \cdot 10^{-10}$.

Since λ is large the gaussian is a very good approximation to the poisson. However, since 9487 lies pretty far out in the tails I find it safer (and just as easy) to keep using the poisson distribution.

1.3.2

The null hypothesis is that the number of signals in the hour follow a poisson distribution with $\lambda = 241089/24$. The alternative hypothesis is that this hour contains something else than the "background" process. Because I have no theoretical reasons for doing a one sided test, a two sided test is appropriate. This means that the *local* p-value, p_{local} is:

$$p_{local} = P(k \leq 9487) + P(k \geq 10603) = 2.667 \cdot 10^{-8} \tag{7}$$

where the probabilities are obtained from scipy's built in poisson cdf and survival function.
To obtain the *global* p-value, p_{global}, I need to correct with a trial factor because I assume that the telescope have been looking for "something" for the extent of 9 weeks. This means that they have been looking for $N = 1512$ hours. This is the trial factor. Because p_{local} is very small, I can obtain p_{global} by:

$$p_{global} = N \cdot p_{local} = 4.03 \cdot 10^{-5} \tag{8}$$

The p-value of the experiment observing a signal which is this far - or further - from the mean is therefore $p_{global} = 4.03 \cdot 10^{-5}$. The observation is definitely extraordinary assuming the null hypothesis. Whether the null hypothesis should be rejected depends on the confidence level which again depends on the nature of the discovery. If a "5σ" confidence level is chosen (critical p-value of $p \approx 3 \cdot 10^{-7}$). However, using all other standard confidence levels, the null hypothesis should be rejected.

Problem 1.4

This problem could also be solved using simulation (see “bad shooters” below).

Problem 1.4.1

Since the probability is 3% and not ~ 0 the number of hits will follow the binomial distribution. Furthermore the number of triales is known, N.

Problem 1.4.2

The probability of getting the first hit after 20 shots is equal to the probability of not hitting anything in the first 20 shots:

$$P_{no\,hit} = 1 - 3\% = 97\%$$
$$P = (P_{no\,hit})^{20} = \underline{\underline{54\%}}$$

Problem 1.4.3

The probability of needing more than 4000 shots to hit a 100 times is the summed probability of hitting between 0 or 99 times using 4000 shots:

$$\sum_{r=0}^{99} P\left(r, \frac{3}{100}, 4000\right) = \underline{\underline{2.61\%}} \tag{6}$$

Not being completely sure of this calculation a simulation is run. For each step of the simulation 4000 shots are made i.e. 4000 samples from a uniform distribution between 0 and 1 is drawn. If the "shot" is below 3% it counts as a hit. If the number of hits at 4000 is 99 or below, implies that the shooter would need more than 4000 shots to hit a 100 times. Running each step 400000 times and count the number of "bad shooters" gives:

$$Bad\ shooters = 10472$$
$$Probability = \frac{10472}{400000} = \underline{\underline{2.62\%}}$$

Now seeing that both methods gives approximately the same result i accept both to be correct.

Problem 2.1

Gaussians should not be fits, but the values obtained by error propagation formula.

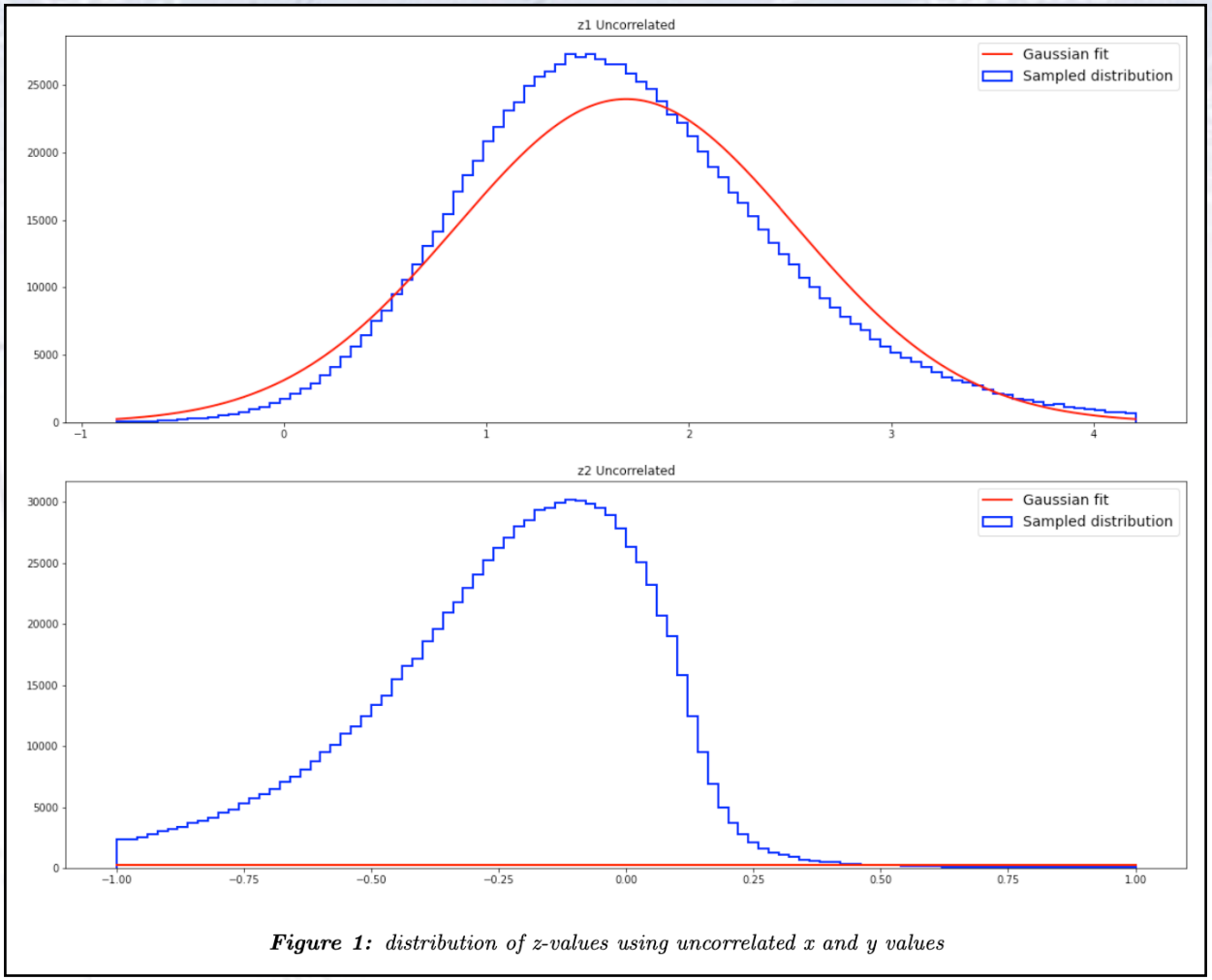


Figure 1: *distribution of z-values using uncorrelated x and y values*

To sum up it seems that the uncertainty on x widens the central part of the distribution the most while y produces extremely long tails

Problem 2.2

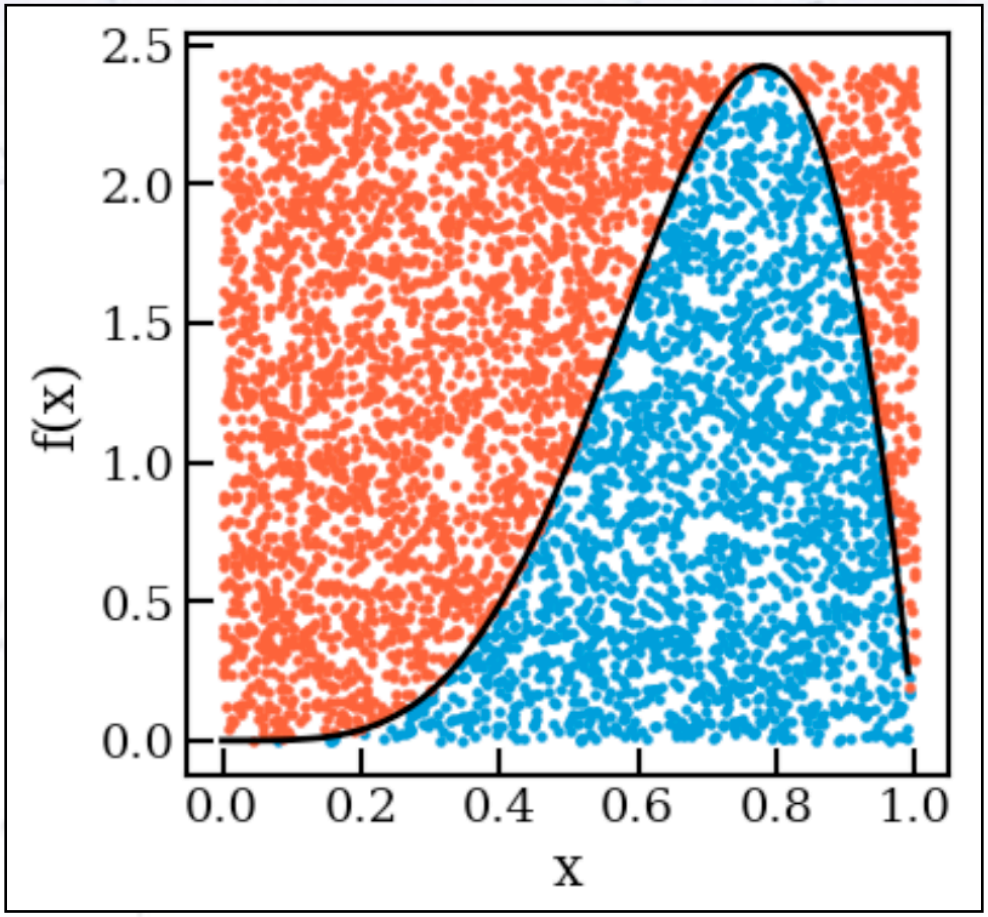
This is actually original data from Gosset ("student"), and requires both the "N-1" when calculating Std. and a t-test when comparing. We did not require the last of the two, but gave bonus points for doing so.

Thanks to Mathias for digging this paper/data out - it is a very beautiful paper, which might go on a reading list next year.

The null hypothesis is that the distributions from which the placebo and drug group are drawn have the same means. The alternative hypothesis is that the mean of the drug group is larger than the mean of the placebo group. Because I assume data to be distributed according to a gaussian, the sample size is small and I estimate the standard deviation from the sample, a two sample t-test is appropriate for testing if there is any difference between the drug group and the placebo group. I compute the t-score according to equation (8.13) in Barlow. This yields $t = 0.73$. This value is distributed according to a t-distribution with $5 + 5 - 2 = 8$ degrees of freedom. Because the alternative hypothesis is that the drug group sleeps longer, a one-sided test is appropriate. To perform the integral from 0.73 to ∞ of the t-distribution, I use scipy's built in t-distribution survival function. This gives the p-value $p = 0.24$.
Therefore the probability of obtaining a t-score this large or larger is 24% using a one sided test assuming that the placebo and drug group have equal means. Therefore the null hypothesis cannot be rejected. The effect of the drug on sleeping times is not significant.

Problem 3.1

This is an obvious case for the accept-reject method. The transformation method fails, as the inversion can only be done numerically, and in any case, it does not save one much in speed, as the A-R is fairly efficient here.

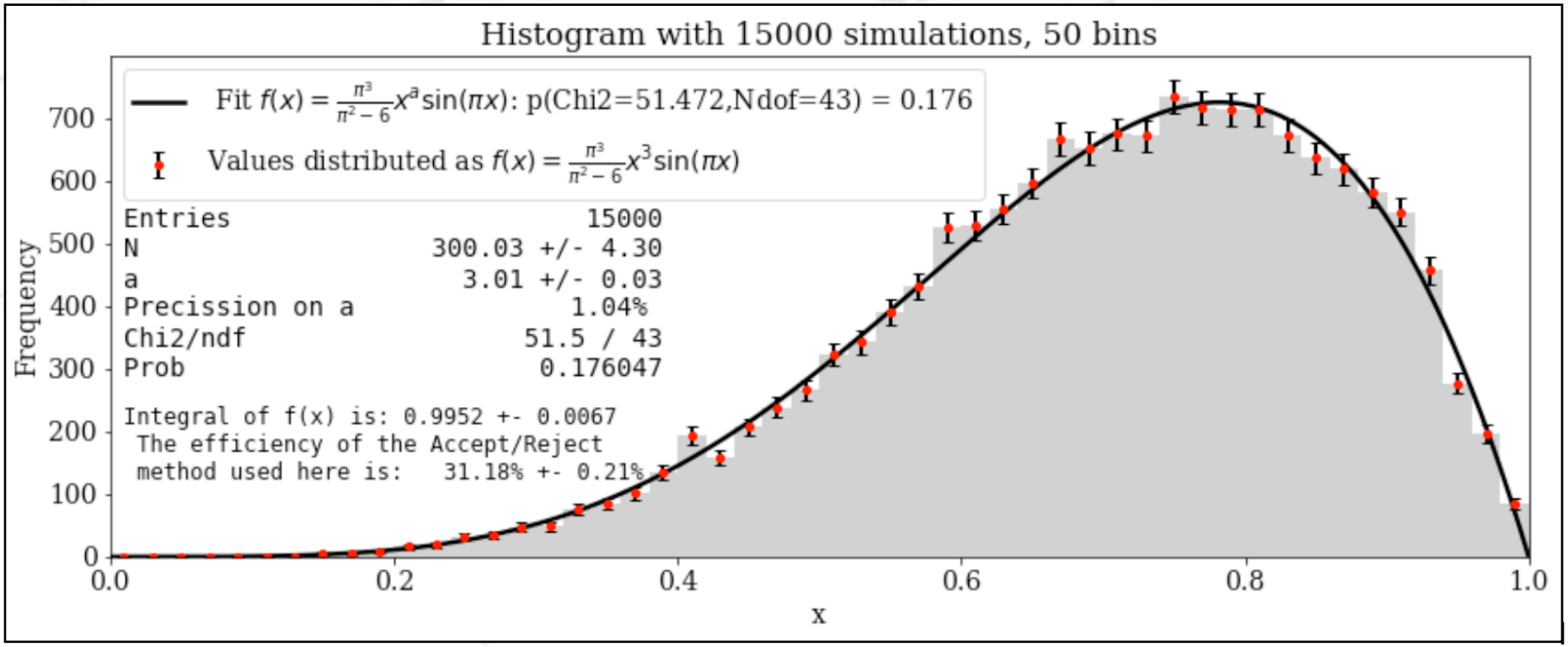


Problem 3.1

The fitting required two considerations:

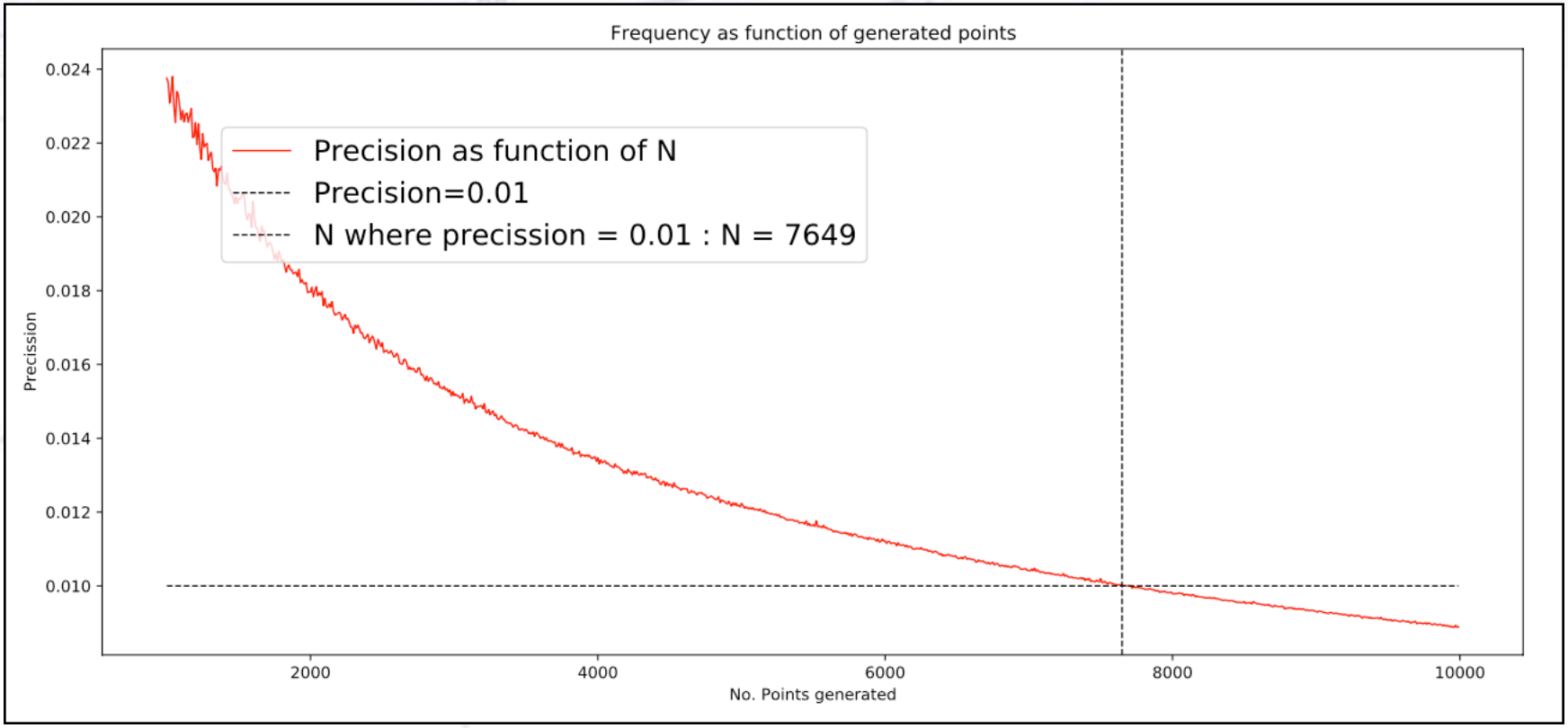
- Is the statistics high enough in the lower bins to fit with a ChiSquare?
- Is the binning fine enough to not distort the sharp right edge?

Some did both (bonus points)…



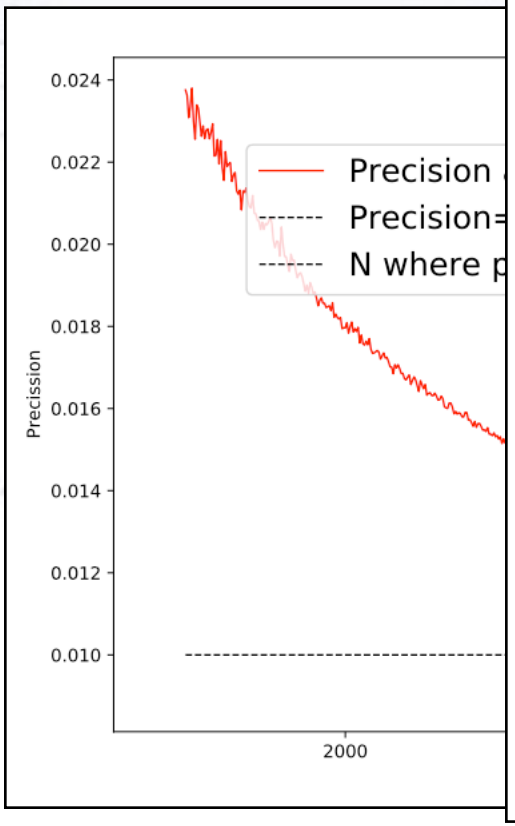
Problem 3.1

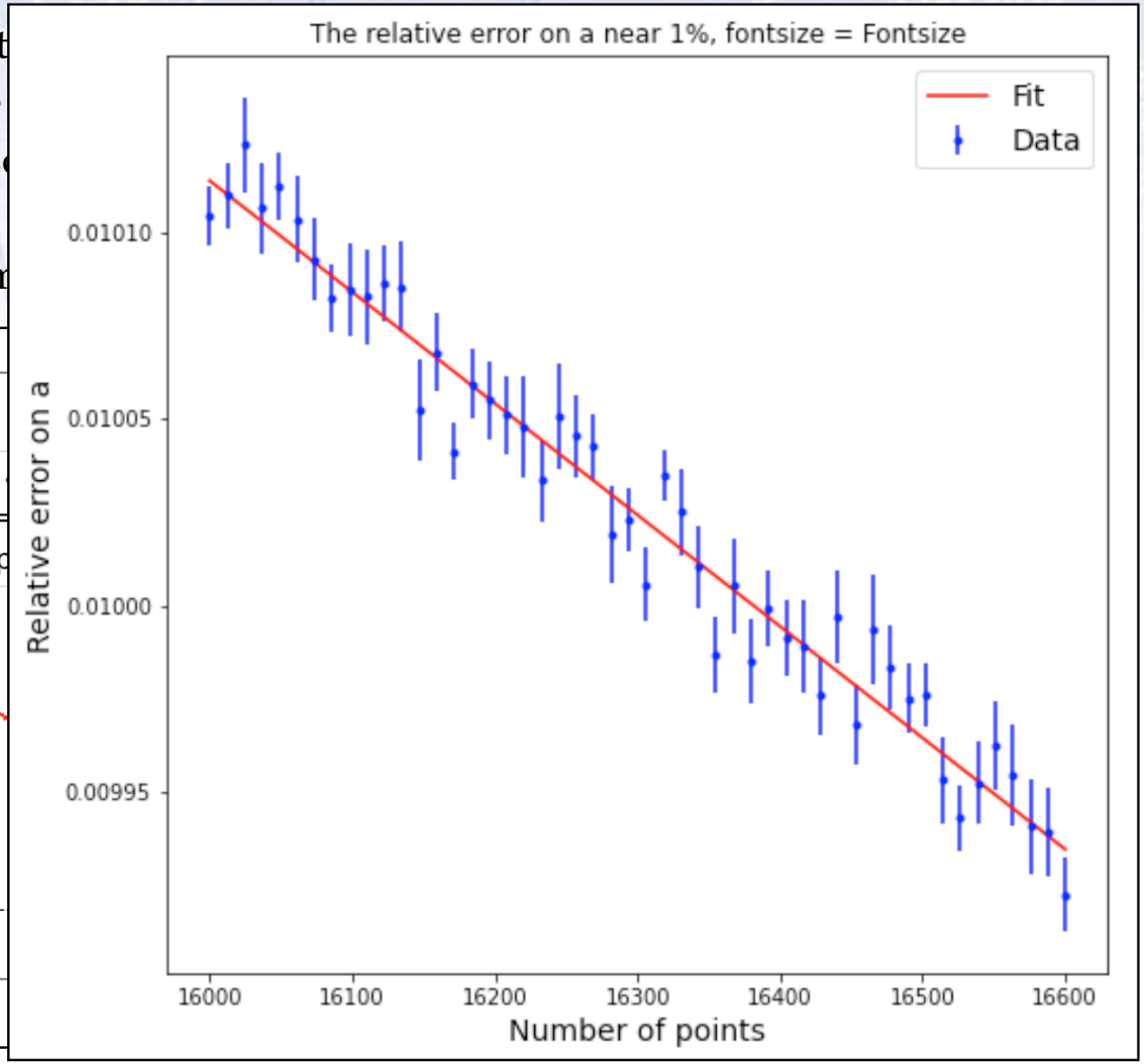
In order to investigate how many events would be needed to measure a with 1% precision, people tried to repeat the process with different statistics, which nicely shows the 1/sqrt(N) behaviour of uncertainties. A few even gave a range of possible values.
Curiously, there seem to be convergence towards two ranges.



Problem 3.1

In order to investigat
1% precision, people
nicely shows the 1/s
of possible values.
Curiously, there seem





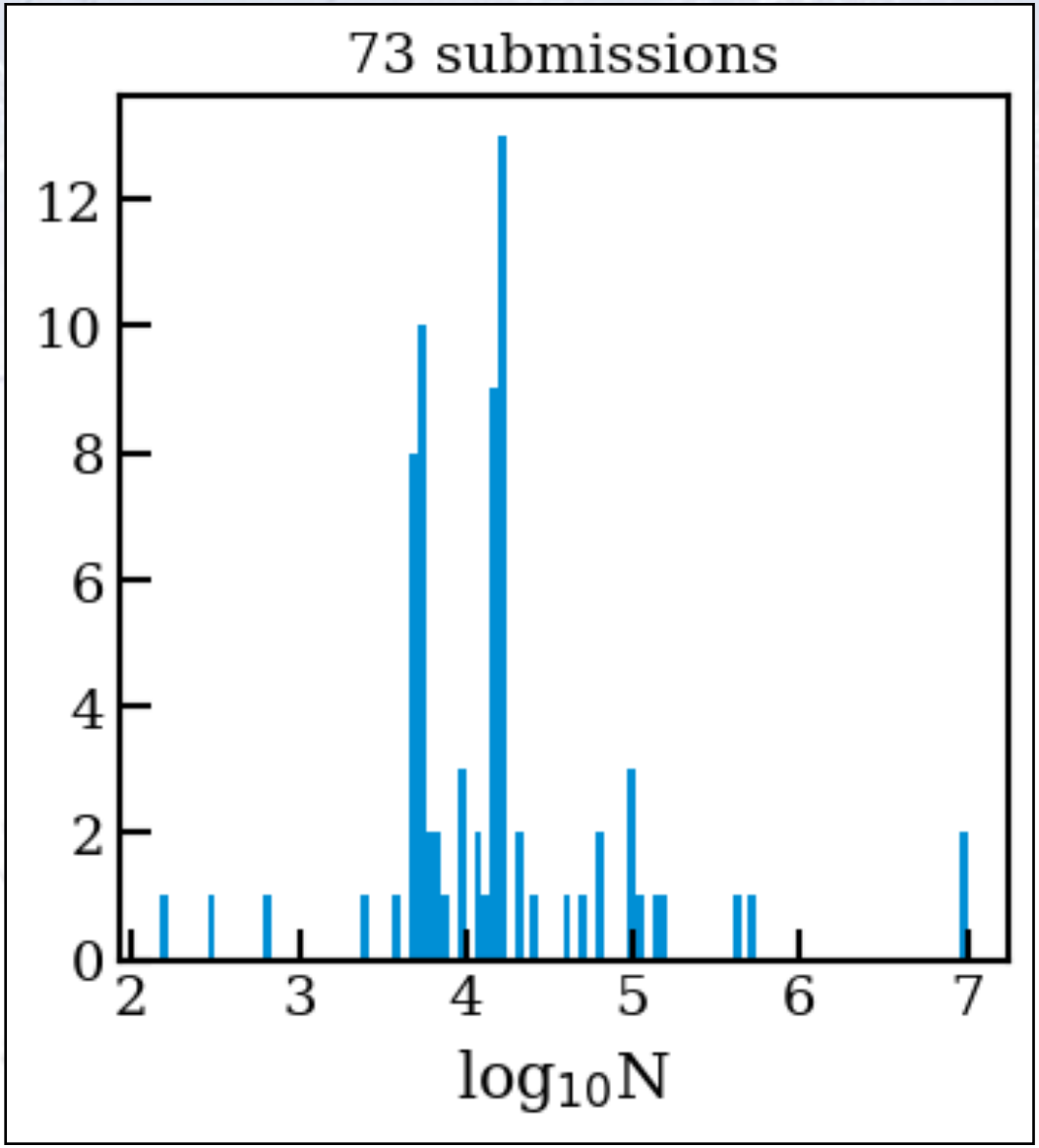
Problem 3.1

We counted up the different estimates of N, and two "camps" are visible.

We of course allowed for many values, especially if the arguments/principles were in place.

However, below 1000 and above 50000 is probably outside a good range.

Imagine, that your measurements had some uncertainty in x…?



Problem 4.1

This problem was meant to illustrate the power of paired tests!

- When not pairing, the distribution is wide, due to strong and weak persons.
- When pairing, the variation in strength cancels out…

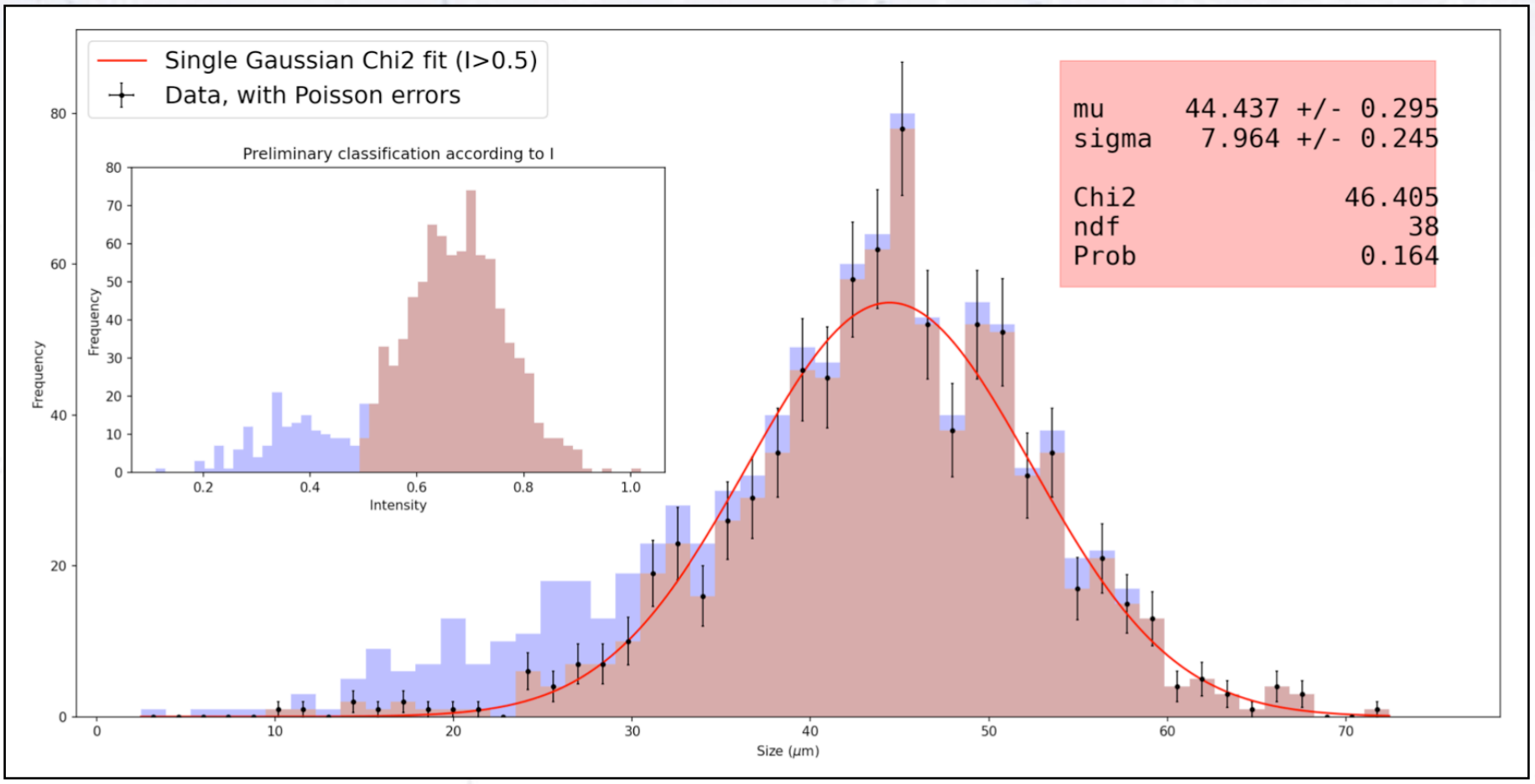
A detail that several referred to: The distribution of G is not Gaussian.
But what is important is, that the mean is (due to CLT), so it is OK to compare means with a z-test.

A simple alternative solution is to do a KS test! This would detect a shift, but also a difference in distributions, which is not exactly what we are looking for!

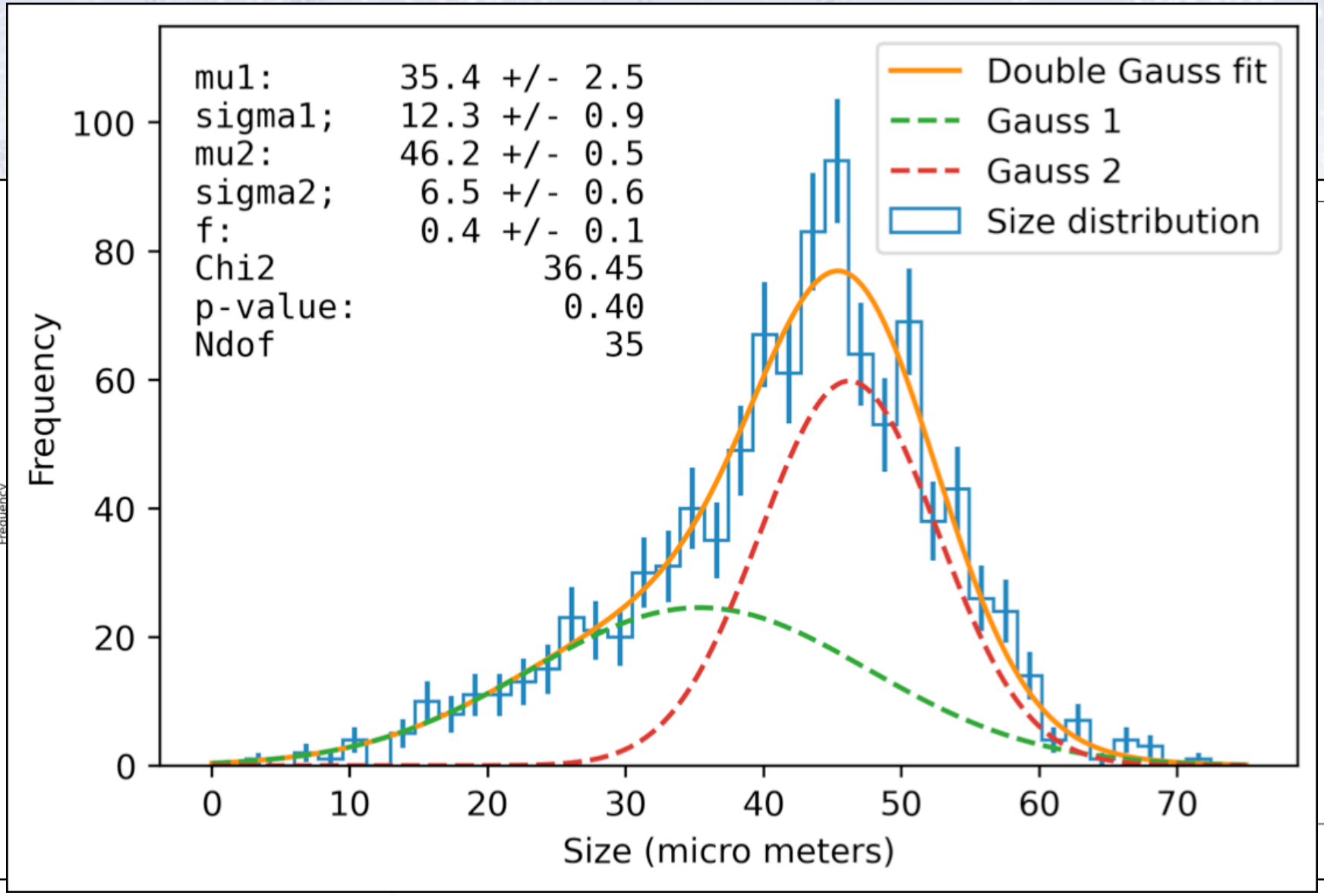
Notice that "dominant hand" is defined by which you e.g. write with.

Problem 4.2

This problem had no labels, and could very well be a real world problem. There were many great plots of this problem - thank you for those. They were very much appreciated.



Problem 4.2

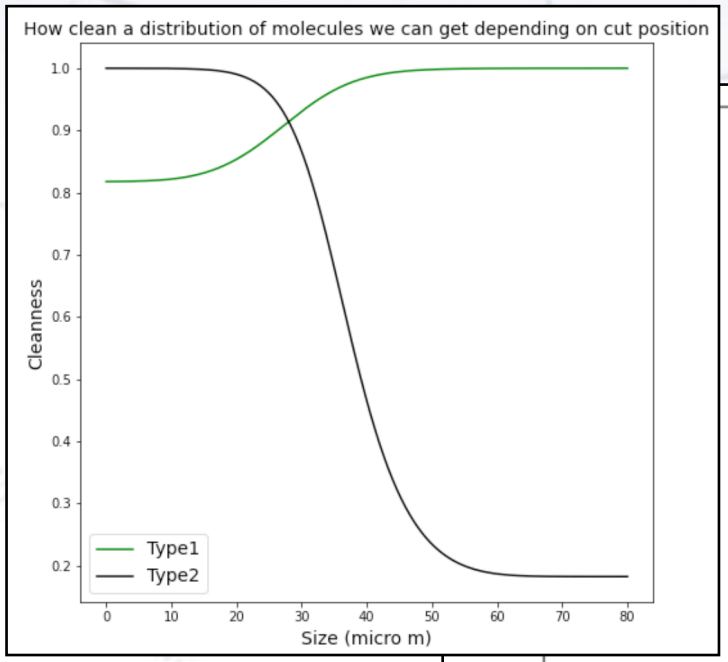

mu1: 35.4 +/- 2.5
sigma1; 12.3 +/- 0.9
mu2: 46.2 +/- 0.5
sigma2; 6.5 +/- 0.6
f: 0.4 +/- 0.1
Chi2 36.45
p-value: 0.40
Ndof 35
Double Gauss fit
Gauss 1
Gauss 2
Size distribution
Frequency
Size (micro meters)
0
20
40
60
80
100
0
10
20
30
40
50
60
70

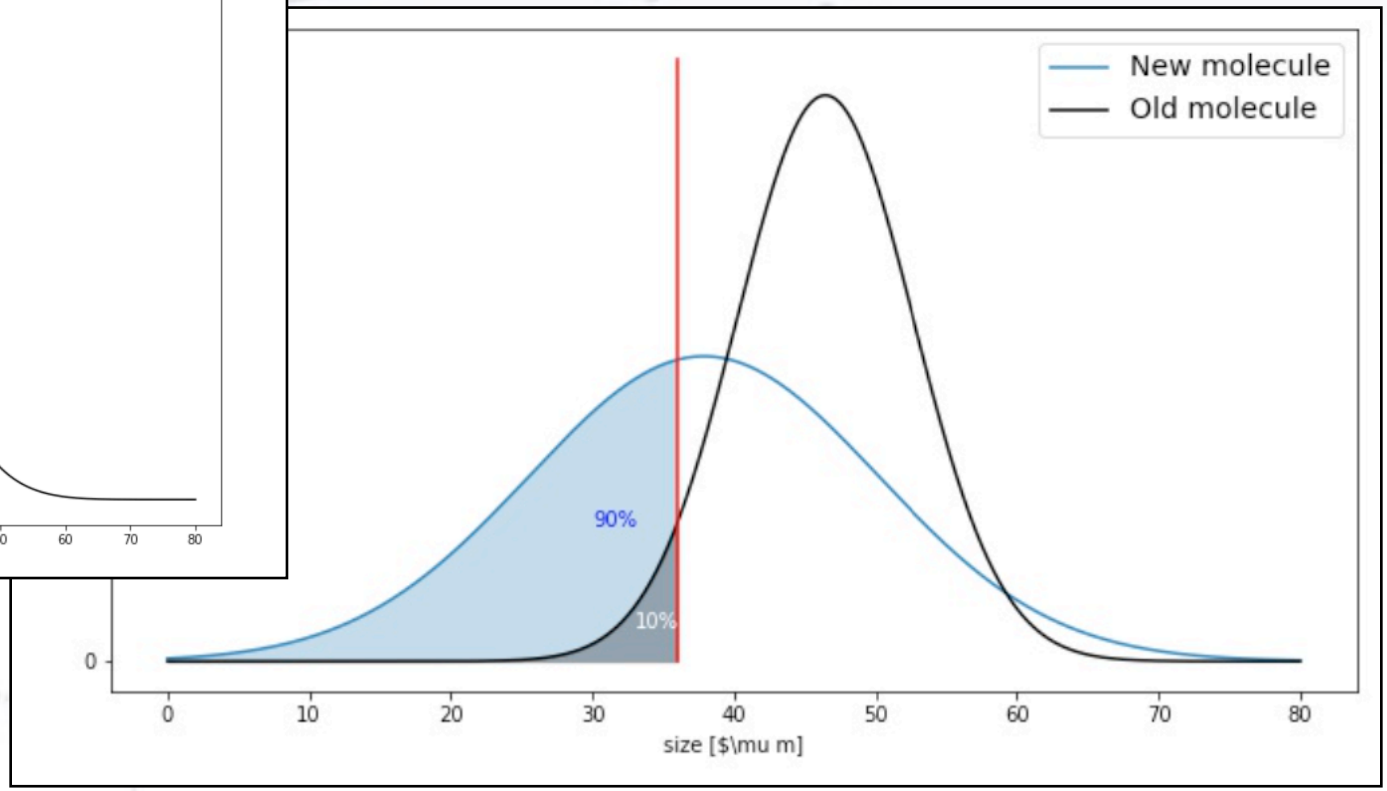
Problem 4.2

The result of the double Gaussian fit depends on several things:

- Functional form (NG+NG or N(fG+(1-f)G))
- Fit type (ChiSquare vs. LLH, later preferred due to low statistics)
- Initial values (of course!)

…and we didn't care what molecule you chose!





Problem 4.2

Two double Gaussian fits of size and intensity actually gives you the parameters for a Fisher discriminant. Alternatively, one can project using a PCA or "by eye".

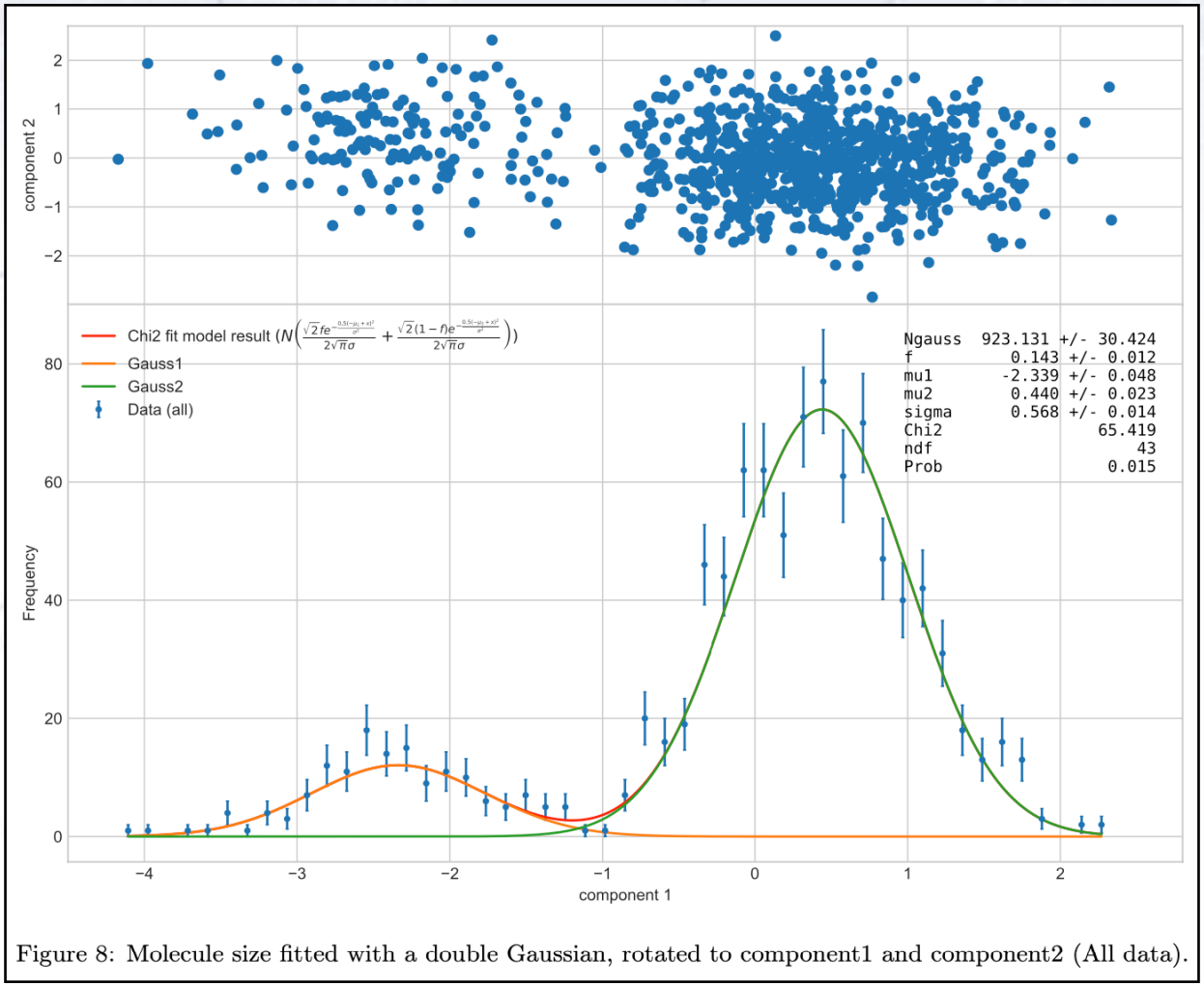


Figure 8: Molecule size fitted with a double Gaussian, rotated to component1 and component2 (All data).

Problem 5.1

The 3rd degree polynomial fit and WW runs test worked nicely for almost all. Several also commented on the fact, that the residuals did not have a Gaussian distribution - smart (admittedly, I didn't think of this).

> It is unclear whether the oscillation should be multiplied or added to the fit, since the words 'term' and 'multiplicative' seem conflicting. We have tried both, but only adding the oscillation works. This does

Problem 5.1

The 3rd degree polynomial fit and WW runs test worked nicely for almost all. Several also commented on the fact, that the residuals did not have a Gaussian distribution - smart (admittedly, I didn't think of this).

> It is unclear whether the oscillation should be multiplied or added to the fit, since the words 'term' and 'multiplicative' seem conflicting. We have tried both, but only adding the oscillation works. This does

Most added a term (corresponding to a fit of the residuals):

$$f(t) = at^3 + bt^2 + ct + d + e \cdot sin(f \cdot t + g).$$

Some multiplied an oscillation function on (as the data was generated):

$$f(x) = \left(ax^3 + bx^2 + cx + d\right)\left[1 + A\sin\left[2\pi x + \delta\right)\right].$$

$$g(x) = (ax^3 + bx^2 + cx + d) \cdot (1 + g\sin(ex + f)).$$

Problem 5.1

The 3rd degree polynomial fit and WW runs test worked nicely for almost all. Several also commented on the fact, that the residuals did not have a Gaussian distribution - smart (admittedly, I didn't think of this).

It is unclear whether the oscillation should be multiplied or added to the fit, since the words 'term' and 'multiplicative' seem conflicting. We have tried both, but only adding the oscillation works. This does

Most added a term (corresponding to a fit of the residuals):

$$f(t) = at^3 + bt^2 + ct + d + e \cdot sin(f \cdot t + g).$$

Some multiplied an oscillation function on (as the data was generated):

$$f(x) = \left(ax^3 + bx^2 + cx + d\right)\left[1 + A\sin\left[2\pi x + \delta\right)\right].$$

$$g(x) = (ax^3 + bx^2 + cx + d) \cdot (1 + g\sin(ex + f)).$$

We see the fit gives f=6.301 which corresponds to a frequency of 24 hours in units of the data, so that is quite satisfactory. We now get a p-value=1, which suggests overfitting. In fact a chi2-value 4 times as large, would still give a p-value above 0.99.

Problem 5.1

The 3rd degree polynomial fit and WW runs test worked nicely for almost all. Several also commented on the fact, that the residuals did not have a Gaussian distribution - smart (admittedly, I didn't think of this).

It is unclear whether the oscillation should be multiplied or added to the fit, since the words 'term' and 'multiplicative' seem conflicting. We have tried both, but only adding the oscillation works. This does

Most added a term (corresponding to a fit of the residuals):

$$f(t) = at^3 + bt^2 + ct + d + e \cdot sin(f \cdot t + g).$$

Some multiplied an oscillation function on (as the data was generated):

$$f(x) = \left(ax^3 + bx^2 + cx + d\right)\left[1 + A\sin\left[2\pi x + \delta\right)\right].$$

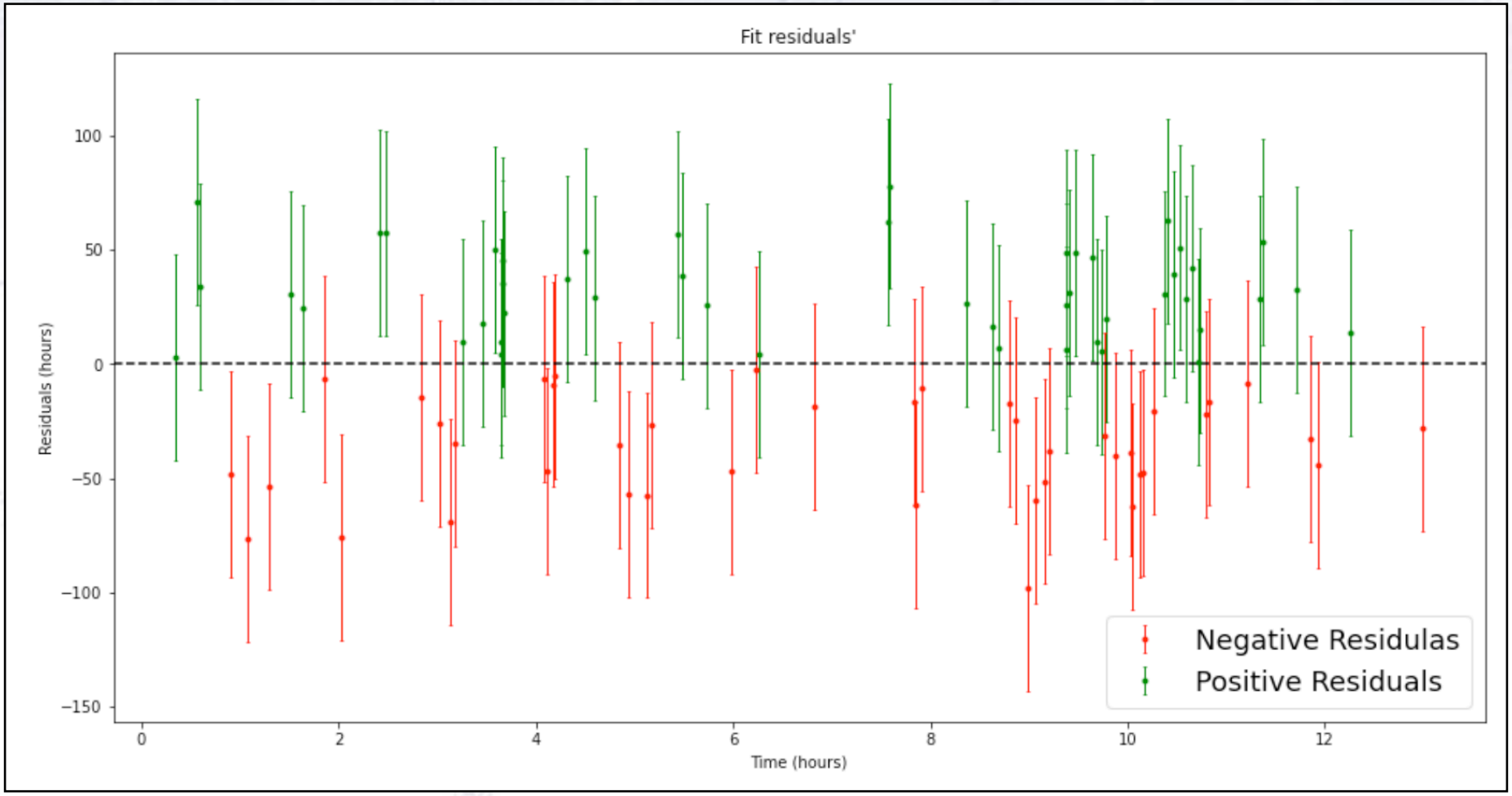
$$g(x) = (ax^3 + bx^2 + cx + d) \cdot (1 + g\sin(ex + f)).$$

We see the fit gives f=6.301 which corresponds to a frequency of 24 hours in units of the data, so that is quite satisfactory. We now get a p-value=1, which suggests overfitting. In fact a chi2-value 4 times as large, would still give a p-value above 0.99.

It is good to keep a cross check, if possible!

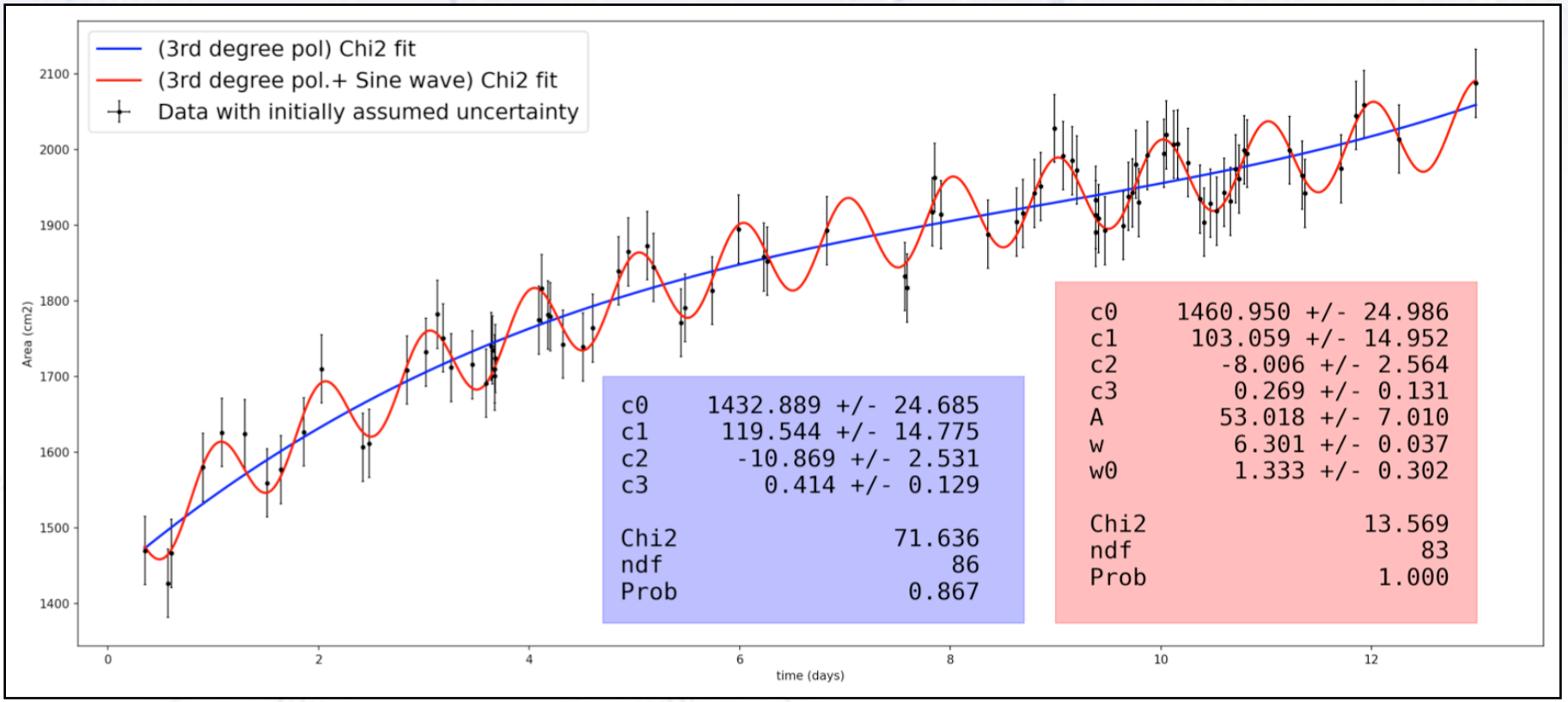
Problem 5.1

Very nice plot of residuals…



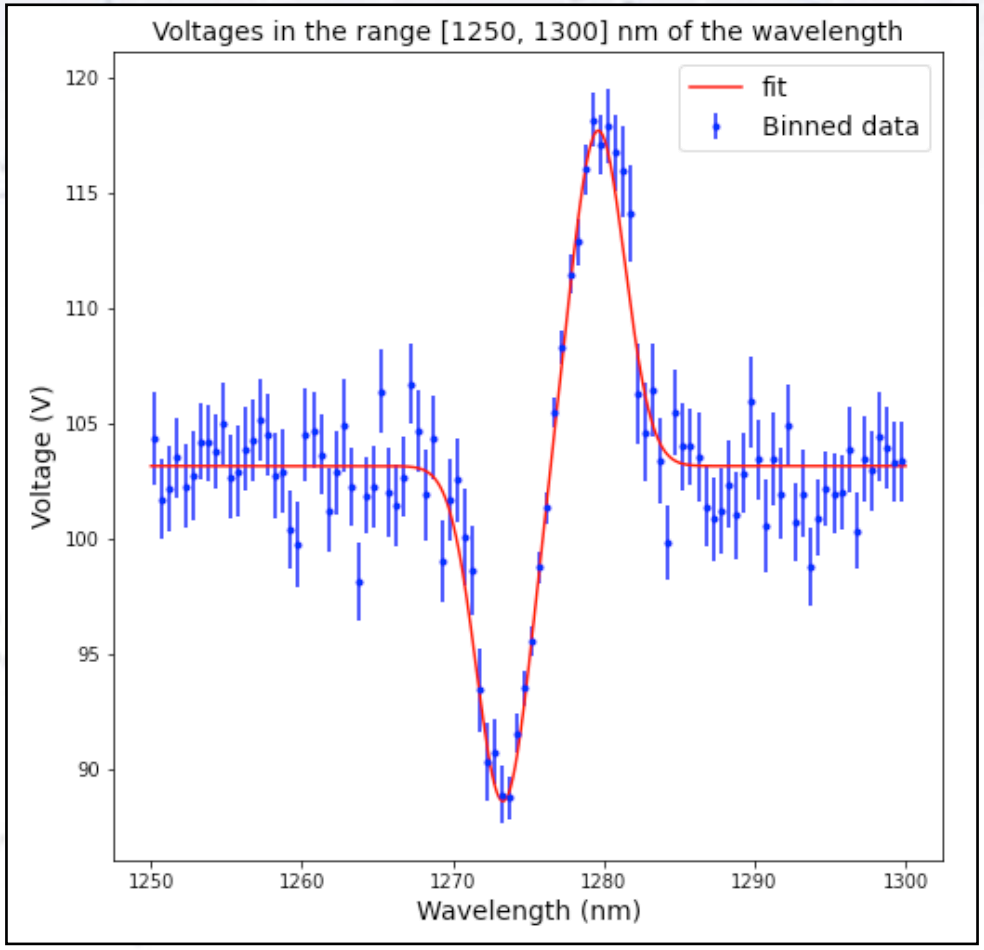
Problem 5.1

Very nice plot of the two fits… impressive to have the time and surplus for this…



Problem 5.2

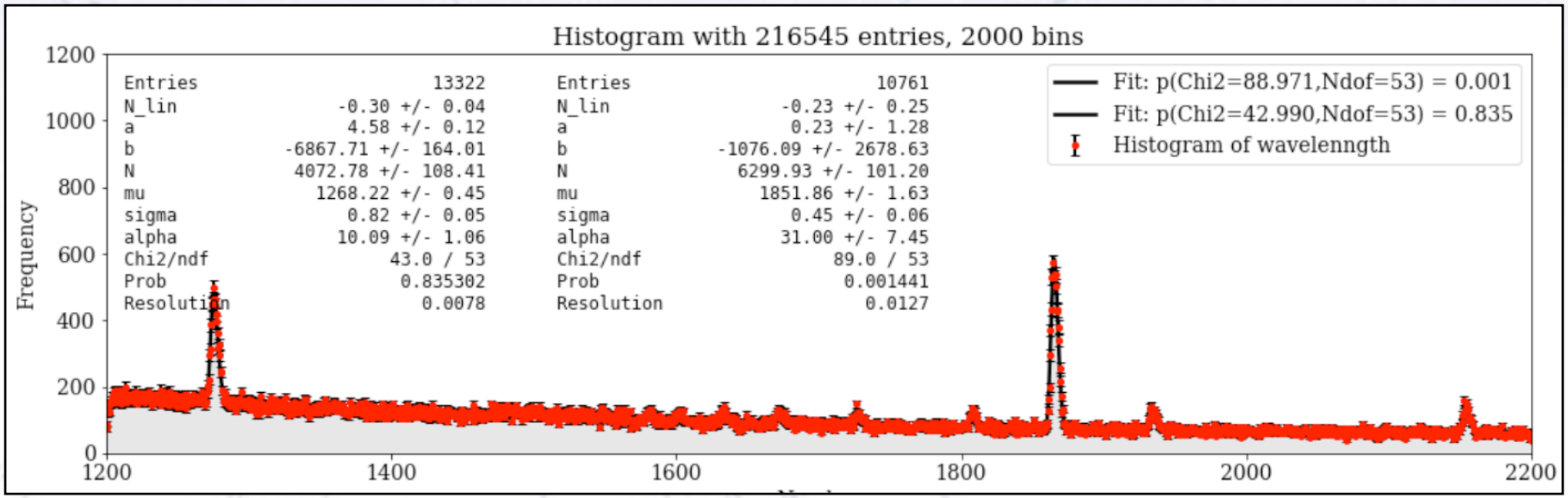
One student started wrongly, and got this plot. Impressively, the student carried on, and did 4/6 questions, before giving into the fact, that something was not as expected! Admirable… (and giving some points)



Problem 5.2

Plotting the range (with many bins) gives a nice overview of what is to come.

Several fitted the whole range and two peaks, but that is not necessary.



There are actually no Trial Factors here, as the positions of the peaks are known!

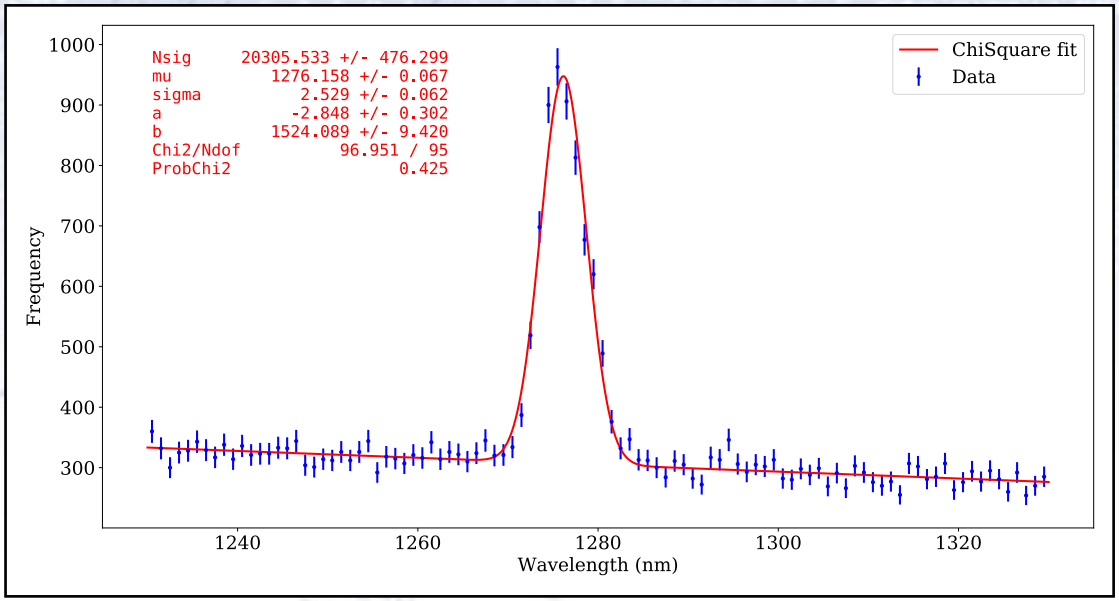
Only exception is the first two peaks used for calibration, but they are very significant.

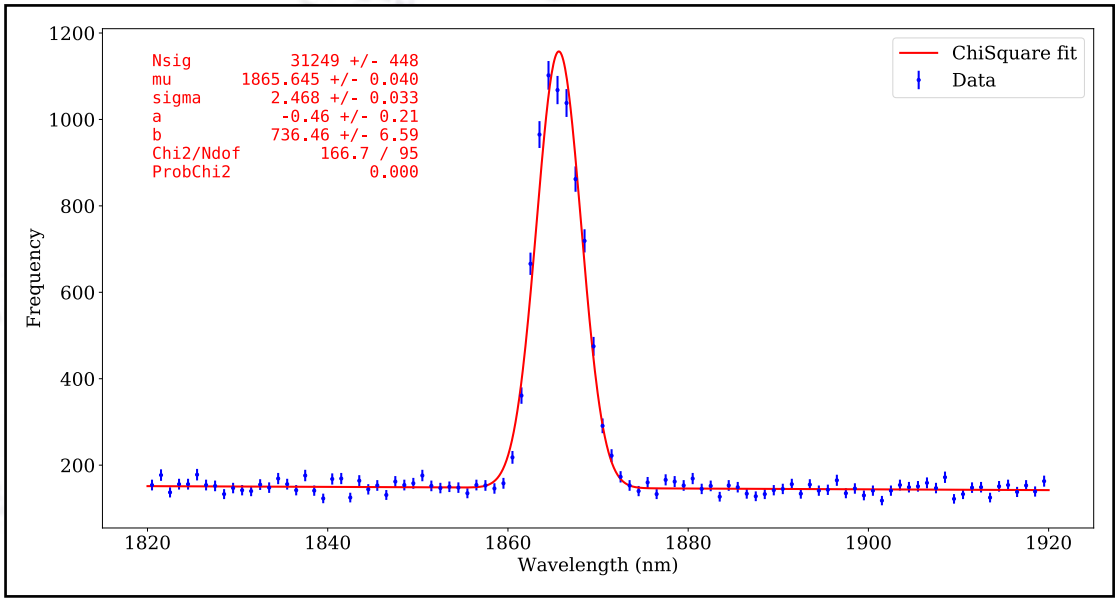
Problem 5.2

Since the rest of the spectrum has no influence, it is an advantage to fit only in the relevant range.

It is good to include some “sideband” (i.e. range outside peak) to establish a good background functions, here a pol1 (i.e. line).

Note how the second peak has a very low p-value (not Gaussian!).





Problem 5.2

Since the two dominant peaks are slightly shifted, they can used for correcting the scale (assuming that their true position is known). Written in many ways:

$$\lambda_{calib} = (\lambda - \mu(1)_{obs}) \cdot \frac{\mu(2)_{true} - \mu(1)_{true}}{\mu(2)_{obs} - \mu(1)_{obs}} + \mu(1)_{true}$$

Problem 5.2

Since the two dominant peaks are slightly shifted, they can used for correcting the scale (assuming that their true position is known). Written in many ways:

$$\lambda_{calib} = (\lambda - \mu(1)_{obs}) \cdot \frac{\mu(2)_{true} - \mu(1)_{true}}{\mu(2)_{obs} - \mu(1)_{obs}} + \mu(1)_{true}$$

$$\hat{\lambda}_1 = \lambda_{1,true} = a\lambda_{1,obs} + b$$
$$\hat{\lambda}_2 = \lambda_{2,true} = a\lambda_{2,obs} + b,$$

which yields the solution

$$a = \frac{\lambda_{2,true} - \lambda_{2,true}}{\lambda_{1,obs} - \lambda_{2,obs}}; \quad b = \lambda_{1,true} - \frac{\lambda_{2,true} - \lambda_{2,true}}{\lambda_{1,obs} - \lambda_{2,obs}} \lambda_{1,obs}.$$

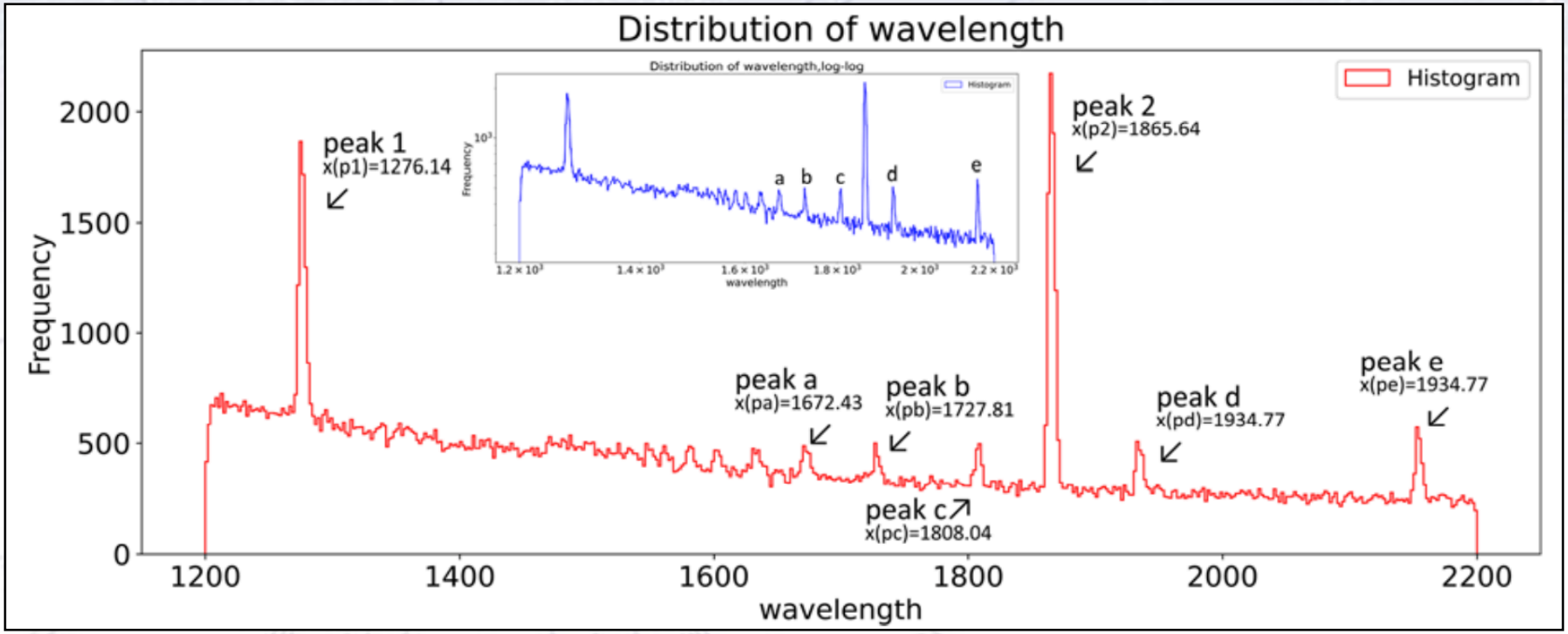
The calibrated estimates for the wavelength is then

$$\hat{\lambda} = a\lambda_{obs} + b \pm \sqrt{\left(\frac{\partial \hat{\lambda}}{\partial \lambda_{obs}}\right)^2 \sigma^2_{\lambda,obs}} = a\lambda_{obs} + b \pm a\sigma_{\lambda,obs}.$$

Including the uncertainties is actually how you would assign a systematic uncertainty to the fact, that you shift the measured values.

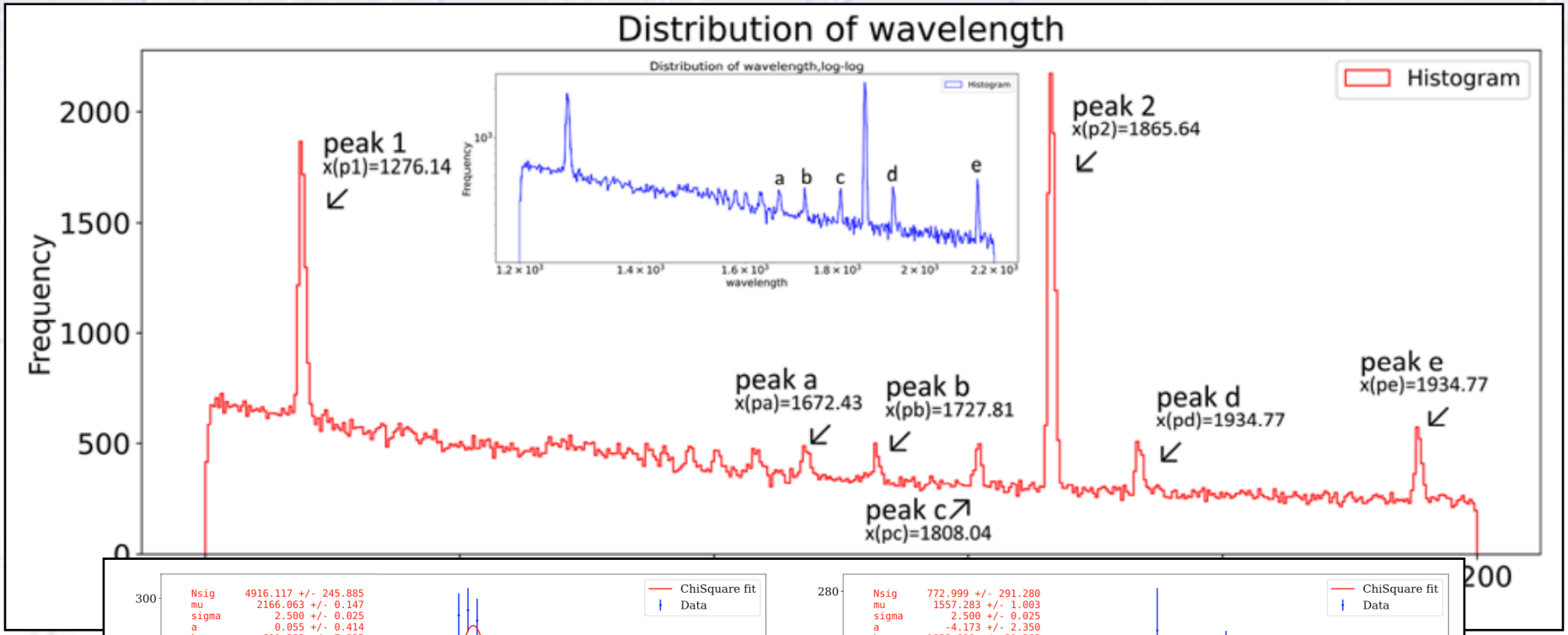
Problem 5.2

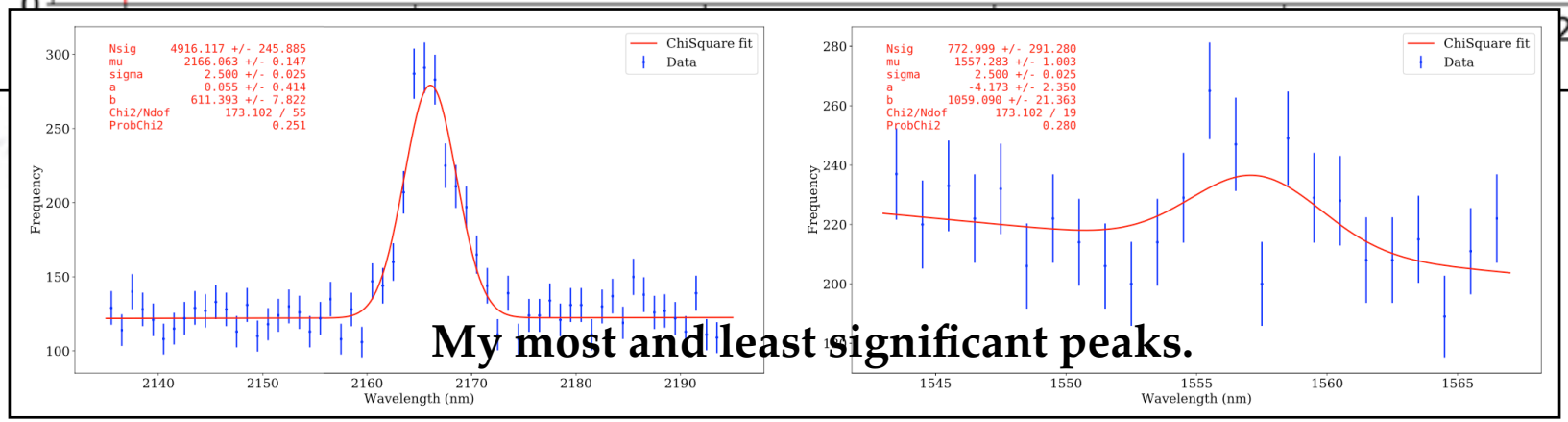
Here is a really nice figure… impressive! Maybe the insert is not needed…
Most people found between 3 and 8 peaks. Beyond that (uncalibrated) it is hard!



Problem 5.2

Here is a really nice figure… impressive! Maybe the insert is not needed…
Most people found between 3 and 8 peaks. Beyond that (uncalibrated) it is hard!

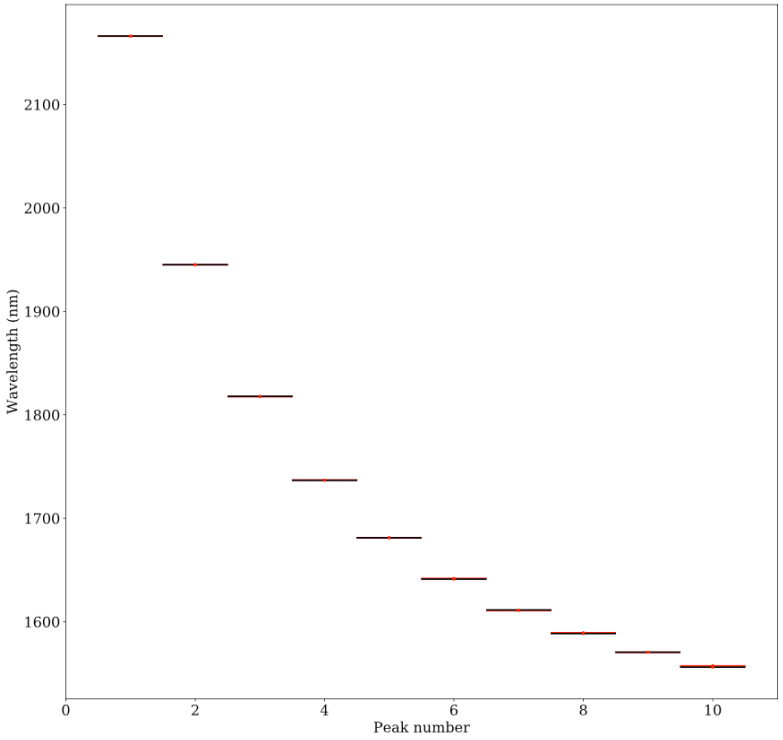




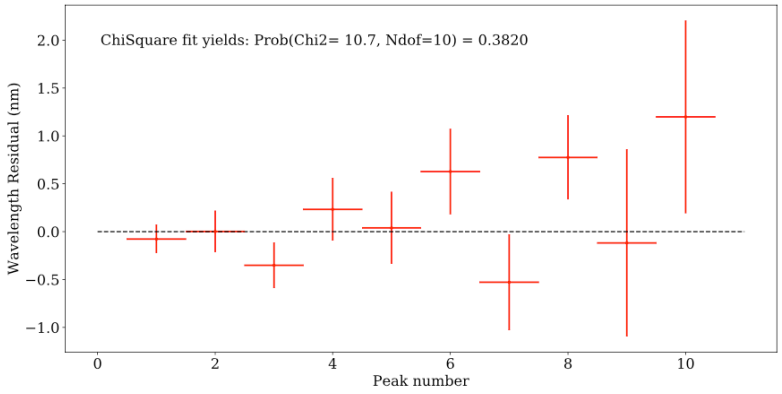
My most and least significant peaks.

Problem 5.2

Comparing several CALIBRATED peak positions should generally be done with a ChiSquare! There are no fitting parameters, but it is still an over constrained system of equations.



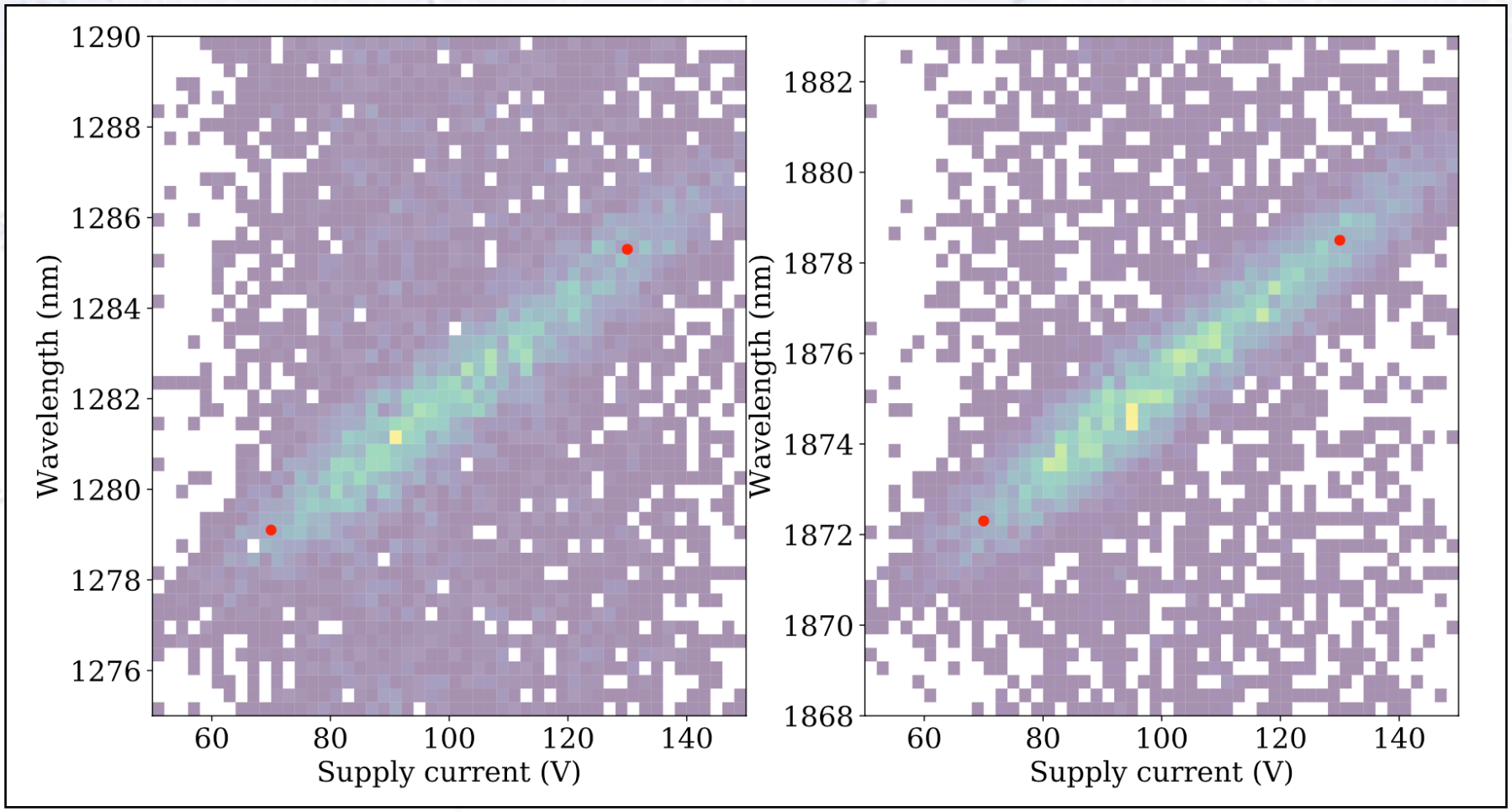
Admittedly, the problem was designed to confirm Bohr's theory… how could one otherwise?



Problem 5.2

Damn… calibration, probably many thought! Yes, it was a tough end problem.

But the two main peaks show clear signs of (the same) linear shift with voltage.

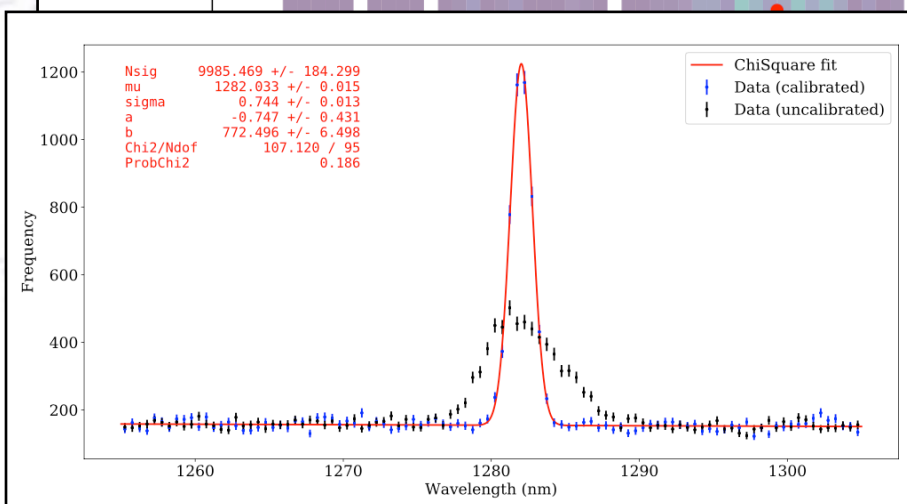


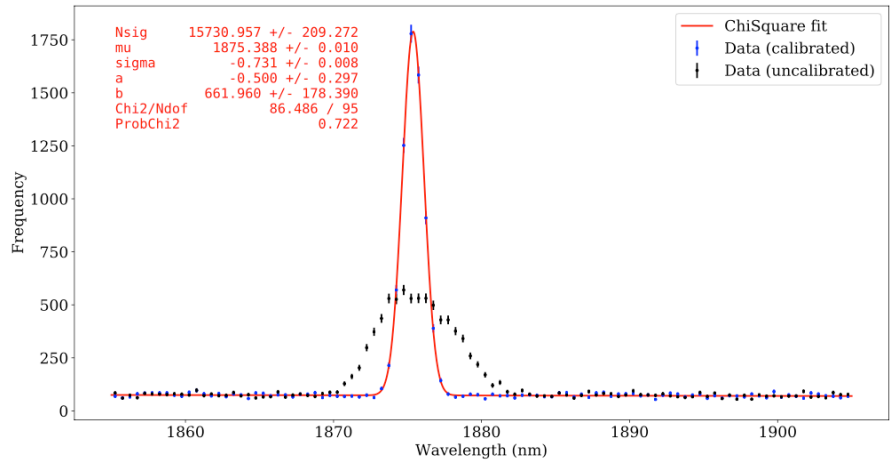
Problem 5.2

Damn… calibration, probably many thought! Yes, it was a tough end problem.

But the two main peaks show clear signs of (the same) linear shift with voltage.

The calibration not only renders the peaks much sharper (reduces resolution by factor ~3), but also very Gaussian. The damn thing works!!!



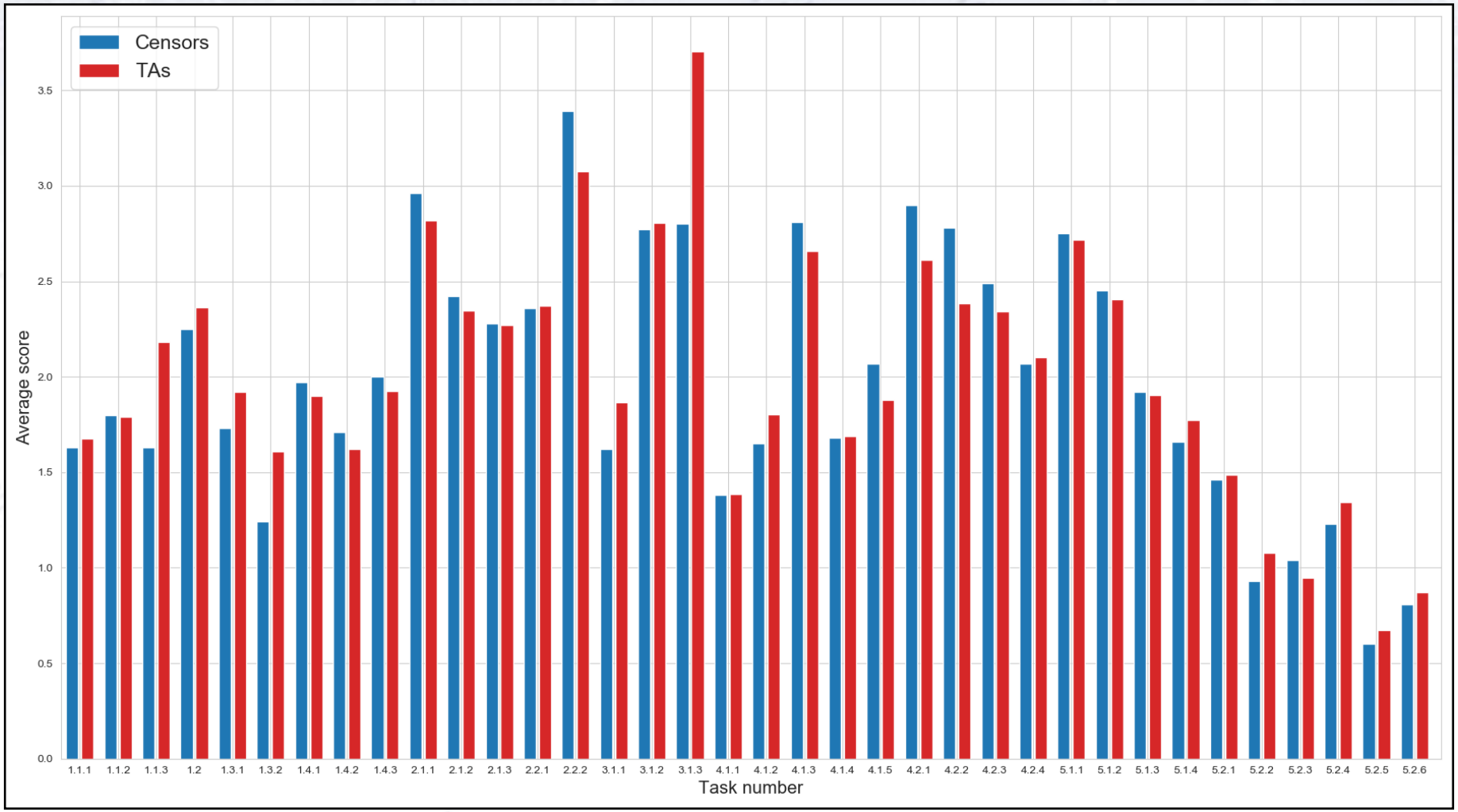


Some statistics

…of course!

Average score per problem

The following figure summarises the average score per problem, divided between Censors (blue) and TAs (red).



Average fraction per problem

The following figure summarises the average fraction per problem, divided between Censors (blue) and TAs (red). Is there a downward trend?!?