Appstat Problemset

Philip Kofoed-Djursner
Tkv976

January 2023

Discussion of some problems was done with Ludvig Marcussen.

1 Distributions and probabilities

1.1 Your friend tells you, that a bag contains 3 white, 6 black, and 7 grey marbles

Taking marbles without putting them back can be thought of in many ways. In this case, we want to know the probability of drawing at least a white ball in two draws with no putbacks given the information about the different balls present. This is equivalent to 1 - the probability of drawing no white balls.

$$P(Drawing\,a\,white\,ball) = 1 - P(Drawing\,no\,white\,ball) = 1 - \frac{13}{16} \cdot \frac{12}{15} = \frac{7}{20} \tag{1}$$

This kind of distribution is known as a hypergeometric distribution, which formula could be used to get the same result in all cases where the answer is less trivial than this.
If the marbles are put back into the bag the chance of getting a result follows a binomial distribution.

$$P(binomial, n, r, p) = p^r(1-p)^{n-r}\frac{n!}{r!(n-r)!} \tag{2}$$

In this case, I want to calculate the probability of drawing 18 gray balls in 25 draws.

$$P(binomial, n = 25, r = 18, p = 7/16) = 0.00295 \tag{3}$$

For distributions with many different outcomes, any given outcome has a low probability of occurring. Therefore it is more telling to calculate the sum of the probability of that outcome and all more unlikely outcomes.

$$P(binomial, n = 25, r \geq 18, p = 7/16) = 0.00404 \tag{4}$$

If I continue under the null hypothesis that the information given to me is correct the result of getting 18 grey marbles or more from the bag has a probability much lower than my accepted 5% significance interval. Therefore I would not trust my friend's information

1.2 The lifetime L of a certain component is exponentially distributed: $L(t) = 1/\tau \cdot \exp(-t/\tau)$

A different way of posing the given question is to ask: What value of τ gives a 96% probability of lasting less than or equal to 500 hours? This can be solved analytically

$$\int_0^{500} L(t)\,dt = 0.96 \Rightarrow -\exp(-500/\tau) + 1 = 0.96 \Rightarrow \tau = 250/\ln(5) \tag{5}$$

I did the integral in python numerically to check and found the value to be true. So if $\tau = 250/\ln(5)$[hours] there is a 4% chance of the lifetime being longer than 500 hours.

1.3 A radio telescope detects 241089 signals/day, based on a 9-week observation campaign

The probability of getting any given result in a setting like this would follow Poisson statistics.

$$P(Possion, \lambda, r) = \frac{\exp(-\lambda)\lambda^r}{r!} \tag{6}$$

In this case $\lambda = 10045$ and r = 9487. If this is computed in python I got an overflow error. As lambda is large the Poisson distribution is well approximated by a Gaussian distribution.

$$P(gaussian, x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{7}$$

With $\mu = \lambda = 10045.375$ and $\sigma = \sqrt{\lambda} = 100.227$. The observation of 9487 signals is beyond 5 sigmas out, and the gaussian is very flat, so the probability of that result is very close to the value of the gaussian distribution at that point.

$$P(gaussian, x = 9487, \mu = 100.227, \sigma = 10045.37) = 7.248 \cdot 10^{-10} \tag{8}$$

The more proper way to calculate the probability is to do an integral around the r value.

$$P(gaussian, r = 9487, \mu = 100.227, \sigma = 10045.37) = \int_{r-0.5}^{r+0.5} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx = 7.249 \cdot 10^{-10} \tag{9}$$

This integration is done numerically with 10000 points, which makes it correct to the 7th digit. With 1000 points it is correct to the 6th digit. This is the probability of this happening in just one 1 hour of observation, but the observations were done for 9 weeks which means that 1512 hours passed. Therefore the actual chance of observing exactly 9487 one or more times over a 9-week period.

$$P(binomial, n = 1512, r \geq 1, p \approx 7.249 \cdot 10^{-10}) \approx 1.096 \cdot 10^{-6} \tag{10}$$

To find out if this is significant I can't just focus on measuring 9487 signals, as every observation is unlikely when the distribution is this wide. Therefore, I ask what is the chance to get 9487 or fewer signals during 1 hour over the 9 weeks.

$$P(gaussian, r \leq 9487, \mu = 100.227, \sigma = 10045.37) = \int_{-\infty}^{r+0.5} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) dx = 1.302 \cdot 10^{-8}$$

$$P(binomial, n = 1512, r \geq 1, p \approx 1.302 \cdot 10^{-8}) \approx 1.97 \cdot 10^{-5} \tag{11}$$

This is 1 in 50000 even after taking into account the length of the observation. In the context given, the result is extraordinary.

1.4 Shooting with a bow, you have 3% chance of hitting a certain target

This problem follows a binomial distribution, equation 2, with n=N and p=0.03
To calculate the probability of hitting the first shot after 20 shots one could ask the alternative question: What is the probability of not hitting in the first 20 shots?

$$P(binomial, n = 20, r = 0, p = 0.03) = 0.544 \tag{12}$$

So the probability is 0.544
For the probability that it will take more than 4000 shots to hit 100 times, again, one can ask: What is the probability of hitting less than 100 times in 4000 shots? This can be calculated with the cumulative distribution function (CDF) for the binomial distribution.

$$P(binomial, n = 4000, r < 100, p = 0.03) = 0.0261 \tag{13}$$

The probability of not hitting 100 times in the first 4000 shots is only 0.0261 which is very low.

2 Error propagation

2.1 Let x = 1.92 ± 0.39 and y = 3.1 ± 1.3, and let z1 = y/x, and z2 = cos(x) * x/y

To calculate the uncertainties in this problem I use the general formular

$$\sigma_f^2 = \sum_{i,j}^{n} \left[\frac{\partial f}{\partial x_i}\right] \left[\frac{\partial f}{\partial x_j}\right] cov(x_i, x_j) \tag{14}$$

If i=j the covariance equals the variance on that parameter. When computing this I get

$$z_1 = 1.6 \pm 0.8$$

$$z_2 = -0.2 \pm 0.3$$

If the covariance is included, and the correlation between x and y is 0.95, one can compute a variance matrix with $i, j \in \{x, y\}$:

$$\boldsymbol{V}_{z_1} = \begin{bmatrix} 0.108 & -0.211 \\ -0.211 & 0.458 \end{bmatrix} \tag{15}$$

The final result for z_1 is

$$z_1 = 1.6 \pm 0.4 \tag{16}$$

So the correlation between x and y reduced the error The same can be done for z_2 to find which variable contributes the most to the final result

$$\boldsymbol{V}_{z_2} = \begin{bmatrix} 0.0729 & 0 \\ 0 & 0.00790 \end{bmatrix} \tag{17}$$

This shows that the error on x is the largest contributor to the error on z_2

2.2 Five patients were given a drug to test if they slept longer (in hours). Their results were: +3.7, -1.2, -0.2, +0.7, +0.8. A Placebo group got the results: +1.5, -1.0, -0.7, +0.5, +0.1.

For the drug groups the result are: $\bar{x} = 0.8 \pm 0.8, \hat{\sigma} = 1.8$
To find the probability of the two group's results being different, one can perform the student t test. For the student t test to apply the two distributions variance need to be the same. The F test can show if this is true. So the null hypothesis is that the variance is the same within 5% significance.

$$F = 3.38 \tag{18}$$

Consulting Table 8.2 in Barlow the critical value for 4-4 degrees of freedom is 6.39. This value is below that and therefore we assume the variance is the same. Doing a student t test on the data from the two groups I find t = 0.729. This can be calculated to a probability that they are different (one-tailed) I.e that drug patients slept longer P = 0.757. This is not significant compared to a 5% significance interval, so one could not say that the drug improved the hours slept

3 Simulation / Monte Carlo

3.1 Assume $f(x) = C \cdot x^a sin(\pi x), x \in [0; 1]$ and a = 3 is a theoretical distribution

For this distribution, I used the accept/reject method. The distribution is neither easy to integrate nor invert, therefore it is the logical choice.
To determine C I did a numerical integration of the distribution with 100000 points. I found C = 8.013.
The next problem I found a bit ambiguous, but I interpreted it as meaning that the error on the fitted a

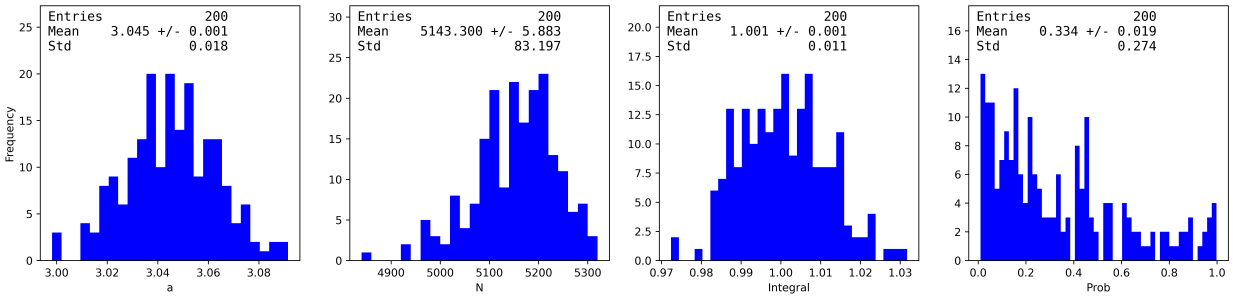

Figure 1: Results form 200 runs of fitting monte carlo data from the given distribution

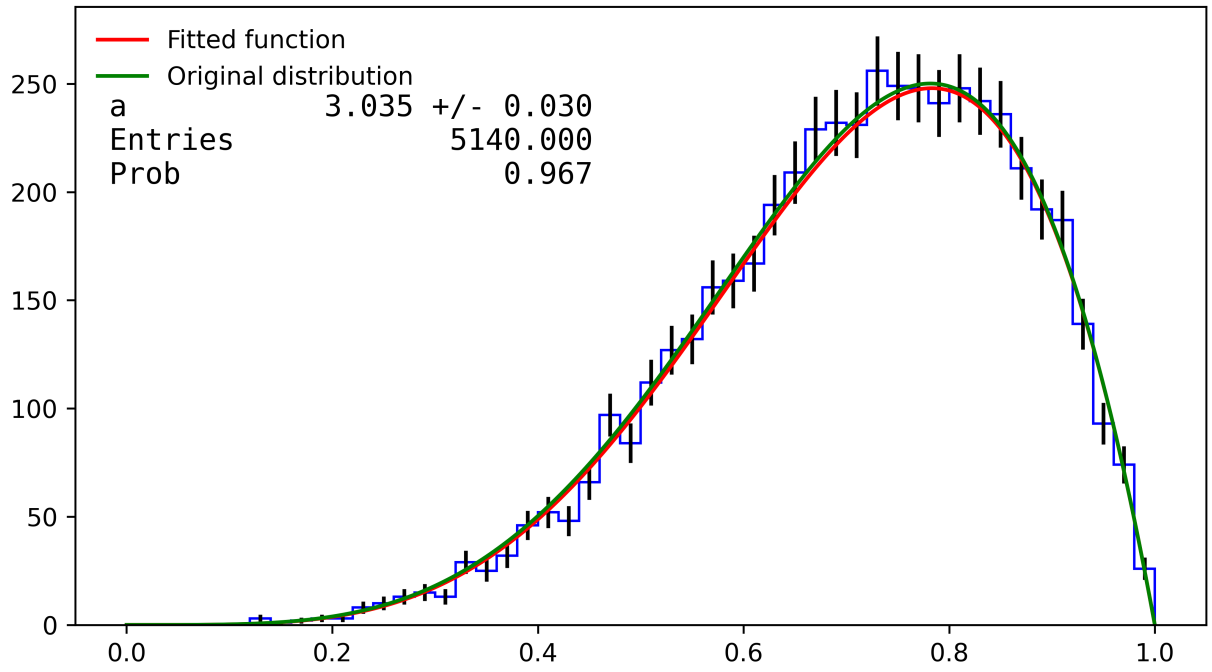

Figure 2: 1 of the 200 fits used to produce the data in Figure 1

value should be less than 1% of that value. I choose to bin the data into 50 bins between 0 and 1. The errors on the bins are Poisson errors. Furthermore, I choose to include a condition that the chi2 probability should be greater than 0.01. I started at N = 4500 and incremented with 20 more data points if the conditions were not met. Discarding the old data points and generating new ones each time. I did 200 runs of this to get an estimation of the number of data points needed. Results are shown in Figure 1. It can be seen that the average number of runs needed was $\bar{N} = 5143 \pm 6$. An example of a fit is shown in Figure 2. This was the 200th run and just so happened to be a very good fit.
The problem with this approach is that higher fitted a values are favored as they allow a larger error, but this, I would guess, is still the right approach if one had no prior knowledge of the real distribution.

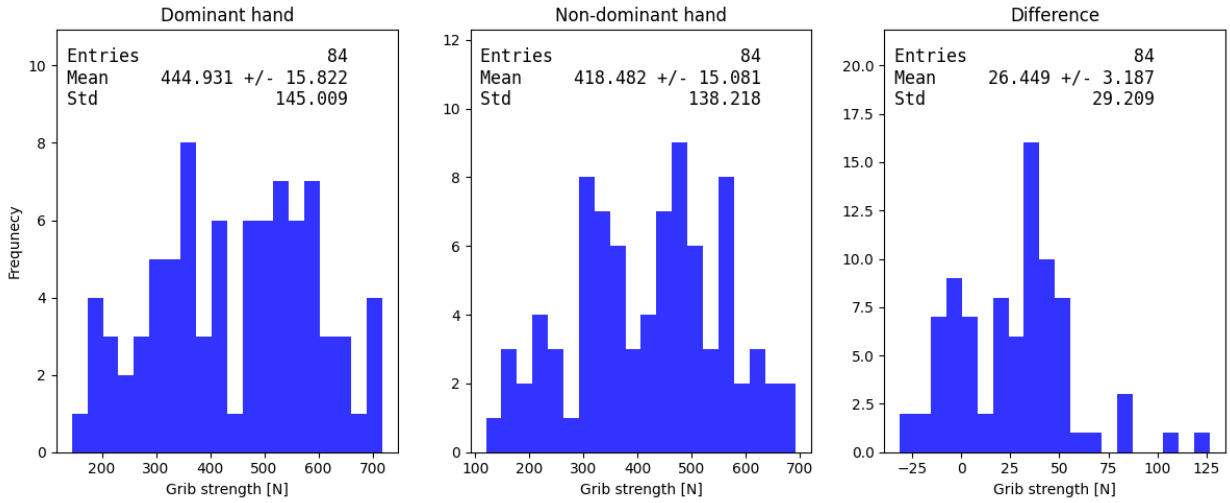

Figure 3: Data about the grip strength

4 Problem 4

4.1 You measure the grip strength (G in Newton N) in the dominant and non-dominant hands (based on writing) of 84 persons, to determine if there is a difference

I found that 68/84 are right-handed. The mean and standard deviation of the two distribution is found in Figure 3. I tested first with a F distribution to compare the variances. F = 1.10 for the two distributions. This is easily within the 5% significance threshold, so no reason to think they are different. The t-test gives t = 1.21. This is also way within the 5% significance threshold, which means that the probability that the two distributions have the same mean. So yes, their means are compatible.
The mean and std of the difference are also shown in Figure 3. The samples in the strength of the dominant and non-dominant hands are highly correlated $\rho = 0.98$. If you start with the null hypothesis that there is no difference and conduct a t-test, you get t = 8.30. This is far beyond the 5% significance cutoff (two-tailed) and therefore the null hypothesis must be rejected. So there is a difference in grip strength between dominant and non-dominant hands.

4.2 From microscope images, you measure size (S in μm) and intensity (I) of large molecules in a sample.

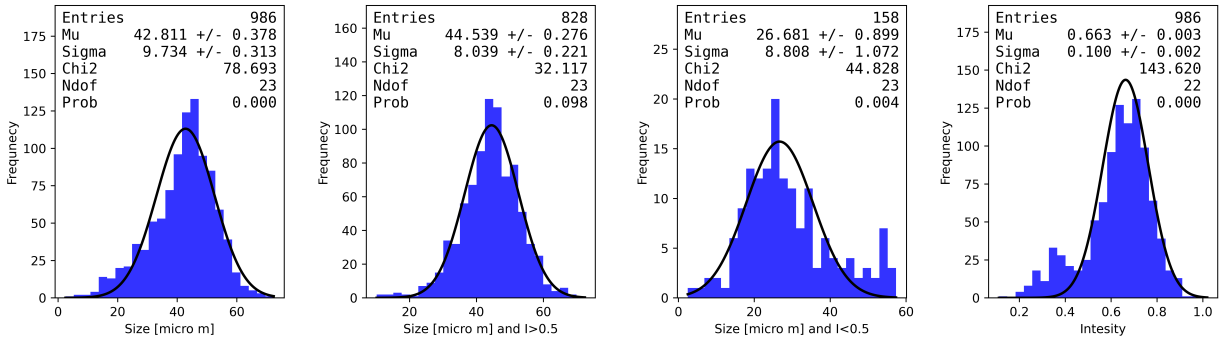

Figure 4: Data of the molecule size and intensity and a fitted Gaussian on top. Fitted and plotted with 25 bins using χ^2

The data and their fits are shown in Figure 4. They are fitted and plotted with 25 bins and using the χ^2-method. The full-sized dataset does not follow a Gaussian. The data with I $>$ 0.5 is acceptable within the usual χ^2 probability interval: 0.99 $>$ prob $>$ 0.01.

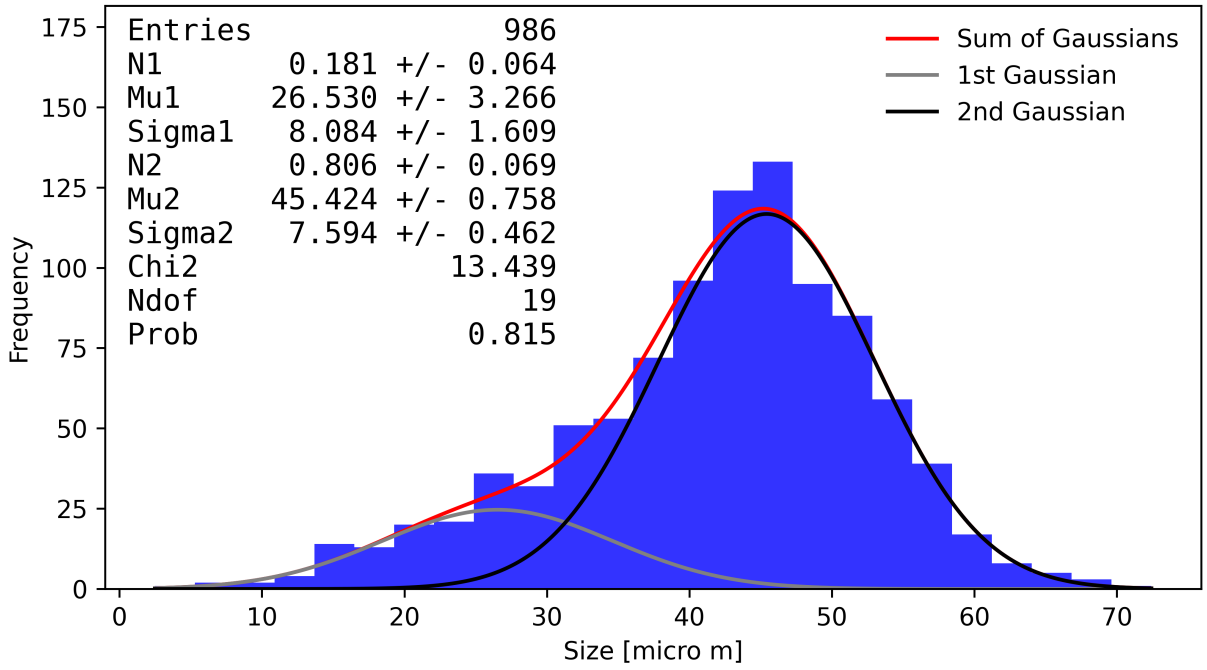

Figure 5: Data of the molecule size with a fitted double Gaussian on top. Fitted and plotted with 25 bins using χ^2

When including two Gaussians to fit the full dataset I get a very good fit with p = 0.815 (Figure 5. I want to have the ratio that the second Gaussian contributes to the total area equal to 0.9. I did first a rough search and afterward a fine search. With my parameters, I found that if I have only sizes $>$ 26.8, 0.9 of them would result from the 2nd Gaussian (The one I interpret as "New molecule"). This means that 886

molecules are included. I would have loved to do error propagation on this result, but I couldn't get python to do the definite integral of the Gaussian and differentiate. It should be possible to still use equation 14.

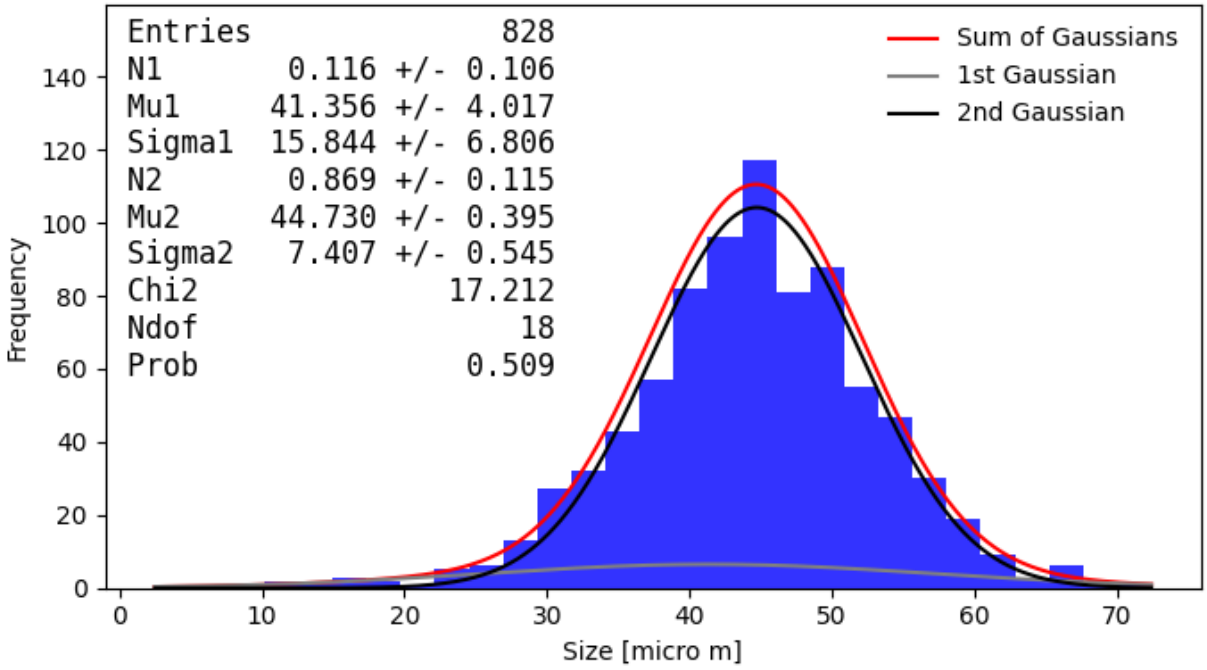

Figure 6: Data of the molecule size (I > 0.5) with a fitted double Gaussian on top. Fitted and plotted with 25 bins using χ^2

The data in Figure 6 is the size when only including I > 0.5. It can be seen that the second Gaussian fit very weirdly. I would have loved to keep the μ and σ from the last plot and just let N fit but I thought of that too late. If I ignore this fact and continue with the same analysis as before I get that the sizes include need to be over 26.9 and 807 molecules are now included. This does not make much sense as we both get a smaller interval of sizes to include and fewer molecules included with more data.

5 Problem 5

5.1 You are studying the growth of an algae type, by considering the area it covers (A in cm2) as a function of time (t in days)

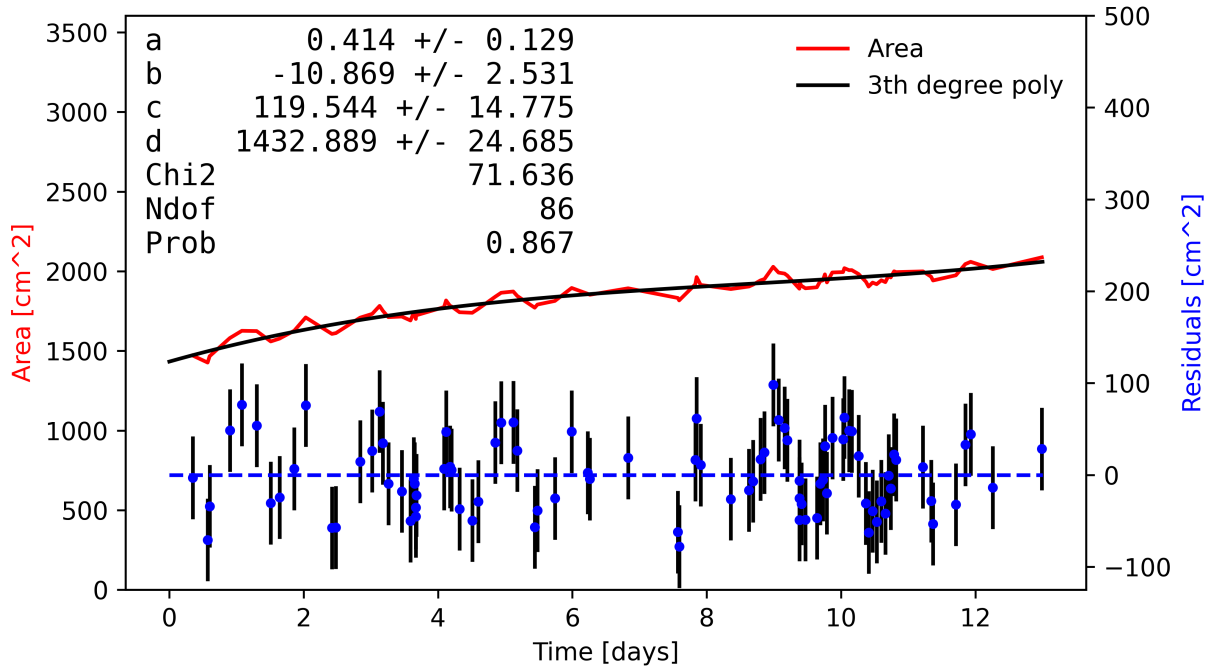

Figure 7: Area data fitted to a 3th degree polynomial

With the given errors the fit is very good. But, doing a run test reveals that only 27 runs happens in the data. In this data I would expect 46 ± 5. This result is 4 sigma out and therefore I conclude that the fit is not good and more details need to be included. Such a detail could be to included a small oscillation. The function is fit to is now

$$f(x) = (ax^3 + bx^2 + cx + d)(1 + A\sin(\pi x + \phi)) \tag{19}$$

As the time is in days and the oscillation is expected to be the day-night variation, I have set the frequency of the oscillation to π and included an angle to let the data start at any point doing the day. I figure **??**, The fit and parameters are shown. The run test is also much more in agreement with the expected value. I get 42 runs. This is now only 0.8σ away and is acceptable.

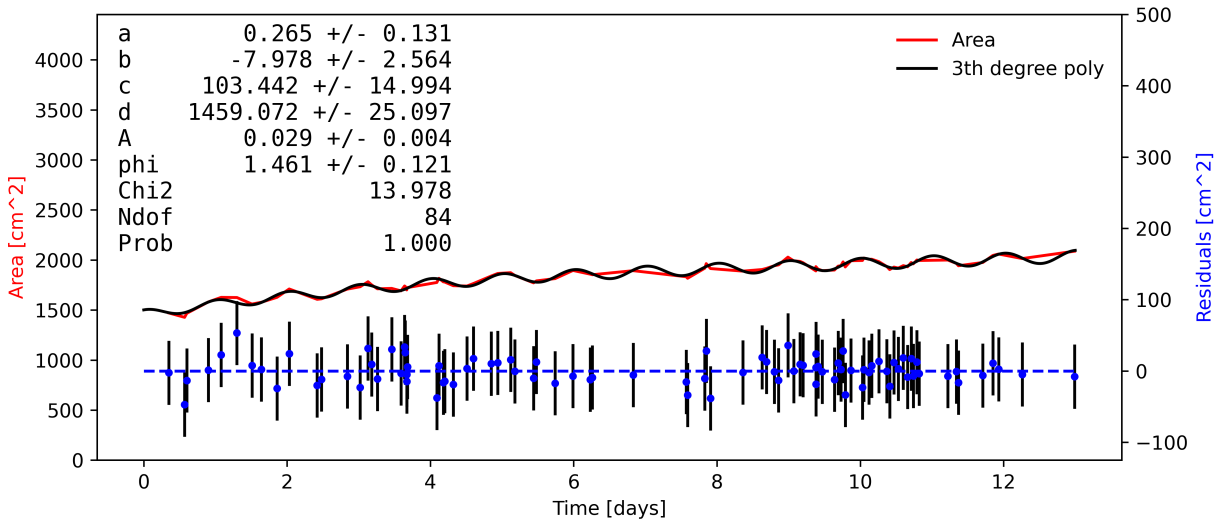

Figure 8: Area data fitted to a 3th degree polynomial and a small oscillation

The p on the fit seems to be too good, which suggests that the errors have been overestimated. Even if $\sigma_A = 22.5$ then $p = 0.992$. This is still too high, so a better estimate of the real error is the standard deviation of the residuals which is 18.

5.2 In the centennial of Bohr's Nobel prize, you decide to test his atomic model, and measure the spectral lines of hydrogen in the infrared spectrum 1200-2200nm, where you would expect to see some of the n1 = 3 and n1 = 4 lines.

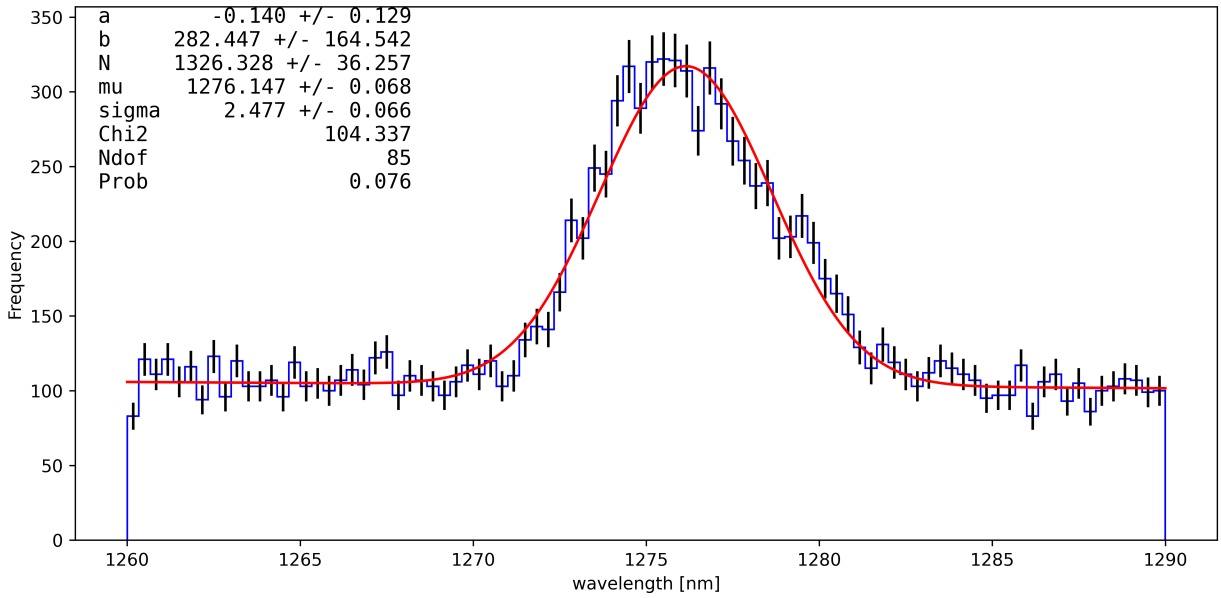

Figure 9: First peak fitted with 90 bins using χ^2

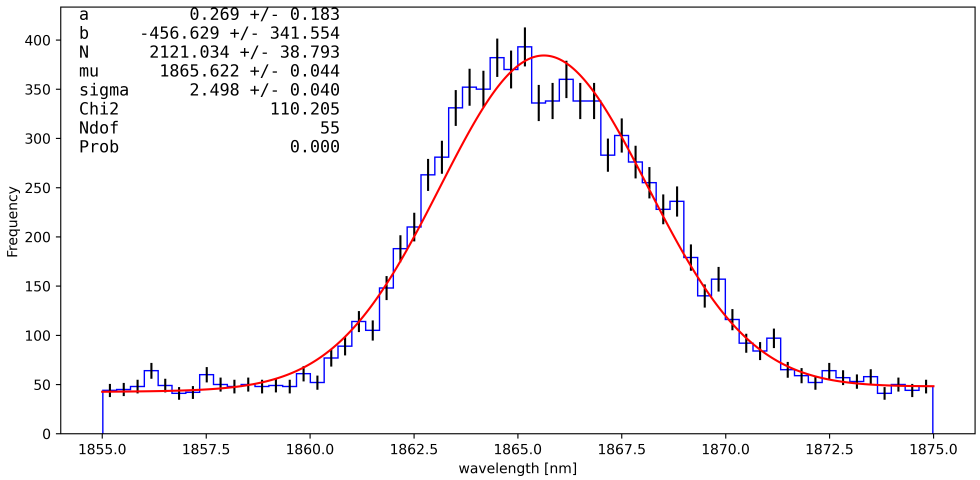

Figure 10: Second peak fitted with 60 bins using χ^2

The two peaks are fitted with a Gaussian plus a linear function (Figure 9 and 10). The linear function was included to simulate the noise. Both fits are bad with the use of Poisson errors, which might be because the real errors are larger. I expect this as the data looks quite noisy. Even with the low p, the σ of the two fits are almost identical, and would easily pass an F test for them being the same.
The peaks lie at a little different position compared to the theoretical value. This can be made up for by calculating a linear calibration which would shift the peaks into the correct position. So by adding the linear curve in Figure 11, the peaks would be put back into their correct position.

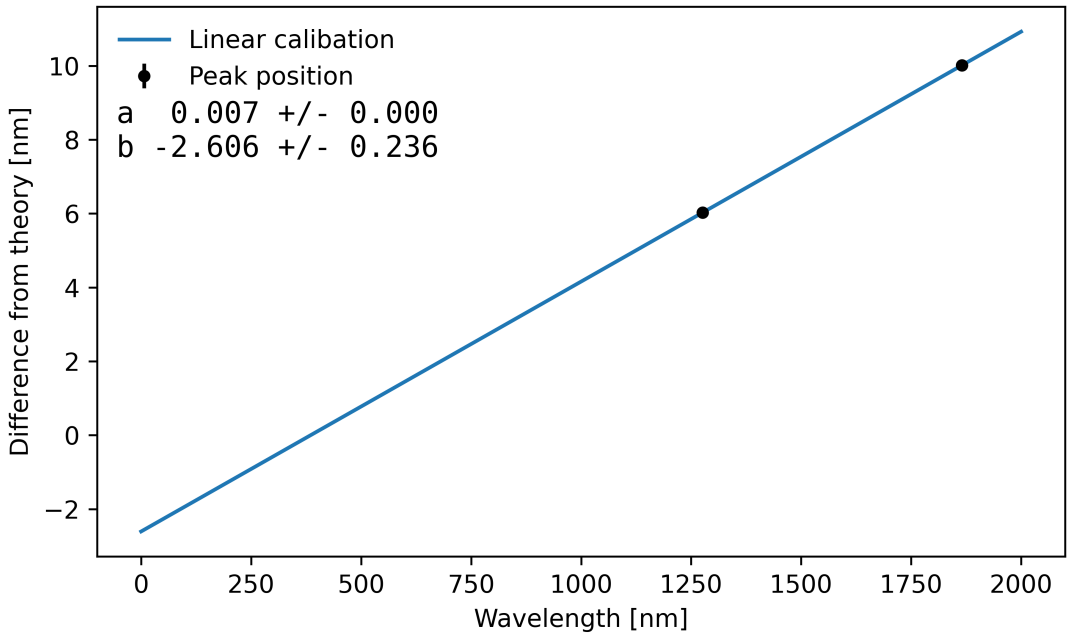

Figure 11: Linear calibration needed to put the peaks into the correct positions

I didn't have time to finish the last questions, but by looking at the full data it seems like there are a few more peaks. The peak at 2170 is the most significant. This is the n=4 to n=7 transition. The rest of

the peaks are n=4 to higher shells than 7. It does by eye look like they follow as, n=4 to n=7 is 2166 nm , n=4 to n=8 is 1945 nm and n=4 to n=9 is 1818 nm. This matches the data well.