AppStat2023: Exam

Niels August Davidsen , phx657

January 2024

1 - Distributions and probability

1.1 An electronic device with 3 components

1.1.1 Probability of the device NOT failing in the first year

For the device to not fail, each component has to last the year. This can be expressed like this where the subscript denotes the components

$$P_{1,\text{no fail}} = 1 - 0.009$$

$$P_{2,\text{no fail}} = 1 - 0.016$$

$$P_{3,\text{no fail}} = 1 - 0.027$$

Then the total probability for the device not to fail is

$$P(\text{no fail}) = P_1 \times P_2 \times P_3 = 0.95$$

1.1.2 When does the device have 50% fail probability?

For each year the chance of each component not failing is $1 - P_i$ which means that for n years the chance of that same component not failing is $(1 - P_i)^n$ which makes the total probability for not failing:

$$P(n, \text{no fail}) = (P_1)^n \times (P_2)^n \times (P_3)^n$$

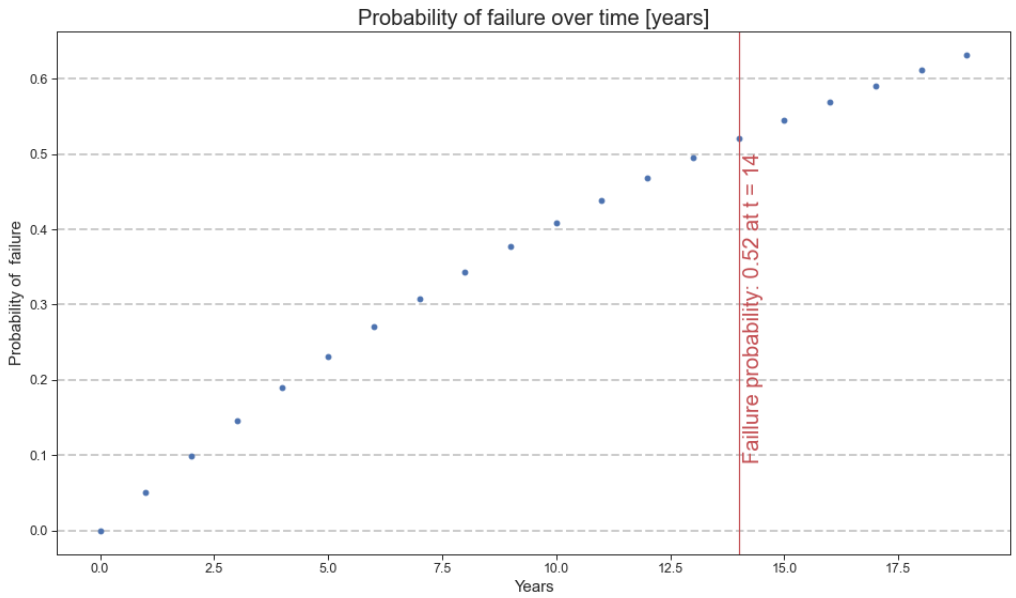

Figure 1: Plotted here is $1 - P_{Poisson}$ for the equation above. The year where the probability of failure is above 50% is $t = 14$ years

It is clear from the plot, that the chance of failure is above 50% after 14 years.

1.2 A busy store with a rate of customers

1.2.1 Distribution of customers pr. day

When dealing with rates of discrete events happening in a continuum, we should always think of the Poisson distribution. Barlow example with thunderstorms explains is pretty well: we can ask what is the rate of lightning strikes, but it is meaningless to ask, how often it doesn't occur (as opposed to the binomial distribution).
The Poisson distribution is written:

$$P_{Poisson} = \frac{e^{-\lambda}\lambda^r}{r!}$$

where r is and the number of events (customers per day) and lambda is the expected number of events (i.e. the rate)

1.2.2 What number of customers define a busy day?

If we use the above equation and find the distribution for a range of r, we can then integrate it in some range $r = [r, \infty]$ to get 20% of the area under the distribution. This means, that a day with the amount of customers in that integration range is considered a busy day.
The specific number of customers can be found by using the Poiosson CDF in $P(r) = 80\%$ to find r.

This r is found to be $r = 59$

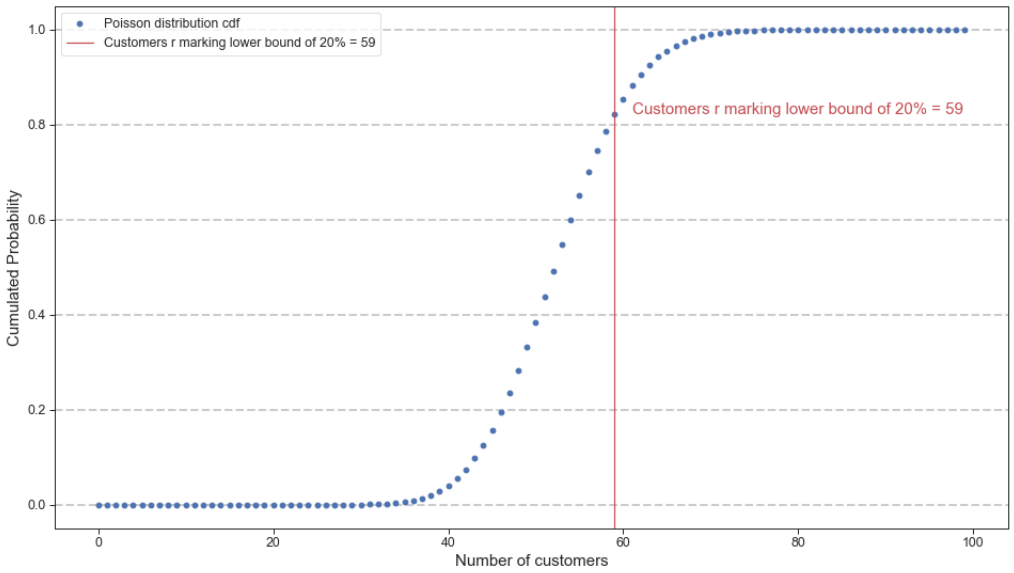

Figure 2: Showing the CDF distribution of Probability density as a function of customers on a random day. It is shown that 59 customers marks the lower bound for what denotes a busy day.

1.2.3 Average number of customers on a busy day

For this question i decided to use simulation: I decided to draw 10000 random numbers from the Poisson distribution and then take the average of the top 20% (i.e. the 2000 biggest numbers)

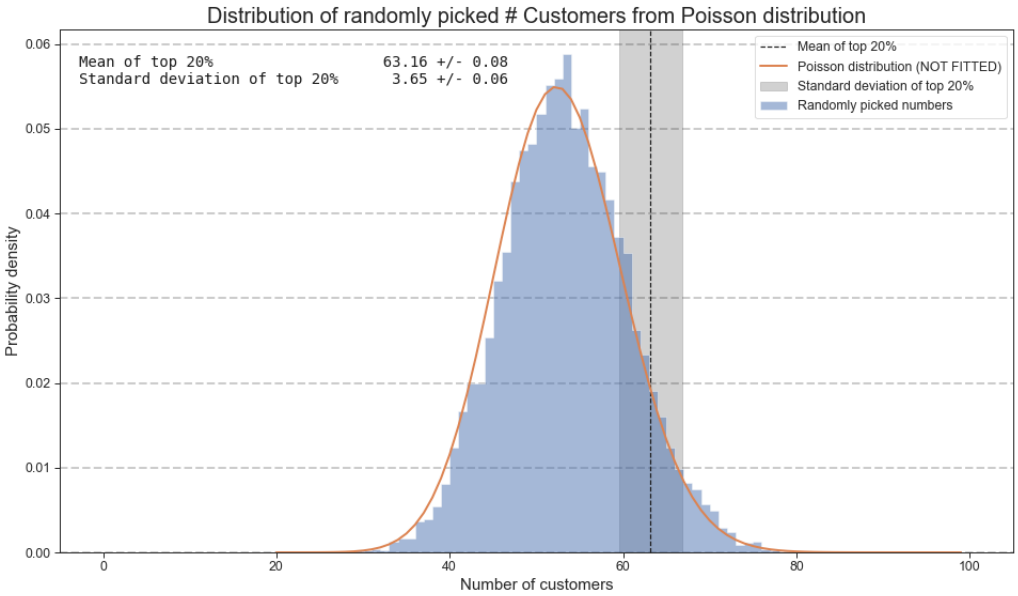

Figure 3: Plot of Poisson distribution with the mean of the top 20% marked along with the standard deviation in the top 20%

As written in the plot, the average number of customers on the 20% busiest days is 63.07±0.08 customers

2 - Error propagation

2.1 Measuring the speed of sound in water

2.1.1 Combined result and uncertainty

I used equation (4.6) and (4.7) to calculate the weighted mean and the combined error. The results were:

$$V_{comb} = 1488 \pm 7$$

2.1.2 Adding measurements

I assume that we are supposed to add the points once more, as if someone were to measure the exact same values.
As expected from equation (4.7) the overall error changes more when adding the last 4 points, as they have the smallest error - we then see the uncertainty on our result drop to ±5. The error when adding the first five points again, doesn't change that much (only about 0.14) but as our precision in the experiment is with no decimals, the error doesn't seem to change at all from the original ±7.

2.1.3 Consistency between measurements

Two answer this I calculated a χ^2 for all the measurements in relation to the weighted mean and the combined error. I then calculated a p-value for this χ^2 with $N_{dof} = N_{measurements} = 9$

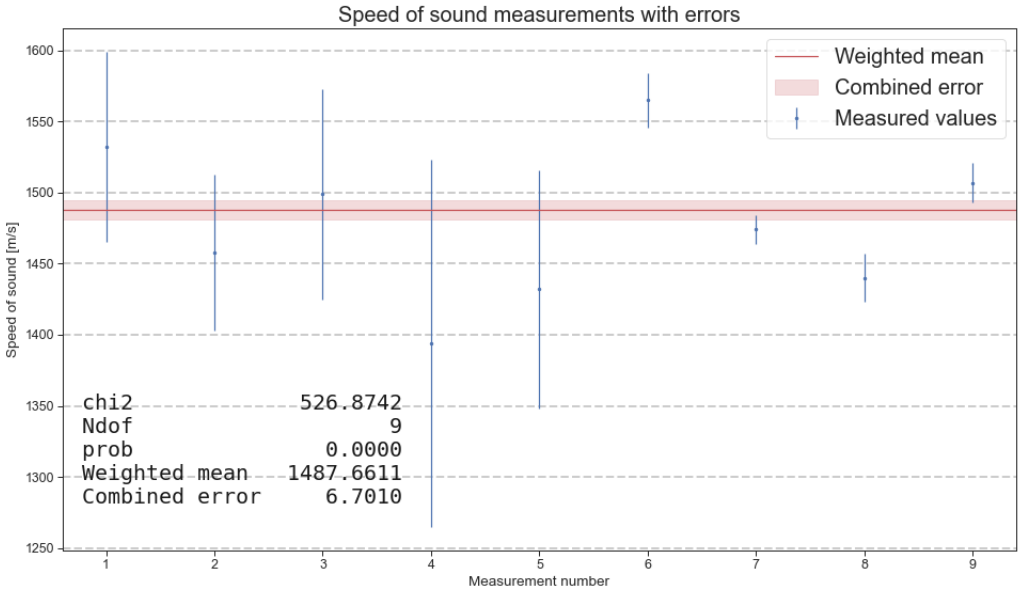

Figure 4: Plot of the nine measurements with errors. Plotted along side is the weighted mean and the combined error. A χ^2 and p-value is shown, showing no consistency between the measurements and the weighted mean.

With $p(\chi^2 = 526.87, N_{dof} = 9) \approx 0$ we must reject that the measurements are consistent with each other.

To improve this, we could remove some of points with the smallest errors, resulting in a larger combined error, which would improve my χ^2 and also my p-value. This would in practice NOT be a good idea, but for the sake of improving a p-value, I will try. This is shown in the plot below, and as you can see, the χ^2 is much improved. The p-value is now also a lot better at $p = 0.0375$ which is within a 97% significance level.

I also tried removing points the furthest away from the weighted mean using a t-test as reference, and i saw some improvements in the χ^2 value, but little to none in the p-value

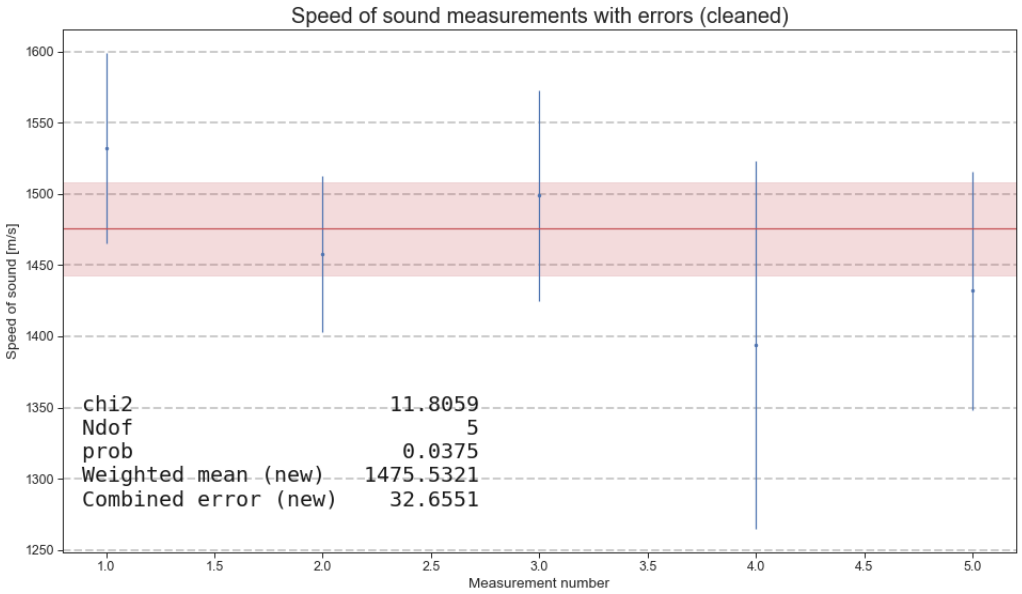

Figure 5: Almost the same plot as before except now some points have been taken away; clearly improving the χ^2 and p-value, but also changing the weighted mean and combined error a lot!

2.1.4 Comparing to real value of 1481 m/s

Without cleaning my data, the actual value is just within one sigma of the weighted mean. Specifically a $z = 0.99$.
For the cleaned data, i have a huge uncertainty, so naturally the actual value is within one sigma of the new weighted mean. Here the z-value is $z = 0.17$ so well within one sigma.

2.2 Damped harmonic oscillator

2.2.1 Uncertainty of x at t=1.

I used standard error propagation to solve this problem resulting in the following equations.

$$\frac{\partial x}{\partial A} = e^{-\gamma t} cos(\omega t)$$

$$\frac{\partial x}{\partial \gamma} = -tAe^{-\gamma t} cos(\omega t)$$

$$\frac{\partial x}{\partial \omega} = -tAe^{-\gamma t} sin(\omega t)$$

And the equation for the error on $x(t)$

$$\sigma_x = \sqrt{(\frac{\partial x}{\partial A})^2\sigma_A^2 + (\frac{\partial x}{\partial \gamma})^2\sigma_\gamma^2 + (\frac{\partial x}{\partial \omega})^2\sigma_\omega^2}$$

As far as I can tell, none of the terms have any problems around the values og A, γ and ω so the i dont see the equation breaking down.
I get the value and uncertainty of

$$x(t = 1) = 0.41 \pm 0.16$$

2.2.2 Uncertainty for different x

The uncertainty of $x(t)$ is shown in the plot below along with the individual terms as a function of t aswell. At small t, the A-term clearly dominates the uncertainty, and then the γ and the ω starts oscillating between who is dominating. Ultimately the error goes towards 0 which makes sense with the decaying $e^{-\gamma t}$ in all terms.

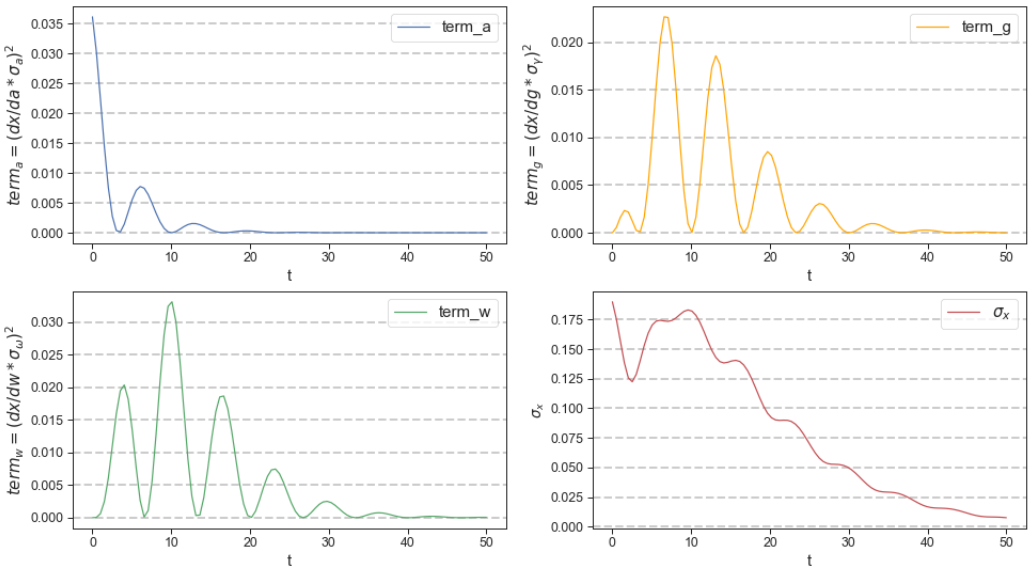

Figure 6: Uncertainty on x, as well of all terms in the uncertainty as a function of time t

3 - Simulation / Monte Carlo

3.1 Penalty statistics

3.1.1 Chance of scoring when aiming at 2.5m

If I understand the question correctly, I assume that every shot I take, is Gaussianly distributed with $\mu = 2.5$ and $\sigma = 1$. For my simulation, I take $N_{total} = 1000$ shots. The I

calculate the probability of me scoring using the given formula, but if $|x| > 4$ I set the probability to 0. And finally the probability of scoring is then calculated as the average scoring probability for a specific x, here $x = 2.5$

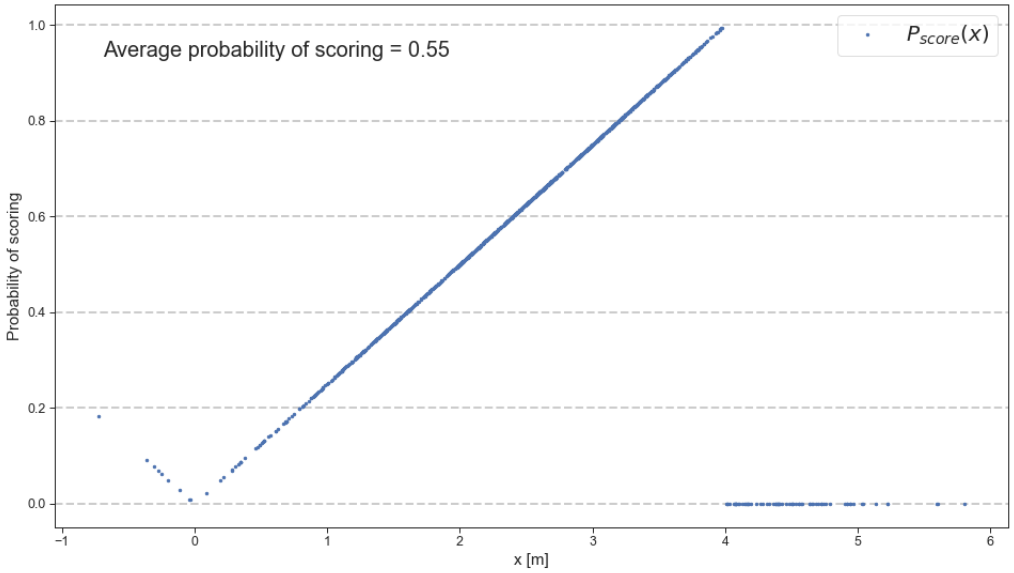

Figure 7: Probability of scoring when aiming at $x = 2.5$m. A distribution of probabilities as a function of x where the average probability is displayed in the top left corner

The average probability of scoring for $x = 2.5$m is then 0.55

3.1.2 Best x for high probability of scoring

Here i essentially followed the same procedure, but i now looped over x-values between $x = [-4, 4]$. For every x-value i looped 100 times and then took the mean of those ten points as well as the standard deviation. This gave me a distribution of average probabilities as a function of x with an error on each point.

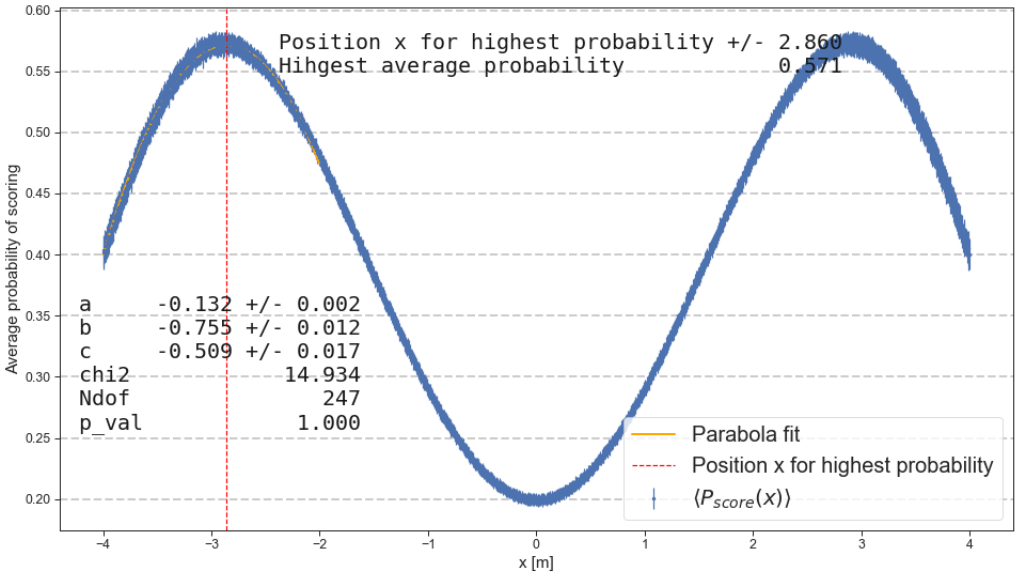

Figure 8: Distribution of average probability as a function of x. Fitted to the distribution is a parabola which help determining the position of x with the highest average probability of scoring.

For finding the position x with the highest average probability, i fitted the distribution with a parabola on one of the two sides, and from the fit values I then calculated the $x_{max} = \frac{-a}{2b}$ where a and b are the fit parameters. The results are shown in the plot and the location with the highest probability of scoring was found to be:

$$x = \pm 2.860\text{m}$$

The fit has a p-value of 1 which is due to the large errors on each point and the many points contributing to the fit values.

3.2 A challenging PDF

3.2.1 Finding C_{PDF} and drawing 100 numbers

Finding C_{PDF} i numerically integrated the PDF from -3 to 3 and used that

$$C_{PDF} = \frac{1}{\int f(x)} = \frac{1}{3\pi} \approx 0.106$$

I then used the Accept/Reject Monte Carlo method for sampling my distribution as both x and y was bound in the interval. I sampled 100 points and used 10 bins in the specified x-range. I got the following distribution:

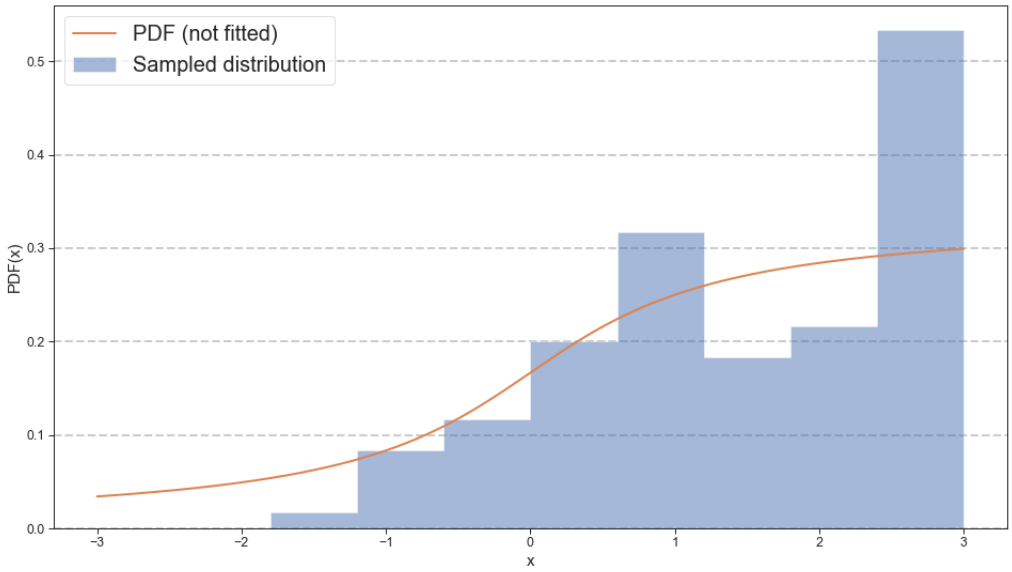

Figure 9: Plot of $f(x)$ distribution along with sampled distribution using Accept/Reject.

3.2.2 Fit the distribution a check C

To fit this data, I would preferably use an Unbinned Likelihood fit, as the data count for some of the bins, are not very high, and thud I cannot assume Poisson distributed errors on the bins. If this somehow doesn't go as planned (foreshadowing) I will do a χ^2 fit and see what that gives me overlooking the low statistics for a bit. So I did exactly that:

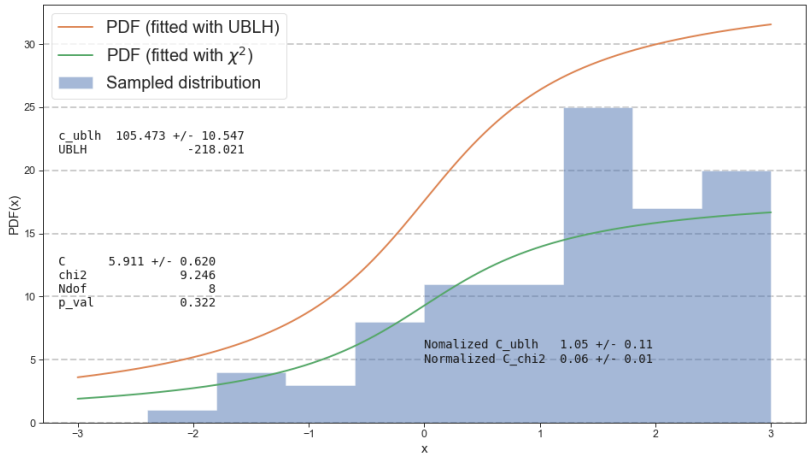

Figure 10: The probability distribution with both an UBLH fit and a χ^2 fit. Fit values are printed as well.

The two fit values I got from the fits were bot nothing close to my estimated $C_p df$.I tried normalizing them but I still ended up with values 10 times greater and half as big as $C_p df$. Not even within one sigma of their error, did they match my estimated value.

If I should improve this, I would try to sample a larger amount of numbers and see if that helped.

4 - Statistical tests

4.1 Indian vs Chinese populations size

4.1.1 Linear fit to Indian data

I started off by fitting a linear fit to the specified data just using sicpy.optimize.curve_fit. From this plot, i calculated the residuals from each point to the plot. I then did an Unbinned Likelihood fit with a Gaussian distribution to the residuals to estimate the standard deviation of the residuals.

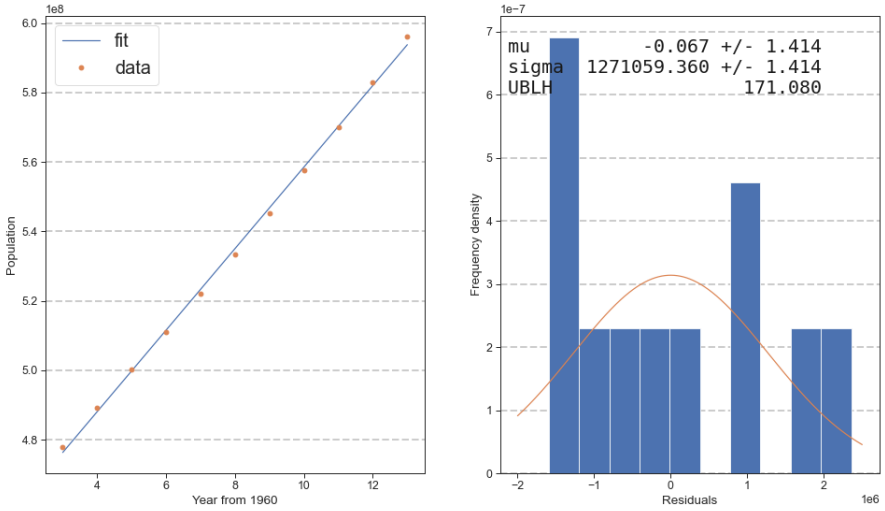

Figure 11: Left plot: The first plot using scipy's curve_fit without errors. Right plot: the Unbinned Likelihood plot for the residuals. On the figure the estimated error is shown.

Lastly i used this standard deviation on each point as the error on population. I then fitted again, this time using Minuit and a χ^2 linear fit.

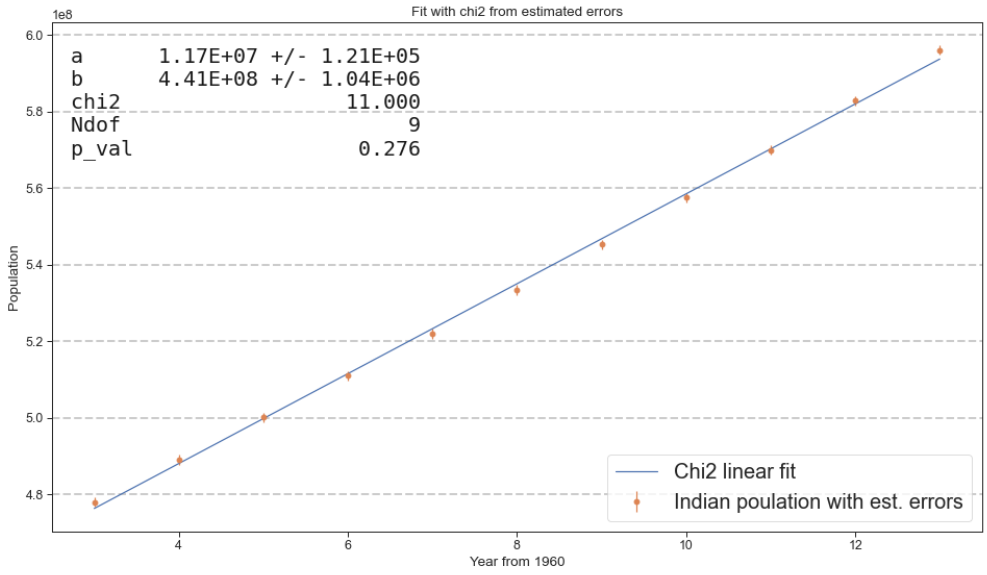

Figure 12: Second time plotting: with iMinuit and a χ^2 linear fit. The fit parameters along with the χ^2 value and the p-value are printed in the plot

Looking at the χ^2 value and the p-value, we must say, that the data points with errors result in a pretty good fit with a p-vale of 0.276 and a χ^2 of 11.000

A short note, is that i changed the x-axis so that, on all the figures it is years from 1960 instead of the specific year. This makes sure, that the fitting constants make sense, when i print those on the plot (see fig. 11)

4.1.2 When does India overtake chine (pop. size)

I decided that the linear fit worked well for estimating the Indian population, so i stuck with that for this problem, but expanded to the whole data set of course. For the Chinese population, i used a parabolic fit with 3 fit parameters; a, b and c.

I then fitted the two data sets with their own function respectfully and then plotted to lines from these fits. Lastly i calculated the intersection between the two fits, to find out where India would overtake China (population wise).

This was at $t = 62 =$ year 2022

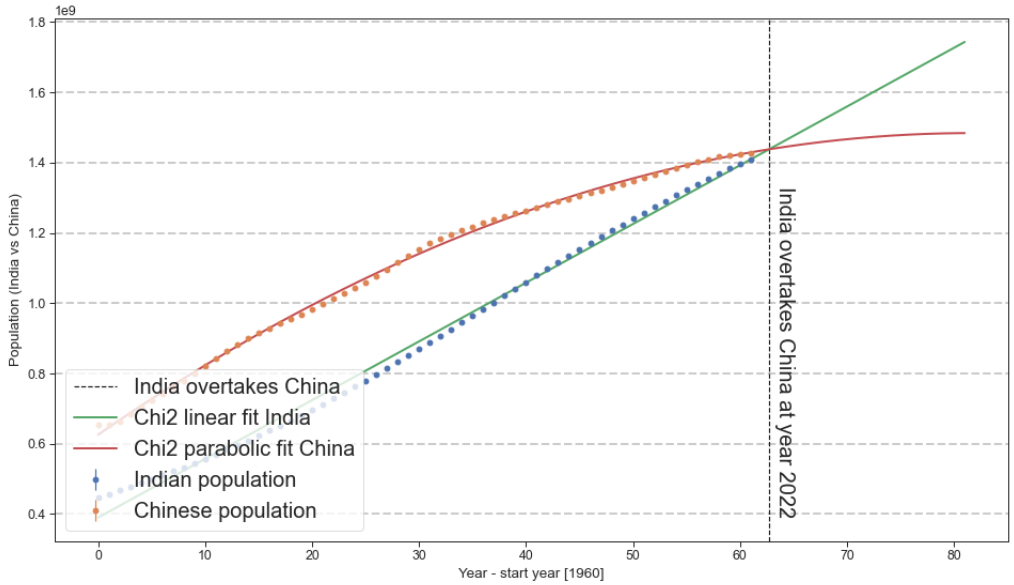

Figure 13: A plot of each population size as a function of time [years]. On top is fitted a first order polynomial for India, and a second order for China. The intersection is marked, and this is where India's populations becomes larger than China's

Some notes about the plot though. I *did* do a χ^2 fit for both of the countries, but the values (p and χ^2) didn't add anything to the goal of finding the intersection between fits. Just as an optimization, i tried fitting with higher order (up to fifth) polynomials, and this *did* reduce the χ^2 value quite a bit, but didn't do much to the p-value. And even worse was, that now my two fits didn't intersect, so the model didn't describe the actual situation very well at higher order polynomials. Therefor i decided to keep to my first and second order poly. and not using the p-value and χ^2 even though they told me, that my specific model wasn't very good at describing my data.

4.2 Medical experiment

This problem can be written as a contingency table and we can then use Fisher's Exact Test on it

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{A!B!C!D!}$$

where A is the entry $1,1$ in the table, B is $1,2$ and so forth.

	Side Effect	No Side Effect	Total
Drug Group	10	14	24
Placebo Group	4	20	24
Total	14	34	48

When doing the test, we have to keep track of the null hypothesis:

H_0: There is no correlation between experiencing a side effect and taking the drug.

With the calculated p-value of 0.043 we must then reject this hypothesis if we set the usual signifiance level to 95%. It would seem like there is some relation between taking the drugs, and experiencing a side effect.

4.3 Smartphone Companies

For finding out weather or not the battery lifetimes are significantly different I used a Two-sample test:

$$z = \frac{\hat{\mu_1} - \hat{\mu_2}}{\sqrt{\hat{\sigma}^2_{\mu,1} + \hat{\sigma}^2_{\mu,2}}}$$

Notice the hats denoting the estimator of μ and σ as we dint have that many data points. Notice as well, that the errors are the standard error on the mean.
Fist of all the null hypothesis:
H_0: There is no significant difference between sample A and sample B
From the z-value i calculated a p-value by evaluating z in the normal normal survival function. I got a p-value of:

$$p = 0.35$$

which mean that with a significance level of 95% we cannot reject our null hypothesis (i.e. the battery lifetime is not significantly different). The p-value states that $\approx 35\%$ of the time, we will get a z-value like this or worse here meaning lower.

5 Fitting data

5.1 A signal has been recorded with phase P, resistance R and frequency ν

5.1.1 Plotting the control sample (first 100000 points), and fitting the peak at $\nu = 1.42GHz$

This was straight forward plotting and fitting. I binned my data with $N_{bins} = 1000$ and a range $\nu \in [0, 8]$GHz and as i had a high count number (above ¿ 5) for most of my data, i assumed Poisson errors on the bins.
For finding the peak, I did two separate χ^2 fits with the functions $f(x)$ and $g(x)$

$$f(x) = \frac{N_{1,f}}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} + N_2 e^{-\frac{x}{\tau_f}}$$

$$g(x) = N_{1,g} e^{-\frac{x}{\tau_g}}$$

where everything expect x is fitting parameters $(N_{1,f}, N_{1,g}, N_2, \mu, \sigma, \tau_f, \tau_g)$

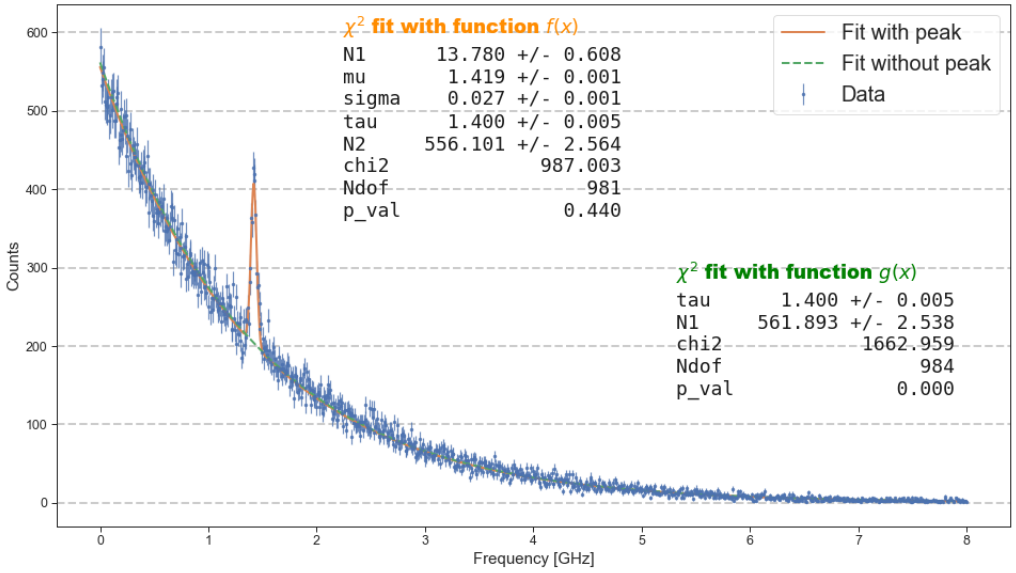

Figure 14: Control data (frequency) with two fits $f(x)$ and $g(x)$. The p-value is way better for $f(x)$, thankfully

As we see from the fit, the p-value is way better for $f(x)$ which we also expected. The peak is therefor significant compared to the rest of the data which can be looked at as background. Looking at the fitted value for μ we see that it indeed matches the given value for $\nu = 1.42$ within one sigma.

5.1.2 Separating signal/noise using P and R

I used Fishers linear discriminant to best separate the data. I wont go over the steps, but my results are plotted below.

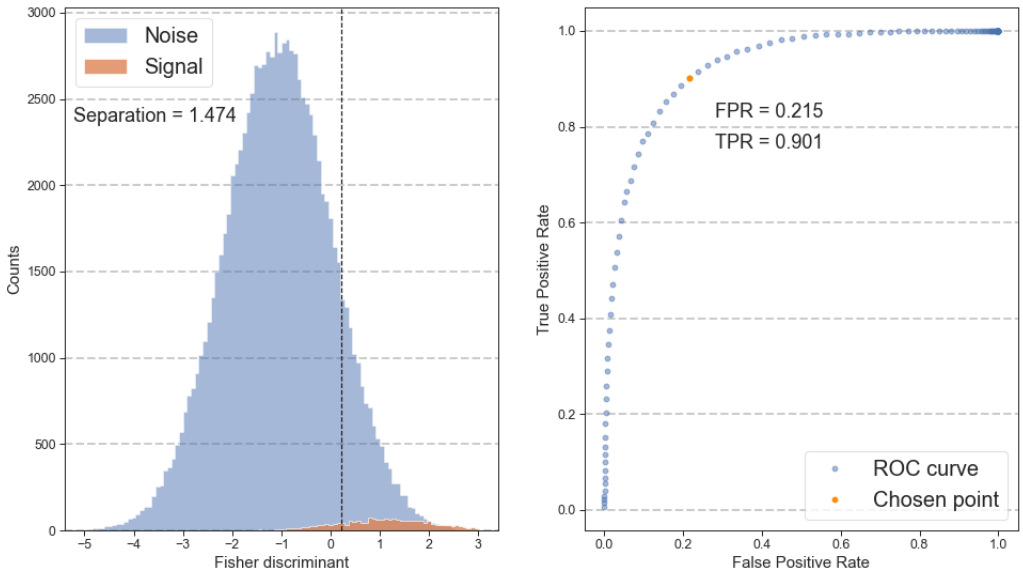

Figure 15: On the left: signal and noise separated as much as possible using Fishers Linear discriminant. On the right: A ROC curve showing TPR and FPR. I chose the point with TPR = 0.901 and FPR = 0.215

On the left plot the separation is shown with a value of 1.474. Though looking at the two histograms, they still don't seem too separated. I used the ROC curve, to decide where i wanted the cut-off to be meaning; how much signal vs how much noise. The TPR here refers to how much of the total signal i end up getting with this cut, and the FPR then is how much of the total noise i get. They are so to speak fractions of the original amount of signal/noise.

5.1.3 Improving the peak

I started of by making a scatter plot of R vs P for both signal and noise. I then decided on some cut off values from this scatter plot, and sliced my Frequency data using these values. This procedure is shown in the following two plots:

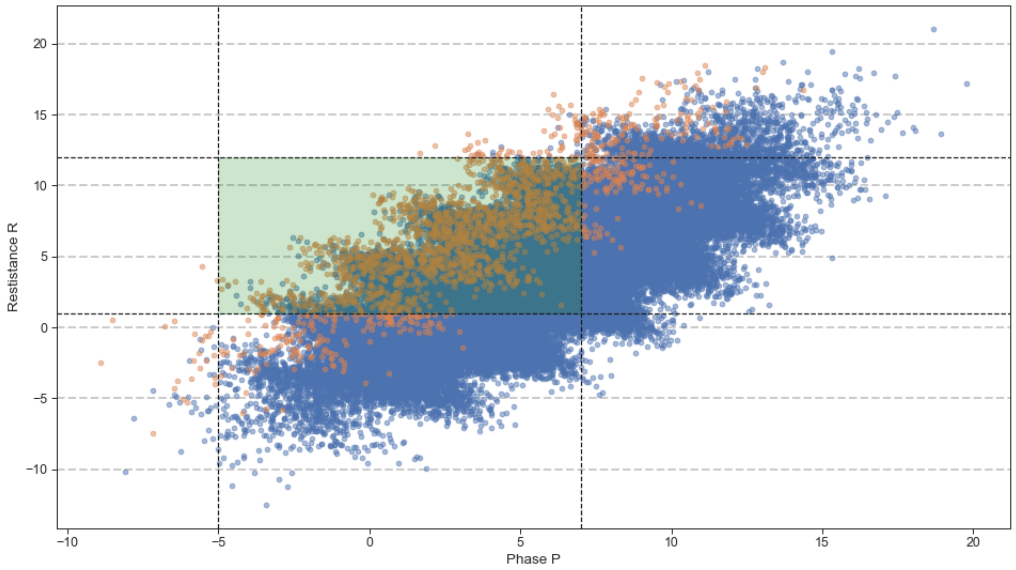

Figure 16: Scatter plot of signal and noise with P on x-axis and R on the y-axis. The chose cut-off values are shown and the green area is where I accepted data

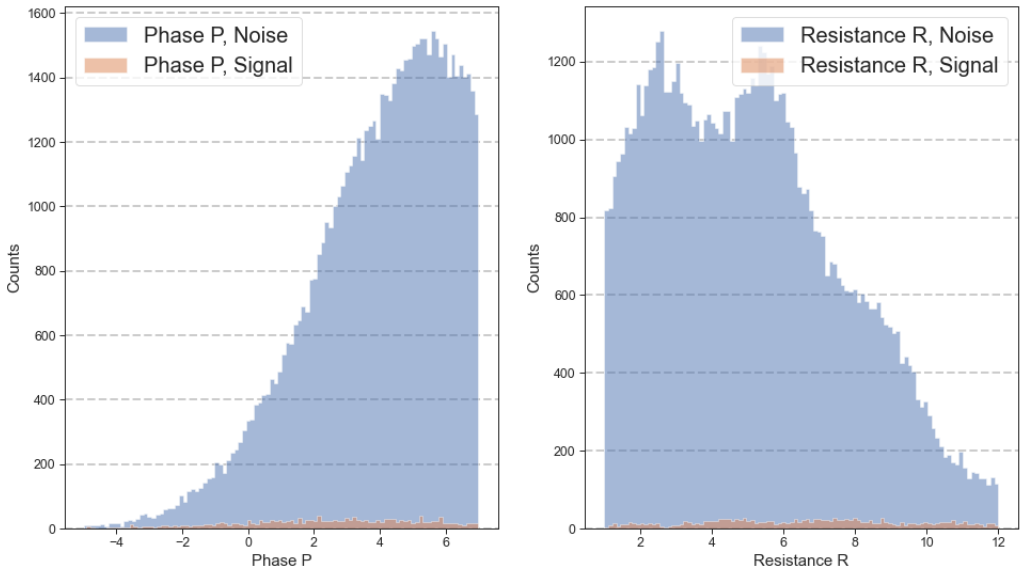

Figure 17: Distributions of noise and signal again. This time each histogram had been sliced with the chosen cut-off values.

I then did exactly what i did in problem 5.1.1 and fitted the sliced frequency data with the

two function $f(x)$ and $g(x)$. I again assume Poisson errors on the bins. I used the same binning and range, but of course removed bins with 0 counts. This is causing the change in N_dof.

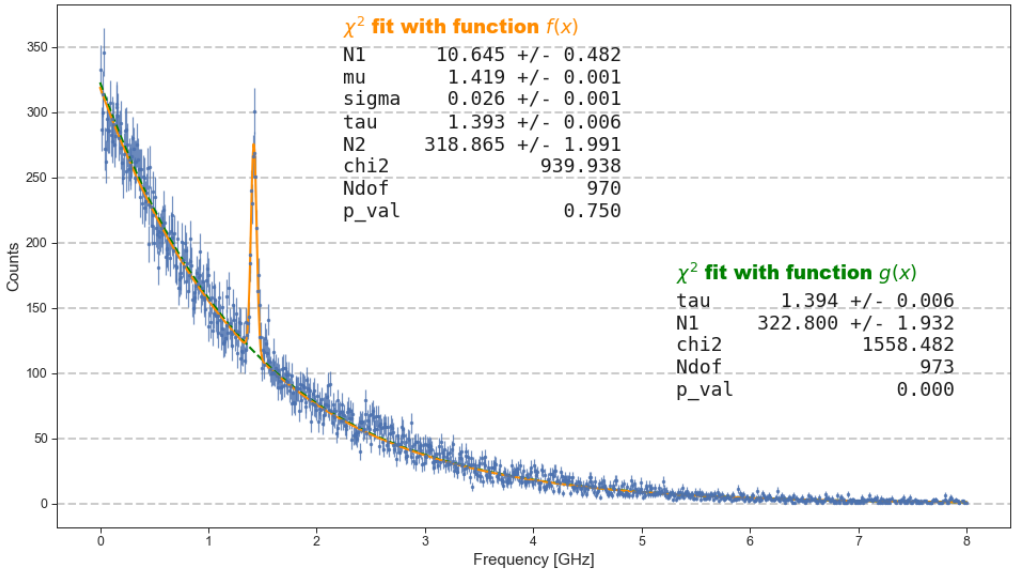

Figure 18: Sliced frequency data with the same two fits. Now the p-value is quite a bit higher than before for $f(x)$ while still remaining 0 for $g(x)$

After slicing the Frequency data using P and R I now got a better p-value for my fit whilst the p-value for $g(x)$ stil remains at 0. So the peak is even more significant now.

5.1.4 Plotting real frequency data and showing peak

Here the plot pretty much speaks for itself. Though I will say, that it was fairly hard even spotting a peak, and it was very dependent on the binning. I ended up using 50 bins in the range [0.1, 1] as specified.

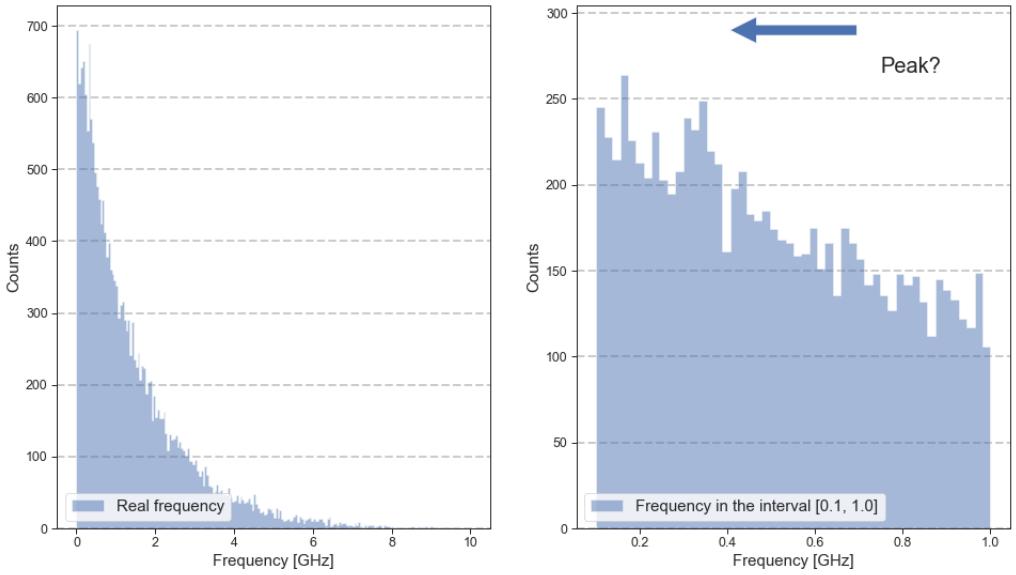

Figure 19: Plot of real frequency data. Left: the whole distribution of data. Right: the specified interval with the peak marked

5.1.5 Estimating signal entries in peak

To estimate entries in the peak, i wanted to see i could fit the peak first. This would visualize which bins were within the peak. I did the exact same fit as in 5.1.1 and 5.1.3 but for the new data. Again i assumed Poisson distributed errors on the bins.

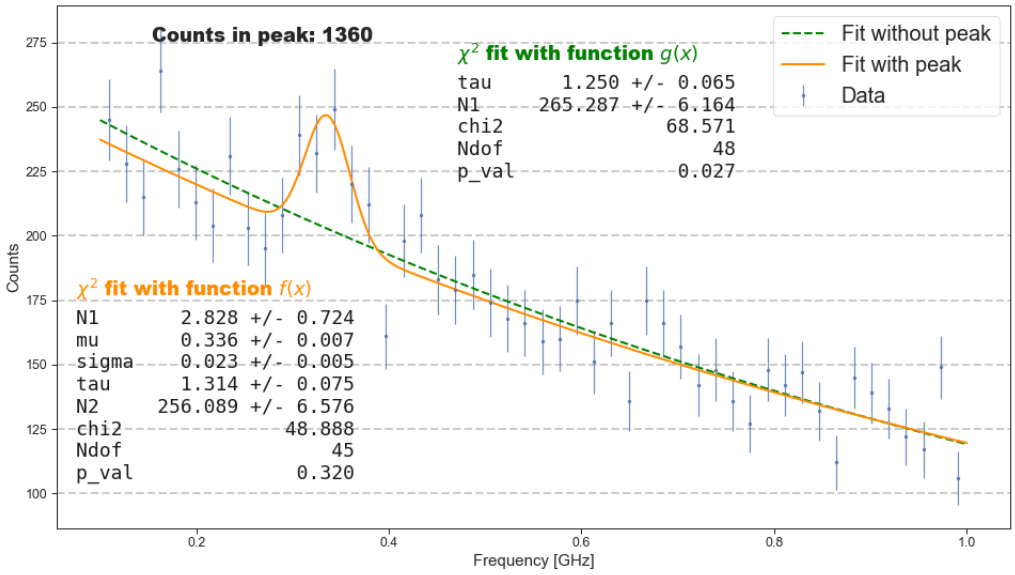

Figure 20: Figure of fitted to the small peak. Again both fitting with $f(x)$ and $g(x)$. An estimate of entries in the peak is printed on the plot as well.

In the figure the two p-values are shown again, only this time, the p-value from $f(x)$ is 0.320 and for $g(x)$ 0.027. At a confidence level of 95% $g(x)$ is still not matching the data, but at least it is closer now. The conclusion was still, that the peak was significant, and I then proceeded to count the number of entries in the peak. This value is also printed in the plot and was found to be

$$N_{entries} = 1360$$

5.1.6 Estimating number of entries in peak if no noise

I am not sure about what is mean with P and R but i used the fit of $g(x)$ to estimate how many counts there would be in each bin, if there wasn't a peak. This of course is only reliable, if the whole exponential decay could be assumed to background which might not be the case. Nevertheless I then subtracted entries in the peak bins estimated from $g(x)$ to the estimated number of entries from last problem. This gave me a total of

$$peak_{signal} = 141\text{entries}$$

but again. Not all of the assumed background might be background...

5.2 Bohrium isotope decay

5.2.1 Plotting the decay and calculating the mean and the median with errors

I used the given pandas.dataframe to compute the mean, standard, deviation, and median. I then used the classic formula for the SEM to get the uncertainty on the mean (surprise). For the error on the median, i used a bootstrapping method. I drew as many numbers as there was entries in the data set, and i drew them from the original dataset. I then found the median for this new distribution, and i did this 1000 times. I then finally calculated the standard deviation of all the medians, and used this as the error on the median. This gave me the results:

$$\mu = 0.963 \pm 0.030$$

$$median = 0.700 \pm 0.030$$

The fact that the errors are the same, was a surprise to me, but i realized that this is a consequence of a large sample size, and some Gaussianity in the dataset.

The plot of the data is shown below, with the mean and the median plotted on top with their errors respectfully.

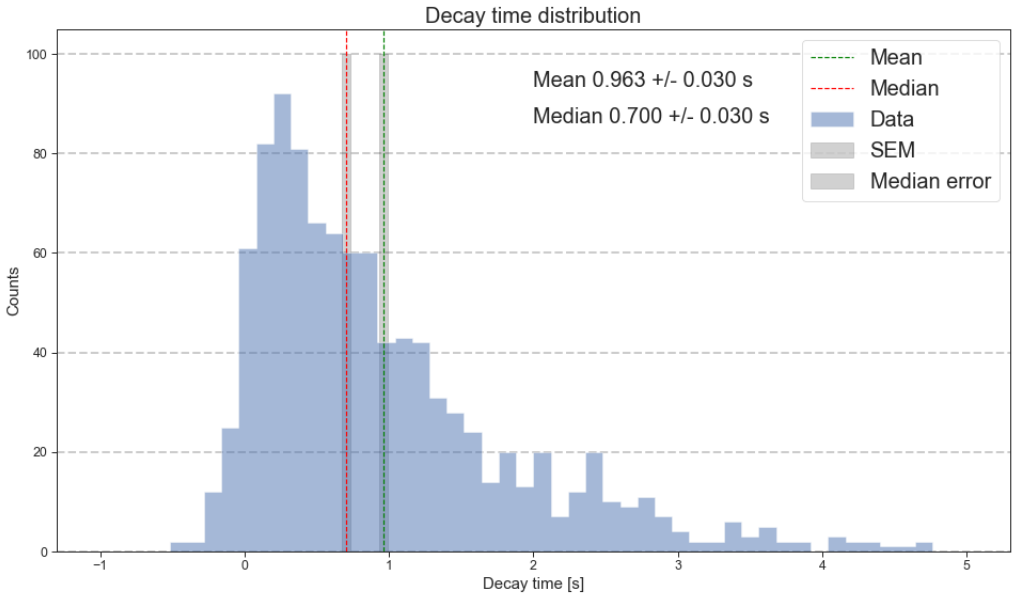

Figure 21: Plot of decay times for the Bohrium isotope. On top is shown the median and the mean with their errors.

5.2.2 A rough estimate on τ using the tail at high t

Fitting the tail I used classic exponential decay function and i defined late t as t ¿ 1.5. I tried using both Unbinned Likelihood and a χ^2 fit, but the UBLH had some problems. I have shown them both in the figure below, so you can judge for yourself. I think what happened was, that there originally is a lot of outliers far out in the tail. This skews the UBLH fit, as it takes every point into account. The same is not the case for the χ^2 fit, where the data is binned before fitting. Here all the outliers are binned into the final bin making it easier to fit despite the low statistics. Because there is certainly low statistics for many of the bins, and thus in that case i shouldn't assume Poisson errors and do the χ^2 fit. But the plot below just goes to show the power of the χ^2 as even with low statistics we get something reasonable.

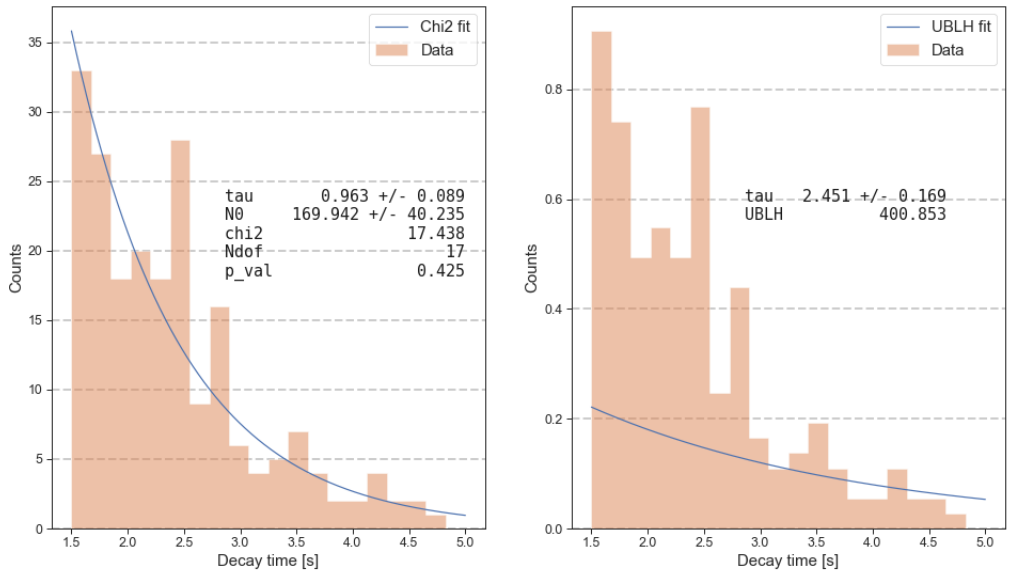

Figure 22: Left: χ^2 fit with the standard exponential decay function for estimating τ. Right: the UBLH fit with a skew towards outliers in the tail.

It is pretty obvious which of the two fits i will use to estimate τ from. And so i get an estimate

$$\tau = 0.963 \pm 0.089$$

5.2.3 Fitting to the whole distribution and determining τ and σ

I tried fitting the distribution to four different functions that i won't write here for the sake of reading. In summary it was a Gauss, an exponential, a mix of both, and a double Gauss with an exponential. Bottom line is: none of them gave anything significant, and didn't help estimating τ and σ. The fits are shown below, and their relative chi^2 values and p-values are shown in the plot too. I will not comment on them, as they are all insignificant, though

i will say, that the more parameters i ended up using, the lower the χ^2 value, which points towards some nasty underlying function.

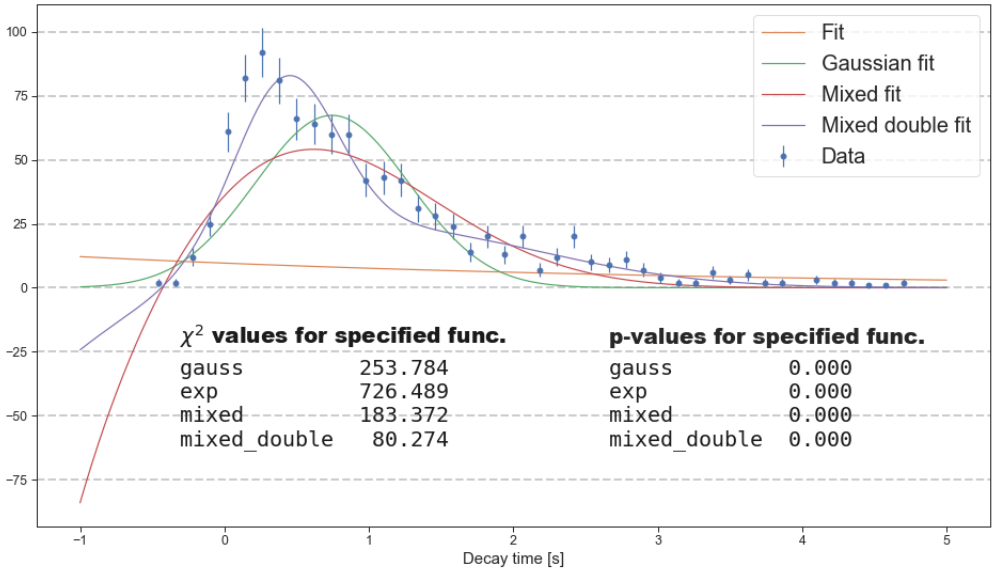

Figure 23: Several fits attempts with related fitting values. None of them are helpful in estimating τ and σ.

To me this distribution look very much like a Poisson distribution. This also makes sense when talking about decay and rates, but Poisson distributions are only for discrete events, and the times we have, are not discrete.