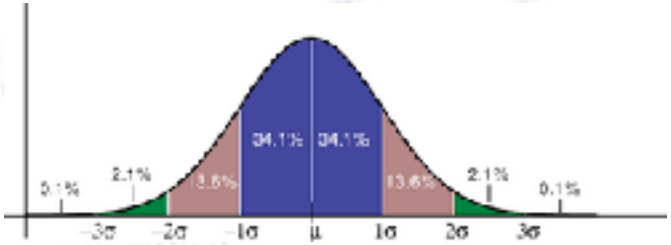
Applied Statistics

Problem Set Solution and Discussion






Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

Overall comments

The problem set is hard!

The problem set is hard, and this one was no exception. If anything, on the contrary.

So if you had a hard time, then there should be no surprise. But the point of the problem set is of course also to give problems, so that every student will be challenged. This problem set (also) managed that…

It closely resembles what to expect for the exam, so you should be well prepared by now.

The solutions

Problem 1.1

1.1 (6 points) An electronic device depends on three components each with independent probabilities 0.009, 0.016, and 0.027 of failing per year.

- What is the probability that the device will **not** fail in the first year?
- After how many years is the probability of failure greater than 50%?

1.1.1: Since the probabilities are independent (no correlation), the probability of the device not failing the first year is the multiplication of the three probabilities.

But since the probabilities take into account the failing probability we need to do $P(\textit{success first year}) = (1 - P(1)) * (1 - P(2)) * (1 - P(3)) = 0.9488$.

1.1.2: To compute the amount of years N it needs to get a probability of failure (P) greater than 50%, we can use $P = P^{N_{years}} < 0.5$. **This gives $N_{years} > 13.19$**.
If an integer number is required (it is not), the answer is 14 years.

Problem 1.1

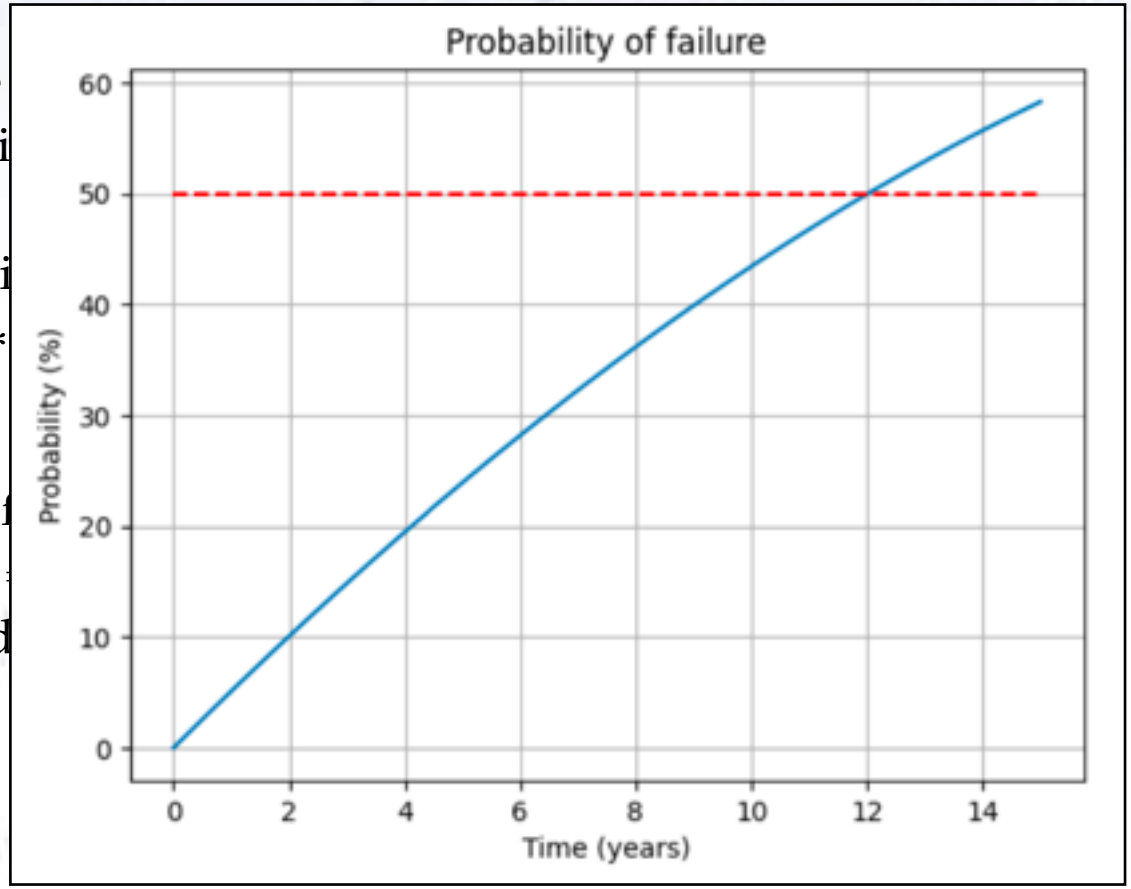
1.1 (6 points) An electronic device depends on three components each with independent probabilities 0.009, 0.016, and 0.027 of failing per year.

- What is the probability that the device will **not** fail in the first year?
- After how many years is the probability of failure greater than 50%?

1.1.1: Since the probabilities are
device not failing the first year i

But since the probabilities take i
$P(\textit{success first year}) = (1 - P(1)) *$

1.1.2: To compute the amount of
greater than 50%, we can use P
If an integer number is required



Problem 1.2

1.2 (8 points) A store has 52.8 customers/day, and considers the top 20% busiest days to be... busy!

- What distribution should the number of daily customers follow and why?
- Discuss what number of customers exactly constitutes a busy day.
- What is the average number of customers on a busy day?

1.2.1: It follows a **Poisson distribution**, as it is a suitable distribution for N being high and p small. Also, a Poisson describes a rate and the number of costumers is integer.

1.2.2: Using the Poisson distribution with $\lambda = 52.8$, we can search for which number the probability or area under the curve is 20%. Doing that we get that a a busy day is when we have 58.68, so **from 59 or more customers, and some portion of 58**.
It should be discussed (explicitly) how this does not fit 80% (but above at ~82.3%).

1.2.3: One takes the "complicated" average, either by formula or simulation.

$$\lambda_{busy} = \frac{\sum_{n=59}^{\infty} n \cdot \frac{\lambda^n}{n!} e^{-\lambda}}{1 - \sum_{n=0}^{58} \frac{\lambda^n}{n!} e^{-\lambda}} = \frac{1 - \sum_{n=0}^{58} n \cdot \frac{\lambda^n}{n!} e^{-\lambda}}{1 - \sum_{n=0}^{58} \frac{\lambda^n}{n!} e^{-\lambda}} \approx 62.9 \text{customers/day}$$

Problem 1.2

1.2 (8 points) A store has 52.8 customers/day, and considers the top 20% busiest days to be... busy!

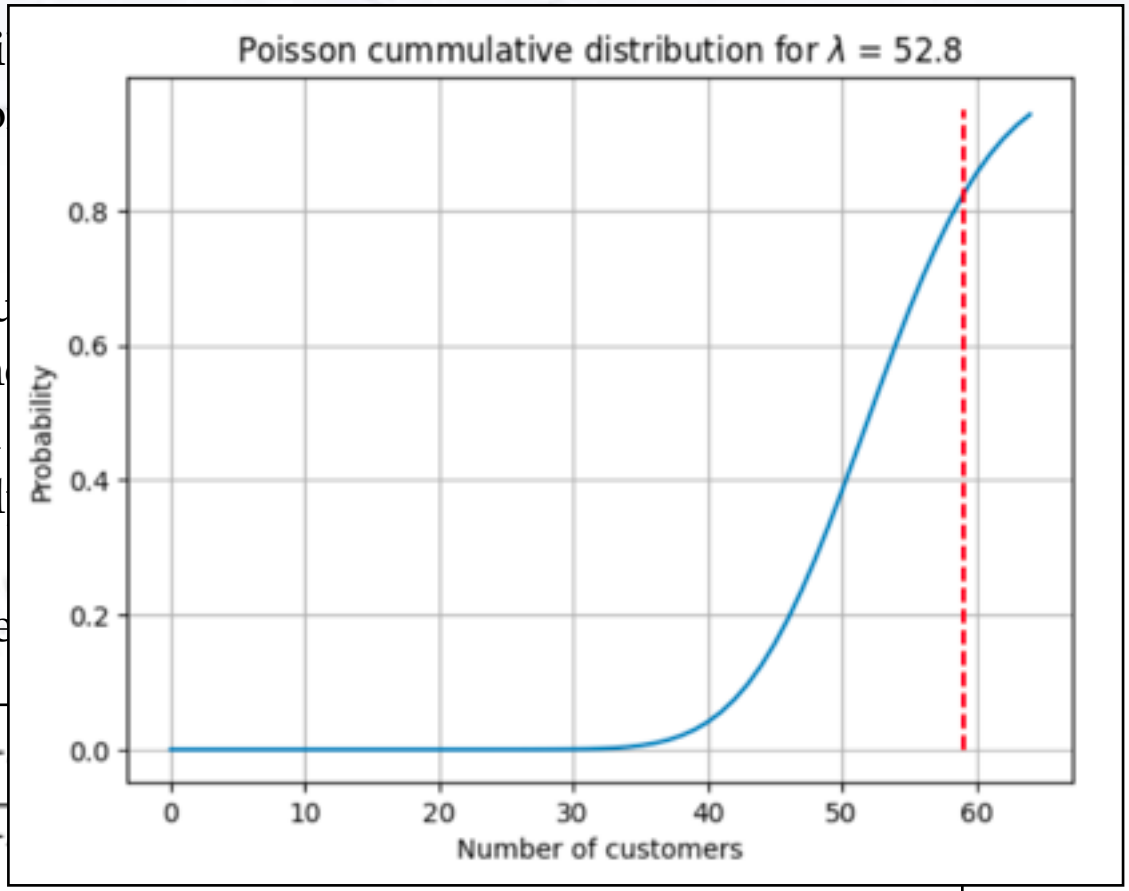
- What distribution should the number of daily customers follow and why?
- Discuss what number of customers exactly constitutes a busy day.
- What is the average number of customers on a busy day?

1.2.1: It follows a **Poisson distri**
high and p small. Also, a Poisso
integer.

1.2.2: Using the Poisson distribu
the probability or area under th
is when we have 58.68, so **from**
It should be discussed (explicitl

1.2.3: One takes the "complicate

$$\lambda_{busy} = \frac{\sum_{n=59}^{\infty} n \cdot \frac{\lambda^n}{n!} e^-}{1 - \sum_{n=0}^{58} \frac{\lambda^n}{n!} e^-}$$



Problem 2.1

II – Error propagation:

2.1 (10 points) You make nine measurements of the speed of sound in water, and obtain as follows:

Speed of sound (in m/s)	1532	1458	1499	1394	1432	1565	1474	1440	1507
Uncertainty (in m/s)	67	55	74	129	84	19	10	17	14

- What is the combined result and uncertainty of all your measurements?
- How much does adding the first five measurements improve the precision compared to the last four?
- Are your measurements consistent with each other? If not, argue for an updated estimate.
- The speed of sound in water is 1481m/s. Does your result agree with this value?

2.1.1: A weighted average gives **1488 ± 7 m/s**.
You should of course test this with a Chi2: Prob(chi2=29.9, Ndof=8) = 0.00002
Thus, **the values are not compatible**.

2.1.2: Using the first five measurements gives **1476 ± 33 *m*/*s***,
while the last four measurements gives **1488 ± 7 *m*/*s*.**
So **the first five measurements do not significantly improve the result**.

2.1.3: If we remove the sixth measurement we get Prob(χ^2=11.2, N_{dof}= 7), ***p* = 0.202**.

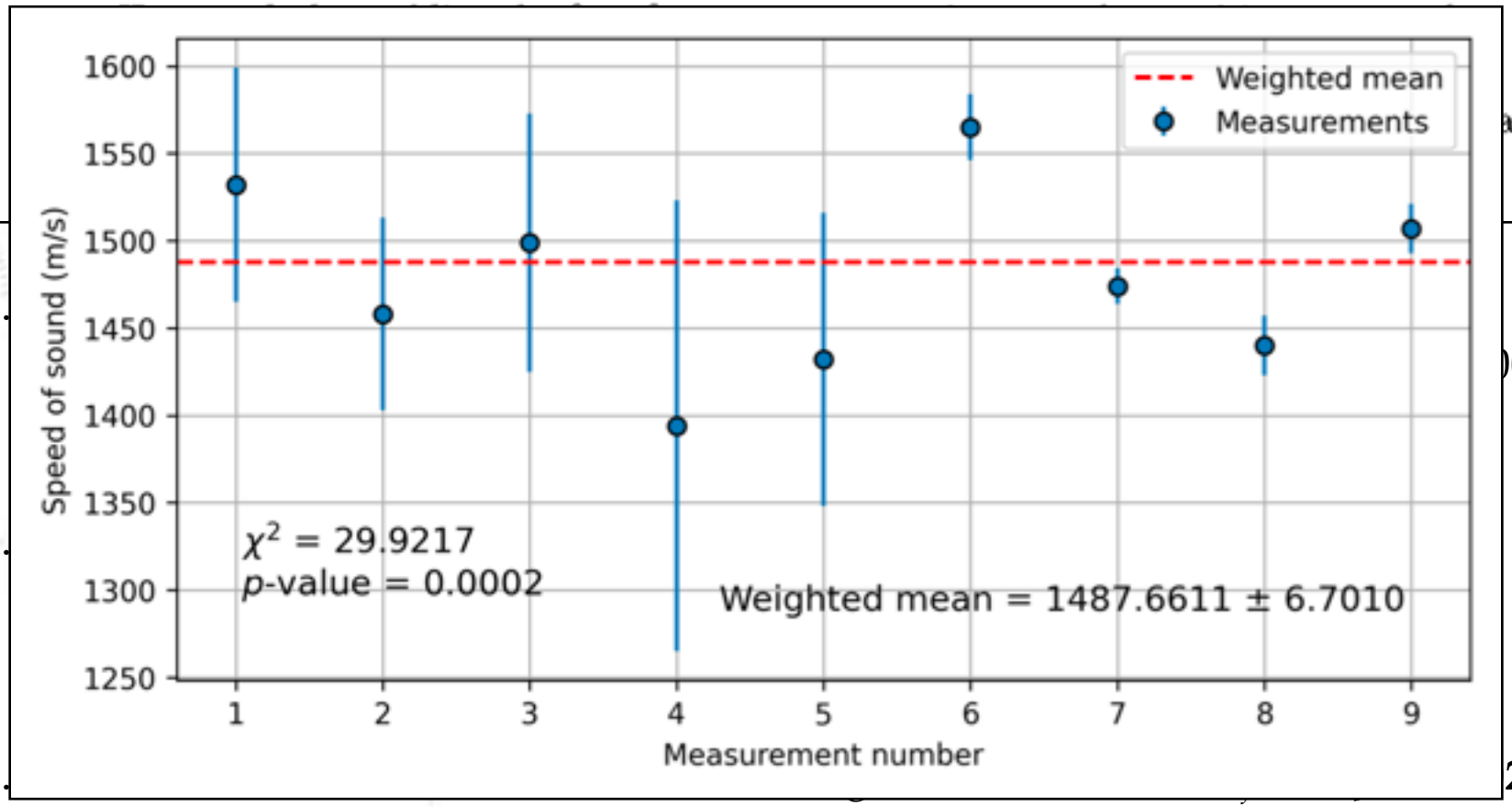
Problem 2.1

II – Error propagation:

2.1 (10 points) You make nine measurements of the speed of sound in water, and obtain as follows:

Speed of sound (in m/s)	1532	1458	1499	1394	1432	1565	1474	1440	1507
Uncertainty (in m/s)	67	55	74	129	84	19	10	17	14

- What is the combined result and uncertainty of all your measurements?



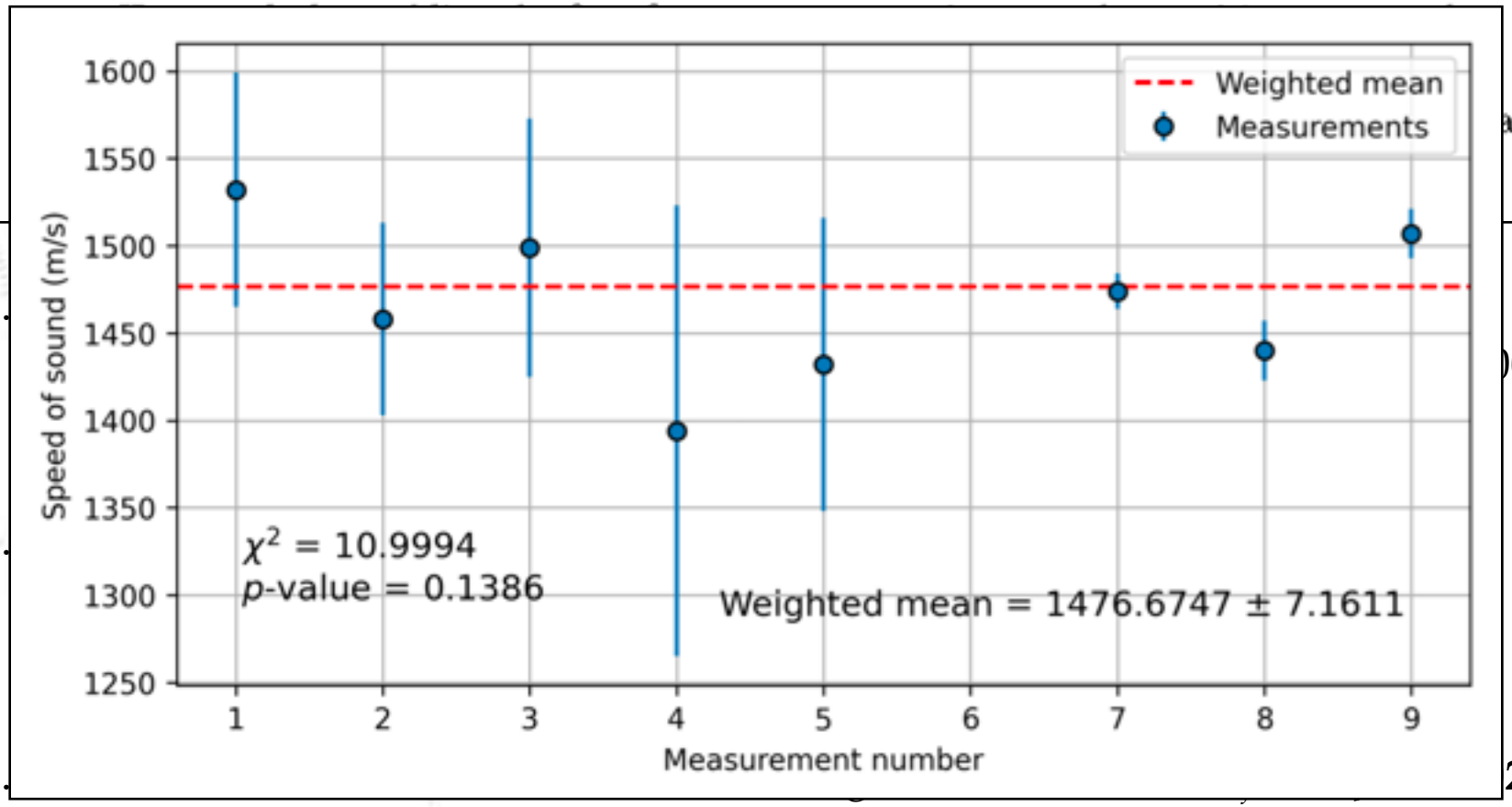
Problem 2.1

II – Error propagation:

2.1 (10 points) You make nine measurements of the speed of sound in water, and obtain as follows:

Speed of sound (in m/s)	1532	1458	1499	1394	1432	1565	1474	1440	1507
Uncertainty (in m/s)	67	55	74	129	84	19	10	17	14

- What is the combined result and uncertainty of all your measurements?



Problem 2.1

II – Error propagation:

2.1 (10 points) You make nine measurements of the speed of sound in water, and obtain as follows:

Speed of sound (in m/s)	1532	1458	1499	1394	1432	1565	1474	1440	1507
Uncertainty (in m/s)	67	55	74	129	84	19	10	17	14

- What is the combined result and uncertainty of all your measurements?
- How much does adding the first five measurements improve the precision compared to the last four?
- Are your measurements consistent with each other? If not, argue for an updated estimate.
- The speed of sound in water is 1481m/s. Does your result agree with this value?

2.1.4: We can compare the results using the z-score. Using all measurements we get $z = 0.99$, while removing the sixth measurement we get $z = 0.60$, in both cases showing consistency from our result to the real speed.

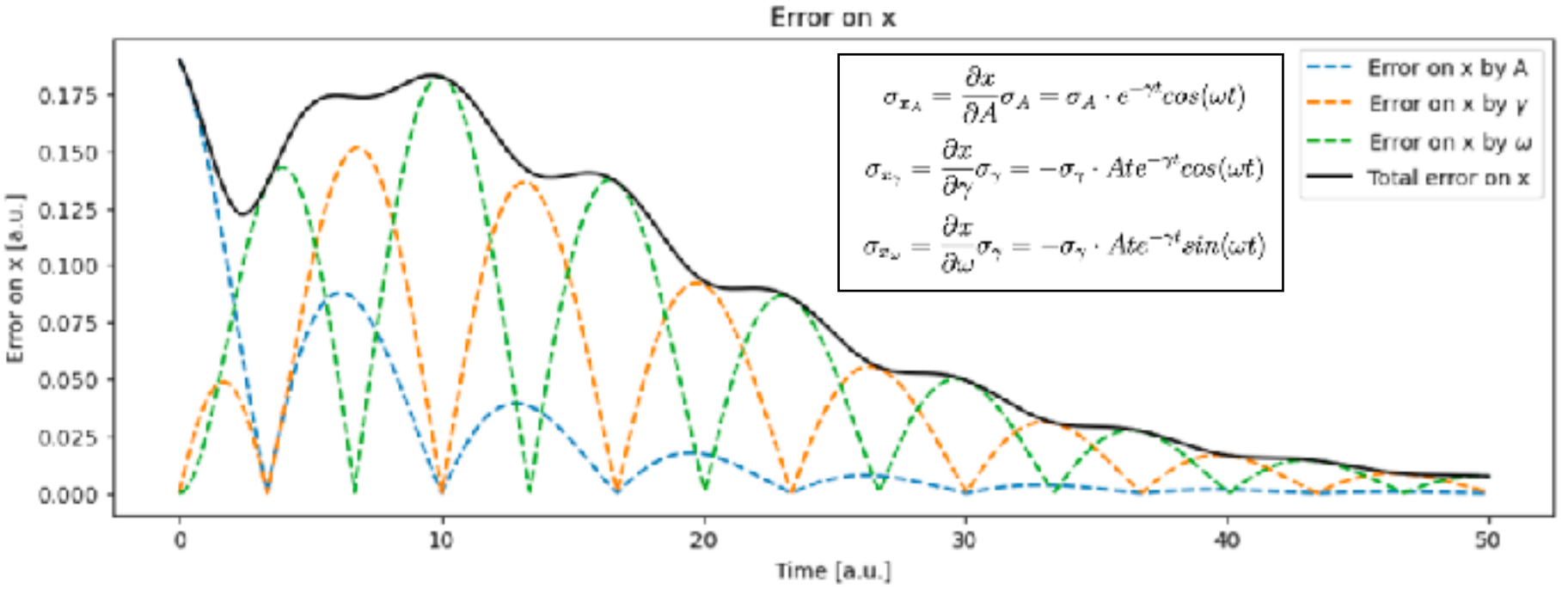
Problem 2.2

2.2 (8 points) A mass is moving in a damped harmonic oscillator with position $x(t) = A\exp(-\gamma t)\cos(\omega t)$ as a function of time t, where $A = 1.01 \pm 0.19$, $\gamma = 0.12 \pm 0.05$, and $\omega = 0.47 \pm 0.06$.

- At $t = 1$, calculate the position and its uncertainty in x position.
- Calculate the uncertainty in x as a function of t for each of the three variables, and comment on which variables dominate the uncertianty during which periods in time.

2.2.1: Doing the error propagation, we get x(1) = 0.80 ± 0.16.

In the very beginning, it is the uncertainty in the Amplitude that dominates.



Problem 3.1

3.1 (10 points) You shoot a penalty, and the probability of scoring depends on the position x (in m) you hit, as $p_{\text{score}} = |x|/4$ m for $|x| < 4$ m and zero otherwise (outside goal). Assume the ball hits the goal where you aim with an uncertainty of one meter.

- What is the chance of scoring, if you aim at $x = 2.5$m?
- Where should you aim to have the highest probability of scoring?

This fun problem was conceived by Mathias, and is "near perfect for simulation", though it can in fact also be solved analytically (with erfc function).

3.1.1: Define a specific aim (2.5m), and compute the average of probabilities:
Aiming at x = 2.5 m, we get p = 0.551.

3.1.2: Now we repeat the process for all aims (also outside goal).
The curve should be lowest (but > 0) in the middle and symmetric outwards, dropping off at the ends.

$$p_{goal}(x) = \int_{-4}^{4} \frac{|x|}{4\sqrt{2\pi}} e^{\frac{1}{2}(x'-x)^2} dx$$

Problem 3.1

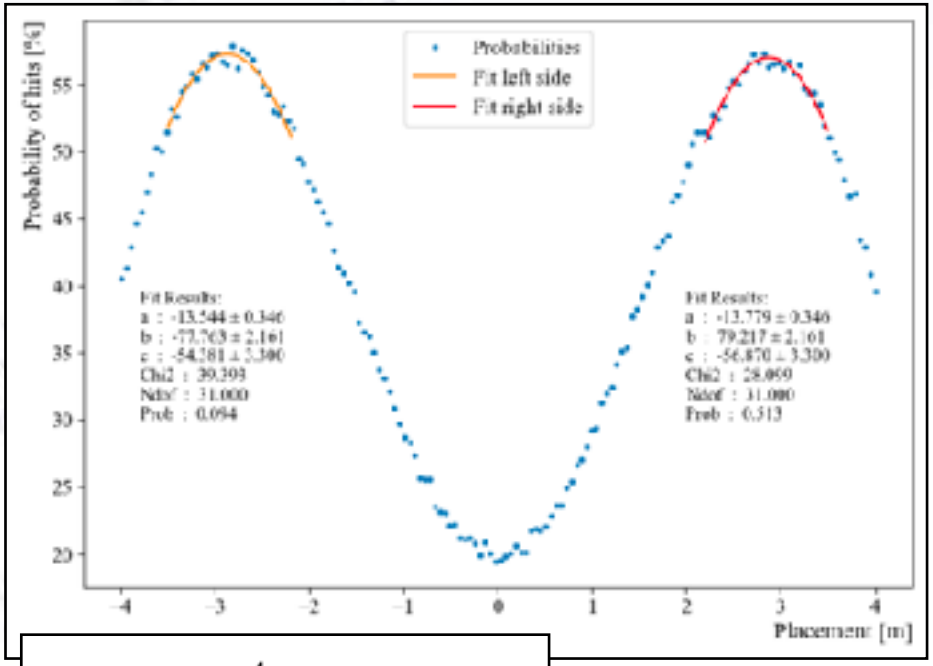
3.1 (10 points) You shoot a penalty, and the probability of scoring depends on the position x (in m) you hit, as $p_{\text{score}} = |x|/4$ m for $|x| < 4$ m and zero otherwise (outside goal). Assume the ball hits the goal where you aim with an uncertainty of one meter.

- What is the chance of scoring, if you aim at $x = 2.5$m?
- Where should you aim to have the highest probability of scoring?

This fun problem was conceived by Mathias, and is "near perfect for simulation", though it can in fact also be solved analytically (with erfc function).

3.1.1: Define a specific aim (2.5m), and compute the average of probabilities:
Aiming at x = 2.5 m, we get p = 0.551.

3.1.2: Now we repeat the process for all aims (also outside goal).
The curve should be lowest (but > 0) in the middle and symmetric outwards, dropping off at the ends.



$$p_{goal}(x) = \int_{-4}^{4} \frac{|x|}{4\sqrt{2\pi}} e^{\frac{1}{2}(x'-x)^2} dx$$

Problem 3.2

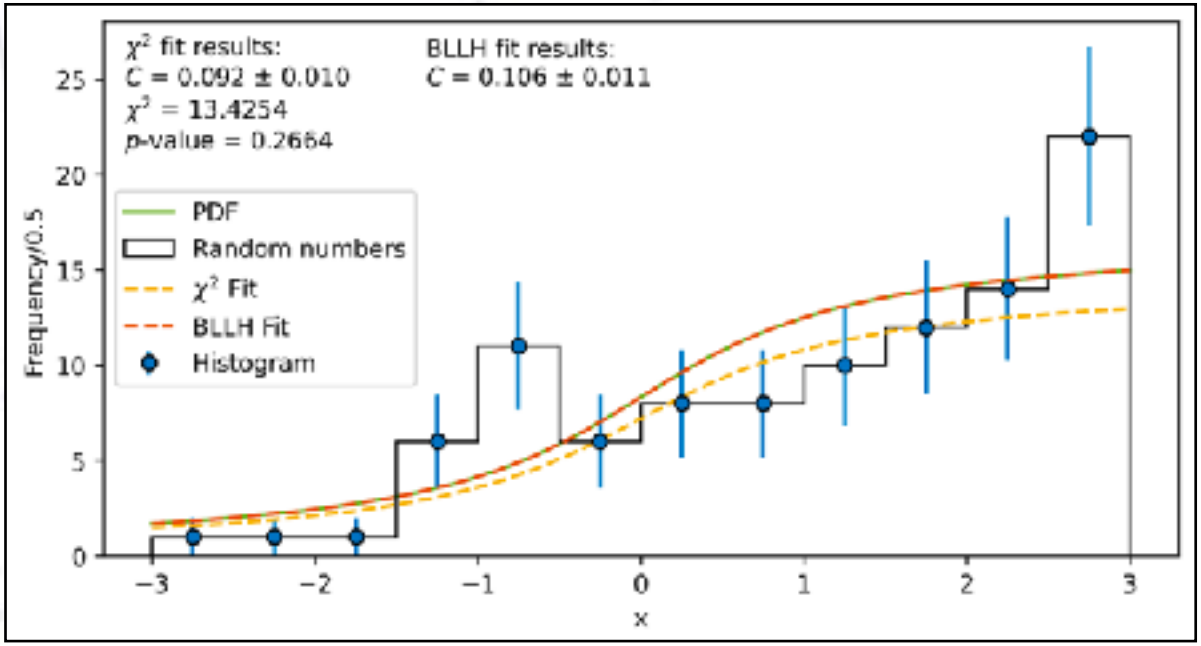
3.2 (10 points) Consider the PDF $f(x) = C_{\text{PDF}}(tan^{-1}(x) + \pi/2)$ with $x \in [-3, 3]$.

- Determine C_{PDF} and generate 100 random numbers following $f(x)$.
- Explain how you would fit these data and do so. Does your fit values for C match C_{PDF}?

3.2.1: In this case, the transformation method doesn't works (integral not easily investable), while accept-reject works well. C = 1 / (3π) = 0.106 from integrating.

3.2.2: The data is low statistics, and should thus be fitted with a (binned) LLH fit. Doing a ChiSquare fit gives a biased result, though not entirely wrong.

Of course you know the PDF, and you should let C be a fit parameter, and get the same value.

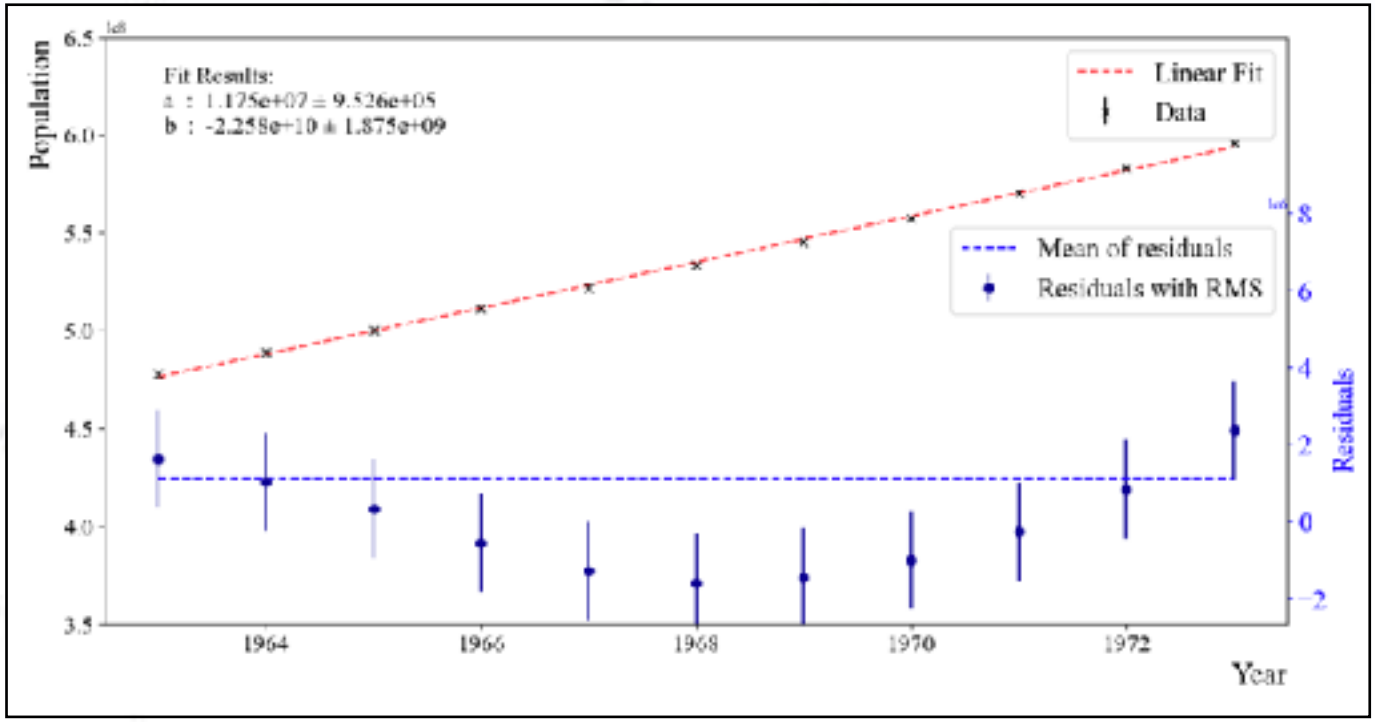


Problem 4.1

4.1 (10 points) The file **www.nbi.dk/~petersen/data_LargestPopulation.csv** contains data on the Indian and Chinese population each year in the period 1960-2021.

- **Linearly fit the Indian population 1963-1973, and estimate the data point uncertainty.**
- **Assuming an uncertainty of ±1000000 on all data points, model the population developments and give your best estimate of when the Indian population overtakes the Chinese.**

4.1.1: This is a simple linear fit, letting the residuals define the Std. on points.
It is probably an overestimate, as their is a clear pattern (pol2 better!).



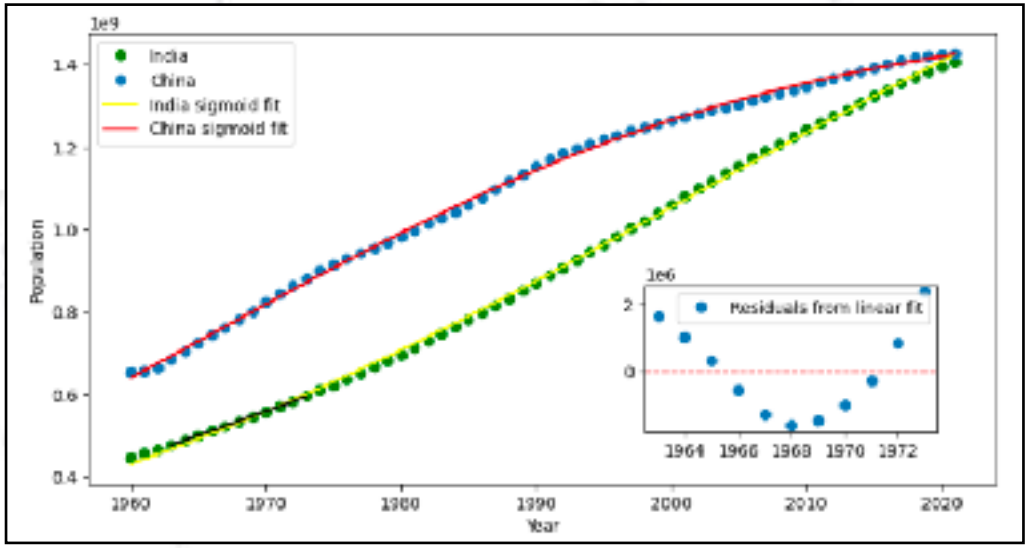
Problem 4.1

4.1 (10 points) The file **www.nbi.dk/~petersen/data_LargestPopulation.csv** contains data on the Indian and Chinese population each year in the period 1960-2021.

- Linearly fit the Indian population 1963-1973, and estimate the data point uncertainty.
- Assuming an uncertainty of ± 1000000 on all data points, model the population developments and give your best estimate of when the Indian population overtakes the Chinese.

4.1.1: This is a simple linear fit, letting the residuals define the Std. on points.
It is probably an overestimate, as their is a clear pattern (pol2 better!).

4.1.2: There is no "requirement" that all data is taken into account! Rather, it is important to focus on the most recent data, and get this part right.



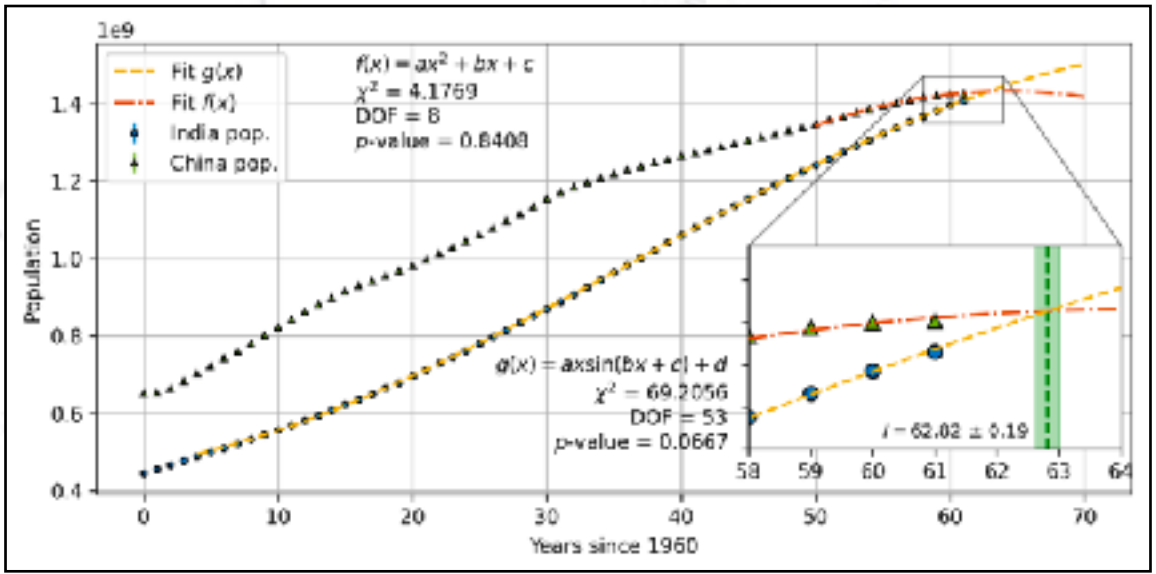
Problem 4.1

4.1 (10 points) The file **www.nbi.dk/~petersen/data_LargestPopulation.csv** contains data on the Indian and Chinese population each year in the period 1960-2021.

- **Linearly fit the Indian population 1963-1973, and estimate the data point uncertainty.**
- **Assuming an uncertainty of ±1000000 on all data points, model the population developments and give your best estimate of when the Indian population overtakes the Chinese.**

4.1.1: This is a simple linear fit, letting the residuals define the Std. on points.
It is probably an overestimate, as their is a clear pattern (pol2 better!).

4.1.2: There is no "requirement" that all data is taken into account! Rather, it is important to focus on the most recent data, and get this part right.

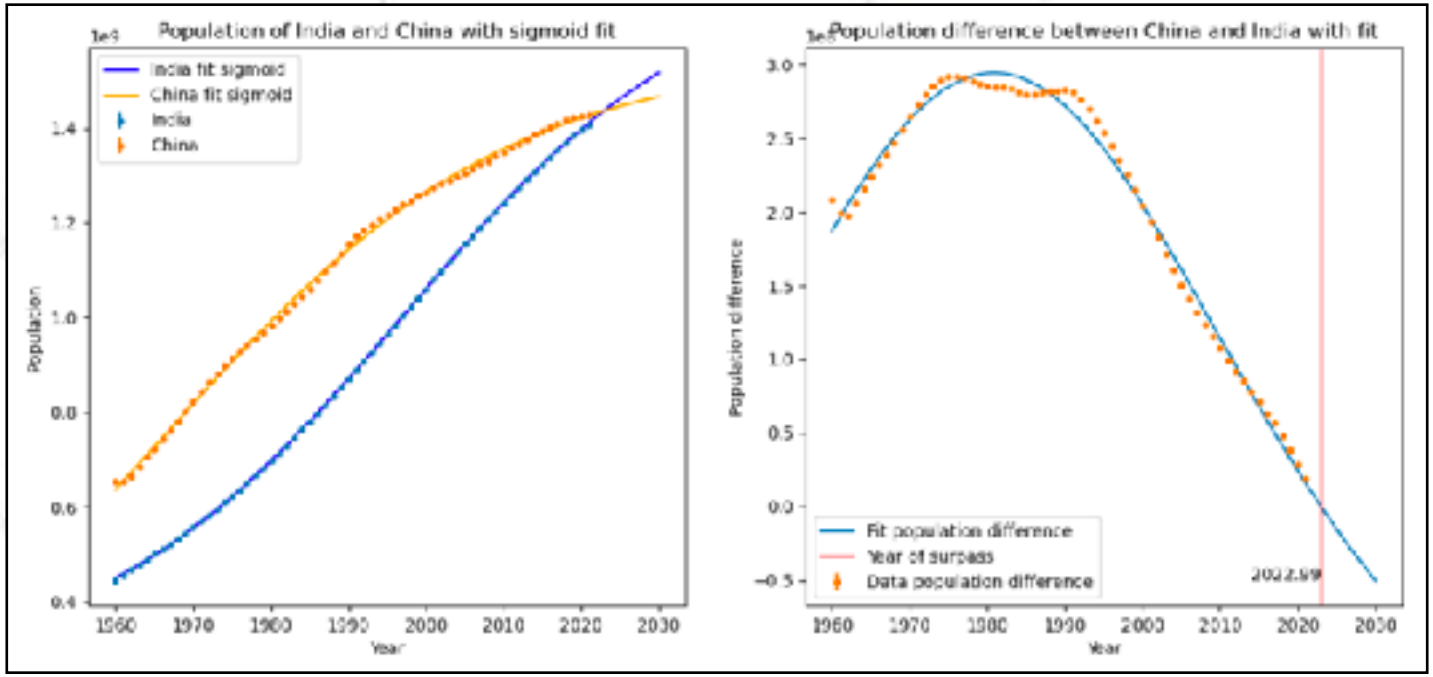


Problem 4.1

4.1 (10 points) The file **www.nbi.dk/~petersen/data_LargestPopulation.csv** contains data on the Indian and Chinese population each year in the period 1960-2021.

- Linearly fit the Indian population 1963-1973, and estimate the data point uncertainty.
- Assuming an uncertainty of ± 1000000 on all data points, model the population developments and give your best estimate of when the Indian population overtakes the Chinese.

4.1.2: There is no "requirement" that all data is taken into account! Rather, it is important to focus on the most recent data, and get this part right.
The best solution is not to fit two populations, but only their difference!



Problem 4.2

4.2 (5 points) A medical experiment is testing if a drug has a specific side effect. Out of 24 persons taking the drug, 10 had the side effect. For 24 other persons getting a placebo, only 5 had the side effect. Would you claim that the drug has this side effect?

4.2.1: For this problem, one should use the Fisher's Exact Test, as the data is a contingency table of 2x2.

The result is $p = 0.076$, thus we can NOT claim that the drug has a side effect, even if it is more likely than not.

	Drug	Placebo	Row total
Side effect	10	5	15
No side effect	14	19	33
Column total	24	24	48

A rare alternative is to use the ChiSquare, but the numbers are too low here.

Problem 4.3

4.3 (5 points) Smartphone producer claims that their phones (A) have a battery lifetime that is significantly longer than that of a rival phone (B). You measure the lifetime of the batteries (in hours) five times for each brand (table below). Test if the claim is reasonable.

A:	28.9	26.4	22.8	27.3	25.9	B:	22.4	21.3	25.1	24.8	22.5

4.3.1: This is a classic hypothesis test. As the data is low statistics, a t-test (instead of a z-test) is in order.
The result is a test statistic $t = 2.44$ and a corresponding $p = 0.0407$.
Thus here, A could possibly claim, that their battery lifetime was longer.

An alternative solution is the K-S test, though the result is two-sided.

Problem 5.1

5.1 (18 points) The file **www.nbi.dk/~petersen/data_SignalDetection.csv** contains 120000 entries with values of measured phase (P), resonance (R), frequency (ν), and type (signal/noise). In the first 100000 entries (control sample) it is known if the measurements are signal (1) or noise (0). In the last 20000 entries (real sample) this is unknown.

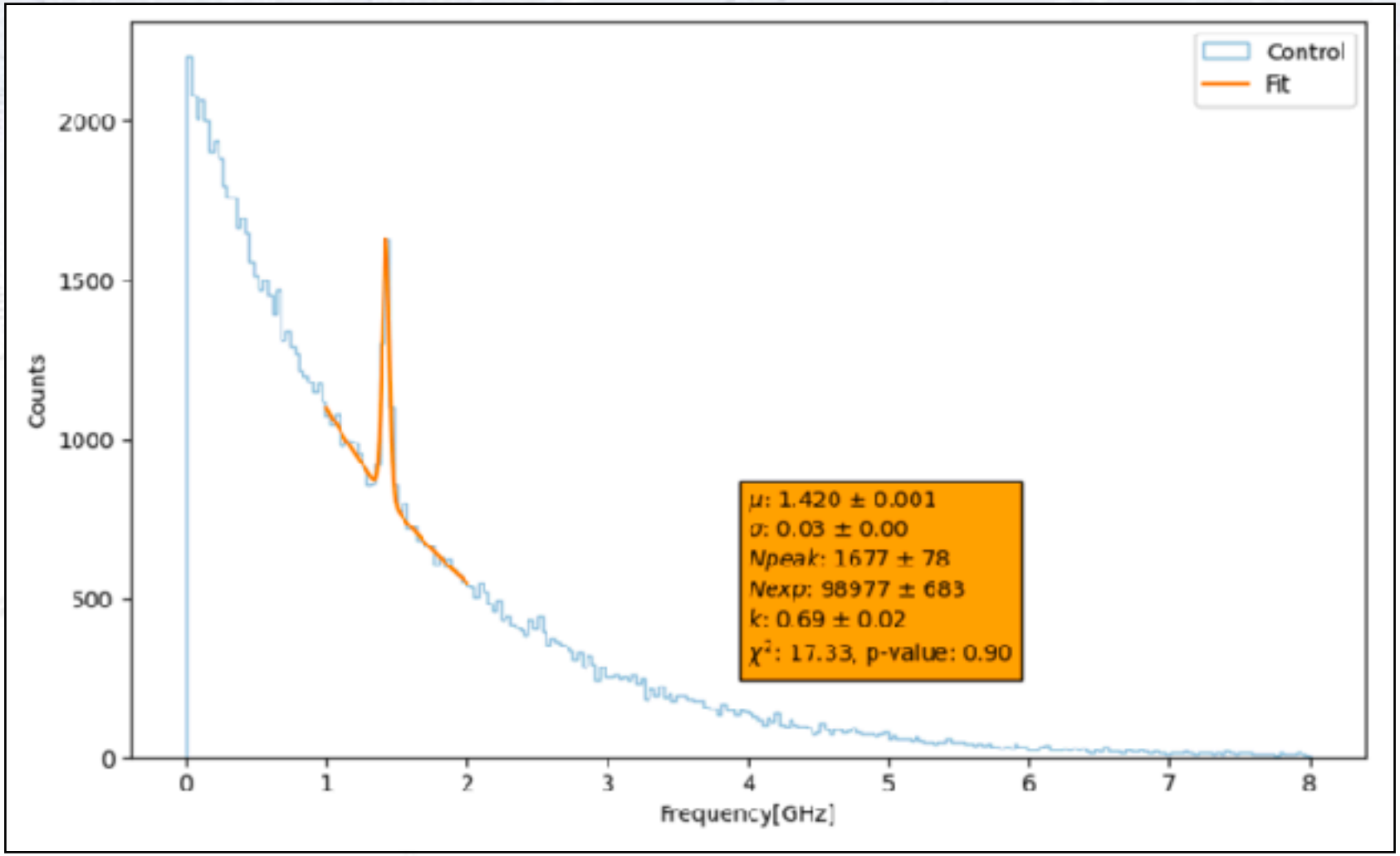
- Plot the control sample frequency distribution. Fit the observed H-peak at $\nu = 1.42$ GHz.
- Quantify how well you can separate signal from noise using the variables P and R.
- Selecting entries based only on P and R, how significant can you get the H-peak fit to be?
- Plot the real data frequency distribution, and search for a peak in the range [0.1,1.0] GHz.
- How many signal entries do you estimate there to be in the peak? Do you find it significant?
- Correcting for the signal selection efficiency when selecting events baseed on P and R, how many signal entries do you estimate there was in the data originally?

5.1: This is the largest problem in the set, and one should thus try an attempt at a quick-and-dirty solution early. **Start by plotting each part of the data set**.

Statistics is high, so ChiSquare fits are in order all the way through.

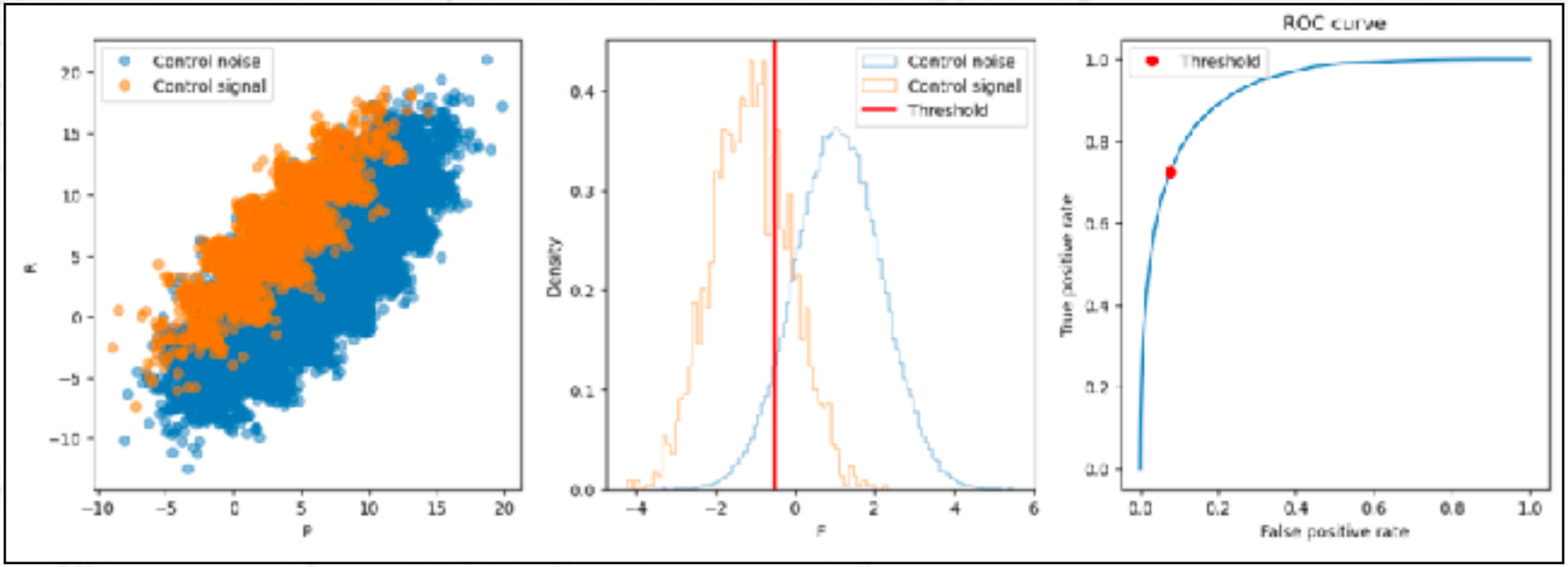
Problem 5.1

5.1.1: The plot easily shows a single peak, which can be fitted nicely.
In reality, it is in fact a double peak, with a smaller wider shifted Gaussian!



Problem 5.1

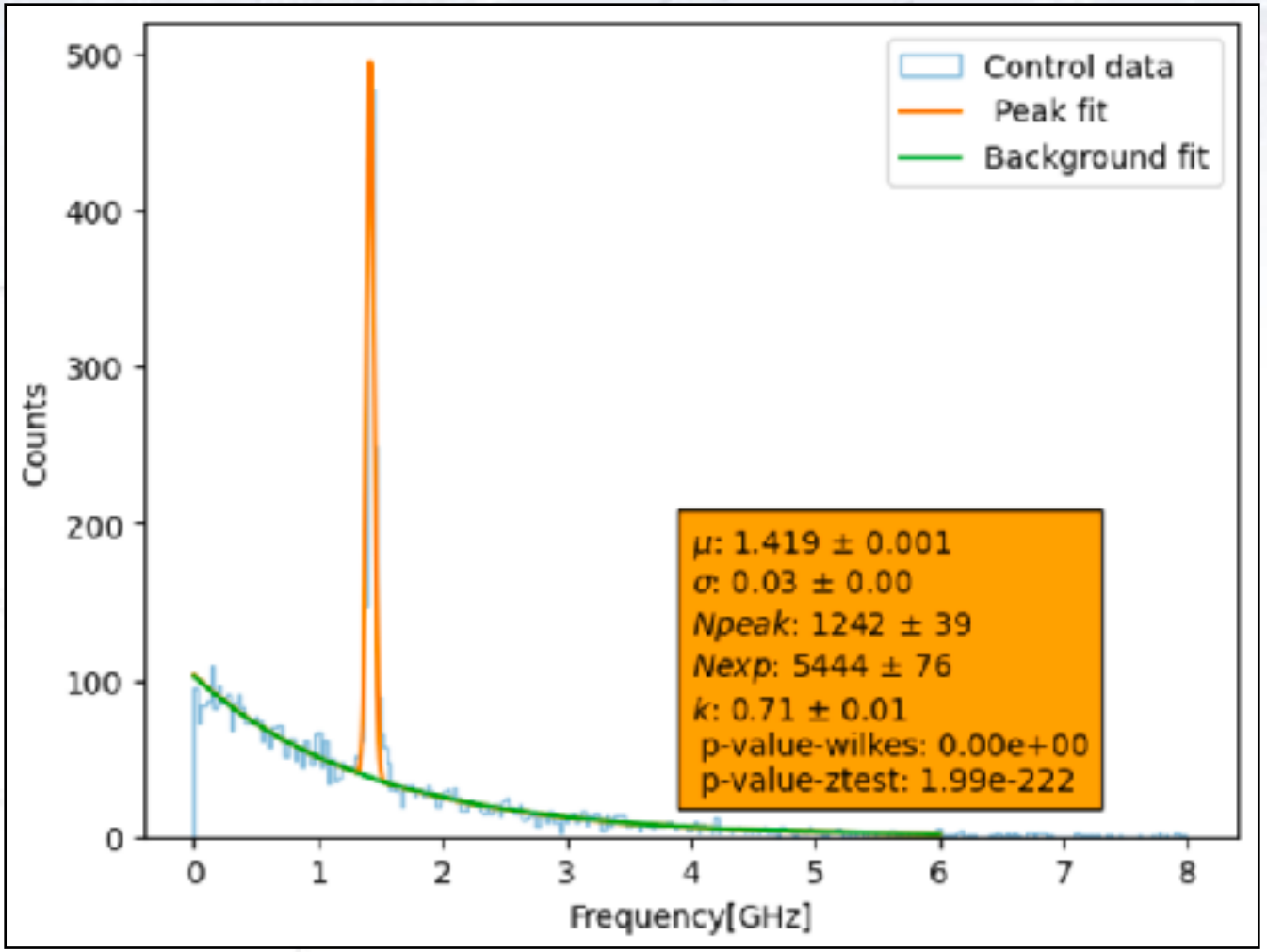
5.1.2: The P and R variables clearly show a separation, especially in 2D.
One can separate either by a cut on each separately (not too good), or a Fisher discriminant.
It is also possible to do "by eye", simply requiring R-P > 0.



The separation can be quantified in terms of a histogram for each type, and a corresponding ROC curve.
An ML algorithm can do even better, given the funny wiggles.

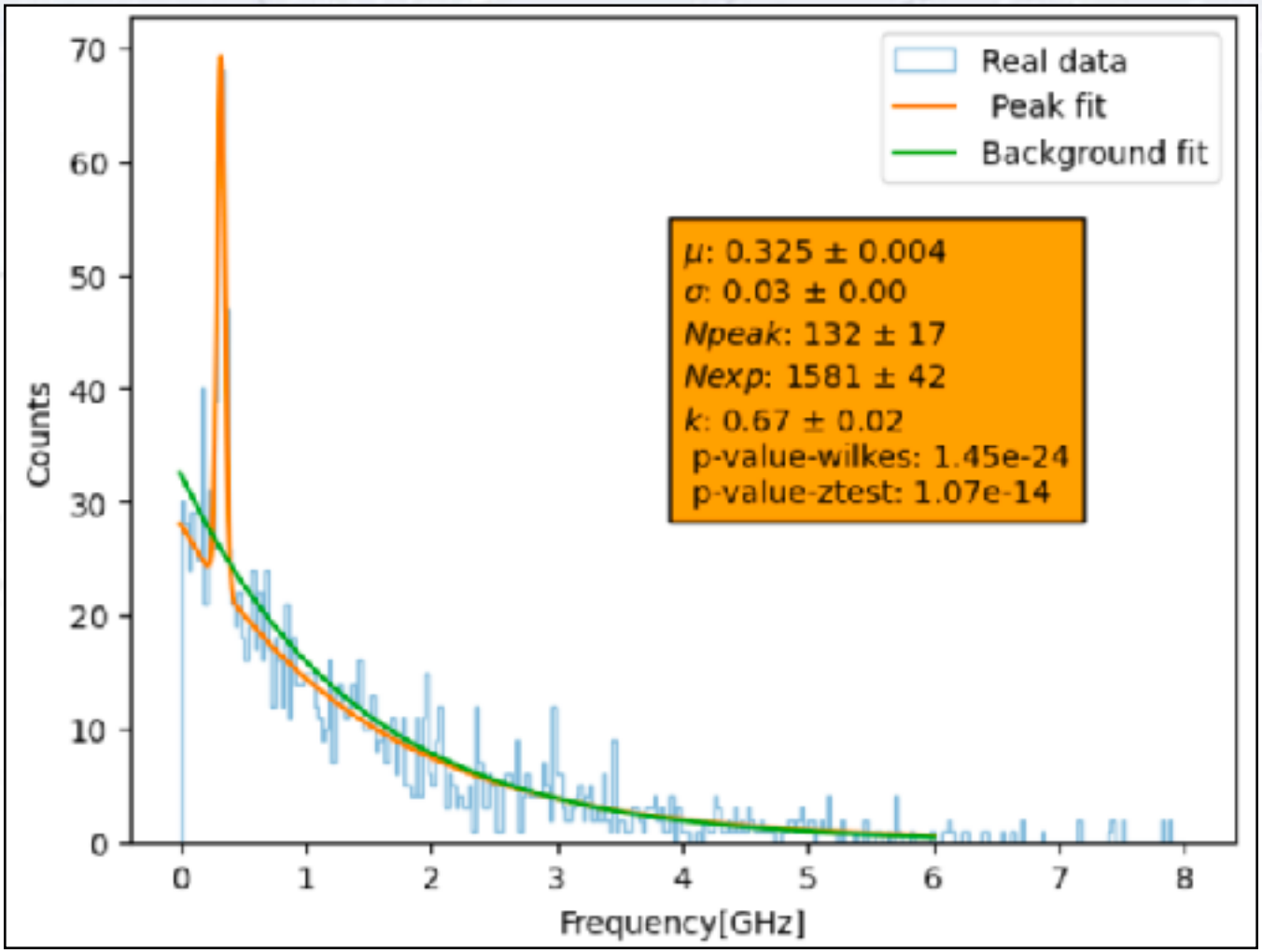
Problem 5.1

5.1.3: After a selection (here Fisher), the peak indeed becomes much more clear!



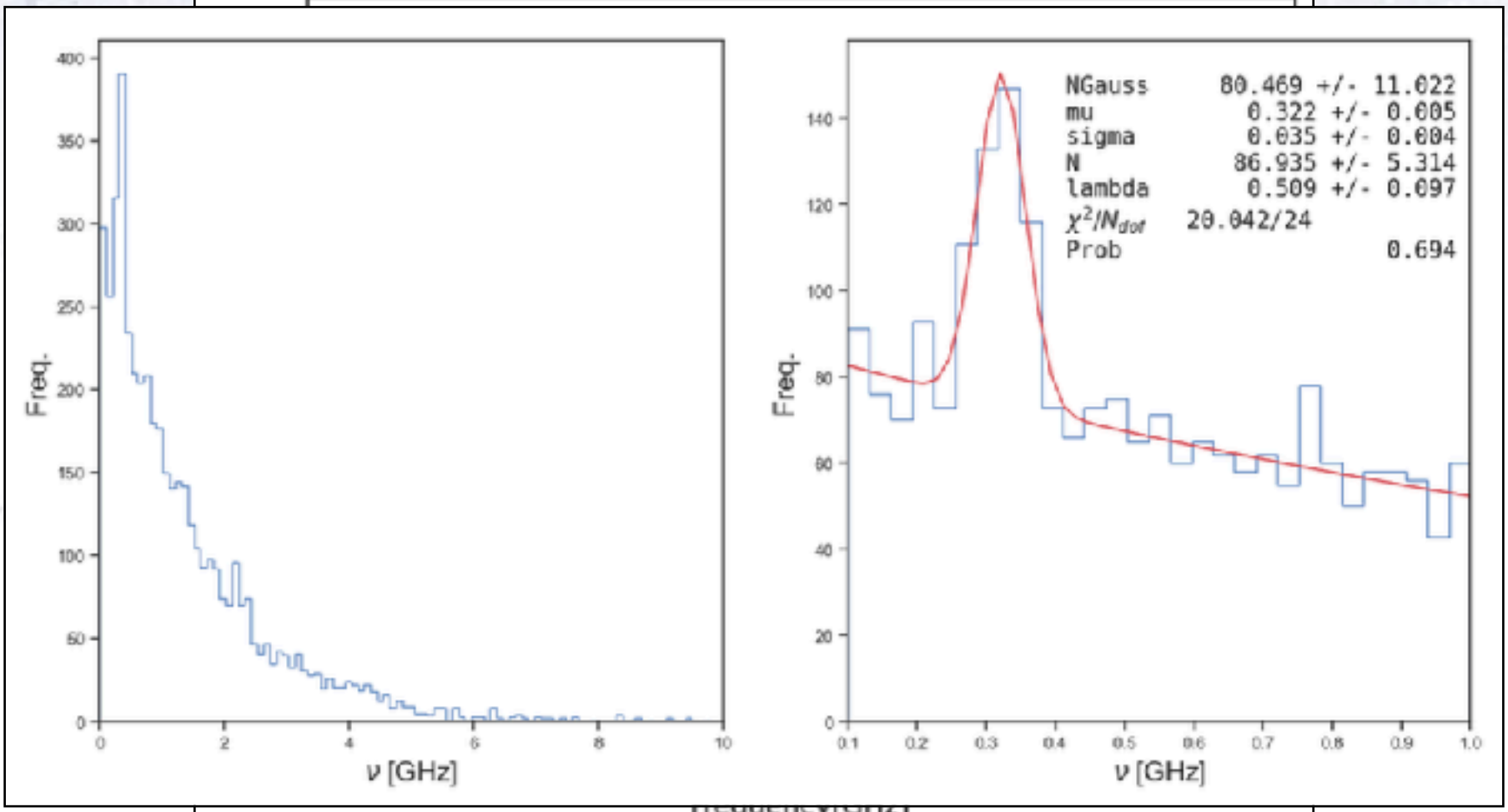
Problem 5.1

5.1.4: Applying the same selection to the real data now gives a much more clear peak, that can be fitted nicely. Lesser selections (R-P > 0) also do the trick.



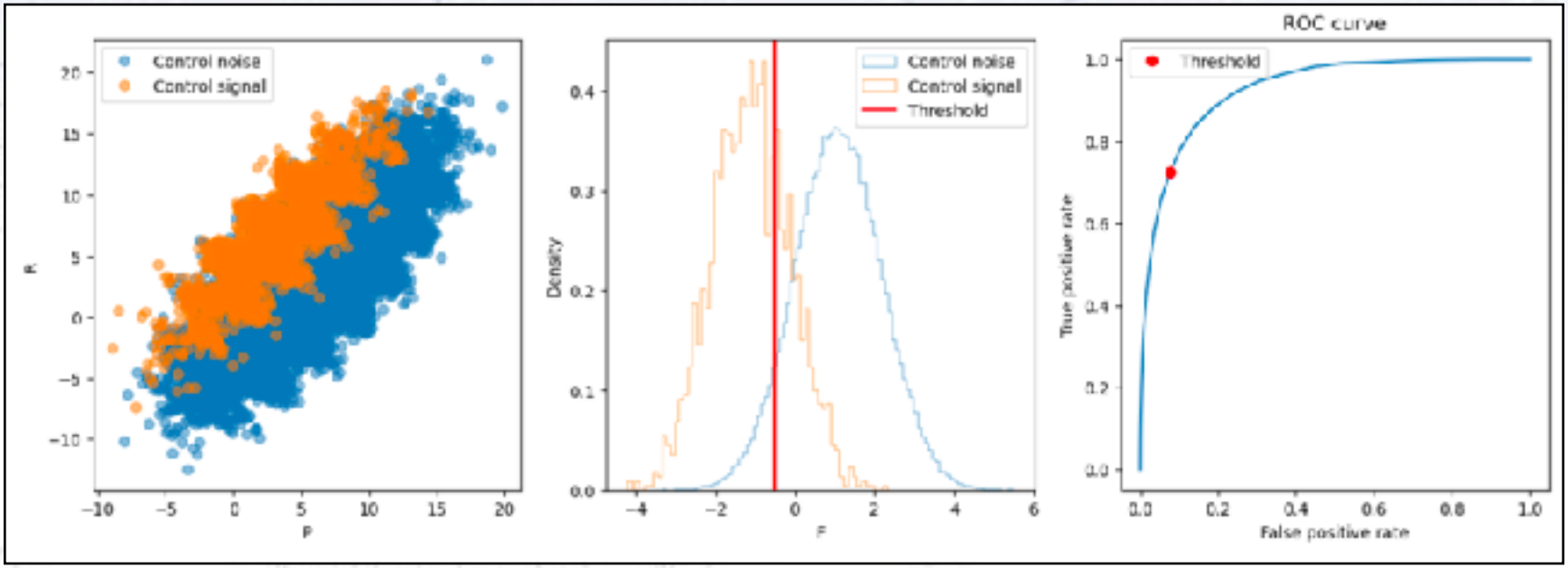
Problem 5.1

5.1.4: Applying the same selection to the real data now gives a much more clear peak, that can be fitted nicely. Lesser selections (R-P > 0) also do the trick.



Problem 5.1

5.1.6: From the middle plot it can be seen that the selection (red line) selects about 60%. Given about 132 events in the small peak just fitted, this yields around 210 events in total.



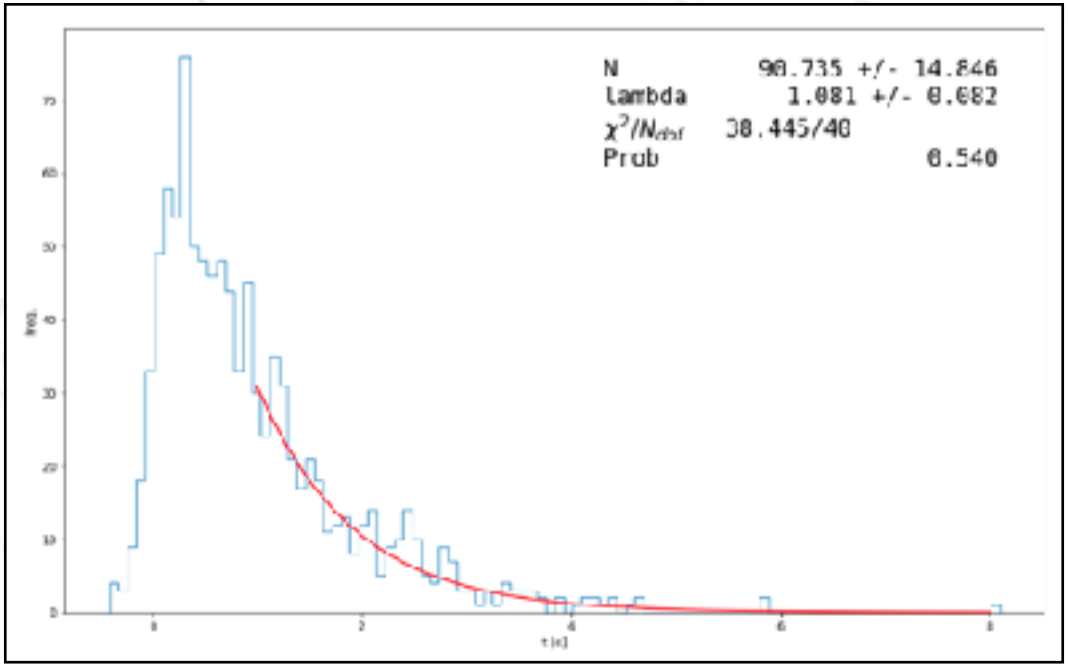
I created the dataset with 200 points in.

Problem 5.2

5.2 (10 points) The file **www.nbi.dk/~petersen/data_DecayTimes.csv** contains the measured decay times (t_i in s) of a Bohrium isotope. The true decay times follow an exponential function, but the measurement of the decay times given have a Gaussian resolution $G(0, \sigma)$ (thus no bias).

- Plot the distribution of decay times, and calculate the mean and median with uncertainty.
- Give a rough estimate of the decay time τ from fitting the high-t tail of the distribution.
- Fit the entire distribution, and (re-)assess the estimated values of τ and σ.

5.2.1: The exponential fit to the tail should yield a lifetime around 1. Deciding on the limits is the challenge. Also, it is clearly best to do a likelihood fit.

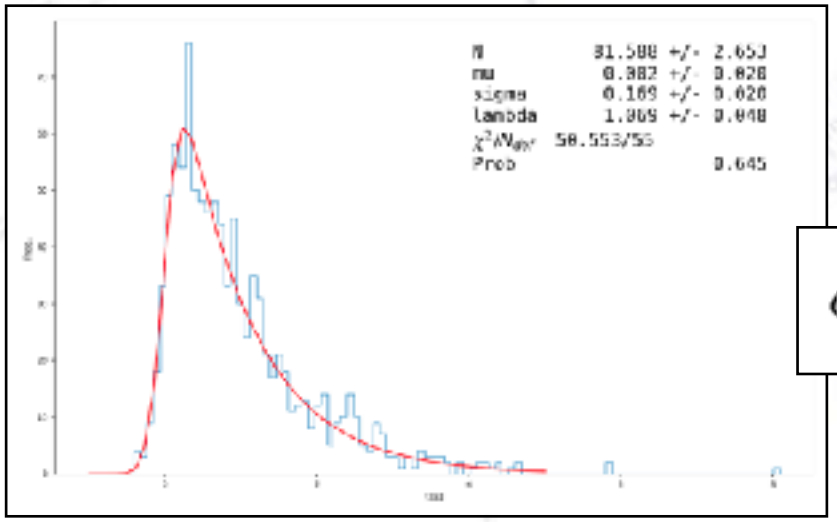


Problem 5.2

5.2 (10 points) The file **www.nbi.dk/~petersen/data_DecayTimes.csv** contains the measured decay times (t_i in s) of a Bohrium isotope. The true decay times follow an exponential function, but the measurement of the decay times given have a Gaussian resolution $G(0, \sigma)$ (thus no bias).

- Plot the distribution of decay times, and calculate the mean and median with uncertainty.
- Give a rough estimate of the decay time τ from fitting the high-t tail of the distribution.
- Fit the entire distribution, and (re-)assess the estimated values of τ and σ.

5.2.2: The function can be expressed in terms of the erfc function, as shown. Alternatives are to fit it by using simulation, as the starting values are well known. One can also write this function numerically, using simply the exponential and a Gaussian in a loop doing the integral. Notice "re-assess"…



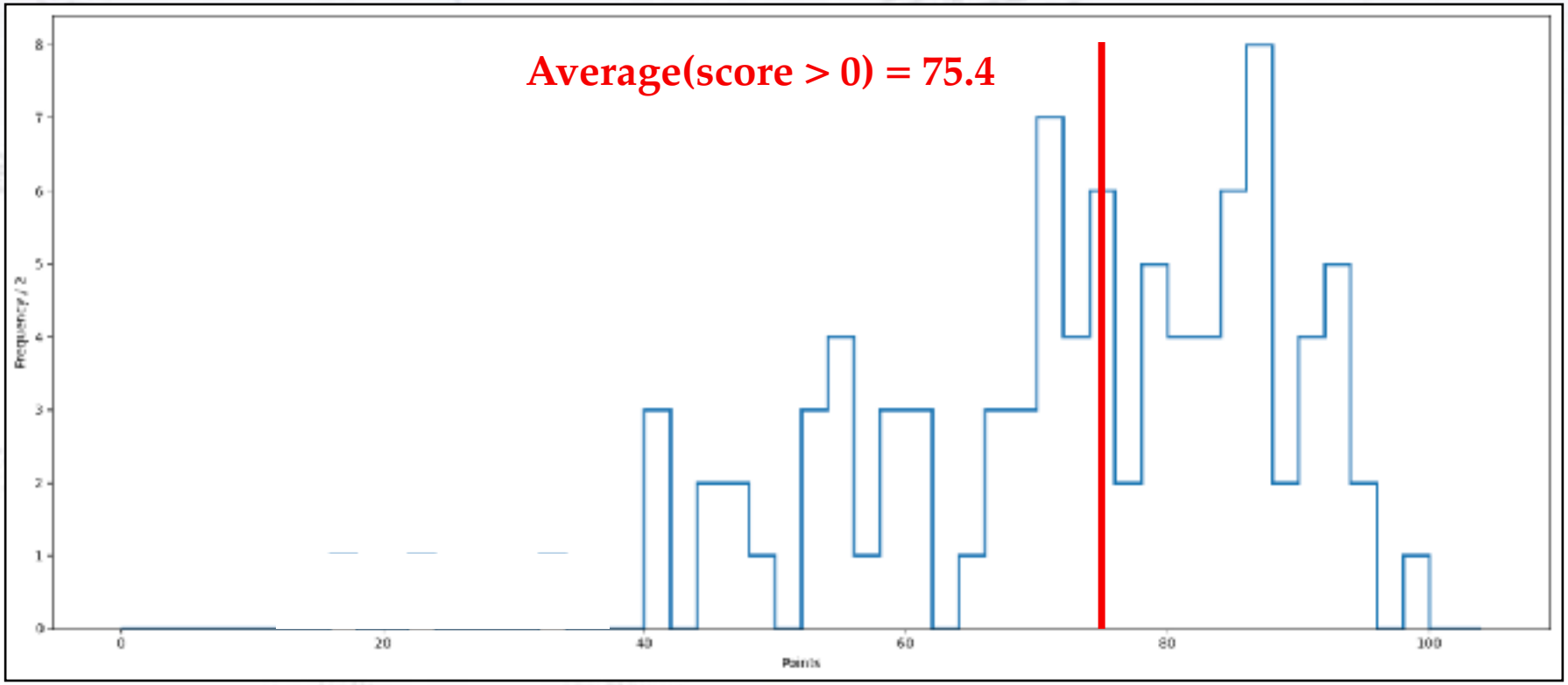
$$G(x) = \frac{1}{2\tau}\exp\left(\frac{1}{2\tau}\left(2\mu + \frac{\sigma^2}{\tau} - 2t\right)\right) \cdot \mathrm{erfc}\left(\frac{\mu + \frac{\sigma^2}{\tau} - t}{\sqrt{2}\sigma}\right)$$

Your scores

General distribution

The distribution of points in the Problem Set was 75.4.
Last year, it was 70.0 and the year before 70.0, so “better than normally”.

Notice, that the grading scale is not fixed, so nothing is “absolute”.



Problem:	1.1.1	1.1.2	1.2.1	1.2.2	1.2.3	2.1.1	2.1.2	2.1.3	2.1.4	2.2.1	2.2.2	3.1.1	3.1.2	3.2.1	3.2.2	4.1.1	4.1.2	4.2	4.3	5.1.1	5.1.2	5.1.3	5.1.4	5.1.5	5.1.6	5.2.1	5.2.2	5.2.3	Total:
Full Points	3	3	2	3	3	2	2	3	3	4	4	5	5	5	5	5	5	5	5	3	3	3	3	3	3	3	4	100
	0.904	0.884	0.876	0.926	0.814	0.86	0.71	0.721	0.6	0.823	0.746	0.826	0.794	0.761	0.652	0.734	0.788	0.80	0.81	0.658	0.753	0.127	0.528	0.370	0.342	0.804	0.693	0.529	75.35852