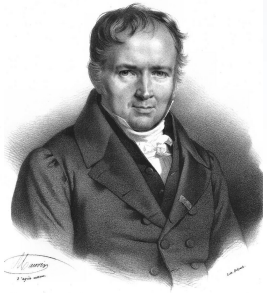
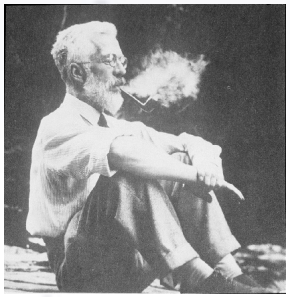
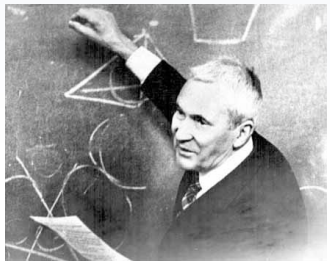
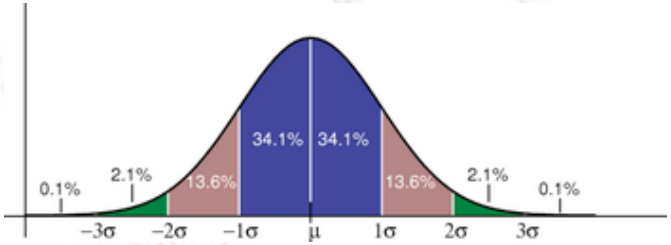
Applied Statistics

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

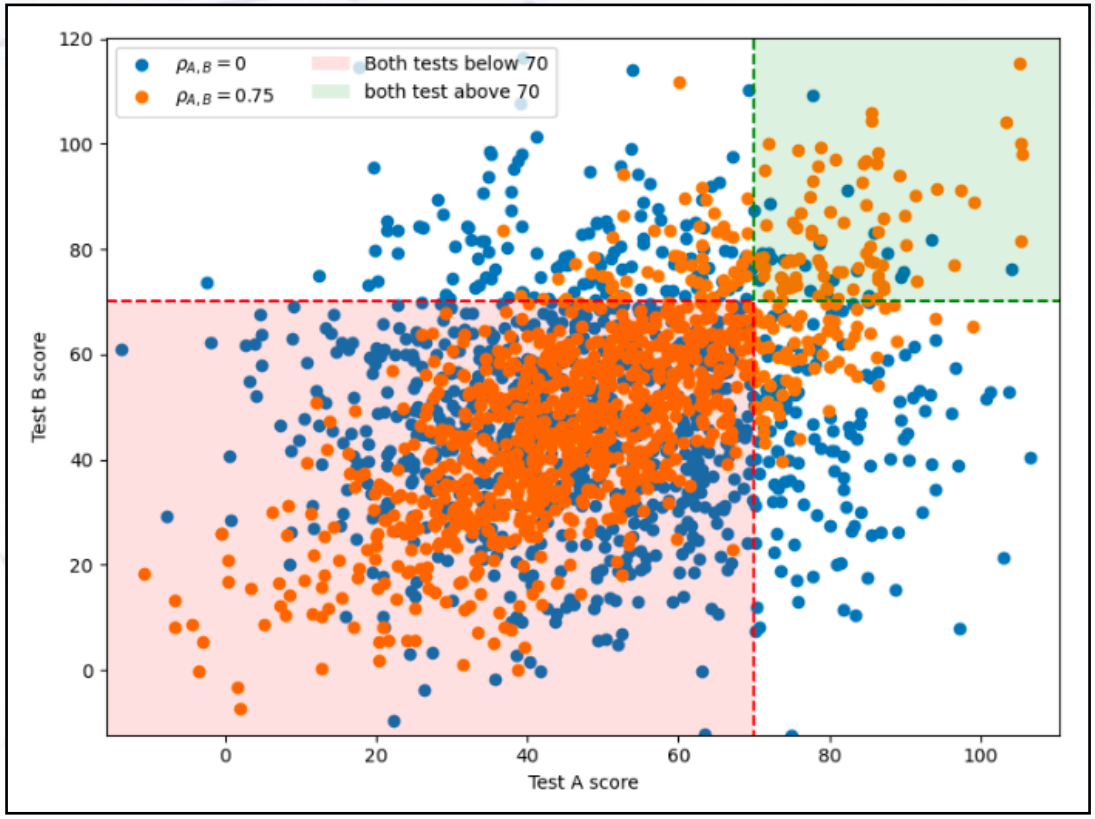
1.1 (8 points) The scores of two tests (A & B) are both Gaussianly distributed with $\mu = 50$, $\sigma = 20$.

- What fraction of students will get a score in test A in the range [55,65]?
- What uncertainty on the mean score do you obtain from 120 B test scores?
- What fraction should get a score above 70 in both tests if $\rho_{A,B} = 0$? If $\rho_{A,B} = 0.75$?

1.1.1: This is an integral of the Gaussian: 0.175

1.1.2: Error on the mean: 1.83

1.1.3: Through simulation (or 2D Gaussian integral), one obtains the value of 0.090 (Significantly greater than if non-correlated).



Problem 1.2

1.2 (4 points) At the roulette you get 12/37 winning chances if you play *douzaine* (e.g. 1-12).

- If you play *douzaine* 20 times, what is the chance that you will win 8 or more times?

1.2.1: The distribution is binomial (N = 20, p = 12/37) and the probability is 30.7%.

Problem 2.1

2.1 (8 points) Let $x = 1.043 \pm 0.014$ and $y = 0.07 \pm 0.23$, and let $z_1 = xye^{-y}$ and $z_2 = (y+1)^3/(x-1)$.

- Which of the (uncorrelated) variables x and y contributes most to the uncertainty on z_1?
- What are the uncertainties of z_1 and z_2, if x and y are correlated with $\rho = 0.4$?
- Plot $z_1 \in [-2, 2]$ against $z_2 \in [-10, 90]$. In this range, what is the z_1 vs. z_2 correlation?

2.1.1: The uncertainty on z_1 is by far dominated by y:

$$(\sigma_{z_1}^2)_x = (0.07 \cdot e^{-0.07} \cdot 0.014)^2 = 8.3410^{-}7$$

$$(\sigma_{z_1}^2)_y = ((1.043 - 1)e^{-0.07} \cdot 0.23)^2 = 0.0433$$

2.1.2: For z_2, this could look like…

Because x and y are correlated with $\rho = 0.4$, σ_{z_1} can be given by 8

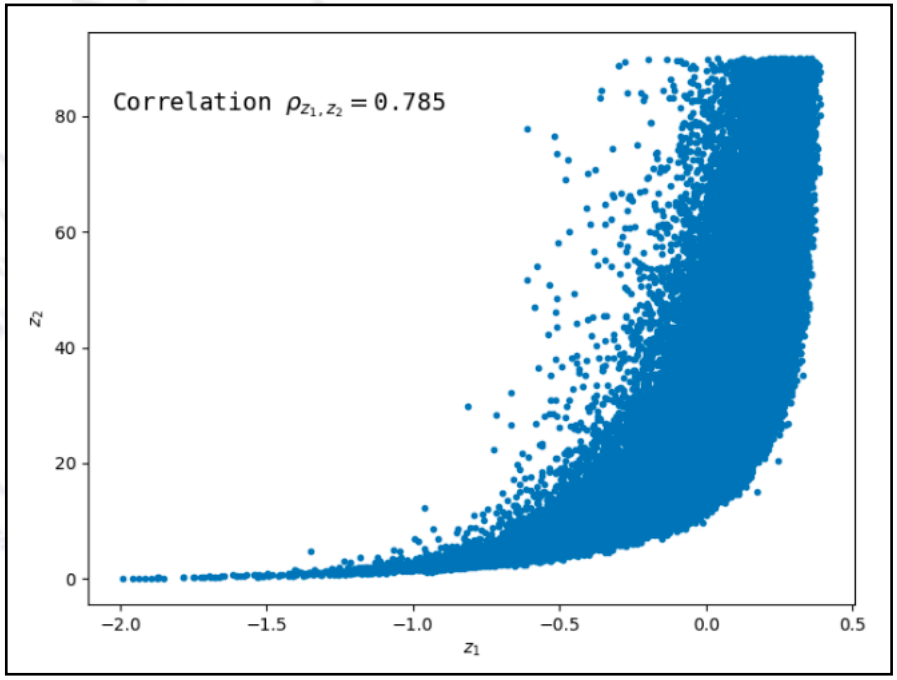
$$\sigma_{z_1} = \sqrt{(\frac{\partial z_1}{\partial x}\sigma_x)^2 + (\frac{\partial z_1}{\partial y}\sigma_y)^2 + 2\rho(\frac{\partial z_1}{\partial x}\sigma_x)(\frac{\partial z_1}{\partial y}\sigma_y)}$$
$$= 0.21$$

Similarly,

$$\sigma_{z_2} = \sqrt{(\frac{\partial z_2}{\partial x}\sigma_x)^2 + (\frac{\partial z_2}{\partial y}\sigma_y)^2 + 2\rho(\frac{\partial z_2}{\partial x}\sigma_x)(\frac{\partial z_2}{\partial y}\sigma_y)}$$

$$\sigma_{z_2} = \sqrt{(-\frac{(y+1)^3}{(x-1)^2}\sigma_x)^2 + (\frac{3(y+1)^2}{x}\sigma_y)^2 + 2\rho(-\frac{(y+1)^3}{(x-1)^2}\sigma_x)(\frac{3(y+1)^2}{x}\sigma_y)}$$
$$= 16.95$$

2.1.3: Scatter plot shows quiet "vivid" functions (so check error formula!)



Problem 2.1

2.1 (8 points) Let $x = 1.043 \pm 0.014$ and $y = 0.07 \pm 0.23$, and let $z_1 = xye^{-y}$ and $z_2 = (y+1)^3/(x-1)$.

- Which of the (uncorrelated) variables x and y contributes most to the uncertainty on z_1?
- What are the uncertainties of z_1 and z_2, if x and y are correlated with $\rho = 0.4$?
- Plot $z_1 \in [-2, 2]$ against $z_2 \in [-10, 90]$. In this range, what is the z_1 vs. z_2 correlation?

2.1.1: The uncertainty on z_1 is by far dominated by y:

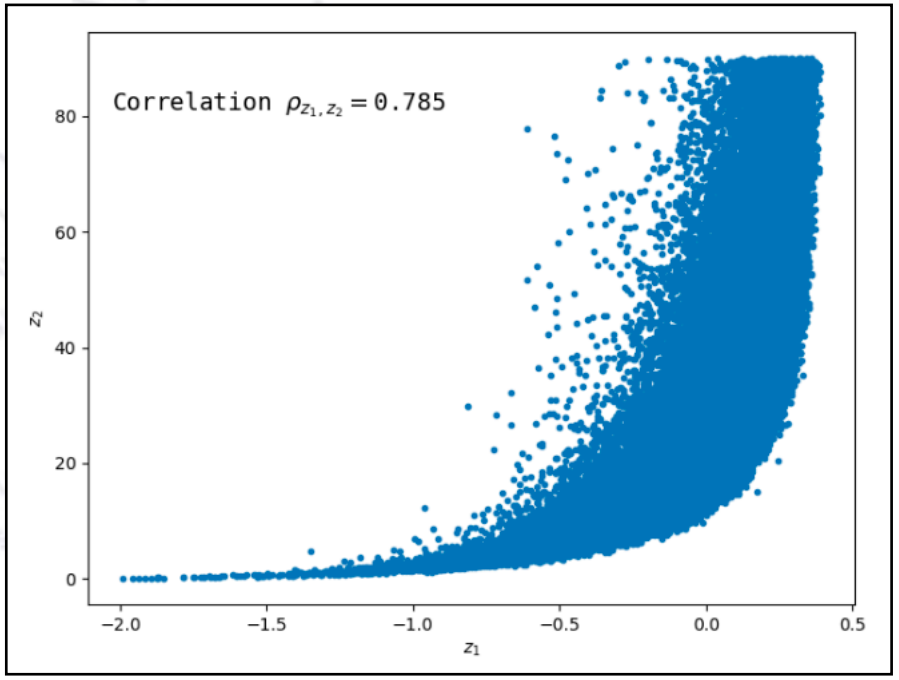
$$(\sigma^2_{z_1})_x = (0.07 \cdot e^{-0.07} \cdot 0.014)^2 = 8.3410^{-}7$$

$$(\sigma^2_{z_1})_y = ((1.043 - 1)e^{-0.07} \cdot 0.23)^2 = 0.0433$$

2.1.2: For z_2, this is…

"A complete and utter breakdown of the error propagation formula"

2.1.3: Scatter plot shows quiet "vivid" functions (so check error formula!)



Problem 2.2

2.2 (7 points) In a (Cavendish) experiment, you have made five measurements of Earth's density ρ:

Observation	1	2	3	4	5
Result (in g/cm^3)	5.50 ± 0.10	5.61 ± 0.21	4.88 ± 0.15	5.07 ± 0.14	5.26 ± 0.13

- What is the combined result and uncertainty of these five measurements?
- Are your measurements consistent with each other? If not, what is then your best estimate?
- The precise value is 5.514 g/cm^3. How consistent is you measurement with this number?

2.2.1: Weighted mean = 5.28 ± 0.06

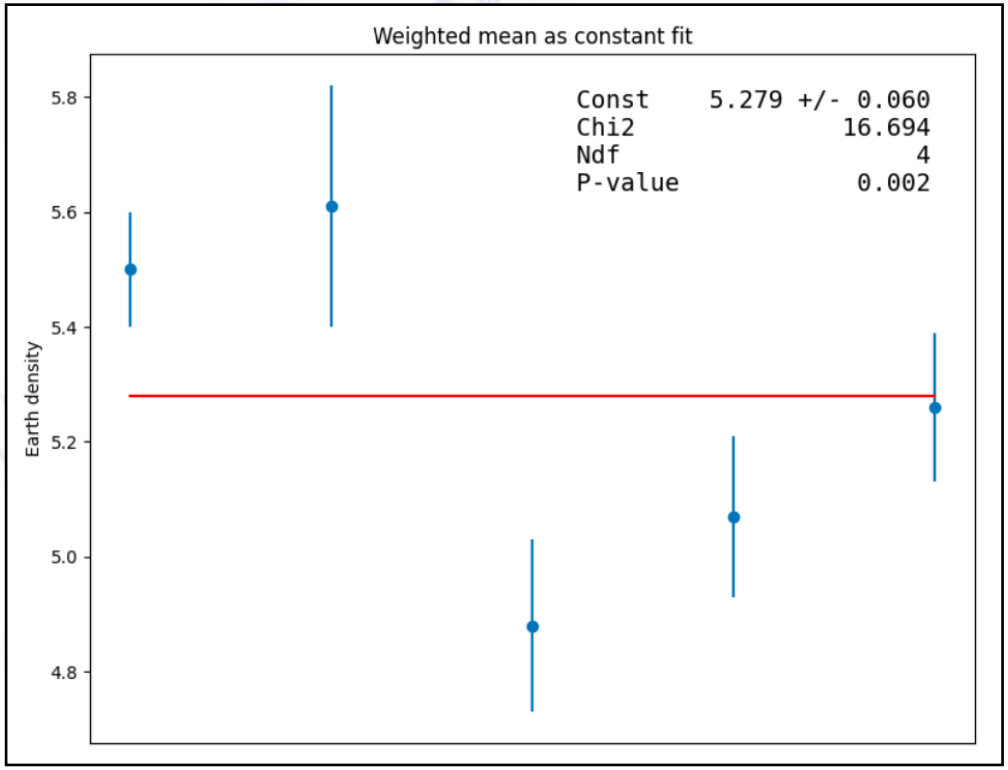
2.2.2: With P(16.7, 4) = 0.002, no!
Check ChiSquare contributions:

Result[g/cm^3]	5.50	5.61	4.88	5.07	5.26
χ^2	4.87	2.48	7.09	2.24	0.02

Rejecting 3rd measurement gives:
P(8.26, 3) = 0.041… acceptable?

One could also drop uncertainties.

2.2.3: 1-sided test (dep. on 2.2.2)



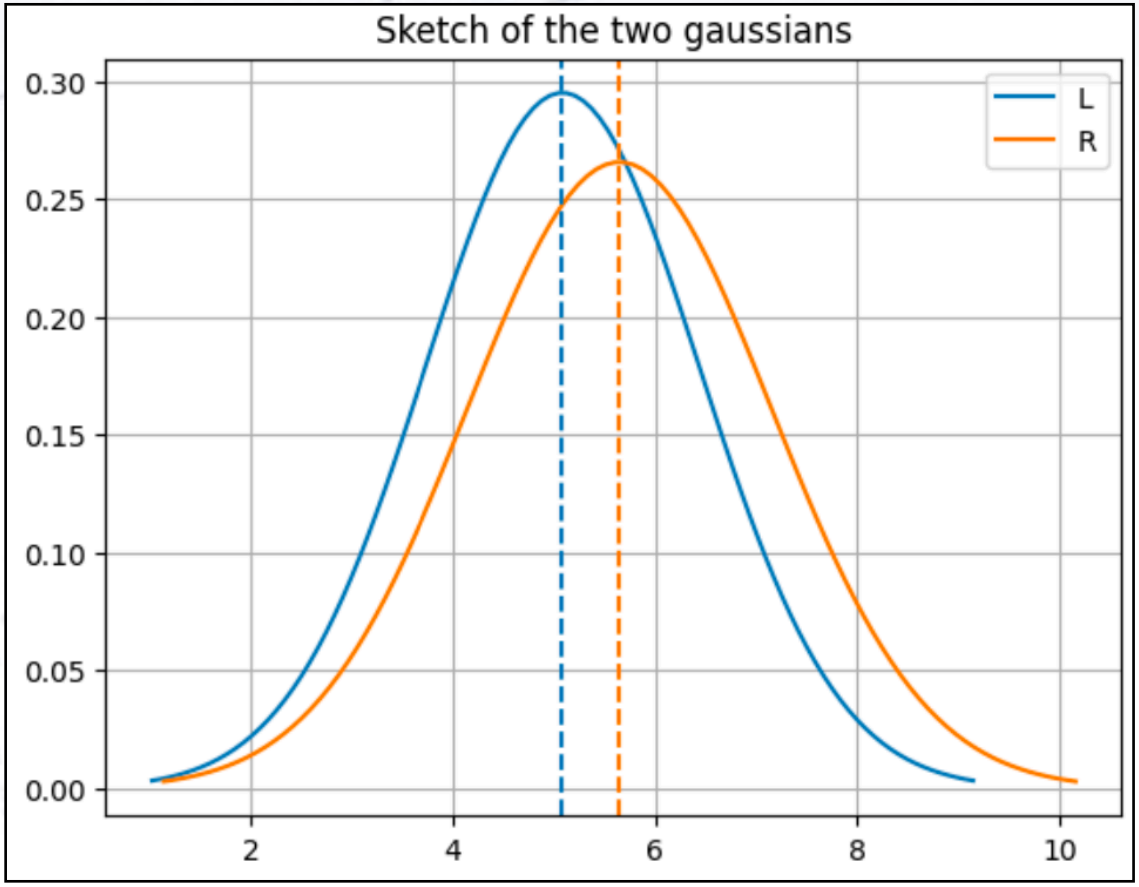
Problem 2.3

2.3 (7 points) An ellipse E has semi-major axis $a = 1.04 \pm 0.27$ and eccentricity $e = 0.71 \pm 0.12$.

- The area A of an ellipse is generally $A = \pi a^2 \sqrt{1 - e^2}$. What is the area of the ellipse E?
- The circumference C has no formula but can be bounded as $4a\sqrt{2 - e^2} < C < \pi a \sqrt{4 - 2e^2}$. What value and uncertainty for C would you give?

2.3.1: A = 2.39 ± 1.31
Thus a very large error!

2.3.2: This was harder, but the illustration sums it up nicely:
The Left and Right limits are close compared to their (correlated) uncertainty.
Thus, take the difference as an error, and add it in quadrature with the widths shown: C ~ 5.3 ± 1.5

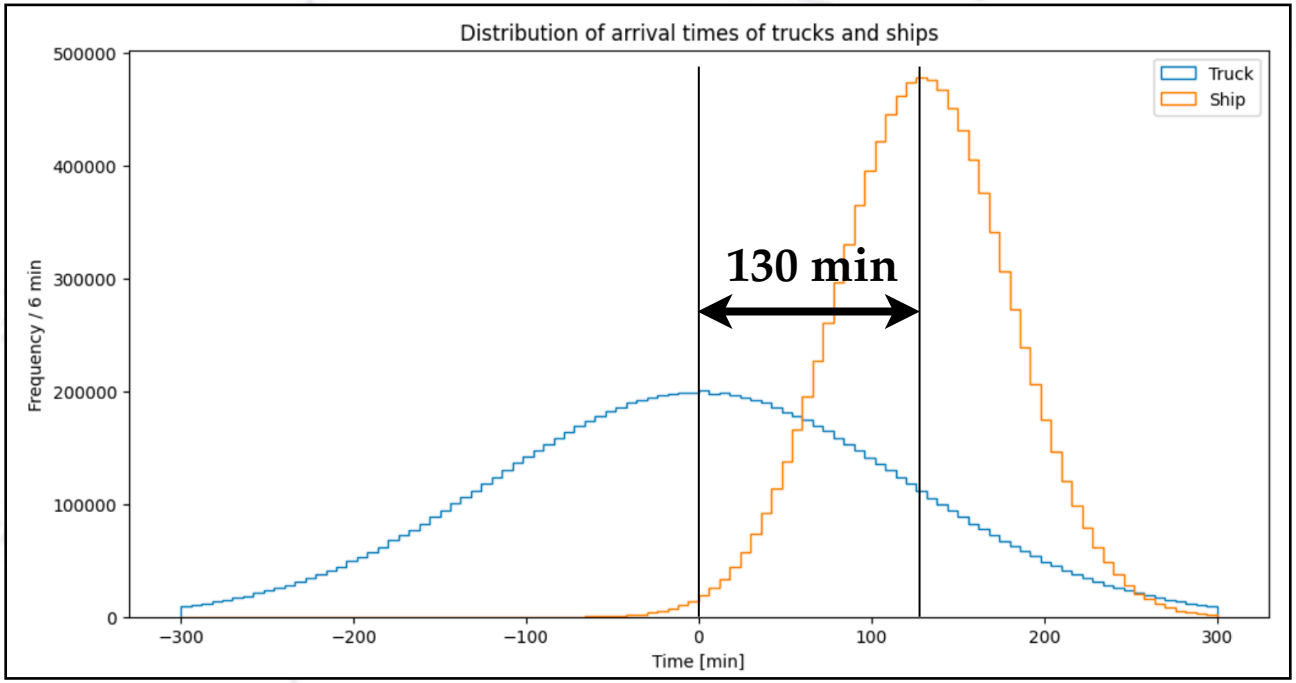


Problem 3.1

3.1 (8 points) You are optimising container transport, in particular the time, Δt, between the daily truck arrivals (120 minutes uncertainty) and the ship departure (50 minutes uncertainty).

- If $\Delta t = 130$ minutes, what fraction of containers will have to wait to the next day?
- For what value of Δt do containers, on average, have the least waiting time?

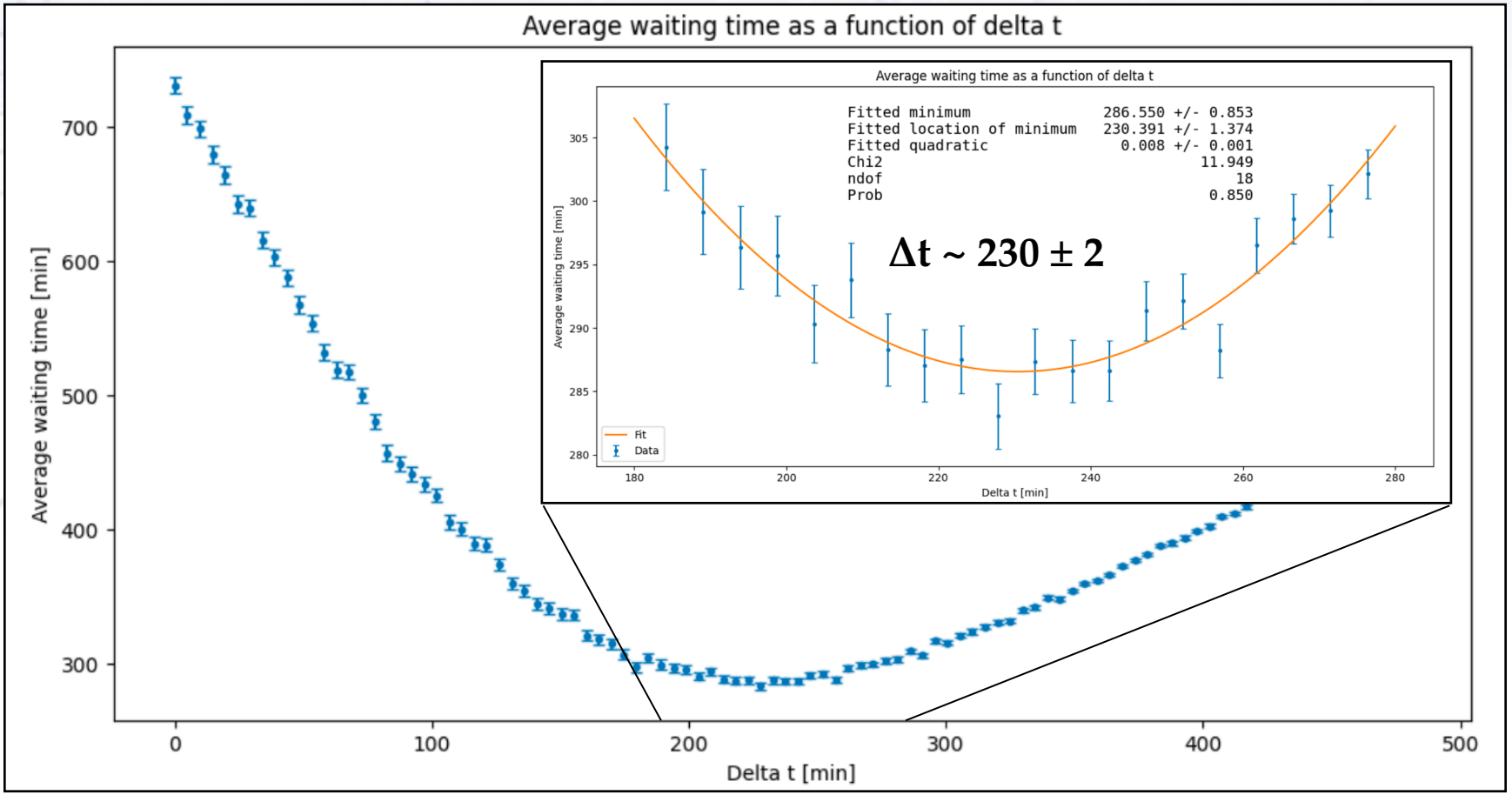
With 130 min, this corresponds exactly to 1 sigma single-sided, so 16%.
The second problem is most easily solved by simulation, though it can be done analytically, using the error function (integral of Gaussian).



Problem 3.1

3.1 (8 points) You are optimising container transport, in particular the time, Δt, between the daily truck arrivals (120 minutes uncertainty) and the ship departure (50 minutes uncertainty).

- If $\Delta t = 130$ minutes, what fraction of containers will have to wait to the next day?
- For what value of Δt do containers, on average, have the least waiting time?



Problem 3.2

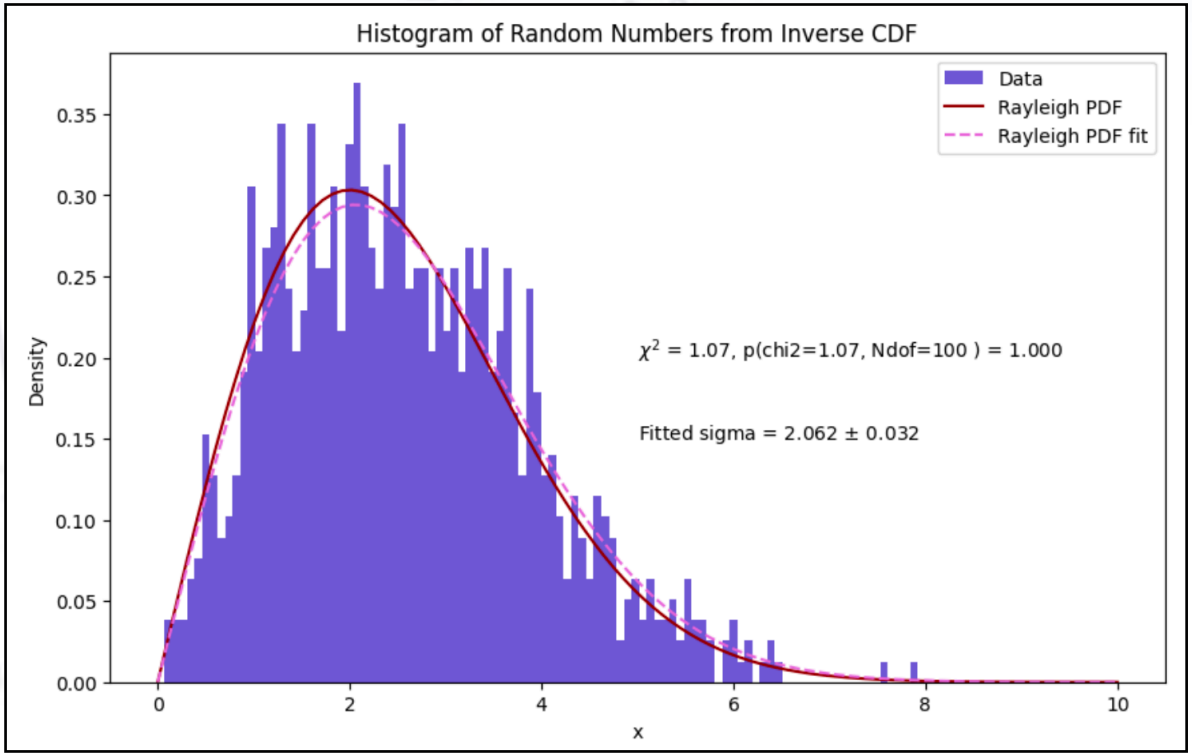
3.2 (13 points) The Rayleigh distribution is a PDF given by: $f(x) = \frac{x}{\sigma^2}\exp(-\frac{1}{2}x^2/\sigma^2)$, with $x \in [0, \infty]$.

- By what method(s) would you generate random numbers (from uniform) according to $f(x)$?
- Generate N=1000 random numbers according to $f(x)$ for $\sigma = 2$, and plot these.
- Fit this distribution of random numbers. How well can you determine σ from the fit?
- Test the $1/\sqrt{N}$ scaling of the σ fit uncertainty for $N \in [50, 5000]$.

3.2.1: In this case, the transformation method works, while accept-reject can do.

3.2.2: This and the next problem are standard. However, be careful of binning vs. fit type!

3.3.3: Repeating the fit for many N shows the 1/sqrt(N) law, though only for likelihood fit, as 50 entries is too little for the ChiSquare!

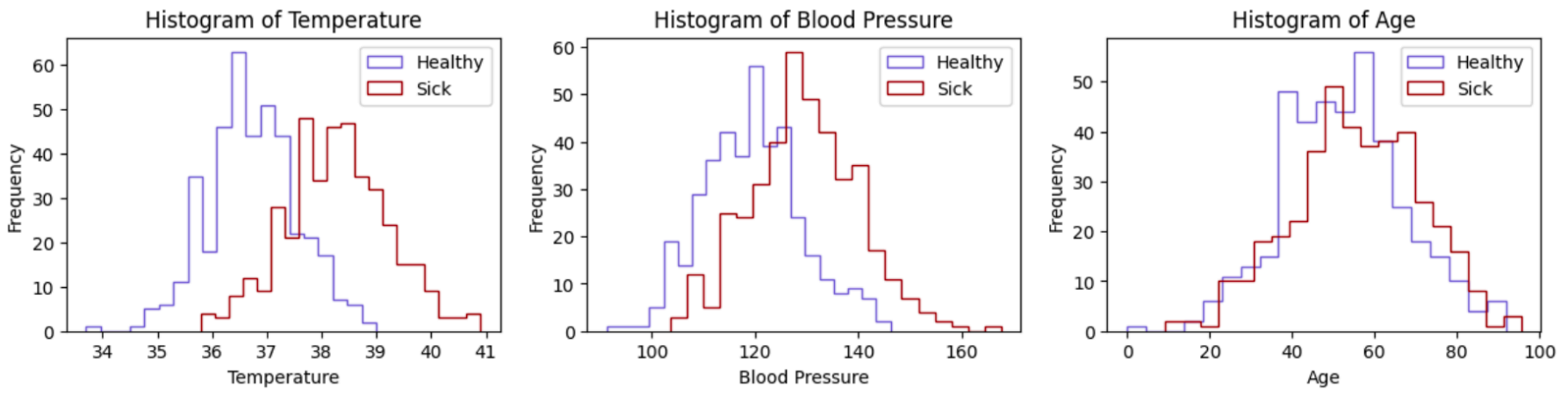


Problem 4.1

4.1 (15 points) Patients are either healthy or infected with Anoroc disease and their temperature, blood pressure and age is found in **www.nbi.dk/~petersen/data_AnorocDisease.csv**. For patients 1-800 (control) the outcome in known, while it is unknown for patients 801-1000 (unknown).

- Using the control sample, plot the three distributions for healthy and sick, respectively. Which of the three single measures gives the highest separation between healthy and sick?
- Test if the age distribution is statistically the same between healthy and sick.
- Given any combination of all three variables, separate the two groups as well as possible and estimate the number of infected patients in the unknown group.
- Assuming a prior probability of $p = 0.01$ of being ill, what is the probability that a new patient with $T = 38.5$ C° is ill?

4.1.1: Temperature separates best. Remember to control the binning (unlike below).
4.1.2: KS-test (mean, Chi2) gives p-values of 0.0018 (0.0027, 0.0081), so not the same.

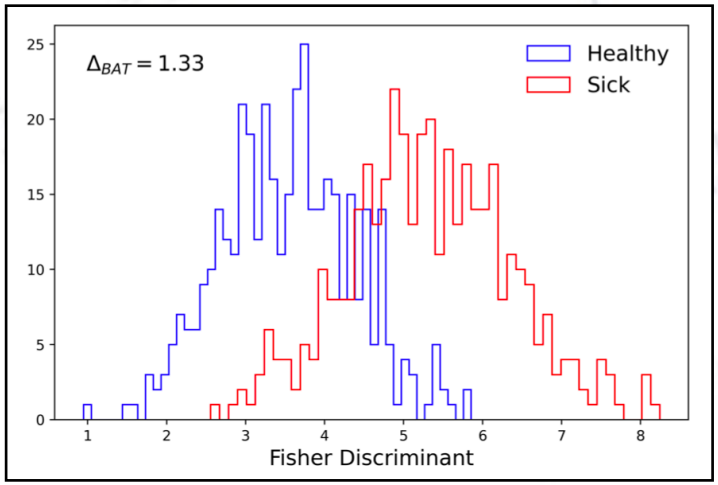


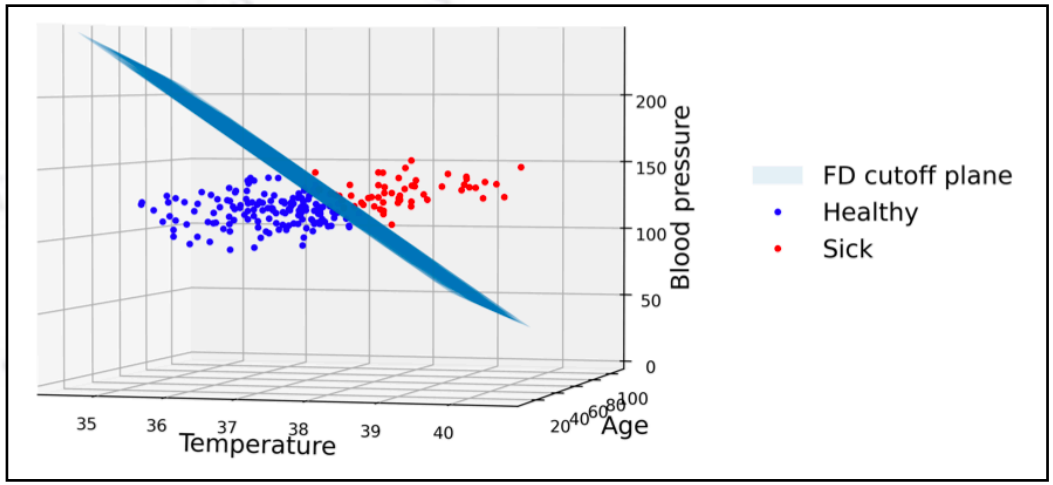
Problem 4.1

4.1 (15 points) Patients are either healthy or infected with Anoroc disease and their temperature, blood pressure and age is found in **www.nbi.dk/∼petersen/data_AnorocDisease.csv**. For patients 1-800 (control) the outcome in known, while it is unknown for patients 801-1000 (unknown).

- Using the control sample, plot the three distributions for healthy and sick, respectively. Which of the three single measures gives the highest separation between healthy and sick?
- Test if the age distribution is statistically the same between healthy and sick.
- Given any combination of all three variables, separate the two groups as well as possible and estimate the number of infected patients in the unknown group.
- Assuming a prior probability of $p = 0.01$ of being ill, what is the probability that a new patient with $T = 38.5$ C° is ill?

4.1.3: Though not required, most people used the Fisher, and did well.
The separation comes out to be better than for temperature (my "proof"!).





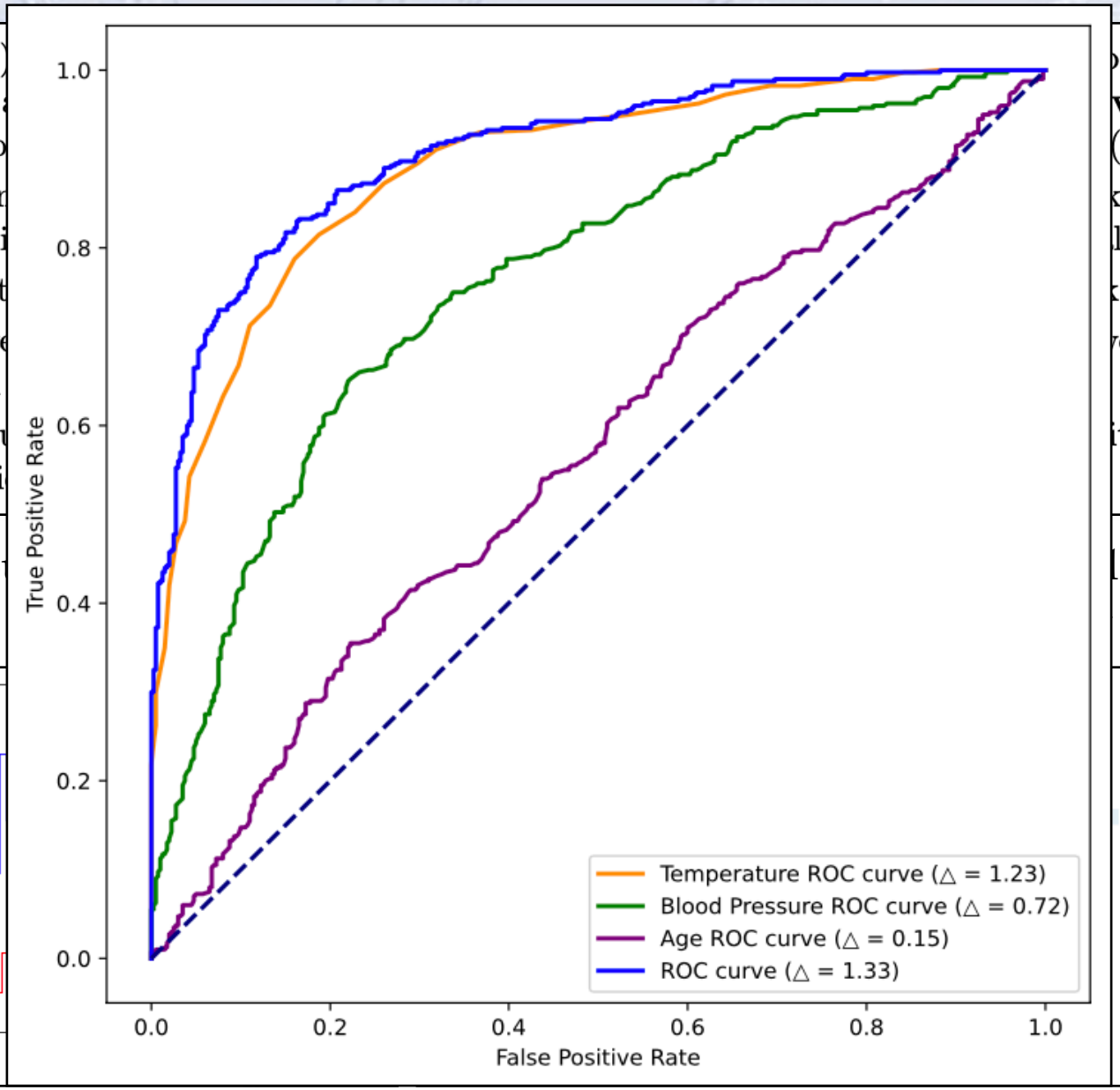
Problem 4.1

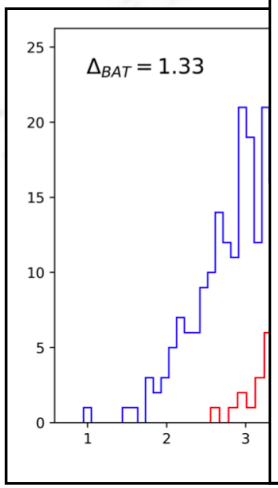
4.1 (15 points)
pressure
1-800 (co
- Usi
Whi
- Test
- Giv
and
- Ass
pati

perature, blood
v. For patients
(unknown).
k, respectively.
lthy and sick?
k.
ell as possible
ity that a new

4.1.3: Tho
The

l.
"proof"!).





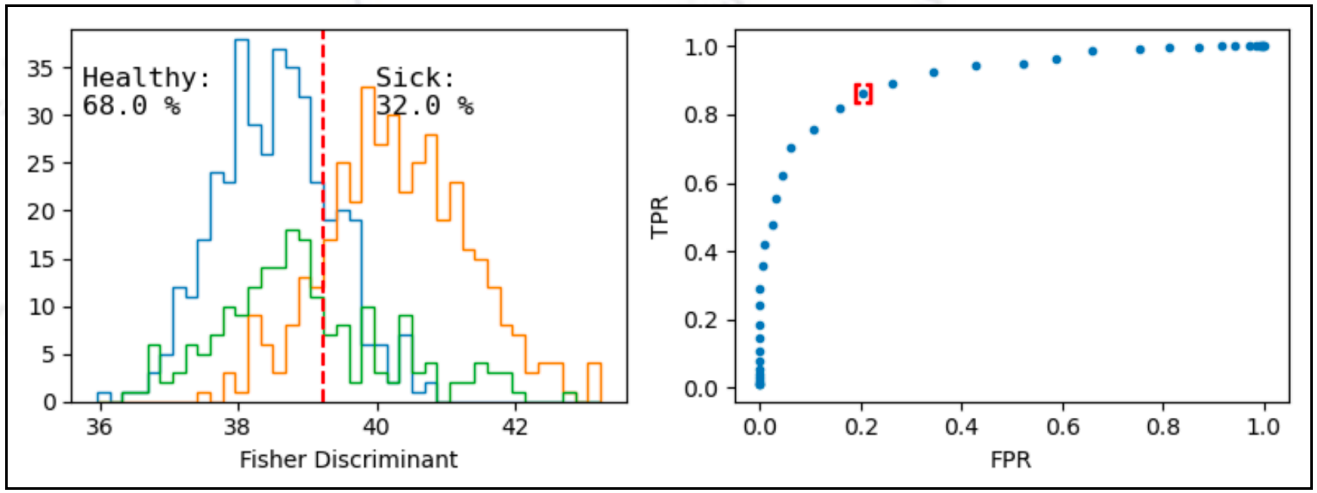
FD cutoff plane
Healthy
Sick

Problem 4.1

4.1 (15 points) Patients are either healthy or infected with Anoroc disease and their temperature, blood pressure and age is found in **www.nbi.dk/~petersen/data_AnorocDisease.csv**. For patients 1-800 (control) the outcome in known, while it is unknown for patients 801-1000 (unknown).

- Using the control sample, plot the three distributions for healthy and sick, respectively. Which of the three single measures gives the highest separation between healthy and sick?
- Test if the age distribution is statistically the same between healthy and sick.
- Given any combination of all three variables, separate the two groups as well as possible and estimate the number of infected patients in the unknown group.
- Assuming a prior probability of $p = 0.01$ of being ill, what is the probability that a new patient with $T = 38.5$ C° is ill?

4.1.3: Though not required, most people used the Fisher, and did well.
Most people (correctly) estimated the number of sick to be around 60.



Problem 4.1

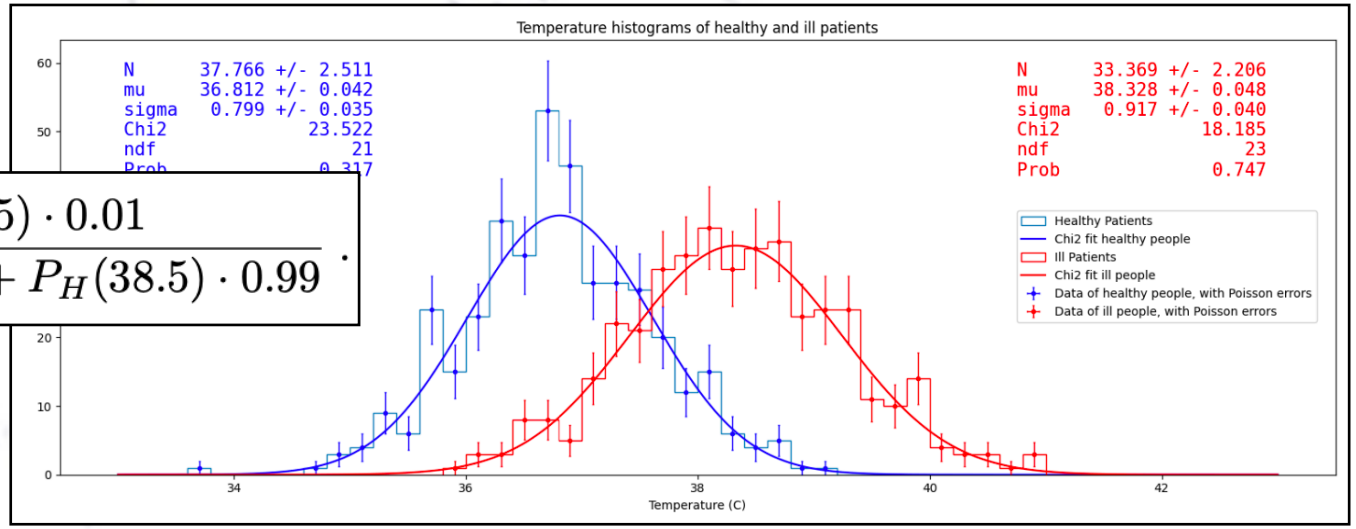
4.1 (15 points) Patients are either healthy or infected with Anoroc disease and their temperature, blood pressure and age is found in **www.nbi.dk/~petersen/data_AnorocDisease.csv**. For patients 1-800 (control) the outcome in known, while it is unknown for patients 801-1000 (unknown).

- Using the control sample, plot the three distributions for healthy and sick, respectively. Which of the three single measures gives the highest separation between healthy and sick?
- Test if the age distribution is statistically the same between healthy and sick.
- Given any combination of all three variables, separate the two groups as well as possible and estimate the number of infected patients in the unknown group.
- Assuming a prior probability of $p = 0.01$ of being ill, what is the probability that a new patient with $T = 38.5$ C° is ill?

4.1.4: This problem is about Bayes' Theorem (give-away word: "prior"), and so the PDF for temp. is needed at 38.5 degrees.

$$P = \frac{P_I(38.5) \cdot 0.01}{P_I(38.5) \cdot 0.01 + P_H(38.5) \cdot 0.99} .$$

Even if P(38.5) is high, P is not really!



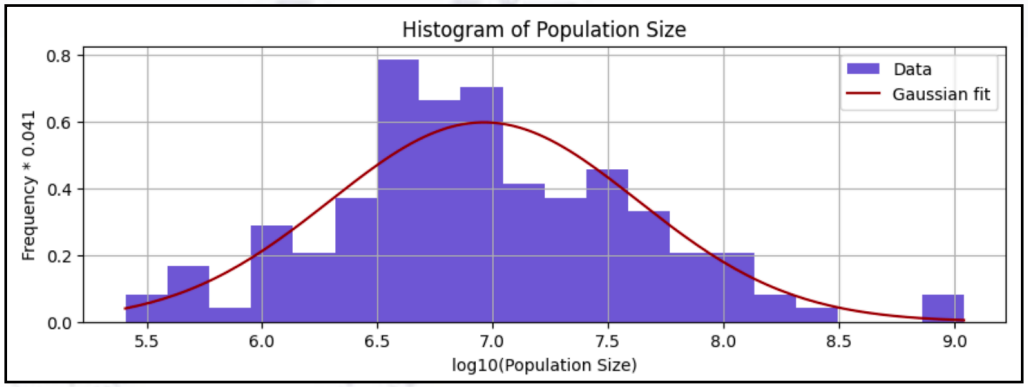
Problem 4.2

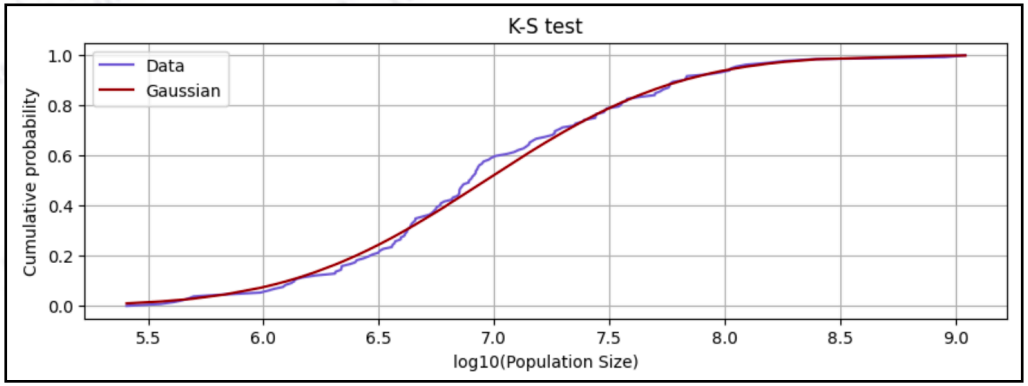
4.2 (14 points) The file **www.nbi.dk/~petersen/data_CountryScores.csv** contains a list of countries along with several key numbers and indices.

- Determine the mean, median, standard deviation, 15.87%, and 84.13% quantiles of the GDP.
- Does the distribution of $\log_{10}$(PopSize) follow a Gaussian distribution?
- What are the Pearson and Spearman correlations between happiness and education indeces?
- Plot the Happiness-Index as a function of GDP, and fit the relation between the two. From this fit, what would you estimate the uncertainty to be on the Happiness-index?

4.2.1:

Country GDP	
Variable	Value
Mean	17362
Median	6677
Standard deviation	23750
15.87% quantile	1187
84.13% quantile	40850





4.2.2: The simple method is to fit distribution with a ChiSquare.

However, this is only a weak test - the KS-test is the right way!

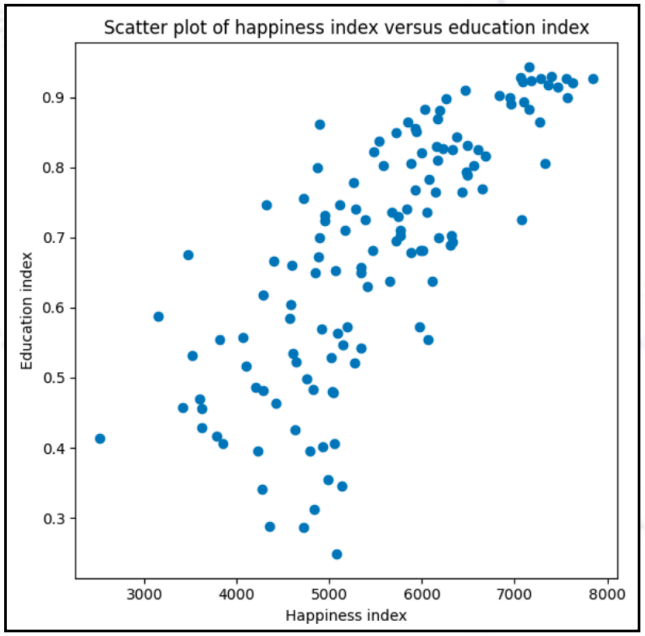
Problem 4.2

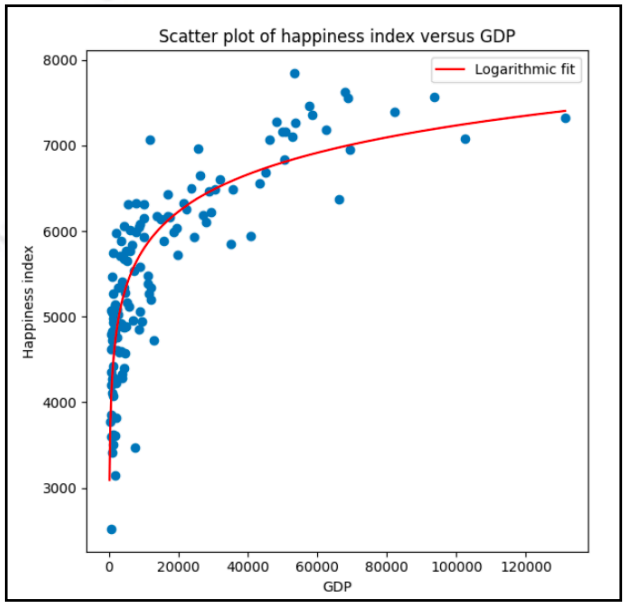
4.2 (14 points) The file **www.nbi.dk/~petersen/data_CountryScores.csv** contains a list of countries along with several key numbers and indices.

- Determine the mean, median, standard deviation, 15.87%, and 84.13% quantiles of the GDP.
- Does the distribution of $\log_{10}$(PopSize) follow a Gaussian distribution?
- What are the Pearson and Spearman correlations between happiness and education indeces?
- Plot the Happiness-Index as a function of GDP, and fit the relation between the two. From this fit, what would you estimate the uncertainty to be on the Happiness-index?

4.2.3: The correlation is high (0.765, 0.804), as is clear from a plot.

4.2.4: There are many fits to be made, giving residuals around 600.





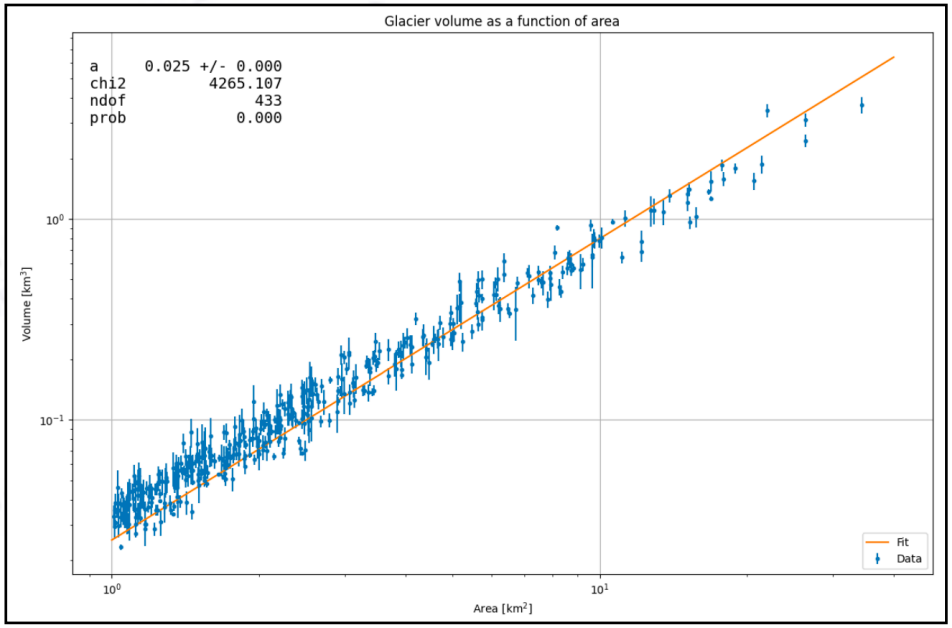
Problem 5.1

5.1 (16 points) The file **www.nbi.dk/~petersen/data_GlacierSizes.csv** contains the estimated area and volume including uncertainties of 434 glaciers with an area above 1 km^2.

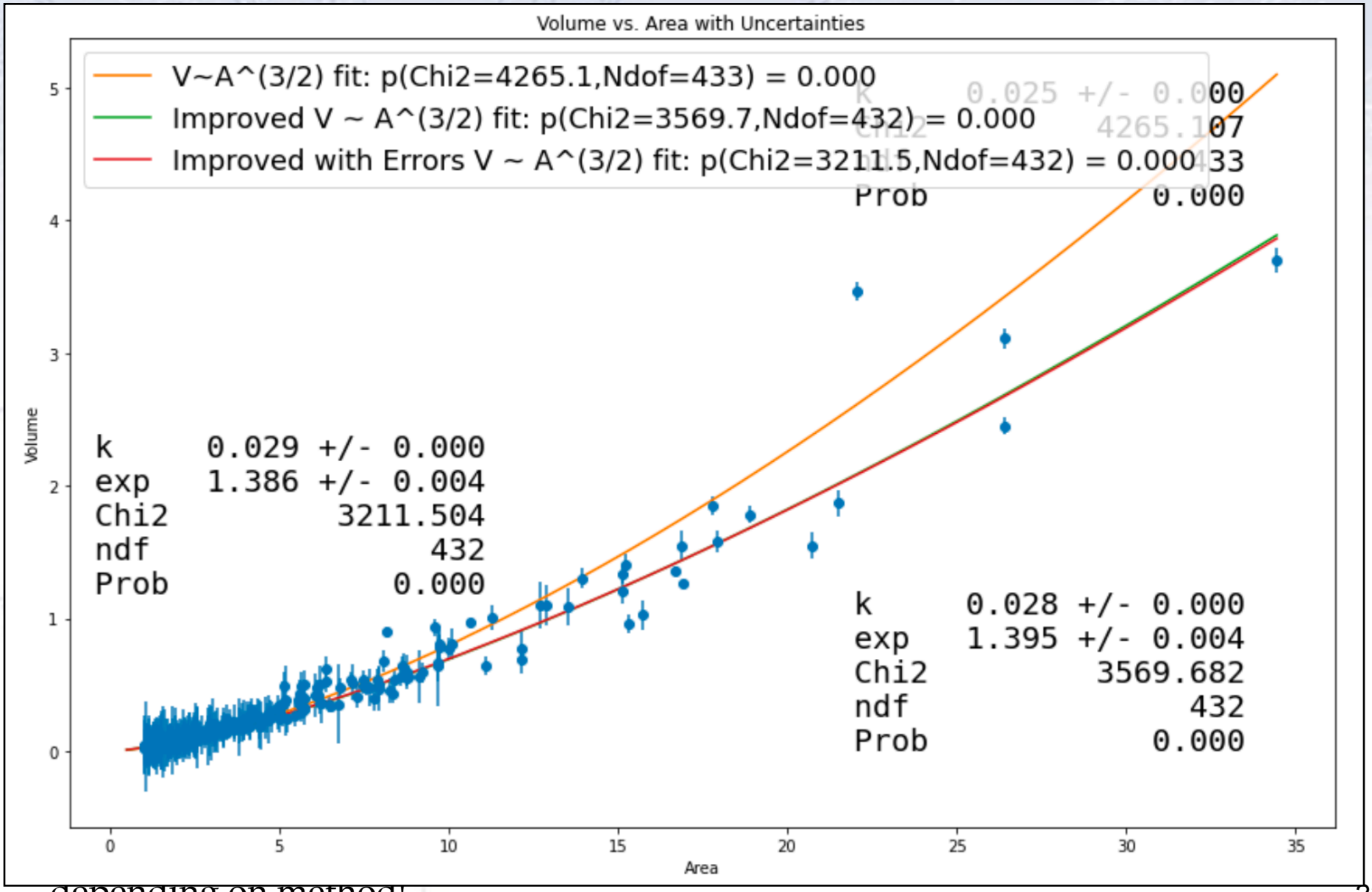
- Plot volume as a function of area. Which of the two have largest relative uncertainties?
- Fit data with the expected Area-Volume relation $V \sim A^{3/2}$. Assume no area uncertainties.
- Are you satisfied with the fit? And if not, point out its specific deficiencies.
- Fit again with improved functional form(s), and quantify the improvements.
- Redo this fit including the uncertainties in area. How large is the effect of including these?
- What volume and with what uncertainty would you expect a glacier of area 0.5 km^2 to have?

5.1.1: It is clearly the volume that contributes the most.

5.1.2: The first fit has 1 parameter.
5.1.3: But it misses the slope!
5.1.4: Floating the 3/2 is one way, but there are many.
5.1.5: Small effect, though visible.
5.1.6: Volume is about 0.01-0.02, but depending on method!



Problem 5.1



depending on method!

Your scores

General distribution

The distribution of points in the Problem Set was 70.0.
Last year, it was 70.2, so “exactly the same”.

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