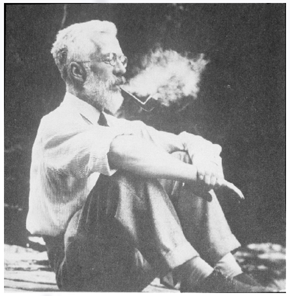
Applied Statistics

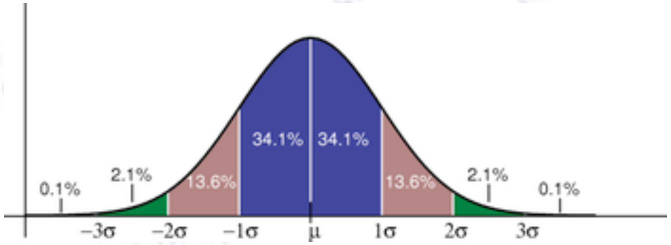
Problem Set Solution and Discussion








Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense"

Overall comments

This problem set was 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…

The solutions

Problem 1.1

Problem 1.1.1:

- The appropriate distribution is **binomial**, as N and p are fixed. Poisson is not a good approximation (N=20 is not large, and p=1/6=16% is not small).

Problem 1.1.2:

- The probability to get 7 or more 3s is:

$$P(k = 7 + |N, p) = \sum_{7}^{20} \left(\frac{20!}{(20-k)!k!} \right) (1/6)^k (1 - 1/6)^{N-k} = 0.0371$$

Problem 1.2

Problem 1.2.1:

- The fraction of positives can for each test be calculated as (binomial fractions with uncertainties):

$$f_{PCR} = \frac{N^{+}_{PCR}}{N^{all}_{PCR}} \pm \sqrt{\frac{N^{+}_{PCR}/N^{all}_{PCR}(1 - N^{+}_{PCR}/N^{all}_{PCR})}{N^{all}_{PCR}}} = 0.0239 \pm 0.0005$$

$$f_{AG} = \frac{N^{+}_{AG}}{N^{all}_{AG}} \pm \sqrt{\frac{N^{+}_{AG}/N^{all}_{AG}(1 - N^{+}_{AG}/N^{all}_{AG})}{N^{all}_{AG}}} = 0.0188 \pm 0.0008$$

These uncertainties can safely be regarded as Gaussian, as the number of positives is high (>50).

Problem 1.2

Problem 1.2.2:

- The false negative rate (FNR) is defined as the ratio $\frac{False\ negative}{Condition\ positive}$, where *Condition positive* stands for all positive people. Since we assume PCR tests have no errors, total # of people that were tested with AG tests and were positive is $f_{PCR} \times N_{AG}^{all} = 624$. Then $FNR = \frac{624 - N_{AG}^{+}}{624} = 0.213 \pm 0.018$.

Problem 1.2.3:

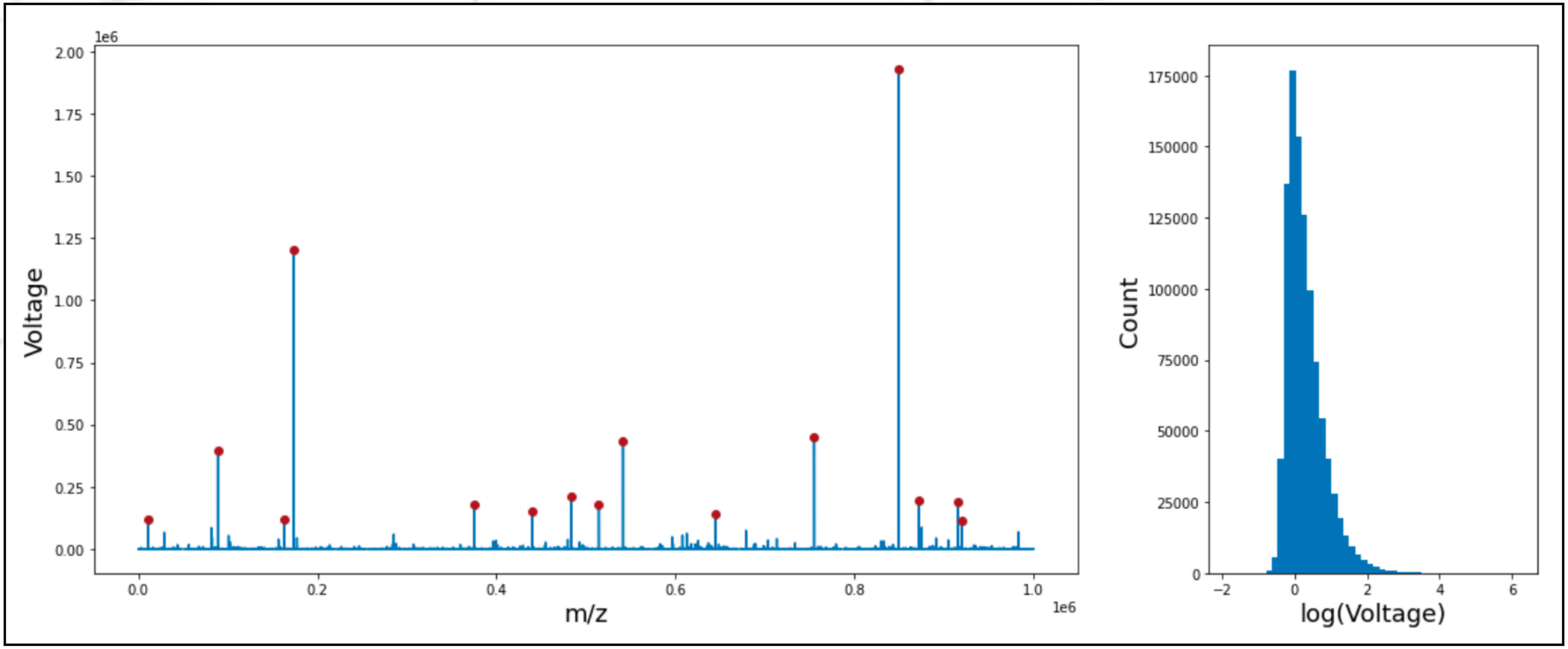
- The fraction of the Danish population truly infected can be calculated from the following equation:

$$(50.000 n_{\text{infected}}) \times 0.0002 + n_{\text{infected}}) \times (1 - 0.2) = 47 \quad \longrightarrow \quad 0.093 \pm 0.013\%$$

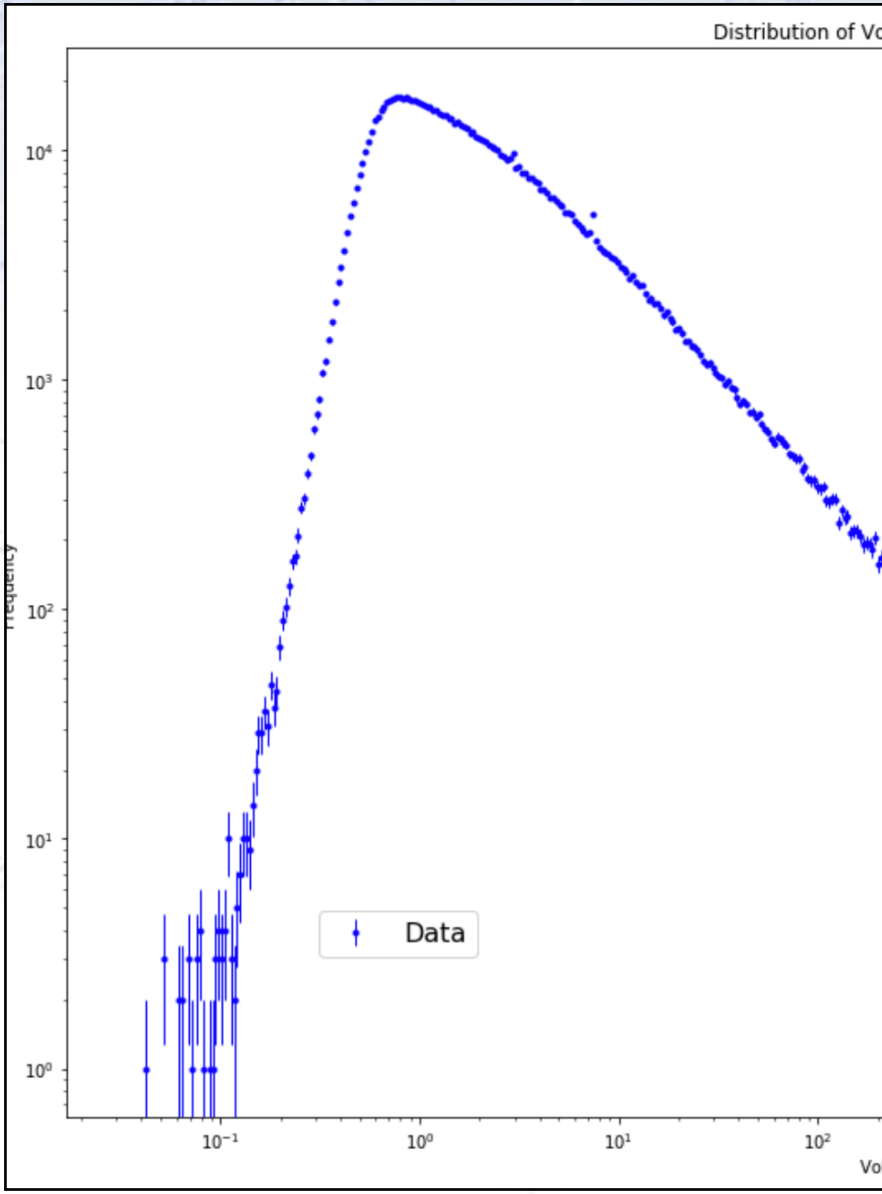
Problem 1.3

This was a hard problem for several, who did not plot a histogram.
And even those who did a histogram, did not all see the (minor) peaks, as the quality of the histogram was poor (make them large!).

In general, given many measurements, **always plot a histogram simply to get an idea of the distribution of values** (even if you don't use this afterwards).
We decided to give points for many different attempts…



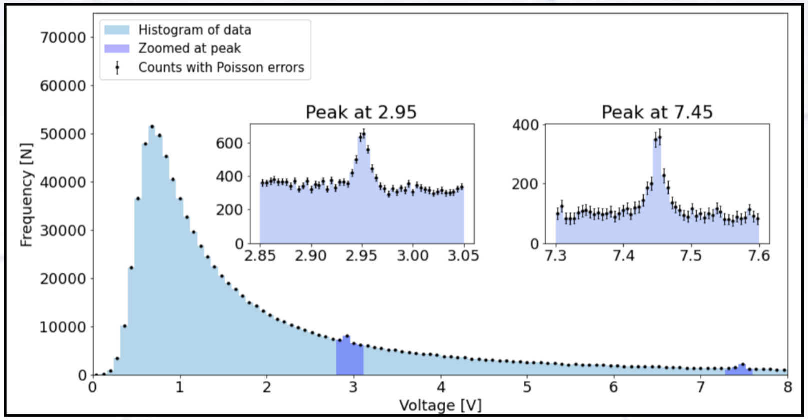
Problem 1.3

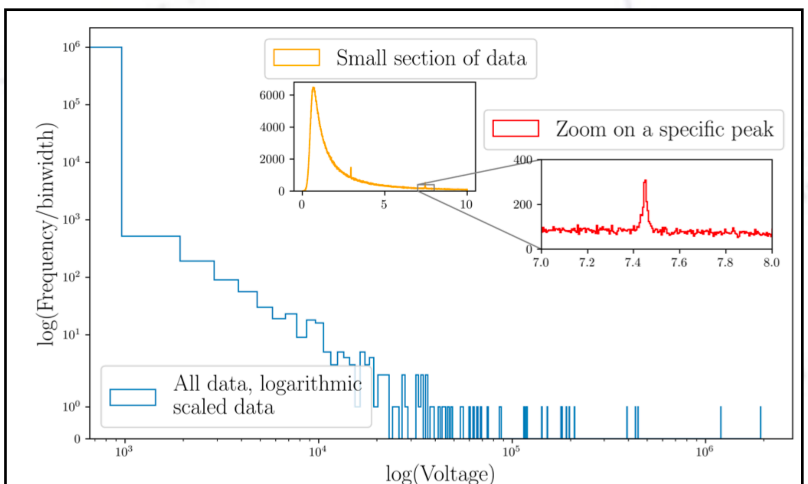


The simple good solution for a plot, is a Log-Log histogram with many bins and showing errors.
Here, the two largest peaks can also be (faintly seen).

Problem 1.3

There were (fortunately) also some very nice solutions…





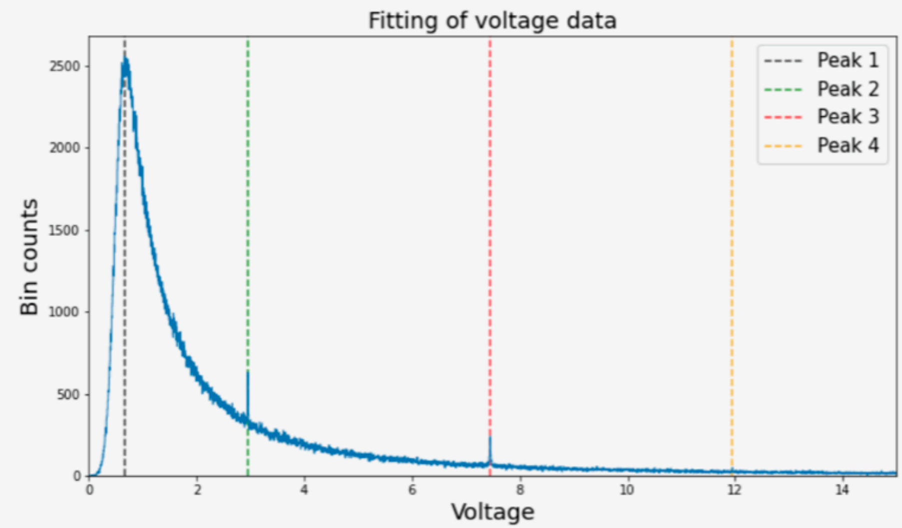
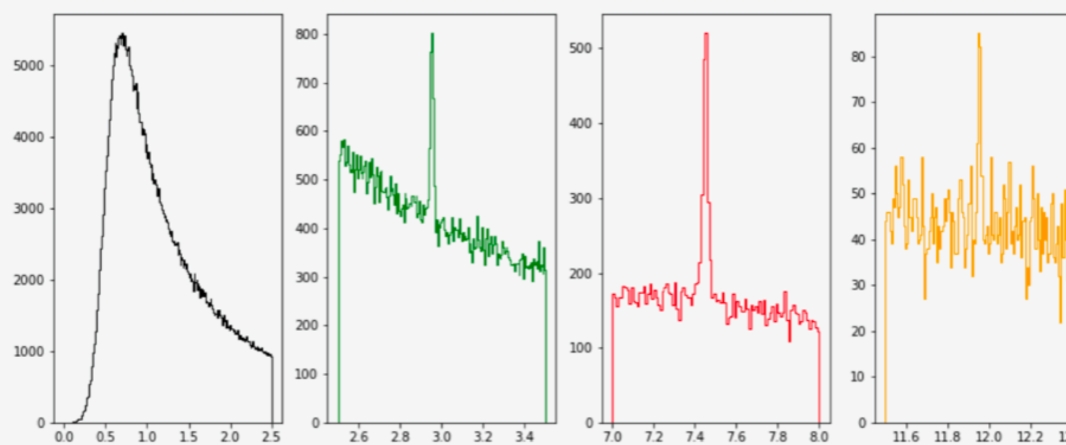
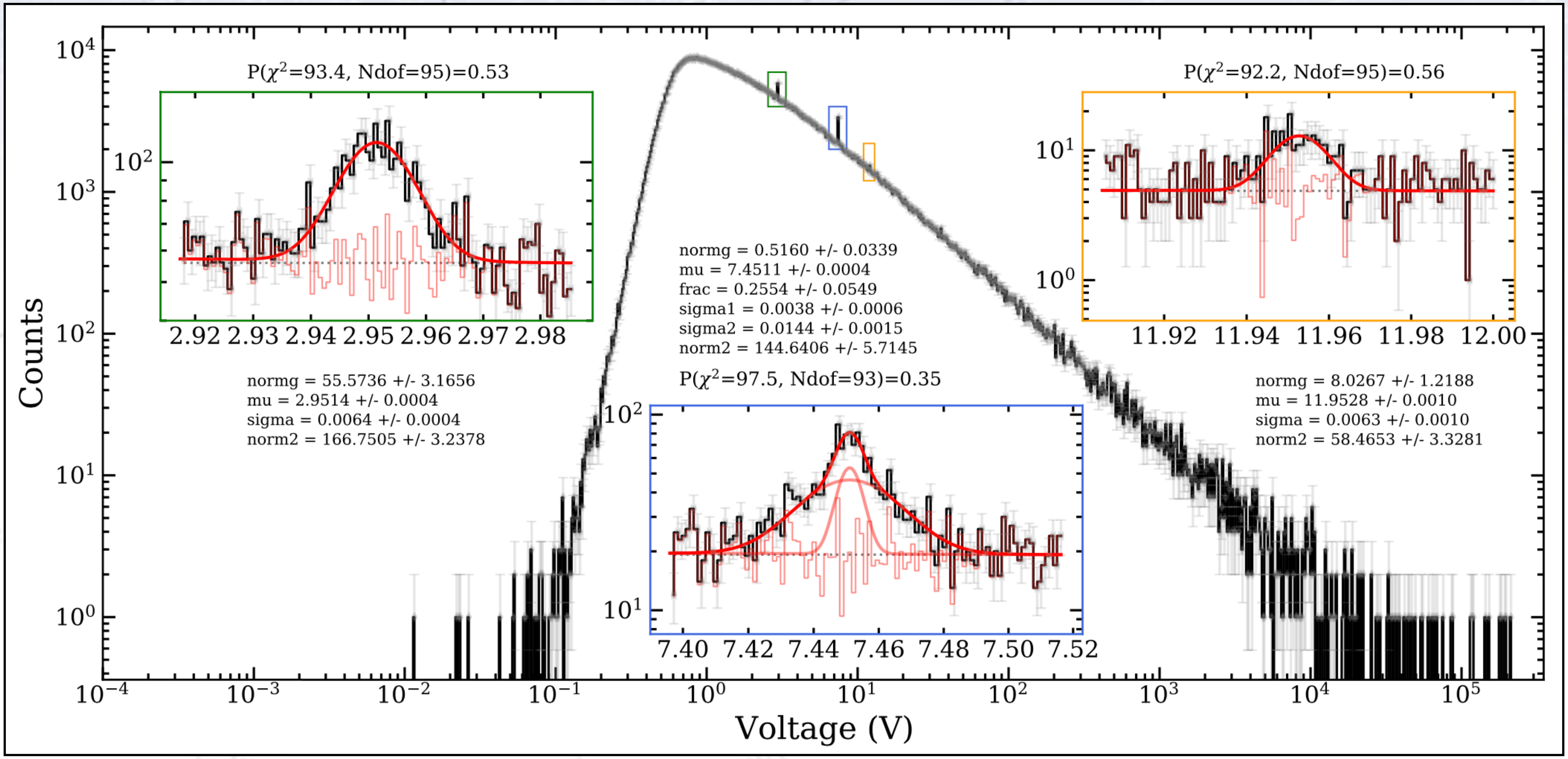


Figure 2: Whole peak



Problem 1.3

The nicest plot of them all was this:



This plot is closing in on "publishing quality"….

Problem 2.1

Problem 2.1.1:

$$\sigma(y) = \sigma(x)\sqrt{\frac{x^2}{(x^2+1)^4}}$$

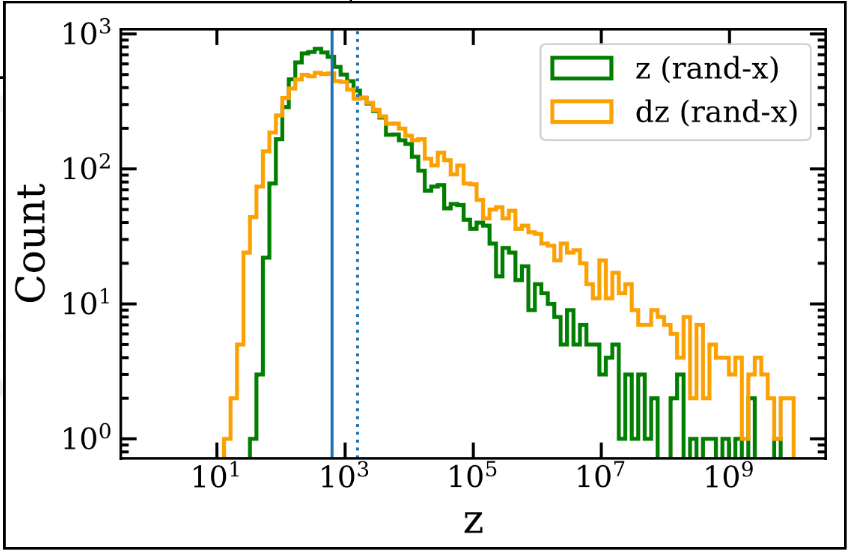
$$\sigma(z) = \sigma(x)\sqrt{\frac{1}{(1-x)^6}}$$

- This is a classic error propagation exercise. The first part is straight-forward: $y = 0.207 \pm 0.005$ and $z = 1.09 \pm 0.07$.

Problem 2.1.2:

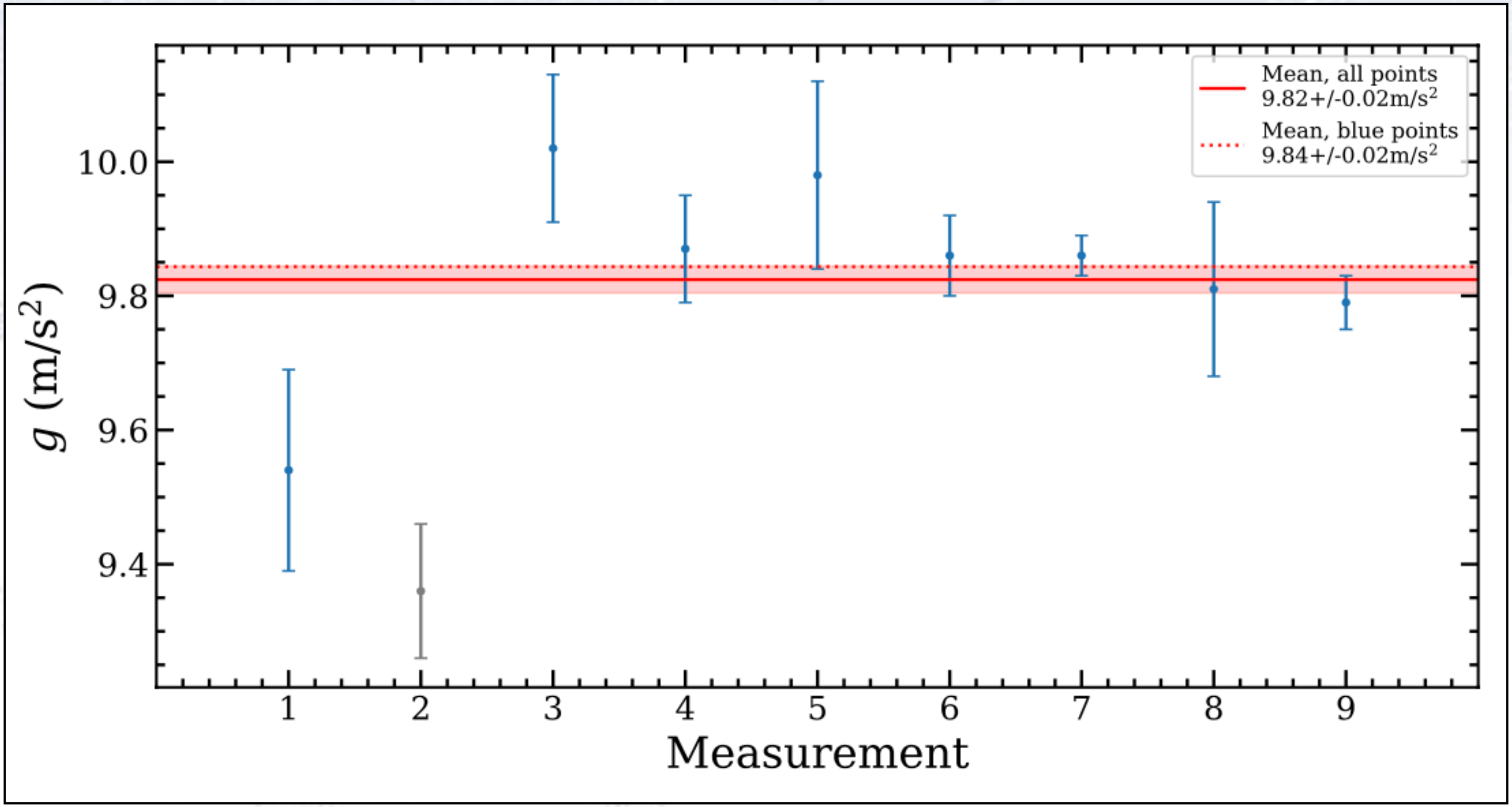
- The next part has a bit of a hiccup. While $y = 0.52 \pm 0.02$, the error propagation formula for z breaks down as the derivative is highly non-constant, as the denominator approaches 0 as x approaches 1. While the result is $z = 625 \pm 937$, it is not accurate. This must be commented on for full points.

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



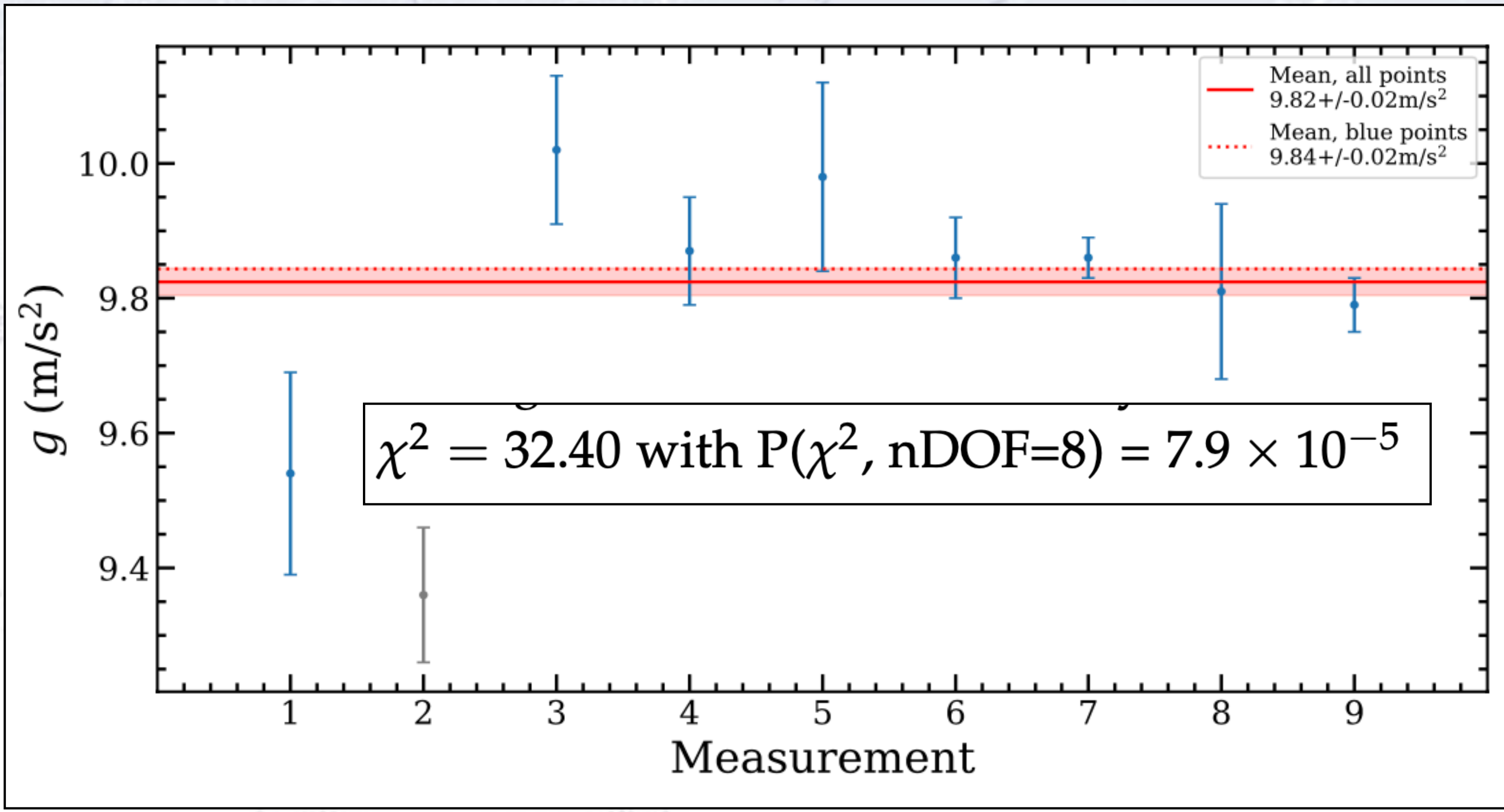
Problem 2.2

The weighted mean gives an average of 9.82 ± 0.02 m/s², but…



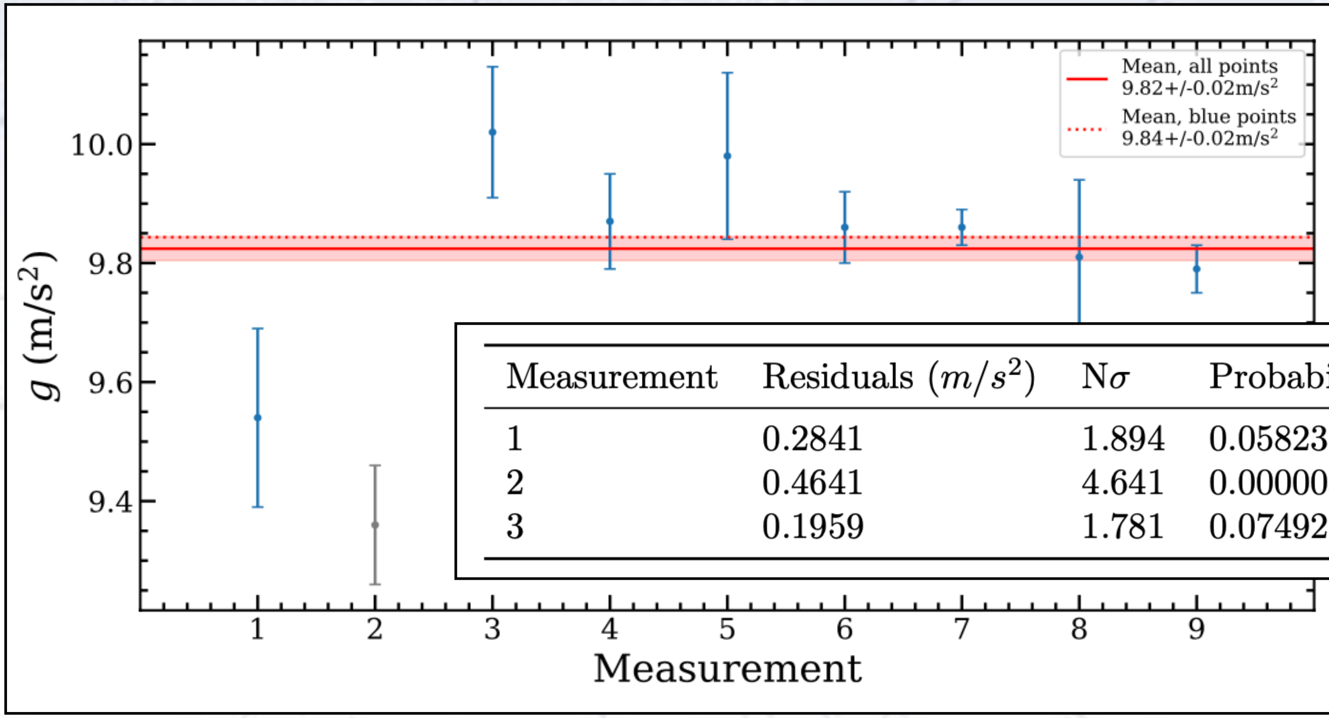
Problem 2.2

The weighted mean gives an average of 9.82 ± 0.02 m/s^2, but a very poor Chi2!



Problem 2.2

The only measurement, which is inconsistent, is measurement 2.



Measurement	Residuals (m/s^2)	Nσ	Probability	Conclusion
1	0.2841	1.894	0.05823	Accepted
2	0.4641	4.641	0.000003468	Rejected
3	0.1959	1.781	0.07492	Accepted

After removing second measurement, everything is consistent and great:

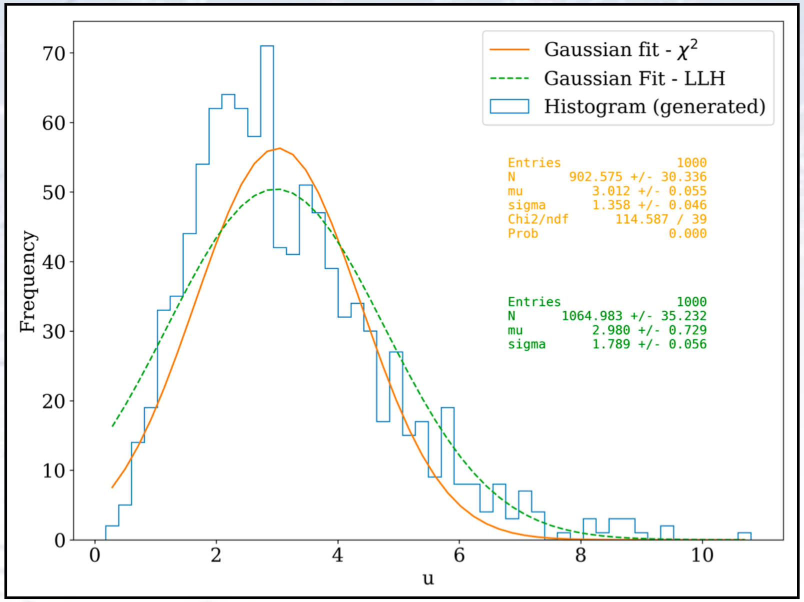
$$\chi^2 = 9.96 \text{ and a } P(\chi^2, \text{nDOF=8}) = 0.19$$

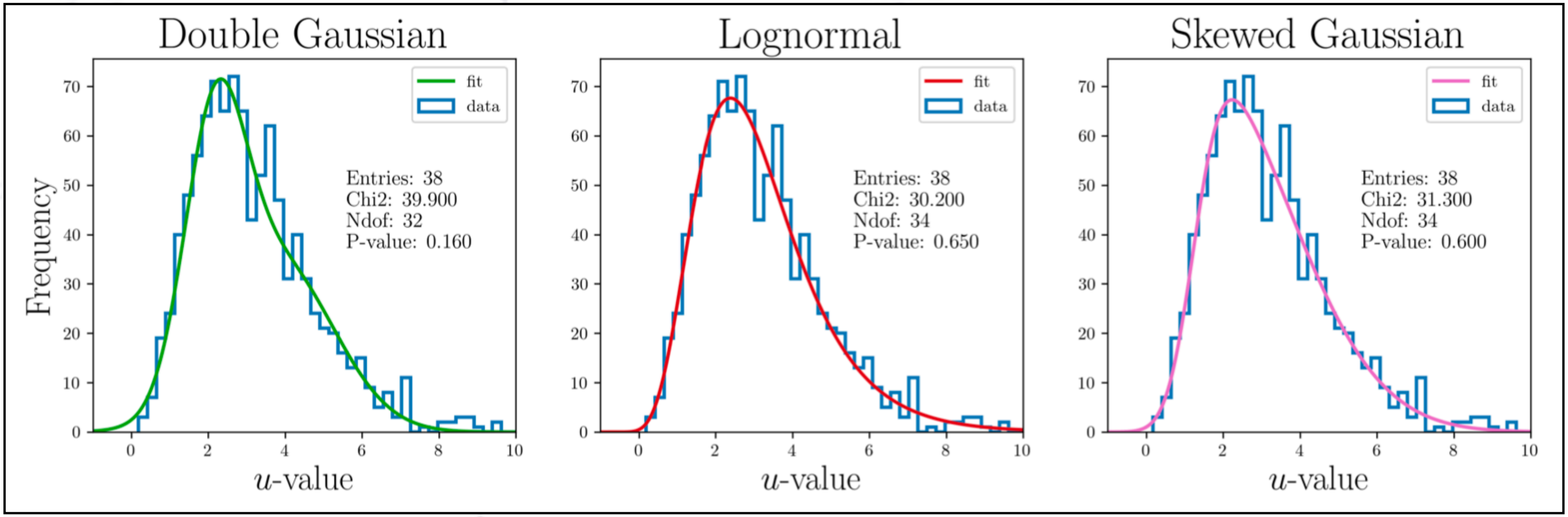
Problem 3.1

The generation of exponential numbers and thus u-values was done by ~all.

Also, fitting to a Gaussian was done by the vast majority. Few did a KS or AD test.

Many functions fits the distribution, which is in fact a Gamma distribution (and E time).





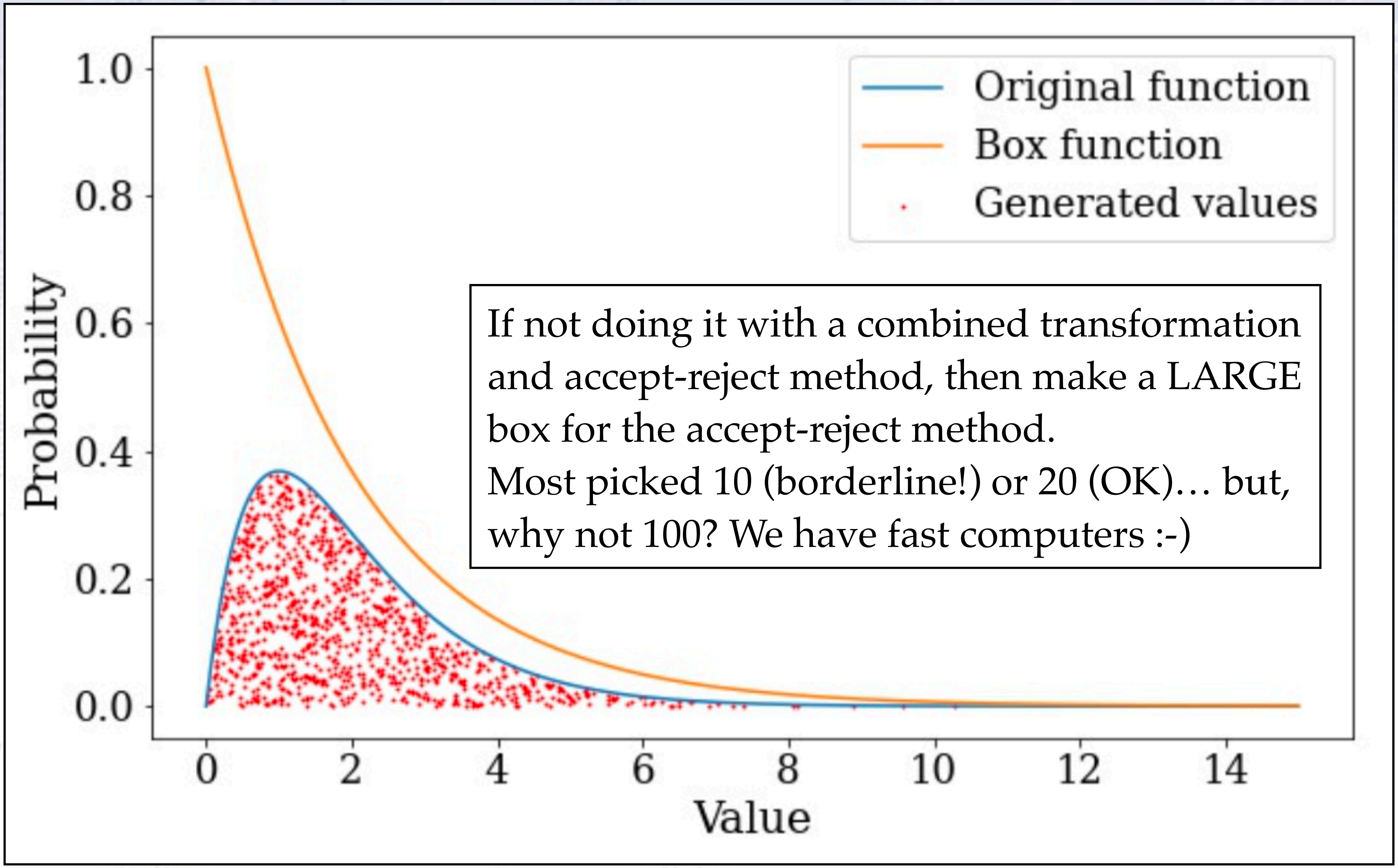
Problem 3.2

In principle, the problem can be solved with the transformation method, and the hard inversion can be solved with Labert's W function…

$$\boxed{F^{-1}(x) = -W((x-1)/e) - 1}$$

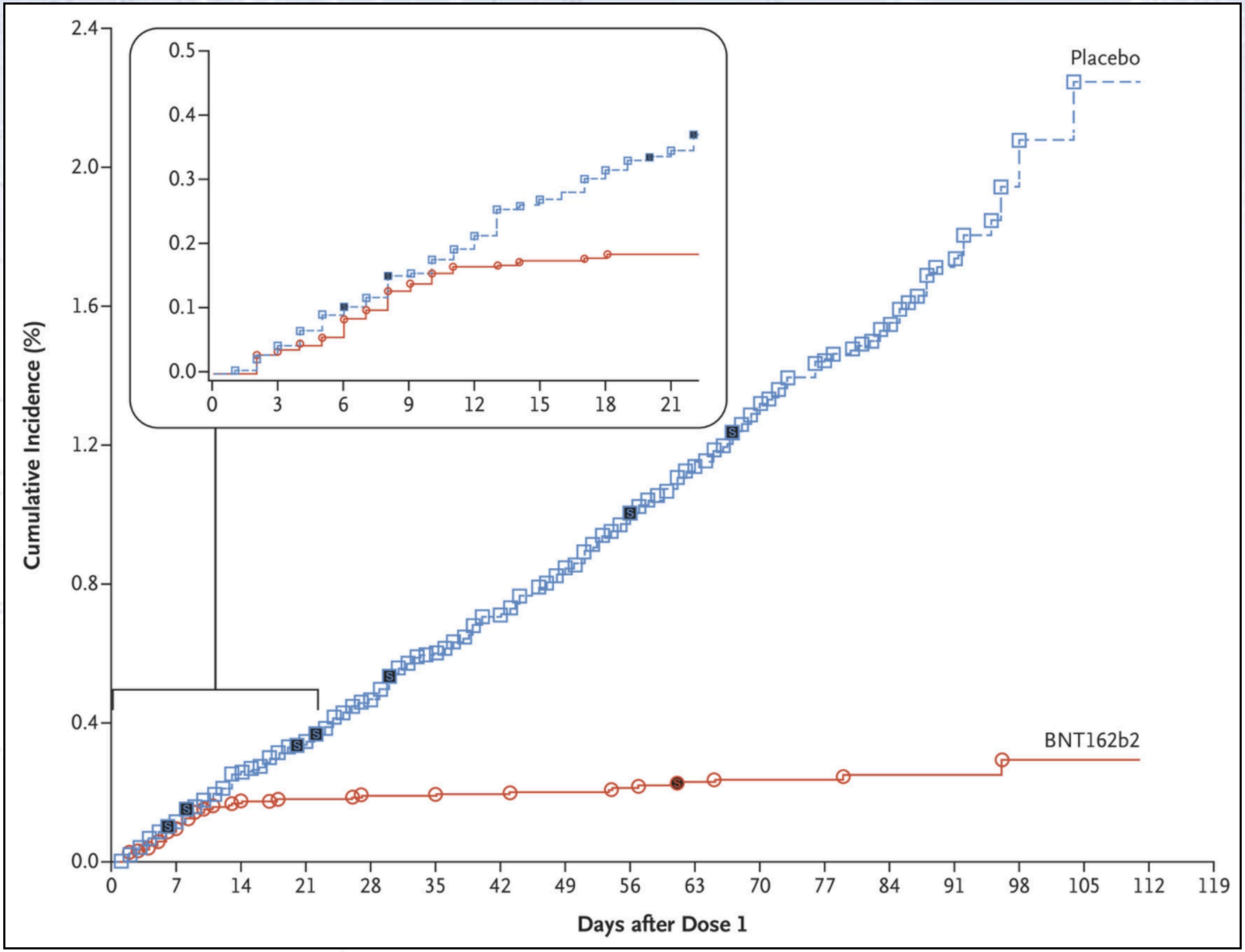
But that might be slightly beyond the math of most of us!

Problem 3.2



Notice, that since there is an EVEN number of entries, the median is not perfectly well defined. Possibly, one could take the average of the 500th and 501st entry.

Problem 4.1 - inspiration



Problem 4.1

Problem 4.1.1:

- It is a good assumption, that the sample size is equal for the two experiments. The two numbers are to a very good approximation Poisson distributed (i.e. the uncertainty is the square root), and to a rough approximation, these errors are Gaussian.

The null-hypothesis H_0 ="the vaccine has no effect" implies that the two experiments were drawn from the same (Poisson) distribution, which has a mean between 8 and 162. With the Gaussian assumption, we thus conduct a two-sample test:

$$z_{positives} = \frac{168 - 8}{\sqrt{168 + 8}} = 12 \tag{3}$$

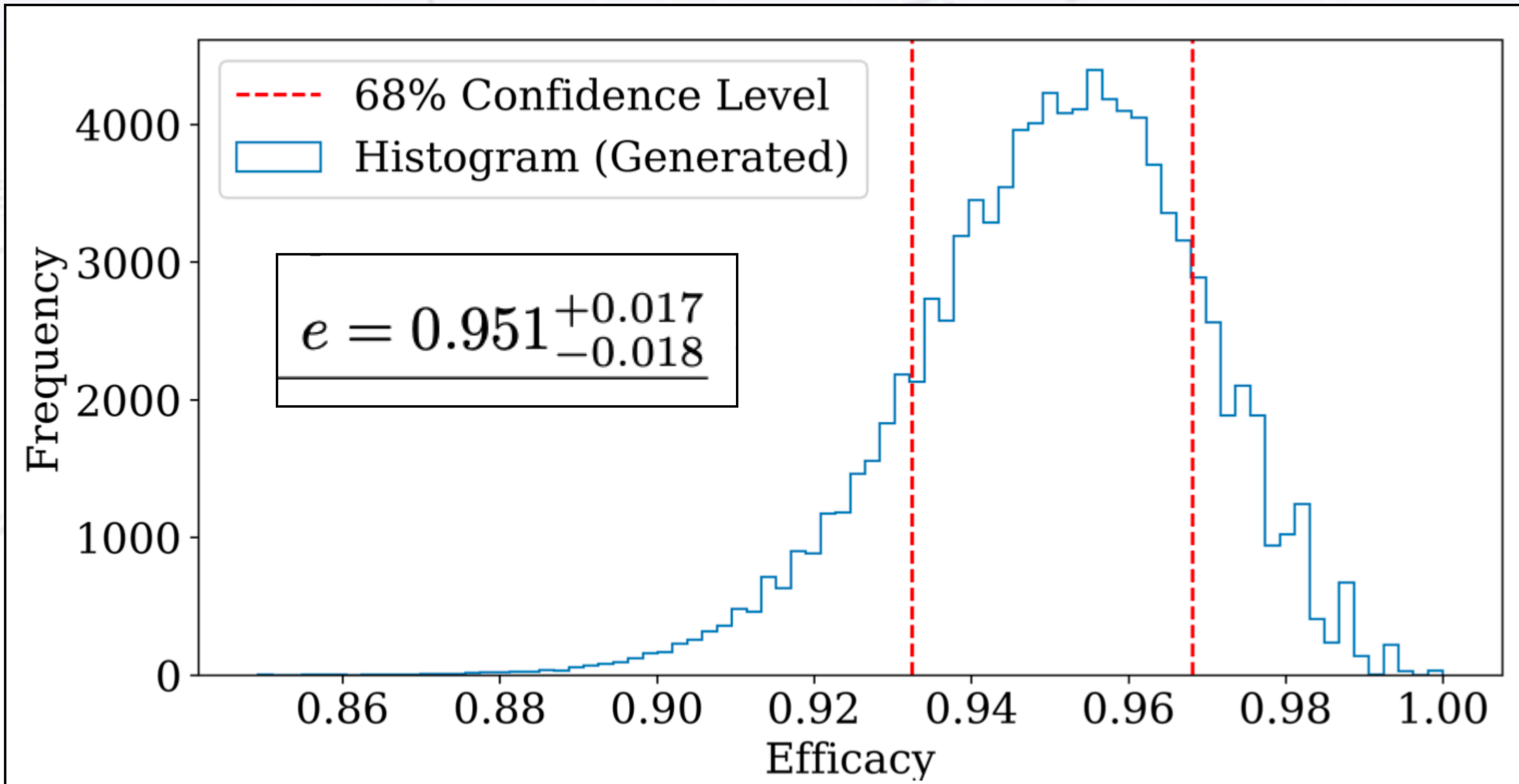
The p-value of this separation is the double-sided integral of the unit-gaussian, computed outside the $z_{positives}$ boundary:

$$p - value(z_{positives}) = 2 * \int_{-\infty}^{z_{positives}} pdf_{gauss}(x)\, dx = 7.4e^{-34} << 1 \tag{4}$$

Problem 4.1

The confidence interval (CI) can be approximated using the Gaussian approximation, which gives almost full points.

To get a precise CI, simulation is the easiest. Since the Poisson is asymmetric (especially for λ=8), so is the resulting CI.



Problem 4.1

The Fisher's exact test can actually be used for both 4.1.1, and 4.1.3, but especially in the latter case, it is really useful:

$$p = \frac{\binom{21720}{8}\binom{21728}{162}}{\binom{43448}{170}} = \frac{170!43278!21720!21728!}{8!162!21712!21566!43448!} \approx 7.6663 \times 10^{-39}$$

Again, the result is VERY clear - the vaccine works!!!

For the severe cases (i.e. low statistics), this test is really useful, as the Gaussian approximation is…. well, an approximation:

$$p = \frac{\binom{21720}{1}\binom{21728}{9}}{\binom{43448}{10}} = \frac{10!43438!21720!21728!}{1!9!21719!21719!43448!} \approx 7.6663 \times 10^{-39} \approx 0.00977$$

Problem 4.1

The Fisher's exact test can actually be used for both 4.1.1, and 4.1.3, but especially in the latter case, it is really useful:

$$p = \frac{\binom{21720}{8}\binom{21728}{162}}{\binom{43448}{170}} = \frac{170!43278!21720!21728!}{8!162!21712!21566!43448!} \approx 7.6663 \times 10^{-39}$$

Again, the result is VERY clear - the vaccine works!!!

For the severe cases (i.e. low statistics), this test is really useful, as the Gaussian approximation is…. well, an approximation:

$$p = \frac{\binom{21720}{1}\binom{21728}{9}}{\binom{43448}{10}} = \frac{10!43438!21720!21728!}{1!9!21719!21719!43448!} \approx 7.6663 \times 10^{-39} \approx 0.00977$$

We of course recognise copy-and-paste-errors :-)

Problem 4.2

Problem 4.2.1:

- The number of aces will follow a binomial distribution with $p = 4/52$ and $n = 4$, as displayed in figure 4. The chance of getting 3 aces or more is obtained by summing the probabilities for 3 aces and 4 aces: $1.7 \cdot 10^{-3} + 3.5 \cdot 10^{-5} = 0.0017$.

Problem 4.2.2:

- Drawing cards without replacement corresponds to a hypergeometric distribution, with the total number of objects $M = 52$, the total number of aces $n = 4$ and number of draws $N = 4$. The chance of getting 3 aces or more is obtained by summing the probabilities for 3 aces and 4 aces: $7.1 \cdot 10^{-4} + 3.7 \cdot 10^{-6} = 0.00071$. The problem can also be calculated using a combination of binomials.

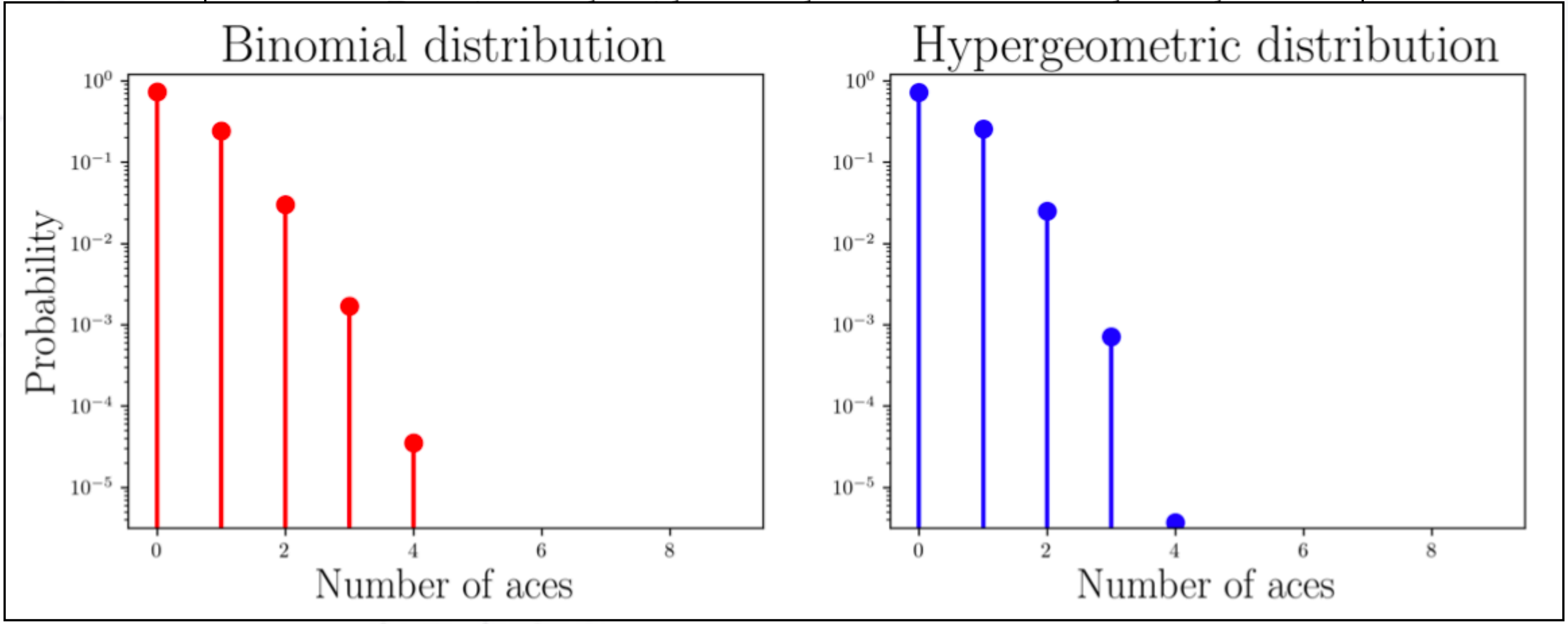
draw number is not ace/draw prob	draw 1	draw 2	draw 3	draw 4	total
4	4/52	3/51	2/50	48/49	1.773×10^{-4}
3	4/52	3/51	48/50	2/49	1.773×10^{-4}
2	4/52	48/51	3/50	2/49	1.773×10^{-4}
1	48/52	4/51	3/50	2/49	1.773×10^{-4}
all aces	4/52	3/51	2/50	1/49	3.694×10^{-6}
Result	7.129×10^{-4}				

Problem 4.2

Problem 4.2.1:

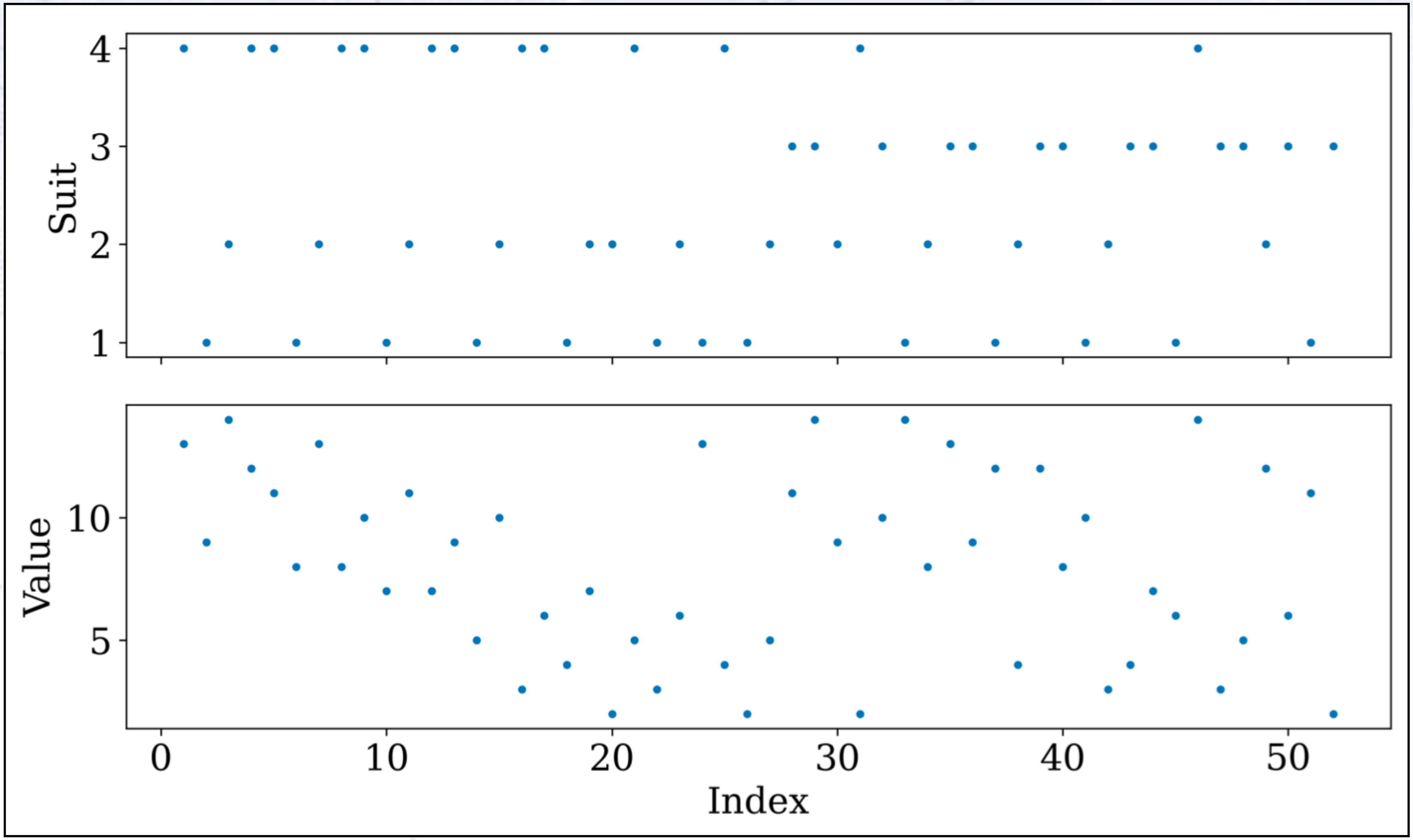
- The number of aces will follow a binomial distribution with $p = 4/52$ and $n = 4$, as displayed in figure 4. The chance of getting 3 aces or more is obtained by summing the probabilities for 3 aces and 4 aces: $1.7 \cdot 10^{-3} + 3.5 \cdot 10^{-5} = 0.0017$.

Problem 4.2.2:



Problem 4.2

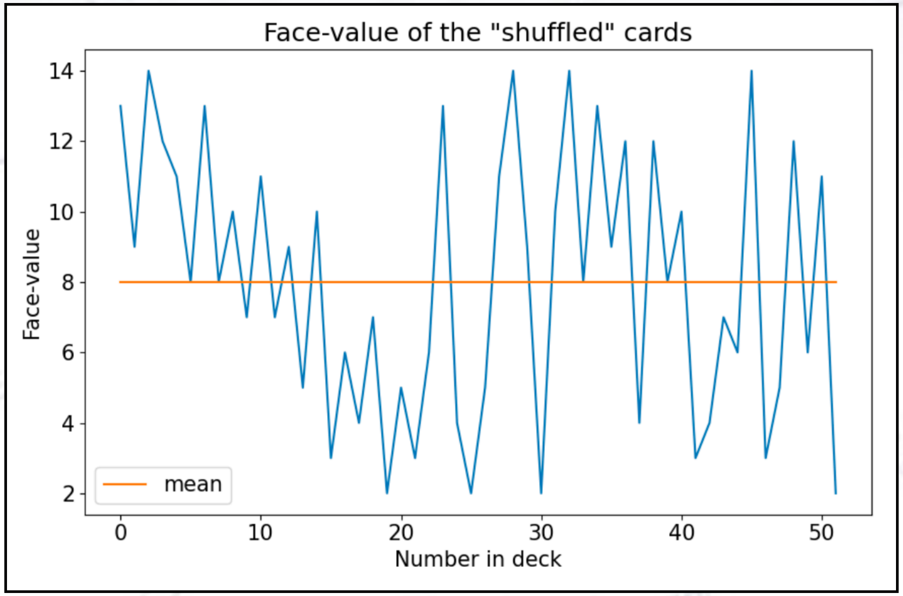
Plotting the data is **always** a good idea, as your eyes are very good at seeing patterns in low (< 3) dimensions. Looking, there are clearly patterns.

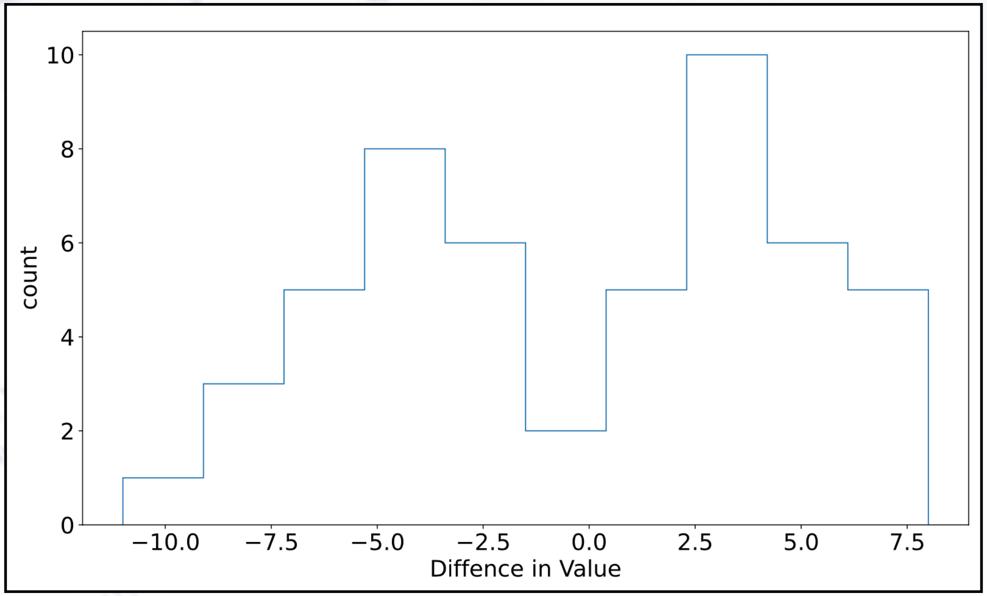


Problem 4.2

Plotting the values, it seems that every second card is higher than the next one. How to test if this is more pronounced than in a shuffled deck?

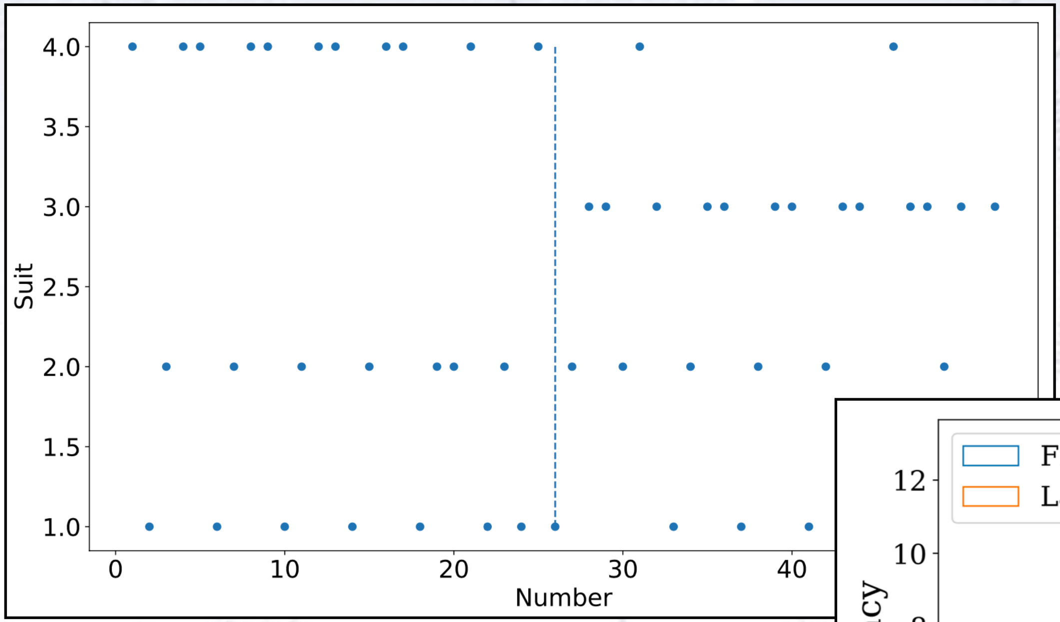
Well, plotting the distribution of differences, one gets a histogram, the distribution of which is known for a shuffled deck. From here, it is a KS test!

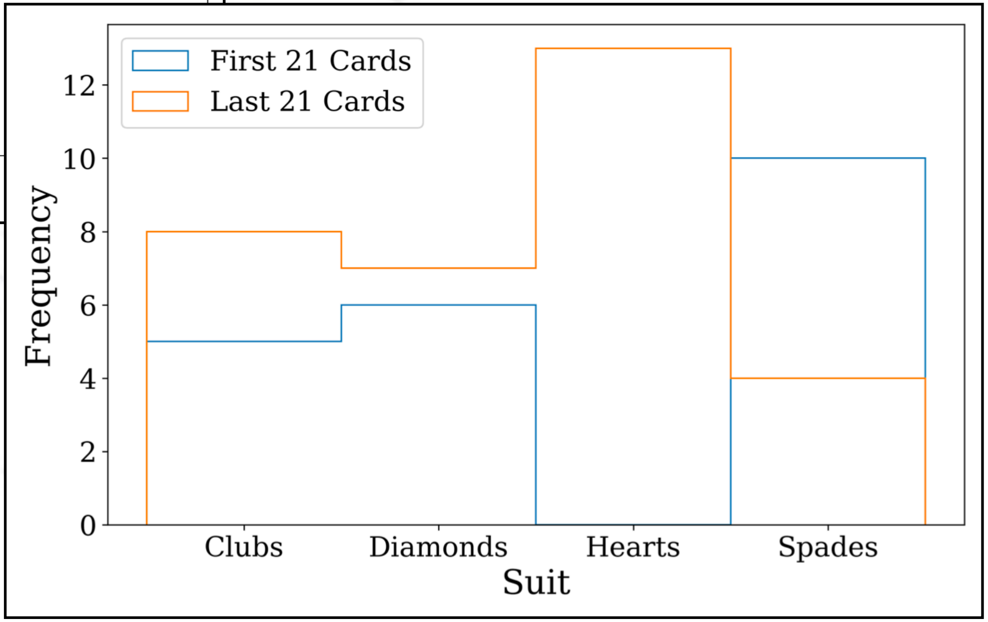




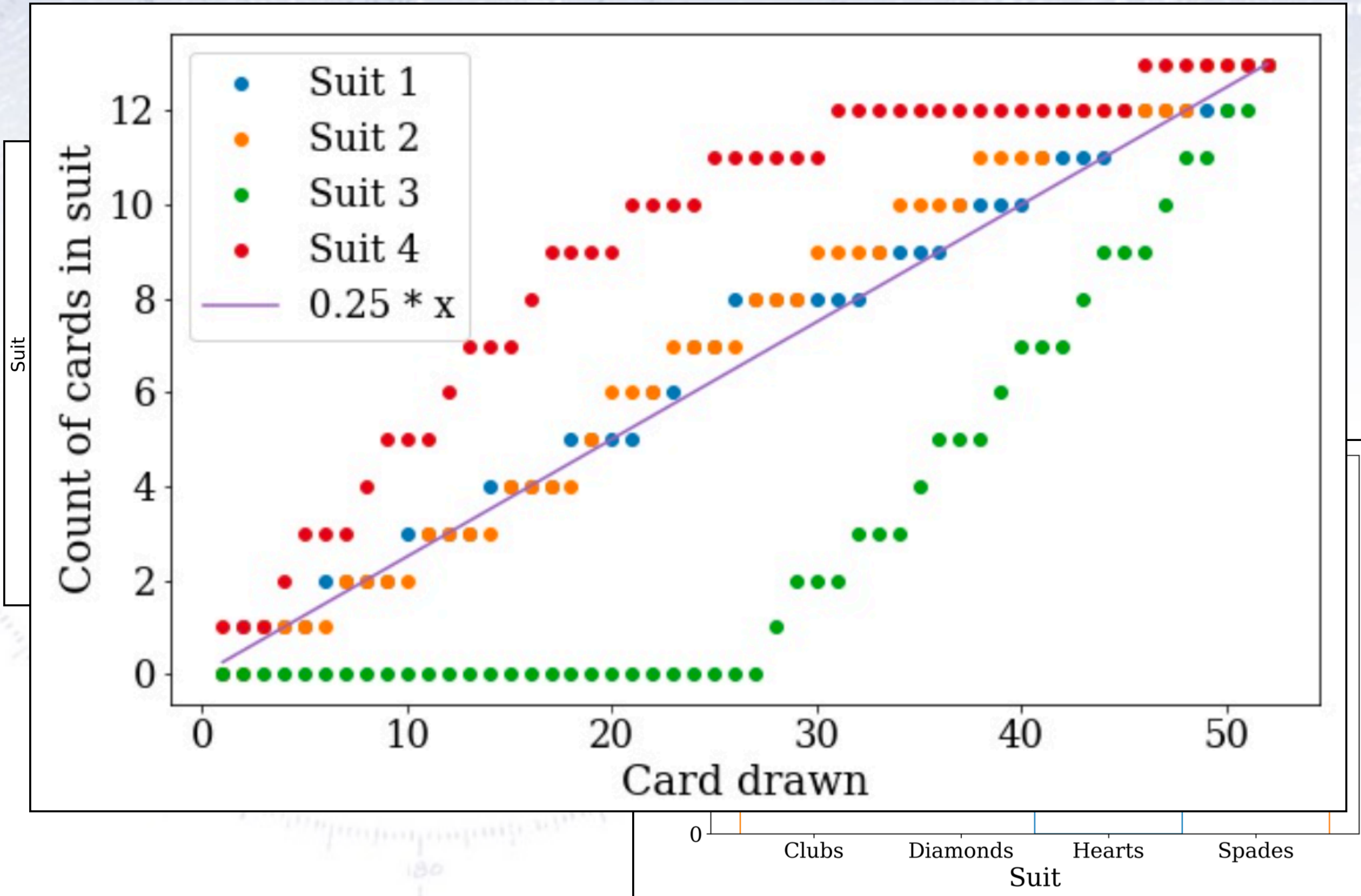
Problem 4.2

The suits can be tested by showing the distribution of the first and last half of the deck. As it happens, there are 0 hearts in the first half, which can be tested!



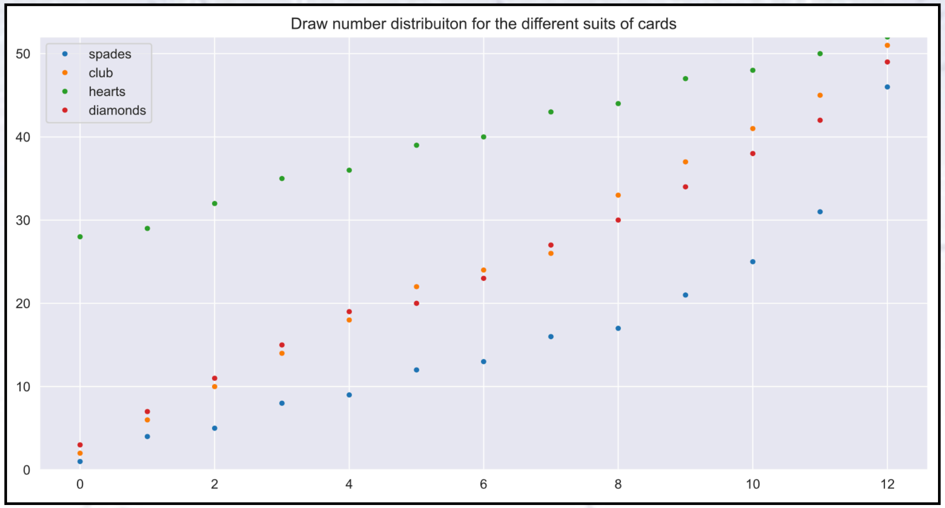


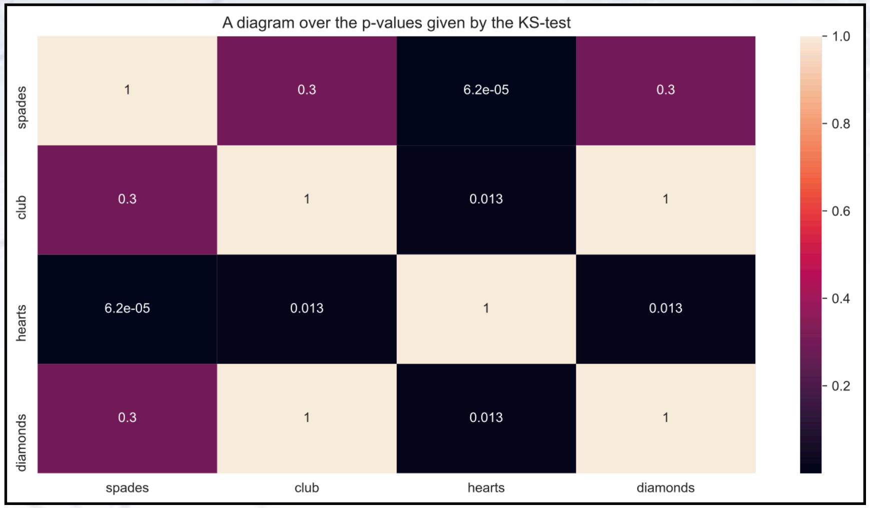
Problem 4.2

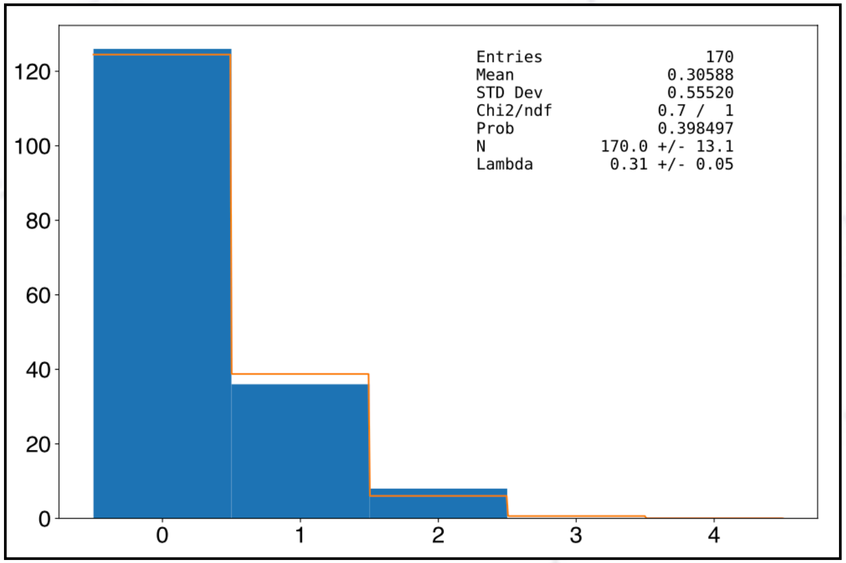

Suit 1
Suit 2
Suit 3
Suit 4
0.25 * x
Count of cards in suit
Card drawn
0
2
4
6
8
10
12
0
10
20
30
40
50
Suit
0
Clubs
Diamonds
Hearts
Spades
Suit

Problem 4.2

More tests…





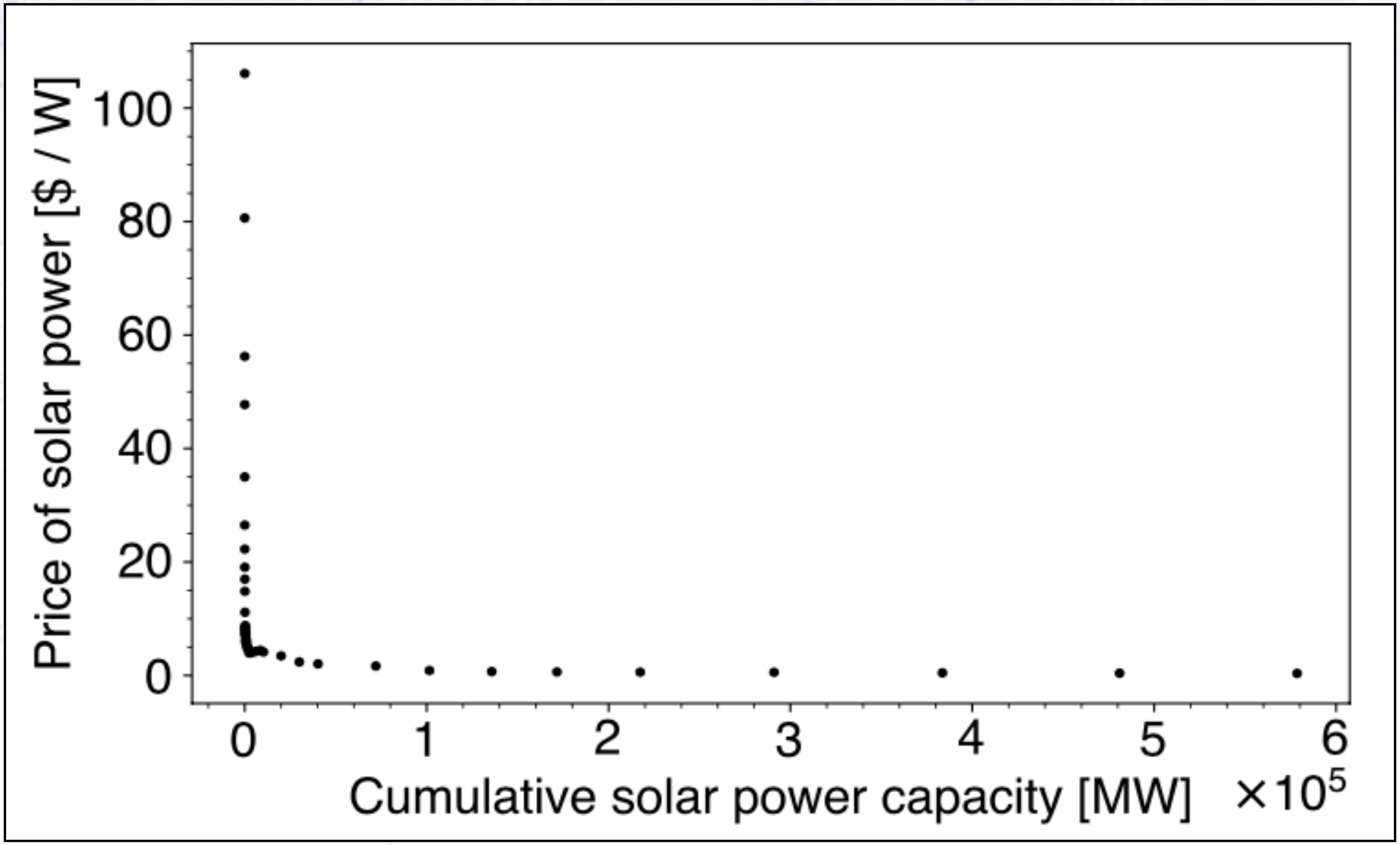


Not all tests yields a result showing “unshuffled”. Here, the value of two consecutive cards are considered. There are 52 such with 170 possibilities, so most should be zero, and only few should be two or more. And that is how they are distributed.

Problem 5.1

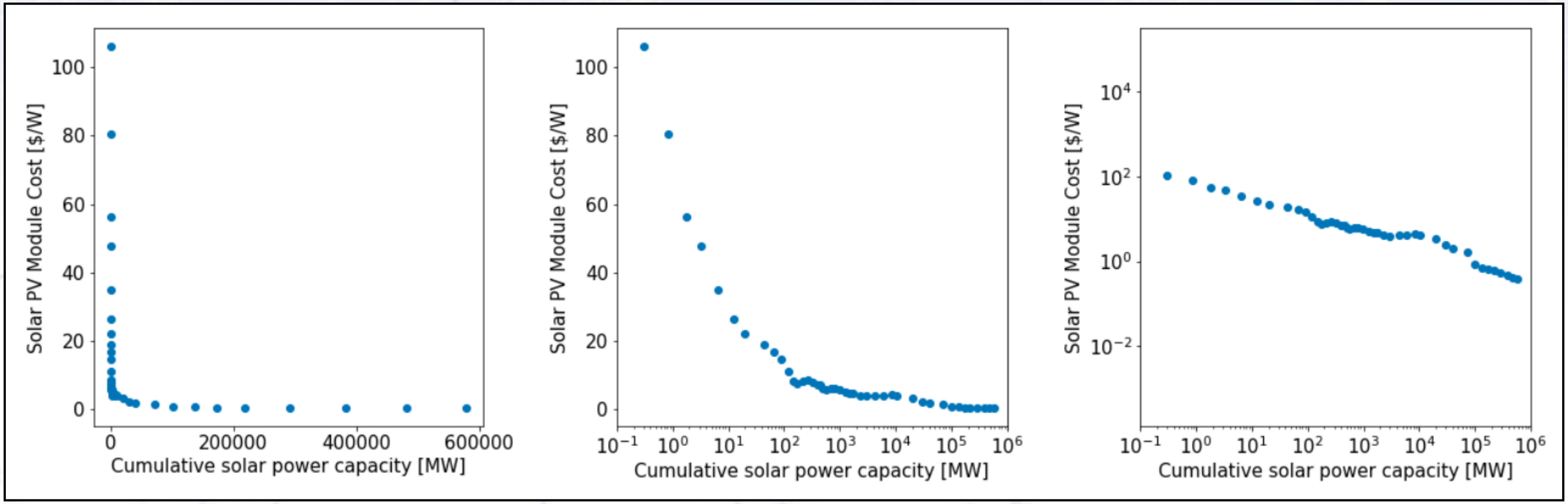
Plotting is an art, and you should give it a least a little thought.

The below example has nice labels, but a poor choice axis…



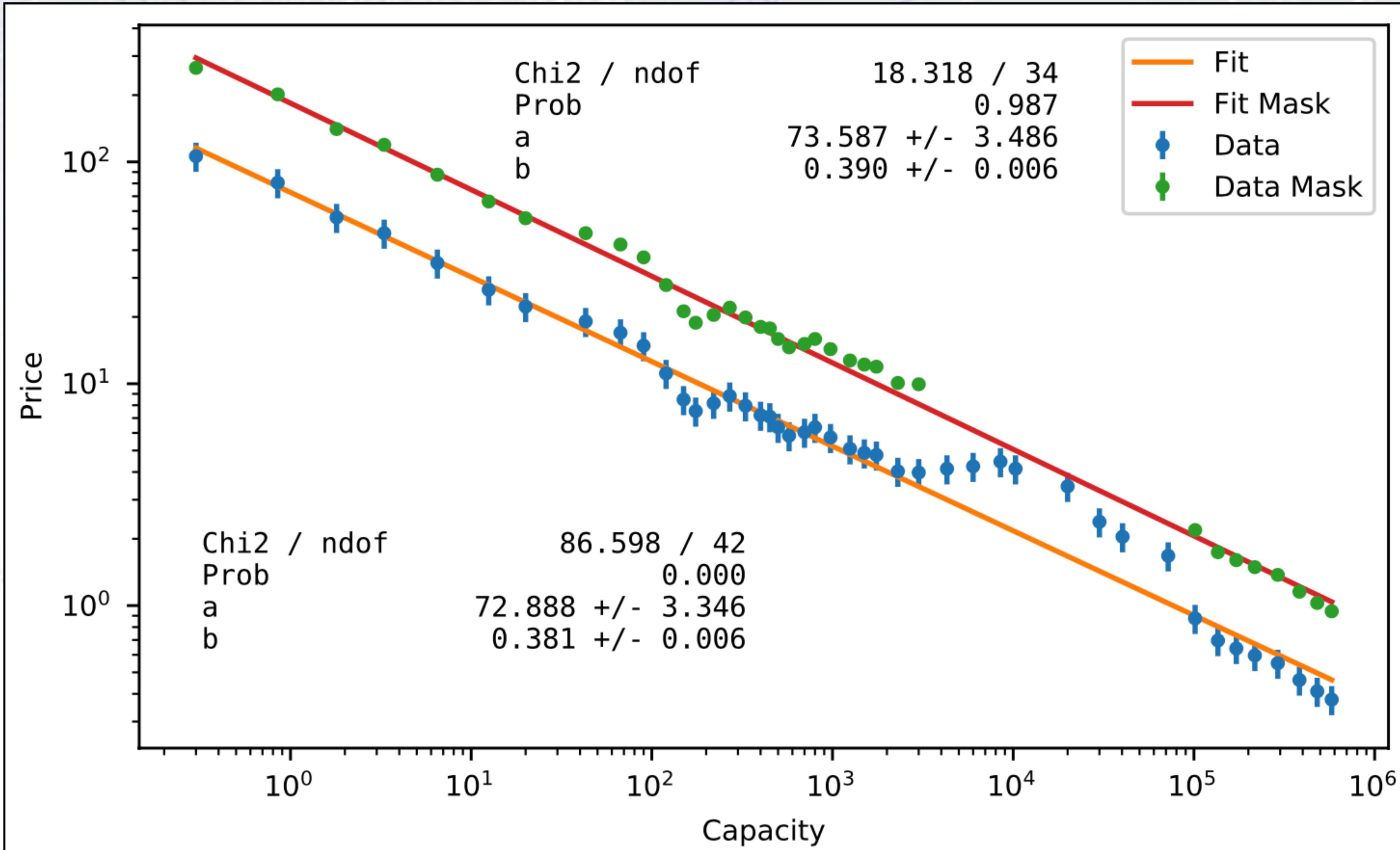
Problem 5.1

Here is a quick test of different types of axis, and given a power low fit, the log-log plot is clearly preferable.



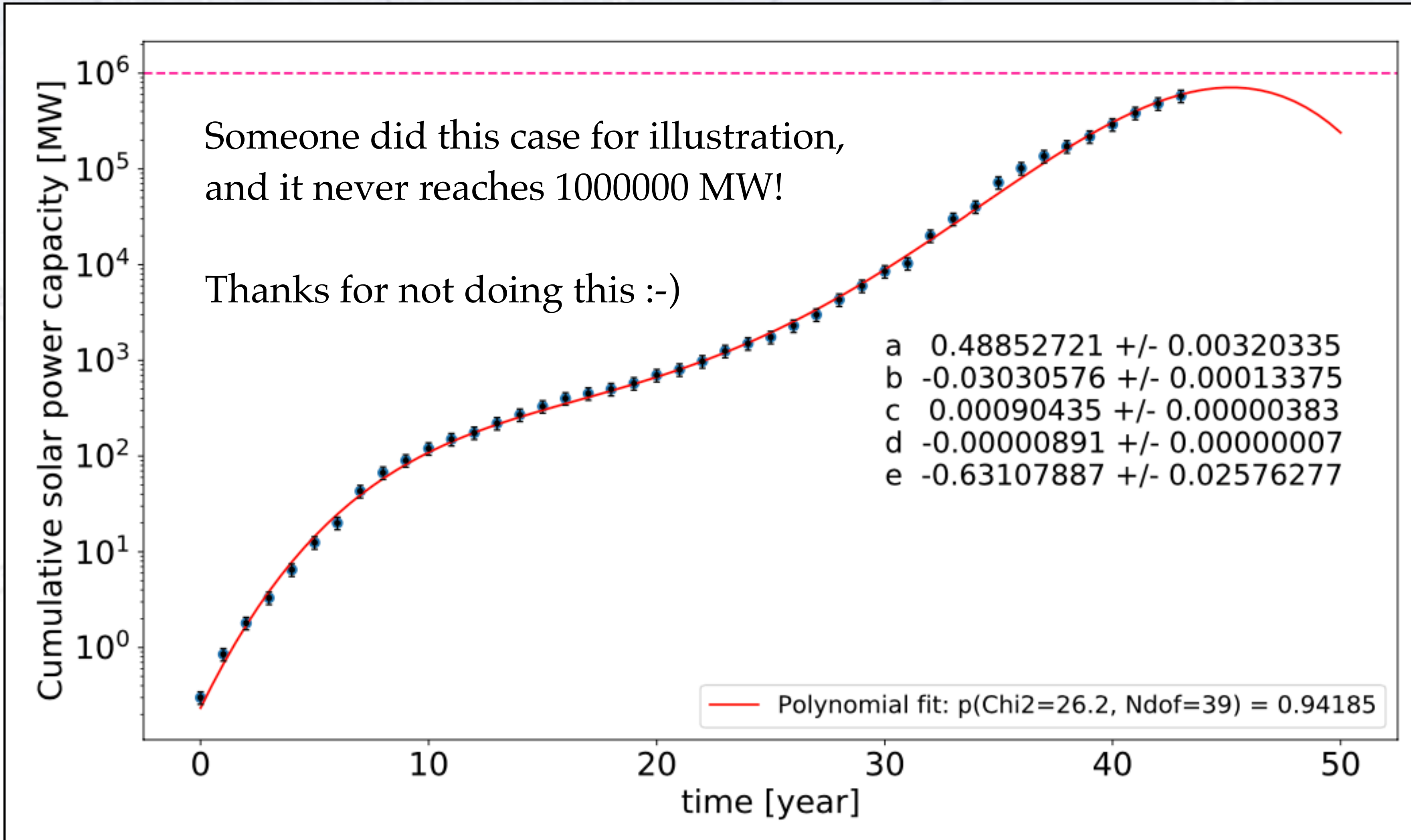
Problem 5.1

The fit is poor, except if you exclude the years 2003-2010 (oil prices high!):



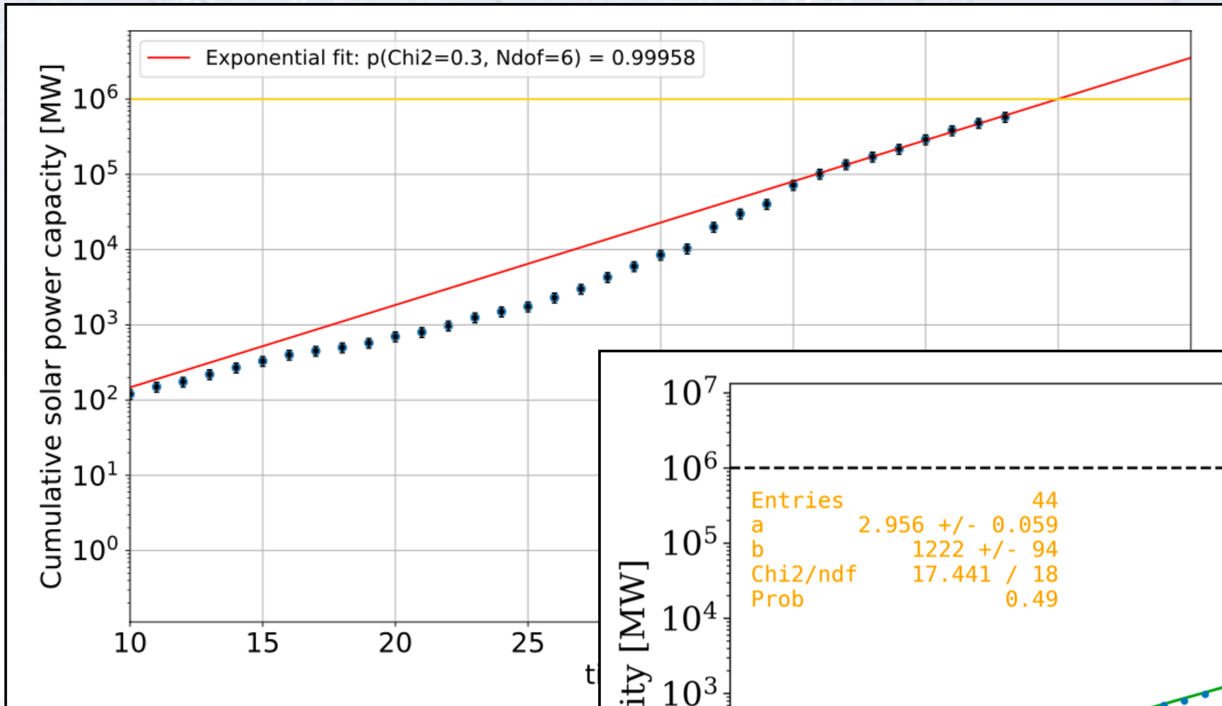
Problem 5.1

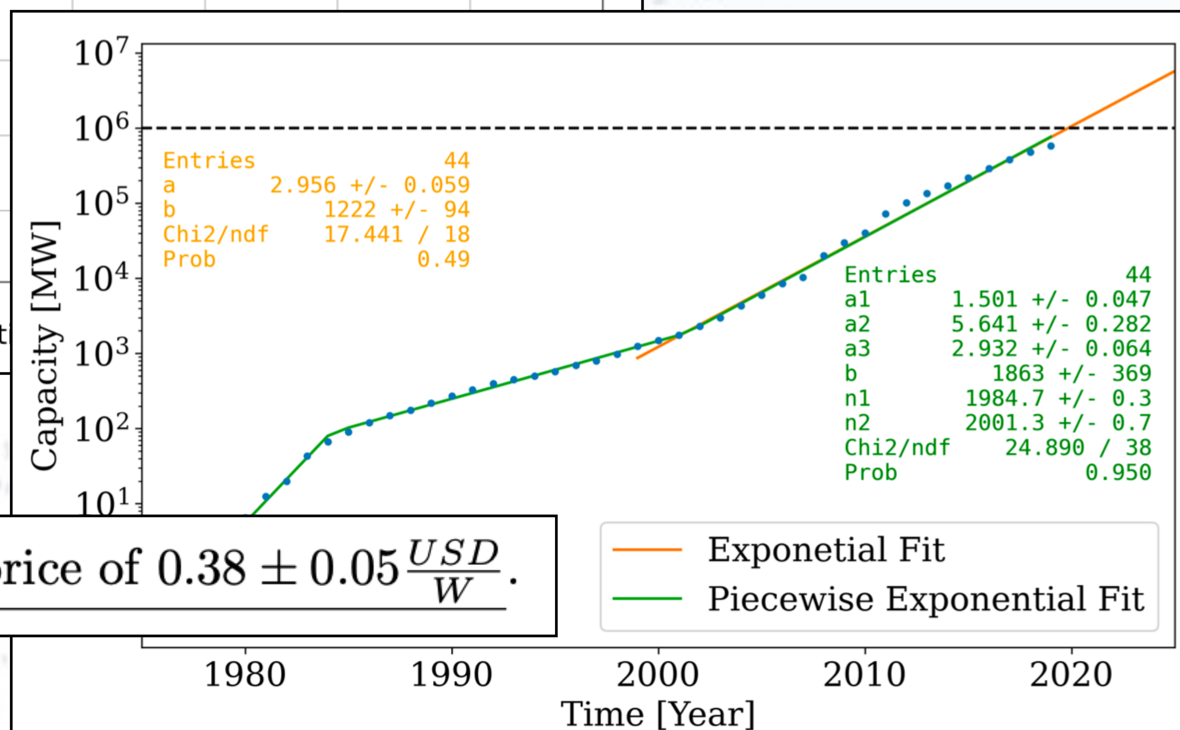
Careful with extrapolating models into the future…. don't use a polynomial!



Problem 5.1

Several ways of extrapolation…

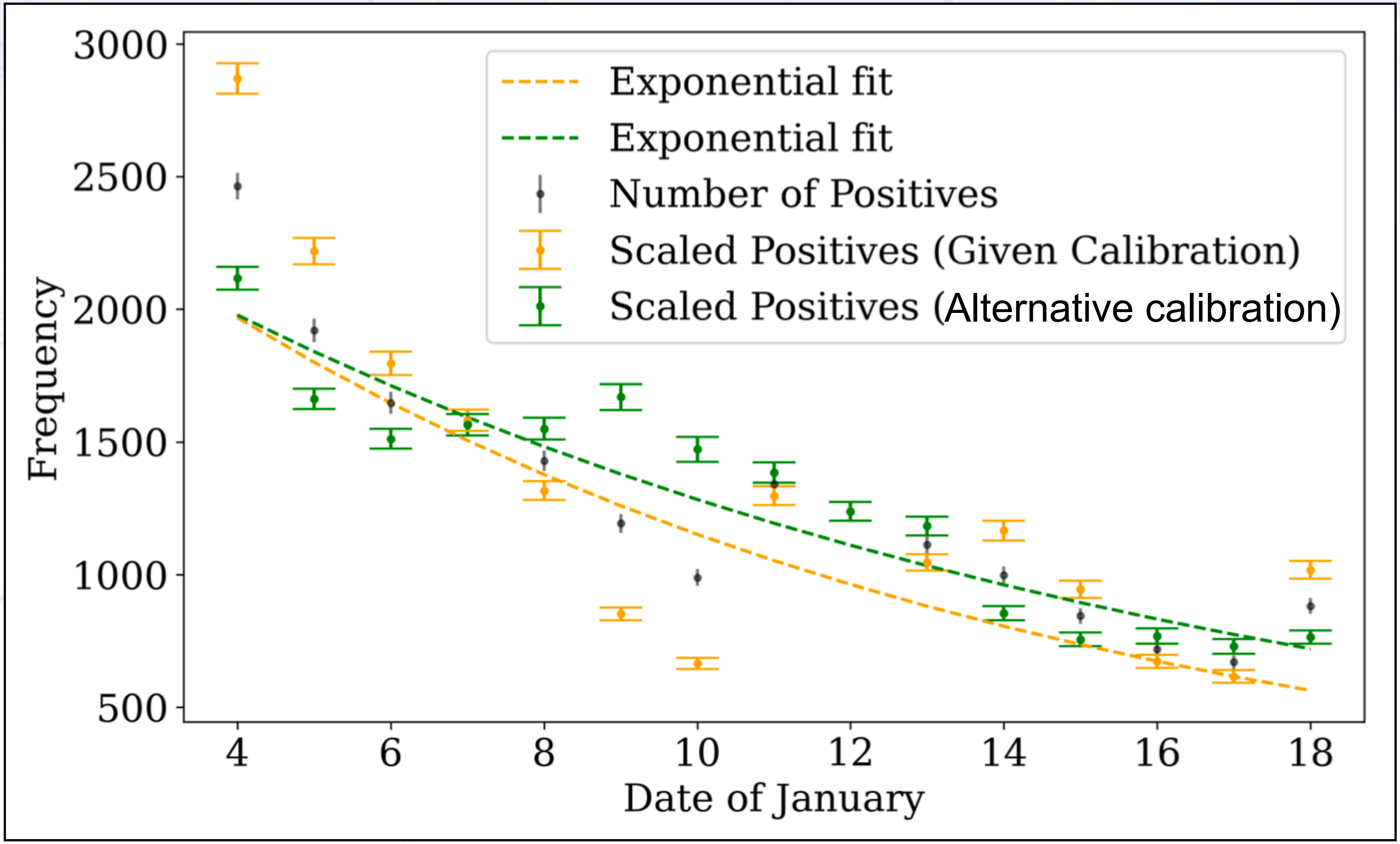




This yields an expected price of $0.38 \pm 0.05 \, \frac{USD}{W}$.

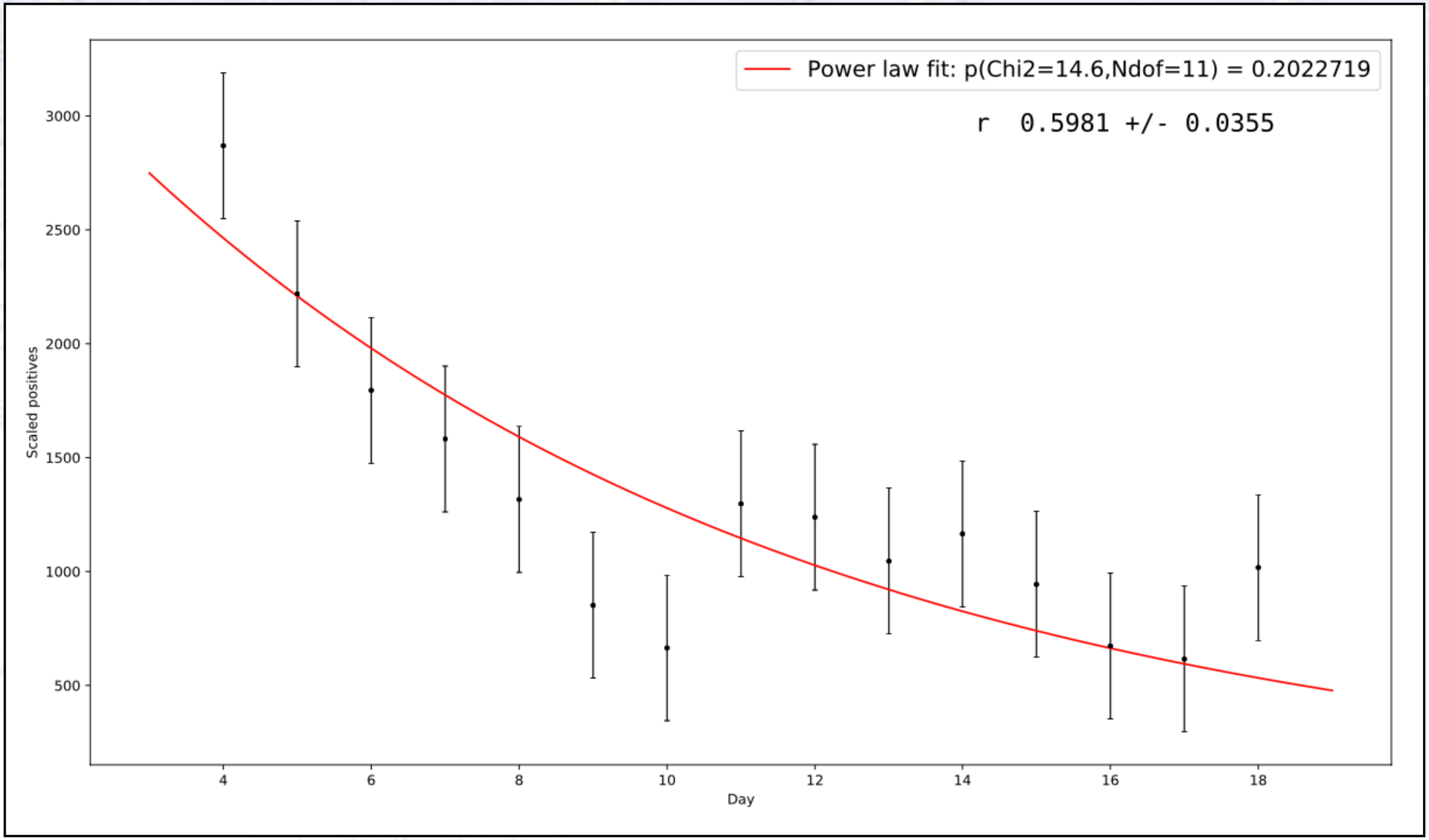
Problem 5.2

Alas - I put a wrong sign in the scaling of positives given tests. It doesn't change the problem, but it would have been nice to be closer to reality!



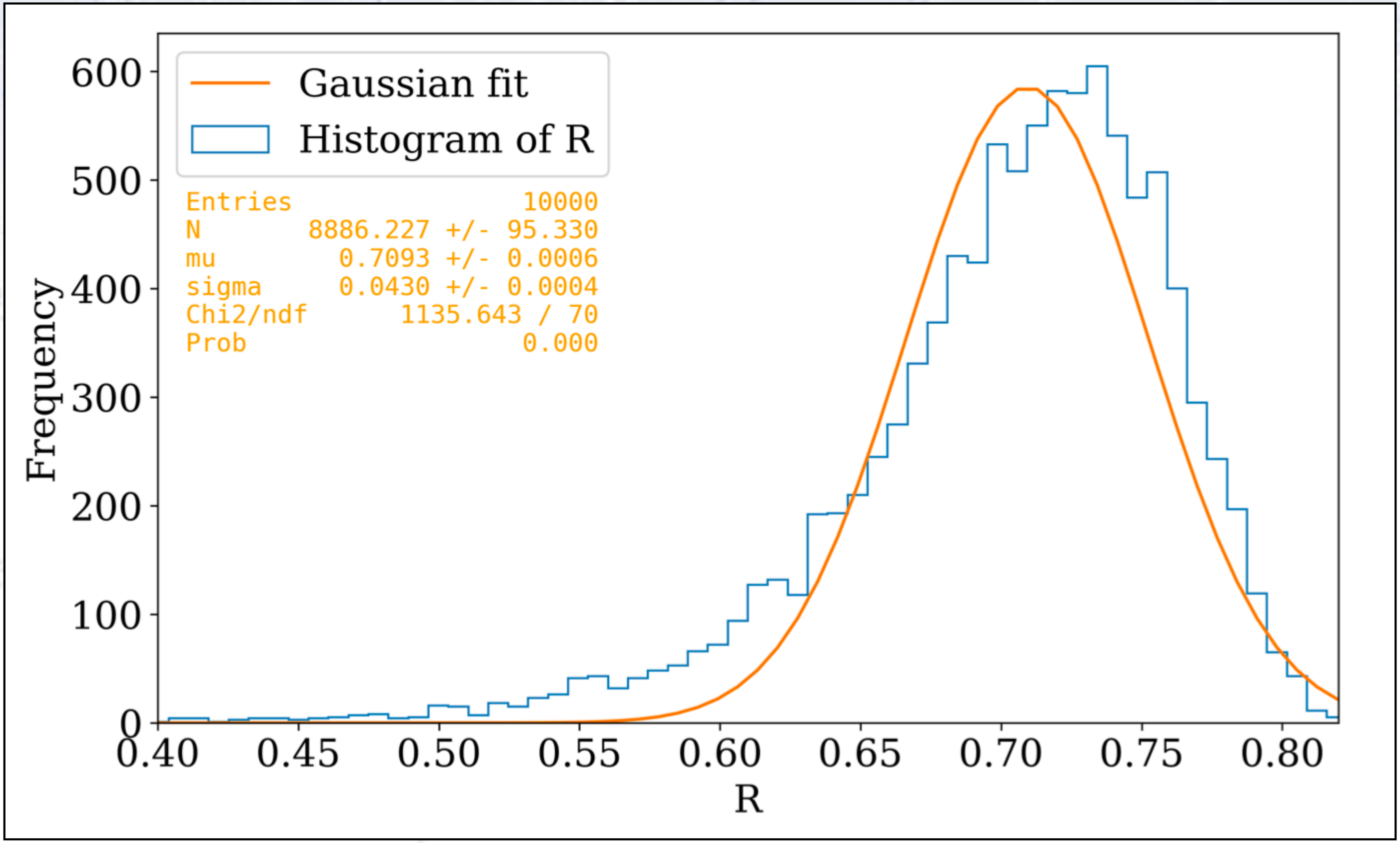
Problem 5.2

Adding a (large) systematic uncertainty makes the fit good:



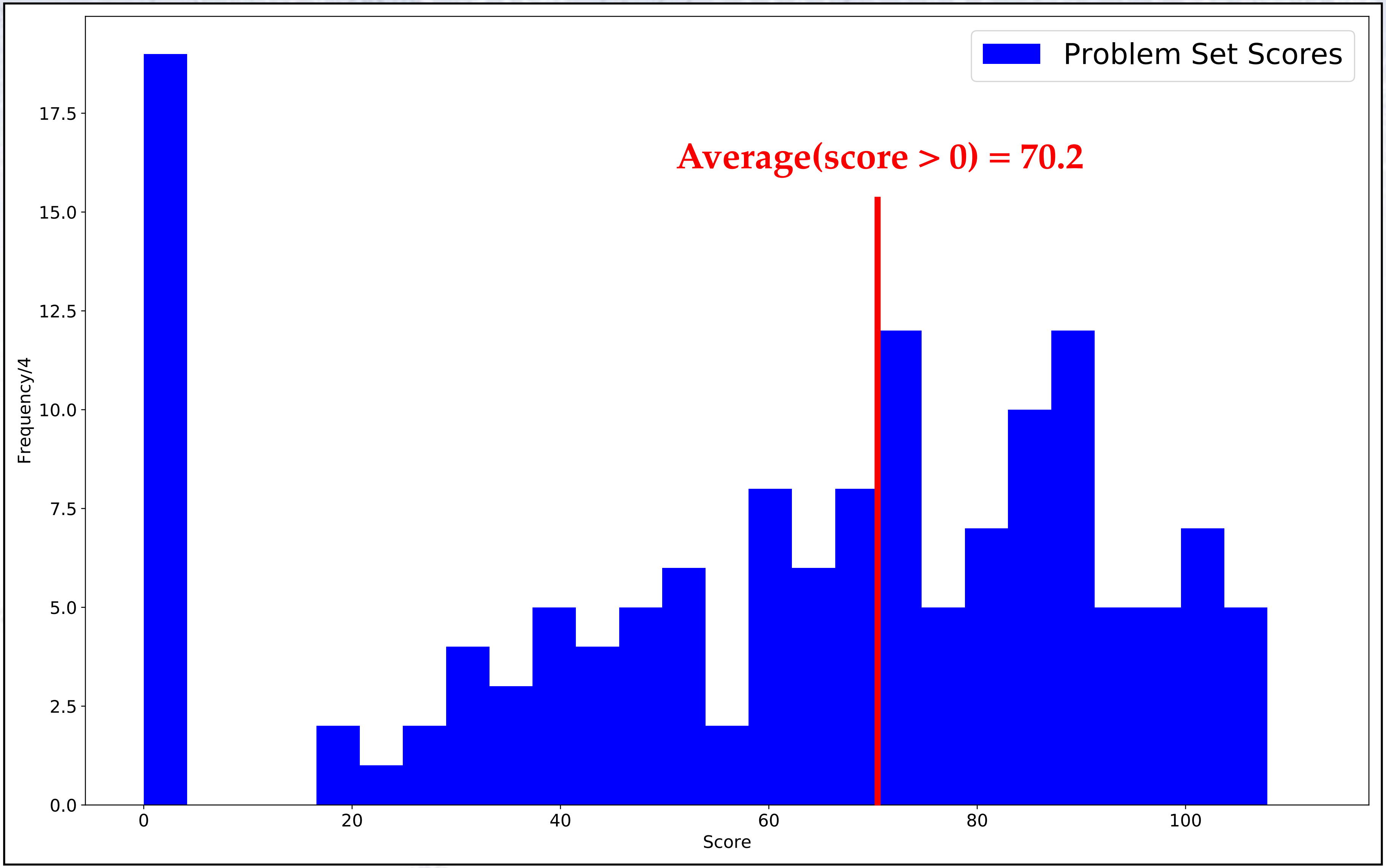
Problem 5.2

The impact of not knowing the generation time gives an asymmetric error on R.



Your scores

General distribution



Individual scores

84	nqg109@
29.925	qlc889@
51.975	skv830@
79.275	fbm531@
58.8	rbg812@
0	gpl151@
61.425	srl902@
68.775	kcn791@
75.6	brd230@
0	hgv994@
43.05	vjb896@
0	mcj576@
0	dvj716@
73.5	wph463@
0	bct232@
48.825	lmz220@
76.65	lgr243@
53.025	xtg390@
37.8	vrs764@
100.8	bqc703@
51.45	wxm206
27.3	sfz419@
87.675	prk312@
99.225	glf136@
48.3	hnr909@
69.825	fkr155@
102.9	vgr442@

79.8	rdl821@
74.025	zft162@
79.275	dkz419@
107.1	whj419@
65.1	lbq747@
50.925	wgm492
0	btq171@
55.65	tsk240@
64.05	jpg878@
94.5	cjb924@
71.925	nld314@
72.45	wmc573
50.4	qrg977@
71.4	wfg813@
70.35	bnv384@
89.25	pln924@
72.975	zts164@
87.675	kxm508
52.5	qgf305@
100.275	dhm160
88.2	hjr420@
89.25	ktj250@
67.2	zls129@
0	mbn681
85.05	fzv545@
90.3	cpr181@
87.675	wsv419

34.65	ckh739@
32.55	vrw703@
96.6	zbp392@
75.6	hjm764@
41.475	wbk841
18.9	ckb742@
73.5	bnv186
58.8	wtj465@
49.35	mzx643
63.525	lkt259@
74.55	fwb590@
95.55	qlc506@
38.85	lxp458@
80.85	mjx381
61.425	ncg400
103.95	kdj269@
69.3	tmn232
0	mgz624
86.625	xmg125
83.475	jvr822@
76.125	lmd295
84	csr396@
42	kvd970
85.05	sbj276@
63	zpj359@
0	zwp315
82.95	hcj888@

0	bkg618@
102.9	xfk351@
99.75	vgj755@
87.675	hcp742@
91.875	kvh318@
84.525	vld975@
40.425	jhm321@
71.925	fqt156@
82.425	vpq602@
97.65	bzr977@
0	xvg720@
104.475	bdp274@
60.375	rnh697@
93.975	gbc493@
90.825	wzg980@
86.1	wlq622@
46.725	zfb849@
31.5	qkf986@
94.5	tlh938@
87.15	gwn174
79.8	qhd476@
103.95	bvr391@
78.75	tkz648@
71.925	rwx433@
41.475	cxk387@
64.05	frm511@
66.675	prk161@

94.5	bmx788
88.725	hkl224@
70.875	rcg963@
61.425	fbc382@
61.425	cbk696@
100.8	qmd636
97.125	qjr103@
83.475	hwg245
87.675	gzl687@
0	vwk284
30.975	phq140@
0	xtw854@
64.575	wkh276
26.775	dmh708
46.2	mbn723
0	lmr494@
69.825	dfv249@
23.1	fxw690@
35.7	mks336
57.225	pwv995@
43.05	xgh688@
58.275	nqs117@
103.95	ldr934@
43.575	gpd492
86.1	vbd402@
0	bxj754@

16.8	dzx335@