University of Copenhagen — Niels Bohr Institute, December 28, 2021

Applied Statistics

Exam in applied statistics 2020/21

Science is not truth. It is the current summary of our experiences. [Jens Martin Knudsen, 1930-2005]

I – Distributions and probabilities:

1.1a

We can reword this to be a situation of two possible outcomes: a die has exactly three eyes facing up (success) or it doesn't (failure). Given that every trial is independent and non-correlated, this is just a binomial, where the probability of success is $p = 1/6$.
N_3 **will follow a binomial distribution.**

1.1b

The probability of getting 7 or more 3s in a roll with 20 normal dice, is the sum of the probability of getting i 3s from $i = 7$ to $i = 20$.

$$\sum_{i=7}^{20} p(r = i; n = 20, p = 1/6) = \sum_{i=7}^{20} p^i (1-p)^{n-i} \frac{n!}{i!(n-i)!} = 0.0371$$

with uncertainty

$$\sigma(f) = \sqrt{\frac{f(1-f)}{N}} = 0.05.$$

And the result is

$$\mathbf{p(7 \text{ or more threes}) = 0.04 \pm 0.05}$$

1.2a

	PCR	Antigen
Tests	103261	26162
Positives	2464	491
Fraction of positives	0.0239	0.0188

Table 1: 1.2a results.

The probability that these two fractions are the same is the probability that come from the same distribution. We'll do a z test.

$z = 4.927$ This is pretty high, and to reject this hypothesis on a 95% confidence level, you would just need z to be larger than ≈ 1.5. This is above some 99% confidence level.

1.2b

If we assume that the two tests are sampling the same population, we would expect the same positive rate. For the AG test to have the same positive rate, it would need 625 positives, of which, still only the original 491 would actually be true positives, and the false positive rate would be

$$1 - \frac{491}{625} = 0.2144 = 21.44\%$$

1.2c

If we start by assuming that ≈ 50 people are sick (the exact value doesn't matter), that leaves 49,950 healthy people. If they all get tested, we would expect $49,950 \cdot 0.0002 \approx 10$ people to be false positives. Now we know that of the 47 positive people, 10 are actually negative, so there are actually only 37 positive people. A 20% false negative rate would mean that the 37 positive people only make up 80% of the positives, meaning the true number is 46.25 rounded down to 46. Through the power of statistics, we have managed to cure 1 person.

We expect the ratio of infected to non-infecteds to be the same in Denmark as in our sample.

$$ratio = \frac{46}{50,000} = 0.00092$$

$$\sigma_f = \sqrt{\frac{f(1-f)}{N}} = 0.00013$$

Meaning we estimate that the ratio of infected in Denmark is $(\mathbf{0.00092 \pm 0.00013})$.

1.3

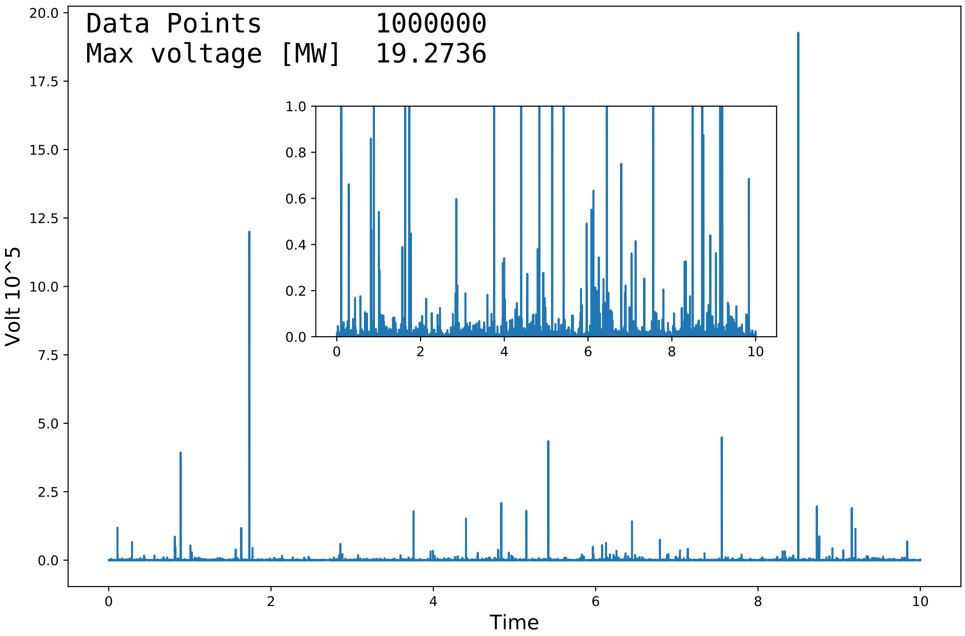


Figure 1: A very informative plot, with an added zoomed in plot, so it is easier to see what are peaks and what aren't, while still keeping the overview with the original plot.

II – Error propagation:

2.1a

I find the error on y and z from the error propagation formula

$$\sigma_y = \sqrt{\left(\frac{dy}{dx}\right)^2 \sigma_x^2} = \sqrt{\left(-\frac{2x}{(1+x^2)^2}\right)^2 \sigma_x^2}$$

Results

$$\mathbf{y = 0.2065 \pm 0.005}$$

$$\mathbf{z = 1.085 \pm 0.0678}$$

2.1b

If $x = 0.96 \pm 0.03$ the procedure does not change:

$$\mathbf{y = 0.520 \pm 0.0156}$$

$$\mathbf{z = 625 \pm 937.5}$$

2.2a

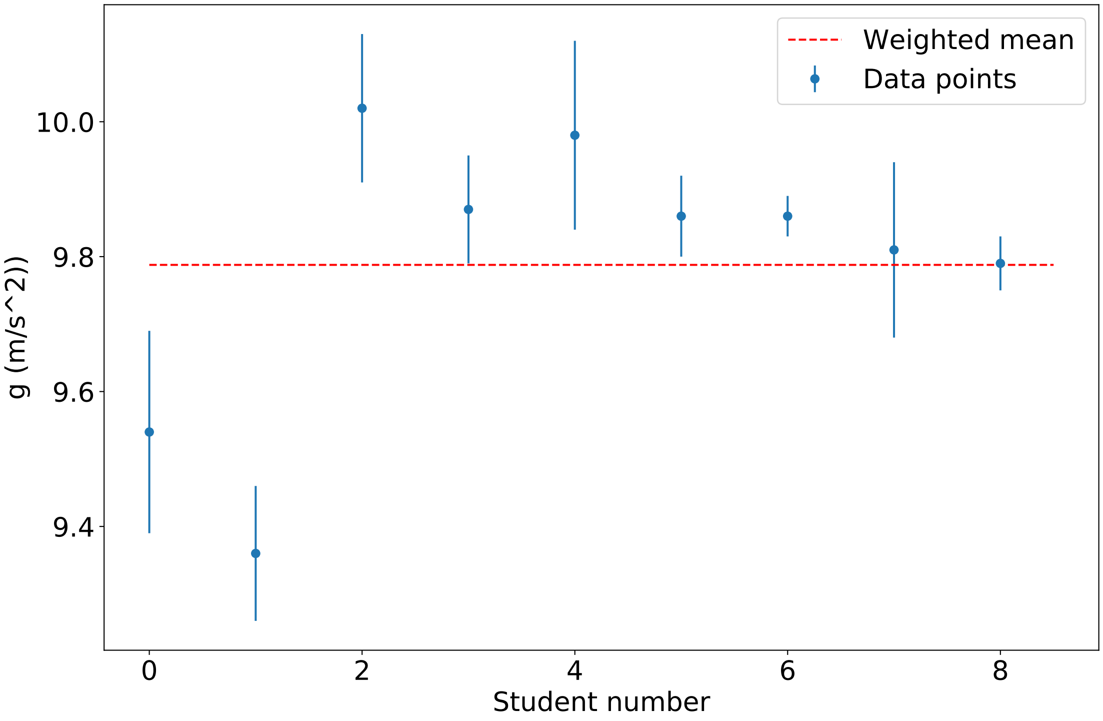


Figure 2: Error bar plot of g estimates plotted with the weighted mean.

The best estimate, given that we have the uncertainties, is a weighted mean.

$$\underline{\underline{\mathbf{\hat{g} = (9.82 \pm 0.02) m/s^2}}}$$

2.2b

I found that $\chi^{\mathbf{2}} \mathbf{= 3.336}$ and $\mathbf{p = 0.911}$.

The second data point lies 4.6σ away from the weighted mean, which is unlikely. Otherwise the rest of the points were within 2σ and were no cause of suspicion.

2.2c

My best estimate was $g = (9.82 \pm 0.02)m/s^2$. Given that $g = (9.8158 \pm 0.0001)m/s^2$ is well within one sigma, **my best estimate of g does agree with the precision measurement**.

III – Monte Carlo:

3.1a

Since the function cannot be enclosed in a box (as it is defined for $x \in [0, \infty[)$, we have to use the transformation method. First we check that it is normalized, which it is, because

$$\int_0^\infty f(t)dt = 1.$$

Then we integrate it.

$$F(t) = \int_0^t f(t)dt = 1 - e^{-1.25t}$$

Then invert it

$$F^{-1}(t) = 0.8\ln(1-t)$$

Now we can generate random numbers from f, by making $u = F^{-1}(t)$.
If t is uniformly distributed between 0 and 1, then so is $1 - t$.

$$F^{-1}(t) = -0.8\ln(t)$$

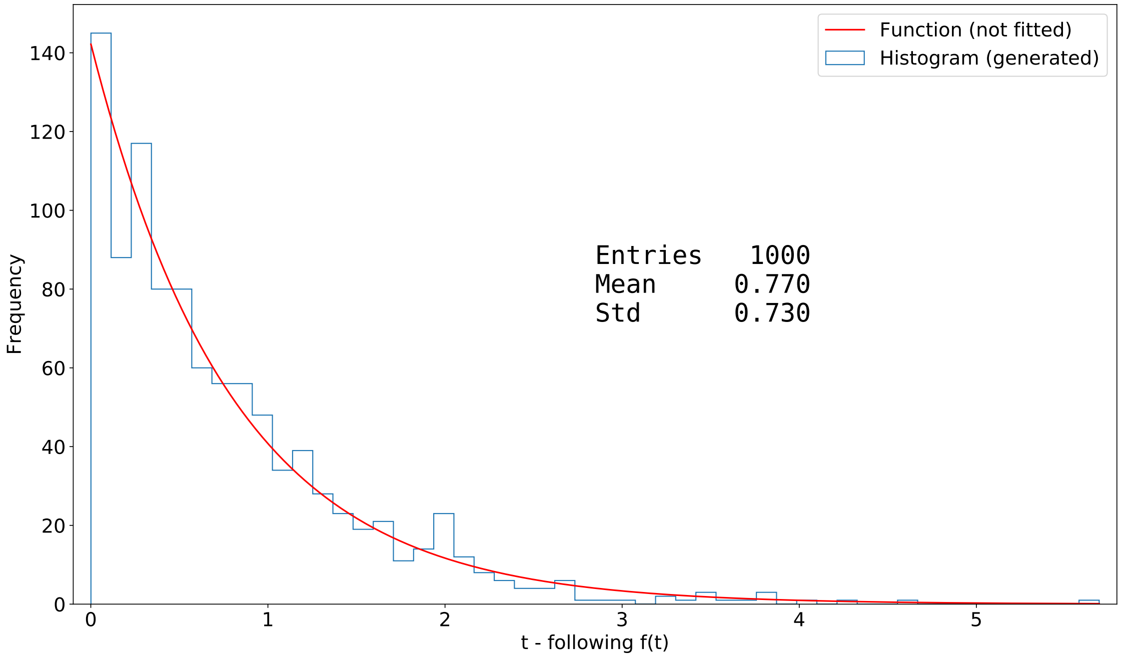


Figure 3: The generated numbers t following $f(t)$.

3.1b

Figure 5: Caption

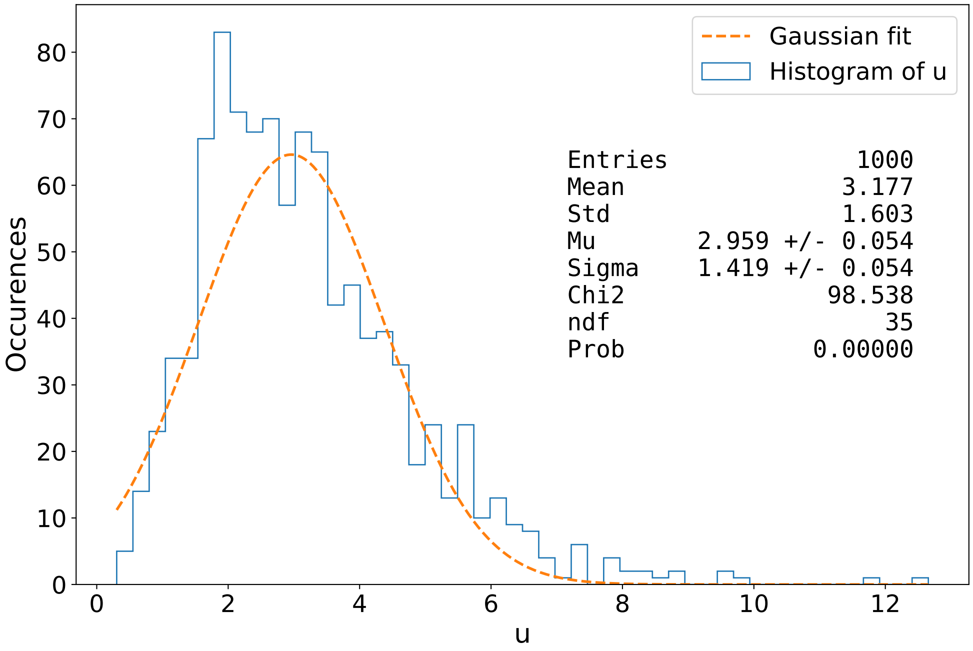


Figure 4: The distribution of u, fitted with a Gaussian.

We get a probability of about 10^{-5} (calculated using random numbers, so it will fluctuate), so no, it probably is not Gaussian. which is not a surprise. From the central limit theorem, we would expect the distribution to be Gaussian if u was a sum of more numbers. If, e.g., u was a sum of 50 numbers, the probability would be ≈ 0.26, which is consistent with it being a Gaussian distribution.

3.1c

The ideal fit would have been the Irwin-Hall distribution, as it is a distribution of sums of uniformly generated numbers. A skewed double Gauss would also fit better than a regular Gauss.

3.2

I can't invert this function without getting something strange, but otherwise the generating process is the same as in the one above. Once the numbers are generated, it is easy to arrange my values from low to high, and find the middle value, the median.

IV – Statistical tests:

4.1a

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

and $\hat{p_{1,2}}$ is our best estimate of $p_{1,2}$, which is the ratio of Covid-19 cases for the vaccinated and placebo groups respectively.

This formula yields $z = -11.83$. From looking in a Z-table, we know that we need to reject the null hypothesis on a 5% significance level, if $|z| > 1.6$, which it very much is.

The chance that BNT162b2 had no effect is $\ll 0.05\%$

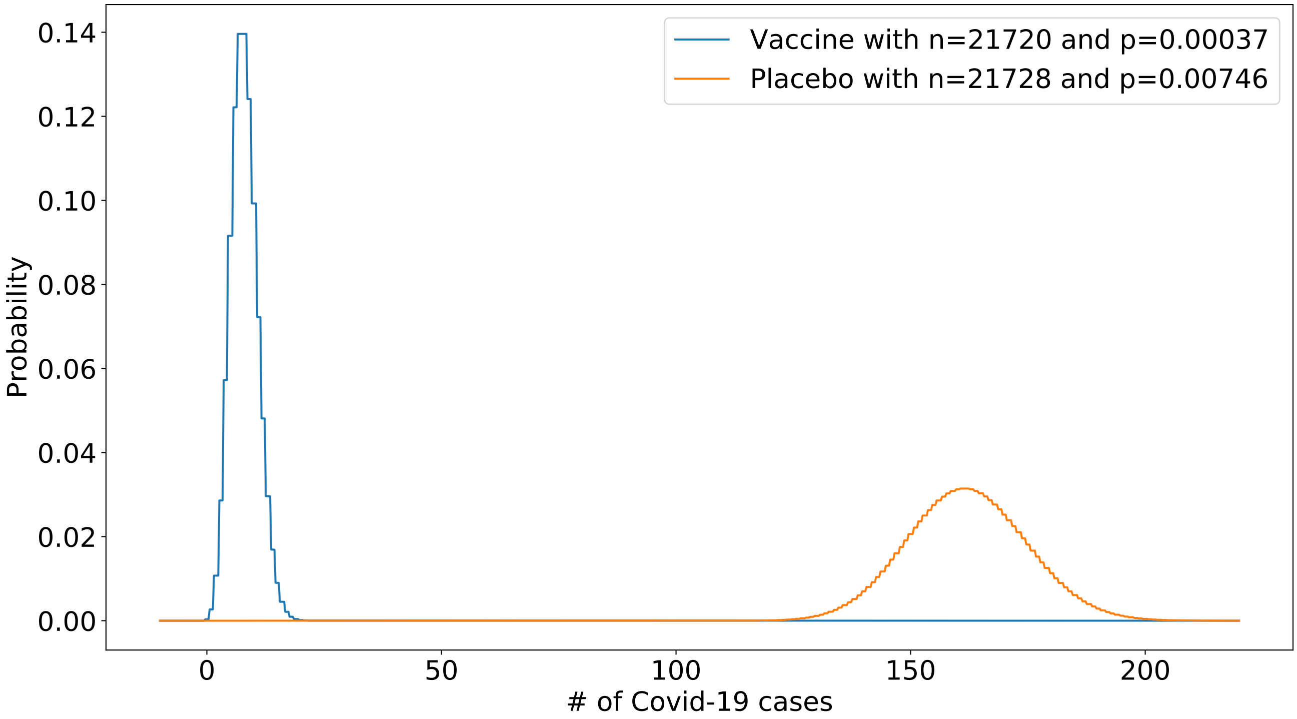


Figure 6: Distribution of Covid-19 cases among vaccinated (blue) and placebo-receivers (orange).

Just by giving this figure a quick glance, it comes as no surprise that the two distributions are not the same, and that the vaccine does have an effect.

4.1b

The efficacy is $\epsilon = 0.951$. A 68% confidence interval is the interval of $[-\sigma; \sigma]$.

$$\sigma(\epsilon) = \sqrt{\frac{\epsilon(1-\epsilon)}{N}} = \sqrt{\frac{0.951(1-0.951)}{170}} = 0.00166$$

The 68% confidence interval for the efficacy ϵ is $[-0.00166; 0.00166]$

4.1c

AKA. what is the chance that chance of getting a severe case, is the same for the vaccinated and unvaccinated group?

We still have two binomial distributions, one with $n_{vacc} = 21720$, $r_{vacc} = 8$, and one with $n_{placebo} = 21728$ and $r_{placebo} = 162$. What is the probability that the probability of success (getting a severe case of Covid-19) in both groups is the same?

Now we get $z = -2.53$ which is significantly less, and it is more likely, that the vaccine doesn't have an effect, but we can still reject that hypothesis with a 99% confidence, since the reject region has an area of $\approx 0.005\%$

4.2a

There are two possible outcomes, ace and not ace, and with replacement, the trials are independent, with the probability of success being constant for all trials.
The distribution of the number of aces is binomial.

The chance of getting 3 aces or more is the chance of getting 3 or 4 aces.

$$p(3 \text{ or } 4 \text{ aces}) = p(r = 3; n = 4, p = 4/52) + p(r = 4; n = 4, p = 4/52)$$

The chance of getting 3 aces or more is 0.001716.

4.2b

Without replacement you have (number of aces left in the deck)/(number of cards left in the deck) chance of drawing an ace.

There are four different ways of drawing 3 aces, and they all have the same probability, so we'll consider the scenario where the first card is not an ace, and the remaining three are, and multiply that with 4.

$$p(3 \text{ aces}) = 4 \cdot \frac{48}{52} \cdot \frac{4}{51} \cdot \frac{3}{50} \cdot \frac{2}{49}$$

$$p(4 \text{ aces}) = \frac{4!(52-4)!}{52!}$$

The total probability of drawing 3 or 4 aces is then the sum of these two probabilities.

$$p(3 \text{ or } 4 \text{ aces}) = 0.00072.$$

4.2c

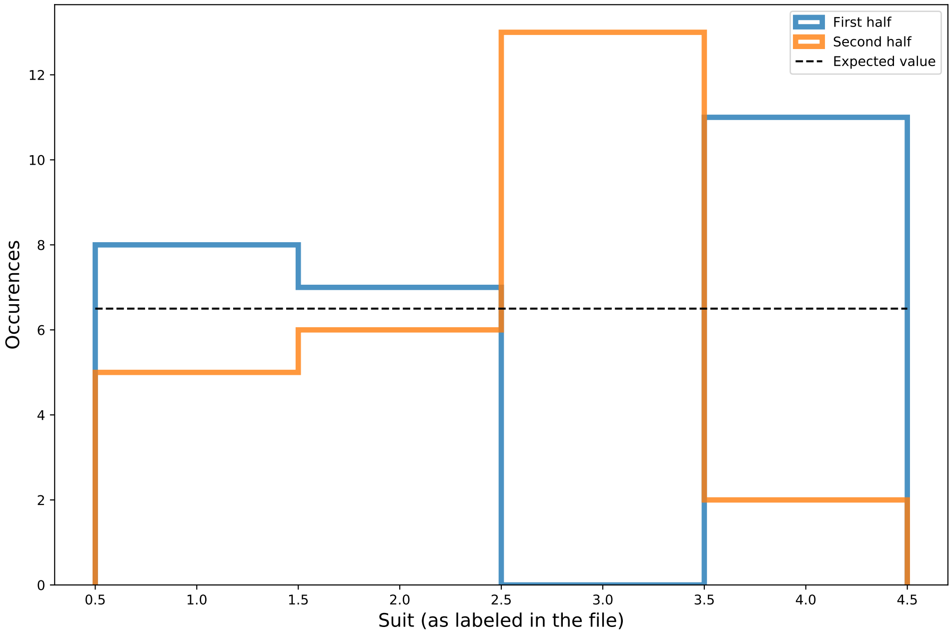


Figure 7: Histogram of the distribution of the suits within the first 26 cards (blue) and last 26 cards (orange).

If the cards were well shuffled, the suits would be uniformly distributed in the first half and second half. The histogram of the distribution of the suits in the second half had a $\chi^2 = 14.9$ and a probability of 0.005% of being uniform, so I conclude that on a 99% confidence level, that the cards are not well shuffled.

I also noted that the suits in the first half follow a specific pattern (4124) repeating, and that the value of the cards, seemed to follow a piecewise downwards linear trend.

V – Fitting data:

5.1a & b

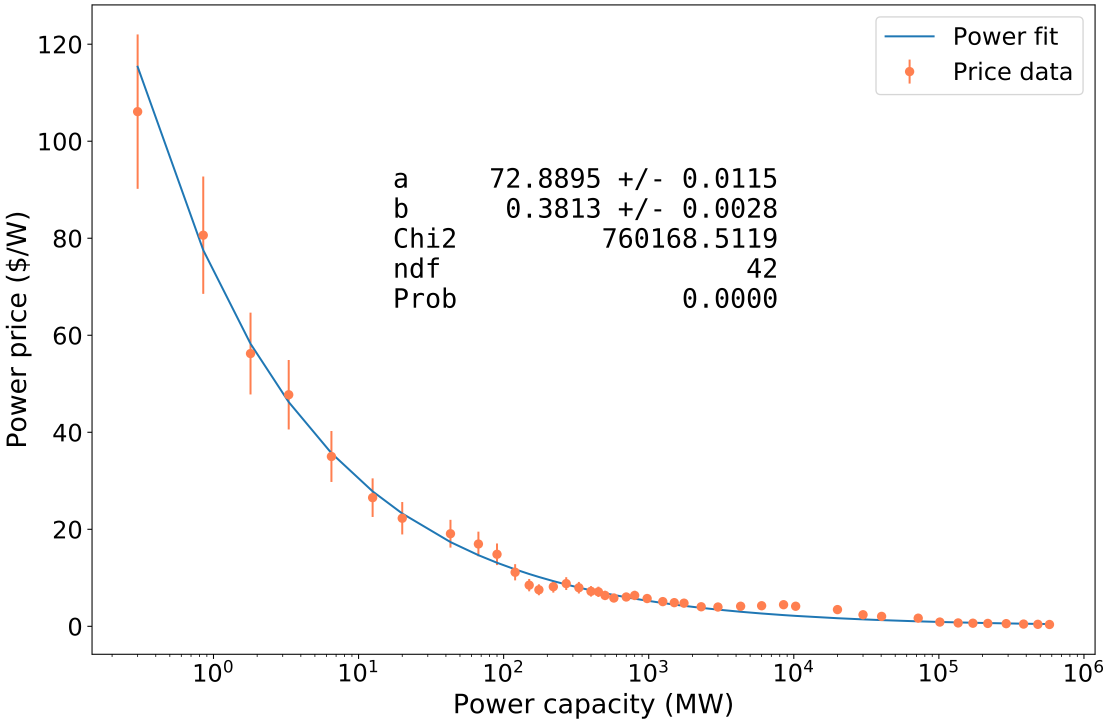


Figure 8: The price of power ($/W), plotted as a function of power capacity (MW) on a logarithmic x-scale, and fitted with a power law.

5.1c

I fitted the relations with the function

$$f(x) = a \cdot \exp\{x/b\}$$

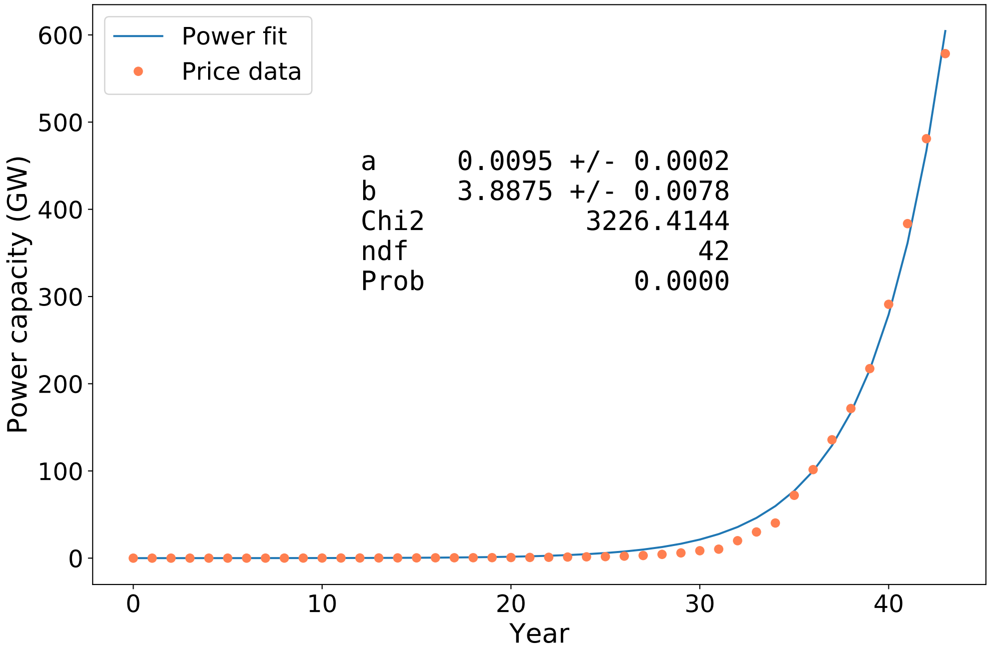


Figure 9: Power capacity plotted as a function of years after 1980, with fit.

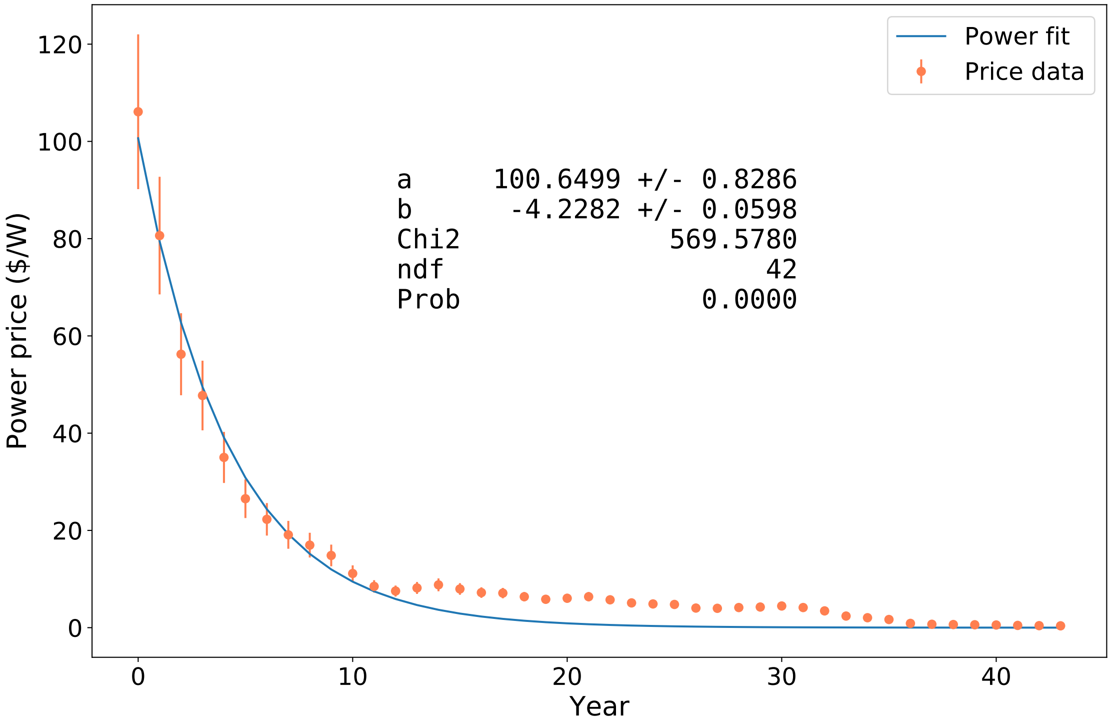


Figure 10: The power price plotted as a function of years after 1980 with fit.

I scaled the data by dividing the capacity with 1000, so the unit became GW, and subtracted 1980 from the years, so the x axis became "years after 1980". Using my fit parameters, i can isolate x in

$$a \cdot \exp\{x/b\} = 1\,TW \implies x = 44.96$$

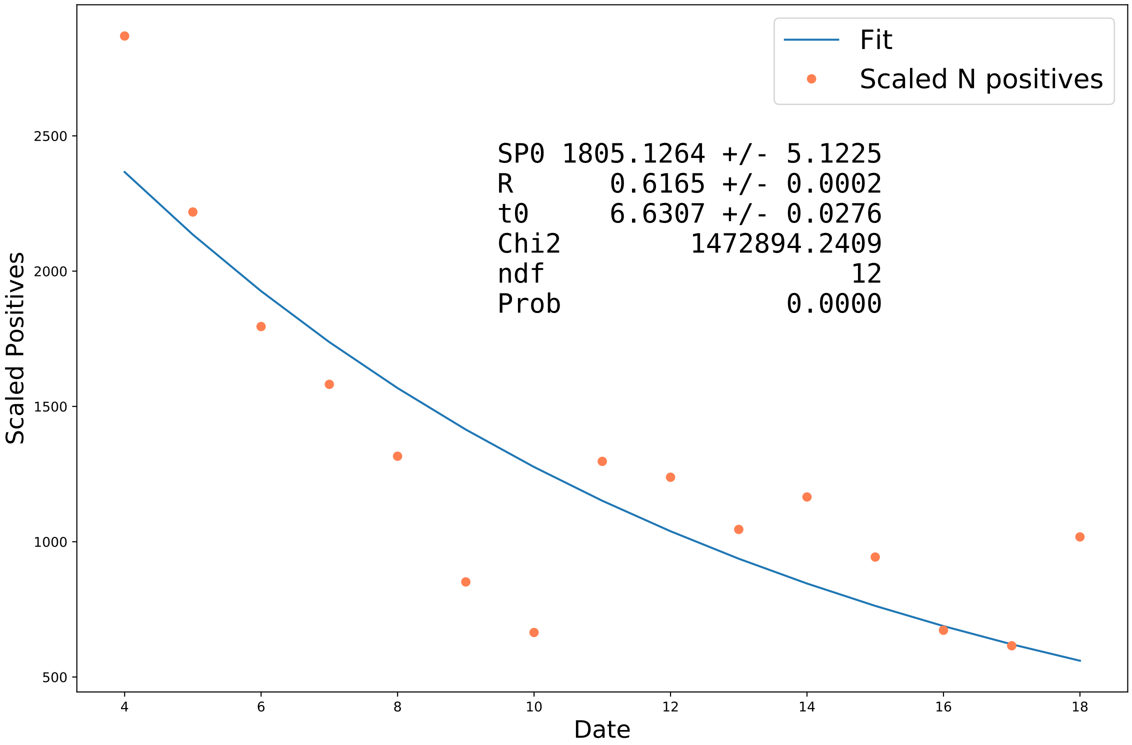
Meaning that after 45 years, in 2025, this model predicts that the capacity will become 1 TW.

Then I made another exponential fit of the power price as a function of the year (after 1980), and calculated with my new parameters and $x = 45$. Via the error propagation formula, i got an uncertainty from the parameters of 0.8, which I then added in quadrature with the (in relation insignificant) 15% uncertainty. The result is that that power will cost \$$(0.0024 \pm 0.8)$ in 2025. The uncertainty is much greater than the value, due to the poor fits the value was generated from.

5.2a

The average number of tests in the period is

5.2b



```
SP0 1805.1264 +/- 5.1225
R      0.6165 +/- 0.0002
t0     6.6307 +/- 0.0276
Chi2        1472894.2409
ndf                   12
Prob              0.0000
```

Figure 11: Scaled positives as a function of the date, and fit.

5.2c

I added the lowest integer value for systematic uncertainty, such that the p value was above 0.05, and that was $\sigma_{sys} = 265$.

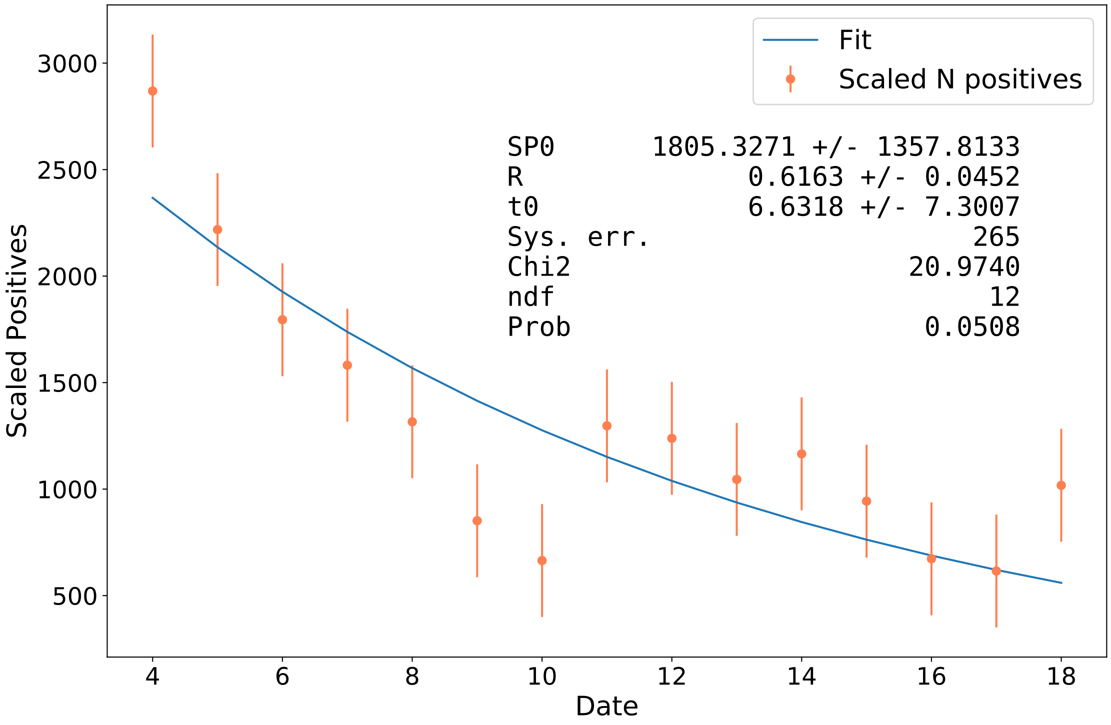


Figure 12: Same plot, but with added systematic erros of 265.

5.2d