Kian Gao

shp593

Final exam on AppStat2020

I – Distributions and probabilities:

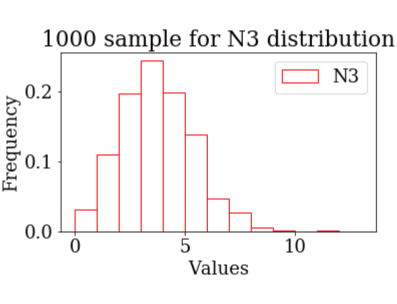
1.1.1

The probability of N_3 follows binomial distribution. First, for each roll, the N3 distribution is binomial distribution N(20, 1/6).

Another property is that according to central limit theorem, if number of dice is large enough, the final distribution of N3 will tend to be gaussian.

1.1.2

The probability of getting 7 or more 3s in one roll with 20 dice is calculated by $1-\sum_{n=0}^{6} C_{20}^{n}(\frac{1}{6})^n(\frac{5}{6})^{20-n} \approx 0.0371$




```
With random sampling:
Getting 7 or more 3s in a roll is 0.03714
```

1.2.1

The positive fraction of PCR is $\frac{2464}{103261} \approx 0.0239$, and the positive rate of AntiGen is $\frac{491}{26162} \approx 0.0188$

Using confidence interval between portion. Variance of p1-p2 is calculated by $\frac{p1(1-p1)}{n1}+\frac{p2(1-p2)}{n2}$

```
p1-p2 is 0.0051156826
The variance of p1-p2 is 0.0000009279
Give 75% confidence interval, p1-p2 +/- 1.15*SE(p1-p2) where SE = sqrt(variance)

Therefore, 75% confidence interval of p1-p2 should be
0.0047753 < 0.0051157 < 0.0069908
So we have at least 75% confidence that they are statistically the same
```

1.2.2

Assuming that PCR test is totally accurate. Then the sample AntiGen use supposes to have $2464 \times 26192 \div 103261 \approx 625$ people infected. Therefore, the false negative rate is $(625-491) \div 625 = 21.44\%$

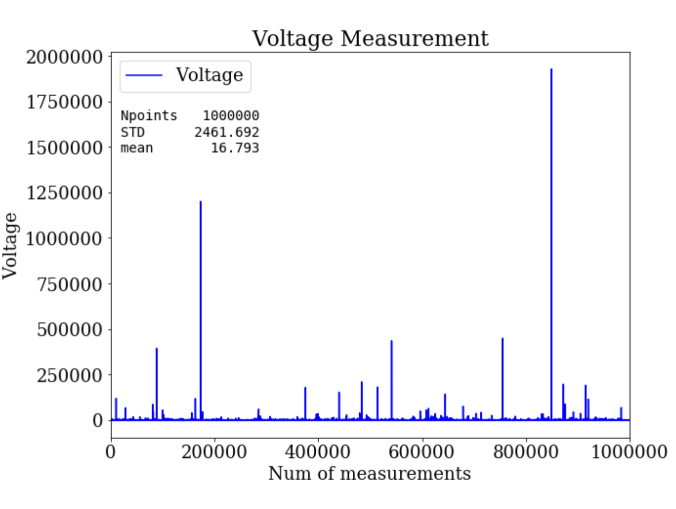
1.2.3

First we have 0.02% false positive and 20% false negative, given:

$$47 = \mathrm{P}(1-20\%) + (50000-\mathrm{P}) \times (0.02\%)$$

where P is the positive in reality. Result in $\mathrm{P} = 37 \div 0.7998 \approx 46.2615654$. Therefore, the positive rate is $\mathrm{P}/50000$, approximately 0.0925%

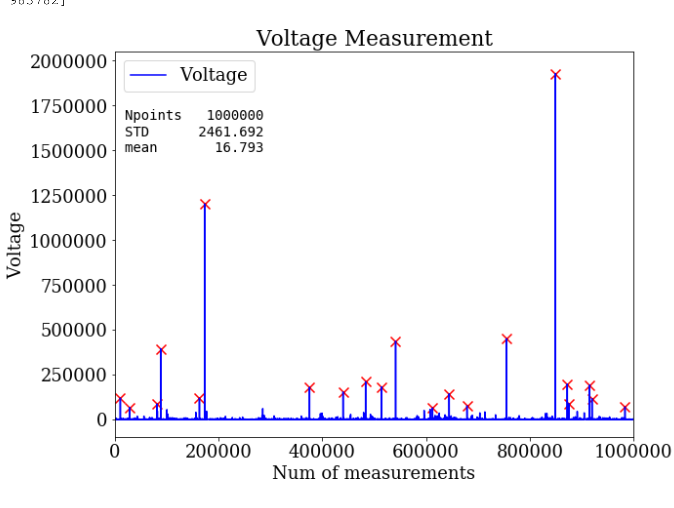
1.3.1



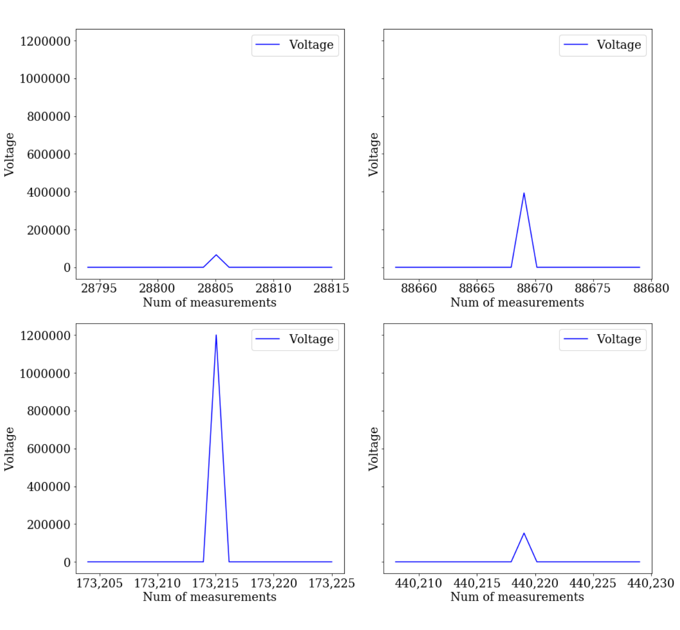
1.3.2

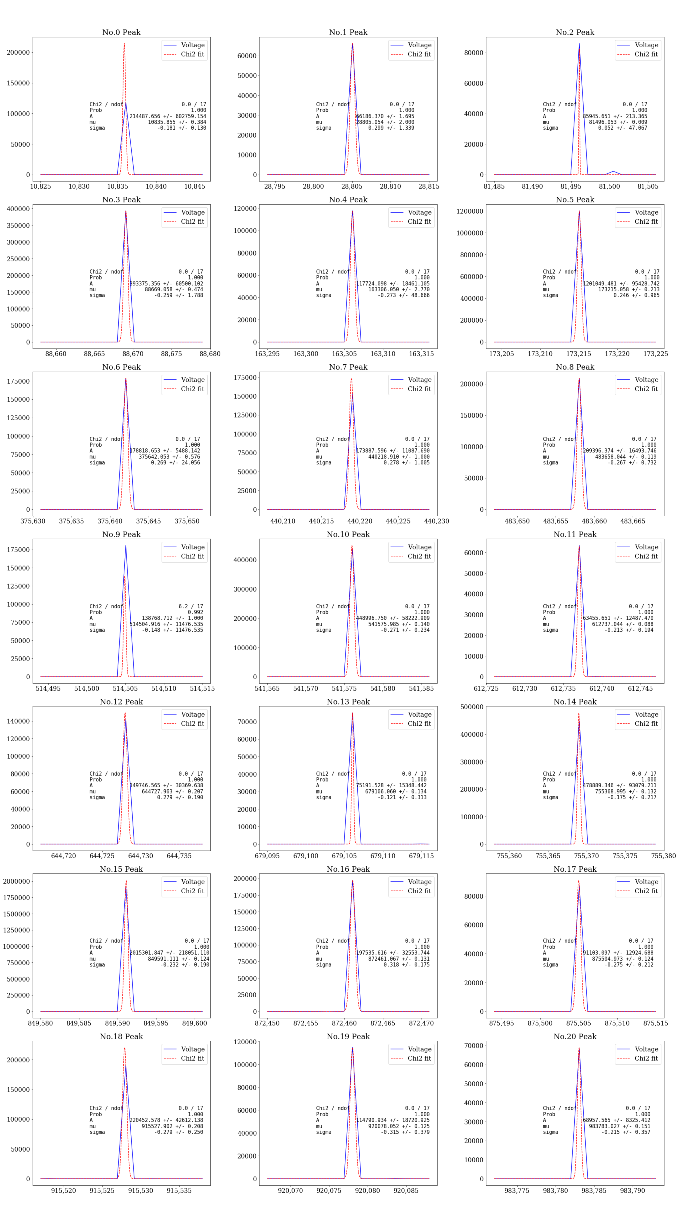
In order to fit the peak in the diagram, first we need to seach all the peaks out, and then zoom in to see what the peaks looks like.

```
The peaks is
 [ 10835  28804  81495  88668 163305 173214 375641 440218 483657 514504
 541575 612736 644727 679105 755368 849590 872460 875504 915527 920077
 983782]
```



Zoom in one peak to seek for the distribution. Notice that it follows a gaussian distribution





The fit is given relative uncertainty 10% to obtain a better fitting. Nevertheless, some of the fitting is still problematic because the peak is so sharp that gaussian distribution is hard to catch up. Almost every peak lasts only for one frame.

II – Error propagation:

2.1.1

$x = 1.96 \pm 0.03$, therefore $\sigma_x = 0.03$, $\sigma_y^2 = (\frac{\partial y}{\partial x})^2\sigma_x^2$, $\sigma_z^2 = (\frac{\partial z}{\partial x})^2\sigma_x^2$.

$$\frac{\partial y}{\partial x} = \frac{-2x}{(1+x^2)^2}$$
$$\frac{\partial z}{\partial x} = \frac{2}{(1-x)^3}$$

```
The value of y is 0.207, and the uncertainty(sigma y) is 0.0050
The value of z is 1.085, and the uncertainty(sigma z) is 0.0678
```

2.1.2

```
when x = 0.96 +/- 0.03
The value of y is 0.520, and the uncertainty(sigma y) is 0.0156
The value of z is 625.000, and the uncertainty(sigma z) is 937.5000
```

2.2.1

Weighted mean is $\frac{\sum x_i/\sigma^2}{\sum 1/\sigma^2}$, and the related SD is $\sqrt{\frac{1}{\sum 1/\sigma^2}}$

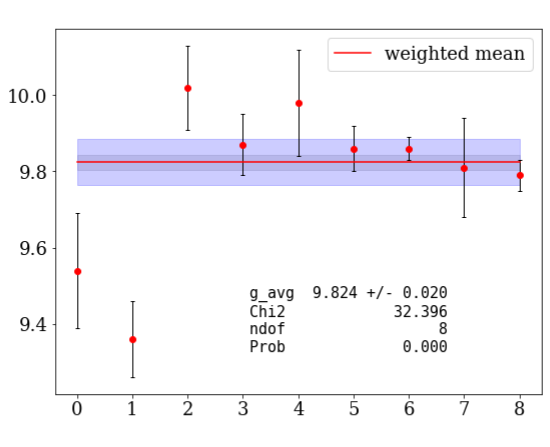
```
Directly look at the percentage uncertainty
Group 1: 1.572%
Group 2: 1.068%
Group 3: 1.098%
Group 4: 0.811%
Group 5: 1.403%
Group 6: 0.609%
Group 7: 0.304%
Group 8: 1.325%
Group 9: 0.409%

While looking at the percentage uncertainty,
the best group is group7 with g = 9.86 m/s^2, and sigma = 0.03 m/s^2
Meanwhile, they also own the best absolute uncertainty

The best estimation from me is weighted average
The calculation result of weighted mean of g is 9.824 m/s^2
And the mean of uncertainty 0.020 m/s^2
```

2.2.2

$\chi^2 = 32.396$ and p-value = 0. From the graph, we can say that the measurement group 1 and 2, 3, 5 are outside of $3\sigma_g$. Besides, although group 5 is also out of range, its uncertainty is large enough to touch the range of our estimation. Hence, group 1, 2, 3s' measurements are unlikely.



2.2.3

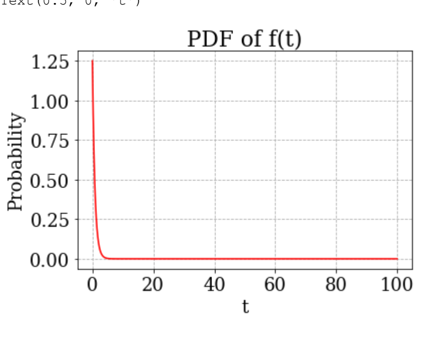
Yes! The value $9.8158 \pm 0.001 m/s^2$ is less than one sigma away from my best estimate $9.824 \pm 0.020 m/s^2$! So they agree with each other.

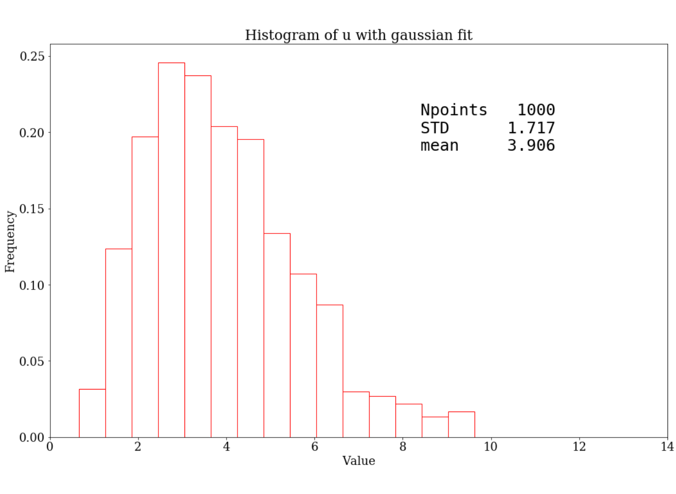
III – Monte Carlo

3.1.1

First, inspect the PDF. We found that the highest probability is 1.25. Then, we sample with accept and reject, with t = [0, 10] (Since ft(10) = 4.66 * 1e-6, small enough!), and a random number = [0, 1.25].

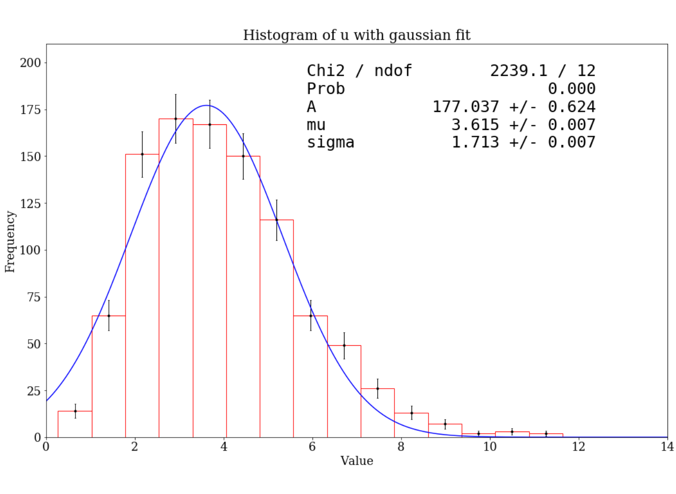
```
Text(0.5, 0, 't')
```





3.1.2

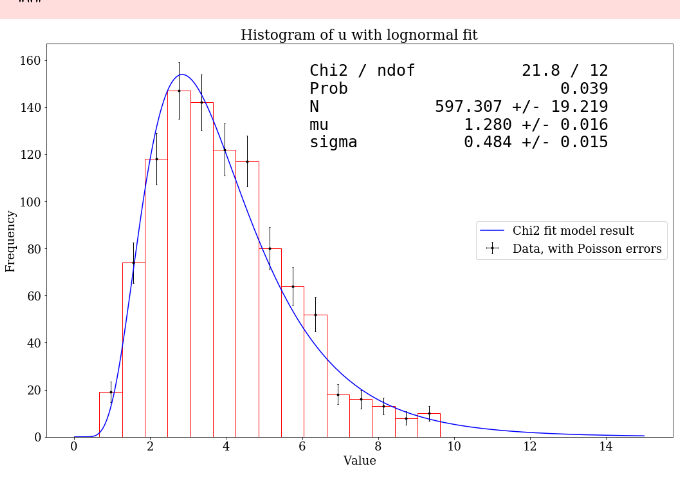
Assume that the bin count is Poisson distributed. The p-value is zero, which makes sense because the plot has a long tail.Therefore, the distribution is not normal distribution



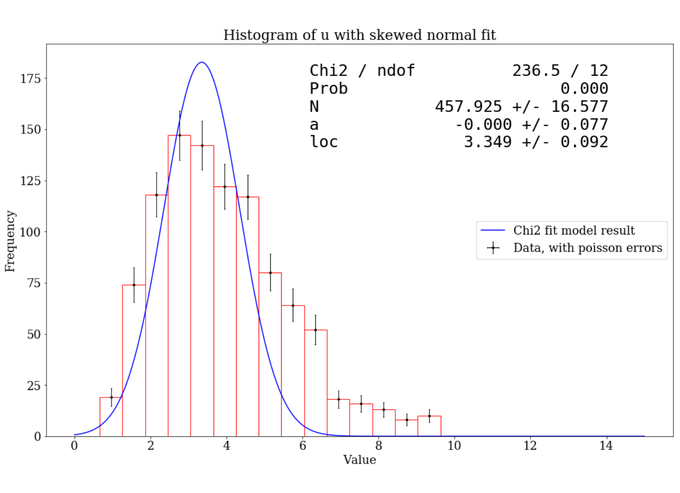
3.1.3

I will try to fit with some distributions here.

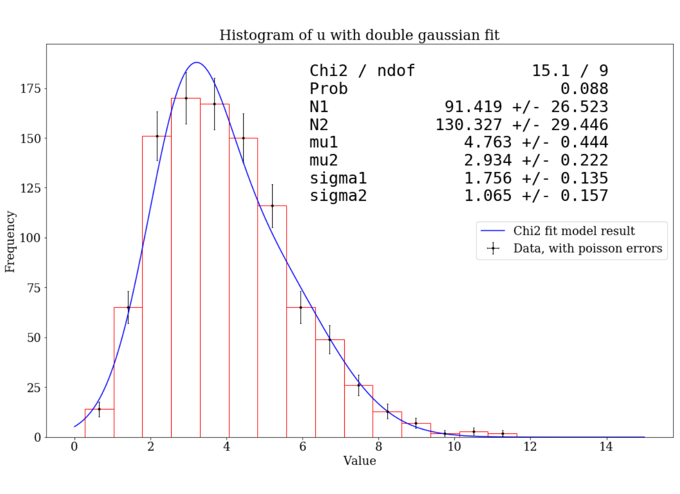
```
c:\users\gaozheming\appdata\local\programs\python\python37\lib\site-packages\ipy
kernel_launcher.py:5: RuntimeWarning: divide by zero encountered in log
  """
c:\users\gaozheming\appdata\local\programs\python\python37\lib\site-packages\ipy
kernel_launcher.py:5: RuntimeWarning: invalid value encountered in true_divide
  """
```



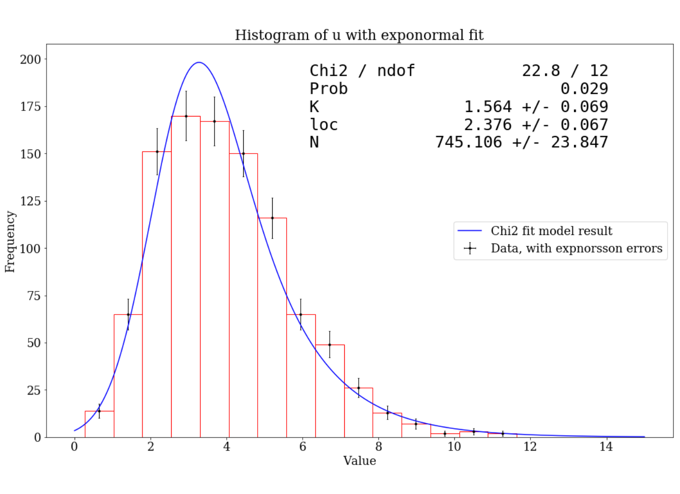
Seems it works well, try the skewed normal distribution



Not good, try double normal distribution.



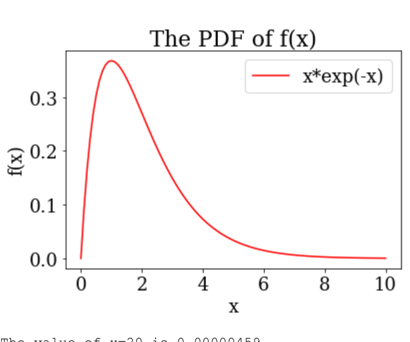
The double gaussian fit gives a p-value of 0.09, which is rather good!



After χ^2 fit of the exponential normal distribution, the p-value is 0.03, which also make sense. Therefore, double gaussian fit and exponormal fit might be options.

3.2.1

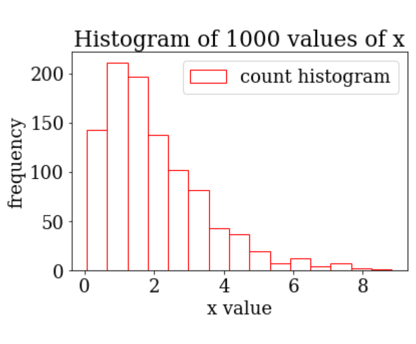
The F(x) of PDF is $-Ce^{-x}(x+1)$. Since $-e^{-x}(x+1)|_0^{\infty}$ is already nomalized, C can be 1.Take a look at the PDF distribution, the largest value of f(x) is no higher than 0.4.




```
The value of x=20 is 0.00000459
```

Therefore, the upper boundary can be set to 20. Next, generate random numbers with accept and reject method. Plot one of the experiment with histogram.

```
No 0 's experiment x median value is 1.66781
No 1 's experiment x median value is 1.62487
No 2 's experiment x median value is 1.72595
No 3 's experiment x median value is 1.66640
No 4 's experiment x median value is 1.57138
No 5 's experiment x median value is 1.65675
No 6 's experiment x median value is 1.63958
No 7 's experiment x median value is 1.71817
No 8 's experiment x median value is 1.60462
No 9 's experiment x median value is 1.66333
```



From sampling, the median of x is around 1.6. Anallytically, when $\mathrm{CDF} = 0.5, x \approx 1.67835$.

IV – Statistical tests

4.1.1

Use fisher exact test to seek the significant level of the vaccine experiment.

```
The p-value of Fisher exact test is 1.5664e-38
Which is much lower than 5%

Since the number is large, chi2 test in contingency is also a choice
The p-value of Chi2 test is 6.6419e-32
Which is also much lower than 5%
Therefore, probability that the vaccine has worse effect is less than 1e-30

On the other hand, a binomial test can also offer a result
Given the affect rate 0.0075, assume the distribution is binomial, then do the
     binomial distribution test to see if we need to accept that BNT162b2 has no
effect
P-value is 3.3355309441616695e-58
Therefore, the probability of vaccine has no effect is low

At last, we do a null hypothesis test

68% confidence interval of p1-p2 should be
0.0064895 < 0.0070875 < 0.0076854
So we have at least 68% confidence that the vaccine is useful
But less than 75% confidence (75% overflows)
```

4.1.2

Use risk ratio $\mathrm{RR} = p_{vaccine}/p_{placebo}$, then the 68% confidence interval of the efficacy is:

$$(1-\mathrm{RR}\exp(1\cdot\sqrt{1/N_v+1/N_p}), 1-\mathrm{RR}\exp(-1\cdot\sqrt{1/N_v+1/N_p}))$$

```
Therefore, the interval is (0.92904, 0.96561)
We see the BNY162b2 efficacy is in 68 confident interval:
 0.92904 < 0.95062 < 0.96561
```

4.1.3

Assume that vaccine is useless, do a binomial test for the severe cases. Assuming the p=0.5, and do two sides test.

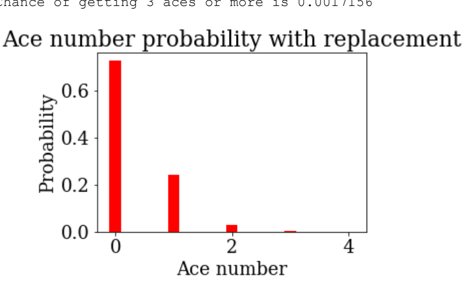
```
That the probability that BNT162b2 had no effect is 0.02148
```

4.2.1

The distribution of the number of aces follows binomial distribution B(4, 1/13). Number of 0 has probability $C_4^0(\frac{12}{13})^4$, 1 has $C_4^1(\frac{12}{13})^3(\frac{1}{13})^1$ and so on.

```
Check if the card is intact
4 4 4 4 4 4 4 4 4 4 4 4 4

Chance of getting 3 aces or more is 0.0017156
```



4.2.2

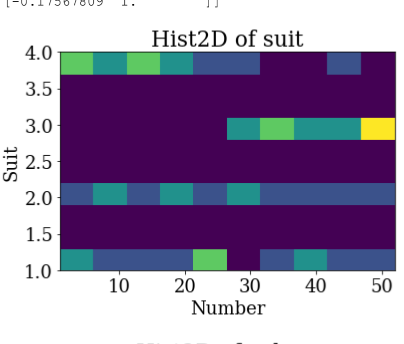
Without replacement, probability of drawing more than 3 aces is $C_4^1(\frac{4}{52})(\frac{3}{51})(\frac{2}{50})(\frac{48}{49})+(\frac{4}{52})(\frac{3}{51})(\frac{2}{50})(\frac{1}{49}) = 0.000713$

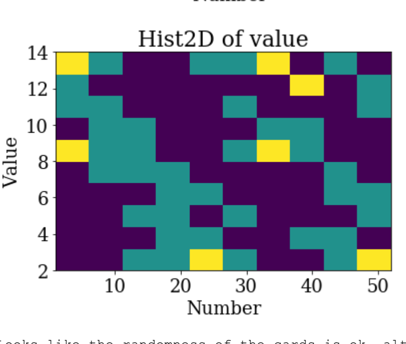
4.2.3

```
Test the number of odd after even and so on
The number of odd_even, odd_odd, even_even, even_odd is
15 9 13 14

Correlation coefficient with suit is
 [[ 1.         -0.09053914]
 [-0.09053914  1.        ]]

Correlation coefficient with value is
 [[ 1.         -0.17567809]
 [-0.17567809  1.        ]]
```

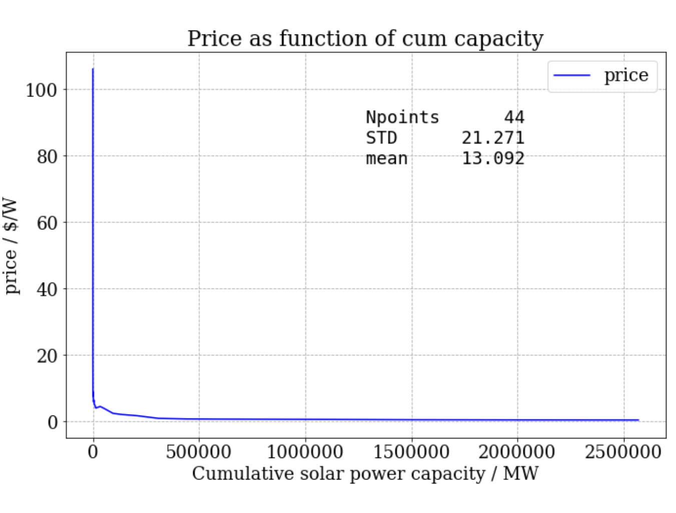





```
Looks like the randomness of the cards is ok, although there might be weak relat
ion between value and the sequence of cards. So I can say it is not really well
shuffled, but ok shuffled.
```

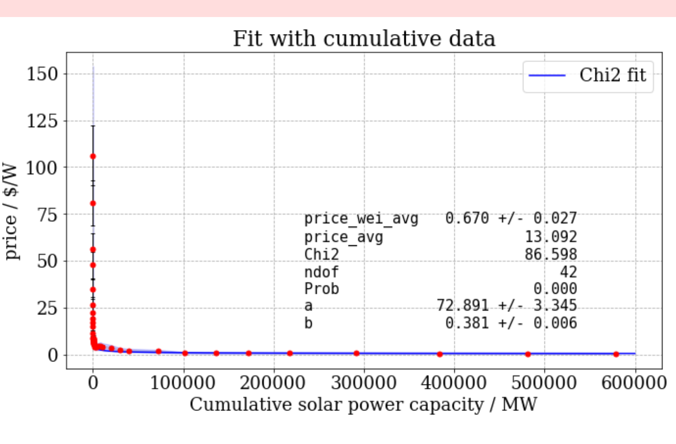
V – Fitting data:

5.1.1



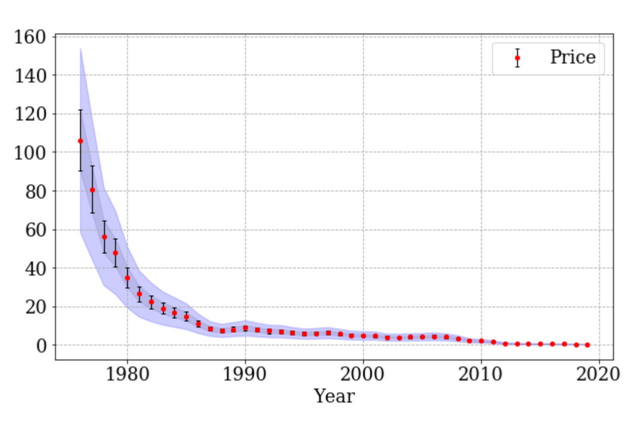
5.1.2

```
c:\users\gaozheming\appdata\local\programs\python\python37\lib\site-packages\ipy
kernel_launcher.py:2: RuntimeWarning: divide by zero encountered in power
```



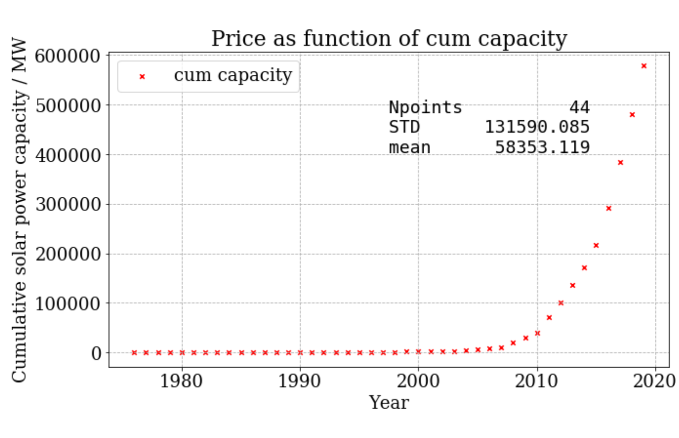
The best χ^2 fit still can't get a decent p value. so the power law fit might not be a good fit.

Inspect price as function of year.

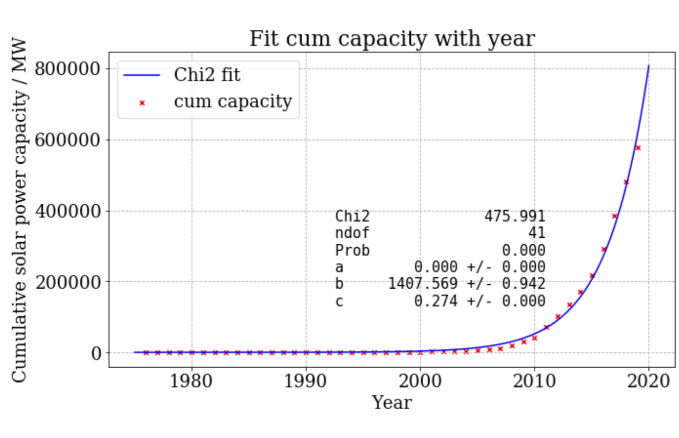


5.1.3

First, plot cumlative solar power capacity with respect to year.

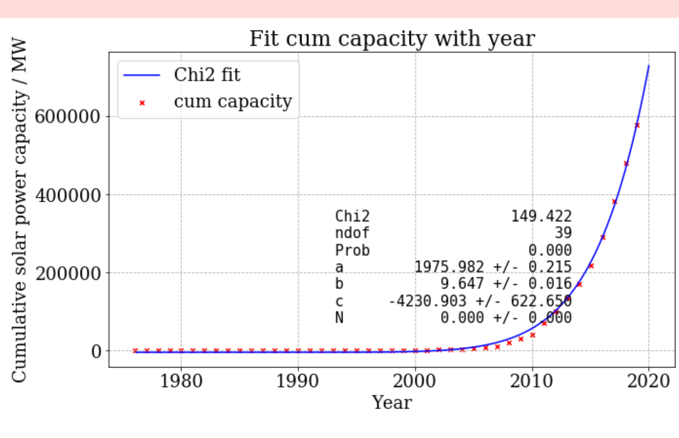


Fit with $a\exp(c \cdot (x-b))$, with absolute uncertainty 3000.

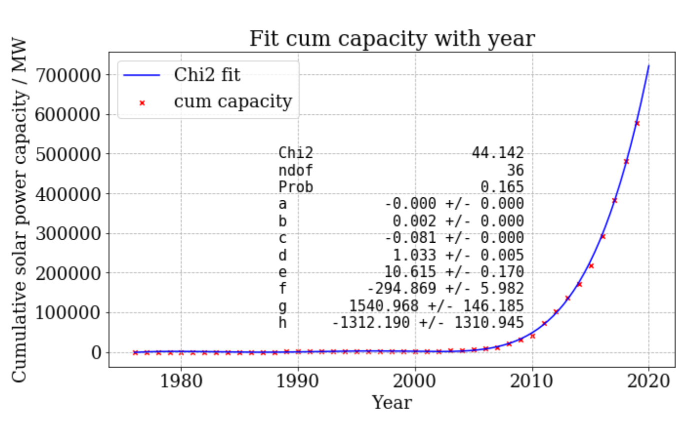


Fit with power function $N*(x-a)^b+c$

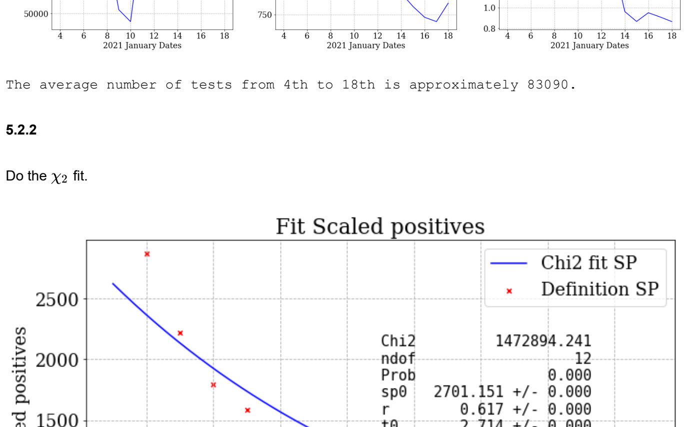
```
c:\users\gaozheming\appdata\local\programs\python\python37\lib\site-packages\ipy
kernel_launcher.py:8: RuntimeWarning: invalid value encountered in power
```



Fit with 7 order polynomial function, set year 1976 as 0 point (Only shown in code).

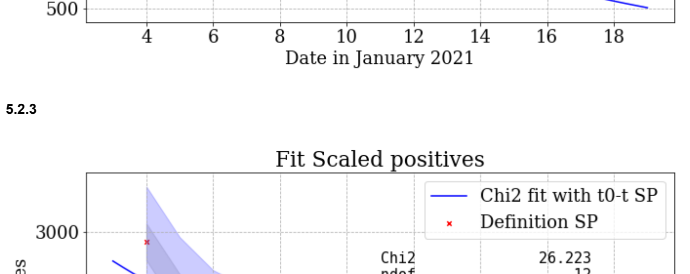



```
Decent fit, using the polynomial fit to estimate the price
Do a price predict in the next 10 years
```




```
The cumulative capacity is over a million MW when at 2022.
When the price is 0.37533 $/W.
Approximately equals to the real price around 2019.
```

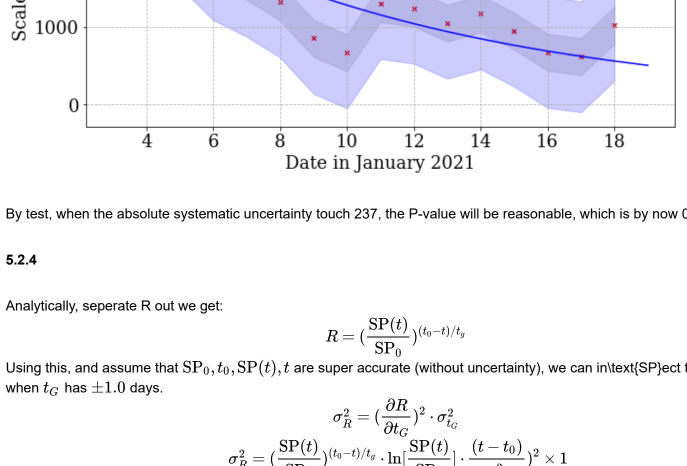
5.2.1



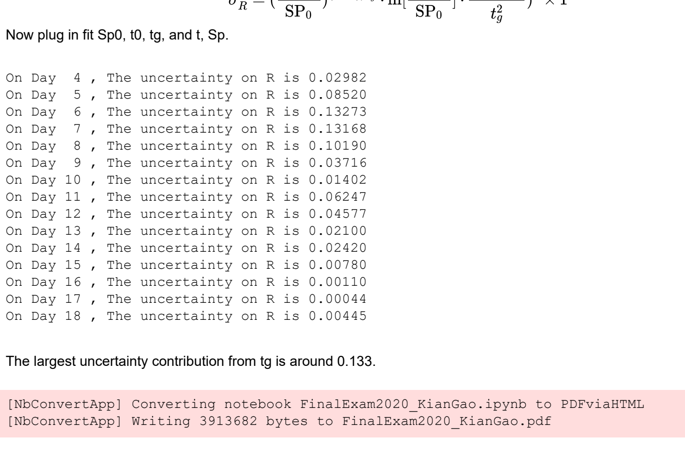

```
The average number of tests from 4th to 18th is approximately 83090.
```

5.2.2

Do the χ_2 fit.



5.2.3



By test, when the absolute systematic uncertainty touch 237, the P-value will be reasonable, which is by now 0.01.

5.2.4

Analytically, seperate R out we get:

$$R = (\frac{\mathrm{SP}(t)}{\mathrm{SP}_0})^{(t_0-t)/t_g}$$

Using this, and assume that $\mathrm{SP}_0, t_0, \mathrm{SP}(t), t$ are super accurate (without uncertainty), we can in\text{SP}ect the contribution when t_G has ± 1.0 days.

$$\sigma_R^2 = (\frac{\partial R}{\partial t_G})^2 \cdot \sigma_{t_G}^2$$

$$\sigma_R^2 = (\frac{\mathrm{SP}(t)}{\mathrm{SP}_0})^{(t_0-t)/t_g} \cdot \ln[\frac{\mathrm{SP}(t)}{\mathrm{SP}_0}] \cdot \frac{(t-t_0)}{t_g^2})^2 \times 1$$

Now plug in fit Sp0, t0, tg, and t, Sp.

```
On Day  4 , The uncertainty on R is 0.02982
On Day  5 , The uncertainty on R is 0.08520
On Day  6 , The uncertainty on R is 0.13273
On Day  7 , The uncertainty on R is 0.13168
On Day  8 , The uncertainty on R is 0.10190
On Day  9 , The uncertainty on R is 0.03716
On Day 10 , The uncertainty on R is 0.01402
On Day 11 , The uncertainty on R is 0.06247
On Day 12 , The uncertainty on R is 0.04577
On Day 13 , The uncertainty on R is 0.02100
On Day 14 , The uncertainty on R is 0.02420
On Day 15 , The uncertainty on R is 0.00780
On Day 16 , The uncertainty on R is 0.00110
On Day 17 , The uncertainty on R is 0.00044
On Day 18 , The uncertainty on R is 0.00445
```

The largest uncertainty contribution from tg is around 0.133.

```
[NbConvertApp] Converting notebook FinalExam2020_KianGao.ipynb to PDFviaHTML
[NbConvertApp] Writing 3913682 bytes to FinalExam2020_KianGao.pdf
```