Applied Statistics: Problem Set

Nathanael van den Berg

January 3, 2022

1 Distributions and probabilities

1.1 5 points

You roll 20 normal dice, count the number of 3s, N3, and repeat this 1000 times.

- What distribution will N3 follow? Why?
- What is the probability of getting 7 or more 3s in a roll with 20 normal dice?

Answers

- There are $N_{dice} = 20$ independent dice each with a chance of $p_3 = \frac{1}{6}$ to land on 3. The n number of dice landing on 3 should follow a binomial distribution: $f(n;, N_{dice}, p) = \frac{N_{dice}!}{n!(N_{dice}-n)!}p^n(1-p)^{N_{dice}-n}$
- The probability of getting 7 or more 3s is equal to 1 - probability of getting less than 7 3s. Use the CDF to calculate this second probability: $F(n; N_{dice}, p) = \sum_{i=0}^{n} \frac{N_{dice}!}{n!(N_{dice}-i)!}p^i(1-p)^{N_{dice}-i}$. Here we just sum to $n = 6$ and find that $p(N_3 \geq 7) = 1 - F(6; 20, \frac{1}{6}) \approx 1 - 0.963 = 0.037$

1.2 7 points

On the 4th of January 2021, the number of Danish Covid-19 tests and positives in 24 hours were: PCR: 103261, with 2464 positives and AntiGen: 26162 with 491 positives.

- Assuming both tests are accurate (i.e. have no errors), what is the fraction of positives in each test? And what is the probability that these fractions are statistically the same?
- If the two tests are sampling the same population, what is the false negative rate (i.e. rate of positive testing negative) of the AntiGen test, assuming no other test errors?
- A test has a 0.02% false positive rate and 20% false negative rate. You test 50000 persons, finding 47 positives. What fraction of the Danish population would you estimate are infected?

Answers

- $\alpha_{CPR} = \frac{2464}{103261} \approx 0.02386$ and $\alpha_{AntiGen} = \frac{491}{26162} \approx 0.01877$
  Assume Poisson distribution Probability that they come from the same distribution?
- Assume that the CPR test has no errors, then α_{CPR} is the real positive rate. The real positive rate is also equal to $\alpha_{AntiGen}$ + the false negative rate $\rightarrow$ False negative rate $= \alpha_{CPR} - \alpha_{AntiGen} \approx 0.0067$.
- 99.906% of the tested persons receive a negative result. 20% of those tests are false negatives, meaning that $\approx 20\%$ of the population is actually infected.

1.3 7 points

The file `www.nbi.dk/~petersen/data_VoltagePeaks.txt` contains voltages from spectrometer measurements. Most of the data are from random noise, but some corresponds to masses, and thus give consistent peaks on top of the noise.

- Plot all the data in as illustrative, informative, and illuminating a manner as you can.
- Fit the peaks that you can find in the spectrum, and comment on their characteristics.

Answers

- See figure 1.

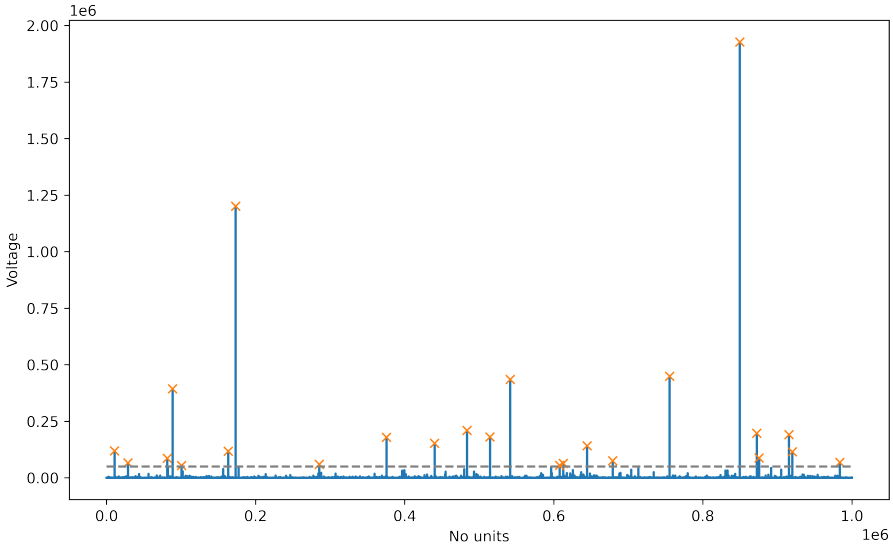


Figure 1: Result for problem 1.3

- I inspected some of the bigger and some of the smaller peaks by zooming in on them. All of them consisted of only one data point sticking out of the noise below. For this reason I chose to 'fit' the peaks by selecting all peaks of a voltage above $5 \cdot 10^4$. The grey, dashed line in figure 1 indicates this criterion.

2 Error Propagation

2.1 6 points

You measure $x = 1.96 \pm 0.03$ to be used in a further calculation of y and z.

- Given x, what are the values of and uncertainties on $y = (1 + x^2)^{-1}$ and $z = (1 - x)^{-2}$?
- What are the values of and uncertainties on y and z, if $x = 0.96 \pm 0.03$ instead?

Answers

- First calculate the error propagation formulas:
  - For y: $\sigma_y = \frac{\partial y(x)}{\partial x}\sigma_x = -\frac{2x}{(1+x^2)^2}\sigma_x$
  - For z: $\sigma_z = \frac{\partial y(x)}{\partial x}\sigma_x = \frac{2}{(1-x)^3}\sigma_x$

  Then use the given value of x to calculate $y = 0.207 \pm 0.005$ and $z = 1.09 \pm 0.07$
- Use the same equations as before to find $y = 0.520 \pm 0.016$ and $z = 625 \pm 938$

2.2 7 points

Students in a statistics class have measured the gravitational acceleration g as follows:

Result (m/s^2)	9.54	9.36	10.02	9.87	9.98	9.86	9.86	9.81	9.79
Uncertainty (m/s^2)	0.15	0.10	0.11	0.08	0.14	0.06	0.03	0.13	0.04

- Assuming independent measurements, what is the best estimate of g and its uncertainty?
- What is the χ^2 and its p-value? Do you find any measurements to be unlikely?
- Does your best estimate of g agree with the precision measurement $9.8158 \pm 0.0001 m/s^2$?

Answers

- Calculate the weighted mean and uncertainty using $\mu = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2}$ and $\sigma_\mu = \sqrt{\frac{1}{\sum 1/\sigma_i}}$ where the x_i are the measured values for g and the σ_i are the uncertainties on those measurements. **This results in** $g = 9.82 \pm 0.02$.
- **I find** $\chi^2 = 32.4$ **with a p-value of 0.000079**. This low p-value isn't very surprising considering that the mean is more than 1 sigma away for 5 out of 9 measurements. For one of those 5 (the second measurement) it's more than 4.6 sigma away. This second measurement should not be used in further analysis. Removing the value results in $\chi^2 = 10$ with a p-value of 0.19 for $g = 9.84 \pm 0.02$.
- **Yes**, the precision measurement is less than one sigma away from my best estimate.

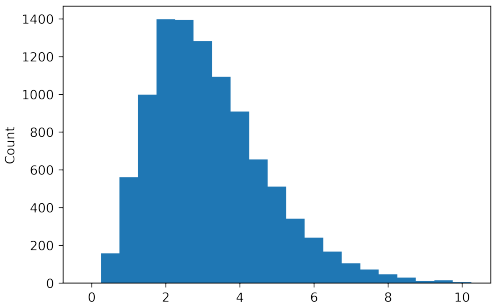
3 Monte Carlo

3.1 11 points

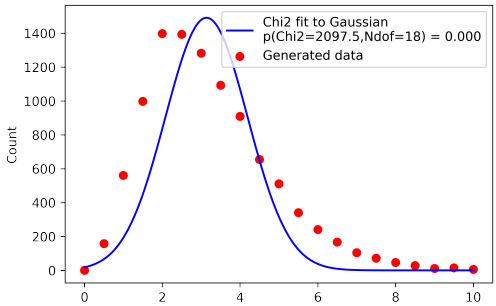
Let u be the sum of 4 exponentially distributed numbers t, with PDF $f(t) = \frac{1}{\tau}\exp(-t/\tau)$ for $t \in [0, \infty]$. Let $\tau = 0.8$.

- Generate 10000 values of u and plot these.
- Try to fit the distribution of u with a Gaussian and comment on the result.
- Try other functional forms to see how well you can match the distribution of u.

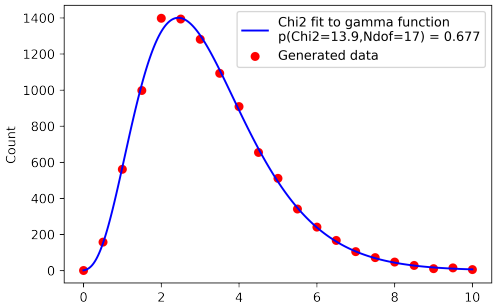
Answers



(a) Histogram of generated random numbers



(b) Fit to normal distribution



(c) Fit to gamma function

Figure 2: Results for problem 3.1

- Integrate $f(t) = \frac{1}{\tau}\exp(-t/\tau)$ to $F(t) = \int_{-\infty}^{t} f(t')dt' = \int_{0}^{t} \frac{1}{\tau}\exp(-t'/\tau)dt' = 1 - \exp(-t/\tau)$, then invert $F^{-1}(p) = -\tau\log(1-p)$ with $p \in [0, 1]$. Finally use uniform random numbers to generate random numbers from the given PDF. **See figure 2a** for the distribution of the generated values.
- See figure 2b for the fit results. The Gaussian distribution does not fit well to the generated data. This is clear from the plot and the p-value.
- I briefly considered fitting a Poisson or binomial distribution, but neither makes sense, because u consists of continuous values. **The gamma function**, on the other hand, does fit well to the distribution as can be seen in **figure 2c**. Here I fitted `N * gamma.pdf(x, a, loc,`

scale) where x is the binned data and the fit parameters are N=4.99, a=4.41, loc=-0.13 and scale=0.752.

3.2 5 points

Let x following the PDF $f(x) = Cx\exp(-x)$ for $x \in [0, \infty]$.

- Generate 1000 values of x, plot these, and determine the median of your x values.

Answer
See figure 3. I used the accept-reject method 10000 times and used the first 1000 accepted values for the histogram and calculations.

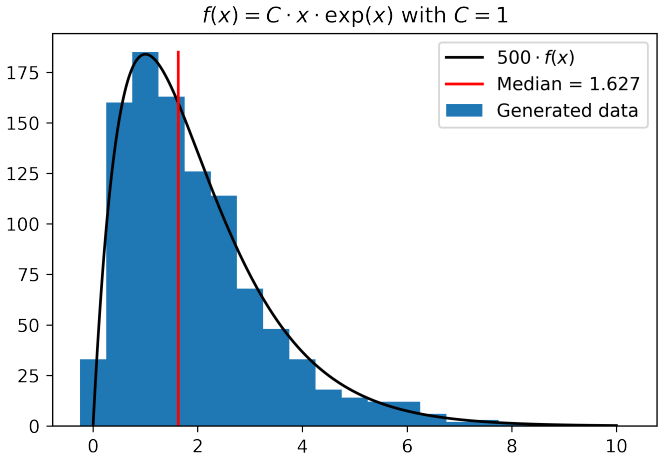


Figure 3: Result for problem 3.2

4 Statistical tests

4.1 12 points

In an observer-blinded study, 21720 persons were given two doses of the Covid-19 vaccine candidate BNT162b2 and 21728 persons two doses of placebo.

- In this study, the *total* number of Covid-19 cases were $N_{vaccine} = 8$ among participants who received BNT162b2 and $N_{placebo} = 162$ among those recieving the placebo. What is (approximately) the probability that BNT162b2 has no effect on being infected?
- Based on the *total* number of Covid-19 cases above, calculate a 68% confidence interval of the BNT162b2 vaccine efficacy, $\epsilon = (N_{placebo} - N_{vaccine})/N_{placebo}$.
- In the study, there were 10 *severe* Covid-19 cases, out of which 9 were in the placebo group. With only this data, what would then be the probability that BNT162b2 had no effect?

Answers

- The probability that the vaccine has no effect is the same as the probability that $N_{vaccine}$ is part of the same distribution as $N_{placebo}$. Assume that the number of Covid-19 cases is Poisson distributed, because there is a large number N of test participants and a small chance p of being infected for both groups.
  Here $\lambda_{placebo} = 162$ (assumption!) and the uncertainty is $\sqrt{\lambda} = 12.7$. $N_{vaccine}$ is clearly more than 10 standard deviations away from the center of the assumed distribution for $N_{placebo}$. This means that there is **approximately 0% chance** that they come from the same distribution and that BNT162b2 has no effect on being infected.
- Error propagate to find the uncertainty on ϵ, here it is assumed that $N_{vaccine}$ and $N_{placebo}$ are the means of two independent Poisson distributions. $68\% \approx 1\sigma \rightarrow$ The confidence interval is $\epsilon = 0.951 \pm 0.018$.
- Use the same assumptions as in the first part of the questions question. $\lambda_{placebo} = 9$. $P(k = 1, \lambda) = 1.2 \cdot 10^{-3}$.

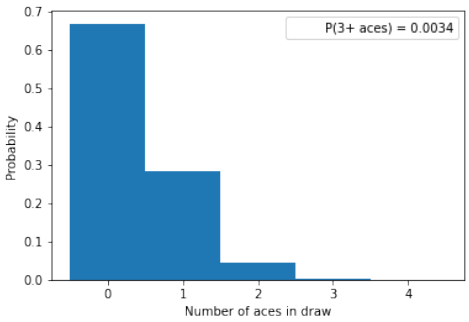
4.2 12 points

The file `www.nbi.dk/~petersen/data_ShuffledCards.txt` contains 52 entries representing a deck of cards.

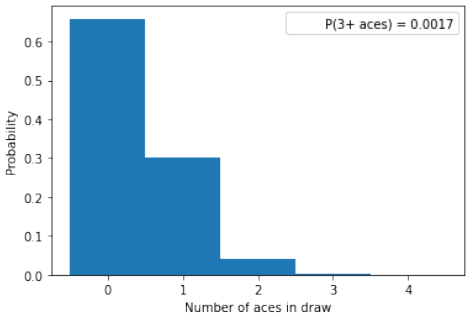
- Drawing 4 cards *with* replacement, what distribution does the number of aces follow? What is the chance of getting 3 aces or more?
- Drawing 4 cards *without* replacement, what is the probability of getting 3 aces or more?
- Are the cards are well shuffled? Perform at least one hypothesis test to check.

Answers

- Numerically generated distribution (1,000,000 trials), see figure 4a. The probability of finding 3 or more aces is the sum of the last two (normalized) bins.
- As before, see figure 4b.
- If the deck is well shuffled, then the distributions of `value` and `suit` should be similar in the first and second half of the deck. From figure 5a it seems to be a fairly well shuffled deck, both halves follow a similar, flat distribution. However, the suits are not so well distributed, in figure 5b it can be seen that all hearts fall in the second half of the deck. This is not a well shuffled deck.

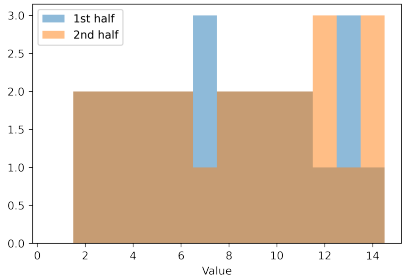


(a) Distribution of aces in a draw of four cards from a deck with replacement.

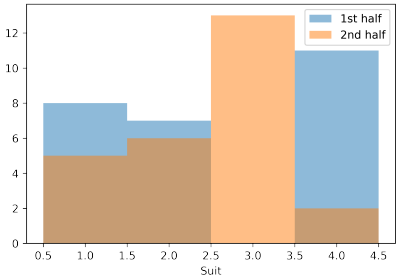


(b) Distribution of aces in a draw of four cards from a deck without replacement.

Figure 4: Results for the first part of problem 4.2



(a) Distribution of card values in the first and second half of the deck. (14 is the ace)



(b) Distribution of the suits in the first and second half of the deck. (1=Clubs, 2=Diamonds, 3=Hearts, 4=Spades)

Figure 5: Results for the second part of problem 4.2

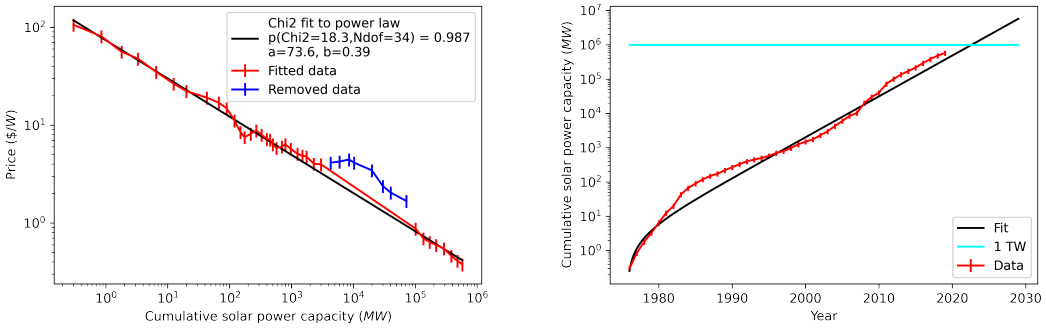
5 Fitting data

5.1 14 points

The cumulative solar power capacity (in MegaWatts) and price of solar power ($\$/W$) from 1976-2019 is listed in the file: `www.nbi.dk/~petersen/data_SolarPower.txt`.

- Plot the price of solar power as a function of cumulative solar power capacity.
- Assuming a *relative* price uncertainty of 15%, fit the data with a power law: $f(x) = ax^{-b}$.
- Fit the cumulative solar power capacity as a function of year, and determine when you expect it to reach a million MW. What do you estimate the price per W to be then?

Answers



(a) (b)

Figure 6: Results for problem 5.1

- See figure 6a.
- A power law fit to all data gave poor results (p-value of 0.00006). After this Chauvenet's criterion was used to remove unlikely values. The removed values are shown in blue in figure 6a. The data points used for the fit are shown in red. This new fit's p-value is sufficiently high (see legend of figure 6a) to be confident that the data follows a power law. The data points of the years 2004-2011 were discarded by Chauvenet's criterion. A possible reason for the relatively high price in this time period might be the commercialisation of solar power. Around 2006 it became profitable to invest in solar power, leading to a great increase in capacity.
- An exponential was fitted to the data as can be seen in figure 6b. This fit was used to calculate the intersection of the exponential with the $1TW$ mark. This is expected to happen in the second half of 2022. The expected price per W at a capacity of $1TW$ is $0.336\$/W$, calculated from the fit used for figure 6a.

5.2 14 points

The number of daily Covid-19 PCR tests and positve cases can for the period 4th-18th of January 2021 be found in the data file `www.nbi.dk/~petersen/data_Covid19tests.txt`.

- Given the number of daily tests T_i, what is the average number of tests $\bar{T}$ in the period?
- Define the number of scaled positives (SP_i) as the number of positives (P_i) times $(T_i/\bar{T})^{-0.7}$, and fit the number of scale positive tests with $SP(t) = SP_0 \cdot R^{(t-t_0)/t_G}$, where $t_G = 4.7$ days.

- How large a systematic uncertainty must be applied, for the fit to give a reasonable p-value?
- How large an uncertainty do you find on R, if t_G has an uncertainty of ± 1.0 days?

Answers

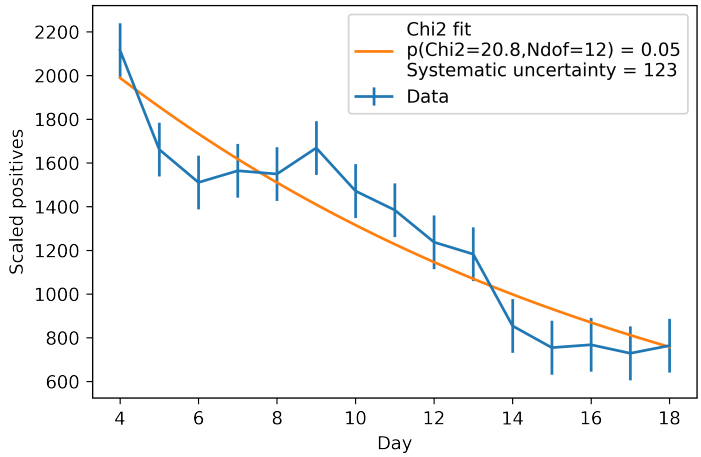


Figure 7: Fit for problem 5.2

- $\bar{T} = 83090$
- See figure 7. Note: iminuit states that the covariance matrix was forced to be positive definite. Fit results: $SP_0 = 3.2 \cdot 10^3$, $R = 0.72$ and $t_0 = -3$ days.
- See figure 7, I considered a p-value of 0.05 or greater a reasonable p-value for this data and fit function.
- I computed this numerically by simulating the effect of different t_G values. t_G was drawn from a normal distribution with a mean of 4.7 and a standard deviation of 1.0, it was then used in the same fitting method as before. I ran 1000 trials and found $R = 0.72 \pm 0.05$.