Applied Statistics

Problem Set in applied statistics 2021/22

This is the problem set for Applied Statistics 2021/22. A solution in PDF format must be submitted on Absalon by 22:00 on Monday the 3rd of January 2022. Links to data files along with code to read the data can be found on the course webpage. Working in groups and discussing the problems with others is allowed. However, you should produce and submit your own solution and state your collaboration(s).

Happy solving, Troels, Mathias, Clara, Kate, Vadim, Iren & Ronja.

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

I – Distributions and probabilities:

1.1 (5 points) You roll 20 normal dice, count the number of 3s, N_3, and repeat this 1000 times.

- What distribution will N_3 follow? Why?
- What is the probability of getting 7 or more 3s in a roll with 20 normal dice?

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?

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.

II – 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?

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 ± 0.0001 m/s^2?

III – 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.

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.

IV – 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?

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.

V – 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?

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 $\overline{T}$ in the period?
- Define the number of scaled positives (SP_i) as the number of positives (P_i) times $(T_i/\overline{T})^{-0.7}$, 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?

"Coincidences, in general, are great stumbling blocks in the way of that class of thinkers who have been educated to know nothing of the theory of probabilities [and statistics] - that theory to which the most glorious objects of human research are indebted for the most glorious of illustration."

[Edgar Allan Poe, "The Murders in the Rue Morgue", 1841]