Applied Statistics

Problem set in applied statistics 2017/18

This problem set was distributed 4th of December 2017, and a solution in PDF format sent by email to `petersen@nbi.dk` must be handed in by Friday the 22nd of December 2017. Links to data files can be found on the course webpage. Working in groups or discussing the problems with others is allowed.

Good luck and thanks for all your hard work so far – Troels, Daniel, Christian & Vojtech.

Statistics has been the most successful information science. Those who ignore Statistics are condemned to reinvent it.

[Attributed to Bradley Efron by Jerome H. Friedman (2001)]

I – Distributions and probabilities:

1.1 (8 points) Two games of dice, each giving the player equal odds are played. In the first game played with one dice, the player wins if getting at least one six in 4 rolls, while in the second game played with two die, the player wins if getting at least once two sixes (at the same time) in 24 rolls (with two die).

- Calculate the odds of winning in both games. Which game would you play, if either?

1.2 (7 points) The IceCube experiment has been running for 1930 days, and has for a particular analysis found an average random background rate of 19.3 event per day.

- What distribution should the daily number of (background) events follow? Why?
- If the experiment in a single day saw 42 events, would that signify a statistically significant excess for the entire data taking period as a whole?

1.3 (8 points) Assume that the height of Danish women follows a Gaussian distribution with a mean of 1.69 m and a standard deviation of 0.06 m.

- What fraction of women are taller than 1.85 m?
- Find (possibly numerically) the average height of the 20% tallest women?

II – Error propagation:

2.1 (5 points) The resistance of a cylindrical resistor is proportional to the length L, and inversely proportional to the cross sectional area $A = \pi r^2$.

- What should the relation between the relative uncertainties on L and r be for them to contribute equally to the uncertainty on the resistance?

2.2 (9 points) A group of ten students have each measured the speed of a bullet, with results as follows:

Measurement	1	2	3	4	5	6	7	8	9	10
Result (10^2 m/s)	3.61	2.00	3.90	2.23	2.32	2.48	2.43	3.86	4.43	3.78

- Assuming independent measurements, what is the average speed and its uncertainty?
- Given $m_{\text{bullet}} = 8.4 \pm 0.5$ g, what is the average kinetic energy, E_{kin} and its uncertainty?
- How much does each uncertainty contribute to $\sigma(E_{\text{kin}})$? And if the speed and mass uncertainties are to contribute evenly, what "suitable" number of speed measurements is needed in total?

III – Monte Carlo:

3.1 (15 points) Let $f(x) = Cx^{-0.9}$ be proportional to a PDF for $x \in [0.005, 1]$.

- In order to fulfill the normalization criteria of a PDF, what value should C have?
- What method would you use to produce random numbers according to $f(x)$? Why? What would your answer be, if the allowed range was $x \in [0, 1]$?
- Produce 10000 random numbers distributed according to $f(x)$ and plot these.
- Let t be a sum of 12 random values from $f(x)$, and generate 1000 values of t. Given this statistics, does t follow a Gaussian distribution?

IV – Statistical tests:

4.1 (17 points) Medical scientists are working hard to develope a cure for the dreaded Fisher Syndrome, a rare but debilitating condition that causes severely reduced ability to enjoy statistics problems. Preliminary research indicates that the disease may be correlated to levels of substances (A, B, and C) in the blood. Data with the levels of these substances from 3000 healthy (index 0) and 2000 ill (index 1) people can be found at **www.nbi.dk/∼petersen/data_FisherSyndrome.txt**.

- What distribution does A seem to follow for ill people? Quantify your statements.
- What is the linear correlation between variables B and C for ill people?
- Using either of the three variables and the combination $F = -1.33 \times A + 0.63 \times B + 2.1 \times C$, what is the separation between healthy and ill people? State your answer in terms of error rates of type I and II (i.e. α and β) for a selection criteria of your choice.

V – Fitting data:

5.1 (15 points) Luke Lightning Lights is a manufacturer of solar powered flashlights. Their monthly income since startup y (in M$) has been recorded as a function of month x in the file **www.nbi.dk/∼petersen/data_LukeLightningLights.txt**. The uncertainty on y has been estimated by the accountant to be $\sigma_y = 0.11$ M$.

- Consider the first twelve months, and test if the monthly income/deficit was constant.
- Assume an initial linear relation between x and y and do a χ^2 fit to first twelve months. Is this hypothesis good? Extending the range, for how many months can this hypothesis be maintained?
- Due to a minor disruptive cost change following the 31st month, the income fell. Estimate the size of the change in income and its uncertainty.
- Try to fit the entire time range with one (or more) hypotheses and discuss its (their) validity.

5.2 (15 points) A class of students have been timing a pendulum, and the residuals of these time measurements in seconds can be found at **www.nbi.dk/∼petersen/data_TimingResiduals.txt**.

- What is the typical timing uncertainty on one single measurement? And is the mean of the residuals consistent with zero?
- Based on the above answers and the size of the sample, do you find any of the residuals suspicious?
- Fit the distribution with a Gaussian and comment on the quality of the fit.
- State alternative PDF hypotheses of your choice, test their validity compared to the data, and quantify to what extent they compare to the Gaussian hypothesis.

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read or write.

[Attributed to H. G. Wells]