Applied statistics

Problem set in applied statistics 2022/23

This is the problem set for Applied Statistics 2022/23. A solution in PDF format must be submitted on Absalon by 22:00 on Tuesday 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 your own code, submit your own solution, and state your collaboration(s).

Thanks for all your hard work, Kate, Rajeeb, Emma, Ting-Yi, Malthe, Mathias, & Troels.

Science may be described as the art of systematic oversimplification.
[Karl Popper, Austrian/British philosopher 1902-1994]

I – Distributions and probabilities:

1.1 (7 points) Your friend tells you, that a bag contains 3 white, 6 black, and 7 grey marbles.

- If you take two random marbles without putting them back, what is the probability that at least one of them is white?
- If you pick a marble, record its color, and then put it back 25 times independently, what is the probability of getting exactly 18 grey marbles? At least 18 grey marbles?
- If you got 18 grey marbles out of 25 picks, would you trust your friend's information?

1.2 (3 points) The lifetime L of a certain component is exponentially distributed: $L(t) = 1/\tau \exp(-t/\tau)$. If there is a 4% chance of this component lasting more than 500 hours, what is the value of τ?

1.3 (5 points) A radio telescope detects 241089 signals/day, based on a 9 week observation campaign.

- One hour, it receives 9487 signals. What is the chance of observing *exactly* this number?
- Is this observation extraordinary, based on an estimate of its *general* probability?

1.4 (7 points) Shooting with a bow, you have 3% chance of hitting a certain target.

- What distribution is the number of hits going to follow, given N shots?
- What is the probability that the first hit will come after 20 shots?
- What is the probability that it will take more than 4000 shots to hit the target 100 times?

II – Error propagation:

2.1 (9 points) Let $x = 1.92 \pm 0.39$ and $y = 3.1 \pm 1.3$, and let $z_1 = y/x$, and $z_2 = cos(x) \cdot x/y$.

- What are the uncertainties of z_1 and z_2, if x and y are uncorrelated?
- If x and y were highly correlated ($\rho_{xy} = 0.95$), what would be the uncertainty on z_1?
- Which of the (uncorrelated) variables x and y contributes most to the uncertainty on z_2?

2.2 (7 points) Five patients were given a drug to test if they slept longer (in hours). Their results were: $+3.7, -1.2, -0.2, +0.7, +0.8$. A Placebo group got the results: $+1.5, -1.0, -0.7, +0.5, +0.1$.

- Estimate the mean, standard deviation, and the uncertainty on the mean for drug group.
- What is the probability that the drug group slept longer than the placebo group?

III – Simulation / Monte Carlo:

3.1 (10 points) Assume $f(x) = Cx^a \sin(\pi x)$ for $x \in [0, 1]$ and $a = 3$ is a theoretical distribution.

- By what method(s) would you generate random numbers according to $f(x)$?
- Determine (possibly numerically) the value of C for $f(x)$ to be normalised.
- Fit a histogram with values from $f(x)$ and determine how many measurements (i.e. values of x) you need in an experiment to determine the value of a with 1% precision.

IV – Statistical tests:

4.1 (12 points) You measure the grip strength (G in Newton N) in the dominant and non-dominant hands (based on writting) of 84 persons, to determine if there is a difference. The data is summarised in the file **www.nbi.dk/∼petersen/data_GripStrength.txt**.

- From this sample, What fraction of persons are right (dominant hand = 1) handed?
- What is the mean and standard deviation of the dominant and non-dominant grip strengths?
- Are the means of the two distributions compatible or different?
- What is the mean and standard deviation of the individual differences in grip strengths?
- Is there a statistically significant difference in grip strengths between hands?

4.2 (14 points) From microscope images, you measure size (S in μm) and intensity (I) of large molecules in a sample, contained in the file **www.nbi.dk/∼petersen/data_MoleculeTypes.txt**.

- Does the molecule size follow a Gaussian distribution? How about when requiring $I > 0.50$?
- Suspecting two different type of molecules, fit the size distribution with two Gaussians.
- Assuming that the double Gaussian fit is good, what size should you require, to get a 90% clean sample of the new molecule? And how many molecules would you then have?
- Including the intesity information, how large a 90% pure sample of the new molecule do you think, that you can obtain?

V – Fitting data:

5.1 (12 points) You are studying the growth of an algae type, by considering the area it covers (A in cm^2) as a function of time (t in days): **www.nbi.dk/∼petersen/data_AlgaeGrowth.txt**. The initially assumed uncertainty on A is $\sigma_A = 45\text{cm}^2$.

- Plot the data, and fit it with a third degree polynomial. Is the fit good?
- Do a runs test on the fit residuals. Does the data seem randomly distributed around the fit?
- You suspect, that there is a day/night variation. Include a multiplicative small oscillation term in your fit, and see if you can improve on the fit and the runs test p-value.
- Is it realistic that the uncertainties are in reality about half of those stated?

5.2 (14 points) In the centennial of Bohr's Nobel prize, you decide to test his atomic model, and measure the spectral lines of hydrogen in the infrared spectrum 1200-2200nm, where you would expect to see some of the $n_1 = 3$ and $n_1 = 4$ lines for $\frac{1}{\lambda} = R_\infty \left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$, where $R_\infty = 1.09677 \times 10^7 \text{m}^{-1}$. Your measurements are in the file **www.nbi.dk/∼petersen/data_BohrHypothesis.txt**, where you have recorded wavelength (λ in nm) and supply voltage (U in V).

- Fit the two prominent known $n_1 = 3$ peaks, which should be at 1875.637nm and 1282.174nm.
- Are the two peaks Gaussian? And if you assume so, are their resolutions (σ) consistent?
- Given your observed peak positions, how would you (linearly) calibrate the scale?
- See how many (significant) peaks beyond the two $n_1 = 3$ peaks you can find.
- Test if their (calibrated) positions follow the Bohr hypothesis.
- You find that your measurements were affected by the supply voltage. Using the two n_3 peaks, calibrate for the variation in supply voltage.