University of Copenhagen Niels Bohr Institute, 2. november 2020

Applied Statistics

Problem Set in applied statistics 2020/21

This is the problem set for Applied Statistics 2020/21. A solution in PDF format must be submitted on Absalon by 22:00 on Sunday the 3rd of January 2021. 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 submit your own solution and state your collaboration(s).

Happy solving, Troels, Giulia, Zuzana, Anna, John & Nikki.

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

I – Distributions and probabilities:

1.1 (4 points) Assuming the "El Clasico" football match is an even game ($p = 0.5$), what is the probability, that the score after 144 non-draw league games is exactly even?

1.2 (4 points) Brad Pitt and Edward Norton are shooting golf balls at a window with $p_{\text{hit}} = 0.054$ chance of hitting. How many golf balls do they need to be 90% sure of hitting the window?

II – Error propagation:

2.1 (10 points) The Hubble constant h has been measured by seven independent experiments: 73.5 ± 1.4, 74.0 ± 1.4, 73.3 ± 1.8, 75.0 ± 2.0, 67.6 ± 0.7, 70.4 ± 1.4, and 67.66 ± 0.42 in (km/s)/Mpc.

- What is the weighted average of h? Do the values agree with each other?
- The first four measurements are based on a different method than the last three. Do the values from the same method agree with each other?

2.2 (10 points) Using Coulomb's law you want to measure a charge, $q_0 = Fd^2/k_eQ$. Assume that Coulomb's constant $k_e = 8.99 \times 10^9$ Nm^2/C^2 and the instrument charge $Q = 10^{-9}$ C are known.

- Given force $F = 0.87 \pm 0.08$ N and distance $d = 0.0045 \pm 0.0003$ m, what is q_0?
- Where does the largest contribution to the uncertainty on q_0 come from? F or d?
- If you could measure F and d with uncertainties $\pm\,0.01$ N and $\pm\,0.0001$ m, respectively, at what distance should you expect to measure the charge in question q_0 most precisely?

2.3 (12 points) Sub-saharan humans tend not to have any Neanderthal DNA, while all others have a few percent. The file: **www.nbi.dk/~petersen/data_DNAfraction.txt** contains the fraction of Neanderthal DNA for 2318 Danish high school students.

- Plot the distribution of Neanderthal DNA fraction, and calculate the mean and RMS.
- Do you find any mismeasurements or outliers from the main population in the data?
- Fit the main population data with distributions of your choice, and comment on the fits.

Statistics like veal pies, are good if you know the person that made them, and are sure of the ingredients.
[Harvard President Lawrence Lowell, 1856-1943]

III – Monte Carlo:

3.1 (15 points) Assume that the outcome of an experiment can be described by first drawing a random number x from the distribution $f(x) = C(c_1 + x^{c_2})$ for $x \in [1, 10]$, where $c_1 = 5$ and $c_2 = 2$ and then using this x value to calculate $y = x\exp(-x)$.

- What is the value of C? And what is the mean and RMS of $f(x)$?
- What method(s) can be used to produce random numbers according to $f(x)$? Why?
- Produce 5000 random pairs (x, y) and calculate the correlation(s) between the (x, y) values.
- Fit the distribution of the produced x values to $f(x)$, with c_1 and c_2 as free parameters.
- How many measurements of x would you need, in order to determine c_1 and c_2, respectively, with a precision better than 1% of their values?

IV – Statistical tests:

4.1 (15 points) The length (l in μm) and transparency (T) of two types of cells (P and E) can be found for 4690 cells in the file: **www.nbi.dk/∼petersen/data_Cells.txt**.

- Selecting P-cells by requiring $l < 9\,\mu$m what is the rate of type I and type II errors?
- Which of the two variables l and T is best at distinguishing between P and E cells?
- Separate P and E cells using l and/or T, and draw a ROC curve of your result.

V – Fitting data:

5.1 (15 points) Kepler's third law states that "the square of the orbital period (T) of a planet is directly proportional to the cube of the semi-major axis (a) of its orbit".
The table lists values for T in days (known very precisely) and a in AU (= 149597870700 m) at the time of the first measurement (in 1778) of the gravitational constant $G_{1778} = (7.5 \pm 1.0) \times 10^{-11} m^3 kg^{-1} s^{-2}$.

Planet	T (days)	a (AU)
Mercury	87.77	0.389 ± 0.011
Venus	224.70	0.724 ± 0.020
Earth	365.25	1 (definition)
Mars	686.95	1.524 ± 0.037
Jupiter	4332.62	5.20 ± 0.13
Saturn	10759.2	9.51 ± 0.34

- Plot the five non-Earth values and fit these to Kepler's third Law: $a = C \times T^{2/3}$.
- In this fit, which planet seems to follow this relation least well? Is it critical?
- From the value you obtain for C and G_{1778} estimate the solar mass $M = 4\pi^2 C^3/G$ in kg.
- Expand the fit to Kepler's third law by further adding two parameters: $a = C \times (T^{c_1} + c_2)$. Does this formula match the data well? Are the two additional parameters necessary?

5.2 (15 points) Searching for slow moving (compared to speed of light) particles at CERN's LHC accelerator, you are calibrating the speed measurement $\beta = v/c$ of the candidate particles, using a control sample of particles known to (effectively) travel at the speed of light, i.e. $\beta = 1$.
The file **www.nbi.dk/∼petersen/data_BetaCalibration.txt** contains 4000 control sample measurements of initial speed estimate (β_{init}), energy (E) in GeV, angle with respect to the beam axis (θ) in radians, and time since start of experiment (T) in seconds, respectively.

- What is the resolution of β_{init}? And is it consistent with a Gaussian distribution?
- Is the distribution in θ consistent with being symmetric around $\pi/2$?
- Test if the mean of β_{init} is constant as a function of energy.
- Due to shifts in timing, the central value of β_{init} shifted with time T, smearing the resolution. Calibrate β_{init} with respect to T and determine the obtained resolution on $\beta_{\text{T-calib}}$.
- Using all information available, what is the best calibration of β you can produce?