Applied Statistics

Problem Set in Applied Statistics 2023/24

This is the problem set for Applied Statistics 2023/24. A solution in PDF format must be submitted on Absalon by 22:00 on Thursday the 4th of January 2024. Links to data files along with code to read the data can be found on the **course webpage** and **GitHub**. Working in groups and discussing the problems with others is allowed. However, you should produce your own code, write your own solution up, and state your collaboration(s).

Thank you for all your hard work so far, Malthe, Azzurra, Arnau, Thomas, Gaia, 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 (8 points) The scores of two tests (A & B) are both Gaussianly distributed with $\mu = 50$, $\sigma = 20$.

- What fraction of students will get a score in test A in the range [55,65]?
- What uncertainty on the mean score do you obtain from 120 B test scores?
- What fraction should get a score above 70 in both tests if $\rho_{A,B} = 0$? If $\rho_{A,B} = 0.75$?

1.2 (4 points) At the roulette you get 12/37 winning chances if you play *douzaine* (e.g. 1-12).

- If you play *douzaine* 20 times, what is the chance that you will win 8 or more times?

II – Error propagation:

2.1 (8 points) Let $x = 1.043 \pm 0.014$ and $y = 0.07 \pm 0.23$, and let $z_1 = xye^{-y}$ and $z_2 = (y+1)^3/(x-1)$.

- Which of the (uncorrelated) variables x and y contributes most to the uncertainty on z_1?
- What are the uncertainties of z_1 and z_2, if x and y are correlated with $\rho = 0.4$?
- Plot $z_1 \in [-2, 2]$ against $z_2 \in [-10, 90]$. In this range, what is the z_1 vs. z_2 correlation?

2.2 (7 points) In a (Cavendish) experiment, you have made five measurements of Earth's density ρ:

Observation	1	2	3	4	5
Result (in g/cm^3)	5.50 ± 0.10	5.61 ± 0.21	4.88 ± 0.15	5.07 ± 0.14	5.26 ± 0.13

- What is the combined result and uncertainty of these five measurements?
- Are your measurements consistent with each other? If not, what is then your best estimate?
- The precise value is 5.514 g/cm^3. How consistent is you measurement with this number?

2.3 (7 points) An ellipse E has semi-major axis $a = 1.04 \pm 0.27$ and eccentricity $e = 0.71 \pm 0.12$.

- The area A of an ellipse is generally $A = \pi a^2\sqrt{1-e^2}$. What is the area of the ellipse E?
- The circumference C has no formula but can be bounded as $4a\sqrt{2-e^2} < C < \pi a\sqrt{4-2e^2}$. What value and uncertainty for C would you give?

III – Simulation / Monte Carlo:

3.1 (8 points) You are optimising container transport, in particular the time, Δt, between the daily truck arrivals (120 minutes uncertainty) and the ship departure (50 minutes uncertainty).

- If $\Delta t = 130$ minutes, what fraction of containers will have to wait to the next day?
- For what value of Δt do containers, on average, have the least waiting time?

3.2 (13 points) The Rayleigh distribution is a PDF given by: $f(x) = \frac{x}{\sigma^2} \exp(-\frac{1}{2}x^2/\sigma^2)$, with $x \in [0, \infty]$.

- By what method(s) would you generate random numbers (from uniform) according to $f(x)$?
- Generate N=1000 random numbers according to $f(x)$ for $\sigma = 2$, and plot these.
- Fit this distribution of random numbers. How well can you determine σ from the fit?
- Test the $1/\sqrt{N}$ scaling of the σ fit uncertainty for $N \in [50, 5000]$.

IV – Statistical tests:

4.1 (15 points) Patients are either healthy or infected with Anoroc disease and their temperature, blood pressure and age is found in **www.nbi.dk/∼petersen/data_AnorocDisease.csv**. For patients 1-800 (control) the outcome in known, while it is unknown for patients 801-1000 (unknown).

- Using the control sample, plot the three distributions for healthy and sick, respectively. Which of the three single measures gives the highest separation between healthy and sick?
- Test if the age distribution is statistically the same between healthy and sick.
- Given any combination of all three variables, separate the two groups as well as possible and estimate the number of infected patients in the unknown group.
- Assuming a prior probability of $p = 0.01$ of being ill, what is the probability that a new patient with $T = 38.5$ C$^\circ$ is ill?

4.2 (14 points) The file **www.nbi.dk/∼petersen/data_CountryScores.csv** contains a list of countries along with several key numbers and indices.

- Determine the mean, median, standard deviation, 15.87%, and 84.13% quantiles of the GDP.
- Does the distribution of $\log_{10}$(PopSize) follow a Gaussian distribution?
- What are the Pearson and Spearman correlations between happiness and education indeces?
- Plot the Happiness-Index as a function of GDP, and fit the relation between the two. From this fit, what would you estimate the uncertainty to be on the Happiness-index?

V – Fitting data:

5.1 (16 points) The file **www.nbi.dk/∼petersen/data_GlacierSizes.csv** contains the estimated area and volume including uncertainties of 434 glaciers with an area above 1 km^2.

- Plot volume as a function of area. Which of the two have largest relative uncertainties?
- Fit data with the expected Area-Volume relation $V \sim A^{3/2}$. Assume no area uncertainties.
- Are you satisfied with the fit? And if not, point out its specific deficiencies.
- Fit again with improved functional form(s), and quantify the improvements.
- Redo this fit including the uncertainties in area. How large is the effect of including these?
- What volume and with what uncertainty would you expect a glacier of area 0.5 km^2 to have?

Don't worry too much about statistics! Just tell us what you do, and do what you tell us.

[Roger Barlow, ICHEP conference 2006, Moscow]