Applied Statistics

Problem Set for Applied Statistics 2024/25

This is the problem set for Applied Statistics 2024/25. A solution in PDF format must be submitted on Absalon by 22:00 on Friday the 3rd of January 2025. 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, Beatrice, Rashmi, Marcela, 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 (6 points) An electronic device depends on three components each with independent probabilities 0.009, 0.016, and 0.027 of failing per year.

- What is the probability that the device will **not** fail in the first year?
- After how many years is the probability of failure greater than 50%?

1.2 (8 points) A store has 52.8 customers/day, and considers the top 20% busiest days to be... busy!

- What distribution should the number of daily customers follow and why?
- Discuss what number of customers exactly constitutes a busy day.
- What is the average number of customers on a busy day?

II – Error propagation:

2.1 (10 points) You make nine measurements of the speed of sound in water, and obtain as follows:

Speed of sound (in m/s)	1532	1458	1499	1394	1432	1565	1474	1440	1507
Uncertainty (in m/s)	67	55	74	129	84	19	10	17	14

- What is the combined result and uncertainty of all your measurements?
- How much does adding the first five measurements improve the precision compared to the last four?
- Are your measurements consistent with each other? If not, argue for an updated estimate.
- The speed of sound in water is 1481m/s. Does your result agree with this value?

2.2 (8 points) A mass is moving in a damped harmonic oscillator with position $x(t) = A\exp(-\gamma t)\cos(\omega t)$ as a function of time t, where $A = 1.01 \pm 0.19$, $\gamma = 0.12 \pm 0.05$, and $\omega = 0.47 \pm 0.06$.

- At $t = 1$, calculate the position and its uncertainty in x position.
- Calculate the uncertainty in x as a function of t for each of the three variables, and comment on which variables dominate the uncertianty during which periods in time.

III – Simulation / Monte Carlo:

3.1 (10 points) You shoot a penalty, and the probability of scoring depends on the position x (in m) you hit, as $p_{\text{score}} = |x|/4$ m for $|x| < 4$ m and zero otherwise (outside goal). Assume the ball hits the goal where you aim with an uncertainty of one meter.

- What is the chance of scoring, if you aim at $x = 2.5$m?
- Where should you aim to have the highest probability of scoring?

3.2 (10 points) Consider the PDF $f(x) = C_{\text{PDF}}(tan^{-1}(x) + \pi/2)$ with $x \in [-3, 3]$.

- Determine C_{PDF} and generate 100 random numbers following $f(x)$.
- Explain how you would fit these data and do so. Does your fit values for C match C_{PDF}?

IV – Statistical tests:

4.1 (10 points) The file **www.nbi.dk/∼petersen/data_LargestPopulation.csv** contains data on the Indian and Chinese population each year in the period 1960-2021.

- Linearly fit the Indian population 1963-1973, and estimate the data point uncertainty.
- Assuming an uncertainty of ±1000000 on all data points, model the population developments and give your best estimate of when the Indian population overtakes the Chinese.

4.2 (5 points) A medical experiment is testing if a drug has a specific side effect. Out of 24 persons taking the drug, 10 had the side effect. For 24 other persons getting a placebo, only 5 had the side effect. Would you claim that the drug has this side effect?

4.3 (5 points) Smartphone producer claims that their phones (A) have a battery lifetime that is significantly longer than that of a rival phone (B). You measure the lifetime of the batteries (in hours) five times for each brand (table below). Test if the claim is reasonable.

A:	28.9	26.4	22.8	27.3	25.9	B:	22.4	21.3	25.1	24.8	22.5

V – Fitting data:

5.1 (18 points) The file **www.nbi.dk/∼petersen/data_SignalDetection.csv** contains 120000 entries with values of measured phase (P), resonance (R), frequency (ν), and type (signal/noise). In the first 100000 entries (control sample) it is known if the measurements are signal (1) or noise (0). In the last 20000 entries (real sample) this is unknown.

- Plot the control sample frequency distribution. Fit the observed H-peak at $\nu = 1.42$ GHz.
- Quantify how well you can separate signal from noise using the variables P and R.
- Selecting entries based only on P and R, how significant can you get the H-peak fit to be?
- Plot the real data frequency distribution, and search for a peak in the range [0.1,1.0] GHz.
- How many signal entries do you estimate there to be in the peak? Do you find it significant?
- Correcting for the signal selection efficiency when selecting events baseed on P and R, how many signal entries do you estimate there was in the data originally?

5.2 (10 points) The file **www.nbi.dk/∼petersen/data_DecayTimes.csv** contains the measured decay times (t_i in s) of a Bohrium isotope. The true decay times follow an exponential function, but the measurement of the decay times given have a Gaussian resolution $G(0, \sigma)$ (thus no bias).

- Plot the distribution of decay times, and calculate the mean and median with uncertainty.
- Give a rough estimate of the decay time τ from fitting the high-t tail of the distribution.
- Fit the entire distribution, and (re-)assess the estimated values of τ and σ.

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]