Applied Statistics

Exam in Applied Statistics 2025/26

This take-home exam was distributed Thursday the 15th of January 2026 at 08:00. A solution in PDF format must be submitted at **www.eksamen.ku.dk by 20:00 Friday the 16th of January**, along with all code used to work out your solutions (as appendix). Links to data files can also be found on the course webpage and github. Working in groups or discussing the problems with others is **NOT** allowed.

Thank you for all your hard work, Nina, Janni, Gabriela, Preet, Clara, Marc, 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) You flip a fair coin 20 times and count the number of tails N_1. You also shoot at a distant target 2500 times ($p_{hit} = 0.004$) and count the number of hits N_2.

- What is the probability of getting $N_1 > N_2$? And vice versa?
- What is the probability of getting $25 \leqslant N_2 \leqslant 250$? How about for N_1?
- What is the probability of getting $N_1 = 10$ and $N_2 = 20$? And vice versa?

II – Error propagation:

2.1 (7 points) Let $z = \cos(x^2)/\ln(xy)$, with $x = 1.71 \pm 0.05$ and $y = 10.1 \pm 0.3$.

- If x and y are uncorrelated, what is the value and uncertainty on z?
- What is the result, if x and y are linearly correlated by $\rho_{xy} = 0.87$?

2.2 (15 points) The file **www.nbi.dk/~petersen/data_WaterDensity.csv** contains 20 measurements of the density of water (in g/cm^3) at five different temperatures (100 in total). You suspect that some of the measurements might not be correct.

- What is the mean density value and its uncertainty for each temperature?
- Would you exclude any measurements as unlikely? List excluded measure, argue quantitatively, and recalculate means and uncertainties, if you exclude measurements.
- At which of the five temperatures does water have the highest density? How confident are you of this?
- Fit the five densities as a function of temperature, and determine the temperature at which water has the highest density.

2.3 (8 points) A particle is measured to have a speed of $\beta = 0.50 \pm 0.02$ (i.e. half the speed of light).

- What is the Lorentz factor $\gamma = 1/\sqrt{1-\beta^2}$ of the particle and its uncertainty?
- What would the answer be, if the speed was $\beta = 0.95 \pm 0.02$? Is the uncertainty symmetric?

III – Simulation / Monte Carlo:

3.1 (16 points) Consider values x from the random harmonic series $x = \sum_{j=1}^{N} \epsilon_j / j$, where ϵ_j can take the values ± 1 with $P(\epsilon_j = -1) = P(\epsilon_j = 1) = 0.5$

- Plot 10000 values of x for $N = 25$ (i.e. $j \in [1, 25]$) as nicely as you can.
- Test if this distribution of x symmetric around 0?
- Calculate 10000 values of x using $N = 250$ terms. Are the two distributions consistent?
- Test if the maximum PDF value of x is consistent with 1/4.

IV – Statistical tests:

4.1 (16 points) The file **www.nbi.dk/∼petersen/data_BloodPressure.csv** contains 2498 systolic blood pressure (SBP in mmHg) measurements from Healthy (H_0), HyperTension (H_1, high blood pressure), or HypoTension (H_2, low blood pressure) patients.

- What is the mean SBP $\bar{s}$ and the 95% confidence interval of this mean for the healthy patients?
- Consider $d = |\text{SBP} - \bar{s}|$ as your test statistic for separating healthy (H_0) from non-Healthy ($H_1 + H_2$) patients. Draw a ROC curve for the separation between the two based on d.
- For a patient with SBP = 98, what is the probability of H_2, assuming equal priors for H_0 and H_2?
- Using priors obtained from the data given, how does the above answer change?
- Try to fit the distribution of SBP values with various PDFs. Do you manage to get a reasonable fit?

V – Fitting data:

5.1 (11 points) The file **www.nbi.dk/∼petersen/data_RunningTimes.csv** contains running time (t) and uncertainties ($\sigma(t)$) data for 21 distances (d) from 50m to 5000m. The uncertainties are determined from the variation in best results observed (i.e. not the individual races).

- Fit the average running times with the function $f_0(t) = vt$, where v is velocity, and comment. Note that the uncertainties given are on t, and should be converted into uncertainties on d.
- Try to best model the running times for all distances. Which record(s) do you find most likely to be improved?

5.2 (19 points) The file **www.nbi.dk/∼petersen/data_RadioSignals.csv** contains $N = 99868$ frequency measurements in the narrow range 12-13 GHz along with two noise discriminating variables disc and electronics temperature index D and E, from a (hypothetical) precursor to the Square Kilometer Array (SKA). The telescope resolution is independent of the frequency.

- Plot the frequency measurements in a histogram, and locally fit the two major peaks and their backgrounds.
- Are the peaks consistent with a Gaussian distribution? And with each other?
- Use the noise discriminating variables to remove some of the noise while preserving the two signal peaks best possible. Refit the two peaks.
- Is there another small signal peak midway between the two large peaks? And if so, how significant can you get it?

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]