University of Copenhagen — Niels Bohr Institute, 9. april 2025

# Applied Statistics

Re-Exam in Applied Statistics 2024/25

This take-home exam was distributed Thursday the 10th of April 2025 at 08:00. A solution in PDF format must be submitted at **www.eksamen.ku.dk by 20:00 Friday the 11th of April**, 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, Malthe, Beatrice, Rashmi, Marcela, Mathias, & Troels.

---

*The knowledge of certain principles easily compensates the lack of knowledge of certain facts.*
[Claude Adrien Helvétius, 1759]

---

**I – Distributions and probabilities:**

**1.1** (9 points) You play two games. In the first game, you flip a fair coin 20 times and count the number of tails $N_1$. In the second, you shoot at a target 2500 times ($p_{hit} = 0.004$) and count the number of hits $N_2$.

- What distribution(s) should $N_1$ and $N_2$ follow?
- What is the chance of getting $N_1 > N_2$? And vice versa?
- What is the chance 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** (16 points) The file **www.nbi.dk/∼petersen/data_WaterDensity.csv** contains 20 measurements of the density of water (in mg/cm$^3$) at five different temperatures (100 in total). You suspect that about 10% 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? 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 with a parabola, 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** (15 points) Consider values $x$ from the random harmonic series $x = \sum_{j=1}^{\infty} \epsilon_j/j$, where $P(\epsilon_j = 0) = P(\epsilon_j = 1) = 0.5$

- Plot 20000 values of $x$ using only 25 terms of the series (i.e. $j \in [1, 20]$).
- To what extend is this distribution of $x$ symmetric around 0?
- Test if the maximum PDF value of $x$ is consistent with 1/4.
- Calculate 20000 values of $x$ using 250 terms. Are the two distributions consistent?

## IV – Statistical tests:

**4.1** (18 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 median SBP $\tilde{s}$ and the 95% confidence interval of SBP for the healthy patients?
- Consider $d = |\text{SBP} - \tilde{s}|$ as your test statistic for separating healthy ($H_0$) from non-Healthy patients. Draw a ROC curve for the separation between the two based on $d$.
- For a patient with SBP of 95, what is the chance of $H_2$, assuming equal priors for $H_0$ and $H_2$?
- If it is known that only 7% of the full population has $H_2$, how does the above answer change?
- Try to fit the distribution of SBP values with various PDFs. Do you manage to get a reasonable PDF?

## V – Fitting data:

**5.1** (15 points) The Taylor expansion $e^{-x} = \sum_{k=0}^{\infty}(-x)^k/k! = 1 - x + x^2/2! - x^3/3! + \ldots$ is known to converge fast. Initially, let $N = 10$.

- For each integer $x \in [0, N]$, generate a random Gaussian value $y$ with a central value of $e^{-x/10}$ and an uncertainty $\sigma_y$ of 1% of the $y$ value, thus $3N$ numbers in total. Plot these illustratively.
- Fit these numbers with a polynomial function representing the first five terms of the Taylor expansion, but with a fit parameter in front of each term (i.e. six parameters). Do you get a good fit?
- What is the maximal value of $N$ for which the fit matches the data to an acceptable degree?

**5.2** (12 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 $f_0(t) = vd$, where $v$ is velocity, and comment.
- Adding a "starting" offset $t_0$, how many distances (starting from 50m) can be modelled well with $f_1(t) = t_0 + vd$?
- Try to best model the running times for all distances. Which record(s) do you find most likely to be improved?

---

*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]