**Statistics Exam 2025-10-03. 08.00-12.00 (13.00).**

**5 Questions. Max points: 24.**

**Pass ≥ 12P.**

Anonymous digital hand-in via Studium (Assignments: ”Stats Exam”).

All study material handed out during the course can be used as help.

You are not allowed to use the internet other than for down/up-loading exam.

The exam is an individual task and suspected cheating will be reported.

You need to answer the questions clearly in written English, providing:

***i) your own motivations for your chosen analysis strategy***

***ii) clear documentation of the essential R-code used to solve the problem***

***iii) your own interpretations and conclusions based on the analysis output***

Add your answers to **a separate document** (word or pdf) that does not include any text from the exam questions found in this document (but note down the number of the question you have answered). Send it in on **Studium** (*Assignments: “Stats Exam”)* as original word docx or as a pdf.

If hand-in fails, send your exam to: [it-kansli@it.uu.se](mailto:it-kansli@it.uu.se), within the time-limit for the hand-in. Send a copy of the email to yourself to know that your email has been sent, and to me ([david.berger@ebc.uu.se](mailto:david.berger@ebc.uu.se)) so I can see that there has been a problem.

*Good luck!*

***David***

**Question 1 (4P)**

An experiment was performed to see how paternal age was related to the number of de novo mutations found in haploid sperm cells. Each of a total of 713 males, of ages spanning 18-56 years, had 10 haploid sperm cells sequenced. De novo germline mutations were then identified and counted by comparing haploid sperm sequences to DNA sequences of hair and skin for respective male.

1. *Motivate your chosen analysis for this data. State the null hypothesis (1P)*
2. *What is your approach to analyzing this data, do you need to do anything before analysis? What is the sample size for estimating the effect of age on germline mutation rate? (1P)*
3. *Partitioning variation in germline mutation rate, SSreg = 36 and SSerror = 97. With your chosen analysis strategy, can you calculate the F-value, degrees of freedom, P-value, as well as the R2 value for the effect of age on germline mutation rate? (2P)*

**Question 2 (5P).**

SciLife lab have carried out mRNA sequencing of 25 libraries for you, but the results are a bit disappointing, with few reads per library. You performed the RNA extractions before sending the RNA to SciLife who prepared the libraries from your samples. Now you are a bit worried that you provided too little RNA.

To investigate if too little sample tissue was the reason for the low read counts, you decide to investigate if there is any relationship between the amount of tissue used for each extraction and the number of mRNA reads you got for the libraries.

You will find the data in the file “***library\_size.txt***”, where mRNA (kilo)reads and the amount of tissue (in mg) are reported for each sample.

1. *Choose an appropriate statistical test and state the null hypotheses (1P)*
2. *Deal with your data in an appropriate way before interpreting your final analysis (motivate all steps and provide a valid interpretation of residual plots) (1P)*
3. *Report the P-value and the appropriate test-statistics of the test (1P)*
4. *Provide a figure that illustrates the main result(s) (1P)*
5. *Provide your interpretations and conclusions based on the results of the test (1P)*

**Question 3 (6P)**

You are interested in understanding responses to temperature of ectothermic organisms and have therefore estimated “fitness” in the laboratory (essentially a combination of traits that correspond to “offspring production/unit time”) across two temperatures (25˚C & 35˚C) for four different ectothermic organisms: baker’s yeast, *E. coli* bacteria, the flour beetle *Tenebrio molitor*, and the plant rockcress (*Arabidopsis thaliana*). You have saved your data in the file “***temp\_growth3.txt***”.

1. *Choose an appropriate statistical test and state the null hypotheses (1P)*
2. *Deal with your data in an appropriate way before interpreting your final analysis (motivate all steps and provide interpretation of residual plots) (1P)*
3. *Report the P-value and the appropriate test-statistics of the test, along with your interpretations and conclusions based on them. (2P)*
4. *Provide a figure that illustrates the result (boxplot using, for example, lattice package or effects package is fine). (1P)*
5. *Do you think your study says anything about differences between fungi, bacteria, plants and insects in terms of differences in tolerance to high temperature? Motivate! (1P)*

**Question 4. (5P)**

You have measured 318 female seed beetles for five traits during their first 24h of adult reproduction (body mass, metabolic rate, water loss, weight loss, and eggs laid). Females were measured at two assay temperatures; either at 23°C or 35°C. The females had either been evolving for 60 generations at cold or hot conditions, and are presumably adapted to their local conditions.

Now you want to know how the five traits have changed due to evolutionary adaptation (i.e. show differences between the cold- and heat-adapted beetles), and if certain traits evolve in concert (i.e. show correlated responses during adaptation). You want to inspect this at the 23°C and 35°C assay temperature separately, so create two subsets of the data.

You want to start by taking a multivariate approach and run PCA that you learnt about on the course to explore variation and the effect of adaptation. You will find the data in the file “***LHdata.txt***” in the exam folder. Note that the five traits are measured on very different scales.

1. *Inspect the five variables to get a feel for the data and see whether some sort of transformation is necessary before continuing with the analysis. Motivate your approach (1P)*
2. *Present and summarize/interpret the important results from the analysis in table format (2P)*
3. *Make plots to illustrate the effect of adaptation at each assay temperature separately – Interpret the plots and explain in what way adaptation has changed the five traits at each temperature, using your own words. Are the results for the two assay temperatures qualitatively different? (2P)*

**Question 5: David’s grapes (4P)**

Your teacher is interested in which grape variety to grow. His daughter, Lovis, likes grapes but does not love the seeds that are inside. Data on the number of seeds found inside 80 grapes were collected in 2025 from two plants of the grape varieties “Spulga” and “Sukribe”. 20 seeds were collected from two stems from each plant (see below). Based on this data and the questions on the next page, can you help your teacher to figure out if one of the varieties produce significantly less seeds than the other?

**Sukribe**

****

**Spulga**

****

1. *The data distributions are illustrated in the figure below. Based on this information, which test would you initially perform, and which distribution would you initially choose for your response variable? (1P)*
2. *Outline an alternative analysis strategy to test your hypothesis that is more robust to assumptions about the data and test criteria. (1P)*

En bild som visar text, diagram, skärmbild, linje

AI-genererat innehåll kan vara felaktigt.

1. *You can inspect the data and look at the variables by reading in the dataset “****grapes.txt****” into R. How many confounding factors in the study design can you identify that potentially may have influenced the conclusions from your analysis? Each suggestion must be explained/motivated (1P)*
2. *How many independent observations are there in your opinion for testing the hypothesis? Can you suggest a better sampling strategy for testing the hypothesis, still measuring 80 grapes in total? (1P)*