**Statistics Exam 2024-10-29. 09.00-13.00.**

**4 Questions. Max points: 24.**

**Grade 3 (Pass) ≥ 12P.**

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

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

If needed, you are allowed to use the internet for searching for R-help, but nothing more.

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** **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 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.

***Make sure to add your anonymity code to the document handed in (which can be used in case there are problems with hand-in on Studium).***

***Good luck!***

***David***

**Question 1. (7P)**

You investigate the frequency of three color-morphs (dark brown, light brown, and green) of larvae of the butterfly *Pararge xiphia*, that feeds on several different grass species. You collect data in the spring and summer of 2024.

You have the hypothesis that the three color-morphs may differ in their frequency depending on the predation they experience throughout the season by birds, where the grass is greener during spring and becomes dry and browner towards the end of summer. You predict that this will affect the camouflage effect that the larvae get from their background when feeding on the grass.

In your sample, you identify 17 dark brown, and 34 light brow, and 53 green larvae in spring. In summer you find 113 dark brown, 133 light brown, and 73 green larvae.

1. *Choose an appropriate statistical test and state the null hypotheses (1P)*
2. *Report the P-value and the appropriate test-statistics of the test (1P)*
3. *Based on your test results, what interpretations can you make, and what is your conclusion regarding your original hypothesis (1P)*
4. *With the data given, can you perform additional tests to learn more about potential differences between morphs? Explain and motivate your strategy? (2P)*
5. *If you were allowed to improve on the experimental design and collect more data, what would you do, and why? (2P)*

**Question 2. (6P)**

You are interested in analyzing the effect of an artificial fertilizer on growth of spruce, which is an economically important group of species in the forestry industry.

You have measured the increase in height (in cm) for four different species of spruce trees, grown with and without the fertilizer. You have saved your data in the file “plants.txt” (found in the exam-folder).

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. *Provide a figure that illustrates the main result(s)* (plot using, for example, *lattice* package or *effects* package is fine) *(1P)*
4. *Report P-values and appropriate test-statistics of the test (1P)*
5. *Provide your interpretations and conclusions based on the results of the test - Should the fertilizer be put in use in forestry of spruce species? (2P)*

**Question 3. (5P)**

A study explored whether (self-reported) exposure to sunlight was related to the incidence of skin cancer reported on a scale from 0-4 (with 0 = no skin cancer, and 1-4 in increasing severity of identified skin-cell changes). The study included 2113 voluntary participants that had gone through check-up and diagnosis.

Partitioning variation in the incidence of skin cancer gave SSreg = 110.7 and SSerror = 2440.2 across the sample.

1. *Can you calculate the F-value, P-value, as well as the R2 value for the effect of sun exposure? (2P)*
2. *What do you conclude about the risks of sun-bathing? (1P)*
3. *What type of analysis was the most appropriate for estimating SSreg and SSerror? –* ***motivate*** *(1P).*
4. *Is the study an example of an experimental study or observational study? What is the main difference between the two study types, and are there any pros/cons associated with them? (1P)*

**Question 4. (6P)**

You have measured 160 female fruit flies for four traits describing an individual’s “pace of life” (body mass, metabolic rate, locomotor activity, and lifespan). Half of the flies were reared at 20°C, and the other half at 28°C, which represents cool and warm temperature for the flies.

Now you want to know if the four traits are correlated to each other (what is the main pattern of variation in these, presumably correlated, traits?). You also want to see if temperature affects these traits. You take a multivariate approach and run a PCA that you learnt about on the course to explore variation and the effect of temperature. You will find the data in the file “pace.txt” in the exam folder.

1. *Inspect the four variables to get a feel for the data and see whether some sort of transformation is necessary before continuing with the analysis.* ***Motivate your decisions*** *(1P)*
2. *Summarize and* ***interpret*** *the important results from the analysis in table format (2P)*
3. *Make a plot to illustrate potential differences between temperatures – in what way do individuals raised at the two temperatures differ?* ***Interpret the plot with your own words****. (1P)*
4. *Analyze each temperature separately and compare the PCAs – is there any fundamental difference in how the traits are related to the PCs (and to each other) now compared to the first analysis, and why does it look different? (2P)*

Hint 1: you can create subsets for the two temperatures from the dataset like this (for example):

subset(myDATA, myDATA$temp==*"20"*) -> Temp20

subset(myDATA, myDATA$temp==*"28"*) -> Temp28

Hint 2: you can create nice plots using the ggfortify package:

library(ggfortify)

autoplot(myPCA, data=myDATA, colour=*"myTREATMENT"*)