**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**

**Summary of results:**

PCA on results:

**En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning**

En bild som visar text, diagram, linje, skärmbild

Automatiskt genererad beskrivning

*Somewhat counterintuitive, high scores on Quizzes were only weakly associated with the final exam score and total grade (see also next page). How come? Because you all more or less had full points on the quizzes, so there was very little variation among students in quiz-scores that in the end could affect differences in the final grade (rather, the good quiz results contributed to overall good grades).*

En bild som visar text, diagram, skärmbild, Plan

Automatiskt genererad beskrivning

*You did really well on the exam, with notably a skewed distribution with 12 students in the end scoring high enough to earn themselves top grade (5).*

*You can calculate your own grade by summing your exam score with the points awarded from quizzes (>70% = 2P/quiz, >80% = 3P/quiz, >90% = 4P/quiz). Below are the cut-offs for different grades.*

|  |  |
| --- | --- |
| Grades | Total score |
| U | <18 |
| 3 | 18-26 |
| 4 | 27-31 |
| 5 | 32-36 |
| half-points rounded upwards – e.g. 31.5P = grade 5. |  |

**For those who did not make the exam, a re-exam is scheduled for February 1st, 09.00-13.00. It will take place with hand-in on Studium.**

**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)*

*We want to analyze frequency data, so a chi-square test seems appropriate. The null-hypothesis is that the three color-morphs do not differ in their relative frequencies across the two seasons – season and color-morph are independent.*

1. *Report the P-value and the appropriate test-statistics of the test (1P)*

*En bild som visar text, Teckensnitt, skärmbild, algebra

Automatiskt genererad beskrivning*

*There is a significant difference in the relative morph-frequencies across the spring and summer season (Chi-2: 31.6, df = 2, p < 0.001).*

1. *Based on your test results, what interpretations can you make, and what is your conclusion regarding your original hypothesis (1P)*

*We reject the null-hypothesis and conclude that the green morph is more common in spring when grasses are green and the dark brown morph is more common in summer when grasses are more often dry, but this is not strong proof of the hypothesis, just patterns that are consistent with it.*

1. *With the data given, can you perform additional tests to learn more about potential differences between morphs? Explain and motivate your strategy? (2P)*

*The relative differences between morphs: Green>dark brown in spring, and Green<dark brown in summer is according to hypothesis. To show this in detail you could, for example, do pairwise comparisons, comparing all groups to each other in chi-2 tests, while controlling for multiple testing. Other suggestions could be to extract observed and expected frequencies for the test and plot these. The data is limited, so also acknowledging this in your answer can give you a point.*

1. *If you were allowed to improve on the experimental design and collect more data, what would you do, and why? (2P)*

*You could sample more seasons to see if the pattern is consistent across years.*

*You could control for variables that could influence the patterns, like rainfall, temperature or predator presence at different localities if you sample more localities.*

*You could look at the frequencies of larvae that sit on green and brown grasses within each season.*

*More examples are possible. The main point is to realize that the results describe a correlation and are not necessarily causational as more factors vary with season. More samples from multiple seasons, and accounting for confounding variables would be necessary.*

**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)*

*We want to test for effects of treatment on growth (increased spruce height in cm), which likely is normally distributed. You also have four different species, so a two-way ANOVA seems appropriate.*

*The main null hypothesis is that there is no difference in mean increase in height between spruce grown with and without fertilizer.*

*As you are also interested in whether the spruce respond the same to the fertilizer, the interaction is also interesting, and the null hypothesis for this effect is that all four species respond in the same way to the fertilizer.*

*There is a third null hypothesis – that the four species do not differ in their mean growth (across the two treatments), but we are not so interested in this effect.*

1. *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)*

*lm(growth~Species\*treatment, data = plants)->model.1*

*par(mfrow=c(2,2))*

*plot(model.1)*

**

*There is a trend that residuals increase with the fitted values of the model (top and bottom left). Log-transformation usually helps here. Logging data might also have another advantage – we will compare effects on log-scale, which correspond to relative differences between treatments, which makes sense if the four spruce species are of very different size – we can expect that such species will show different responses to treatments on raw data scale, but only as a result of them showing different capacity for growth due to their size. Comparing their growth in relative terms (log-scale) therefore makes more sense when interpreting the effect of treatment. Let’s log-transform and see if it improves model fit!*

*lm(log(growth)~Species\*treatment, data = plants)->model.2*

*plot(model.2)*

**

*Only a slight improvement, but OK to continue with the ANOVA*

1. *Provide a figure that illustrates the main result(s)* (plot using, for example, *lattice* package or *effects* package is fine) *(1P)*

*library(lattice)*

*bwplot(growth ~ treatment|Species, data=plants)*

En bild som visar text, skärmbild, diagram, linje

Automatiskt genererad beskrivning

1. *Report P-values and appropriate test-statistics of the test (1P)*

*En bild som visar text, Teckensnitt, skärmbild

Automatiskt genererad beskrivning*

*There is an effect of the fertilizer (F1,16 = 13.5, P = 0.002), and no significant difference in how the four species react to the fertilizer (F3,16 = 2.0, P = 0.15).*

1. *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)*

*There seems to be an effect of the fertilizer, but it has a negative effect on growth in three out of four species! There must be something added to the fertilizer that hampers spruce growth. We should definitely NOT use this product if we want to grow spruce! We can also see that even if the interaction term is not significant, the effect is completely absent in one of the species, so we cannot talk about a general effect until more species are studied, but that seems pointless given the negative outcome for ¾ species.*

**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)*

*F = 110.7/(2440.2/(2113-2))*

*[1] 95.7658*

*P = pf(95.8, df1=1, df2=2111, lower.tail=F)*

*[1] 3.731954e-22*

*R2 = 110.7/(110.7+2440.2) = 0.043*

1. *What do you conclude about the risks of sun-bathing? (1P)*

*There is an increased risk of skin cancer when sun-bathing. The effect is very significant due to the large sample size, but the effect is small; variation in the extent of sun bathing only explains ca 4% of the variation in incidence of skin cancer. Other factors, like genetics and variation in accuracy of self-reporting likely play a role.*

1. *What type of analysis was the most appropriate for estimating SSreg and SSerror? –* ***motivate*** *(1P).*

*As the incidence of skin cancer is scored on a ordinal scale with strength increasing from 0->4 and sun bathing is continuous/ranked variable, a simple regression would be most powerful.*

1. *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)*

*It’s an observational study. It’s typically not as powerful as an experimental study as you cannot control for a range of potentially confounding factors – in this case – bias in self-reporting that can depend on how much the person likes sun bathing, and the fact that people that are sensitive to sun may expose themselves to sun less and vice versa. Other factors that covary with skin cancer and sun bathing, so that the observed relationship is not causal but merely a correlation, The participants may not be a random sample of the population, etc. Experimental studies apply a treatment to subjects in randomized fashion and therefore do a better job of controlling for confounding factors. Observational studies are weaker in this aspect, but are easier to do as data can be collected from existing information.*

**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)*

*I inspect the four traits for each temperature first. Start with 20C:*

*plot(log(pace[pace$temp=="20",2:5]))*

En bild som visar text, diagram, mönster

Automatiskt genererad beskrivning

*For 28C:*

*plot(log(pace[pace$temp=="28",2:5]))*

**

*Everything together:*

*plot(log(pace[,2:5]))*

En bild som visar text, mönster

Automatiskt genererad beskrivning

*Logging helped a lot with some data at 28C, mainly because some outliers and non-linearities between different variables, so I decided to look at logged variables. Here is also the distribution of variables at 28C when logged relative to raw data:*

*Log-transformed:*

En bild som visar text, diagram, skärmbild, linje

Automatiskt genererad beskrivning

*Raw data:*

En bild som visar text, diagram, skärmbild, Plan

Automatiskt genererad beskrivning

1. *Summarize and* ***interpret*** *the important results from the analysis in table format (2P)*

*En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning*

*There seems to be significant variance along the first three PCs (45%, 28% and 21.5% respectively), and it is hard to throw out more than one PC. Focusing on PC1, that still explains almost half of the variation, we see that it loads heavily negatively on mass and metabolism, meaning that females with low scores on PC1 are heavy and have high metabolism. PC2 is interesting as it suggests a trade-off between being active and living long; females scoring high on PC2 are moving around more but die younger. Remember, we have note statistically verified these relationships with proper hypothesis testing, but this is what the PCA suggests. 2 points were awarded if you explained both variance explained and loadings (how original variables relate to PCs).*

1. *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)*

*library(ggfortify)*

*factor(pace$temp)->pace$temp*

*autoplot(PCAtot, data=pace, colour="temp")*

En bild som visar text, skärmbild, diagram, Graf

Automatiskt genererad beskrivning

*Interestingly, temperature separates individuals along both PC1 and PC2. Individuals reared at hot temperature have higher scores on PC2, meaning that they are more active but die young, and they also score low on PC1, meaning that they mature at a larger size and have higher metabolism.*

1. *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)*

En bild som visar text, skärmbild, Teckensnitt, dokument

Automatiskt genererad beskrivning

The PCAs are similar, but the loadings on PC1 and PC2 have been rotated so the signs are reversed. The eigenvalues of the PCs are similar - the dimensionality of the studied response is similar (hard to throw out dimensions except for the fourth dimension).

In addition to this, we can see that at 20C (top), there is a pattern quite similar to the overall analysis, describing trade-offs (negative correlations) between being active and living long, and then a separate axis for variation in metabolism and mass.

En bild som visar text, skärmbild, diagram, Graf

Automatiskt genererad beskrivningAt 28C (bottom), the covariation between the four traits is better described by positive covariation between mass, metabolism and longevity, which could indicate that there is more general differences in condition between individuals exposed at stressful hot temperature. PC2 instead describe variation in locomotor activity that is more or less independent of variation in the other variables.

The results from the analyses with each temperature separately are fundamentally different compared to the first analysis, as the former captures effects of temperature on trait variation, whereas the altter two describe (co)variation in traits within each temperature.