**BIFX 503: Biostatistics in R**

**Midterm Examination**

**October 6, 2021**

*Instructions:*

*This is an open-book, open-note exam … books being the course textbook (Hothorn & Everitt) and notes being your personal notes taken for class, homework assignments, and course slides.*

*Permitted technology for calculations is limited to R on your desktop. No internet use (other than Blackboard) or smartphone use is permitted during the exam.*

*Type your answers into this Word document, and paste in results from R as needed. You may use RMarkdown if your prefer. Include graphs and results from statistical procedures (ANOVA, regression). When you finish the exam, submit it via Blackboard. All exams are due at* ***11:59pm*** *today.*

*The old cliché is true – you miss 100% of the shots you don’t take ☺ Please don’t leave any questions blank! Partial credit will be given for incomplete responses.*

*Good luck!*

Nitrofen is an herbicide that was assumed to be safe, but was found to be teratogenic and was the first pesticide withdrawn from the market due to teratogenic effects. The dataset **Nitrofen.csv** contains data demonstrating the reproductive toxicity of nitrofen in zooplankton (*Ceriodaphnia dubia*). In this study, 50 animals were randomized into batches of 10, and each batch was put in a solution with a different concentration of nitrofen. The number of live offspring born to each animal was recorded. The dataset contains the following variables: conc=nitrofen concentration, offspring=total number of live offspring.

1. Generate a box plot that shows how the total number of offspring varies by nitrofen concentration. Calculate summary statistics (mean and SD) for total offspring for each concentration of nitrofen. Does the reproductive toxicity of nitrofen seem to vary by concentration?

Based on the graph and summary statistics, the reproductive toxicity does appear to vary by concentration. Larger concentrations seem to be increasingly toxic.

Chart, box and whisker chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

1. Test the hypothesis that different concentrations of nitrofen result in different numbers of live offspring in *Ceriodaphnia dubia* using a one-way ANOVA. Do the ANOVA results corroborate the evidence from the box plot and confidence intervals? Use post-hoc tests if warranted.

The ANOVA results support what was seen in the box plots and summary statistics. The p-value of 1.04e-15 confirms that there is a significant difference between offspring of each concentration. The post-hoc test shows us that the 0, 80, and 160 categories do not have significant differences. But the 235 and 310 definitely do.

Text

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

1. Explain what a *sampling distribution* is, and how it is different from a *sample distribution*.

Sampling Distribution is the distribution of sample statistics from all samples of the same size, while Sample Distribution just shows how the measured data are distributed. Sampling distribution might be measured by a histogram of means from several samples of the same size. While Sample Distribution might be measured by a histogram of a single sample.

1. Explain what a non-parametric method is, and how the non-parametric methods described in the text overcome distributional issues such as skewness and outliers.

Many procedures assume data is normally distributed. But it is hard to determine whether this is actually the case when working with a small data set. Non-parametric methods, such as the Wilcox test, help us get around this problem by ranking data to remove assumptions like normality. This allows us to work on smaller data sets with more precision.

The dataset **CPUs.csv** contains data on the relative performance of 209 CPUs. The variables include manufacturer/model, cycle time, minimum and maximum main memory, cache size, minimum and maximum number of channels, and log10 of performance. You are going to use correlation and linear regression to determine the correlates of CPU performance.

1. Make six scatterplots to visualize the relationship between log\_perf and each of the other quantitative variables (cycle\_time, min\_memory, max\_memory, cache, min\_channels, max\_channels).

Diagram

Description automatically generated

1. Create a subset that consists of the seven quantitative variables (i.e., all except manufacturer/model). Calculate the correlation matrix for these variables. Which variables are positively correlated? Negatively correlated? Highly correlated (i.e., correlation exceeds 0.5)?

It appears that that cycle time is negatively correlated to everything. All other variables are positively correlated. Pairs with high correlation include cycle\_time and log\_perf, min\_memory and cache\_size, min\_memory and min\_channels, max\_memory and everything but cycle\_time, cache\_size and min\_channels, cache\_size and log\_perf, min\_channels and max\_channels, max\_channels and log\_perf.

A screenshot of a computer

Description automatically generated with medium confidence

1. Fit a multiple linear regression model having log\_perf as the dependent variable and all other quantitative variables as independent variables. (This is one model, not a series of models as in the homework.) Evaluate the fit of the model by looking at the residuals plot and R square value. What are your conclusions, regarding how well the model fits the data?

The multiple r-squared value is 0.813 which indicates that our model explains 81% of the variation. This is a sign of good model fit. However, the residuals plot does not seem very random and has a distinctive arc shape which may be indicative of poor model fit.

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

1. Report the results in terms of regression coefficients and 95% confidence intervals. Which variables are significantly (p<0.05) associated with performance? What does the model tell you about predictors of CPU performance?

For cycle\_time, the regression coefficient was -3.629e-04 and the confidence interval was 4.779519e-04 to -2.478726e-04. Since the p-value was 2.81e-09, cycle\_time is significantly associated with log\_perf.

For min\_memory, the regression coefficient 1.176e-05 and the confidence interval was -2.325813e-07 to 2.375905e-05. Since the p-value was 0.0546, min\_memory is not significantly associated with log\_perf.

For max\_memory, the regression coefficient is 1.854e-05 and the confidence interval was 1.432446e-05 to 2.275335e-05. Since the p-value was 1.38e-15, max\_memory is significantly associated with log\_perf.

For cache\_size, the regression coefficient was 3.348e-03 and the confidence interval was 2.431509e-03 to 4.264382e-03. Since the p-value was 1.14e-11, cache\_size is significantly associated with log\_perf.

For min\_channels, the regression coefficient was 2.919e-03 and the confidence interval was -2.699916e-03 to 8.537655e-03. Since the p-value was 0.3069, min\_channels is not significantly associated with log\_perf.

For max\_channels, the regression coefficient was 9.322e-05 and the confidence interval was -1.351668e-03 to 1.538107e-03. Since the p-value was 0.8989, min\_channels is not significantly associated with log\_perf.

The most significant predictors of CPU performance are cycle\_time, max\_memory, and cache\_size. For every 1 unit increase in cycle\_time, we except to see a decrease of log\_perf by

-3.629e-04 units. For every 1 unit increase in max\_memory and cache\_size, we expect to see respective increases of log\_perf by 1.854e-05, and 3.348e-03 units.

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated with medium confidence

Text

Description automatically generated