Final Lab Project – Biostatistics for Life Science

Primary Biliary Cirrhosis Analysis

You are to perform required data analysis ***on your own*** and write a written report by yourself to be evaluated and graded. ***This is an individual project, and it is <u>not</u> a collaborative exercise***. You are expected to work on your own on this project and you are not allowed to share your project, results, electronic files, or documentation with anyone.

You are not allowed to communicate or discuss approaches, code, questions, clarifications, or anything at all involving the project with other students in any capacity. This is to ensure that each student has a basic understanding of biostatistics, how to apply statistical measures, and how to work with R. If you have any questions, please create an office hour appointment with Dr. Boies. www.calendly.com/lboies

You can think of this like a take-home exam to test your understanding of R and statistical measures.

Description and Instruction of Individual Project:

- Data to be used: Mayo_Hepato.csv
- Natural history of 418 subjects with primary biliary cirrhosis (PBC)
- Please refer to the PDF for the information on all variables.
- Each person will be assigned a random number in order to create their own unique dataset. You will get your number from Dr. Boies, once you have that number, you will use it where the number "122" is in the following code. You will name your new generated dataset something (I named it Boies), and you will do all of your analysis on your random dataset.
- `install.packages("dplyr")`
- `library(dplyr)`
- `Boies <- sample_n(Mayo_Hepato, 122)`

Write-Up:
Introduction:

- This section is expecting to answer, "What is the rationale for the scientific question asked?" The rationale needs to be based on medical significance of the relationships between the different variables in the dataset. Describe the relationships of interest and the purpose of the analysis. Conduct a small PubMed literature search (5 or more studies/sources) to support the scientific question asked in the project and provide a brief summary of the issues in this section.

- Evaluation of this section will be based upon your ability to clearly state the problem, its relevance to medicine, supported work of at least 5 peer review articles from primary journals, and clearly state the purpose of the current research question.
    - You will be required to use proper in-text citation and have a works cited page. You are welcome to use MLA or APA format.

Hypothesis:
Students are asked to create their own hypothesis based upon the dataset and background information from reading the primary literature. The hypothesis does not have to be completely novel, but something the student has put thought into. As scientists, it is an important skill to be able to read the primary literature, reflect upon it, and develop our own questions.

As with any questions that arise during the course of the project, please contact Dr. Boies if you need help!

Methods:

- This section should describe what steps and statistical methods you did to analyze the data and how you applied them to solve the questions asked. You need to provide a description of what statistical methods were used and the rationale or purpose for it. Please describe any statistical methods used for testing assumptions of the test if applicable. If you created new variables for your analyses, you need to provide the rationale for creating the variable and describe or define the method you used to create the new variable.

- You will not put code in this section, but you should reference your statistical software (R) and any packages that you utilized for this analysis. (This is what you would do if you were writing an article for peer-review publication).

- Evaluation will be based on the extent to which the analytic steps involved in each component of the results is described. It is typical to describe the analytical steps in the order of which it is present in the results so that the reader can follow and knows what to expect in the results.

Results:

Part I:

- Describe the data in terms of the descriptive statistics that you calculated (see Model Table 1 below).
    - Generate new categorical variables as needed so that they match the categories of the variable in the given table template (you may have to do this for status).
    - For each continuous variable in the table, compute the appropriate measure of location and dispersion (i.e., for variables normally distributed compute the mean and SD, and for variables non-normally distributed compute the median and IQR). Indicate in the table which measures you computed.
    - For each set of variables in the table, compute the appropriate parametric or nonparametric test statistic for continuous outcomes or perform a chi square/fisher's exact test for categorical outcomes to assess the simple association between each independent variable and the deceased and not deceased status.
        - Describe the statistics used above in the Methods section and report the P-value in the table.
        - For categorical variables report only the P-value for the overall chi-square/Fisher's exact test on the line with the category label. Don't compute separate chi-square/Fisher's statistics for the individual levels.
    - Summarize your findings based on the initial descriptive statistics in a brief paragraph.

Part II:

- Generate a new variable defining categories of triglycerides:
    - Less than 150: Normal
    - 150 – 199: Mildly High
    - 200 – 499: High
    - 500 or Higher: Very High
- Evaluate if there are differences in the:
    - Mean serum bilirubin levels and the triglyceride categories described above
    - Mean albumin levels and the triglyceride categories described above
    - Mean alkaline phosphatase levels and the triglyceride categories described above
- For all of the above, assume that the data are normally distributed, and test the above-mentioned hypotheses. Report the p-values for the overall tests.
    - If the overall tests is statistically significant, use an appropriate pairwise multiple comparisons procedure to test for differences in means to identify specific differences between categories of triglyceride level.
- Summarize your findings in two to three paragraphs.

Use this table below to report your descriptive statistics.

- For deceased, you will use the "status" variable. Deceased will be status = 2 and not deceased will be status 0 and 1.
- You should add units where applicable (I did not put it in all places below).
- For sex, pick one sex (and indicate that in table) and report the percentage for deceased and not deceased
- For some data points you will report percentages (and then the number), others you will report the mean ± the standard deviation or the median (IQR).
    - I have helped by putting some of this information in the table, but I did not put it in for all parts. You should be able to analyze the data and report on your own.
- P-values will be calculated utilizing the different tests we have learned throughout the semester.

Model Table 1: Descriptive Statistics

	Deceased (N =)	Not Deceased (N =)	p-Value
Age			0.xxx[a]
Sex			
Albumin Levels (units)			
Alkaline Phosphatase Levels			
Aspartate Aminotransferase Levels			
Serum Bilirubin Levels			
Serum Cholesterol Levels			
Triglycerides Levels			
Urine Copper Levels			
Platelet Count			
Edema			0.xxx[b]
No Edema (% (n))			
Untreated or Successfully Treated (% (n))			
Edema Despite Diuretic Therapy (% (n))			
Presence of Hepatomegaly (% (n))			
Presence of Ascites			
Presence of Blood Vessel Malformations in the Skin			
Histologic Stage of Disease			
Stage 1			
Stage 2			
Stage 3			
Stage 4			

a – p-values from ---- test
b – p-values from ---- test
c – p-values from ---- test

Discussion:
In this section you need to describe what the results mean in the context of the scientific question integrating all the questions asked in the project. Discussion should include a summary of overall findings (but **not** a restatement of your actual numerical results), reference back to the original introduction and material cited, implications of the findings to medicine, and the conclusions. You may include limitations if any from your analysis/dataset that you observe that may have influenced your findings.

Submission:
You will upload the written report (a word file or PDF) as well as your .R file outlining all statistical procedures that you may have used with appropriate comments. (Please do ***not*** copy and paste your code into the word file).
Your .R file **must** be executable without error and I should be able to replicate your results without error (meaning, if I highlight all of your text and run it, I should get all of the work that you did and no errors). If your .R file does not run without error, *it will be graded as thought your* ***results are incorrect***.

Finally, the purpose of this project is for you to gain real life experience in both conducting and reporting statistical analyses. Although you are not writing a full scientific paper, you are still expected to utilize professional language and formatting as if your report would be included in a paper.

You do not need to show calculations, step-by-step procedures, or all R output, but you should show relevant output (graphs/tables).

For those of you who may need help with the technical writing aspects, I highly encourage you to utilize the Writing Center in the Rattler Success Center. You also likely have a writing guide from Biology 1401/1402 Lab.
Additionally, you can use the following guide (focus primarily on the methods, results, and discussion):
https://www.bates.edu/biology/files/2010/06/How-to-Write-Guide-v10-2014.pdf

Overall, your report should be professional. Refer to a published manuscript for appropriate writing style and display. Points will be deducted for unprofessional reports, failure to follow reporting instructions, or unsuitable formatting.

Your written report will likely average between 4 – 11 pages depending on how succinctly you write. Please use the following formatting:

- Calibri (11 point font) or Times New Roman (12 point font)
- 1.5 spacing
- 1 inch margins

Examples of different sections (these examples are just 2 paragraphs from each section; meaning this is a truncated example to just help give you an idea/guide you. It is not an extensive, all-encompassing example):

Introduction

Early detection and identification of cells with the potential to be positive for breast cancer and additionally predict recurrence is an important factor when considering effective treatments and long-term patient outcomes. At the cellular level, there are many changes that will signal if the cell is benign or cancerous. Through the use of fine needle aspiration (FNA), samples of breast tumors can be obtained and their cells investigated using a variety of criteria including tumor size, cell radius, cell-nuclei area, cell concavity, compactness, smoothness, symmetry, fractal dimension, and cell-nuclei texture. Utilizing machine learning to analyze prepared histological slides may provide diagnostic information free from user-bias to counsel patients on effective treatments and long-term outcomes of their malignancies (Wolberg, *et. al.,* 1995).

Larger tumor volume size is associated with an increased likelihood of malignancy, but the best predictor of tumor outcome is through genotyping and observation of additional features such as changes to nuclear morphology (Kasangian, *et. al.,* 2017). Malignant cells have shown evidence of enlarged nuclear areas ($59\mu^2$ for malignant cells compared to $25\mu^2$ for benign cells) (Laishram, 2017). Furthermore, changes in the nuclear morphology is evidenced by the presence of concavity resulting in distortions of the nuclear boundary in cells. Presence of concavity (including the number of concave points) are important characteristics when categorizing the malignant cells (Das, *et. al.,* 2020). Studies have shown cell-nuclei texture can be successfully used as a prognostic indicator between malignant and benign cells (95.8%) and can differentiate between grades (93.3%) (Doyle, *et. al.,* 2008). Texture changes can be attributed to nuclei being vacuolated and a significant decrease in interstitial fibrous tissue (Yu, *et. al.*, 2019).

Methods (Don't forget, if you load any R libraries to reference those!)

From the Wisconsin Prognostic Breast Cancer (WPBC) dataset of 500 samples, a random sample of 120 observations was used for statistical analysis (28 recurrent and 92 non-recurrent). The statistical program R was used for all statistical tests and data analysis. Data for each tumor category for recurrent and non-recurrent were assessed for normal distribution through graphical measures (histogram, box plot, and QQ-plot) and the Shapiro-Wilk test for normality. For normally distributed data, means with standard deviations were reported and for data not normally distributed, medians and interquartile

ranges were calculated for both recurrent and non-recurrent categories. Tumor characteristics included: cell radius, diameter of excised tumor, number of positive axillary lymph nodes, severity of concave portions of the contour, cell-nuclei perimeter, cell-nuclei area, concave points, compactness, smoothness, cell-nuclei texture, symmetry, and fractal dimension. For data normally distributed, the difference in means between the recurrent and non-recurrent tumor characteristics were calculated using a t-test and p-values reported. The difference in means for data not normally distributed was calculated using the Wilcoxon Rank Sum Test and p-values were reported.

Results

The Wisconsin Prognostic Breast Cancer (WPBC) dataset provides information on 500 breast cancer samples gathered by fine needle aspiration (FNA) and a score of different measurements/characteristics taken from histological preparations. For each tumor characteristic studied only diameter of excised tumor (in centimeters) resulted in a statistically significant p-value of 0.002 when comparing recurrent vs. non-recurrent breast carcinomas. All other tumor characteristics studied did not exhibit a statistical significance between recurrent and non-recurrent tumors as seen in Table 1: cell radius ($p = 0.5249$), number of positive axillary lymph nodes ($p = 0.077$), severity of concave portion of the contour ($p = 0.5158$), cell-nuclei perimeter ($p = 0.5249$), cell-nuclei area ($p = 0.5514$), concave points ($p = 0.3048$), compactness ($p = 0.6101$), cell-nuclei texture ($p = 0.6484$), symmetry ($p = 0.4727$), and fractal dimension ($p = 0.4564$).

Delving into the relationship between mean cell-radius and tumor size per category (less than or equal to 2 cm, between 2 and 3 cm, and larger than 3 cm), an ANOVA test was performed to compare the difference in the mean cell radius across the three categories. The resulting p-value from the ANOVA statistical test was $p = 0.0782$ which fails to reject the null hypothesis that there is no difference in mean cell-radius and tumor size categories; the differences are not significant. A Bartlett's test for equal variances resulted in $p = 0.982$ so an ANOVA statistical test was appropriate in this scenario.

```
. regress concavity tumlog
```

Source	SS	df	MS	Number of obs	=	120
				F(1, 118)	=	0.84
Model	.00374361	1	.00374361	Prob > F	=	0.3607
Residual	.524655555	118	.004446234	R-squared	=	0.0071
				Adj R-squared	=	-0.0013
Total	.528399165	119	.004440329	Root MSE	=	.06668

concavity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tumlog	-.0088283	.0096212	-0.92	0.361	-.0278808	.0102242
_cons	.1619472	.0097856	16.55	0.000	.1425691	.1813254

Table 2: Regression analysis of the association between cell concavity and tumor size.

Discussion

Utilizing machine learning to categorize tumor staging and predict potential recurrence of breast cancer has the appeal of providing clinicians with unbiased guidelines to increase favorable patient outcomes. Evidence in the literature has indicated tumor characteristics relating to changes to the cell nucleus could have indicative powers to predict recurrence of the tumor. Cell-nuclei texture has been reported to be a reliable prognostic indicator between benign and malignant lesions (95.8%) and grading of the tumor (93.3%); analysis did not report any statistically significant differences between non-recurrent and recurrent tumors or a relationship between tumor size and changes in cell-nuclei texture (Doyle, *et. al.,* 2008). While the cell-nuclei texture may be useful for initial diagnosis, it may not relay enough information regarding the probability of the tumor recurring. The severity and number of concave points have been used to classify malignancies in previous studies (Das, *et. al.,* 2020).