Final Exam

MBP 1201H - Biostatistics

Due November 14, 2023

Exam Description

This is a take home final that you will have 2 weeks to complete. This final is worth 50% of your mark.

Submission Format

This final assessment should be submitted online on Quercus by Tuesday November 14, 11:59 pm EDT. Your submission needs to include a single file:

1. STUDENT_NAME.STUDENT_NUMBER.ipynb containing the answers and code to the assignment questions, along with any figures and explanations that are asked for.

You are allowed to consult your lecture notes, and any static documentation on the internet. You are allowed to read content from Q&A websites such as StackOverflow, but you may not ask questions on such websites about questions on the final. You are not allowed to collaborate with other students for this assessment, but you may email the TA for clarifications.

Note: Unless specified otherwise, $\alpha = 0.05$ is the cutoff significance level for all questions in the final.

Questions

Multiple Choice Questions (5 marks each)

1. Given a measured value x, a p value represents:

   (a) The probability of measuring x or more extreme values given that the null hypothesis is correct.

   (b) The probability of measuring x or more extreme values given that the alternate hypothesis is correct.

(c) The probability that the null hypothesis is correct given a measurement of x or more extreme.

(d) The probability that the alternate hypothesis is correct given a measurement of x or more extreme.

2. A Pearson r result of 0.34 ($p < 0.05$) denotes:

   (a) a small, positive, nonsignificant linear correlation.

   (b) a small, positive, significant linear correlation.

   (c) a large, positive, significant linear correlation.

   (d) a small, negative, significant linear correlation.

3. What is the sequence of steps of training a deep learning model?

   I. Initialize weights of the model randomly.

   II. Go to the next batch of the dataset.

   III. If the prediction does not match the label, change the weights.

   IV. For a sample input, compute the output.

   (a) I, II, III, IV

   (b) IV, III, II, I

   (c) III, I, IV, II

   (d) I, IV, III, II

4. Which of the following pairings correctly denote a parametric statistical test and its corresponding nonparametric statistical test?

   I. Paired Samples t-Test $\leftrightarrow$ Wilcoxon Signed-Rank Test

   II. Pearson r Coefficient $\leftrightarrow$ Spearman r Coefficient

   III. One-Way ANOVA $\leftrightarrow$ Kruskal-Wallis Test

   IV. Unpaired t-Test $\leftrightarrow$ Mann-Whitney U-Test

   (a) I and III only

   (b) II and IV only

   (c) I, II, and IV

   (d) I, III, and IV

   (e) All of the above

5. Sam attended a beach party in Florida in January 2021 and thinks he has a 50% chance of catching COVID. He does not have symptoms, but takes a rapid antigen test, which comes up positive. To confirm this result, he takes another rapid antigen test, which comes back negative, and goes about his life as normal. What is the probability (to the nearest 0.01) that Sam has COVID, assuming a sensitivity of 0.58 and specificity of 0.99 for rapid antigen tests?

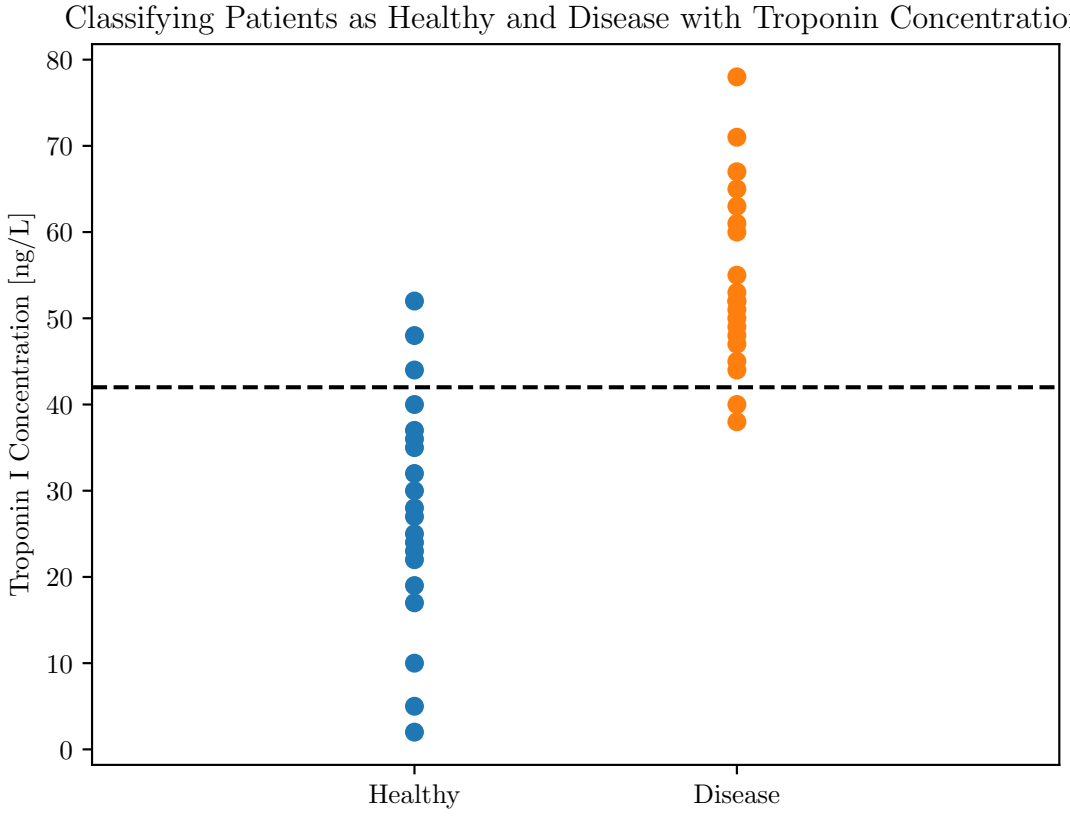
   (a) 50%

   (b) 57%

   (c) 96%

   (d) 98%

Short-Answer Questions (5 marks each)



6. Doctors in Chicago studied troponin I concentration results from 20 healthy patients and 20 heart attack patients (see above figure). A model was trained to determine a

cutoff value of the troponin concentration so that patients with a lower concentration are predicted to be healthy, and patients with a higher concentration are predicted to have had a heart attack. The optimal cutoff value was found to be 42 ng/L, shown as a horizontal line in the figure. *Note*: Each dot on the plot corresponds to one and only one patient.

(a) Define the sensitivity and specificity of this test in terms of $N_{D,+}; N_{D,-}; N_{H,+}; N_{H,-}$. Respectively, these are the number of (diseased/healthy) patients that recieve a (positive/negative) test.

(b) Calculate the sensitivity and specificity of this test.

(c) An alternative statistic for determining the effectiveness of a test are the positive and negative predictive value, defined as:

$$\text{PPV} = \frac{N_{D,+}}{N_{D,+} + N_{H,+}}; \quad \text{NPV} = \frac{N_{H,-}}{N_{D,-} + N_{H,-}}$$

Calculate the positive predictive value and negative predictive value of this test.

(d) What is the difference in interpretation between the sensitivity and positive predictive value? Specificity and negative predictive value?

7. A graduate student Sarah is conducting an experiment to see if Drug X is more effective than Drug Y in treating cancer in a mouse model. Before conducting the experiment, she set a significance level of $\alpha = 0.01$ for statistical significance. After conducting the experiment, she conducted a one-tailed t-test and determined a significance level of $p = 0.02$. This is still considered significant by the widely-accepted threshold of $\alpha = 0.05$ in her research field, so she considered the result as statistically significant. Do you agree with Sarah's reasoning? Why or why not?

8. Overfitting occurs when a deep learning model is able to model our training set well but achieves poor performance on the test set. How would using over-fitted models affect clinical deployment and patient care? Should clinicians base their trust of a deep learning model off of the training set performance alone?

9. What are the four assumptions needed for a (multivariate) linear regression test? Define each term and briefly explain why it is important.

10. Suppose that $X \sim N(\mu, \sigma^2)$ is a random variable. Answer the following questions using the "useful properties of the normal distribution" and "Propagation of errors" slides in the Lecture 1 slides.

(a) Show that $Z = (X - \mu)/\sigma \sim N(0, 1)$.

(b) Using the result from 10a above, show that $XZ \sim N(0, \mu^2)$.

Data Analysis in Python (10 marks each)

11. **Parametric and Nonparametric Statistics**: The template file contains code to generate a Gamma distribution X and a Poisson distribution Y, with approximately the same mean.

    (a) Plot histograms of X and Y on different figures. Upon visual inspection, do they look normally distributed?

    (b) Plot the ECDF response function of X and the ECDF of Y on different figures. Do these agree visually with the theoretical CDF response functions from a normal distribution with the same mean and standard deviation as X and Y?

    (c) Use one unpaired parametric statistical test and one unpaired nonparametric statistical test to evaluate whether or not X is statistically distinct from Y at the level $\alpha = 0.05$. Which of the test(s) are able to distinguish X from Y?

12. **Monte Carlo Approximation**: The goal of this problem is to use stochastic (computational) methods to approximate a well-known constant using probabilistic methods.

    (a) Use `np.random.uniform()` to select 10,000 random x values between 0 and 1. Similarly, randomly select 10,000 separate y values between 0 and 1. Is there positive, negative, or no apparent skewness in this data?

    (b) Make a scatter plot of the (x, y) data, labelling points that satisfy $x^2 + y^2 \leq 1$ as blue and all other points as red.

    (c) Let p denote the number of blue points divided by the total number of points. Calculate p and $4 \cdot p$ from your data.

    (d) What does $4 \cdot p$ approximate? How can you interpret p as a probability?

13. **Regression in Cardiovascular Disease**: For this question and the next, we'll use the `sklearn` diabetes dataset. Sample code has been provided in the template file to load in this data.

    (a) Using `sns.pairplot()`, visually determine if there are any correlations between the `age`, `bmi`, and `bp` variables.

    (b) Fit a linear regression model with `age`, `bmi`, and all `s1`, ..., `s6` keys as independent variables and `bp` as the dependent variable. What is the R^2 value for this model?

    (c) Use a barplot to plot the coefficient weights for your linear regression model in (b). How do you interpret these coefficient weights? Which features are the most predictive in your model? Can you interpret any causality from this regression analysis?

14. **Principal Component Analysis**: Using the `sklearn.decomposition.PCA()` function from the scikit-learn package, perform a dimensionality reduction of the 10 features in the `sklearn` diabetes dataset. For help in implementing PCA and subsequent

analysis, feel free to follow the exposition here: https://saskeli.github.io/data-analysis-with-python-summer-2019/pca.html

(a) Perform PCA with 10 components, and make a scatter plot of the first two components. Is the joint distribution of the first two components normally distributed? Are the marginal distributions normally distributed? Justify your conclusions with appropriate normality tests.

(b) Plot the explained variance as a function of the component number. On a separate axis, plot the sum $S_k = \sum_{i=1}^{k} s_i$, where s_i is the explained variance of component i, as a function of component number k. How many components are needed to capture 90% of the variance?

15. **Edge Detection using Convolution**: Convolutions are one of the main operations in deep learning models. To illustrate the effects of a convolution, we will convolve a test image with different types of Sobel (`scipy.ndimage.sobel()`) filters to detect edges. To read in the test image, use the code: `img = scipy.misc.ascent()`

(a) Plot the test image using `plt.imshow()`. Plot a histogram of the (flattened) pixel values of `img`. What is the 95% confidence interval of the pixel values of `img`?

(b) Apply the sobel filter along the x axis `axis=1`, and plot the resulting image D_x.

(c) Apply the sobel filter along the y axis `axis=0`, and plot the resulting image D_y.

(d) Plot the image $D_{xy} = \sqrt{D_x^2 + D_y^2}$. What are the similarities and differences between D_x, D_y, D_{xy}?

(e) Is the histogram of D_{xy} similar to the histogram of `img`? Why or why not?