IMPERIAL

Probability and Statistics

Coursework 2024

Words and Figures

CID Number: 06005311

Name: Erik Garcia Oyono

WE ARE
THE DEPARTMENT OF
BIOENGINEERING

Question 6

In the numerical questions, you selected statistical tests. For each, explain why you selected the one you did and what the null hypothesis is in words

a) For Question 1c (< 60 words).

Since data are sampled from normal distributions, have small sample size and unknown population standard deviation, t-test was chosen. It is one-sample because we are comparing the mean OD with arginine, and upper-tailed because we only care about an increase. The null hypothesis states that arginine does not increase the mean OD from the baseline value of 0.35.

b) For Question 2b (< 60 words).

Welch's t-test, a two-sample unpaired t-test with unequal variances was selected because we assume data are normally distributed, come from two independent groups, and want to compare their means to determine whether one is better. The null hypothesis is that the mean OD is not statistically different using urea vs arginine.

c) For Question 3e (< 60 words).

The rank-sum test was chosen to compare the distributions of urea and arginine because it is a non-parametric test used for independent data that are not normally distributed. This test does not assume normality, making it appropriate for this data. The null hypothesis states there is no statistical difference between the distributions of OD values for urea and arginine.

d) For Question 4b and 4c (< 60 words).

Since we are dealing with paired samples from normally distributed data with unknown population standard deviations, paired t-test is appropriate as it accounts for the dependency between two samples. The null hypothesis is that there is no statistical difference in the mean OD between day 1 and day 2 for the control and test groups.

Question 7

a) In Question 2, you carried out a null hypothesis test, from which you may or may not have rejected the null hypothesis. Irrespective of your personal result, what could you have concluded from the data if the null hypothesis was not rejected. Be as precise and comprehensive as possible (< 60 words)?

There is insufficient evidence to conclude that one nitrogen source is better than the other. This does not prove they are equivalent, only that the data do not provide strong enough evidence to distinguish their effects. Any observed difference in sample means could be due to random sampling variability rather than a true difference in population means.

b) Reflect on your answers to 4b. How do they compare to what you expected? (< 100 words).

Literature suggests that adding arginine on day 2 has no effect. With a significance level of 5%, we would expect around 5% of p-values to be below 0.05 due to random chance, even if there is no actual difference. Since 4.6% were rejected, the result is slightly below the expected and within the range of random variation, so we cannot conclude that adding arginine on day 2 has an effect. This aligns well with expectations under the null hypothesis, suggesting no strong evidence of a statistical difference between days 1 and 2 in the control group.

c) Reflect on your answers to 4c and 4d. How do they compare to what you expected? (< 100 words).

From literature, we expect a rejection of the null hypothesis due to statistical difference between the test samples on days 1 and 2. The statistical power of 78.4% suggests the test has the potential to detect a meaningful difference, but the higher-than-expected percentage of non-rejections (32.3%) suggests no evidence for a statistical difference between days 1 and 2, which does not align with the expectations. This suggests weak evidence for an increase in bacterial growth on day 2 due to a smaller effect size from urea or greater variability in the experimental data.

d) Reflect on your answers to Question 5. What do the statistics you calculated tell you about the real-world application? (< 100 words).

The interval [0.3167, 1.0382] indicates the range of expected OD values at the antibiotic concentration of 2 µg/ml, with the lower end representing higher effectiveness of the antibiotic. The variability in this range suggests the antibiotic's effect is not uniform across all conditions and shows that predicting the precise effect of an antibiotic concentration is challenging. Hence, this concentration may work well under certain conditions but be less effective under others, which should be considered when determining drug dosage. Further research could help determine a narrower interval to maximise efficacy for targeted treatments.

Figure 1

Plot the data for Question 2 in a way that you think most clearly communicates the data and the statistical analysis you carried out. See coursework brief for marking criteria for plots and captions

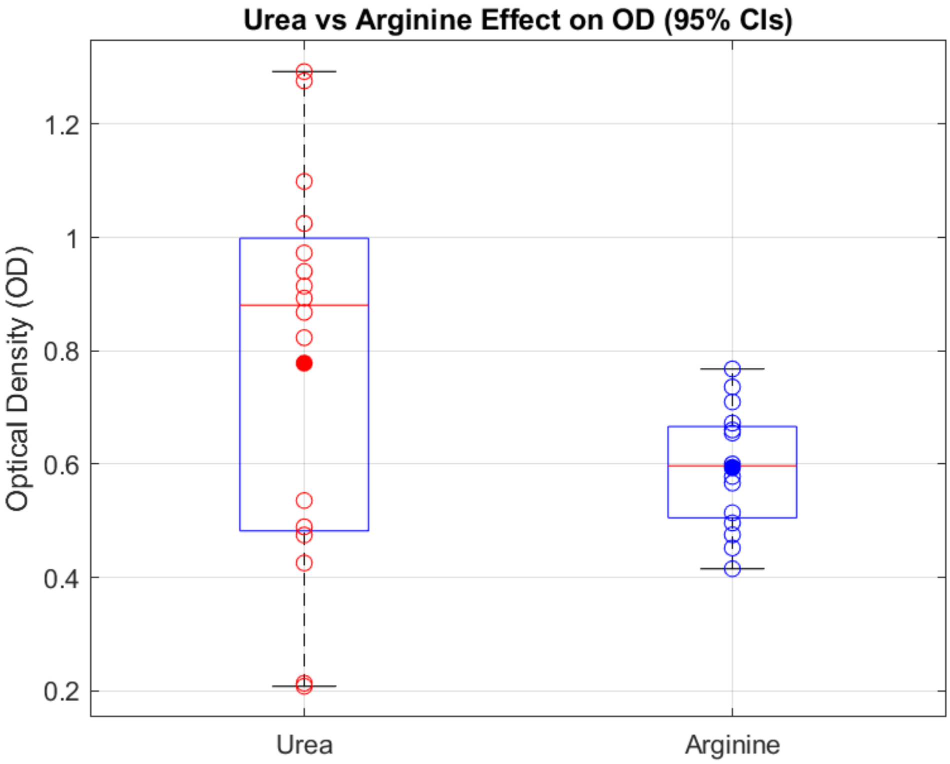

Figure 1: Boxplot of the OD values for urea (red) and arginine (blue) at a 95% confidence interval. The filled dots represent the mean for each group, the horizontal lines within each box represent the median for each group, and the the T-shaped whiskers represent their maximum and minimum values.

Figure 2

Plot the p-values from Questions 4b and 4c on a single graph. Use a log-scale for the p-values. Add annotations, colours etc. to tell the story of your data.

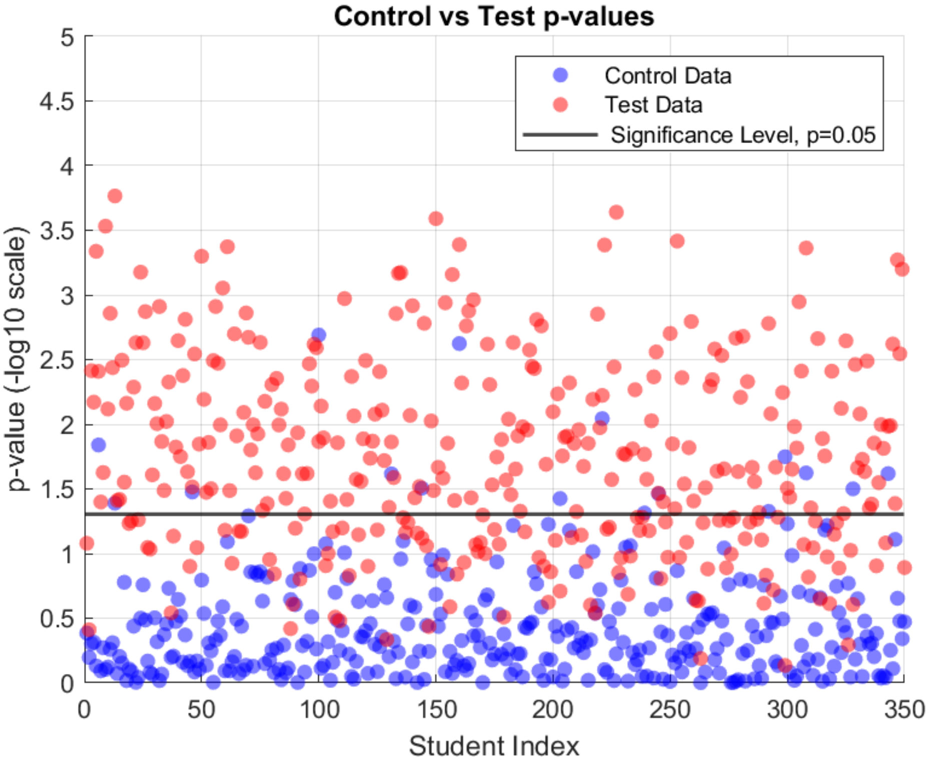

Figure 2: Comparison of *p*-values for control (blue) and test (red) data sets. The *p*-values are plotted on a -log10 scale to illustrate the distribution of significant values across 350 students.

Figure 3

Plot your data for Question 5 and visualise the outcomes of your regression analysis. You may wish to include visualisation of the values calculated in part d.

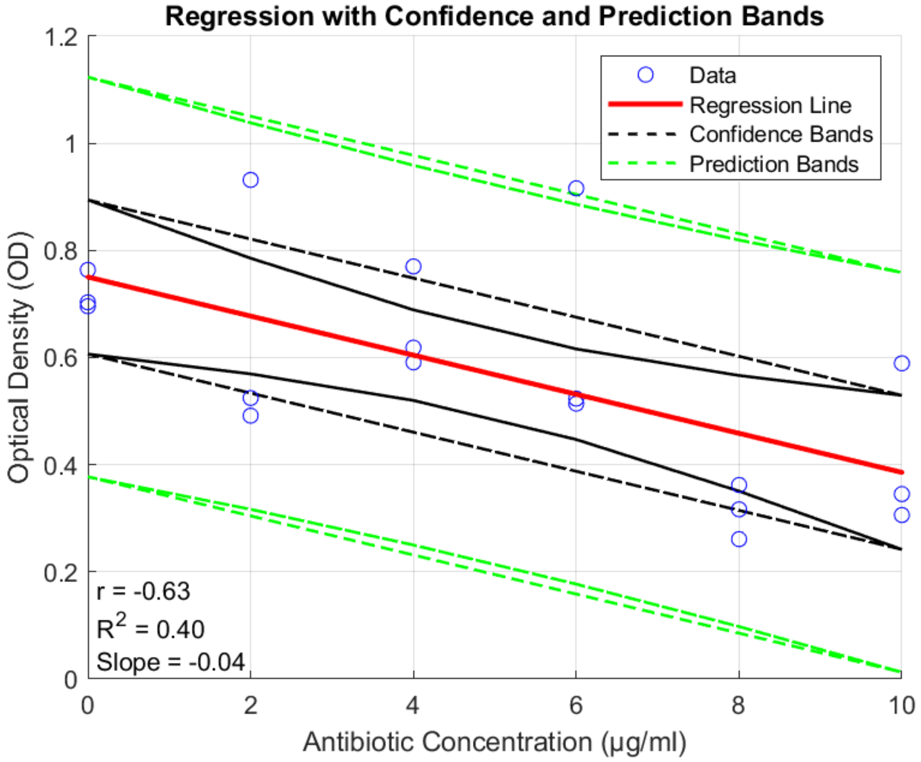

Figure 3: Regression plot for the relationship between OD and antibiotic concentration. The regression line (red) represents the best-fit model, with a slope of -0.037 (≈ -0.04), correlation coefficient of -0.631 (≈ -0.63) and coefficient of determination of 0.399 (≈0.40). The dotted black lines denote the 95% confidence interval, showing uncertainty in the regression line. The dotted green lines represent the 95% prediction interval, indicating expected variability in future observations.