Assignment (Mode 2): Probability and Statistics

Course: CSEG 2036P | *School of Computing Sciences, UPES*
Faculty: Dr. Mrittunjoy Guha Majumdar

Hypotheses are essentially assertions or statements regarding an occurrence (of an event) that we can investigate. Originally, the term *hypothesis* in ancient times described a synopsis of the storyline in a classical play. Fascinatingly, during an early medieval clash between matters of spirituality and science, Cardinal Bellarmine provided a renowned instance of this usage. In the early 17^{th} century, he cautioned Galileo not to consider the motion of the Earth as an established truth but rather as a hypothesis. To qualify as a scientific hypothesis, it must be testable according to the scientific method. Scientific hypotheses are typically formulated based on prior observations that existing scientific theories fail to adequately explain. Statistics provide a method through which we can either support or refute these claims. Hypothesis testing allows us to formulate these issues in a way that statistical data can be used to evaluate and determine the validity of these claims.

What is the Null hypothesis (H_0) and Alternative hypothesis (H_1)? Give an example.

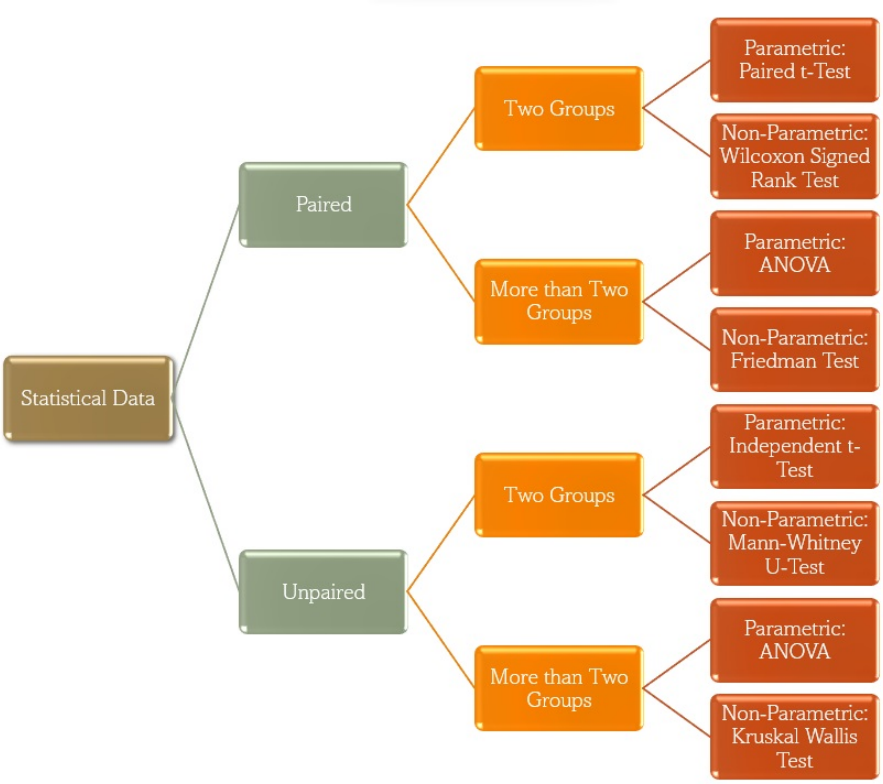

Figure 1: Flow-Chart for Selection of Statistical Tests

1. t-Tests

(a) What is the Student's t-Test? What are the two kinds of t-Tests?

(b) Implement a programming code, using C/C++, Mathematica, MATLAB or a programming language that you are familiar with, to generate two samples of 75 Gaussian random numbers with the same variance of $6.5 + \frac{k}{10}$ and differing means of $50 + k$ and $51 + 2k$ (where k is the last digit of your SAP ID) respectively, and use the t-test to accept or reject the null hypothesis around similarity of the samples. Begin your answer with highlighting the algorithm and/or logical flow-chart.

(c) What happens if we assume the data-sets generated above are paired?

(d) What is the Welch's t-Test? How is it different from the student's t-Test?

(e) Implement a code on the data generated in part (b) with variances $6.5 + \frac{k}{10}$ and $5.7 + \frac{k}{15}$ for the two samples, keeping all other parameters to be the same, and use the Welch's t-Test to check for the null-hypothesis on similarity.

2. Wilcoxon Signed Rank Test

(a) What is the Wilcoxon Signed Rank Test? In which situation is the Wilcoxon Signed Rank Test used instead of the t-Tests? What is a one-sample test and a paired data test, in the context of the Wilcoxon Signed Rank test?

(b) Implement a programming code, using C/C++, Mathematica, MATLAB or a programming language that you are familiar with, for realization of the Wilcoxon Signed Rank Test on the data given in Table 1:

i	1	2	3	4	5	6	7	8	9	10
$x_{2,i}$	$125 + k$	115	$130 + k$	140	$140 + k$	115	$140 + k$	125	$140 + 2k$	135
$x_{1,i}$	$110 + 2k$	122	$125 + 2k$	120	$140 + 2k$	124	$123 + 2k$	137	$135 + 2k$	145

Table 1: Table for Wilcoxon Signed Rank Test

where k is the last digit of your SAP ID.

Given: Significance level to be considered is $\alpha = 0.05$.

3. Mann and Whitney's U-test

(a) What is the Mann and Whitney's U-test? What are the assumptions for the applicability of this test?

(b) Suppose we have a Unit Test conducted for the course CSEG 2036P in Batch 1 (B_1) and

Batch 2 (B_2), and the results are as given in Table 2.

B_1	9	5	$2+\|k-5\|$	4	7	6	$1+2\times\|k-5\|$	8	9	3
B_2	7	6	$3+\|k-4\|$	8	2	5	$1+\|k-6\|$	4	7	10

Table 2: Table for Mann and Whitney's U-test

($|a|$ is the absolute value of a and k is the last digit of your SAP ID.)

Implement a programming code, using C/C++, Mathematica, MATLAB or a programming language that you are familiar with, for realization of the Mann and Whitney's U-test for the given data.

(c) What are the advantages of the Mann and Whitney's U-test?

4. ANOVA

(a) What is ANOVA? What is a one-way ANOVA and a two-way ANOVA?

(b) What is the Total Sum of Squares (SSTO), Error Sum of Squares (SSE) and Regression Sum of Squares (SSR), in the context of ANOVA? What is the relation between SSTO, SSE and SSR?

(c) What is Regression Mean Square (MSR) and Error Mean Square (MSE), in the context of ANOVA?

(d) Define what is the ANOVA coefficient.

(e) Medical researchers aim to determine if there are variations in the mean blood pressure reductions caused by two different medications (M_1 and M_2). To investigate this, they assign 5 patients randomly to each medication for a month and measure the blood pressure (systolic) before and after usage, as given in Table 3:

Before M_1	After M_1	Before M_2	After M_2
142	$125+\|k-5\|$	137	$121+\|k-5\|$
$145+\|k-4\|$	118	$140+\|k-4\|$	120
138	$121+\|k-3\|$	146	$117+\|k-3\|$
$149+\|k-5\|$	123	$142+\|k-5\|$	125
144	$120+\|k-4\|$	139	$123+\|k-4\|$

Table 3: Medical Data for ANOVA

where $|a|$ is the absolute value of a and k is the last digit of your SAP ID.

Implement a programming code, using C/C++, Mathematica, MATLAB or a programming

language that you are familiar with, to use this data in a one-way ANOVA, considering the type of medication as the variable and blood pressure reduction as the result, with significance level $\alpha = 0.05$.

5. Kruskal Wallis Test

(a) What is the Kruskal Wallis Test? Is it a parametric or non-parametric test? What are the elements and assumptions for a Kruskal Wallis Test?

(b) A researcher aims to investigate the potential differences in the effects of three drugs on nociceptive pain. To conduct this study, the researcher enrolls 5 individuals experiencing similar levels of nociceptive pain. These participants are randomly assigned to three groups, each group receiving one of the three drugs: Drug 1 (D_1), Drug 2 (D_2), or Drug 3 (D_3). After a month of taking the assigned drug, the individuals are asked to self-assess and rate their nociceptive pain on a scale from 1 to 100, where 100 indicates the most severe pain they experience, as given in Table 4.

D_1	72	$20 + 2 \times k$	63	$24 + 4 \times k$	66
D_2	88	$31 + 7 \times k$	$18 + 5 \times k$	70	$75 + 2 \times k$
D_3	53	$41 + 2 \times k$	91	$60 + 2 \times k$	44

Table 4: Table for Kruskal Wallis Test

where k is the last digit of your SAP ID.

Implement a programming code, using C/C++, Mathematica, MATLAB or a programming language that you are familiar with, to analyse this data using the Kruskal Wallis Test.

6. Chi-Squared Distribution

(a) What is the χ^2 distribution? Give three properties and two uses of the distribution. Give any three conditions for the validity of the χ^2 test.

(b) What is the relationship of the χ^2 distribution with the standard normal distribution?

(c) A study suggests that a particular species of insects occurs in the ratio of 5:3:2:1 in different habitats - A, B, C, and D. In an investigation, 2000 insects were sampled and categorized into these habitats resulting in counts of $860 + 2k$, $510 - 3k$, 430, and $200 + 4k$ for habitats A, B, C, and D, respectively (where k is the last digit of the SAP ID). Implement a programming code, using C/C++, Mathematica, MATLAB or a programming language that you are familiar with, to analyse the observed data. Given the critical value for the chi-squared test at a significance level of 0.05 with 3 degrees of freedom as $\chi^2_{0.05}(\nu = 3) = 7.82$. Use the code for the chi-squared test to determine if there's a significant difference between the observed and expected distributions.

7. Non-Linear Regression and Hypothesis Testing

(a) What is meant by regression analysis and non-linear regression? Give an example.

(b) How is non-linear regression useful for hypothesis testing?

(c) Implement a programming code, using a programme language that you are familiar with, to fit a least-square curve of the form $y = \frac{a}{x(x+b)}$ (a and b being constants) to the data given in Table 5:

x	$3.3 + 0.5 \times k$	4.5	$6.1 + \frac{k}{3}$	$7.6 + \frac{k^2}{10}$	8.2	$9.3 - \frac{k}{7}$	$10.2 + 0.5 \times k$
y	$0.92 + 0.07 \times k$	0.47	$0.12 + \frac{k}{30}$	$0.11 + \frac{k^2}{50}$	0.03	$\lvert 0.06 - \frac{k}{90} \rvert$	$0.09 + \frac{k}{75}$

Table 5: First Data-Set for Non-Linear Regression

where k is the last digit of your SAP ID.

(d) Implement a programming code, using a programme language that you are familiar with, to fit a least-square curve of the form $y = ae^{bx}$ (a and b being constants) to the data given in Table 6:

x	2	$4 + \left[\frac{\lvert k-5 \rvert}{10}\right]$	7	$10 - 2 \times \left[\frac{\lvert k-5 \rvert}{15}\right]$	13
y	$11 + \left[\frac{k}{10}\right]$	14	$10 + \left[\frac{k}{7}\right]$	16	$23 + \left[\frac{k}{13}\right]$

Table 6: Second Data-Set for Non-Linear Regression

where k is the last digit of your SAP ID, $|a|$ is the absolute value of a and $[b]$ gives the nearest integer to b.

8. Correlations

(a) What is the Karl Pearson Coefficient Correlation and Spearman's Rank Order Correlation? Give simple examples to illustrate your understanding of these concepts.

(b) What are the differences between Karl Pearson Coefficient Correlation and Spearman's Rank Order Correlation?

(c) Write a programming code to calculate Karl Pearson's coefficient correlation between two sets of data for exam scores in Mathematics and Science for 10 students.

Data provided:

Mathematics scores: $80 + k$, 85, 90, 70, $60 + 3k$, 75, $95 - 2k$, $78 + k$, 82, 88
Science scores: 85, $88 - 3k$, 92, $72 + 2k$, 65, 80, $96 - k$, 79, 85, $90 + k$]

where k is the last digit of your SAP ID.

The program should use a programming language of your choice (Python, MATLAB, C++, etc.) to calculate the Karl Pearson's coefficient correlation between the two sets of data.

(d) Write a programming code to calculate the Spearman's rank order correlation for two sets of data giving the ranking of students' scores in two subjects - English and History.

Data provided:

English scores rank: 3, 1, 2, 4, 5
History scores rank: 2, 1, 4, 3, 5

The program should utilize a programming language of your choice to compute the Spearman's rank order correlation between the rankings in the two sets of data.

Note: Writing steps in the code from scratch (first principles) rather than using inbuilt functions and modules will be given higher marks, for all questions.