Statistical Inference

Lecturer: Abdol-Hossein Vahabie

Spring Semester 1401-1402

Writing Assignment II | Deadline 1402/01/28

1 True or False?

Explain your reasoning briefly.

1. In a normal distribution, Q1 and Q3 are less than one standard deviation away from the mean. (3pts)
2. The hair colors found in a random sample of 10 students is a binomial random variable. (3pts)
3. The mean of a Poisson distribution is not always equal to its variance. (3pts)
4. If the binomial probability of success is near 0.5, then the distribution is approximately symmetric. (3pts)
5. The probability of a student randomly guessing answers to a true/false exam is best modeled with a binomial distribution. (3pts)

2 Bayesian Inference

A new language model algorithm is being tested to see if it performs better than the best existing model in the task of Hate Speech Detection. The existing model can correctly predict the answer on 50 percent of the test set, while it is believed initially that the new model has a 2/3 chance of predicting the hate correctly on 60 percent of the test set and a 1/3 chance of predicting the hate correctly on 50 percent of the test set. In a pilot study done with 20 randomly selected sentences, the new model correctly predicts the correct label for 15 sentences. What is the probability that the new model is better than the best existing model? (13pts)

3 Binomial Distribution

In the sentiment analysis task of natural language processing, we aim to determine the sentiment or emotional tone of a piece of text, such as a tweet or a review, using computational methods. Let's consider the following scenario:

"A company has launched a new product and wants to understand the sentiment of its customers towards the product. They collected a sample of 75 reviews from various sources and asked human annotators to label each review as positive or negative. The label of each review is considered as the ground truth sentiment. X is a binary feature that models whether a review is positive or negative, with a probability of 0.8 for a positive review."

1. Find the expectation of the number of positive reviews in the sample. (7pts)

2. Explain why the number of negative tweets can be modeled as a binomial random variable Y with parameters (75,0.2). (7pts)

4 Exponential Distribution

Suppose a customer asks your conversational AI system: "How long does it usually take for my package to arrive?" Your system should be able to provide an estimate based on historical data and the distribution of delivery times. Specifically, if the delivery times follow an exponential distribution with a mean of 3 days.

1. What is the probability that a package will arrive in less than 2 days? (7pts)

2. Similarly, if the delivery times follow an exponential distribution with a mean of 5 days, what is the probability that a package will take more than 7 days to arrive? (7pts)

5 Poisson Approximation

An OCR(Optical Character Recognition) which is a technology that recognizes text within a digital image, generate texts with exactly 100 words per image. If a Spell checker such as Grammarly says that 0.5 percent of the words have spell errors, then what percent of the words will:

1. Have no spell error? (7pts)

2. Have 2 or more spell errors? (7pts)

6 R Programming

Given the input sentence use ggplot to calculate the frequency distribution of unigrams and create different visualizations of this data. The input sentence is:

"The power of words cannot be underestimated. Language is a critical component of human communication, and natural language processing (NLP) is a field that seeks to understand and enhance our ability to communicate through language. NLP involves developing algorithms and computational models that can analyze, interpret, and generate human language. One of the most common tasks in NLP is text classification. In this task, an algorithm is trained to automatically assign predefined categories or labels to a given text. For example, a text classification algorithm may be trained to identify whether an email is spam or not, based on the content of the email. Another important task in NLP is named entity recognition (NER). NER involves identifying and classifying named entities in a text, such as people, organizations, and locations. This task is useful in applications such as information extraction and text-to-speech synthesis. Other tasks in NLP include sentiment analysis, machine translation, and text summarization. Sentiment analysis involves determining the sentiment or emotion expressed in a given text, such as positive or negative. Machine translation involves automatically translating text from one language to another, while text summarization involves generating a brief summary of a longer text. In order to perform these tasks, NLP algorithms typically rely on statistical models and machine learning techniques. These techniques involve training the algorithm on large amounts of data, so that it can learn to recognize patterns and make accurate predictions."

1. Draw a horizontal bar chart of the 10 most frequent unigrams. Don't forget to label x and y axis correctly. (15pts)

2. Draw boxplot based on the character length of each word in each sentence. (You must at first split sentences based on dot and then split the sentence into words and then make an array for each sentence that the arrays must have the length of each word in order) (15pts)

❖ Please note that ChatGPT can be used implicitly for coding the tasks that are outside our area of concentration, but please refrain from using it for other purposes like creating plots!
For example in first question, you can ask it like this: "Help me in finding frequency of unigrams in a paragraph using R language."

Required Document

No additional document is required for this homework!

General Rules

Please upload a file in ZIP format (~~not RAR~~) to the elearn platform.
You are allowed a total grace period of 3 days to submit late assignments for all of your exercises.
It is prohibited to use handwritten material and only material produced through typing in the HW template is permissible.
Utilizing a LaTeXto compose the report will grant an additional 5 points.
Feel free to use any R packages, just ensure that each plot has a title and appropriate legends, and don't forget to label the axes.

Deadline

Monday 23:59. 1402/01/28.

Contact Information

Please direct your questions regarding Homework 2 only to the teaching assistants, Romina Oji and Mahyar Maleki, through the course mail (statistical.inference.ut@gmail.com). Use "HW2" as the subject line.

Good Luck