Observation for tutorial 1

The tutorial code shows how to execute an ANOVA test to compare average of three groups of yield data sets. It starts with the computation of the grand mean and the means of the three groups, determination of the yields of three groups, and graphic data illustration by scatter plots. Vertical lines represent the grand mean and group means. The code then uses formula to calculate the within-group and between group sum of squares to measure the variance, then calculate the degree of freedom and mean square of both. The F-ratio is computed as the ratio of between-group mean square to the within-group mean square, while an F-distribution function calculates probability distribution and p-value. It affords the researcher the opportunity to test hypothesis that seeks to draw relationship between various group means in a bid to find out if such differences are statistically significant.

Observation for Task 1

While applying the code, I wanted to compare yield data gained using various irrigation techniques and estimate the statistical measures tied with these yields. The first process concerned the computation of the means and standard deviations of three groups of yield data. Indeed, I observed that utilizing the mean() and std() functions of NumPy made the computations of these statistics much easier. I was able to store the standard deviations as a list and this make it easy for me to know the ratio of the largest standard deviation to the smallest one. This ratio offered an understanding concerning the fluctuations of the yields across the various datasets so as to reveal which irrigation technique delivered the most stable yield.

As I continued to work through the code, I calculated the means and standard deviations for three additional sets of yield data derived from different irrigation techniques: drip, sprinkler, and flood. Overall, I observed that by averaging these yields as well as standard deviation, it is easy to compare the methods. Here, the patterns of four yields were identified showing on average what method fared best and how much deviation was possible for the every method yields. In conclusion, the evaluation stressed the significance of statistics as a means of interpreting agricultural information, and contributed to the assessment of irrigation techniques in improving crop yield.

Obasrvation for case study task 2

In this task, I compiled a code snippet whose purpose was to determine how various types of fertilizer impacted crop yield using statistical tools. First, I created a 2D NumPy array to store the yield information regarding three types of fertilizer- A, B, C. To compare the efficacy of all the fertilizers applied, I computed the means for each type of a fertilizer as well as the grand mean. I also made a scatter plot to depict the yields where separate horizontal lines represent the grand mean and the group means. This was helpful in terms of bringing about a visuals of the differentiation in yields when comparing between the fertilizers.

Then I proceeded to calculate the ANOVA of the result by determining the sum of squares, the degrees of freedom and the mean square both within and between groups. This helped to compute the F-ratio which compares the ratio of variance between groups to the variance within groups. Furthermore, I used the F-distribution function to calculate the p value which enabled me ascertain if my results were statistically significant. The authors also emphasized providing the F-ratio and p-value as they not only indicated better or worse performance of various fertilizers compared to each other, but also

emphasized the role of statistical analysis in decision-making in the field of agriculture. Altogether, I benefited from this task to improve my overall knowledge of statistical methods that can be used in practical scenarios and how to use statistical tools for actual events such as crop yield rate.generating data on yield for the three types of fertilizers, A, B, and C. Finally after giving the nature yield data a 2D array form, I was able to compute group means and overall grand mean. I also plotted the yields with a scatter plot and then indicated the mean yield for each type of fertilizer with horizontal dashed line. The F-distribution function was used to determine the variance among groups which was also implemented. Using the information I arrived at the degrees of freedom and thereafter computed the mean square in order to arrive at the F-ratio. The results also held an approximate F-ratio of 15.87 for fertilizer yields, highlighting the existence of differences. Also, the obtained p-value equal to 0.0001 demonstrated high sign for rejecting the null hypothesis whereby, at least one type of the fertilizer brought a significantly different yield as compared to the other types. All in all, it was seen that fertilizer selection plays a great role on food production and this proves that right and appropriate fertilizer play a dramatic role in agricultural productivity.

Explain the limitations and assumptions of one-way analysis of variance(anova).

Assumptions of One-Way ANOVA

1. Independence: The samples from different groups have to be disparate. This means that the type of subject for each result should not matter and the selection of subjects of one group should be independent of the selection of subjects of another group.

2. Normality: It is also necessary that the distribution of the data in each group should be approximately normal. Nonetheless, although ANOVA is relatively insensitive to violation of homogeneity of variance– the violations that can be critical include the normality assumption, especially in samples with a small size.

3. Homogeneity of Variance: In other words, the amount of dispersion of the dependent variable (squared residuals) must be similar in all the groups for modeling the data (homoscedasticity). This behaviour can be tested using the Levene test or the Bartlett test. The differences can be unequal which increase the probability of committing a Type I error.

4.Scale of Measurement: Focused on the dependent variable the type of data should be interval/ratio since it measures the degree of the relationship.

Limitations of One-Way ANOVA

1. Only One Factor: With one-way ANOVA, the researcher can only examine one independent variable at a time. In case there are several variables that may influence the dependent variable, one might use the methods like factorial ANOVA or others.

2. Does Not Indicate Which Groups Differ: Although use of ANOVA shows the presence of significant differences among the group means, it does not tell any of the specific group differences. The Type I error rate is always affected by the number of comparisons made and when an ANOVA is significant, post hoc tests such as Tukey's test or Bonferroni correction is used to do multiple pairwise comparisons.

3. Sensitivity to Outliers: Deviations and variations of at least several of the subjects may be summarized by outlier cases, thereby misguiding the results of ANOVA concerning the mean and variance of the groups.

4. Sample Size: In general, ANOVA assumes that the samples are large enough to produce accurate results. Smaller sample sizes pose the danger of failing to have adequate statistical power to establish differences in the means of groups.

5. Assumption Violations: However, disadvantage of using ANOVA is that, when assumptions of normality or homogeneity of variance are not met it produces inaccurate results. In such instances other non-parametric methods are recommended, for instance the Kruskal-Wallis test.

6. Linear Relationship: The limitations of ANOVA also embrace the method of analysis whereby the model assumes a linear relationship between the independent variable and dependent variable. This technique cannot be used to model non-linear relationships in a data set if these are present.

Analysis of second tutorial

The above tutorial provides an extensive example explaining how to conduct ANOVA on given data set of students, with emphasis on the graduates' salaries grouped by their field of study. First, the code loads a dataset and then selects only students who have graduated. The graduated students who sat for the test are then randomly selected in equal groups of fives; five different schools are used and the total sample is split in half a total of five hundred graduated students are used to draw a manageable sample. The code provides probability plots as a way of checking the normality of salaries in each major that is crucial for the ANOVA assumptions. Following that, it builds an ANOVA backbone table where it estimates the sum of square for between-groups (SSTR) as well as within-group (SSE) changes, and registers updated value of degrees of freedom for each category. This structure ensures that one can make a statistical inference as to whether the mean salaries of workers in different majors differ and hence assess the possibility of a salary discrimination within different fields of specialization.

Observation for task 4

In this task, I performed an analysis of variance (ANOVA) on a dataset of graduated students by using Pandas and NumPy. My first step was to input the data and use Sedan to narrow down the data to only the graduating students. Firstly, I excluded a section of students at random to a maximum of 500 different students so as to have accessible data but not compromising on variation of the data set. From this sample, I focused on two variables: I combined 'major' and 'salary,' and excluded records that have numerically nonsensical salary as I carried out the subsequent computations.

I also put together an ANOVA (analysis of variance) skeletal table to show the calculations that would be required for this context. To compare the variability between different majors, I calculated the Sum of Squares for Treatment (SSTR), and to compare the variability within each major, I used the Sum of Squares for Error (SSE). From these, I got the Total Sum of Squares (SST) and worked out between and within groups' degrees of freedom and mean squares (MS). I calculated the F-ratio and its corresponding P-value to find out if there exists a significant difference in total salary between the majors. Lastly, I computed the F-critical value at the 0.05 levels of significance to finish the analysis of variance. I wrapped the final results neatly in a clean cut table to facilitate the understanding of the statistical results presented.