Which Test?

Mini Workshop

Statistical Consulting for Data Scientists

January 26, 2025

Introduction

This is an individual assignment.

Make sure your name appears somewhere on your submission.

Just try these on your own (don't look it up, try to answer it first, then feel free to look up the answers). You are being asked which statistical test would you use to solve the problem presented. Some of these questions have more than one correct answer. I do not need to know your answer (you are not required to turn in your actual answers), I just want to know your overall score (e.g., 8/10) and a brief self-assessment (e.g., I did just fine maybe I need to review a little etc.)

What to turn in

After you have tried to answer the following questions, please grade your answers (ChatGPT, Google, etc.) and write a little bit about how you performed overall and where you think you could use a little more review/practice. Again, I don't need your answers, just your score and it's OK if you have a low score (the typical self-reported score was between 7 and 8), this assignment is graded 2 = done, 1 = partially done, 0 = was not done.

If you scored 10/10 correct or you do not think you need any review/practice and do not wish to do this assignment, please write two addition questions that you think would have added more variety to this basic set of questions.

The questions to be answered (just need your summary paragraph and overall score)

1. In 2024, data regarding people who were identified as living in an unidentified large city in January 2020 and also between the ages of 50 and 59 at that time, were examined retrospectively. Those who had died in the period January 2020 - December 2023 were selected from the data set. The causes of deaths were categorized as suicide and all other causes. The employment status of the ones who had died from suicide were categorized into three classes as employed, unemployed and other (e.g., retired). Considering the information obtained, which statistical test would you use to investigate the possible relationship between cause of death and employment status?
2. For the calculation of the Apgar score, which is based on the physical states of the newborns measured at the 1st and the 5th minutes after birth, the score value is obtained by the sum of

5 components. Each component can take the values 0, 1 or 2 and a total score between 0 and 10 is obtained. Which statistical test would you use to investigate the relationship between the Apgar score values measured at the 1st and 5th minutes for a large random sample of babies? These scores are not distributed as bell-shaped curves, i.e., Apgar scores do not exhibit a normal distribution.

3. In a study investigating whether there is any relationship between the estriol levels of pregnant women and the birth weight of their newborn babies, the estriol levels are measured from the blood samples taken from pregnant women. Later, the newborn babies birth weights were obtained. Which statistical test would you use to investigate whether it is possible to predict newborn birth weights based on maternal estriol levels?

4. In a study which investigates whether passive smoking have a measurable effect with regard to pulmonary health, pulmonary functions were measured for 6 different smoking states. These are non-smokers, passive smokers, non-inhaling smokers, light smokers (1 to 10 per day), moderate smokers (11 to 39 per day) and heavy smokers (40 and over per day). The measurements of pulmonary functions were performed by forced mid-expiratory flow (FEF). Which statistical test would you use to compare whether the variations of the data with regard to each smoking categories are similar?

5. A radiologist and a chatbot trained to evaluate computed tomographic scans are being compared for diagnostic accuracy. A total of 109 scans were randomly selected, some of which contain neurologic pathology. The true disease status of each scan was determined using reliable methods. Both the radiologist and the chatbot rated each scan using a scale from 1 to 5, where 1 = definitely normal, 2 = probably normal, 3 = uncertain, 4 = probably abnormal, and 5 = definitely abnormal. What statistical test would be most appropriate to compare the diagnostic accuracy of the radiologist and the chatbot?

6. The blood pressure values of 50 male patients, between the ages of 35 and 44, who had been selected for a study, were measured as mmHg. After some treatment, the blood pressure values were measured again, and it was seen that the blood pressure values obtained from both measurements were distributed as a bell-shaped curve, i.e. exhibited a normal distribution. Which statistical test would you use in determining whether there is a statistically significant difference between the mean blood pressure values measured before and after the treatment?

7. A researcher wishes to test whether the mean IQ scores of 35 students are different from the mean IQ score value of 100, which had been obtained in the previous studies. If the IQ scores of the students are known to be distributed as a bell-shaped curve, i.e. exhibit a normal distribution, which statistical test would you use to determine whether the mean of the selected sample group is different from the mean value obtained in the previous studies?

8. A researcher desires to quantify the effect of three exercise forms (e.g., Yoga, Zumba, and Traditional Strength Training) on mood improvement scores measured on a scale of 0–100 (0 = no improvement, 100 = maximum improvement). A random sample of 90 patients from a mental health facility were recruited and then randomly assigned into groups of 30 and given one of the three exercise regimens for 6 weeks. At the end of the 6 week study, the patients were asked to score their mood improvement. It was discovered that the distributions were highly non-normal and also contained outliers. What statistical test might be suitable for assessing which regimen provided the highest mood improvement?

9. In a study for determining whether environmental factors cause a rise in serum cholesterol level,

which is an important risk factor for the etiology of the cardiovascular diseases, it is seen that the serum cholesterol levels obtained, from 100 genetically unrelated married couples, distributed as bell-shaped curve, i.e., exhibit a normal distribution. Which statistical test would you use to investigate the relationship between the serum cholesterol levels of husbands and their wives?

10. In a study investigating the effect of diet type with regard to sex on systolic blood pressure, the diet type is categorized into three (strict vegetarian, lacto-vegetarian and normal). It is known that the mean systolic blood pressure values for different diet types with regard to sex are distributed as bell-shaped curve, i.e. exhibit normal distributed, and the variablity of the data is similar. With regard to this information, which statistical test would you use to determine whether there is a difference between mean systolic blood pressure values for different diet types with regard to sex?