# Statistical Inference

*HW Author: Ali Mollahosseini, Pouya Gohari, Parham Sazdar*
*Instructor: Mohammadreza A. Dehaqani*

Spring 2024

HW 3: Part II Solutions

## Problem 1: Sufficient Evidence

The average adult male height in a certain country is 170 cm. We suspect that the men in a certain city in that country might have a different average height due to some environmental factors. We pick a random sample of size 9 from the adult males in the city and obtain the following values for their heights (in cm):
176.2, 157.9, 160.1, 180.9, 165.1, 167.2, 162.9, 155.7, 166.2
Assume that the height distribution in this population is normally distributed. Here, we need to decide between

$$H_0 : \mu = 170$$

$$H_1 : \mu \neq 170$$

Based on the observed data, is there enough evidence to reject $H_0$ at significance level $\alpha = 0.05$?

## Problem 2: Alternate Hypothesis

- Let $X_1, X_2, \cdots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where $\mu$ is unknown and $\sigma$ is known. Design a level $\alpha$ test to choose between

$$H_0 : \mu \leq \mu_0,$$

$$H_1 : \mu > \mu_0.$$

- Let $X_1, X_2, X_3, X_4$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where $\mu$ and $\sigma$ are unknown. Suppose that we have observed the following values 3.58, 10.03, 4.77, 14.66 .We would like to decide between

$$H_0 :\geq 10$$

$$H_1 : \mu < 10$$

  Assuming $\alpha = 0.05$, Do you accept $H_0$ or $H1$?

## Problem 3: Comparing Means

The manager of a lemonade bottling plant is interested in comparing the performance of two production lines, one of which has only recently been installed. For each line she selects 10 one hour periods at random and records the number of crates completed in each hour. The table below gives the results. From past experience with this kind of equipment it is known that the variance in these figures will be 10 for Line 1 and 25 for Line 2. Assuming that these samples came from normal populations with these variances, test the hypothesis that the two populations have the same mean.

| Production line | Number of crates completed per hour | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 (new) | 78 | 87 | 79 | 82 | 87 | 81 | 85 | 80 | 82 | 83 |
| 2 (old) | 74 | 77 | 78 | 70 | 87 | 83 | 76 | 78 | 81 | 76 |

**Table 1:** Production output comparison between new and old production lines

## Problem 4: Significant Difference

The temperature of the earth may be measured either by thermometers on the ground (x), which is an accurate but tedious method, or by sensors mounted in space satellites (y), which is a less accurate method and may be biased. The following table gives readings taken by both methods at eleven sites.

| Site | Ground therm, x | Satellite sensors, y |
|---|---|---|
| 1 | 4.6 | 4.7 |
| 2 | 17.3 | 19.5 |
| 3 | 12.2 | 12.5 |
| 4 | 3.6 | 4.2 |
| 5 | 6.2 | 6.0 |
| 6 | 14.8 | 15.4 |
| 7 | 11.4 | 14.9 |
| 8 | 14.9 | 17.8 |
| 9 | 9.3 | 9.7 |
| 10 | 10.4 | 10.5 |
| 11 | 7.2 | 7.4 |

**Table 2:** Comparison of ground and satellite sensor measurements

Given that all readings are normally distributed, investigate the hypothesis that satellite sensors give, on average, significantly higher readings than the ground thermometers.

## Problem 5: Testing Independance

Suppose that 300 persons are selected at random from a large population, and each person in the sample is classified according to blood type, O, A, B, or AB, and also according to Rh, positive or negative. The observed numbers are given in Table below. Test the hypothesis that the two classifications of blood types are

**Table 3:** Data for Question 5

| | O | A | B | AB |
|---|---|---|---|---|
| Rh positive | 82 | 89 | 54 | 19 |
| Rh negative | 13 | 27 | 7 | 9 |

independent.

## Problem 6: Tests Comparing Variance

Two college instructors are interested in whether or not there is any variation in the way they grade math exams. They each grade the same set of 30 exams. The first instructor's grades have a variance of 52.3. The second instructor's grades have a variance of 89.9. Test the claim that the first instructor's variance is smaller.

The level of significance is 10%. (In most colleges, it is desirable for the variances of exam grades to be nearly the same among instructors.)

## Problem 7: Variance Test(Programming Question)

Suppose we have:

$$X_1, X_2, \cdots, X_n \sim N(0, \sigma_x^2)$$
$$Y_1, Y_2, \cdots, Y_n \sim N(0, \sigma_y^2)$$

1. What method do you suggest for testing the hypothesis $H_0 : \sigma_x = \sigma_y$?
2. Using samples from two normal distributions with variances 3 and 10, examine the above method.

## Problem 8: Contingency Table(Programming Question)

Using the Titanic dataset, create a frequency table to display the association between "sex" and "survive" variables.

1. Part 1: Creating Frequency Table and Visualization.
   - Read the Titanic dataset from the CSV file.
   - Create a contingency table for the variables "sex" and "survive".
   - Save this table as sex_survive_table.
   - Display the contingency table showing the count of individuals based on gender and survival status.
   - Generate a mosaic plot visualizing the relationship between gender and survival, specifying colors and axis labels for better interpretation.
2. Part 2: $\chi^2$ Contingency Test and Fisher's Exact Test.
   - Conduct the $\chi^2$ test on the sex_survive_table contingency table.
   - Verify the assumptions for the $\chi^2$ test.
   - (Optional) Research about Fisher's exact test and then execute it for the same contingency table.
   - Display and analyze the results of the $\chi^2$ test, including the chi-squared value, degrees of freedom, and the p-value indicating the association significance.
   - (Optional) Present the Fisher's exact test results, including the p-value, confirming the association between gender and survival without approximations.

# Problem 9: Goodness of Fit(Programming Question)

Table 4. shows the number of sons in families that have 12 children. Assuming that $X_i$ is the random variable corresponding to the number of sons in these 6115 families,

**Table 4:** Number of sons and their frequency

| Number of sons | Frequency |
|---|---|
| 0 | 7 |
| 1 | 45 |
| 2 | 181 |
| 3 | 478 |
| 4 | 829 |
| 5 | 1112 |
| 6 | 1343 |
| 7 | 1033 |
| 8 | 670 |
| 9 | 286 |
| 10 | 104 |
| 11 | 24 |
| 12 | 3 |

1. Introduce statistics that can be used to test the following hypothesis:

$$H_0 : X_1, X_2, \cdots , X_{6115} \sim Binomial(12, 0.5)$$

2. Using simulation, obtain the distribution of the introduced statistic and draw its histogram.

3. Based on the obtained distribution, calculate the p-value for the statistic calculated in part 1. With $\alpha = 0.05$, is $H_0$ rejected?