Indian Institute of Technology Madras

Department of Data Science and Artificial Intelligence

DA5000: Mathematical Foundations of Data Science

Tutorial VII - Solutions

Problem 1: One-Sample Z-Test for Blood Pressure Data

Step 1: Calculate the Sample Mean and Standard Deviation

Given the blood pressure data, we calculate:

$$\bar{x} = \frac{\sum_{i=1}^{25} x_i}{25} = 118.48\,\text{mmHg}$$

The sample standard deviation is:

$$s = \sqrt{\frac{\sum_{i=1}^{25}(x_i - \bar{x})^2}{24}} = 2.6633\,\text{mmHg}$$

Step 2: Formulate Hypotheses

- Null Hypothesis (H_0): $\mu = \mu_0 = 120$ mmHg
- Alternative Hypothesis (H_a): $\mu \neq \mu_0$

Step 3: Calculate the Z-statistic

The z-statistic is calculated using the formula:

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Where:

$$\begin{aligned}
\bar{x} &= 118.48\,\text{mmHg}\,\text{(sample mean)} \\
\mu_0 &= 120\,\text{mmHg}\,\text{(population mean)} \\
\sigma &= 10\,\text{mmHg}\,\text{(population standard deviation)} \\
n &= 25\,\text{(sample size)}
\end{aligned}$$

Substituting these values into the formula gives:

$$z_{\text{stat}} = \frac{118.48 - 120}{10/\sqrt{25}} = -0.76$$

Step 4: Determine the p-value

The p-value is determined from the z-statistic using a standard normal distribution. Since this is a two-tailed test (as we are testing for any difference, not just an increase or decrease), the p-value is calculated as:

$$p\text{-value} = 2 \times (1 - \Phi(|z_{\text{stat}}|))$$

Calculating from z-table, for our z-statistic of $z_{\text{stat}} = -0.76$, the p-value is approximately:

$$p\text{-value} = 0.4472$$

where Φ is the cumulative distribution function of the standard normal distribution.

Reference: Steps to calculate the p-value for a z-score of -0.76 using a standard normal distribution table, follow these steps:

1. **Locate the Z-Score in the Table:** Standard normal distribution tables typically provide the cumulative probability (area under the curve to the left) for positive z-scores. For a z-score of -0.76, you would first find the cumulative probability for 0.76 (since the distribution is symmetric).
2. **Find the Cumulative Probability:** Look up 0.76 in the z-table. The table will provide the cumulative probability for this z-score. Typically, this value is around 0.7764.
3. **Calculate the P-Value for Two-Tailed Test:**
   - Since you are dealing with a two-tailed test (testing for any difference), you need to consider both tails of the distribution.
   - The cumulative probability for -0.76 is calculated as:
     $$1 - 0.7764 = 0.2236$$
   - For a two-tailed test, multiply this by 2 to account for both tails:
     $$2 \times 0.2236 = 0.4472$$

Thus, the p-value for a z-score of -0.76 in a two-tailed test is approximately 0.4472.

Step 5: Decision

At a significance level of $\alpha = 0.05$, since $0.4472 > 0.05$, we fail to reject the null hypothesis.

Thus, there is not enough evidence to conclude that the new diet affects blood pressure levels in this sample.

Problem 2: F-test

(Given: $F = 3.36$ at 5% level for $\nu_1 = 10$ and $\nu_2 = 8$.)

Solution: Let us take the null hypothesis that the two populations have not the same variance.

Applying F-test:

$$F = \frac{S_1^2}{S_2^2}$$

A X_1	$(X_1\text{-}\bar{X}_1)$ x_1	x_1^2	B X_2	$(X_2\text{-}\bar{X}_2)$ x_2	x_2^2
65	-15	225	64	-19	361
66	-14	196	66	-17	289
73	-7	49	74	-9	81
80	0	0	78	-5	25
82	2	4	82	-1	1
84	4	16	85	2	4
88	8	64	87	4	16
90	10	100	92	9	81
92	12	144	93	10	100
			95	12	144
			97	14	196
$\sum X_1$= 720	$\sum x_1$=0	$\sum x_1^2$-=798	$\sum X_2$=913	$\sum x_2$=0	$\sum x_2^2$=1298

Figure 1: F-Test Info Table

$$\bar{X}_1 = \frac{\sum X_1}{n_1} = \frac{720}{9} = 80$$

$$\bar{X}_2 = \frac{\sum X_2}{n_2} = \frac{913}{11} = 83$$

$$S_1^2 = \frac{\sum x_1^2}{n_1 - 1} = \frac{798}{9 - 1} = 99.75$$

$$S_2^2 = \frac{\sum x_2^2}{n_2 - 1} = \frac{1298}{11 - 1} = 129.8$$

$$F = \frac{S_1^2}{S_2^2} = \frac{99.75}{129.8} = 0.768$$

At the 5% level of significance, for $\nu_1 = 10$ and $\nu_2 = 8$, the table value of $F_{0.05} = 3.36$.

Since the calculated value of F is less than the table value, we accept the hypothesis. Hence, the two populations do not have the same variance.

Problem 3

Mean of sample 1

$$\mu_1 = \frac{\Sigma_{i=1}^{6} X_{1i}}{6}$$

$$\mu_1 = 0.61$$

Mean of sample 2

$$\mu_2 = \frac{\Sigma_{i=1}^{5} X_{2i}}{5}$$

$$\mu_2 = 0.492$$

sample 1 standard deviation

$$s_1 = \sqrt{\frac{\Sigma_{i=1}^{6}(X_{1i} - \mu_1)^2}{5}} = 0.0518$$

similarly sample 2 standard deviation

$$s_2 = 0.0819$$

It is unequal variance question. Confidence Interval=

$$(\mu_1 - \mu_2) \pm t * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$t = 2.2621, dof = n_1 + n_2 - 1 = 9, \alpha = 0.05, \alpha/2 = 0.025$

On solving, we get the confidence interval as:

$$0.118 \pm 0.09567 = [0.02232, 0.2137]$$

The null hypothesis is

$$H_0 : \mu_1 = \mu_2$$

Alternate hypothesis

$$H_0 : \mu_1 > \mu_2$$

$$d_0 = 0$$

The value of t sample is

$$\frac{\bar{X}_1 - \bar{X}_2 - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

On calculating, we get

$$t_s = 2.79$$

degrees of freedom=9

$$t_{crit} = 1.8331$$

As

$$t_2 > t_{crit}$$

the null hypothesis is rejected. Therefore, the fish have higher levels of mercury in the polluted lake as compared with the unpolluted one.

Problem 4: Paired t-test

Question:

We intend to compare several methods for predicting the shear strength for steel items. Data for two of these methods, the M1 and M2, when applied to nine specific items, are shown in Table 1.

Determine if there is a significant difference in shear strength predictions between methods M1 and M2 at a significance level of $\alpha = 0.05$. Address the following while arriving at the conclusion:

1. Formulate the Null and Alternate Hypothesis
2. Calculate the test statistic and compare with the critical value
3. Also compute the p value corresponding to the test statistic

Item	Method 1	Method 2
1	1.186	1.061
2	1.151	0.992
3	1.322	1.063
4	1.339	1.062
5	1.200	1.065
6	1.402	1.178
7	1.365	1.037
8	1.537	1.086
9	1.559	1.052

Table 1: Comparison of predictions of Method 1 and Method 2 for 9 items

Solution:

Step 1: Formulate Hypotheses

- **Null Hypothesis** (H_0): There is no significant difference between the shear strength predictions of M1 and M2, i.e., $\mu_D = 0$, where $D = M1 - M2$.
- **Alternate Hypothesis** (H_1): There is a significant difference between the predictions, i.e., $\mu_D \neq 0$.

Step 2: Calculate the Test Statistic

Let D_i represent the difference $D_i = M1_i - M2_i$ for each item i.

We calculate:

1. The mean of the differences, $\bar{D}$.
2. The standard deviation of the differences, s_D.

3. The test statistic t using the formula:

$$t = \frac{\bar{D}}{s_D/\sqrt{n}}$$

where $n = 9$ is the number of items.

Item	$M1$	$M2$	$D = M1 - M2$
1	1.186	1.061	0.125
2	1.151	0.992	0.159
3	1.322	1.063	0.259
4	1.339	1.062	0.277
5	1.200	1.065	0.135
6	1.402	1.178	0.224
7	1.365	1.037	0.328
8	1.537	1.086	0.451
9	1.559	1.052	0.507

Calculating the mean difference:

$$\bar{D} = \frac{\sum D}{n} = \frac{0.125 + 0.159 + 0.259 + 0.277 + 0.135 + 0.224 + 0.328 + 0.451 + 0.507}{9} = 0.2738$$

Calculating the standard deviation of the differences:

$$s_D = \sqrt{\frac{\sum(D_i - \bar{D})^2}{n-1}} = 0.135$$

Substituting into the test statistic formula:

$$t = \frac{\bar{D}}{s_D/\sqrt{n}} = \frac{0.2738}{0.135/\sqrt{9}} = 6.084$$

Step 3: Determine the Critical Value

For a two-tailed test with $df = n-1 = 8$ degrees of freedom at a significance level of $\alpha = 0.05$, the critical t-value from the t-distribution table is approximately 2.306.

Since our calculated t-value $6.084 > 2.306$, we reject the null hypothesis.

Step 4: Calculate the p-value

Using a t-distribution table or calculator, we find the p-value corresponding to $t = 6.084$ with 8 degrees of freedom, which is $p = 0.000147$.

Conclusion

Since $p = 0.000147 < \alpha = 0.05$, we reject the null hypothesis and conclude that there is a significant difference between the shear strength predictions of M1 and M2.