Indian Institute of Technology Madras

Department of Data Science and Artificial Intelligence

DA5000: Mathematical Foundations of Data Science

Tutorial IX - Solutions

Problem 1

1. We're testing whether there's a relationship between two categorical variables: region (North, South, West) and satisfaction level (Satisfied, Neutral, Dissatisfied).

Step 1: Hypotheses

- **Null Hypothesis** (H_0): Customer satisfaction is independent of region
- **Alternative Hypothesis** (H_1): Customer satisfaction is dependent on region

Step 2: Calculate Row and Column Totals

Region	Satisfied	Neutral	Dissatisfied	Row Total
North	45	30	25	100
South	40	35	25	100
West	50	25	25	100
Column Total	135	90	75	300

Table 1: Observed Frequencies with Row and Column Totals

Step 3: Calculate Expected Frequencies

Expected frequency formula:

$$E_{i,j} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

For each cell:

$$\text{North-Satisfied: } E_{11} = \frac{100 \times 135}{300} = 45$$

$$\text{North-Neutral: } E_{12} = \frac{100 \times 90}{300} = 30$$

$$\text{North-Dissatisfied: } E_{13} = \frac{100 \times 75}{300} = 25$$

Region	Satisfied	Neutral	Dissatisfied
North	45	30	25
South	45	30	25
West	45	30	25

Table 2: Expected Frequencies

Step 4: Calculate Chi-Square Statistic

$$\chi^2 = \sum \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Calculations for each cell:

$$\text{North-Satisfied: } \frac{(45-45)^2}{45} = 0$$

$$\text{South-Satisfied: } \frac{(40-45)^2}{45} = 0.556$$

$$\text{West-Satisfied: } \frac{(50-45)^2}{45} = 0.556$$

Total Chi-Square statistic = 2.778

Step 5: Degrees of Freedom

$$df = (rows - 1)(columns - 1) = (3-1)(3-1) = 4$$

Step 6: Decision Making

At $\alpha = 0.05$:

- Critical value for $df = 4$ is 9.488
- Calculated $\chi^2(2.778) <$ critical value (9.488)
- p-value $= 0.595 > 0.05$

Decision: Fail to reject the null hypothesis.

There is not enough evidence to conclude that customer satisfaction depends on region.

Reference - Using Chi-Square Distribution Table:

(a) Critical value - At $\alpha = 0.05$ significance level with df = 4, looking at the chi-square table: The critical value is 9.488.

(b) Approximate p-value - Find row with df = 4, look for values that our $\chi^2(2.778)$ falls between. Our 2.778 falls somewhere between $\chi^2(0.20) = 5.989$ and $\chi^2(0.90) = 1.064$. Therefore p-value is between 0.20 and 0.90.

Problem 2

The null hypothesis for the given problem is :

$$H_o : \mu_L = \mu_M = \mu_H$$

The alternate hypothesis is given by: H_1 : Atleast 1 of the means is not equal degrees of freedom between groups=k-1=3-1=2

degrees of freedom within groups =n-k=24-3=21

total degrees of freedom=n-1=23

$$\bar{X}_{low} = \frac{\Sigma x_{low}}{8} = 92.875$$

$$\bar{X}_{medium} = \frac{\Sigma x_{medium}}{8} = 82.125$$

$$\bar{X}_{high} = \frac{\Sigma x_{high}}{8} = 75.5$$

$$\bar{X} = 83.5$$

(This is the overall mean).

$$SS_t = \Sigma(x_i - \bar{X})^2 = 1216 + 584 + 986 = 2786$$

$$SS_b = n * \Sigma(\bar{X}_i - \bar{X})^2 = 1230.25$$

$$SS_w = SS_t - SS_b = 1555.75$$

$$MSB = \frac{SS_b}{df_b} = 1230.25/2 = 615.125$$

$$MSE = \frac{SS_w}{df_w} = 1555.75/21 = 74.083$$

Let us now find the F-value.

$$F = \frac{MSB}{MSE} = 615.125/74.0833 = 8.303$$

From F tables for

$$\alpha = 0.01$$

,

$$f_c = F_{2,21} = 5.78$$

. As

$$F > f_c$$

, the differences are statistically significant and we reject the null hypothesis.

2. p-value calculation:

$$df_1 = 2, df_2 = 21, f = 8.303$$

,from f table for probability, we get:

$$p(f > 8.303) = 0.0022$$

which is the required answer.

3. This is a 2 sided confidence interval. For high air voids

$$\bar{X}_{high} = 75.5$$

. Effective standard deviation for distribution with n samples:$s/\sqrt{(}n)$ Here we get

$$s' = \frac{s}{\sqrt{n}} = 8.22/\sqrt{(}8) = 2.909$$

(Use sample standard deviation for s) number of degrees of freedom=n-1=8-1=7 For 95 % confidence, we find $t_{0.025,7} = 2.3646$ as it is a 2 tailed test. The confidence interval is given by:

$$C.I = \bar{X_{high}} \pm t_{\frac{\alpha}{2},df} * s'$$

$$C.I = 75.5 \pm 2.3646 * 2.909$$

Therefore final range is given by

$$[68.621, 82.3786]$$

4. 95 % on difference between low and high:

$$\Delta X = \bar{X}_{low} - \bar{X}_{high} = 92.875 - 75.5 = 17.375$$

$$n_{low} = 8, n_{high} = 8$$

$$s_{high} = 8.22, s_{low} = 8.55$$

(Use sample standard deviation for both s_{low} and s_{high}

$$S.E_{\Delta X} = \sqrt{\frac{s_{low}^2}{n_{low}} + \frac{s_{high}^2}{n_{high}}}$$

$$S.E_{\Delta X} = 4.1933$$

$$df = n_1 + n_2 - 2 = 8 + 8 - 2 = 14$$

Therefore the critical t for the 2 sided test would be given by $t_{0.025,14} = 2.144$ The formula for confidence interval is given by

$$C.I = \Delta X \pm t_{0.025,14} * S.E_{\Delta X}$$

which gives us the range

$$[8.384, 26.365]$$

Problem 3

Let us first get the null hypothesis and the alternative hypothesis: H_o :The mean survival times of all cancers are equal H_a : The mean survival time of atleast 1 cancer isnt the same as the rest. Let us find the mean for different cancers as provided in the data:

$$\bar{X}_{stomach} = 286$$

$$\bar{X}_{bronchus} = 211.588$$

$$\bar{X}_{colon} = 457.4118$$

$$\bar{X}_{ovary} = 884.333$$

$$\bar{X}_{breast} = 1395.909$$

$$\bar{X}_{overall} = 558.625$$

Now the total sum of squares is given by:

$$SS_t = \Sigma(x_i - \bar{X})^2 = 37983905$$

$$SS_b = \Sigma(n_i * (\bar{X}_i - \bar{X})^2) = 11535760.5223$$

$$SS_w = SS_t - SS_b = 26448144.4777$$

$$df_{between} = k - 1 = 5 - 1 = 4$$

$$df_{within} = n - k = 64 - 5 = 59$$

$$MSB = \frac{SS_b}{df_b} = 11535760.5223/4 = 2883940.1306$$

$$MSE = \frac{SS_w}{df_w} = 26448144.4777/59 = 448273.6352$$

The F value is given by

$$F = \frac{MSB}{MSE} = 2883940.1306/448273.6352 = 6.433$$

For df1=4,df2=59, the critical f value at a significance of 0.01 is given as

$$f_c = F_{4,59} = 3.6549$$

As $F > f_c$, the result is significant and hence the null hypothesis is rejected.

Alternatively one may also compute the p value for F=6.433,df1=4,df2=59. p=0.000229. As p¡0.01, the null hypothesis is rejected.

Therefore atleast 1 of the cancer varieties has a different mean survival time as compared with others.

Problem 4

1. **State the null and alternative hypotheses and the level of significance**
   - H_0: Breastfeeding and autism are independent.
   - H_A: Breastfeeding and autism are dependent.
   - Significance level: $\alpha = 0.01$.
2. **State and check the assumptions for the hypothesis test**
   (a) A random sample of breastfeeding time frames and autism incidence was taken.
   (b) Expected frequencies for each cell are greater than or equal to 5 (i.e., $E \geq 5$). See step 3. All expected frequencies are greater than 5.
3. **Find the test statistic and p-value**
   Test Statistic:
   First, calculate the expected frequencies for each cell.

$$E(\text{Autism and no breastfeeding}) = \frac{818 \times 261}{934} \approx 228.585$$

$$E(\text{Autism and ¡ 2 months}) = \frac{818 \times 223}{934} \approx 195.304$$

$$E(\text{Autism and 2 to 6 months}) = \frac{818 \times 191}{934} \approx 167.278$$

$$E(\text{Autism and more than 6 months}) = \frac{818 \times 259}{934} \approx 226.833$$

Similar calculations are done for the other cells. The calculations for the test statistic are shown in Table.

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
241	228.585	12.415	154.132225	0.674288448
198	195.304	2.696	7.268416	0.03721591
164	167.278	-3.278	10.745284	0.064236086
215	226.833	-11.833	140.019889	0.617281828
20	32.4154	-12.4154	154.1421572	4.755213792
25	27.6959	-2.6959	7.26787681	0.262417066
27	23.7216	3.2784	10.74790656	0.453085229
44	32.167	11.833	140.019889	4.352904809
			Total	11.2166432

Table 3: Calculations for Chi-Square Test Statistic

The test statistic is calculated as:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 11.2166432$$

p-value:

Degrees of freedom: $df = (2-1)(4-1) = 3$.

Using a calculator:

$$\text{TI-83/84: } \chi\text{cdf}(11.2166432, \infty, 3) \approx 0.01061$$

Using R:

$$1 - p_{\chi^2}(11.2166432, 3) \approx 0.01061566$$

4. **Conclusion**

   Fail to reject H_0 since the p-value (0.0106) is greater than $\alpha = 0.01$.

5. **Interpretation**

   There is not enough evidence to conclude that breastfeeding and autism are dependent. This means that whether a child is breastfed or not does not provide sufficient information to indicate if the child will be diagnosed with autism.