Indian Institute of Technology Madras

Department of Data Science and Artificial Intelligence

DA5000: Mathematical Foundations of Data Science

Tutorial IX

Problem

1. A company conducted a customer satisfaction survey across three regions (North, South, and West) to determine if customer satisfaction is independent of the region. The responses were categorized as "Satisfied," "Neutral," and "Dissatisfied." The survey results are shown in the table below:

Region	Satisfied	Neutral	Dissatisfied
North	45	30	25
South	40	35	25
West	50	25	25

   (a) State the null and alternative hypotheses for this test.

   (b) Calculate the expected frequencies for each cell in the table under the assumption that satisfaction is independent of region.

   (c) Compute the Chi-Square test statistic using the formula:

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

   where $O_{i,j}$ is the observed frequency and $E_{i,j}$ is the expected frequency.

   (d) Determine the degrees of freedom for this test.

   (e) At a significance level of $\alpha = 0.05$, should you reject or fail to reject the null hypothesis? Use a Chi-Square distribution table or software to find the critical value or p-value.
   (Hint: Critical Value - If test statistic $>$ critical value: Reject H_0; If test statistic $<$ critical value: Fail to reject H_0. P-value - If p-value $< \alpha$: Reject H_0; If p-value $> \alpha$: Fail to reject H_0)

2. In an experiment,asphalt with low levels of air voids(2-4%),with medium level of air voids (4-6 %) and high(6-8 %) are tested.

   Refer to table 1 and answer the following:

   (a) Do the different levels of air voids significantly affect the mean retained strength. Use $\alpha = 0.01$

   (b) Find the p value of the F statistic in part (a)

   (c) Find the 95 % confidence interval on mean retained strength where there is a higher level of air voids.

   (d) Find a 95 % confidence interval on difference in mean retained strength at the low and high level of air voids.

3. Cancer is a terrible disease. Surviving may depend on the type of cancer the person has. To see if the mean survival time for several types of cancer are different, data was collected on the survival time in days of patients with one of these cancer in advanced stage. Do the data indicate that at least two of the mean survival time for these types of cancer are not all equal? Test at the 1% level.Refer table 2.

Table 1: Strength of asphalt at different concentrations

Low	Medium	High
106	80	78
90	69	80
103	94	62
90	91	69
79	70	76
88	83	85
92	87	69
95	83	85

Table 2: Survival Times for Different Cancer Types

Stomach	Bronchus	Colon	Ovary	Breast
124	81	248	1234	1235
42	461	377	89	24
25	20	189	201	1581
45	450	1843	356	1166
412	246	180	2970	40
51	166	537	456	727
1112	63	519		3808
46	64	455		791
103	155	406		1804
876	859	365		3460
146	151	942		719
340	166	776		
396	37	372		
	223	163		
	138	101		
	72	20		
	245	283		

4. Is there a relationship between autism and breastfeeding? To determine if there is, a researcher asked mothers of autistic and non-autistic children to report the time period they breastfed their children. The data is shown in Table 2. Do the data provide enough evidence to show that breastfeeding and autism are independent? Test at the 1% significance level.

	None	Less than 2 months	2 to 6 months	More than 6 months	Row Total
Yes (Autism)	241	198	164	215	818
No (No Autism)	20	25	27	44	116
Column Total	261	223	191	259	934

Table 3: Autism Versus Breastfeeding (Columns are breastfeeding timeline)