20170219_Batch26_CSE7315c_Chi_Anova_Lab_Activity

Objective:

In this lab you will revise the concepts and solve some problem that deal with hypothesis testing for goodness of fit, independence test, and testing the difference between means and variance using ANOVA.

Key takeaways:

- Chi-squared test for goodness of fit and as an independence test
- ANOVA
- Correlations

Problem 1: A survey is conducted by a gaming company that makes three video games. It wants to know if the preference of game depends on the gender of the player. Total number of participants is 1000. Here is the survey result. (Hint: A case of independence test)

	Game A	Game B	Game C	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

Problem 2: A national survey agency conducts a nationwide survey on consumer satisfaction and finds out the response distribution as follows:

Excellent:	8%
Good:	47%
Fair:	34%
Poor:	11%

A store manager wants to find if these results of customer survey apply to customers of super market in her city. So, she interviews 207 randomly selected customers and asked them to rate their responses. The results of this local survey are:

Response	Frequency
Excellent	21
Good	109
Fair	62
Poor	15

Determine if the local responses from this survey are the same as expected frequencies of the national survey, at 95% significance.

Problem 3: A car crash research team wants to examine the safety of compact cars, intermediate and full size cars. Given below are the hypothetical values of the mean pressure applied to the drivers head during the crash test for each of the car types. Check whether means are equal for each type of these cars.

Compact	643	655	702
Intermediate	469	427	525
Full size	484	456	402

Problem 4: Given below is the number of cups of coffee ordered in a restaurant in a week. Do the numbers show that people prefer any one coffee over the other?

Blue Label	Green Label	Red Label
3	6	9
5	7	10
6	9	15
2	7	12
1	11	11
2	6	10

Problem 5: Find the covariance of the eruption duration and waiting time in the data set "faithful" (built-in dataset in R). Observe if there is any linear relationship between the two variables.

Problem 6: Analyzing the linear relation among the attributes in the "Cereals" dataset

a. compute covariance and correlations on the data
b. write it to a file
c. plot the correlations/covariance and obtain the pairs of attributes that are highly correlated

Problem 7: Suppose that a random sample of n = 5 was selected from the vineyard properties for sale in Sonoma County, California, in each of three years. The following data are consistent with summary information on price per acre for disease-resistant grape vineyards in Sonoma County. Carry out an ANOVA to determine whether there is evidence to support the claim that the mean price per acre for vineyard land in Sonoma County was not the same for each of the three years considered. Test at the 0.05 level and at the 0.01 level.

1996: 30000 34000 36000 38000 40000
1997: 30000 35000 37000 38000 40000
1998: 40000 41000 43000 44000 50000

Problem 8: A business owner had been working to improve employee relations in his company. He predicted that he met his goal of increasing employee satisfaction from 65% to 80%. Employees from four departments were asked if they were satisfied with the working conditions of the company. The results are shown in the following table:

	Finance	**Sales**	**Human Resources**	**Technology**
Satisfied	12	38	5	8
Dissatisfied	7	19	3	1
Total	19	57	8	9

Problem 9: Many casinos use card-dealing machines to deal cards at random. Occasionally the machine is tested to ensure an equal likelihood of dealing for each suit. To conduct the test, 1,500 cards are dealt from the machine while the number of cards in each suit is counted. Theoretically, 375 cards should be dealt from each suit. As you can see from the results in the table, this is not the case:

	Spades	**Diamond**	**Clubs**	**Hearts**
Observed	402	358	273	467
Expected	375	375	375	375

Problem 10: In the dataset *mtcars*, is there significant difference in mileage across no. of cylinders of the car.

Problem 11: In the dataset 'airquality', manually compute covariance and correlation of 'Solar.R' & 'Temp'

Problem 12: In the dataset 'pressure', manually compute covariance and correlation of 'temperature' & 'pressure'.

Exercises: In each of the problems, state the Null and Alternate hypothesis, type of test, test statistics, Calculated value and tabulated value, rejection region and your conclusions.

Problem1: The Heavy Metal Corporation produces aluminum sheets which are specified to be 11mm thick. Due to several factors, there is natural variability in the thickness of the finished product. Two machines are used to produce the metal sheets. We wish to estimate whether the variance of machine 1 is same as variance of machine 2. To do this, a sample of metal sheets is selected from each machine for testing. The results are as follows:

Machine 1: n = 10; $\bar{x}$ = 11.02 mm; s_1^2 = 0.0284

Machine 2: n = 12; $\bar{x}$ = 10.9875 mm; s_2^2 = 0.0051

Problem 2: Susan Sound predicts that students will learn most effectively with a constant background sound, as opposed to an unpredictable sound or no sound at all. She randomly divides twenty-four students into three groups of eight. All students study a passage of text for 30 minutes. Those in group 1 study with background sound at a constant volume in the background. Those in group 2 study with noise that changes volume periodically. Those in group 3 study with no sound at all. After studying, all students take a 10 point multiple choice test over the material. Their scores follow:

Constant Sound	Random Sound	No Sound
7	5	2
4	5	4
6	3	7
8	4	1
6	4	2
6	7	1
2	2	5
9	2	5

Should Susan reject the Null Hypothesis?

Problem 3: A manger wants to see if geographical region is associated with ownership of a Dell computer. The manager surveys 100 people and the data breaks down as follows:

	Dell	No Dell	Row total
North East	12	14	26
South West	21	18	39

Mid West	17	18	35
Column Total	50	50	100

Problem 4: A Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans, and 10% are All-Stars.

Suppose a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with the company's claim? Use a 0.05 level of significance.

References:

Datasets:

http://www.idvbook.com/teaching-aid/data-sets/

UCI – datasets