**Agenda:** T-test,Chi-square test – Goodness of fit, Chi-square test – test of independence, ANOVA

**Worked-Out examples:**

1. A marketing director of a large department store wants to estimate the average number of customers who enter the store every five minutes. She randomly selects five-minute intervals and counts the number of arrivals at the store. She obtains the figures 68, 42, 51, 57, 56, 80, 45, 39, 36 and 79. The analyst assumes the number of arrivals is normally distributed. Using this data, the analyst computes a 95% confidence interval to estimate the mean value for all five-minute intervals. What interval value does she get?

Sample is small = 10 <30

Population variance is not given, go for T-TEST.

q1 <- c(68, 42, 51, 57, 56, 80, 45, 39, 36, 79)

t.test(q1, alternative = "two.sided")

#output : probability is 1.614e-06

Probability is less than 0.05

1. In ITC Paper board Limited, Bhadrachalam, a paper board packing plant, a machine packs cartons with cigarettes. It is supposed that a new machine will pack faster on the average than the machine currently used. To test that hypothesis, the times it takes each machine to pack ten cartons are recorded. The results in seconds, are shown in the text file “machine.txt”. Does the data provide sufficient evidence to conclude that, on the average, the new machine packs faster? Perform the required hypothesis test at the 5% level of significance.

H0 : Null hypothesis: new machine is faster.

H1: new machine is not faster.

setwd('E://Insofe//Week4//Day2//Lab//20171224\_Batch37\_CSE7315c\_t-test-ChiSqtest\_Ftest\_\_Lab05')

machines <- read.csv("machines.csv", sep = ",", header = T)

t.test(machines$New.machine, machines$Old.machine,alternative = "less")

p-value = 0.001621 is less than alpha which is 0.05

We can reject the null hypothesis

1. A study was performed to test whether cars get better mileage on premium gas than on regular gas. Each of 10 cars was first filled with either regular or premium gas, decided by a coin toss, and the mileage for that tank was recorded and shown in the below table. The mileage was recorded again for the same cars using the other kind of gasoline. Determine whether cars get significantly better mileage with premium gas. Find the confidence interval for both regular and premium mileages.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Mileage** | **Regular** | 16 | 20 | 21 | 22 | 23 | 22 | 27 | 25 | 27 | 28 |
| **Premium** | 19 | 22 | 24 | 24 | 25 | 25 | 26 | 26 | 28 | 32 |

H0: premius is better than regular.

H1: premium is not better than regular.

reg = c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28)

prem = c(19, 22, 24, 24, 25, 25, 26, 26, 28, 32)

t.test(prem,reg,alternative = "less", paired = T, conf.level = 0.95)

p-value = 0.9992 is greater than alpha = 0.05

we accept the null hypothesis.

1. 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 the 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 given below. Determine if the local responses from this survey are the same as expected frequencies of the national survey, at 95% confidence level.

|  |  |
| --- | --- |
| Response | Frequency |
| Excellent | 21 |
| Good | 109 |
| Fair | 62 |
| Poor | 15 |

expectation is given, and survey response is also given.

We need to find if survey is as per expectation. It is goodness of test.

# HO: expected values are equal to observed values

# H1 : E values not equals observed values

observedFrequencies<- c(21,109,62,15)

expectedProportions<- c(0.08,0.47,0.34,0.11)

chisq.test(observedFrequencies,p = expectedProportions)

p value = 0.1001, alpha = 0.5, pvalue > alpha so accept null hypothesis

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.

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

* 1. State the null hypothesis and alternate hypothesis b. Calculate the degrees of freedom

null hypothesis - no preference for game depends on gender

alternative hypothesis - preference for game depends on gender

#degrees of freedom

(m-1)\*(n-1) = (2-1)\*(3-1) = 2

* 1. Does men's preference is different from women's preference? Check with 0.05 level of significance

genderPreference <- matrix(c(200,150,50,250,300,50),nrow=2,byrow = T)

chisq.test(genderPreference)

p value = 0.03, alpha = 0.05

p < alpha, reject H0

1. A car-crash research team wants to examine the safety of compact, 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 |

use Anova as 3 differnt types of casrs

scores = c(643,655,702,469,427,525,484,456,402)

group <- c(rep(c("compact","inter","fullsize"),c(3,3,3)))

data <- data.frame(scores,group)

aov(scores~group, data=data)

summary(aov(scores~group, data=data))

probabilty 0.00121 < 0.05, we can reject the null hypothesis

1. Given below is the data of hobby preferences for men and women. Determine whether there is an association between gender preference for choosing hobbies

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Hobby | | |  |
|  |  | Music | Sport | Writing | Total |
| Gender | Men | 100 | 120 | 60 | 280 |
| Women | 350 | 200 | 90 | 640 |
|  |  | 450 | 320 | 150 | 920 |

H0: Gender and hobbies are independent and not associated.

H1: Gender and hobbies are associated.

genderPreference\_hobbies <- matrix(c(100,120,60,350,200,90),nrow=2,byrow = T)

genderPreference\_hobbies

chisq.test(genderPreference\_hobbies)

p value is 6.938e-07 which is less than alpha = 0.05

we can reject the null hypothesis

1. # Applying our learning on mtcars data set in R and answer the following

# Q1 Does the variables am and gears are independent?

# Which test should you use?

chisq as am is categorical and gears is numeric

# What is the conclusion?

q8 <- table(mtcars$am, mtcars$gear)

q8

chisq.test(q8) # assuming significance level is 0.05

H0 : Automatic are equal to manual

as p-value < 0.05 , reject null hypothesis

## Q2 Does the mpg varies on the am?

# create the two vectors one for manual and another for automatic.

# apply the test

H0: mpg varies on am.

H1: mpg does not vary with am.

mpg <- c(mtcars$mpg)

am <- c(mtcars$am)

t.test(mpg,am,paired = T)

## Q3 Does the mpg depends on am and gear?

H0: mpg depends on am and gear.

H1: mpg does not depend on am and gear.

aov(formula = mpg ~ am & gear, data = mtcars)

summary(aov(formula = mpg ~ am & gear, data = mtcars))

probability 0.000285 less than 0.05, we reject the null hypothesis.

## Q4 Does the mpg depends on am and cyl ?

H0: mpg depends on am and cyl.

H1: mpg does not depend on am and cyl.

aov(formula = mpg ~ am & cyl, data = mtcars)

summary(aov(formula = mpg ~ am & cyl, data = mtcars))

probability 8.01e-12 is less than 0.05, we reject the null hypothesis.

**### Summary**

# chi-square test is used to check if there are only categorical variables

# t test is to check one or two sample tests

# ANOVA is to apply multiple tests