**KUMARAGURU COLLEGE OF TECHNOLOGY**

**LABORATORY MANUAL**

**Experiment Number: 7**

**Lab Code : U18MAI4201**

**Lab : Probability and Statistics**

**Course / Branch : B.E-CSE,ISE, B.Tech-IT**

**Title of the Experiment : Application of Chi square test**

**STEP 1: INTRODUCTION**

**OBJECTIVES OF THE EXPERIMENT**

1. To apply chi square test for goodness of fit

2. To apply chi square test for independence of attributes

**STEP 2: ACQUISITION**

**Conditions for the validity of -test**

1. The sample observations must be independent of one another.
2. The sample size must be reasonably large, say 50.
3. No individual frequency should be less than 5. If any frequency is less than 5, then it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5. Finally adjust for the d.f lost in pooling.
4. The number of classes must be neither too small nor too large, ie

**-test of goodness of fit**

Tests ofgoodness of fit are used when we want to determine whether an actual sample distribution matches a known theoretical distribution. It enables us to find if the deviation of the experiment from theory is just by chance or it is due to the inadequacy of the theory to fit the data.

**Null Hypothesis**: H0: The difference between the observed and expected frequencies is not significant. ie, the theory fits well into the given data.

**Regular method**: Let be a set of observed frequencies and be the corresponding set of expected frequencies. Then follows Chi- Square Distribution with n – 1 d.f.

(One degree of freedom is subtracted for the constraint )

Compare the calculated **-**value with the tabulated **-**value (with n – 1 d.f) and form the conclustion.

**- test of Independence of Attributes**

**-** test isused for testing the null hypothesis that two criteria of classification are independent. Let the two attributes be *A* and *B*, where *A* has *r* categories and *B* has *s* categories. Thus the members of the population and hence, those of the sample are divided into classes. Let the total number of observations be *N*. The observations are arranged in the form of a matrix, called contingency table .

H0:The attributes *A* and *B* are independent.

**Regular method:**

The expected frequencies for various cells are calculated using the formula:

=

Test statistic is which follows - distribution with degrees of freedom.

**Note**: For a contingency table with cell frequencies , the - value is given by

; ,

Degree of freedom =1

**Procedure for doing the Experiment:**

|  |  |
| --- | --- |
| **1.** | **R-code for testing goodness of fit:**  f=vector of observed frequencies  p= vector of expected ratios (probabilities)  a=chisq.test(f,p=c(p1 ,p2 , ….))  a |
| **2.** | **R-code for testing independence of attributes:**  a = vector of elements in first row of contingency table  b = vector of elements in second row of contingency table  c = ……….  contingency = as.data.frame(rbind(a,b,c,….)) # to create the table  contingency  chisq.test(contingency,simulate.p.value=T) |

**Example: (-test of goodness of fit )**

**The following table gives the number of aircraft accidents that occur during the various days of a week. Find whether the accidents are uniformly distributed over the week.**

**Days Sun Mon Tue Wed Thu Fri Sat**

**No. of accidents: 14 16 8 12 11 9 14**

**Null Hypothesis:** The accidents are uniformly distributed over the week

**Alternative Hypothesis:** The accidents are not uniformly distributed over the week

Level of significance: 5% (say)

**R-code:**

accident=c(14,16,8,12,11,9,14)

p=c(1/7,1/7,1/7,1/7,1/7,1/7,1/7)

a=chisq.test(accident,p=c(1/7,1/7,1/7,1/7,1/7,1/7,1/7))

a

**Output:**

Chi-squared test for given probabilities

data: accident

X-squared = 4.1667, df = 6, p-value = 0.6541

**Table value** of for 6 d.f = 12.59

**Conclusion:** < ,so we accept and conclude that the accidents are uniformly distributed over the week.

**(Or)**

Here *p* value ≥ α value,so we accept and conclude that the accidents are uniformly distributed over the week.

**Task 1**

**The following figures show the distribution of digits in numbers chosen at random from a telephone directory**

**Digits 0 1 2 3 4 5 6 7 8 9 Total**

**Frequency 1026 1107997 966 1075 933 1107 972 964 853 10000**

**Test whether the digits may be taken to occur equally frequently in the directory.**

**Null Hypothesis:** The digits occur equally frequently in the directory.

**Alternative Hypothesis:** The digits does not occur equally frequently in the directory

**Level of significance:** 5%

**R-code:**

frequ=c(1026,1107,997,996,1075,933,1107,972,964,853)

p=c(1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10)

a=chisq.test(frequ,p=c(1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10))

a

**Output:**

Chi-squared test for given probabilities

data: frequ

X-squared = 58.542, df = 9, p-value = 2.558e-09

**Table value** of for 9 d.f = 16.919

**Conclusion:**

> ,so we reject and conclude that the digits does not occur equally frequently in the directory

**Task 2**

**The following is the distribution of the hourly number of trucks arriving at a company’s warehouse:**

**Trucks arriving hour 0 1 2 3 4 5 6 7 8 Total**

**Frequency 52 51 56 47 60 57 59 6157 500**

**Test whether the arrival of trucks is equally distributed at the 0.05 level of significance.**

**Null Hypothesis**: The arrival of trucks is equally distributed at the 0.05 level of significance **Alternative Hypothesis:** The arrival of trucks is not equally distributed at the 0.05 level of significance

**Level of significance:** 5%

**R-code:**

frequ=c(52,51,56,47,60,57,59,61,57)

p=c(1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9)

a=chisq.test(frequ,p=c(1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9,1/9))

a

**Output:**

Chi-squared test for given probabilities

data: frequ

X-squared = 3.1, df = 8, p-value = 0.9279

**Table value** of for 8 d.f = 15.507

**Conclusion:**

< ,so we accept and conclude that the arrival of trucks is equally distributed at the 0.05 level of significance

**Example (- test of Independence of Attributes)**

**A survey of 920 people that ask for their preference of one of three ice cream flavours (chocolate, vanilla, strawberry) gives the following results:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Flavour** | | | | |
| **Gender** |  | **Chocolate** | **Vanilla** | **Strawberry** | **Total** |
| **Men** | **100** | **120** | **60** | **280** |
| **Women** | **350** | **200** | **90** | **640** |
| **Total** | **450** | **320** | **150** | **920** |

**Using test, determine whether or not there is an association between gender and preference for ice cream flavour.**

**R-code**

men = c(100, 120, 60)

women = c(350, 200, 90)

icecream = as.data.frame(rbind(men, women))

chisq.test(icecream,simulate.p.value=T)

**Output:**

V1 V2 V3

men 100 120 60

women 350 200 90

>chisq.test(icecream, simulate.p.value=T)

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: icecream

X-squared = 28.362, df = NA, p-value = 0.0004998

**Table value** of =5.991

**Conclusion**: >, hence we conclude that there is association between gender and preference for ice cream flavour.

**Note:**

The R-code

men = c(100, 120, 60)

women = c(350, 200, 90)

ice.cream.survey = as.data.frame(rbind(men, women))

ice.cream.survey

generates the table

V1 V2 V3

men 100 120 60

women 350 200 90

**Task 3**

**Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas and one from the residents of urban areas. The results are given in the table. Examine whether the nature of the area is related to voting preference in the election**

|  |  |  |  |
| --- | --- | --- | --- |
| **Votes for area** | **A** | **B** | **Total** |
| **Rural** | **620** | **380** | **1000** |
| **Urban** | **550** | **450** | **1000** |
| **Total** | **1170** | **830** | **2000** |

**Null Hypothesis: H0:** The nature of the area is not related to voting preference in the election

**Alternative Hypothesis: H1:** the nature of the area is related to voting preference in the election

**Level of significance: = 5%**

**R-code:**

rural = c(620,380)

urban= c(550,450)

office = as.data.frame(rbind(rural,urban))

chisq.test(office,simulate.p.value=T)

**Output:**

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: office

X-squared = 10.092, df = NA, p-value = 0.002499

**Table value:** = 3.841

**Conclusion:**

>, we reject H0 hence we conclude that the nature of the area is related to voting preference in the election

**Task 4**

**A sample of 200 persons with a particular disease was selected. Out of these, 100 were given a drug and the others were not given any drug. The results are as follows:**

|  |  |  |
| --- | --- | --- |
| **No. of persons** | **Drug** | **No drug** |
| **Cured** | **65** | **55** |
| **Not cured** | **35** | **45** |

**Test whether the drug is effective or not (Use = 0.05)**

**Null Hypothesis: H0:** The Drug is not effective.

**Alternative Hypothesis: H1:** The Drug is effective

**Level of significance: = 5%**

**R-code:**

cured = c(65,55)

notcured= c(35,45)

Drug = as.data.frame(rbind(cured,notcured))

chisq.test(Drug,simulate.p.value=T)

**Output:**

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: Drug

X-squared = 2.0833, df = NA, p-value = 0.1839

**Table value:** = 3.841

**Conclusion:**

<, we accept H0 hence we conclude that the drug is not effective.

**Task 5**

**The following data are collected on two characters.**

|  |  |  |
| --- | --- | --- |
|  | **Smokers** | **Non – Smokers** |
| **Literates** | **83** | **57** |
| **Illiterates** | **45** | **68** |

**Based on this, can you say that there is no relation between smoking and literacy?**

**Null Hypothesis: H0**: There is no relation between smoking and literacy

**Alternative Hypothesis: H1:** There is a relation between smoking and literacy

**Level of significance: = 5%**

**R-code:**

literates = c(83,57)

illiterates= c(45,68)

smoke = as.data.frame(rbind(literates,illiterates))

chisq.test(smoke,simulate.p.value=T)

**Output:**

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: smoke

X-squared = 9.4757, df = NA, p-value = 0.001999

**Table value:** = 3.841

**Conclusion:**

>, we reject H0 hence we conclude that there is a relation between smoking and literacy

**Task 6**

**From the following data, test whether there is any association between intelligence and economic conditions?**

**Intelligence**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Economic condition** | **Excellent** | **Good** | **Medium** | **Dull** | **Total** |
| **Good** | **48** | **200** | **150** | **80** | **478** |
| **Not good** | **52** | **180** | **190** | **100** | **522** |
| **Total** | **100** | **380** | **340** | **180** | **1000** |

**Null Hypothesis: H0:**  There is no association between intelligence and economic conditions

**Alternative Hypothesis: H1:** There is association between intelligence and economic conditions

**Level of significance: = 5%**

**R-code:**

good = c(48,200,150,80)

notgood= c(52,180,190,100)

int = as.data.frame(rbind(good,notgood))

chisq.test(int,simulate.p.value=T)

**Output:**

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: intl

X-squared = 6.2168, df = NA, p-value = 0.1039

**Table value:** 3 df at 5% = 7.815

**Conclusion:**

< , we accept H0 hence we conclude that there is no association between intelligence and economic conditions

**STEP 3: PRACTICE/TESTING**

1. **When is chi-square test used?**

The Chi Square statistic is commonly used for testing relationships between categorical variables. The Chi-Square statistic is most commonly used to evaluate Tests of Independence when using a bivariate table.

1. **State the conditions for the validity of -test**

1. The sample observations must be independent of one another.

2. The sample size must be reasonably large, say ≥ 50.

3. No individual frequency should be less than 5. If any frequency is less than 5, then it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5. Finally adjust for the d.f lost in pooling.

4. The number of classes k must be neither too small nor too large, ie4≤𝑘≤16

1. **When do we use -test of goodness of fit ?**

The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not. It is often used to evaluate whether sample data is representative of the full population.

1. **When do we use - test of Independence of Attributes?**

𝝌𝟐- test is used for testing the null hypothesis that two criteria of classification are

independent. Let the two attributes be A and B, where A has r categories and B has s

categories. The observations are arranged in the form of a matrix, called contingency

table.

**KUMARAGURU COLLEGE OF TECHNOLOGY**

**LABORATORY MANUAL**

**Experiment Number: 8**

**Lab Code : U18MAI4201**

**Lab : Probability and Statistics**

**Course / Branch : B.E-CSE,ISE, B.Tech-IT**

**Title of the Experiment : ANOVA – one way classification**

**STEP 1: INTRODUCTION**

**OBJECTIVES OF THE EXPERIMENT**

To perform analysis of variance for a completely randomized design

**STEP 2: ACQUISITION**

Analysis of variance refers to the separation of variance ascribable to one group of causes from the variance ascribable to the other group. It is used to test the homogeneity of several means.

Three types of variation are present in a data

1. Treatments
2. Environmental
3. Residual or Error

Assumptions for ANOVA test

1. The observations are independent.
2. The parent population is normal
3. Various treatment and environmental effects are additive in nature.
4. The samples have been randomly selected from the population

Null Hypothesis: All the population means are equal

Alternative Hypothesis: Some of the means are not equal.

Three important designs of experiments:

1. Completely Randomised Design (CRD) – One-way classification
2. Randomised Block Design (RBD) – Two-way classification
3. Latin Square Design (LSD) – Three-way classification

**Procedure for doing the Experiment:**

|  |  |
| --- | --- |
| **1.** | aov(response~factor,data=data\_name) |

**Example**

**A drug company tested three formulations of a pain relief medicine for migraine headachesufferers. For the experiment 27 volunteers were selected and 9 were randomly assigned to one of three drug formulations. The subjects were instructed to take the drug during theirnext migraine headache episode and to report their pain on a scale of 1 to 10 (10 beingmaximum pain)**

**Drug A 4 5 4 3 2 4 3 4 4**

**Drug B 6 8 4 5 4 6 5 8 6**

**Drug C 6 7 6 6 7 5 6 5 5**

**R-code:**

pain=c(4,5,4,3,2,4,3,4,4,6,8,4,5,4,6,5,8,6,6,7,6,6,7,5,6,5,5)

drug=c(rep("A",9),rep("B",9),rep("C",9))

data=data.frame(pain,drug)

data

results=aov(pain~drug,data=data)

summary(results)

**Output:**

pain drug

1 4 A

2 5 A

3 4 A

4 3 A

5 2 A

6 4 A

7 3 A

8 4 A

9 4 A

10 6 B

11 8 B

12 4 B

13 5 B

14 4 B

15 6 B

16 5 B

17 8 B

18 6 B

19 6 C

20 7 C

21 6 C

22 6 C

23 7 C

24 5 C

25 6 C

26 5 C

27 5 C

Df Sum Sq Mean Sq F value Pr(>F)

drug 2 28.22 14.111 11.91 0.000256 \*\*\*

Residuals 24 28.44 1.185

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

, so we reject the null hypothesis and conclude that that the means of the

three drug groups are different.

**Task 1**

**Three machines A, B & C gave the production of pieces in 4 days as below is there a significant difference between machines?**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A** | **17** | **16** | **14** | **13** |
| **B** | **15** | **12** | **19** | **18** |
| **C** | **20** | **8** | **11** | **17** |

**Null Hypothesis: H0:** There is no significant difference between machines.

**Alternative Hypothesis: H1:** There is significant difference between machines.

**R-code:**

pieces=c(17,16,14,13,15,12,19,18,20,8,11,17)

machine=c(rep("A",4),rep("B",4),rep("C",4))

data=data.frame(pieces,machine)

data

results=aov(pieces~machine,data=data)

summary(results)

**Output:**

pieces machine

1 17 A

2 16 A

3 14 A

4 13 A

5 15 B

6 12 B

7 19 B

8 18 B

9 20 C

10 8 C

11 11 C

12 17 C

Df Sum Sq Mean Sq F value Pr(>F)

machine 2 8 4.00 3.61 0.764

Residuals 9 130 14.44

**Conclusion:**

, so we accept the null hypothesis and conclude that there is no significant difference between machines.

**Task 2**

**Four machines A,B,C,D are used to produce a certain kind of cotton fabric. Samples of size 4 with each unit as 100 square meters are selected from the outputs of the machines at random and the number of flaws in each 100 square meters is counted with the following result.**

**A B C D**

**8 6 14 20**

**9 8 12 22**

**11 10 18 25**

**12 4 9 23**

**Do you think that there is significant difference in the performance of the four machines?**

**Null Hypothesis: H0:** There is no significant difference between machines.

**Alternative Hypothesis: H1:** There is significant difference between machines.

**R-code:**

cotton=c(8,9,11,12,6,8,10,4,14,12,18,9,20,22,25,23)

machine=c(rep("A",4),rep("B",4),rep("C",4),rep("D",4))

data=data.frame(cotton,machine)

data

results=aov(cotton~machine,data=data)

summary(results)

**Output:**

cotton machine

1 8 A

2 9 A

3 11 A

4 12 A

5 6 B

6 8 B

7 10 B

8 4 B

9 14 C

10 12 C

11 18 C

12 9 C

13 20 D

14 22 D

15 25 D

16 23 D

Df Sum Sq Mean Sq F value Pr(>F)

machine 3 540.7 180.23 25.22 1.81e-05 \*\*\*

Residuals 12 85.7 7.15

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Conclusion:**

, so we reject the null hypothesis and conclude that there is significant difference between machines.

**Task 3**

**Ten varieties of wheat are grown in 3 plots each and the following yields in quintals per acre, obtained.**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Variety** | | | | | | | | | |
|  |  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Plots** | **I** | **7** | **7** | **14** | **11** | **9** | **6** | **9** | **8** | **12** | **9** |
|  | **II** | **8** | **9** | **13** | **10** | **9** | **7** | **13** | **13** | **11** | **11** |
|  | **III** | **7** | **6** | **16** | **11** | **12** | **6** | **12** | **11** | **11** | **11** |

**Test the significance of the differences between variety yields**

**Null Hypothesis: H0:** There is no significant difference between variety yields.

**Alternative Hypothesis: H1:** There is significant difference between variety yields.

**R-code:**

variety=c(7,7,14,11,9,6,9,8,12,9,8,9,13,10,9,7,13,13,11,11,7,6,16,11,12,6,12,11,11,11)

plots=c(rep("I",10),rep("II",10),rep("III",10))

data=data.frame(variety,plots)

data

results=aov(variety~plots,data=data)

summary(results)

**Output:**

variety plots

1 7 I

2 7 I

3 14 I

4 11 I

5 9 I

6 6 I

7 9 I

8 8 I

9 12 I

10 9 I

11 8 II

12 9 II

13 13 II

14 10 II

15 9 II

16 7 II

17 13 II

18 13 II

19 11 II

20 11 II

21 7 III

22 6 III

23 16 III

24 11 III

25 12 III

26 6 III

27 12 III

28 11 III

29 11 III

30 11 III

Df Sum Sq Mean Sq F value Pr(>F)

plots 2 8.87 4.433 1.555 0.533

Residuals 27 186.10 6.893

**Conclusion:**

, so we accept the null hypothesis and conclude that there is no significant difference between variety yields

**Task 4**

**An experiment was conducted to study effect of four different dyes A, B, C, D on the strength of the fabric and following results of fabric strength are obtained.**

**Dye**

**A 8.67 8.68 8.66 8.65**

**B 7.68 7.58 8.67 8.65 8.62**

**C 8.69 8.67 8.92 7.7**

**D 7.7 7.90 8.65 8.20 8.60**

**Null Hypothesis: H0:** There is no significant difference between fabric strength.

**Alternative Hypothesis: H1:** There is significant difference between fabric strength

**R-code:**

strength=c(8.67,8.68,8.66,8.65,7.68,7.58,8.67,8.65,8.62,8.69,8.67,8.92,7.7,7.7,7.90,8.65,8.20,8.60)

dye=c(rep("A",4),rep("B",5),rep("C",4),rep("D",5))

data=data.frame(strength,dye)

data

results=aov(strength~dye,data=data)

summary(results)

**Output:**

strength dye

1 8.67 A

2 8.68 A

3 8.66 A

4 8.65 A

5 7.68 B

6 7.58 B

7 8.67 B

8 8.65 B

9 8.62 B

10 8.69 C

11 8.67 C

12 8.92 C

13 7.70 C

14 7.70 D

15 7.90 D

16 8.65 D

17 8.20 D

18 8.60 D

Df Sum Sq Mean Sq F value Pr(>F)

dye 3 0.6202 0.2067 1.023 0.412

Residuals 14 2.8304 0.2022

**Conclusion:**

, so we accept the null hypothesis and conclude that there is no significant difference between fabric strength.

**STEP 3: PRACTICE/TESTING**

1. **What are the basic principles of Experimental Design?**

1. Replication – Repetition of treatments under investigation

2. Randomisation – Assigning treatments randomly to the experimental units

3. Local control – Making the experimental units homogeneous and reducing the experimental error

1. **Mention the important designs of experiments:**

1. Planning of the experiment

2. Obtaining relevant information from the experiment regarding the statistical hypothesis under study.

3. Making a statistical analysis of the data

1. **Explain a completely randomized design.**

The term Completely Randomized Design refers to the fact that a single variable factor of interest is controlled and its effect on the other elementary units is observed. In other words, the data are classified according to only one criterion in one-way classification.

1. **What is the purpose of analysis of variance?**

Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.