1. (IND)A random sample of 500 U.S. adults is questioned regarding their political affiliation and opinion on a tax reform bill. The results of this survey are summarized in the following contingency table:

	Favor	**Indifferent**	**Opposed**	**Total**
Democrat	138	83	64	285
Republican	64	67	84	215
Total	202	150	148	500

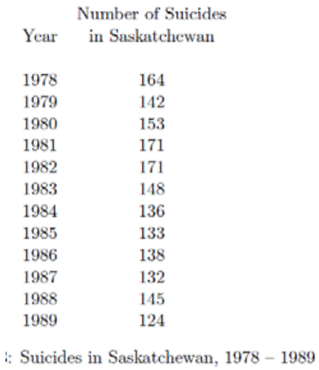
From here, we would want to determine if an association (relationship) exists between Political Party Affiliation and Opinion on Tax Reform Bill. That is, are the two variables dependent?

```
Input=("  democrat       republican
    favor                  138              64
    indifferent   83                 67
    opposed       64                 84
    ")
Mat
chisq.test(Mat,correct=F)

#Another way of reading the data

table3b=matrix(c(138,83,64,64,67,84),nrow=3, ncol=2, dimnames=list(c("Favor", "Indifferent","Opposed"),
                                        c("Democrat", "Republican") ))
table3b
chisq.test(table3b)
```

```
        Pearson's Chi-squared test

data:  table3b
X-squared = 22.152, df = 2, p-value = 1.548e-05
```

Conclusion: reject null hypothesis (of no relationship between Political Party Affiliation and Opinion on Tax Reform Bill).

2. (GoF)The media often seizes on yearly changes in crime or other statistics. A large jump in the number of murders from one year to the next, or a large decline in the the support for a particular political party may become the subject of many news reports and analysis. These statistics may be expected to show some shift from year to year, just because there is variation in any phenomenon. This question addresses this issue by looking at changes in the annual number of suicides in the province of Saskatchewan. 'Suicides in Saskatchewan declined by more than 13% from 1982 to 1983," was the headline in the Saskburg Times in early 1984. The article went on to interview several noted experts on suicide who gave possible reasons for the large decline. As a student who has just completed a statistics course, you come across the article and decide to check out the data. By consulting Table 25 of Saskatchewan Health, Vital Statistics, Annual Report, for various years, you find the figures for 1978-1989. The data is given in Table



Year	Number of Suicides in Saskatchewan
1978	164
1979	142
1980	153
1981	171
1982	171
1983	148
1984	136
1985	133
1986	138
1987	132
1988	145
1989	124

: Suicides in Saskatchewan, 1978 – 1989

Test the hypothesis that the number of suicides reported for each year from 1978 to 1989 does not differ significantly from an equal number of suicides in each year. (0.05 significance).
Based on this result, what might you conclude about the newspaper report, especially in light of the extra information for 1984 { 1989 that is now available to you?

```
#EX2.
observed = c(164,142,153,171,171,148,136,133,138,132,145,124)# observed frequencies
expected = c(rep(1/12, 12))      # expected proportions

chisq.test(x = observed,     p = expected)
```

```
Chi-squared test for given probabilities

data:  observed
X-squared = 18.133, df = 11, p-value = 0.07855
```

Null Hypothesis: Distribution of the observed number of suicides for each year does not differ significantly from an equal number of suicides each year

Alternative Hypothesis: Distribution of the observed number of suicides for each year differs significantly from an equal number of suicides each year

Conclusion: Since the test is not significant, there is insufficient evidence to reject Ho at the 5% level of significance. Therefore we fail to reject Ho.

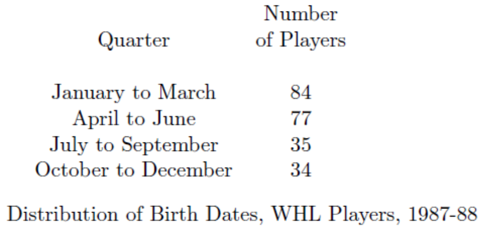
Let's note that the number of suicides is not equal for these 12 years, hence the null hypothesis is not entirely correct (considerable chance of Type II error – fail to reject Ho when H0 is false).

The data from 1983 to 1989 show that for these years (for which data is available) the number of suicides is lower than in the early 1980s. We note a recent peak of 171 suicides in 1981 and 1982, after which the number decreases. This fact points is a different direction than what is claimed in the null hypothesis and may be viewed as having peaked in early 1980s, followed by a drop in number of suicides.

3. (GoF) The Toronto Globe and Mail of November 27, 1987 contained an article, written by Neil Campbell, entitled \NHL career can be preconceived."In the article, Campbell claimed that the organization of hockey has turned half the boys in hockey-playing countries into second class citizens. The disadvantaged are those unlucky enough to have been born in the second half of the calendar year.
Campbell calls this the Calendar Effect, arguing that it results from the practice of age grouping of very young boys by calendar year of birth. For example, all boys 7 years old in 1991, and born in 1984, would be in the same grouping. By 1991, those boys born in the first few months of 1984 are likely to be somewhat larger and better

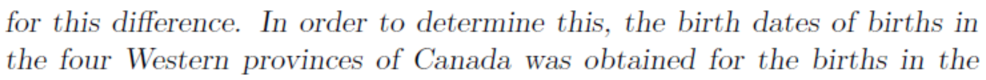
coordinated than those boys born in the later months of 1984. Yet all these players compete against each other. Campbell argues that this initial advantage stays with these players, and may become a permanent advantage by the time the boys are separated into elite leagues at age 9.

In order to test whether this Calendar Effect exists among hockey players who are somewhat older, a statistics student collected data from the Western Hockey League (WHL) Yearbook for 1987-88.

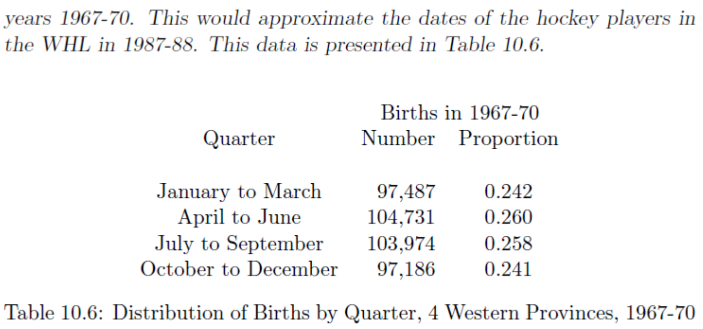


Quarter	Number of Players
January to March	84
April to June	77
July to September	35
October to December	34

Distribution of Birth Dates, WHL Players, 1987-88



for this difference. In order to determine this, the birth dates of births in the four Western provinces of Canada was obtained for the births in the



years 1967-70. This would approximate the dates of the hockey players in the WHL in 1987-88. This data is presented in Table 10.6.

	Births in 1967-70	
Quarter	Number	Proportion
January to March	97,487	0.242
April to June	104,731	0.260
July to September	103,974	0.258
October to December	97,186	0.241

Table 10.6: Distribution of Births by Quarter, 4 Western Provinces, 1967-70

```
> observed = c(84,77,35,34)# observed frequencies
> expected = c(0.242, 0.260,0.258, 0.241)      # expected proportions
> chisq.test(x = observed,
+            p = expected)
Error in chisq.test(x = observed, p = expected) : 
  probabilities must sum to 1.

> chisq.test(x = observed,
+            p = expected, rescale.p = TRUE)  ## see why this was needed!

        Chi-squared test for given probabilities

data:  observed
X-squared = 37.683, df = 3, p-value = 3.299e-08

> a=230*expected
> a
[1] 55.66 59.80 59.34 55.43
```

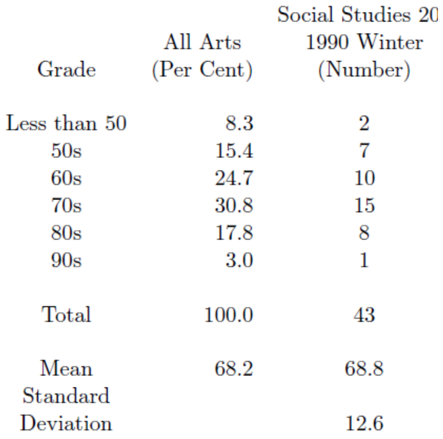
We will use α=0.001 significance level.

Null Hypothesis: The distribution of the observed number of births by quarter does not differ significantly from that of the distribution of births in the Western provinces in 1967-1970.

Alternative Hypothesis: The distribution of the observed number of births by quarter differs significantly from that of the distribution of births in the Western provinces in 1967-1970.

Conclusion: Since p-value=3.299e-08 < α=0.01, the null hypothesis can be rejected at the specified level of significance. Strong evidence exists of the Calendar Effect, based on the birth dates of the Western Hockey League (WHL) Yearbook for 1987-88.

4. The grade distribution for Social Studies 201 in the Winter 1990 semester is contained in Table. Along with the grade distribution for Social Studies 201 is the grade distribution for all the classes in the Faculty of Arts in the same semester.



		Social Studies 20
	All Arts	1990 Winter
Grade	(Per Cent)	(Number)
Less than 50	8.3	2
50s	15.4	7
60s	24.7	10
70s	30.8	15
80s	17.8	8
90s	3.0	1
Total	100.0	43
Mean	68.2	68.8
Standard Deviation		12.6

First test whether the model of a normal distribution of grades adequately explains the grade distribution of Social Studies 201. Then test whether the grade distribution for Social Studies 201 differs from the grade distribution for the Faculty of Arts as a whole. For each test, use the 0.20 level of significance.

```
> x<-c(rep(50,7), rep(60,10), rep(70,15), rep(80,8), rep(90,1))
> ks.test(x,"pnorm", m=68.8, sd=12.7)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.24285, p-value = 0.01588
alternative hypothesis: two-sided

Warning message:
In ks.test(x, "pnorm", m = 68.8, sd = 12.7) :
  ties should not be present for the Kolmogorov-Smirnov test
```

```
> observed<-c(2,7,10,15,8,1)
> x<-c(50,60,70,80,90)
> z_scores<-(x-68.8)/12.7

> z_scores
[1] -1.48031496 -0.69291339  0.09448819  0.88188976  1.66929134
> expected_prop<-c(0.0694,0.1757,0.2908,0.2747,0.1419,0.0475)
>
```

For this test, we need to determine the grade distribution that would exists if the grades were distributed exactly as a normal distribution (formal curve with mean 68.8 and standard deviation 12.7).

The z-scores for this distribution are computed, for the X values of 50,60,70, 80 and 90. The table presents the areas under the normal curve for the computed z's and the second column represent these proportions multiplied by 43 (total number)

Null Hypothesis: The distribution of the observed number of students in each grade category does not significantly differ from what would be expected if the grade distribution would be exactly normally distributed.

Alternative Hypothesis: The distribution of the observed number of students in each grade category significantly differs from what would be expected if the grade distribution would be exactly normally distributed.

We know that the cells should exceed 5 – in order for the test to be properly applied. Hence merge the 80's and the 90's together.

```
> chisq.test(observed,p=expected_prop)

        Chi-squared test for given probabilities

data:  observed
X-squared = 2.85, df = 5, p-value = 0.7231

Warning message:
In chisq.test(observed, p = expected_prop) :
  Chi-squared approximation may be incorrect
>
>
> observed<-c(9,10,15,9)
> expected_prop<-c(0.2451,0.2908,0.2747,0.1894)
>
> chisq.test(observed,p=expected_prop)

        Chi-squared test for given probabilities

data:  observed
X-squared = 1.6767, df = 3, p-value = 0.6421
```

```
Since p-value=0.6421 > 0.20 we fail to reject the null hypothesis that they are
normally distributed.
```

```
###############(CH)
> p<-c(0.083, 0.154, 0.247, 0.308, 0.208)
> arts<-43*p
> arts
[1]  3.569  6.622 10.621 13.244  8.944
> s201<-c(2,7,10,15, 9)
>
> observed<-c(s201,arts)
> observed
 [1]  2.000  7.000 10.000 15.000  9.000  3.569  6.622 10.621 13.244  8.944
>
>
> tab<-matrix(observed,ncol=2)
>
> colnames(tab)=c("S201","arts")
> rownames(tab)=c("<50","50", "60", "70", "80")

> addmargins(tab)
    S201   arts    Sum
<50    2  3.569  5.569
50     7  6.622 13.622
60    10 10.621 20.621
70    15 13.244 28.244
80     9  8.944 17.944
Sum   43 43.000 86.000
```

```
> tab
    S201   arts
<50    2  3.569
50     7  6.622
60    10 10.621
70    15 13.244
80     9  8.944
>
> chisq.test(tab)

        Pearson's Chi-squared test

data:  tab
X-squared = 0.58059, df = 4, p-value = 0.9652

Warning message:
In chisq.test(tab) : Chi-squared approximation may be incorrect
>
>
> tab <- matrix(c(33,153,103,16,29,181,81,14), nrow =4)
> tab
     [,1] [,2]
[1,]   33   29
[2,]  153  181
[3,]  103   81
[4,]   16   14

> addmargins(tab)
                Sum
     33  29  62
    153 181 334
    103  81 184
     16  14  30
Sum 305 305 610
>
>
> chisq.test(table4b,simulate.p.value = TRUE, B = 10000)

        Pearson's Chi-squared test with simulated p-value (based on 10000 replicates)

data:  table4b
X-squared = 0.58059, df = NA, p-value = 0.9947
```

Null Hypothesis: The distribution of the observed number of students in each grade category, in Social Studies 201, does not significantly differ from what would be expected if the grade distribution exactly matched that for all Arts grades.

Alternative Hypothesis: The distribution of the observed number of students in each grade category, in Social Studies 201, significantly differs from what would be expected if the grade distribution exactly matched that for all Arts grades.

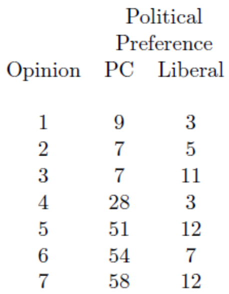
To obtain the number of grades in each category under the null hypothesis, multiply by 43 the percentages from the distribution of grades in the Faculty of Arts. This provides the expected number of cases in each category. Again, for the 90s we get only 1.3 cases, hence we merge it with the 80s.

Again, we fail to reject the null hypothesis at the 20% level of significance.

5. (TI) The table below gives the distribution of the opinions of 214 PC supporters and 53 Liberal supporters in the Edmonton study. The respondents were asked their view concerning the opinion "Unemployment is high because

trade unions have priced their members out of a job." Respondents gave their answers on a 7 point scale, with 1 being strongly disagree and 7 being strongly agree.

Use the data in this table to test whether political preference and opinion are independent of each other or not. Use the 0.01 level of significance.



	Political Preference	
Opinion	PC	Liberal
1	9	3
2	7	5
3	7	11
4	28	3
5	51	12
6	54	7
7	58	12

Distribution of Opinions of Edmonton PCs and Liberals

```
mydata <-"ex5.txt"
datam <- read.table("ex5.txt", row.names = 1, heade=TRUE)
datam

test<-chisq.test(datam)
#warning because many of the expected values will be very small
#therefore the approximations of p may not be right.
test<-chisq.test(datam, simulate.p.value = TRUE)
test
```

```
> test<-chisq.test(datam)
Warning message:
In chisq.test(datam) : Chi-squared approximation may be incorrect
> test

        Pearson's Chi-squared test

data:  datam
X-squared = 28.105, df = 6, p-value = 8.979e-05

> mmm <- read.table("ex6.txt", row.names = 1)
> mmm
            yes no
too_little   40  6
about_right  16 13
too_much      9  7
>
>
> test<-chisq.test(mmm)
Warning message:
In chisq.test(mmm) : Chi-squared approximation may be incorrect
> test

        Pearson's Chi-squared test

data:  mmm
X-squared = 10.996, df = 2, p-value = 0.004095
```

Null Hypothesis: No relationship between political preference and opinion concerning whether trades unions are partly responsible for unemployment.

Alternative Hypothesis: There is a relationship between political preference and opinion concerning whether trades unions are partly responsible for unemployment.

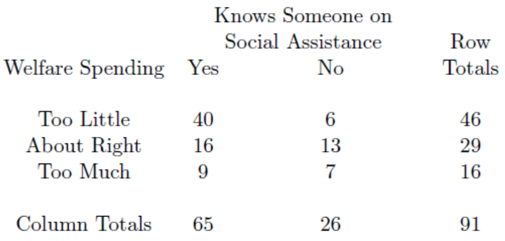
Since at the α=0.01 level of significance the test is significant, the null hypothesis is rejected and we can conclude that there is evidence of a relationship between political preference and opinion concerning whether trades unions are partly responsible for unemployment.

Observation: some cells have few cases …violation if the chi-square. However, the result is highly significant, with the null hypothesis rejected at 1%, hence we do not expect that the slight violation to affect too much the result.

6. **Attitudes to Social Spending in Newfoundland**

(TI) A sample of adults in Eastern and Central Newfoundland was conducted early in 1988 to examine public attitudes toward government cuts in social spending. Some of the results from this study are described in Morris Saldov, \Public Attitudes to Social Spending in Newfoundland," Canadian Review of Social Policy, 26, November 1990, pages 10-14. The data in Table below comes from Table 2 of this article. Concerning this data, the author comments,

Respondents, who knew someone on social assistance, were more likely to feel that welfare rates were too low.



	Knows Someone on Social Assistance		Row
Welfare Spending	Yes	No	Totals
Too Little	40	6	46
About Right	16	13	29
Too Much	9	7	16
Column Totals	65	26	91

```
mydata <-"ex6.txt"
mmm <- read.table("ex6.txt", row.names = 1)
mmm


test<-chisq.test(mmm)
#warning because many of the expected values will be very small
#therefore the approximations of p may not be right.
 test<-chisq.test(mmm, simulate.p.value = TRUE)
 test
test$observed
test$expected
```

```
> test

        Pearson's Chi-squared test

data:  mmm
X-squared = 10.996, df = 2, p-value = 0.004095

> test<-chisq.test(mmm, simulate.p.value = TRUE)
> test

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

data:  mmm
X-squared = 10.996, df = NA, p-value = 0.006997
```

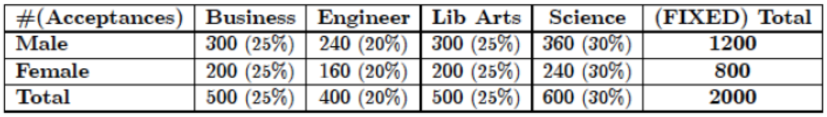
Choose α=0.001 level of significance.

Null Hypothesis: No relation between attitude and knowing someone on assistance.

Alternative Hypothesis: Some relation between attitude and knowing someone on assistance.

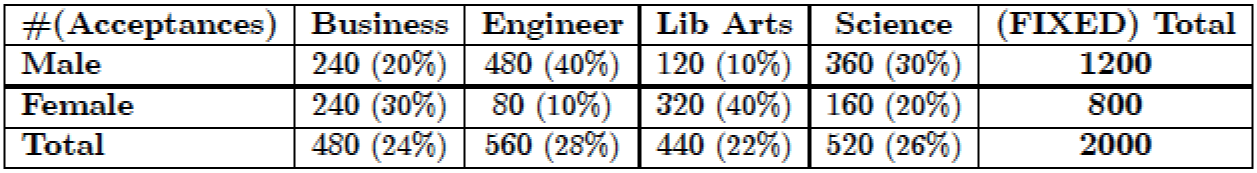
At the given level of significance, the test is not significant; hence the null hypothesis cannot be rejected. However, inspecting the table one can note the nature of the relationship. Of those respondents who know someone receiving social assistance, 40 out of 65 (61.5%) said that there is too little welfare spending. … Therefore, we can conclude that larger percentages of those who know someone on social assistance support more welfare spending than those who do not know anyone receiving such assistance.
The author result shows thaw ether or not an individual knows someone on social assistance appears to be associated with that individual's views concerning welfare spending.

7. (TH) A university admissions officer was concerned that males and females were accepted at different rates into the four different schools (business, engineering, liberal arts, and science) at her university. She collected the following data on the acceptance of 1200 males and 800 females who applied to the university:



#(Acceptances)	Business	Engineer	Lib Arts	Science	(FIXED) Total
Male	300 (25%)	240 (20%)	300 (25%)	360 (30%)	1200
Female	200 (25%)	160 (20%)	200 (25%)	240 (30%)	800
Total	500 (25%)	400 (20%)	500 (25%)	600 (30%)	2000

Are males and females distributed equally among the various schools? ---YES



#(Acceptances)	Business	Engineer	Lib Arts	Science	(FIXED) Total
Male	240 (20%)	480 (40%)	120 (10%)	360 (30%)	1200
Female	240 (30%)	80 (10%)	320 (40%)	160 (20%)	800
Total	480 (24%)	560 (28%)	440 (22%)	520 (26%)	2000

NO

```
> x<-matrix(c(300,200,240,160,300,200,360,240), ncol=4)
>
> colnames(x)=c("Business", "Engineer","Lib Arts", "Science")
> rownames(x)=c("Male", "Female")
> addmargins(x)
       Business Engineer Lib Arts Science  Sum
Male        300      240      300     360 1200
Female      200      160      200     240  800
Sum         500      400      500     600 2000
>
> chisq.test(x)

        Pearson's Chi-squared test

data:  x
X-squared = 0, df = 3, p-value = 1

>
>
>
>
>
> xy<-matrix(c(240,240, 480,80, 120,320, 360,160), ncol=4)
>
```

```
> colnames(xy)=c("Business", "Engineer","Lib Arts", "Science")
> rownames(xy)=c("Male", "Female")
> addmargins(xy)
       Business Engineer Lib Arts Science  Sum
Male        240      480      120     360 1200
Female      240       80      320     160  800
Sum         480      560      440     520 2000
>
> chisq.test(xy)

        Pearson's Chi-squared test

data:  xy
X-squared = 389.11, df = 3, p-value < 2.2e-16
```

Conclusion: Male and female are not distributed equally among the four schools.

8. (TH) Suppose that two colleges, the U and State, are worried about the student drinking behaviors, so they both independently choose random samples of their students. The results of the drinking behaviors are given in the table here:

Drinking Level	The U	State
None	140	186
Low	478	661
Moderate	300	173
High	63	16

The question is, does there appear to be a difference with drinking behaviors between the two colleges? Obviously, those who drink a lot represent the lowest category in both schools, and those who drink a little represent the highest in both schools. Perhaps the schools are not that different. You can run a test, though, to make sure whether that's the case or to dispute whether that's the case.

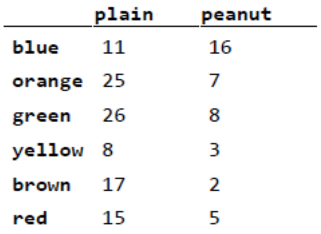
Choose α=0.05 level of significance.

```
>
> datam <- read.table("ex8.txt", row.names = 1, heade=TRUE)
> datam
           U State
None     140   186
Low      478   661
Moderate 300   173
High      63    16
> chisq.test(datam)

        Pearson's Chi-squared test

data:  datam
X-squared = 96.526, df = 3, p-value < 2.2e-16
```

Ho: distribution of drinking levels same for The U as is for the State.
H1: distribution of drinking levels is not the same for The U as is for the State.

Conclusion: Test of homogeneity – reject H0.

9. (TH) Plain M&M's candies come in six colors: orange, yellow, brown, green, blue and red. As do peanut M&M's. But do both types of candies share the same distribution of those colors?
I purchased a king-size package of plain M&M's and counted the number of each color of candy: of 102 candies in the package, 11 were blue, 25 orange, 26 green, 8 yellow, 17 brown and 15 red. I also purchased a king-size bag of peanut M&M's and counted the number of each color: of 41 candies, 7 were orange, 3 yellow, 2 brown, 8 green, 16 blue and 5 red.



	plain	peanut
blue	11	16
orange	25	7
green	26	8
yellow	8	3
brown	17	2
red	15	5

Choose α=0.01 level of significance.

```
> dat <- read.table("ex9.txt", row.names = 1, header=TRUE)
> dat
        plain peanut
blue       11     16
orange     25      7
green      26      8
yeallow     8      3
brown      17      2
red        15      5
> chisq.test(dat)

        Pearson's Chi-squared test

data:  dat
X-squared = 16.716, df = 5, p-value = 0.005071

Warning message:
In chisq.test(dat) : Chi-squared approximation may be incorrect
> chisq.test(dat,simulate.p.value = TRUE, B = 10000)

        Pearson's Chi-squared test with simulated p-value (based on 10000 replicates)

data:  dat
X-squared = 16.716, df = NA, p-value = 0.0048
```

Ho: color of plain M&Ms and the colors of peanut M&Ms candies have the same distribution
H1: color of plain M&Ms and the colors of peanut M&Ms candies do not have the same distribution

Conclusion: Small p-value, reject Ho.

10.(GoF) According to the manufacturer of M&M candy, the color distribution for plain chocolate M&Ms is 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, 16% green. We select a random sample of 300 plain M&M candies to test these hypotheses. In the 300 bags that we buy we have 38,32,51,58,74,47.

If the sample has the distribution of color stated in the null hypothesis, then we expect 13% of the 300 to be brown, 13% of 300 to be red, 14% of 300 to be yellow, 24% of 300 to be blue, and so on. Test the manufacturer claim.

Choose α=0.01 level of significance.

```
> observed<-c(38,32,51,58,74,47)
> expected_prop<- c(0.13,0.13,0.14,0.24,0.2,0.16)
>
> chisq.test(observed,p=expected_prop)

        Chi-squared test for given probabilities

data:  observed
X-squared = 9.2203, df = 5, p-value = 0.1006
```

Ho: color distribution of plain M&Ms is 13% brown, 14% yellow, 24% blue, 20% orange and 16% green
H1: color distribution of plain M&Ms is different from the hypothesized distribution

Conclusion: Fail to reject H0

11. (GoF)Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans but not All-Stars, and 10% are veteran All-Stars.
Suppose a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

```
> observed<-c(50,45,5)
> expected_prop<- c(0.3,0.6,0.1)
>
> chisq.test(observed,p=expected_prop)

        Chi-squared test for given probabilities

data:  observed
X-squared = 19.583, df = 2, p-value = 5.592e-05
```

Ho: proportion of Rockies, veterans and all –stars is 30%, 60^ and 10%
H1: proportion of Rockies, veterans and all –stars is 30%, 60^ and 10% is different (at least one is different)

Conclusion: Highly significant p-value, reject Ho.