STAT 614 CHAPTER 8 CHI-SQUARE TEST OF INDEPENDENCE

We will use the Chisquare Test for Independence in order to determine if two categorical variables are dependent or independent. For example, does Political Party Affiliation depend on Gender? Does you Gender influence your status as being a Democrat or a Republican? Typically the null hypothesis is; the two categories are independent and the alternative hypothesis is; the categories are dependent. We make a decision to reject or fail to reject the null hypothesis based on the size of the p value

Example 1

Is there a relationship between marital status and happiness? The data in the table below shows the marital status and happiness of individuals who participated in the General Social Survey.

H_0: marital status and happiness are independent (no relationship)

H_a: marital status and happiness are dependent (there is a relationship)

	Married	Widowed	Divorced/Separated	Never/Married
Very Happy	600	63	112	144
Pretty Happy	720	142	355	459
Not too Happy	93	51	119	127

Find column totals, row totals, and the table total

	Married	Widowed	Divorced/Separated	Never/Married	Totals
Very Happy	600	63	112	144	919
Pretty Happy	720	142	355	459	1676
Not too Happy	93	51	119	127	390
Totals	1413	256	586	730	**2985**

We can now perform typical basic two way table or (contingency table) calculations:

How many people surveyed are Pretty Happy and Never Married?

What proportion of Not Too Happy people are Widowed?

What proportion of all surveyed are Divorced or Separated?

What proportion of people surveyed are Pretty Happy or Never Married?

Find expected values (row total x column total)/ table total

	Married	Widowed	Divorced/Separated	Never/Married	Totals
Very Happy	600	63	112	144	919
Pretty Happy	720	142	355	459	1676
Not too Happy	93	51	119	127	390
Totals	1413	256	586	730	**2985**

Let's find selected expected values by hand (we will us R to find all of them)

E11 = (919 x 1413)/2985 = 435.02 E12 = (919 x 586)/2985 = 180.41

E32 = (390 x 256)/2985 = 33.45 E34 = (390 x 730)/2985 = 95.376

E23 = (1676 x 586) /2985 = 329.02

Step 4 Find the degrees of freedom (r-1)(c-1) = (3-1)(4-1) = 6

Step 4 Find the Chi Square Statistic

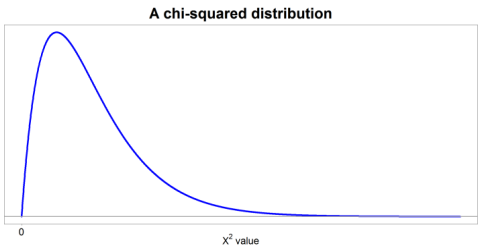
$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = (600-435.02)^2/435.02 + (63-180.41)^2/180.41 + … + (127-95.36)^2/ 95.36

= 224.12 (It is best to use software or R to calculate the chisquare statistic)

Lets find the p value

1 - pchisq(224.12, 6) 0



Conclusion: Since p = 0 which is less that .05. we reject the null hypothesis that there is no relationship between marital status and happiness. We say that marital status and happiness are dependent

	Married	Widowed	Divorced/Separated	Never/Married
Very Happy	600	63	112	144
Pretty Happy	720	142	355	459
Not too Happy	93	51	119	127

```
observed_table <- matrix(c(600,63,112,144,720,142,355,459,93,51,119,127), nrow = 3, ncol = 4, byrow = T)

colnames(observed_table) <- c("Married", "Widowed" , 'Divorced/Separated' , "NeverMarried")

rownames(observed_table) <- c("VeryHappy", "Prettyhappy", "NotToHappy")

observed_table

chisq.test(observed_table)
```

```
Pearson's Chi-squared test

data:  observed_table

X-squared = 224.12, df = 6, p-value < 2.2e-16
```

Conclusion: Since p < .05 , we reject the null hypothesis of independence and conclude that marital status and happiness are dependent.

Example 2 (Textbook)

Are the categorical variables Political Parties and Gender Independent ?

	Party Identification		
Gender	Democrat	Independent	Republican
Females	495	590	272
Males	330	498	265

```
PartyID <- matrix(c(495,590,272,330,498,265), nrow = 2, ncol = 3, byrow = T)

colnames(PartyID) <- c("Democrat", "Independent", "Republican")

rownames(PartyID) <- c("Females", "Males")

PartyID
```

```
chisq.test(PartyID)
```

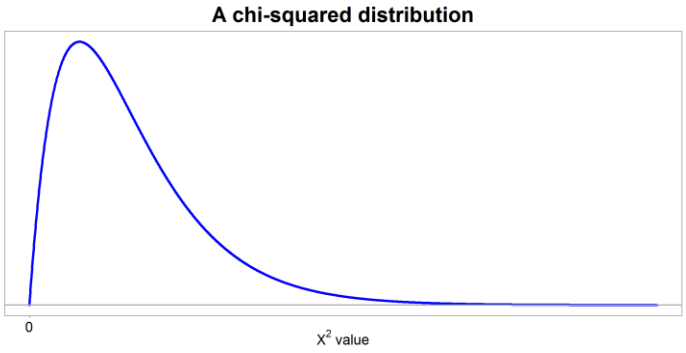
We can also find the expected values by extending our R code as follows:

```
chisq.test(PartyID) ->  ExpectedValues
ExpectedValues
ExpectedValues$expected
```

```
Pearson's Chi-squared test
data:  PartyID
X-squared = 12.569, df = 2, p-value = 0.001865
```

Expected Values

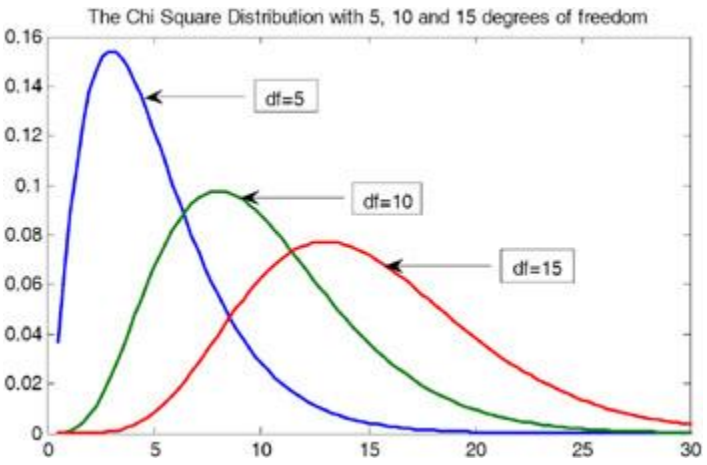
	Democrat	Independent	Republican
Females	456.949	602.6188	297.4322
Males	368.051	485.3812	239.5678



Conclusion: Since p < .05, we reject the null hypothesis of Independence and conclude that Party Identification and Gender are dependent.

CHISQUARE TEST OF INDEPENDENCE AND THE CHISQUARE DISTRIBUTION

FACTS AND OBSERVATIONS



Low degress of freedom results in a χ^2 distribution that is highly skewed to the right

As degress of freedom increases, χ^2 distributions become less skewed. An extremely large degree of freedom measure results in a distribution that is nearly normal.

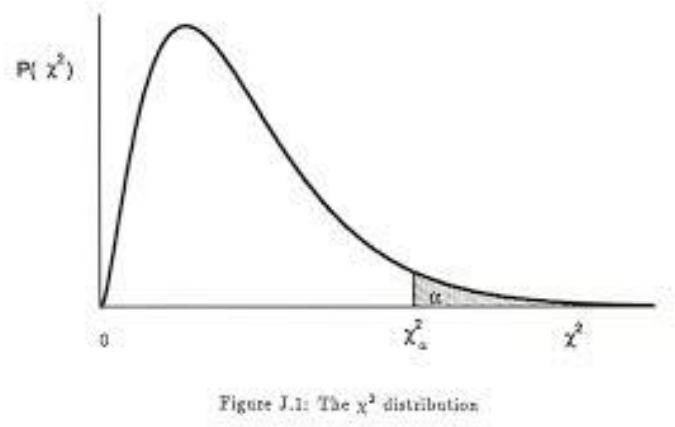


Figure J.1: The χ^2 distribution

The shaded area to right of χ^2 statistic is p-value. The higher the χ^2 value the lower the p value, which leads to a stronger case against the null hypothesis of independence.

Consider again the ChiSquare formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The closer the observed values and the expected values are to each other, the smaller the χ^2 value will be

Additional facts and observations

- The father apart the observed values and the expected values are to each other, the larger the χ^2 value will be
- One primary Chisquare Test requirement is that the expected value for every cell should be larger than 5
- The mean of the χ^2 distribution = df
- The standard deviation of the χ^2 distribution is sqrt(2df)
- The χ^2 test of independence will determine if the two categorical variables are related but it will not determine how strong the relationship is.

Residual Analysis

By definition,

- a residual for a specific cell is $(O_i - E_i)$.
- a standardized residual is $z = (O_i - E_i)/\text{sqrt}(O_i(1 - \text{row proportion})(1 - \text{column proportion})$

We use a standardized residual to determine if the difference $(O_i - E_i)$ has a special meaning or impact. Since the standardized residuals generated a standard normal distribution, if a difference is lower than **-3** or greater than **3** the relationship of the cell is special or unusual given an assumption of independence for the null hypothesis.

For example : Let us consider two categorical variables Gender and Favorite Color. Of course

Gender Male Female

Favorite Color Blue Red Green

For the cell of intersection for Female and Red we have a standardized residual of **3.25 .** This tells us that the Females prefer a color of Red **more** than we would expect if the variables are independent. And conversely if the standardized residual was **-3.25** we would say that Females prefer a color of Red **less** than we would expect if the variables are independent.

Calculating and analyzing Standard Residuals

```
PartyID <- matrix(c(495,590,272,330,498,265), nrow = 2, ncol = 3, byrow = T)
colnames(PartyID) <- c("Democrat", "Independent", "Republican")
rownames(PartyID) <- c("Females", "Males")
PartyID
chisq.test(PartyID)
Democrat Independent Republican
Females      495         590        272
Males        330         498        265
Pearson's Chi-squared test
data:  PartyID
X-squared = 12.569, df = 2, p-value = 0.001865

chisq.test(PartyID)$expected

chisq.test(PartyID)$expected
Democrat Independent Republican
Females  456.949    602.6188   297.4322
Males    368.051    485.3812   239.5678
chisq.test(PartyID)$stdres

chisq.test(PartyID)$stdres
         Democrat Independent Republican
Females  3.272365   -1.032199  -2.498557
Males   -3.272365    1.032199   2.498557
```

Special or unusual relationships assuming the variables are Independent:

Democrat/Females Democrat/Males