

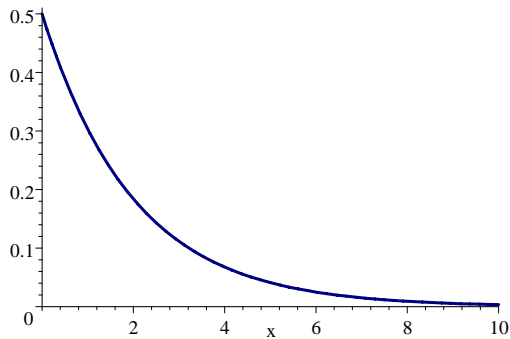
Let us look at another tool that will help as compare population proportions for more than two populations.

## Chi Square Distribution Introduction and Applications

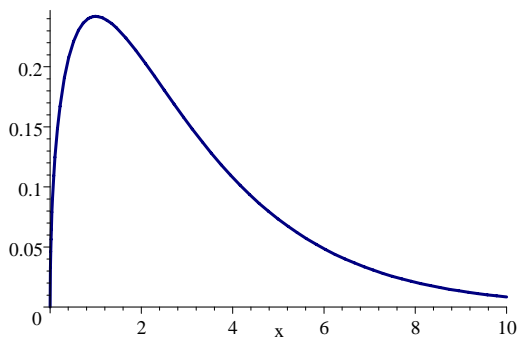
Atul N Roy

The Chi Square distribution is a probability distribution, where the random variable assumes only non negative real values. It is skewed to the right. Its shape depends on a characteristic called the degrees of freedom. Here are a few examples:

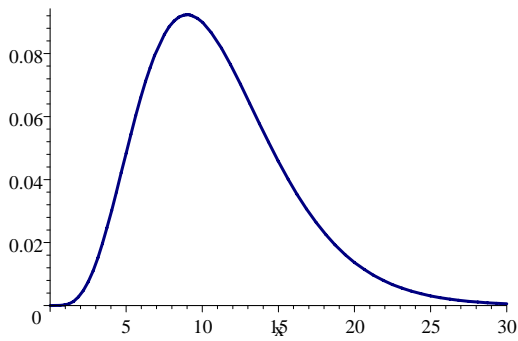
1. Chi Square (also expressed by  $\chi^2$ ) with 2 degrees of freedom



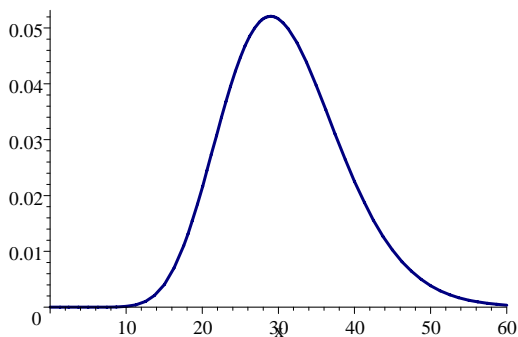
2.  $\chi^2$  with 3 degrees of freedom looks like the following



3.  $\chi^2$  with 11 degrees of freedom looks like



4.  $\chi^2$  with 31 degrees of freedom looks like



Note that the distribution has become quite symmetric with the increase in the degrees of freedom.

Fact:

If the degrees of freedom of a  $\chi^2$  distribution is  $n$ , then its mean is  $n$  and the standard deviation is  $\sqrt{2n}$ .

We shall either use the tables or a computing device to use the probabilities involving the  $\chi^2$  distribution.

I shall write the description of the computer usage later in this section, first let us look at a table that is typically found in most statistics-text books.

	0.5	0.1	0.05	0.025	0.01	0.005
2	1.386294	4.605176	5.991476	7.377779	9.210351	10.59653
3	2.365973	6.251394	7.814725	9.348404	11.34488	12.83807
4	3.356695	7.779434	9.487728	11.14326	13.2767	14.86017
5	4.351459	9.236349	11.07048	12.83249	15.08632	16.74965
6	5.348119	10.64464	12.59158	14.44935	16.81187	18.54751
7	6.345809	12.01703	14.06713	16.01277	18.47532	20.27774
8	7.34412	13.36156	15.50731	17.53454	20.09016	21.95486
9	8.342832	14.68366	16.91896	19.02278	21.66605	23.58927
10	9.341816	15.98717	18.30703	20.4832	23.20929	25.18805
11	10.341	17.27501	19.67515	21.92002	24.72502	26.75686
12	11.34032	18.54934	21.02606	23.33666	26.21696	28.29966
13	12.33975	19.81193	22.36203	24.73558	27.68818	29.81932
14	13.33927	21.06414	23.68478	26.11893	29.14116	31.31943
15	14.33886	22.30712	24.9958	27.48836	30.57795	32.80149
16	15.3385	23.54182	26.29622	28.84532	31.99986	34.26705
17	16.33818	24.76903	27.5871	30.19098	33.40872	35.71838
18	17.3379	25.98942	28.86932	31.52641	34.80524	37.15639
19	18.33765	27.20356	30.14351	32.85234	36.19077	38.58212
20	19.33743	28.41197	31.41042	34.16958	37.56627	39.99686
21	20.33723	29.61509	32.67056	35.47886	38.93223	41.40094
22	21.33704	30.81329	33.92446	36.78068	40.28945	42.79566
23	22.33688	32.00689	35.17246	38.07561	41.63833	44.18139
24	23.33673	33.19624	36.41503	39.36406	42.97978	45.55836
25	24.33658	34.38158	37.65249	40.6465	44.31401	46.92797
26	25.33646	35.56316	38.88513	41.92314	45.64164	48.28978
27	26.33634	36.74123	40.11327	43.19452	46.96284	49.64504
28	27.33623	37.91591	41.33715	44.46079	48.27817	50.99356
29	28.33613	39.08748	42.55695	45.72228	49.58783	52.3355
30	29.33603	40.25602	43.77295	46.97922	50.89218	53.67187

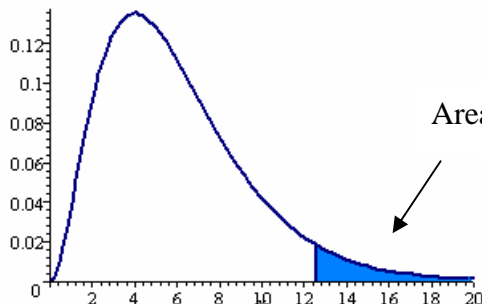
The values in the top row are the right tail probabilities and the values in the first column are the degrees of freedom.

Following is the way to interpret this table.

Note that

	0.05
	↓
6	→ 12.591

means that the area under the graph of the density function to the right of  $\chi^2 = 12.591$  at 6 degrees of freedom is approximately 0.05 as shown in the following graph.



This distribution has many uses in statistical analyses, we shall use it for comparing population proportions.

Example 1.

The following data is based on the information provided by a heartburn medication called Aciphex regarding relief from heartburn symptoms among the patients who took placebo and the patients who took different dosage of Aciphex.

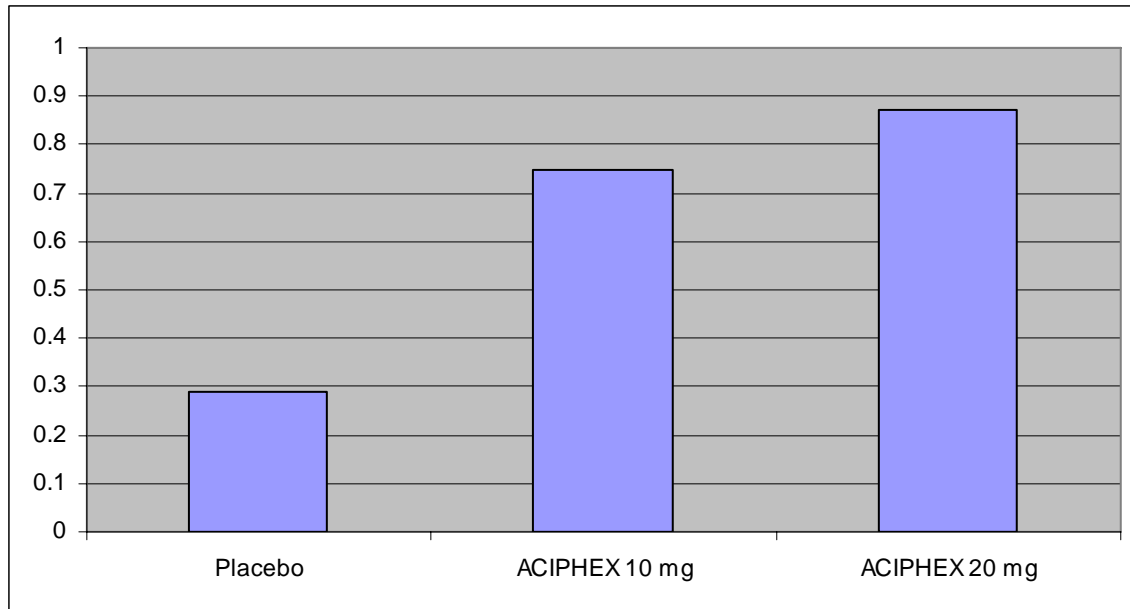
The table shows the number that maintained the healing from ulcers verses the number that did not under two treatment groups and the placebo group 52 weeks after the treatment.

	Maintained Healing	
	YES	NO
Placebo	49	120
ACIPHEX 10 mg	119	40
ACIPHEX 20 mg	139	21

We call such a table a 3x2 table that is a table with 3 rows and 2 columns and having 6 cells overall.

Note that the following shows us the proportion of the subjects who maintained healing for these three groups.

	Maintained Healing
Placebo	0.29
ACIPHEX 10 mg	0.75
ACIPHEX 20 mg	0.87



We would like to test whether the differences that we see in the above table and the bar graph may be attributed to chance.

Note that the above table shows sample proportion of the people who maintained healing after receiving the shown dosage.

If

$p_1$  = the proportion who maintained healing in the placebo group

$p_2$  = the proportion who maintained healing in the ACIPHEX 10 mg group

$p_3$  = the proportion of headache in the population in the ACIPHEX 20 mg group

Note that the null hypothesis here is

$$p_1 = p_2 = p_3$$

that is there is no association between the percentage of subjects who maintained healing and the type of pill that they receive.

If we write the alternative as a negation then it is

$$p_1 \neq p_2 \text{ or } p_2 \neq p_3 \text{ or } p_1 \neq p_3$$

If we did three different hypothesis tests for proportions, then it will be too detailed and at the same time we run into doing multiple testing based on the same data.

We are going to reason this problem in the following manner. A summary of the steps is given at the end of this discussion in the example 2.

First let us look at an expanded version of the above table.

	Maintained Healing		Total
	YES	NO	
Placebo	49	120	169
ACIPHEX 10 mg	119	40	159
ACIPHEX 20 mg	139	21	160
Total	307	181	488

Note that overall 307 out of 488 maintained healing in the above situation.

Therefore if there is no association between the maintenance of healing and type of the dosage, then in the placebo group we should expect  $169 \frac{307}{488} = \frac{169 \times 307}{488} \cong 106.31$  people to maintain healing.

Along the same lines, the expected number for

$$\text{ACIPHEX 10 mg is } \frac{159 \times 307}{488} \cong 100.02$$

$$\text{ACIPHEX 20 mg is } \frac{160 \times 307}{488} \cong 100.66$$

Note that the formula to compute the expected value of a cell is 
$$\frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

The table with the expected values inserted looks like

	Maintained Healing	
	YES	NO
Placebo	Observed=49 Expected=106.31	Observed=120 Expected=62.69
ACIPHEX 10 mg	Observed=119 Expected=100.02	Observed=40 Expected=58.98
ACIPHEX 20 mg	Observed=139 Expected=100.66	Observed=21 Expected=59.34

Note that we see difference between the observed and the expected in the rows for all the groups, specially the Placebos group.

Note that such differences will vary from sample to sample. To check the probability of a sample showing such a difference in the event null is true, we shall use the Chi Square Statistic, which, in this case is computed according to the following rule.

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \text{ at } (r-1)(c-1) \text{ degrees of freedom.}$$

Where r is the number of rows in the data table and c is the number of columns. The sample should be simple random and the expected frequency in at least 80% of the cells should be at least 5.

In the above example, note that

$$\begin{aligned} \chi^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(49 - 106.31)^2}{106.31} + \frac{(120 - 62.63)^2}{62.63} \\ &\quad + \frac{(119 - 100.02)^2}{100.02} + \frac{(40 - 58.98)^2}{58.98} \\ &\quad + \frac{(139 - 100.66)^2}{100.66} + \frac{(21 - 59.34)^2}{59.34} \\ &\cong 132.53 \end{aligned}$$

If you look at the table for  $\chi^2$  in this section, at 2 degrees of freedom, the largest value in the table is 10.59653 and the right tail probability for that value is 0.005. The calculated value of  $\chi^2$  for this sample is

132.53 which is greater than 10.59653, therefore, the P\_value is less than 0.005. Therefore we see a significant evidence of higher proportion of maintenance of healing in the Aciphex, specially the 20 mg group. Note that the Chi Square procedure will only give us an overall picture but we still have to look at the individual cells to see differences that are a mater of practical interest.



Example 2.

Let us work on an example that shows the summary of the above procedure.

The following (hypothetical) data shows the sales of vacation resort that were made by using the three different methods,

I: telephone sale call by a sales associate,

II: customer filling out an electronic form after reading an email about the promotion

III: customer calling a toll free number after reading the details via regular mail

Method	# making a purchase	# not making a purchase	Total
I	67	187	254
II	98	152	250
III	110	140	250
Total	275	479	754

To test, if there is significant difference in the rate of sale by the three different methods,

- i) compute the expected frequency of each cell.
- ii) compute the Chi Square Statistic.
- iii) Write your conclusion at 5% level of significance.

We can state the null and the alternative hypotheses in the following words.

$H_0$  : There is no association between the purchase decision of the customer and the method of approach

$H_A$  : There is an association between the two variables

- i) The expected frequencies are

Method	# making a purchase	# not making a purchase
I	$\frac{254 \times 275}{754} \cong 92.64$	161.36
II	$\frac{250 \times 275}{754} \cong 91.18$	158.82
III	$\frac{250 \times 275}{754} \cong 91.18$	158.82

ii)

$\chi^2$

=

$$\frac{(67 - 92.64)^2}{92.64} + \frac{(187 - 161.36)^2}{161.36} \\ + \frac{(98 - 91.18)^2}{91.18} + \frac{(152 - 158.82)^2}{158.82} \\ + \frac{(110 - 91.18)^2}{91.18} + \frac{(140 - 158.82)^2}{158.82}$$

$\cong 18.088$

iii)

From the table, we can conclude that the P\_value is less than 0.005.

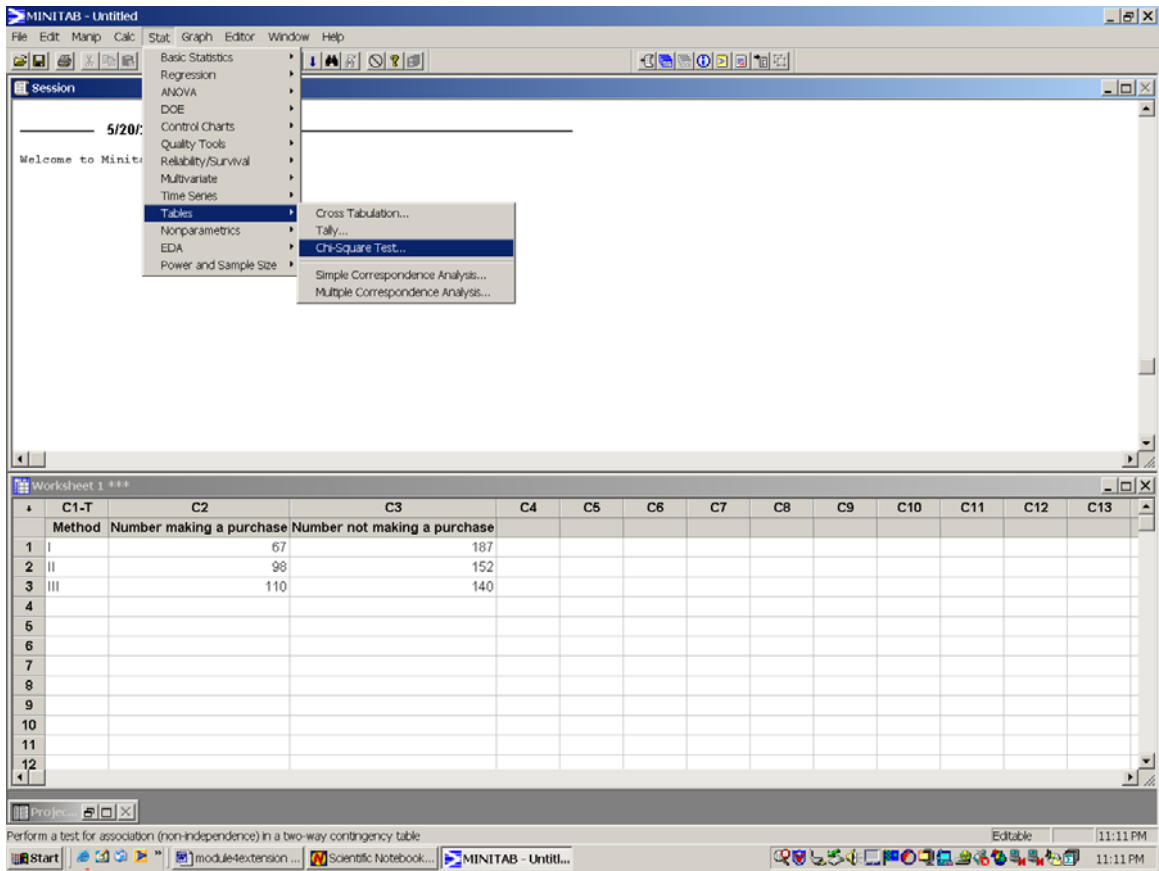
Therefore the data shows a significant evidence of the dependence between the two variables. This samples suggests that the direct telephone calls are least effective for sales.

### Using Technology:

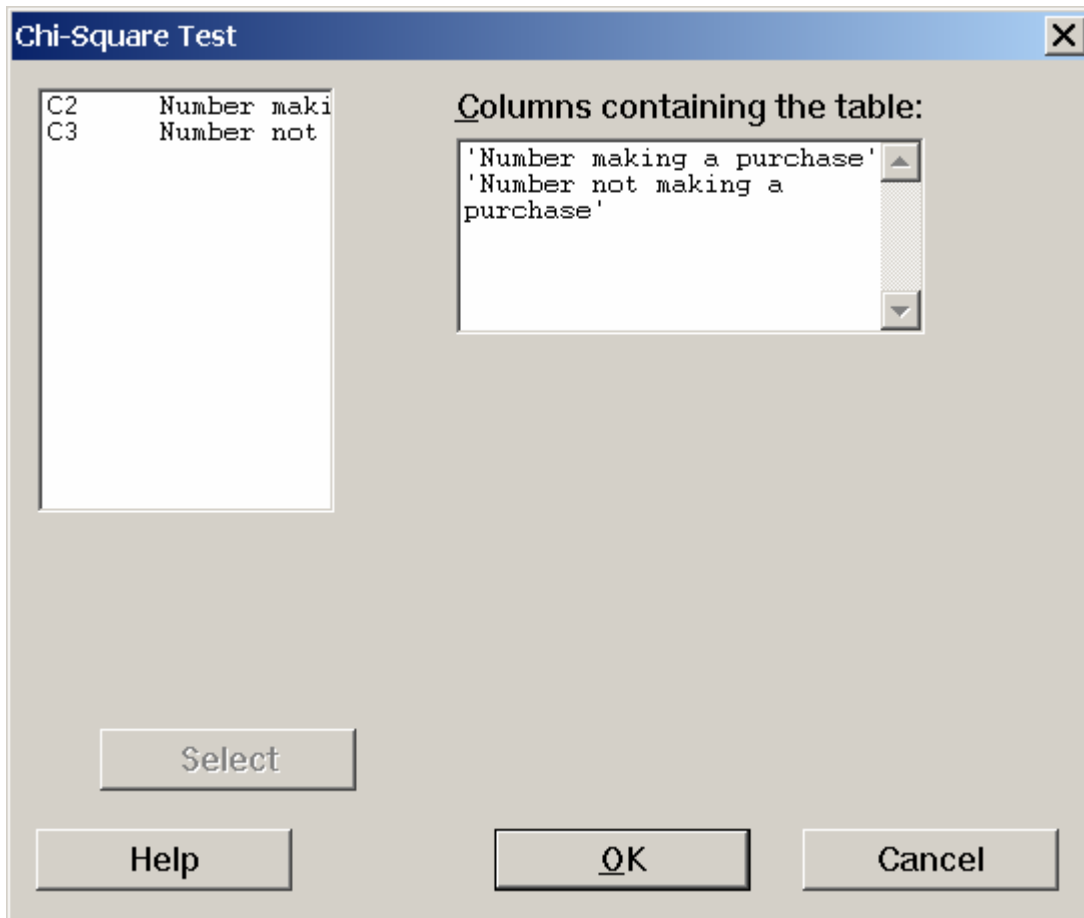
MINITAB:

To use MINITAB for the analysis of the table in the example 2,

Choose STAT→Tables as shown below



Then choose the variables in the dialogue box



Then click on OK to get the output

### Chi-Square Test: Number making a purchase, Number not making a purchase

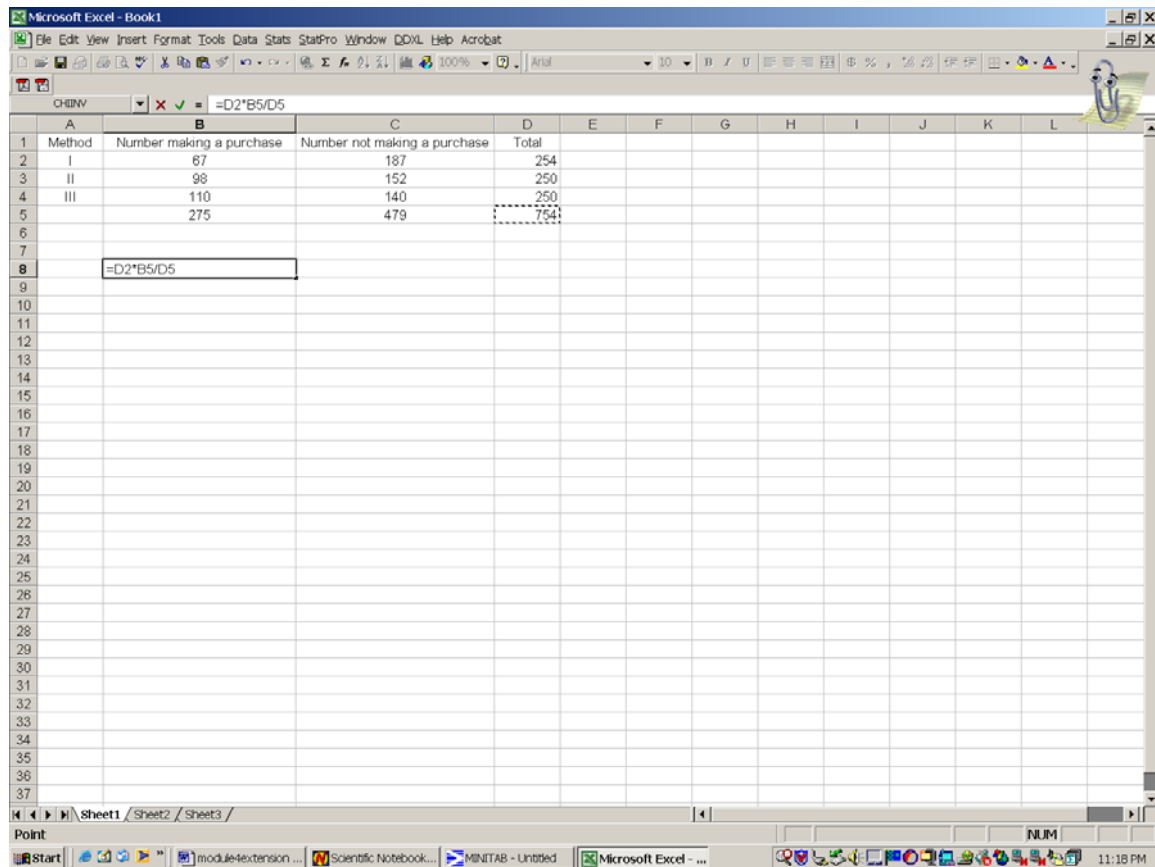
Expected counts are printed below observed counts

	Number m	Number n	Total
1	67	187	254
	92.64	161.36	
2	98	152	250
	91.18	158.82	
3	110	140	250
	91.18	158.82	
Total	275	479	754

Chi-Sq = 7.096 + 4.074 +  
 0.510 + 0.293 +  
 3.884 + 2.230 = 18.087  
 DF = 2, P-Value = 0.000

EXCEL:

First enter ( $=D2*B5/D5$ ) in the cell B8 in the following example to get the expected frequency



The screenshot shows a Microsoft Excel spreadsheet with the following data:

Method	Number making a purchase	Number not making a purchase	Total
I	67	187	254
II	98	152	250
III	110	140	250
	275	479	754

Cell B8 contains the formula  $=D2*B5/D5$ .

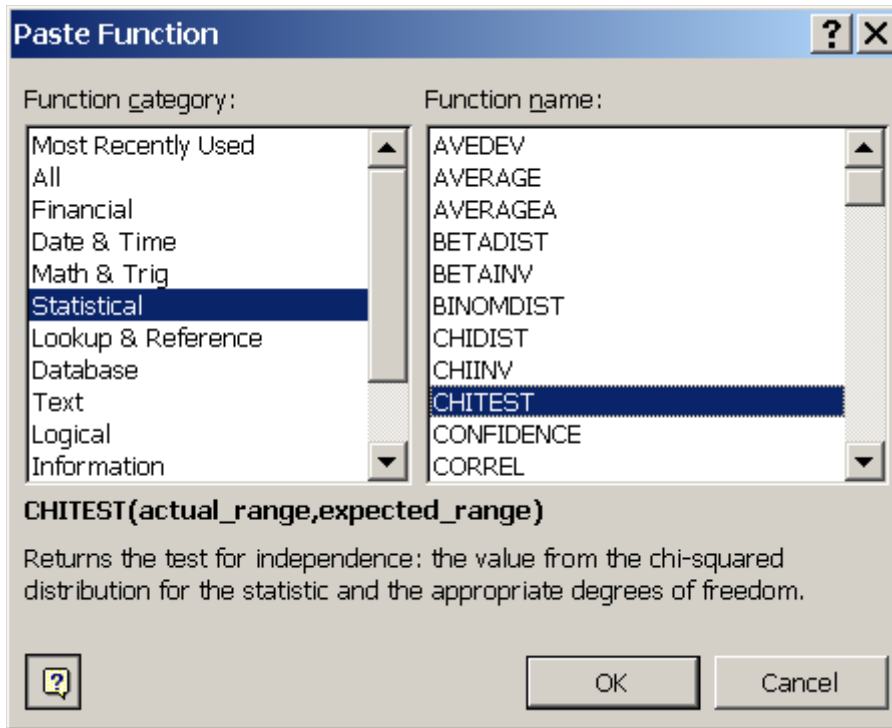
Then complete the table with expected frequencies

The screenshot shows a Microsoft Excel spreadsheet with the following data:

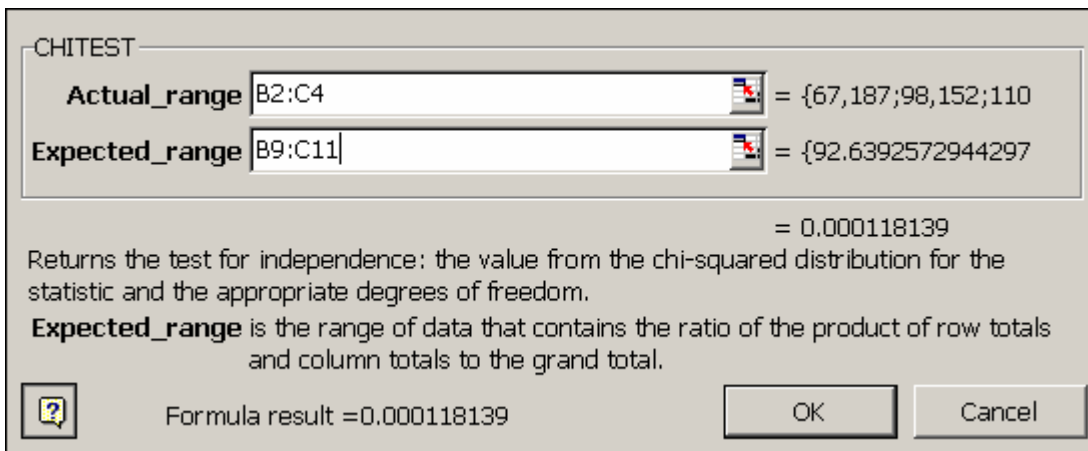
	A	B	C	D
1	Method	Number making a purchase	Number not making a purchase	Total
2	I	67	187	254
3	II	98	152	250
4	III	110	140	250
5		275	479	754
6				
7				
8		Expected		
9		92.63925729	161.3607427	
10		91.18037135	158.8196286	
11		91.18037135	158.8196286	
12				
13				
14				
15				
16				
17				
18				

The status bar at the bottom shows: Ready, Sum=3770, NUM, and the system clock is 11:35 PM.

Then choose f. → statistical



Fill in the dialogue box and say OK to see the P\_value.



Drills:

1. The following data is from the paper, "Female Participation in mathematical degree at English and Scottish Universities.", *Journal of Royal Statistical Society, Series A*, 155, 251-258, Table 7.

The data shows the gender of a student and the class (an analogue of GPA) that the students obtained.

	I	II(i)	II(ii)	III
Male	782	1390	1346	825
Female	343	643	793	364

Run a Chi Square Test to see if there is an association between the gender and the class of degree for such students.

Answer:

Brief:

### Chi-Square Test: I, II (i), II (ii), III

Expected counts are printed below observed counts

	I	II(i)	II(ii)	III	Total
1	782	1390	1346	825	4343
	753.30	1361.29	1432.27	796.15	
2	343	643	793	364	2143
	371.70	671.71	706.73	392.85	
Total	1125	2033	2139	1189	6486

Chi-Sq = 1.094 + 0.606 + 5.196 + 1.045 +  
 2.217 + 1.227 + 10.530 + 2.119 = 24.033  
 DF = 3, P-Value = 0.000

The test does show a significance, **still one must look at the descriptive statistics in the individual cells to see if there is a practical significance.**

2.

The following data has been taken from the text *Introductory Statistics* by Neil Weiss , sixth edition, page 674 published by Addison Wesley. The data is based on *The Lawyer Statistical Report*. The table shows 307 randomly selected U.S. lawyers by status in practice and the size of the city in which they practice.



	Size of City				Total	
	<25,000	250,000 to 499,999	500,000 or more			
Status In Practice	Government/Judicial	20	5	16	41	
	Private Practice	122	31	69	222	
	Salaries	19	7	18	44	
Total		161	43	103	307	

Does this data provide sufficient evidence to conclude that the size of the city and the status in practice are statistically dependent for US Lawyers?

Answer:

#### Brief Answer

Expected counts are printed below observed counts

	<25,000	250,000	500000 o	Total
1	20 21.50	5 5.74	16 13.76	41
2	122 116.42	31 31.09	69 74.48	222
3	19 23.07	7 6.16	18 14.76	44
Total	161	43	103	307

$$\text{Chi-Sq} = 0.105 + 0.096 + 0.366 + 0.267 + 0.000 + 0.403 + 0.720 + 0.114 + 0.710 = 2.781$$

DF = 4, P-Value = 0.595

Note that our null hypothesis here is that that the size of the city and the status in practice are statistically dependent for US Lawyers and the P\_value is quite large.

Therefore, we can not reject the null hypothesis and the data does not show sufficient evidence for dependence.

**Another Example to Review the Procedure:**

**Jim bought Tulips of three different brands and planted them in identical conditions.**

**The following is the data**

	<b>Blossom</b>	<b>Did not blossom</b>	<b>Total</b>
<b>Brand A</b>	<b>141</b>	<b>59</b>	<b>200</b>
<b>Brand B</b>	<b>129</b>	<b>41</b>	<b>170</b>
<b>Brand C</b>	<b>73</b>	<b>27</b>	<b>100</b>
	<b>243</b>	<b>127</b>	<b>470</b>

**To see if there is an association between the "proportion that blossom" and the brand at 5% level of significance.**

	<b>Blossom</b>	<b>Did not blossom</b>	<b>Blossom Rate</b>
<b>Brand A</b>	<b>141</b>	<b>59</b>	$\frac{141}{200} = 0.705$
<b>Brand B</b>	<b>129</b>	<b>41</b>	$\frac{129}{170} = 0.75882$
<b>Brand C</b>	<b>73</b>	<b>27</b>	$\frac{73}{100} = 0.73$

**Have to compute the test statistic,**

**Then find or estimate the P\_value**

**Compare the P\_value with 0.05 as level of significance**

**Reject the null (No association) if the P\_value is less than 0.05**

**How to compute the test statistic here, consult the posting "Chi Square Distribution" in the week 14**

**Discussion: If there is no association between the brand and the rate of blossom,**

**What % of tulips should blossom?**

**Observed counts**

	Blossom	Did not blossom	Total
Brand A	141	59	200
Brand B	129	41	170
Brand C	73	27	100
	343	127	470

Rate of blossom  $\frac{343}{470} = 0.72979$  if there is no association

In this event:

Expected Numbers will be

	Blossom	Did not blossom	Total
Brand A	$200 \left( \frac{343}{470} \right) = \frac{200 \times 343}{470} = 145.96$	$\frac{200 \times 127}{470} = 54.043$ OR take $200 - 145.96 = 54.04$	200
Brand B	$\frac{170 \times 343}{470} = 124.06$	$170 - 124.06 = 45.94$	170
Brand C	$\frac{100 \times 343}{470} = 72.979$	$100 - 72.979 = 27.021$	100
	343	127	470

Test Statistic:

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

$$\text{Brand A } \frac{(141 - 145.96)^2}{145.96} = 0.16855$$

$$\frac{(59 - 54.04)^2}{54.04} = 0.45525$$

$$\text{Brand B } \frac{(129 - 124.06)^2}{124.06} = 0.19671$$

$$\frac{(41 - 45.94)^2}{45.94} = 0.53121$$

$$\text{Brand C } \frac{(73 - 72.979)^2}{72.979} = 6.0428 \times 10^{-6}$$

$$\frac{(27 - 27.021)^2}{27.021} = 1.6321 \times 10^{-5}$$

$$\chi^2 = 0.16855 + 0.45525 + 0.19671 + 0.53121 + 6.0428 \times 10^{-6} + 1.6321 \times 10^{-5} = 1.3517$$

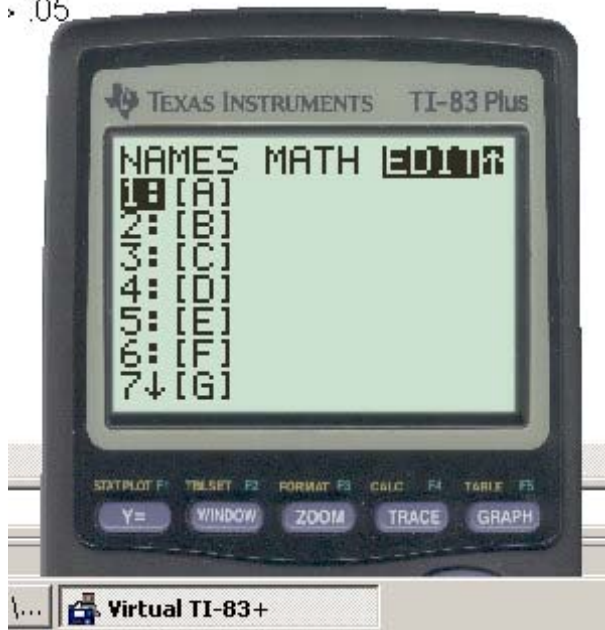
$$\text{Degrees of freedom } (\#rows - 1)(\#columns - 1) = (3 - 1)(2 - 1) = 2$$

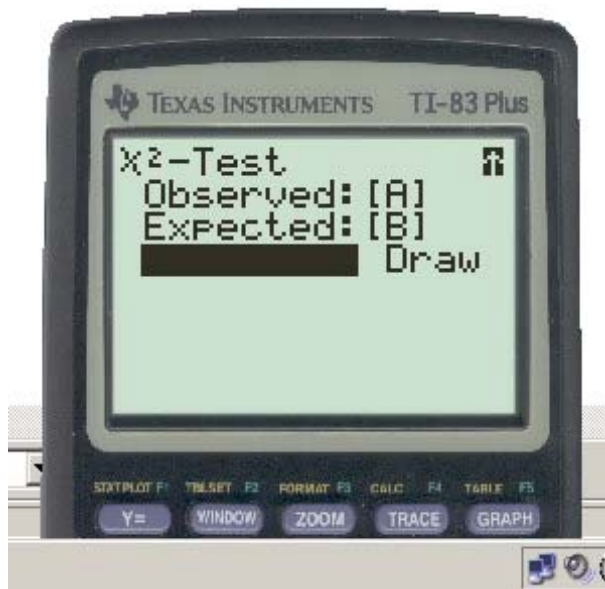
The use of table shows that the  $p\_value > .05$

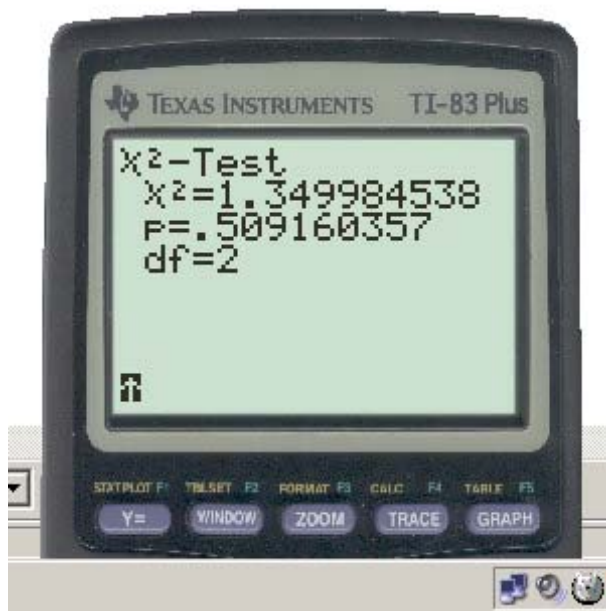
TI83plus

2nd Matrix

> .05







**Do not reject the null**

**No association**