Analysis of Two-Way Tables

IPS Chapter 9

- 9.1: Inference for Two-Way Tables
- 9.2: Formulas and Models for Two-Way Tables
- 9.3: Goodness of Fit

© 2012 W.H. Freeman and Company

Analysis of Two-Way Tables

9.1 Inference for Two-Way Tables

© 2012 W.H. Freeman and Company

Objectives

9.1 Inference for two-way tables

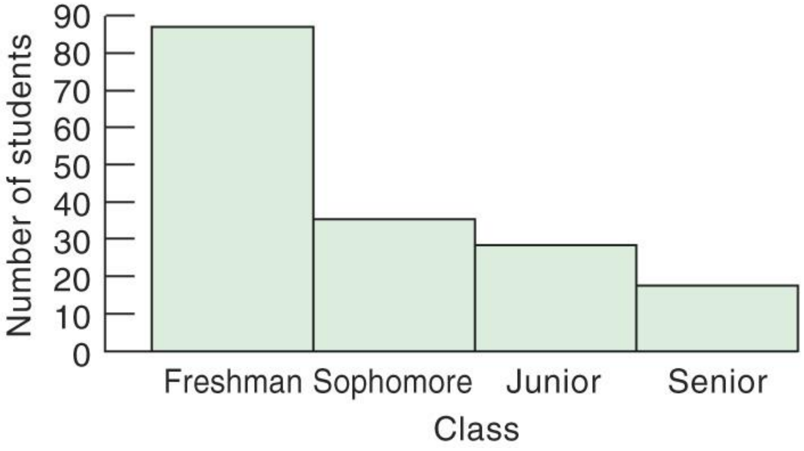
- The hypothesis: no association
- Expected cell counts
- The chi-square test
- The chi-square test and the *z* test
- Using software

Hypothesis: no association

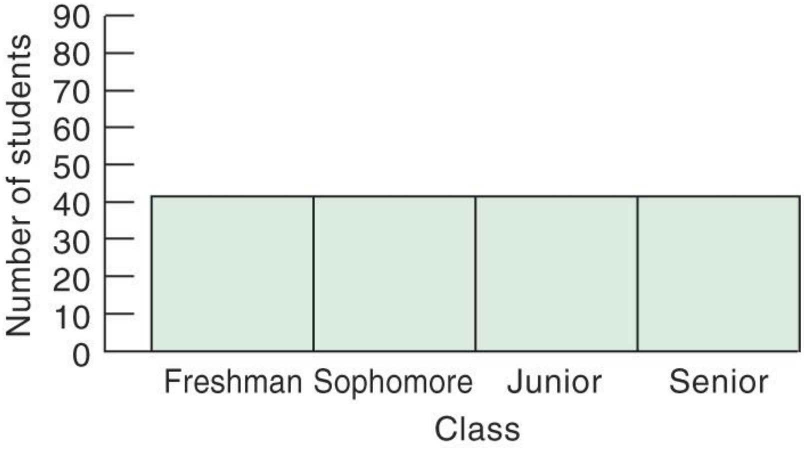
Again, we want to know if the differences in sample proportions are likely to have occurred just by chance due to random sampling.

We use the **chi-square** (χ^2) **test** to assess the null hypothesis of no relationship between the two categorical variables of a two-way table.

Observed Frequency Distribution for Students (*n* = 171)



Expected Frequency Distribution for Students (*n* = 171)



Expected cell counts

Two-way tables sort the data according to two categorical variables. We want to test the hypothesis that there is no relationship between these two categorical variables (H_0).

To test this hypothesis, we compare **actual counts** from the sample data with **expected counts,** given the null hypothesis of no relationship.

The expected count in any cell of a two-way table when H_0 is true is:

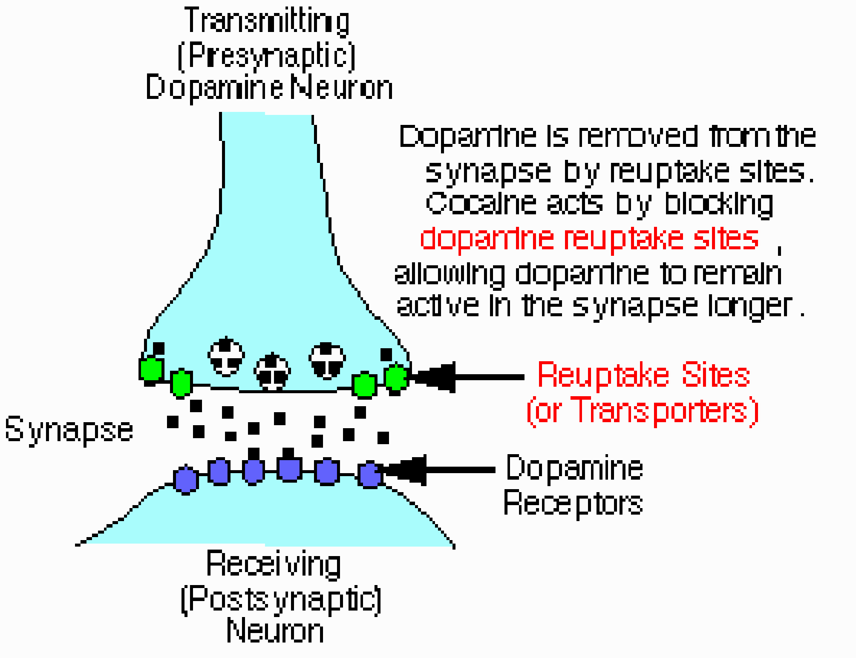
$$\boxed{\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}}$$

Cocaine addiction

Cocaine produces short-term feelings of physical and mental well being. To maintain the effect, the drug may have to be taken more frequently and at higher doses. After stopping use, users will feel tired, sleepy and **depressed.**

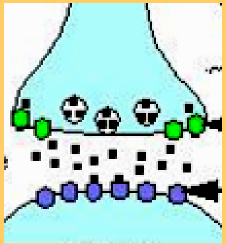
The pleasurable high followed by unpleasant after-effects encourage repeated compulsive use, which can easily lead to dependency.

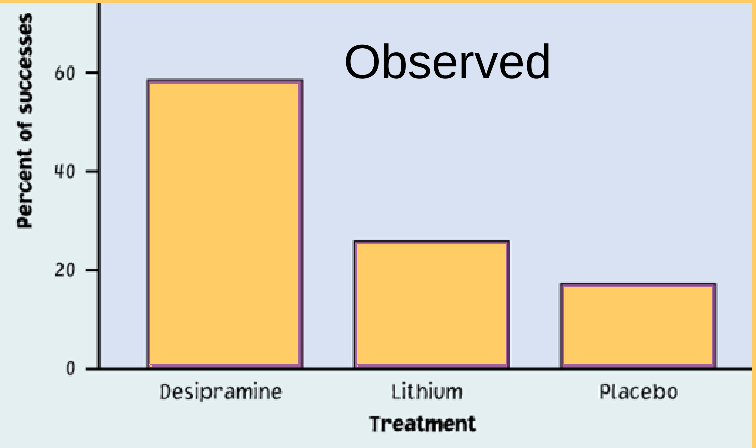
Desipramine is an **antidepressant** affecting the brain chemicals that may become unbalanced and cause depression. It was thus tested for recovery from cocaine addiction.



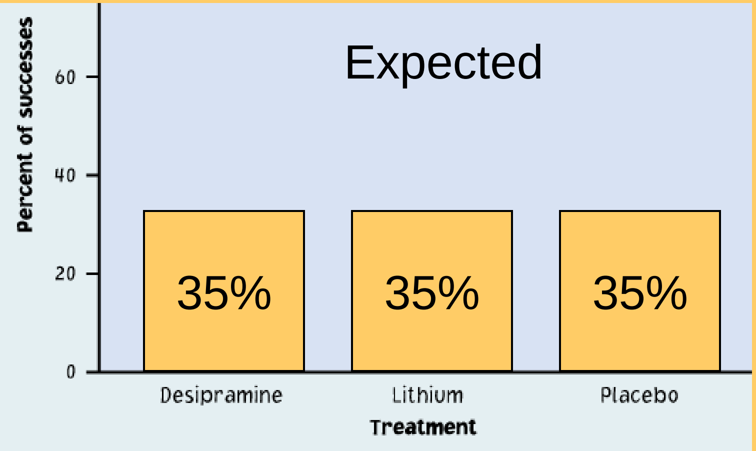
Treatment with desipramine was compared to a standard treatment (lithium, with strong anti-manic effects) and a placebo.

Cocaine addiction





	Relapse		Total
	No	Yes	
Desipramine	15	10	25
Lithium	7	19	26
Placebo	4	19	23
Total	26	48	74



Expected relapse counts

	No	Yes
Desipramine	(25)(26)/74 ≈ 8.78 25 x **0.35**	16.22 25 x 0.65
Lithium	9.14 26 x **0.35**	16.86 25 x 0.65
Placebo	8.08 23 x **0.35**	14.92 25 x 0.65

The chi-square test

The chi-square statistic (χ^2) is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts.

The formula for the χ^2 statistic is:

(summed over all $r \times c$ cells in the table)

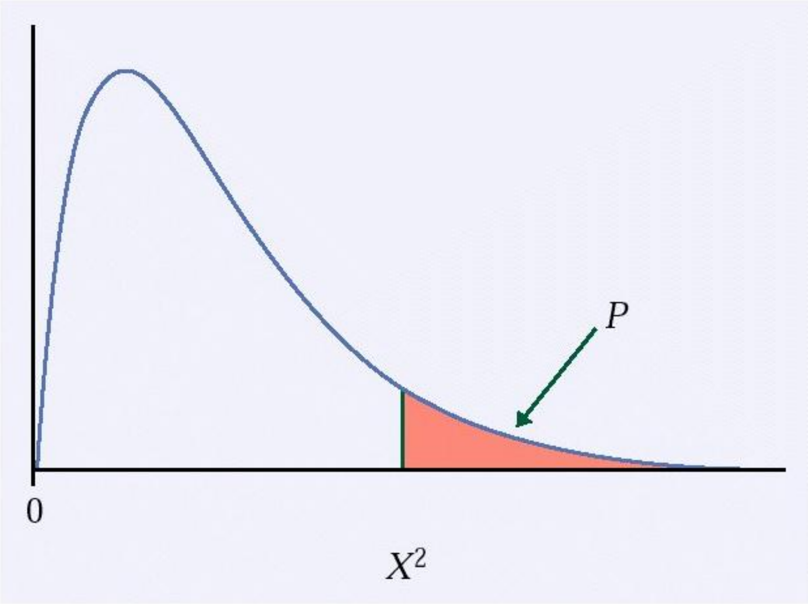
$$\boxed{\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}}$$

Large values for χ^2 represent strong deviations from the expected distribution under the H_0 and provide evidence against H_0.

However, since χ^2 is a sum, how large a χ^2 is required for statistical significance will depend on the number of comparisons made.

For the chi-square test, H_0 states that there is no association between the row and column variables in a two-way table. The alternative is that these variables are related.

If H_0 is true, the chi-square test has approximately a χ^2 **distribution with (r − 1)(c − 1) degrees of freedom**.



The P-value for the chi-square test is the area to the right of X^2 (the computed test value) under the χ^2 distribution with df (r−1)(c−1):

$P(\chi^2 \geq X^2)$.

When is it safe to use a χ^2 test?

We can safely use the chi-square test when:

- The samples are simple random samples (**SRS**)
- All individual **expected counts** are 1 or more (≥1)
- Average of **expected counts** are 5 or more

➔ *For a 2 x 2 table, all four expected counts should be 5 or more.*

Chi-square test vs. *z*-test for two proportions

When comparing only two proportions, such as in a 2x2 table where the columns represent counts of “success” and “failure,” we can test

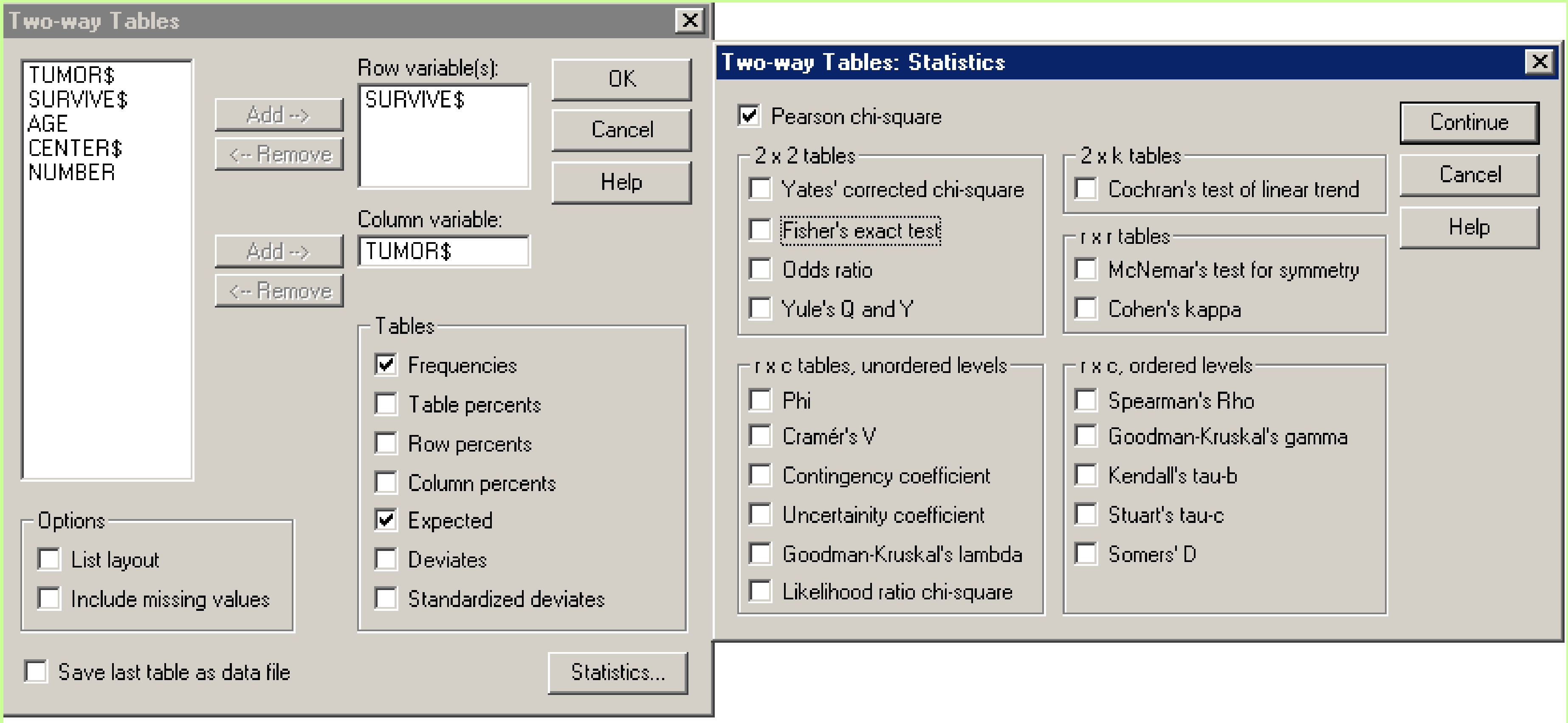
$$H_0: p_1 = p_2 \text{ vs. } H_a: p_1 \neq p_2$$

equally with a two-sided *z* test or with a chi-square test with 1 degree of freedom and get the same p-value. In fact, the two test statistics are related: $\chi^2 = (z)^2$.

Using software

- In Excel you have to do almost all the calculations for the chi-square test yourself, and it only gives you the p-value (not the component).

- This is **Systat**: Menu/Statistics/Crosstabs



Two-way Tables

TUMOR$
SURVIVE$
AGE
CENTER$
NUMBER

Row variable(s): SURVIVE$

Column variable: TUMOR$

Tables
☑ Frequencies
☐ Table percents
☐ Row percents
☐ Column percents
☑ Expected
☐ Deviates
☐ Standardized deviates

Options
☐ List layout
☐ Include missing values

☐ Save last table as data file

Two-way Tables: Statistics

☑ Pearson chi-square

2 x 2 tables
☐ Yates' corrected chi-square
☐ Fisher's exact test
☐ Odds ratio
☐ Yule's Q and Y

2 x k tables
☐ Cochran's test of linear trend

r x r tables
☐ McNemar's test for symmetry
☐ Cohen's kappa

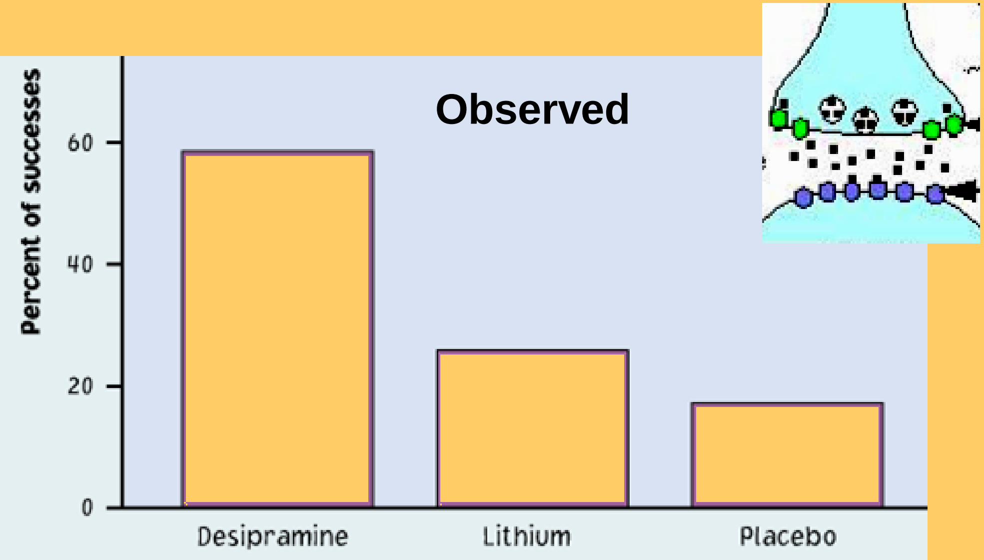
r x c tables, unordered levels
☐ Phi
☐ Cramér's V
☐ Contingency coefficient
☐ Uncertainity coefficient
☐ Goodman-Kruskal's lambda
☐ Likelihood ratio chi-square

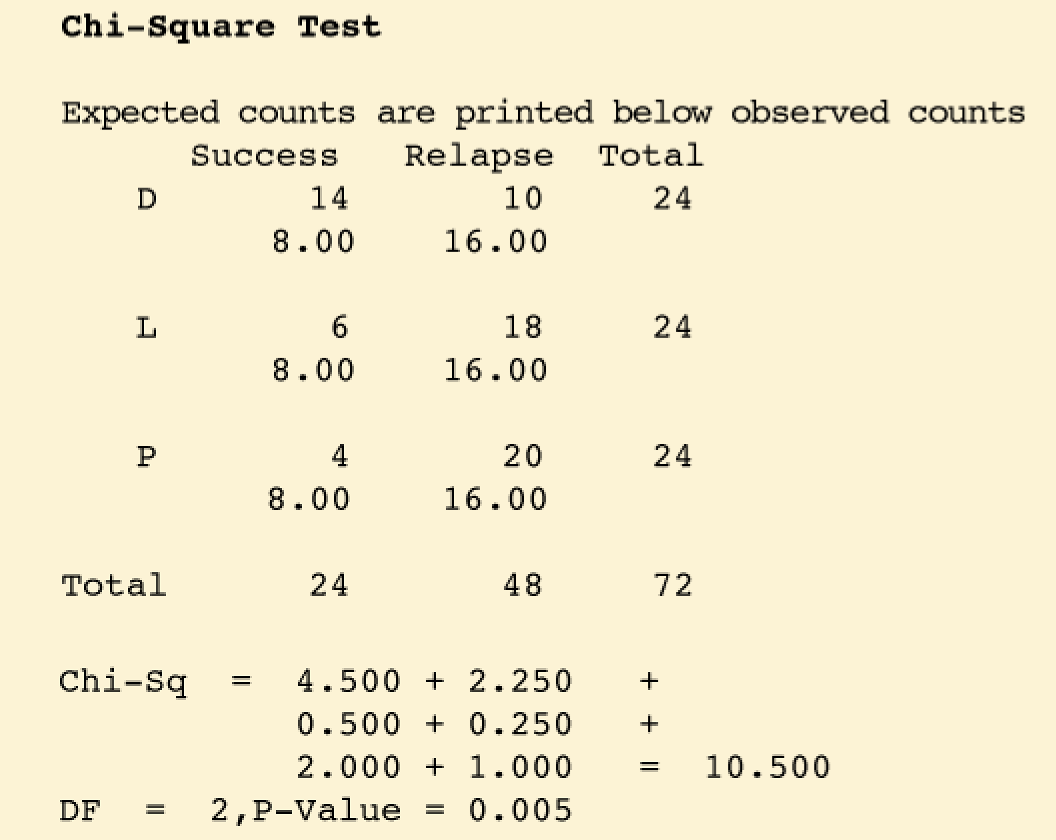
r x c, ordered levels
☐ Spearman's Rho
☐ Goodman-Kruskal's gamma
☐ Kendall's tau-b
☐ Stuart's tau-c
☐ Somers' D

Cocaine addiction

Minitab statistical software output

for the cocaine study





```
Chi-Square Test

Expected counts are printed below observed counts
       Success    Relapse   Total
    D       14         10      24
          8.00      16.00

    L        6         18      24
          8.00      16.00

    P        4         20      24
          8.00      16.00

Total       24         48      72

Chi-Sq  =  4.500 + 2.250   +
           0.500 + 0.250   +
           2.000 + 1.000   =  10.500
DF  =  2,P-Value = 0.005
```

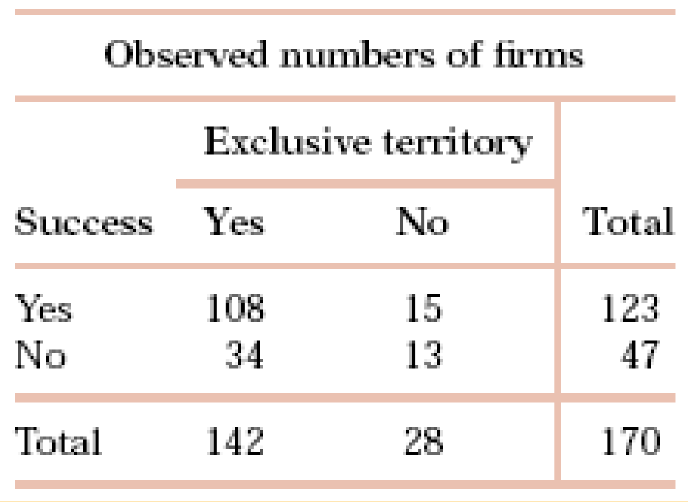
The p-value is 0.005 or half a percent. This is very significant.

We reject the null hypothesis of no association and conclude that there is a significant relationship between treatment *(desipramine, lithium, placebo)* and outcome *(relapse or not)*.

Successful firms

Franchise businesses are sometimes given an exclusive territory by contract. This means that the new outlet will not have to compete with other outlets of the same chain within its own territory. How does the presence of an exclusive-territory clause in the contract relate to the survival of the business?

A simple random sample of 170 new franchises recorded two categorical variables for each firm: (1) whether the firm was successful or not (based on economic criteria) and (2) whether or not the firm had an exclusive-territory contract.



Observed numbers of firms

	Exclusive territory		
Success	Yes	No	Total
Yes	108	15	123
No	34	13	47
Total	142	28	170

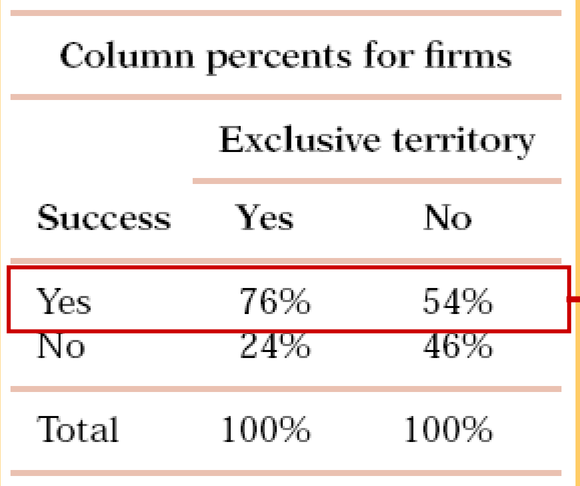
This is a 2x2 table (two levels for success, yes/no; two levels for exclusive territory, yes/no).

➔df = (2 − 1)(2 − 1) = 1

Successful firms

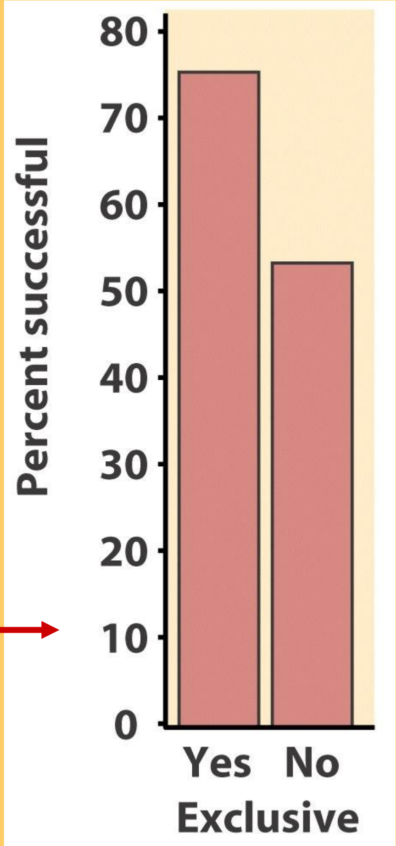
How does the presence of an exclusive-territory clause in the contract relate to the survival of the business?

To compare firms that have an exclusive territory with those that do not, we start by examining column percents (conditional distribution):



Column percents for firms

	Exclusive territory	
Success	Yes	No
Yes	76%	54%
No	24%	46%
Total	100%	100%



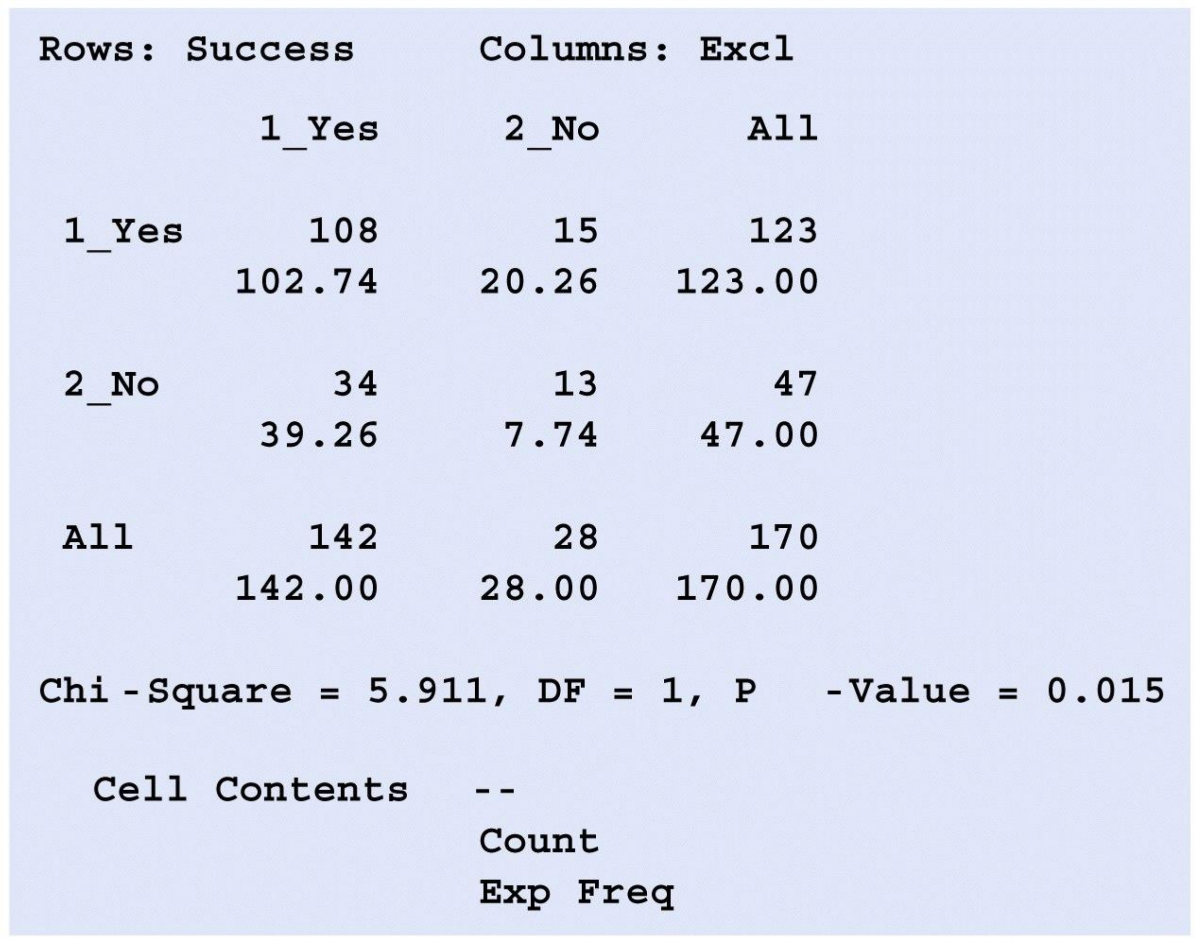
The difference between the percent of successes among the two types of firms is quite large. The chi-square test can tell us whether or not these differences can be plausibly attributed to chance (random sampling). Specifically, we will test:

H_0: No relationship between exclusive clause and success

H_a: There is some relationship between the two variables

Successful firms

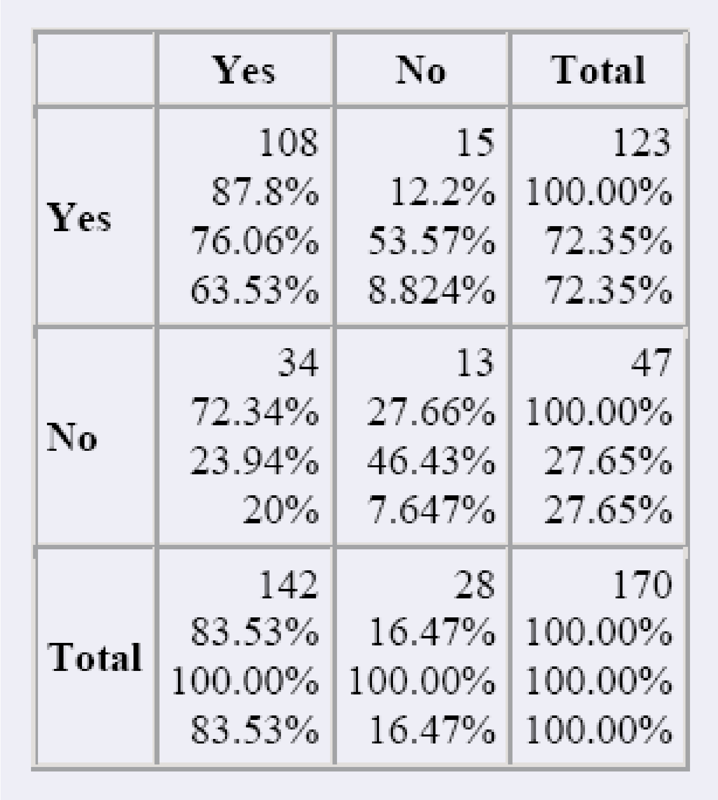
Here is the chi-square output from **Minitab**:



```
Rows: Success      Columns: Excl

          1_Yes      2_No       All

 1_Yes      108        15       123
         102.74     20.26    123.00

 2_No        34        13        47
          39.26      7.74     47.00

 All        142        28       170
         142.00     28.00    170.00

Chi -Square = 5.911, DF = 1, P   -Value = 0.015

  Cell Contents    --
                   Count
                   Exp Freq
```

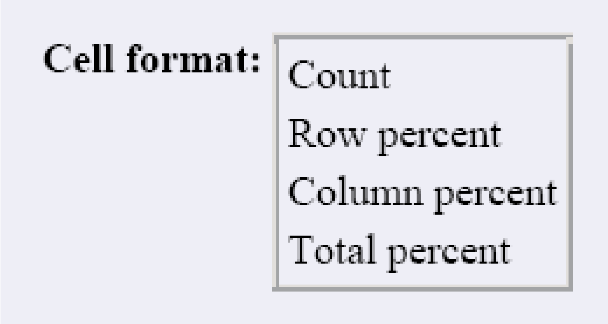
The p-value is significant at $\alpha = 5\%$ ($p = 1.5\%$) thus we reject H_0: we have found a significant relationship between an exclusive territory and the success of a franchised firm.

Successful firms



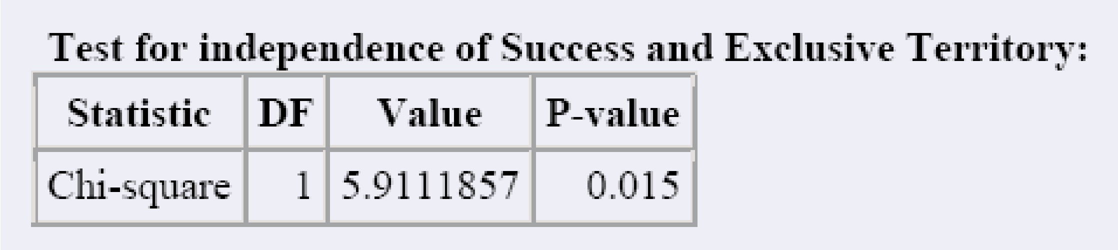
	Yes	No	Total
Yes	108 87.8% 76.06% 63.53%	15 12.2% 53.57% 8.824%	123 100.00% 72.35% 72.35%
No	34 72.34% 23.94% 20%	13 27.66% 46.43% 7.647%	47 100.00% 27.65% 27.65%
Total	142 83.53% 100.00% 83.53%	28 16.47% 100.00% 16.47%	170 100.00% 100.00% 100.00%

Computer output
using **Crunch It!**



Cell format:

Count
Row percent
Column percent
Total percent



Test for independence of Success and Exclusive Territory:

Statistic	DF	Value	P-value
Chi-square	1	5.9111857	0.015

Analysis of Two-Way Tables

9.2 Formulas and Models for Two-Way Tables

9.3 Goodness of Fit

© 2012 W.H. Freeman and Company

Objectives

9.2 Formulas and models for two-way tables
9.3 Goodness of fit

Computations for two-way tables

- Computing conditional distributions
- Computing expected cell counts
- Computing the chi-square statistic
- Finding the p-value with Table F

Models for two-way tables

- Comparing several populations
- Testing for independence

Goodness of fit

Computations for two-way tables

When analyzing relationships between two categorical variables, follow this procedure:

1. Calculate descriptive statistics that convey the important information in the table—usually column or row percents.

2. Find the expected counts and use them to compute the X^2 statistic.

3. Compare your X^2 statistic to the chi-square critical values from Table F to find the approximate *P*-value for your test.

4. Draw a conclusion about the association between the row and column variables.

Computing conditional distributions

The calculated percents within a two-way table represent the **conditional distributions** describing the “relationship” between both variables.

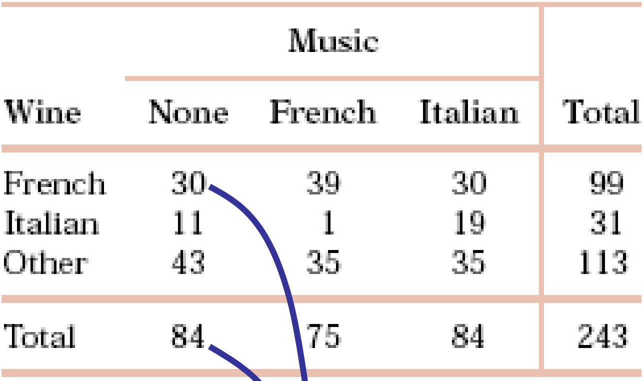
For every two-way table, there are two sets of possible conditional distributions (column percents or row percents).

> For column percents, divide each cell count by the column total. The sum of the percents in each column should be 100, except for possible small roundoff errors.

When one variable is clearly explanatory, it makes sense to describe the relationship by comparing the conditional distributions of the response variable for each value (level) of the explanatory variable.

Music and wine purchase decision

What is the relationship between type of music played in supermarkets and type of wine purchased?



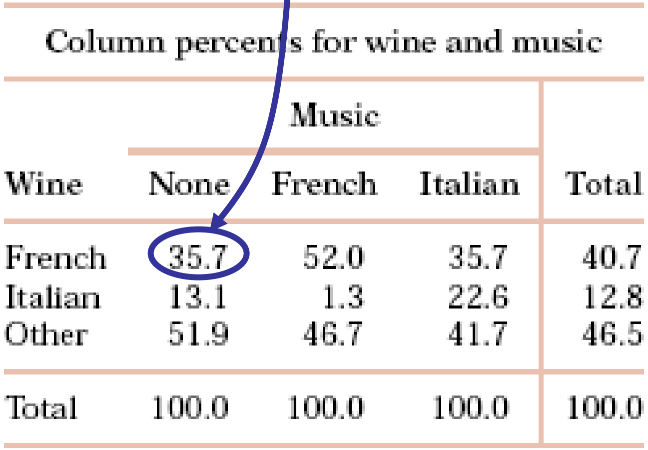
	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

We want to compare the conditional distributions of the response variable (wine purchased) for each value of the explanatory variable (music played). Therefore, we calculate column percents.

Calculations: When no music was played, there were 84 bottles of wine sold. Of these, 30 were French wine. 30/84 = 0.357 ➔ 35.7% of the wine sold was French when no music was played.

$\frac{30}{84} = 35.7\%$

$= \frac{\text{cell total}}{\text{column total}}$



Column percents for wine and music

	Music			
Wine	None	French	Italian	Total
French	35.7	52.0	35.7	40.7
Italian	13.1	1.3	22.6	12.8
Other	51.9	46.7	41.7	46.5
Total	100.0	100.0	100.0	100.0

We calculate the column conditional percents similarly for each of the nine cells in the table:

Computing expected counts

When testing the null hypothesis that there is no relationship between both categorical variables of a two-way table, we compare **actual counts** from the sample data with **expected counts** given H_0.

The expected count in any cell of a two-way table when H_0 is true is:

$$\boxed{\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}}$$

Although in real life counts must be whole numbers, an expected count need not be. The expected count is the mean over many repetitions of the study, assuming no relationship.

Music and wine purchase decision

The null hypothesis is that there is no relationship between music and wine sales. The alternative is that these two variables are related.

	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

What is the expected count in the upper-left cell of the two-way table, under H_0?

Column total 84: Number of bottles sold without music

Row total 99: Number of bottles of French wine sold

Table total 243: all bottles sold during the study

This expected cell count is thus

(84)(99) / 243 = 34.222

Nine similar calculations produce the table of expected counts:

Expected counts for wine and music

	Music			
Wine	None	French	Italian	Total
French	34.222	30.556	34.222	99.000
Italian	10.716	9.568	10.716	31.000
Other	39.062	34.877	39.062	113.001
Total	84.000	75.001	84.000	243.001

Computing the chi-square statistic

The chi-square statistic (χ^2) is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts.

The formula for the χ^2 statistic is:

(summed over all *r* x *c* cells in the table)

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

Tip: First, calculate the χ^2 components, (observed-expected)^2/expected, for each cell of the table, and then sum them up to arrive at the χ^2 statistic.

Music and wine purchase decision

H_0: No relationship between music and wine H_a: Music and wine are related

Observed counts

	Music		
Wine	None	French	Italian
French	30	39	30
Italian	11	1	19
Other	43	35	35

Expected counts

	Music		
Wine	None	French	Italian
French	34.222	30.556	34.222
Italian	10.716	9.568	10.716
Other	39.062	34.877	39.062

We calculate nine X^2 components and sum them to produce the X^2 statistic:

$$X^2 = \sum \frac{(observed - \exp ected)^2}{\exp ected}$$

$$= \frac{(30-34.222)^2}{34.222} + \frac{(39-30.556)^2}{34.222} + \frac{(30-34.222)^2}{34.222}$$

$$+ \frac{(11-10.716)^2}{10.716} + \frac{(1-9.568)^2}{9.568} + \frac{(19-10.716)^2}{10.716}$$

$$+ \frac{(43-39.062)^2}{39.062} + \frac{(35-34.877)^2}{34.877} + \frac{(35-39.062)^2}{39.062}$$

$$= 0.5209 + 2.3337 + 0.5209 + 0.0075 + 7.6724$$

$$+ 6.4038 + 0.3971 + 0.0004 + 0.4223$$

$$= 18.28$$

Finding the p-value with Table F

The χ^2 distributions are a family of distributions that can take only positive values, are skewed to the right, and are described by a specific degrees of freedom.

Table F gives upper critical values for many χ^2 distributions.

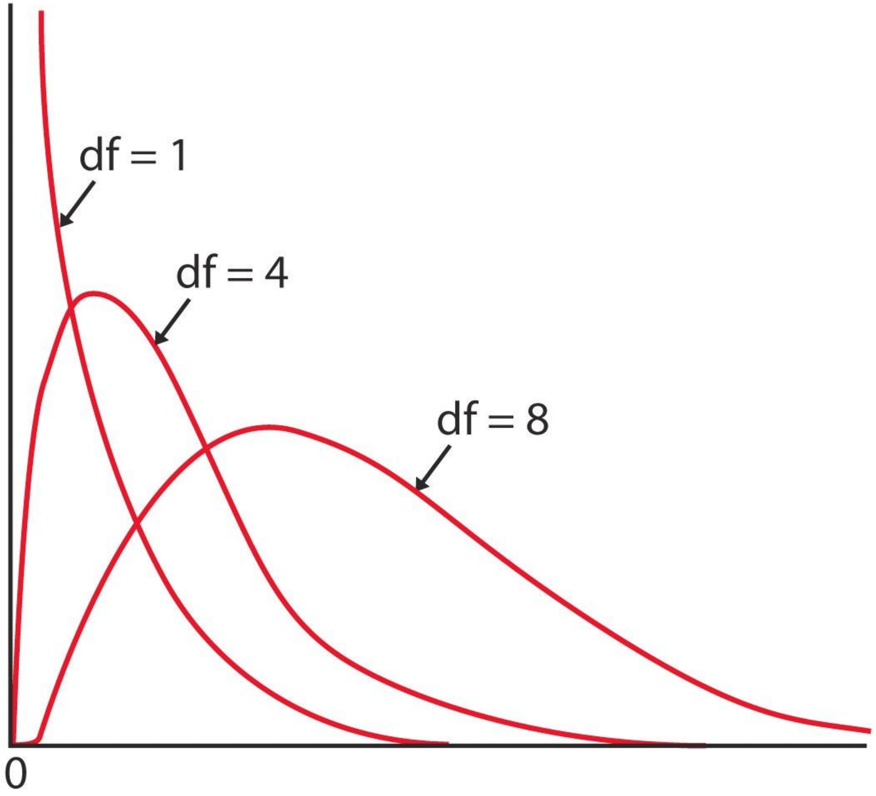
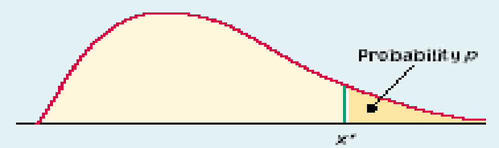


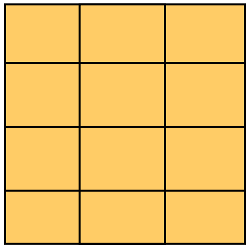
Table F



df = (r−1)(c−1)

					p						
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.70
80	88.13	90.41	93.11	96.58	101.90	106.60	108.10	112.30	116.30	120.10	124.80	128.30
100	109.10	111.70	114.70	118.50	124.30	129.60	131.10	135.80	140.20	144.30	149.40	153.20

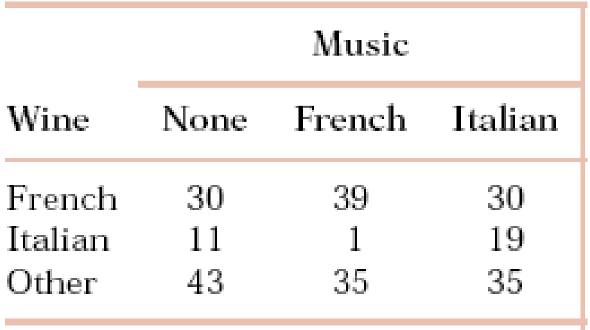
Ex: In a 4x3 table, df = 3*2 = 6



If $X^2 = 16.1$, the p-value is between 0.01−0.02.

Music and wine purchase decision

H_0: No relationship between music and wine H_a: Music and wine are related



	Music		
Wine	None	French	Italian
French	30	39	30
Italian	11	1	19
Other	43	35	35

We found that the χ^2 statistic under H_0 is 18.28.

The two-way table has a 3x3 design (3 levels of music and 3 levels of wine). Thus, the degrees of freedom for the χ^2 distribution for this test is:

$$(r - 1)(c - 1) = (3 - 1)(3 - 1) = 4$$

	p											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00

$$16.42 < X^2 = 18.28 < 18.47$$

$0.0025 > \text{p-value} > 0.001$ ➔ very significant

There is a significant relationship between the type of music played and wine purchases in supermarkets.

Interpreting the χ^2 output

- The values summed to make up χ^2 are called the **χ^2 components.**

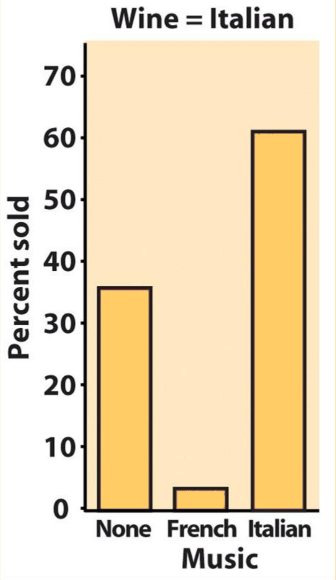
When the test is statistically significant, the **largest components** point to the conditions most different from the expectations based on H_0.

Music and wine purchase decision

χ^2 **components**	Music		
Wine	None	French	Italian
French	0.5209	2.3337	0.5209
Italian	0.0075	7.6724	6.4038
Other	0.3971	0.0004	0.4223

Two chi-square components contribute most to the χ^2 total ➔ the largest effect is for sales of Italian wine, which are strongly affected by Italian and French music.

Actual proportions show that Italian music helps sales of Italian wine, but French music hinders it.



Models for two-way tables

The chi-square test is an overall technique for comparing any number of population proportions, testing for evidence of a relationship between two categorical variables. We can either:

- **Compare several populations:** Randomly select several SRSs each from a different population (or from a population subjected to different treatments) → experimental study.
- **Test for independence:** Take one SRS and classify the individuals in the sample according to two categorical variables (attribute or condition) → observational study, historical design.

Both models use the χ^2 test to test the hypothesis of *no relationship*.

Comparing several populations: the first model

Select independent SRSs from each of *c* populations, of sizes $n_1, n_2, . . . , n_c$. Classify each individual in a sample according to a categorical response variable with *r* possible values. There are *c* different probability distributions, one for each population.

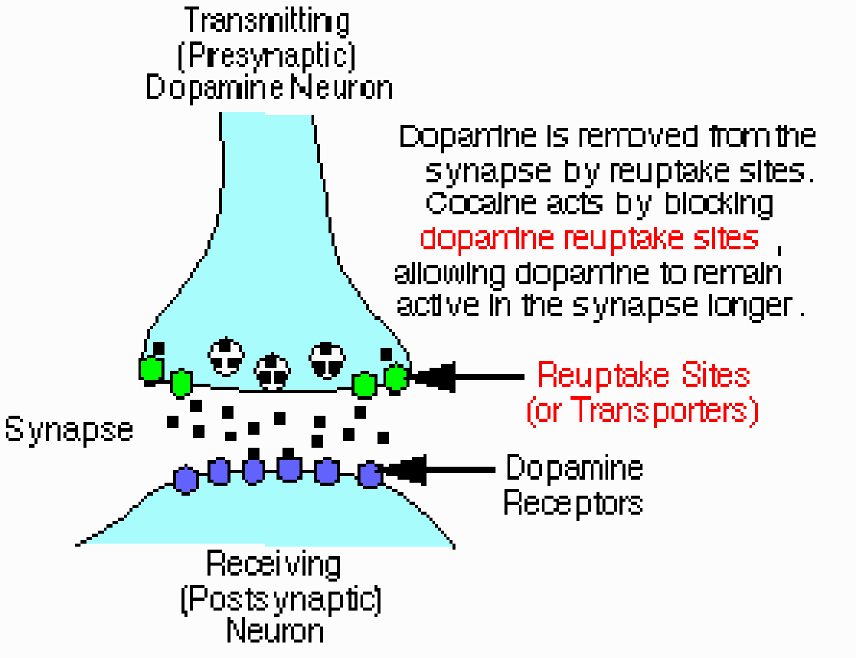
The null hypothesis is that the distributions of the response variable are the same in all *c* populations. The alternative hypothesis says that these *c* distributions are not all the same.

Cocaine addiction

Cocaine produces short-term feelings of physical and mental well being. To maintain the effect, the drug may have to be taken more frequently and at higher doses. After stopping use, users will feel tired, sleepy, and **depressed.**

The pleasurable high followed by unpleasant after-effects encourage repeated compulsive use, which can easily lead to dependency.

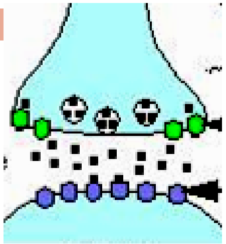
We compare treatment with an anti-depressant (desipramine), a standard treatment (lithium), and a placebo.

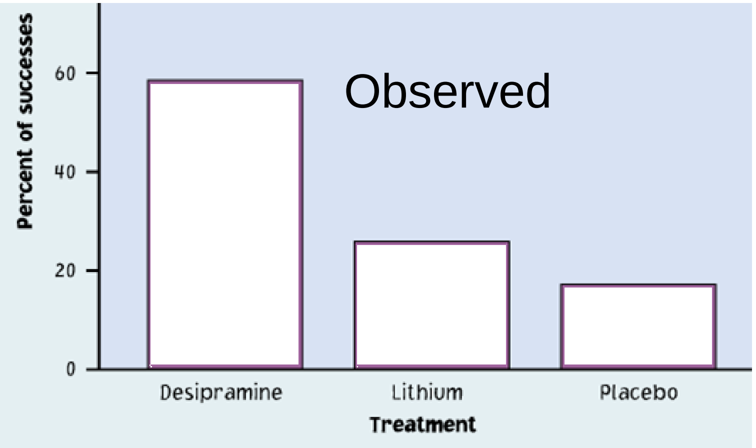


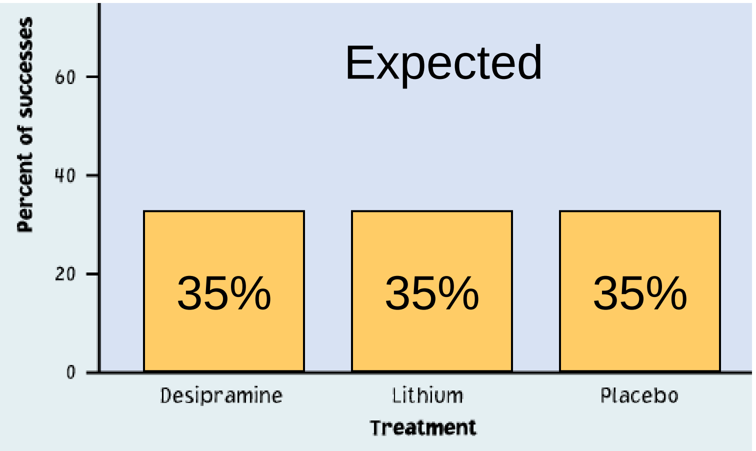
Population 1: Antidepressant treatment (desipramine)
Population 2: Standard treatment (lithium)
Population 3: Placebo (“sugar pill”)

Cocaine addiction

H_0: The proportions of success (no relapse) are the same in all three populations.







	Relapse		Total
	No	Yes	
Desipramine	15	10	25
Lithium	7	19	26
Placebo	4	19	23
Total	26	48	74

Expected relapse counts

	No	Yes
Desipramine	(25)(26)/74 ≈ 8.78 25 x **0.35**	16.22 25 x 0.65
Lithium	9.14 26 x **0.35**	16.86 25 x 0.65
Placebo	8.08 23 x **0.35**	14.92 25 x 0.65

Cocaine addiction

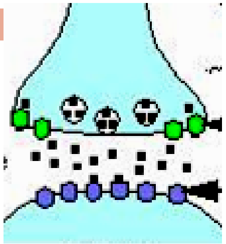


Table of counts:
“actual / **expected**,” with
three rows and two
columns:

	No relapse	Relapse
Desipramine	15 **8.78**	10 **16.22**
Lithium	7 **9.14**	19 **16.86**
Placebo	4 **8.08**	19 **14.92**

df = (3−1)*(2−1) = 2

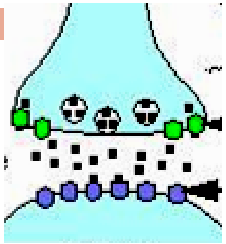
$$X^2 = \frac{(15-8.78)^2}{8.78} + \frac{(10-16.22)^2}{16.22} + \frac{(7-9.14)^2}{9.14} + \frac{(19-16.86)^2}{16.86} + \frac{(4-8.08)^2}{8.08} + \frac{(19-14.92)^2}{14.92} = 10.74$$

$\Longrightarrow$

χ^2 components:

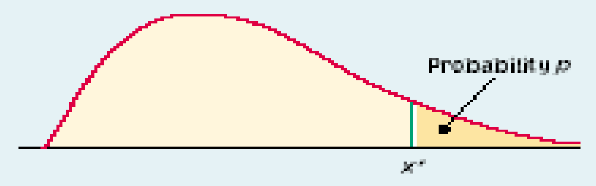
4.41	2.39
0.50	0.27
2.06	1.12

Cocaine addiction: Table F



H_0: The proportions of success (no relapse)

are the same in all three populations.



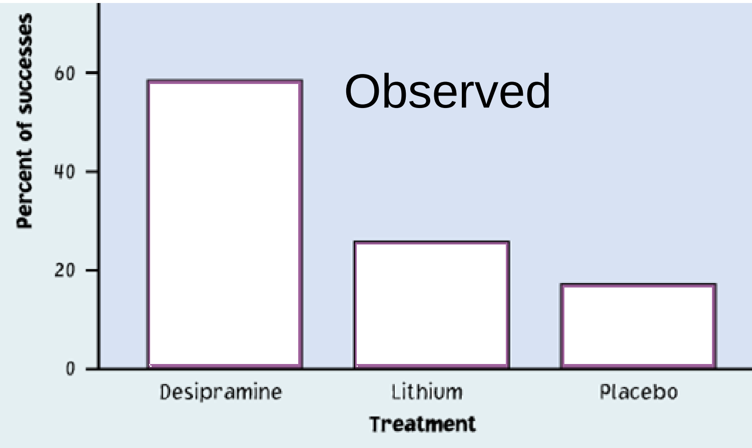
df	p											
	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	★ 11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11

$X^2 = 10.71$ and df = 2

$10.60 < X^2 < 11.98$ ➔ $0.005 < p < 0.0025$ ➔ reject the H_0

➔The proportions of success are not the same in all three populations (Desipramine, Lithium, Placebo).

Desipramine is a more successful treatment →→→



Testing for independence: the second model

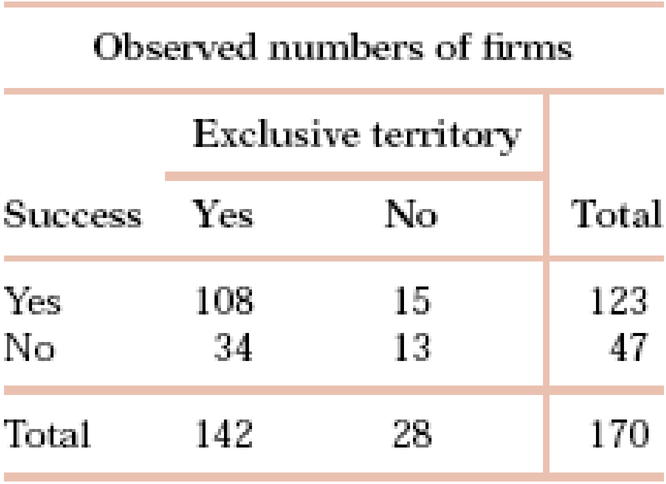
We now have a *single* sample from a *single* population. For each individual in this SRS of size *n* we measure two categorical variables. The results are then summarized in a two-way table.

The null hypothesis is that the row and column variables are independent. The alternative hypothesis is that the row and column variables are dependent.

Successful firms

How does the presence of an exclusive-territory clause in the contract for a franchise business relate to the survival of that business?

A random sample of 170 new franchises recorded two categorical variables for each firm: (1) whether the firm was successful or not (based on economic criteria) and (2) whether or not the firm had an exclusive-territory contract.



Observed numbers of firms

	Exclusive territory		
Success	Yes	No	Total
Yes	108	15	123
No	34	13	47
Total	142	28	170

This is a 2x2, two-way table

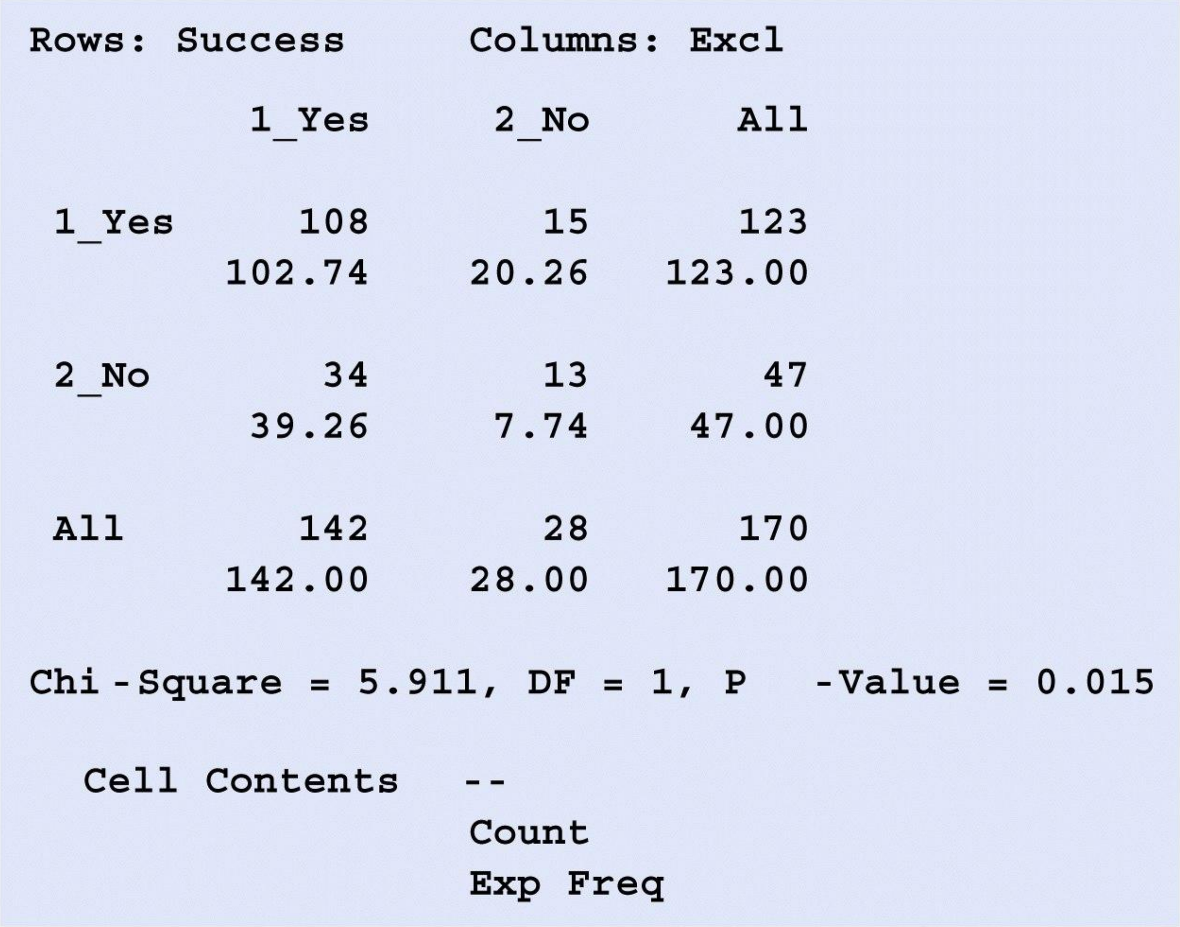
(2 levels for business success, yes/no,

2 levels for exclusive territory, yes/no).

We will test H_0: The variables exclusive clause and success are independent.

Successful firms

Computer output for the chi-square test using **Minitab**:



```
Rows: Success      Columns: Excl

             1_Yes      2_No      All

 1_Yes         108        15      123
            102.74     20.26   123.00

 2_No           34        13       47
             39.26      7.74    47.00

 All           142        28      170
            142.00     28.00   170.00

Chi-Square = 5.911, DF = 1, P  -Value = 0.015

  Cell Contents   --
                  Count
                  Exp Freq
```

The p-value is significant at *α* 5% thus we reject H_0:

The existence of an exclusive territory clause in a franchise's contract and the success of that franchise are not independent variables.

Parental smoking

Does parental smoking influence the incidence of smoking in children when they reach high school? Randomly chosen high school students were asked whether they smoked (columns) and whether their parents smoked (rows).

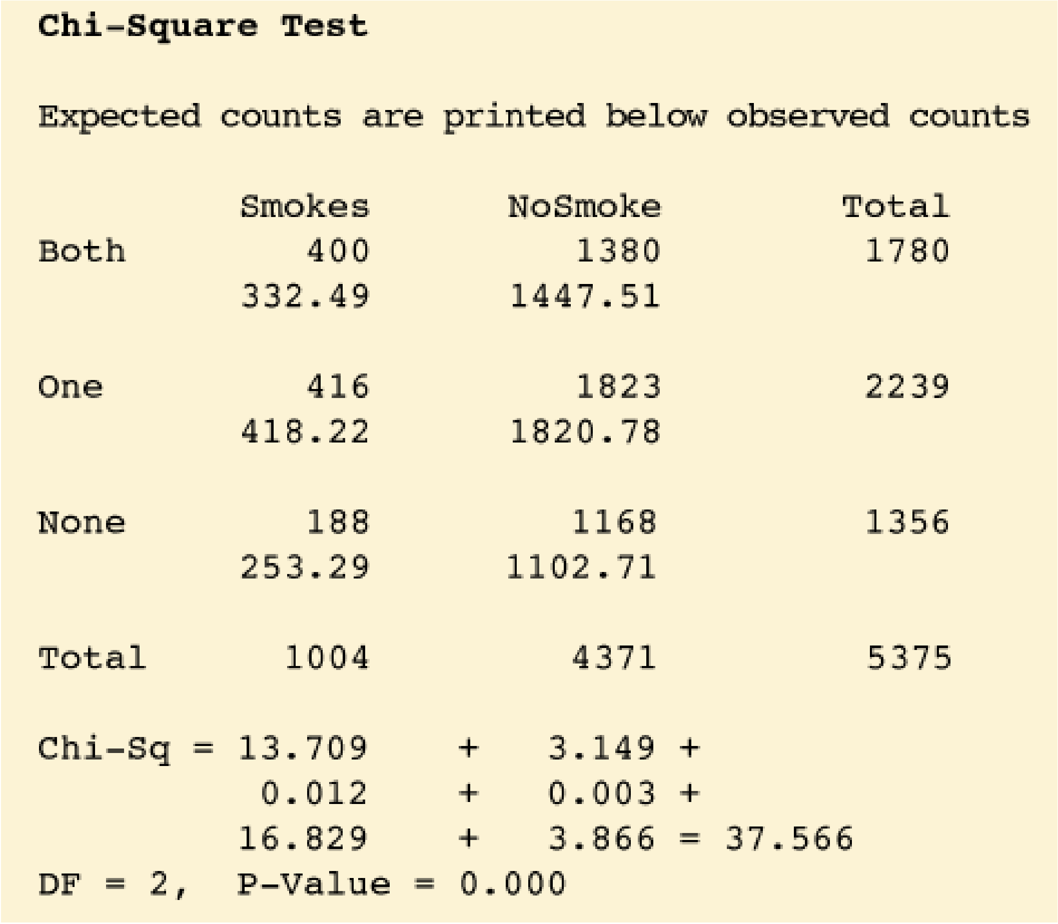
Examine the computer output for the chi-square test performed on these data. What does it tell you?

Sample size?

Hypotheses?

Are data ok for χ^2 test?

Interpretation?



```
Chi-Square Test

Expected counts are printed below observed counts

           Smokes     NoSmoke      Total
Both          400        1380       1780
           332.49     1447.51

One           416        1823       2239
           418.22     1820.78

None          188        1168       1356
           253.29     1102.71

Total        1004        4371       5375

Chi-Sq = 13.709    +   3.149 +
          0.012    +   0.003 +
         16.829    +   3.866 = 37.566
DF = 2,  P-Value = 0.000
```

Testing for goodness of fit

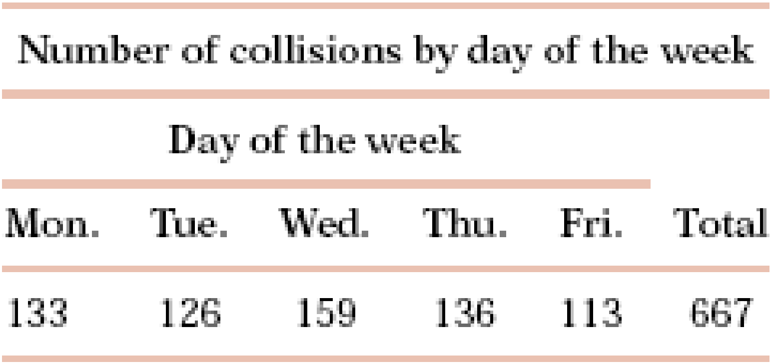
We have used the chi-square test as the tool to compare two or more distributions all based on sample data.

We now consider a slight variation on this scenario where only one of the distributions is known (our sample data observations) and we want to compare it with a hypothesized distribution.

- Data for n observations on a categorical variable with k possible outcomes are summarized as observed counts, $n_1, n_2, . . . , n_k$.
- The null hypothesis specifies probabilities $p_1, p_2, . . . , p_k$ for the possible outcomes.

Car accidents and day of the week

A study of 667 drivers who were using a cell phone when they were involved in a collision on a weekday examined the relationship between these accidents and the day of the week.



Number of collisions by day of the week

Day of the week					
Mon.	Tue.	Wed.	Thu.	Fri.	Total
133	126	159	136	113	667

Are the accidents equally likely to occur on any day of the working week?

H_0 specifies that all 5 days are equally likely for car accidents ➔ each $p_i = 1/5$.

The chi-square goodness of fit test

Data for n observations on a categorical variable with k possible outcomes are summarized as observed counts, $n_1, n_2, \ldots, n_k$ in k cells.

H_0 specifies probabilities $p_1, p_2, \ldots, p_k$ for the possible outcomes.

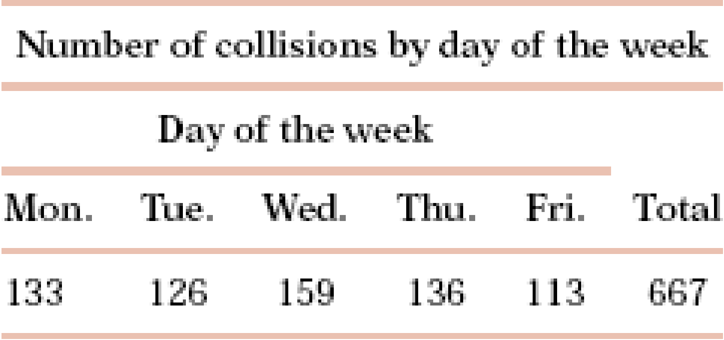
For each cell, multiply the total number of observations n by the specified probability p_i:

$$\text{expected count} = np_i$$

The **chi-square statistic** follows the chi-square distribution with **$k - 1$ degrees of freedom**:

$$X^2 = \sum \frac{(observed - \exp ected)^2}{\exp ected}$$

Car accidents and day of the week



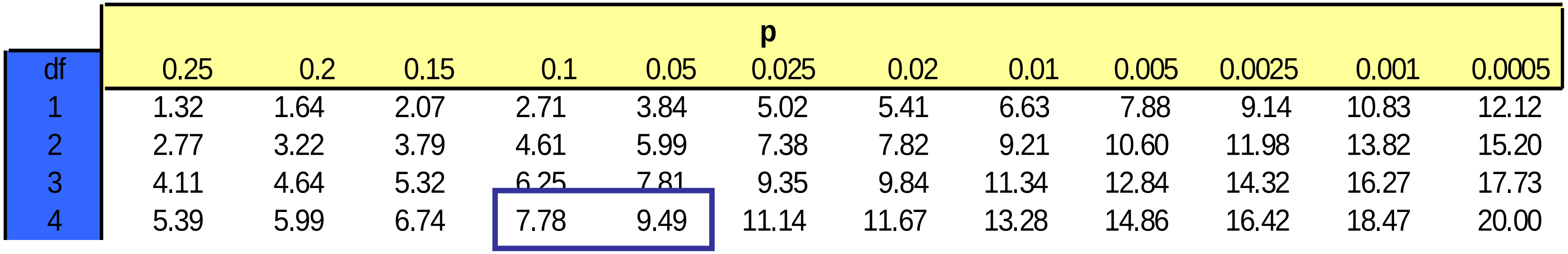
Number of collisions by day of the week					
Day of the week					
Mon.	Tue.	Wed.	Thu.	Fri.	Total
133	126	159	136	113	667

H_0 specifies that all days are equally likely for car accidents ➔ each $p_i = 1/5$.

The expected count for each of the five days is $np_i = 667(1/5) = 133.4$.

$$X^2 = \sum \frac{(observed - expected)^2}{expected} = \sum \frac{(count_{day} - 133.4)^2}{133.4} = 8.49$$

Following the chi-square distribution with 5 − 1 = 4 degrees of freedom.



	p											
df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00

The p-value is thus between 0.1 and 0.05, which is not significant at α 5%.

➔ There is no significant evidence of different car accident rates for different weekdays when the driver was using a cell phone.

Using software

The chi-square function in **Excel** does not compute expected counts automatically but instead lets you provide them. This makes it easy to test for goodness of fit. You then get the test's p-value—but no details of the χ^2 calculations.

=CHI TEST (array of actual values, array of expected values)

with values arranged in two similar *r* * *c* tables

--> returns the p value of the Chi Square test

Note: Many software packages do not provide a direct way to compute the chi-square goodness of fit test. But you can find a way around:

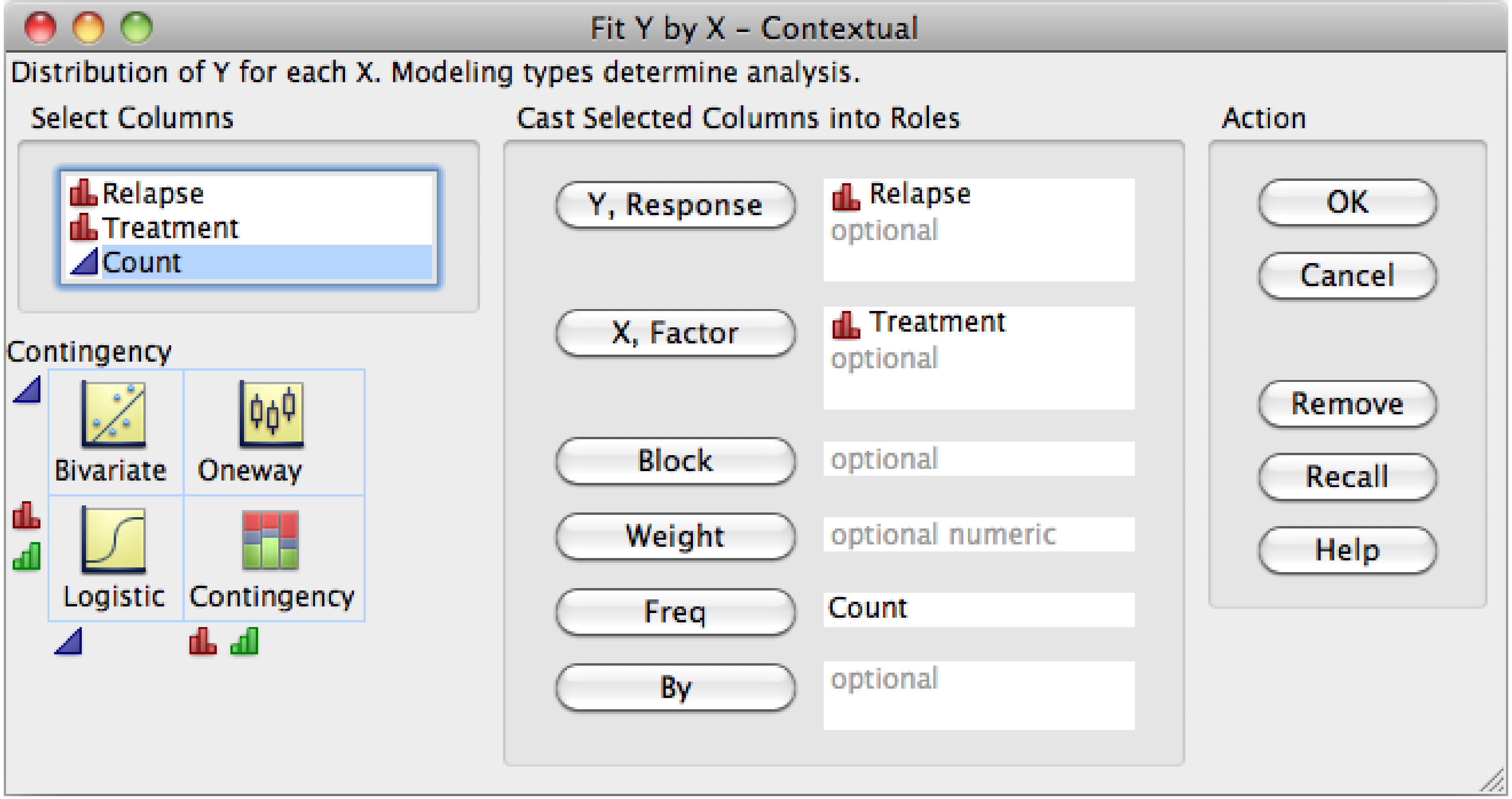
> Make a two-way table where the first column contains *k* cells with the observed counts. Make a second column with counts that correspond *exactly* to the probabilities specified by H_0, using a <u>very</u> large number of observations. Then analyze the two-way table with the chi-square function.

Alternate Slides

The following slides offer alternate software output data and examples for this presentation.

Using software

- In Excel you have to do almost all the calculations for the chi-square test yourself, and it only gives you the p-value (not the component).
- This is ***JMP***: Analyze/Fit Y by X



Fit Y by X – Contextual

Distribution of Y for each X. Modeling types determine analysis.

Y, Response: Relapse

X, Factor: Treatment

Block: optional

Weight: optional numeric

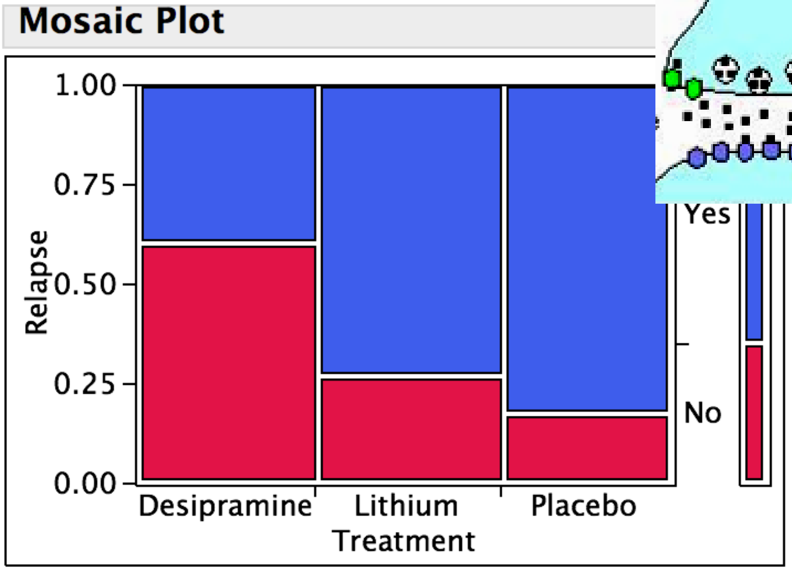
Freq: Count

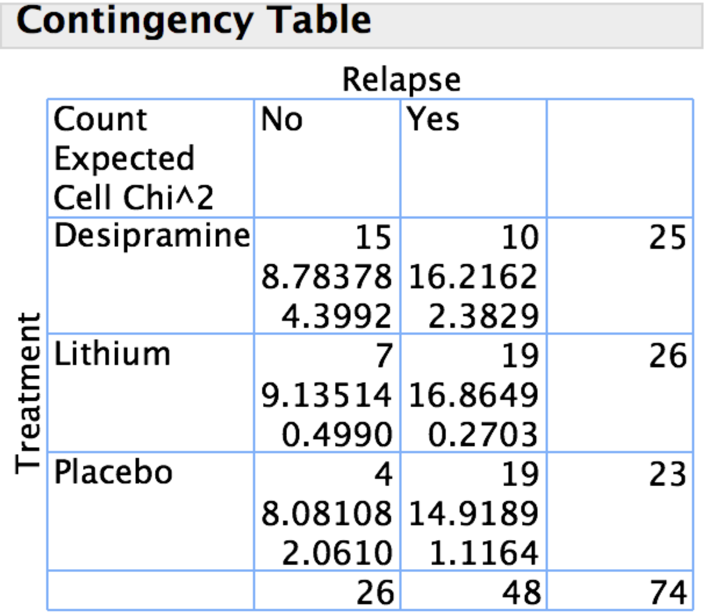
By: optional

Cocaine addiction

JMP statistical software output for the cocaine study

- after de-selecting Total, Column and Row percentages and selecting Expected and Cell Chi Square

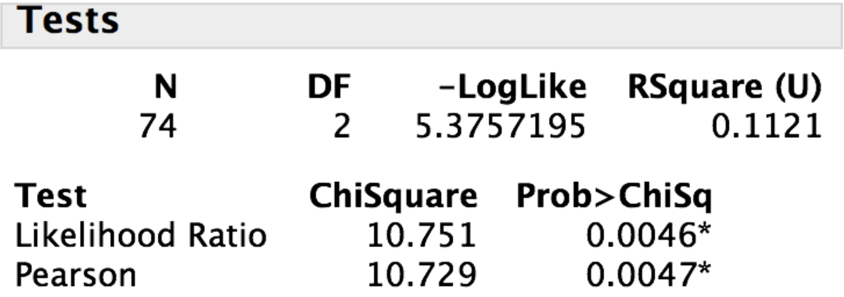




Contingency Table

Relapse (columns), Treatment (rows)

Count Expected Cell Chi^2	No	Yes	
Desipramine	15 8.78378 4.3992	10 16.2162 2.3829	25
Lithium	7 9.13514 0.4990	19 16.8649 0.2703	26
Placebo	4 8.08108 2.0610	19 14.9189 1.1164	23
	26	48	74



Tests

N	DF	-LogLike	RSquare (U)
74	2	5.3757195	0.1121

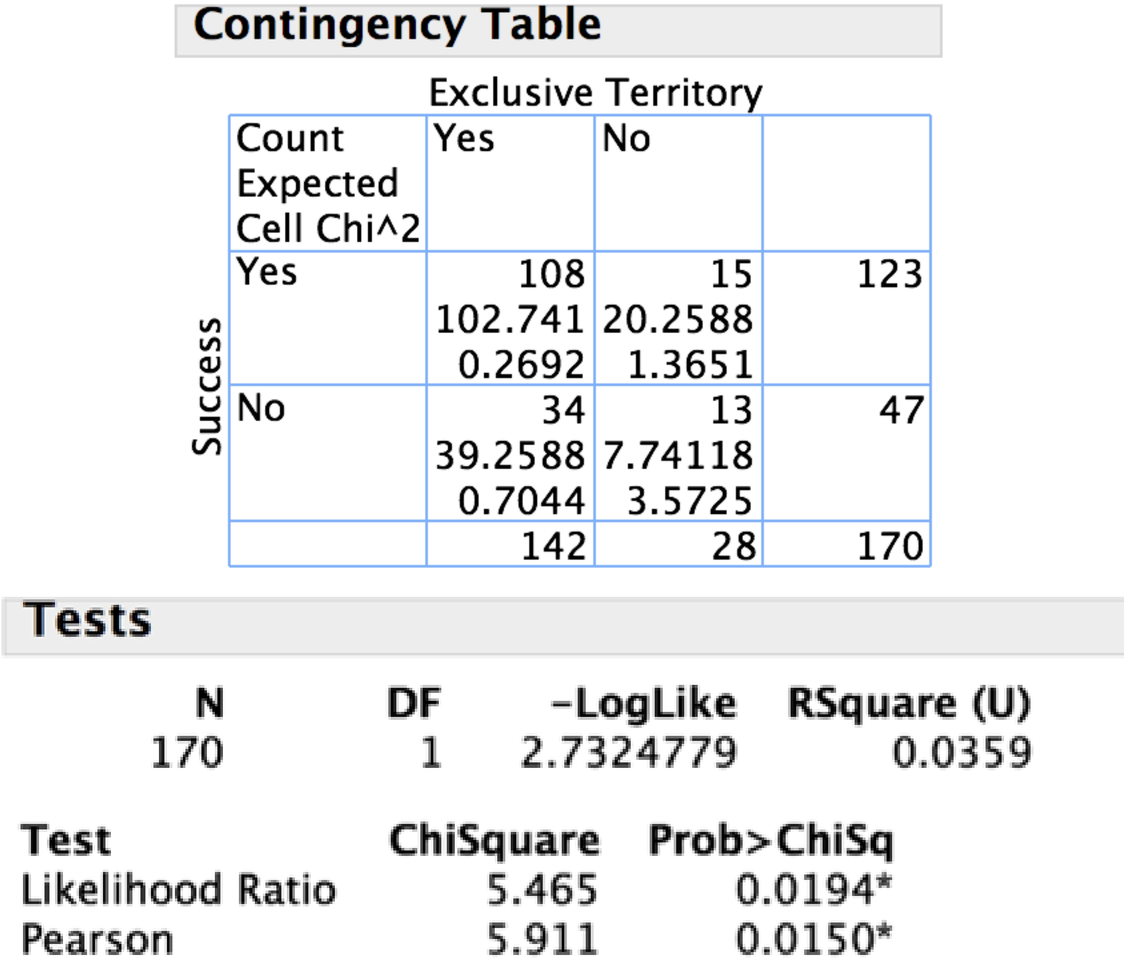
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	10.751	0.0046*
Pearson	10.729	0.0047*

The p-value is 0.0047 or half a percent (use Pearson χ^2). This is very significant.

We reject the null hypothesis of no association and conclude that there is a significant relationship between treatment *(desipramine, lithium, placebo)* and outcome *(relapse or not)*.

Successful firms

Here is the chi-square output from ***JMP***:



Contingency Table

		Exclusive Territory		
	Count Expected Cell Chi^2	Yes	No	
Success	Yes	108 102.741 0.2692	15 20.2588 1.3651	123
	No	34 39.2588 0.7044	13 7.74118 3.5725	47
		142	28	170

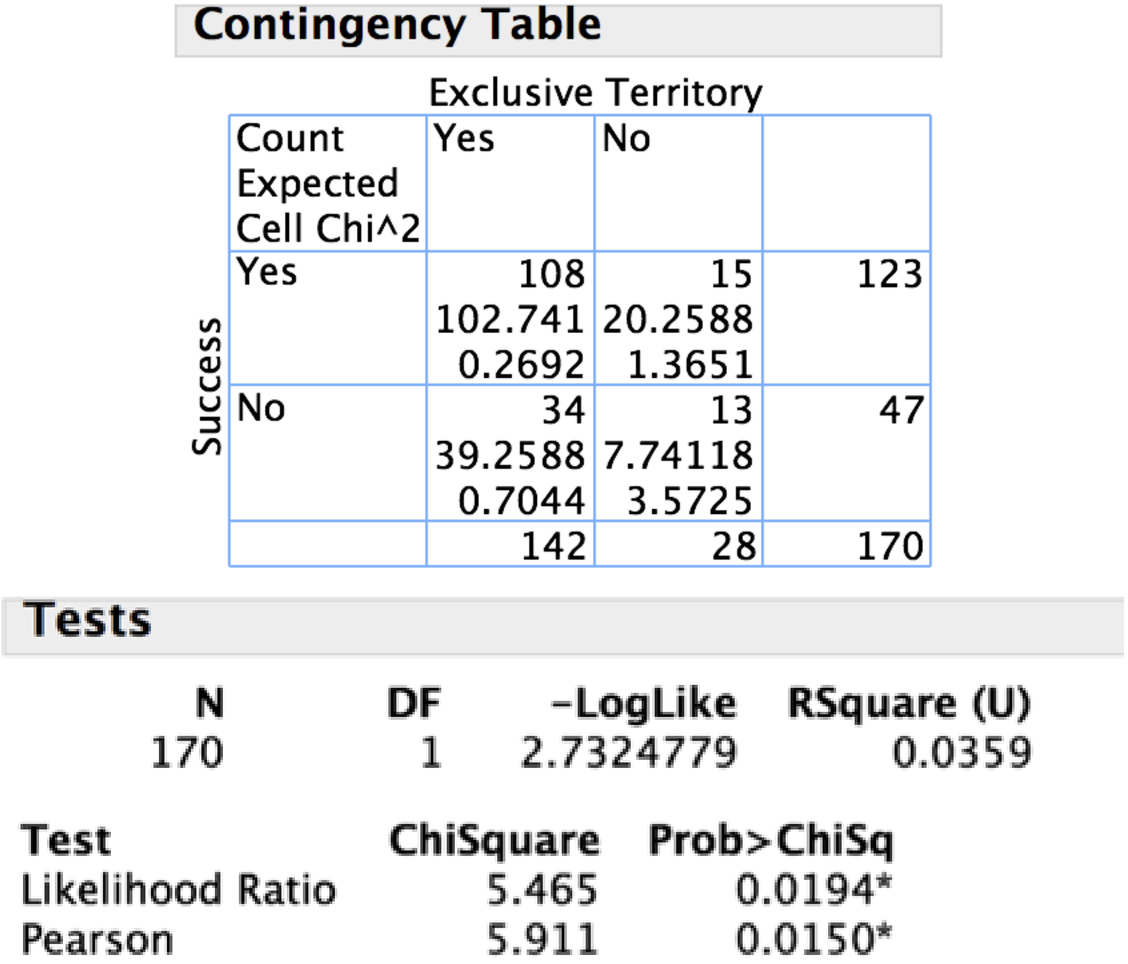
Tests

```
       N          DF       -LogLike   RSquare (U)
     170           1      2.7324779        0.0359

Test               ChiSquare   Prob>ChiSq
Likelihood Ratio       5.465      0.0194*
Pearson                5.911      0.0150*
```

The p-value is significant at $\alpha = 5\%$ ($p = 1.5\%$) thus we reject H_0: we have found a significant relationship between an exclusive territory and the success of a franchised firm.

Successful firms

Here is the chi-square output from ***JMP***:



Contingency Table

		Exclusive Territory		
	Count Expected Cell Chi^2	Yes	No	
Success	Yes	108 102.741 0.2692	15 20.2588 1.3651	123
	No	34 39.2588 0.7044	13 7.74118 3.5725	47
		142	28	170

Tests

N	DF	-LogLike	RSquare (U)
170	1	2.7324779	0.0359

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	5.465	0.0194*
Pearson	5.911	0.0150*

The p-value is significant at α 5% thus we reject H_0:

The existence of an exclusive territory clause in a franchise's contract and the success of that franchise are <u>not independent</u> variables.