Two categorical variables: The chi-square test

BPS chapter 23

Objectives (BPS chapter 23)

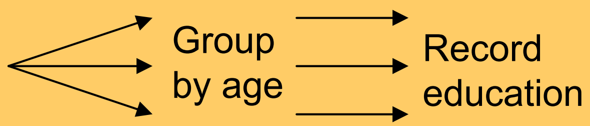
The chi-square test

- Two-way tables
- The problem of multiple comparisons
- Expected counts in two-way tables
- The chi-square test
- Using technology
- Cell counts required for the chi-square test
- Uses of the chi-square test
- The chi-square distributions
- The chi-square test and the z test
- Chi-square test for goodness of fit

Two-way tables

An experiment has a **two-way** design if two **categorical** factors are studied with several levels of each factor.

Two-way tables organize data about two categorical variables with any number of levels/treatments obtained from a two-way, or block, design. (*There are now two ways to group the data.*)



First factor: age

Second factor: education

Years of school completed, by age (thousands of persons)

Education	Age group: 25 to 34	35 to 54	55 and over
Did not complete high school	4,459	9,174	14,226
Completed high school	11,562	26,455	20,060
College, 1 to 3 years	10,693	22,647	11,125
College, 4 or more years	11,071	23,160	10,597

Chi-square hypothesis test

H_0: There is no relationship between categorical variable A and categorical variable B.

H_a: There is some relationship between categorical variable A and categorical variable B.

This alternative hypothesis is not really one-sided (> or <) or two-sided ($\neq$). It can be called “many-sided” because it allows any kind of relationship between variables A and B to count.

The problem of multiple comparisons

Why do we do the “many-sided” chi-square test instead of a proportion test using the two biggest proportions for each variable if we want to see if there is a relationship?

Picking out just the biggest proportions would be considered cheating because the other proportions are important too, even though they are smaller. They come together as one package and need to be tested together.

The steps for multiple comparisons tests

The chi-square test is an example of a **multiple comparisons test**. Statistical methods for dealing with multiple comparisons usually have two steps:

1. An overall test to see if there is good evidence of any differences among parameters that we want to compare.

2. A detailed follow-up analysis to decide which of the parameters differ and to estimate how large the differences are.

Some other multiple comparisons tests:

one-way ANOVA F test (Ch. 25), two-way ANOVA F tests (Ch. 29), multiple regression ANOVA F-test (Ch. 28)

Expected counts in two-way tables

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.

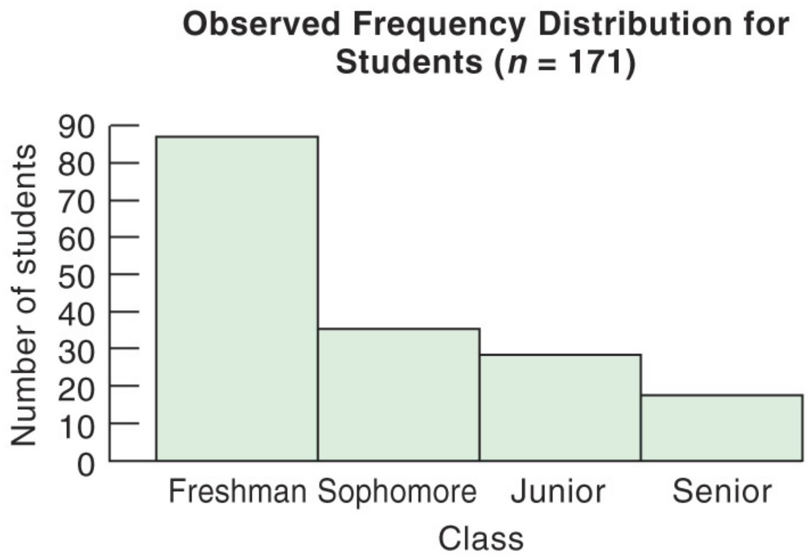
The expected count in any cell of a two-way table when H_0 is true is:

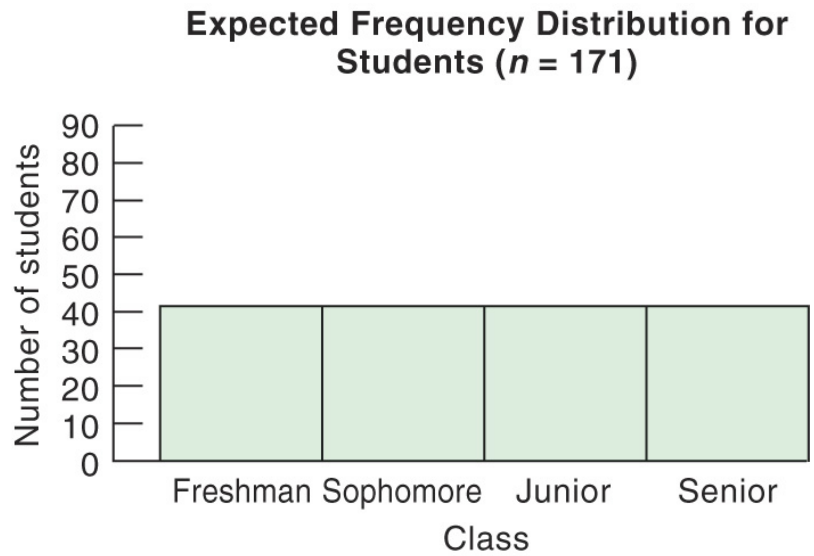
$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

The chi-square test

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





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 (summed over all cells in the table) is:

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

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

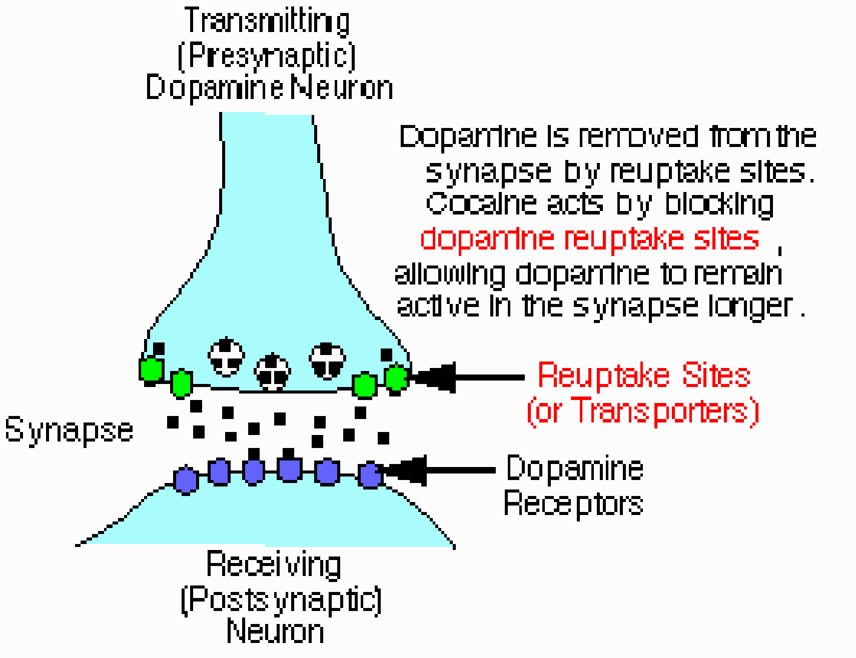
However, since χ^2 is a sum, how large a χ^2 is required for statistical significance will depend on the number of comparisons made.

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.**

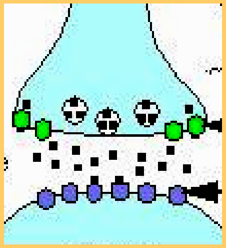
The pleasurable high followed by unpleasant after-effects encourage repeated compulsive use, which can easily lead to dependency.

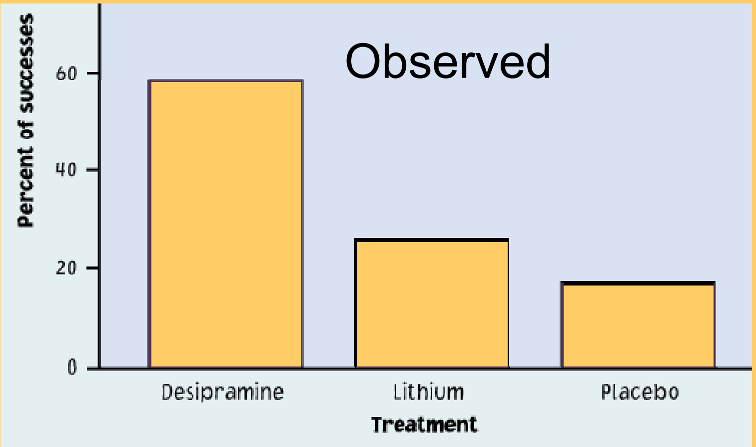
Desipramine is an **antidepressant** affecting the brain chemicals that may become unbalanced and cause depression. It was thus tested for recovery from cocaine addiction.



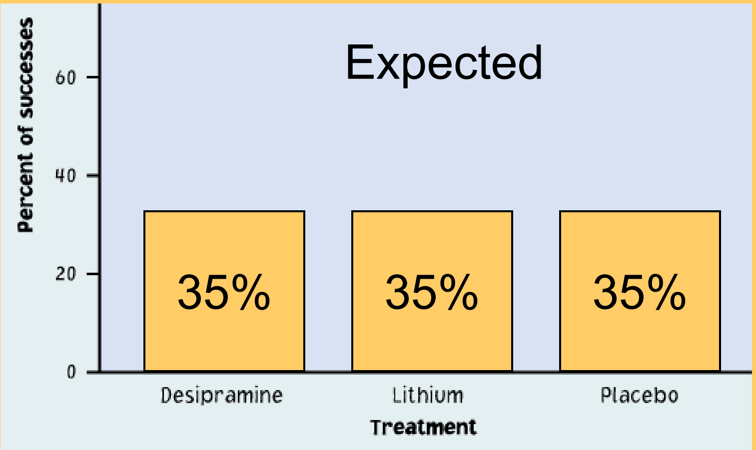
Treatment with desipramine was compared to a standard treatment (lithium, with strong anti-manic effects) and a placebo.

Cocaine addiction





	Relapse: No	Relapse: Yes	Total
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***0.35**	16.22 25*0.65
Lithium	9.14 26***0.35**	16.86 25*0.65
Placebo	8.08 23***0.35**	14.92 25*0.65

The chi-square distributions

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

Table E gives upper critical values for many chi-square distributions.

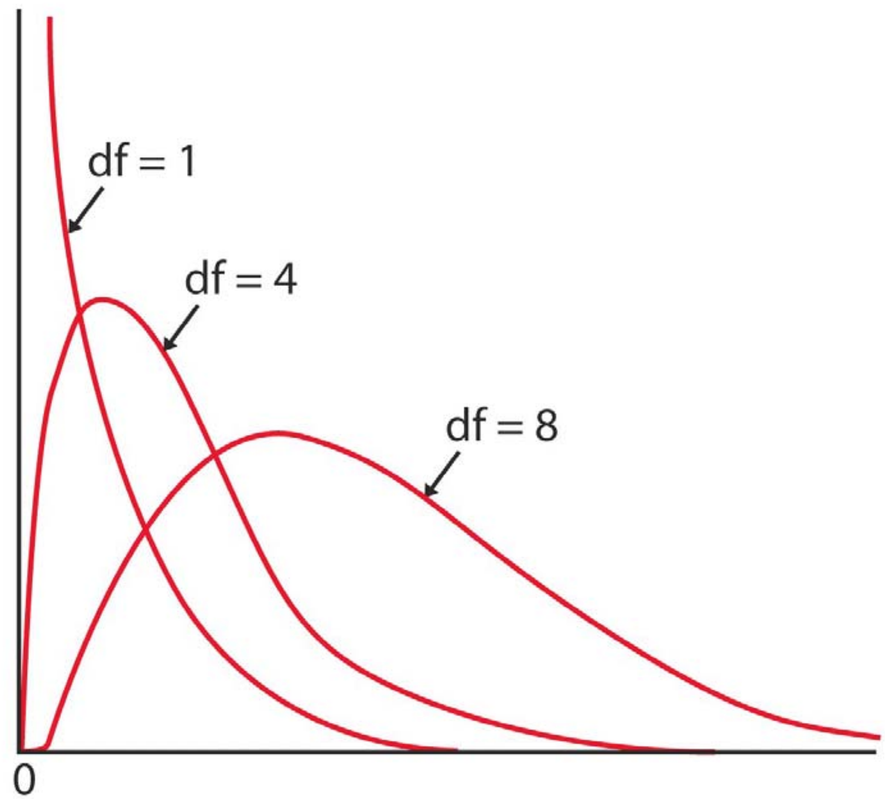
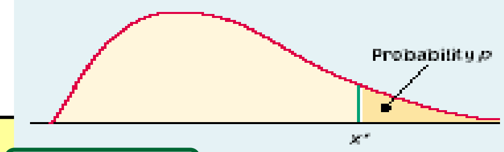
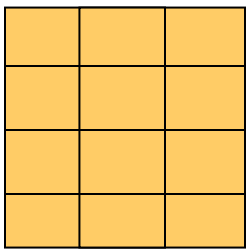


Table E



df = (*r* -1)(*c* -1)

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

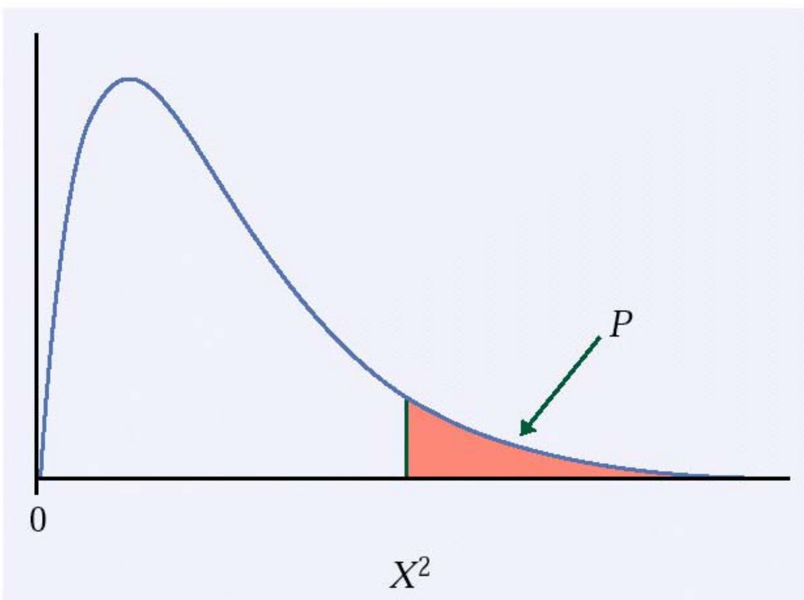


If χ^2 = 16.1 the p-value is between 0.01 −0.02.

df	0.25	0.2	0.15	0.1	0.05	p 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

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 χ^2 : $P(\chi^2 \geq X^2)$.

Cocaine addiction

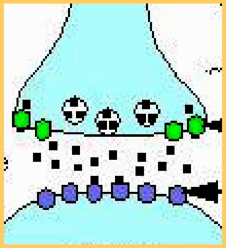


Table of counts: "actual/**expected**," with three rows and two columns:

$$df = (3 - 1)*(2 - 1) = 2$$

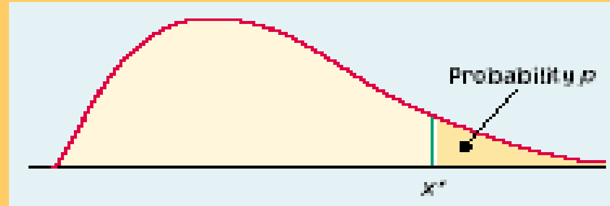
	No relapse	Relapse
Desipramine	15 **8.78**	10 **16.22**
Lithium	7 **9.14**	19 **16.86**
Placebo	4 **8.08**	19 **14.92**

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

χ^2 components:

4.41	2.39
0.50	0.27
2.06	1.12

Cocaine addiction
Table E



	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

$X^2 = 10.71$ and df = 2

$10.60 < X^2 < 11.98$ ➔ $0.005 < P\text{-}value < 0.0025$

The *P*-value is less than one-half percent, thus we reject the null hypothesis.

➔ There is a significant relationship between treatment type (desipramine, lithium, placebo) and outcome (relapse or not).

Interpreting the χ^2 output

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

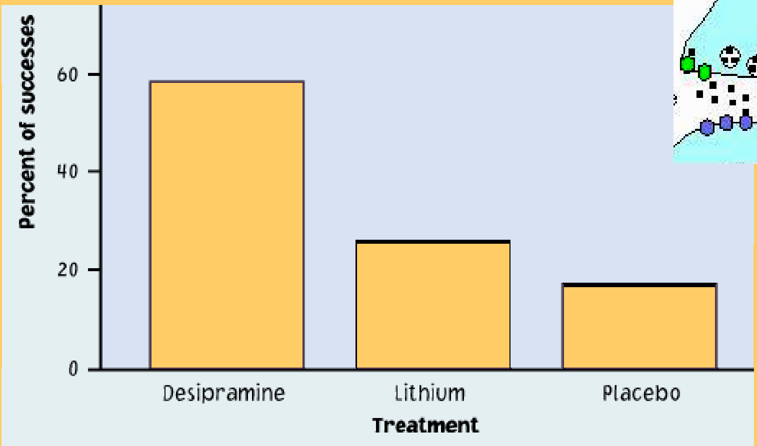
When the test is statistically significant, the **largest components** point to the conditions most different from the expectations based on H_0.

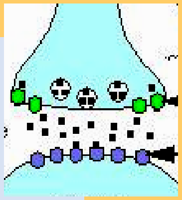
- You can also calculate the actual **proportions** for each condition (instead of the counts) and compare them qualitatively.

Cocaine addiction

	No relapse	Relapse
Desipramine	4.41	2.39
Lithium	0.50	0.27
Placebo	2.06	1.12

χ^2 components





Desipramine stands out from the other treatments.
Actual proportions show it is most beneficial.

Using the chi-square test

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. The samples can be drawn either:

- By randomly selecting several simple random samples each from a different population (or from a population subjected to different treatments) → experimental study

- Or by taking one simple random sample and classifying the individuals in the sample according to two categorical variables (attribute or condition) → observational study, historical design

When is it safe to use a chi-square test?

We can safely use the chi-square test when:

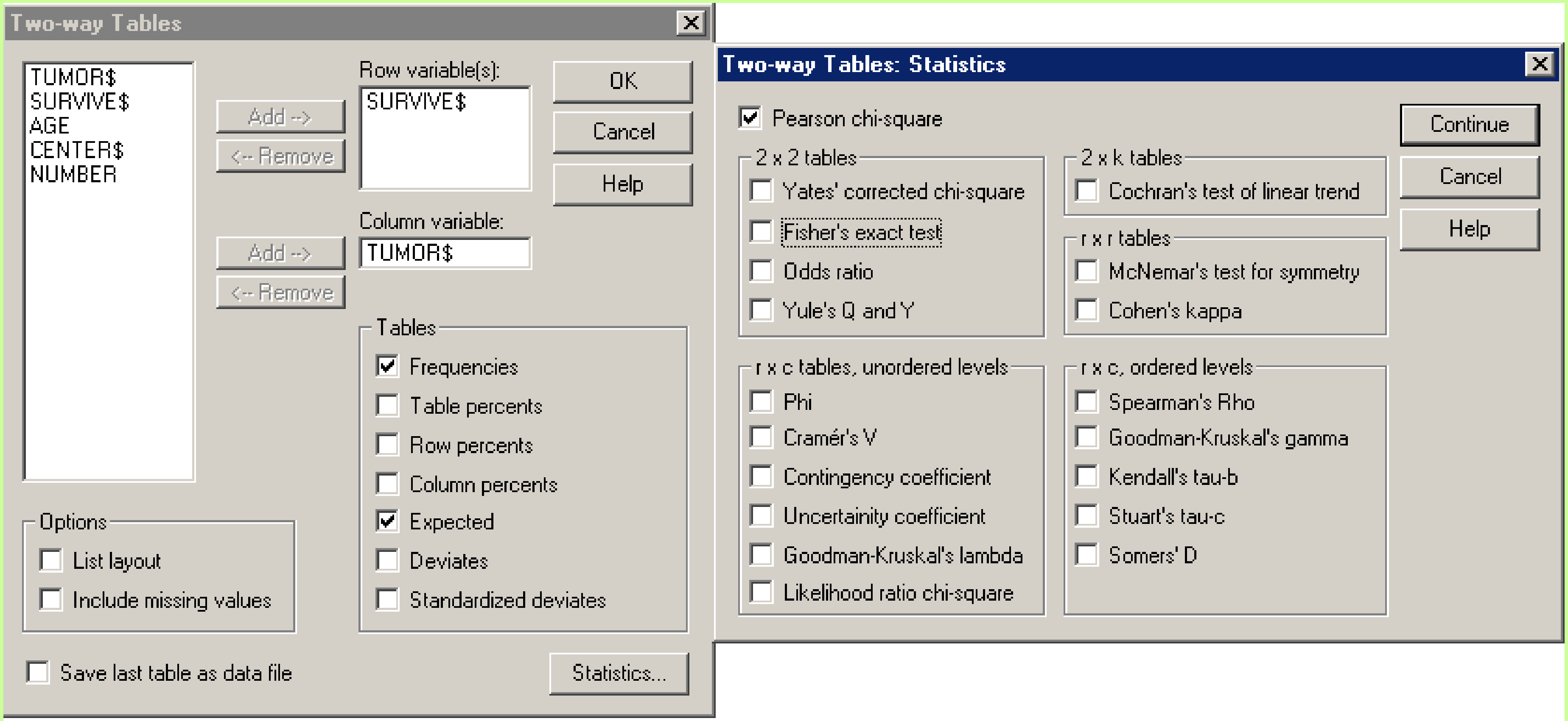
- The samples are simple random samples (SRS).
- All individual **expected counts** are 1 or more (≥1)
- No more than 20% of **expected counts** are less than 5 (< 5)

  For a 2x2 table, this implies that all four expected counts should be 5 or more.

Using software

- In Excel you almost have to do 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.



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 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.

We will test H_0: no relationship between exclusive clause and success.

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, and two levels for exclusive territory, yes/no).

➔df = (2 − 1)(2 − 1) = 1

Successful firms

```
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
```

Test for independence of Success and Exclusive Territory:

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

The *P*-value is significant at α = 5%, thus we reject H_0: We have found a significant relationship between an exclusive territory and the success of a franchised firm.

The chi-square test and the *z* test

If you have a 2 x 2 table, and you are using a 2-sided H_a, then you can use either the chi-square test OR the 2-sample proportion *z* test.

The *P*-values are identical, and the X^2 test statistic = the (z test statistic)^2.

The *z* test for the 2-sample proportions is preferable because you have the one-sided alternative hypothesis and confidence interval options.

Chi-square test for goodness of fit

This variant of the chi-square test is used for comparing a theoretical distribution with the observed data from a sample.

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

The null hypothesis specifies probabilities $p_1, p_2, \ldots, p_k$ for the possible outcomes.

Is the number of car accidents similar on every weekday?

$$H_0: p_{Mo} = p_{Tu} = p_{We} = p_{Th} = p_{Fr} = 1/5$$

→ Use the sample counts and expected counts (from H_0) to calculate X^2

For each cell, we multiply the total number of observations n by the specified probability to determine that expected count:

$$\text{expected count}_{(i)} = np_i$$

The **chi-square statistic** measures how much the observed cell counts differ from the expected cell counts. It follows the chi-square distribution with $k - 1$ degrees of freedom and has for formula:

$$X^2 = \sum \frac{(\text{count of outcome } i - np_{i0})^2}{np_{i0}}$$

The *P*-value is the tail area under the X^2 distribution with df $k - 1$.

Car accidents and day of the week

A study of 667 accidents that happened on a weekday classified these accidents by the day of the week they occurred.

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 $np_i = 667(1/5) = 133.4$.

• Each chi-square component is equal to: $(\text{count}_{day} – 133.4)^2/133.4$, which sums up to $X^2 = 8.49$ with 5 − 1 = 4 degrees of freedom.

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 the days of the week when the driver was using a cell phone.

	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