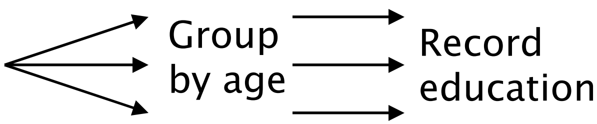
Lecture 12

Analysis of Two-Way tables

Ch 9

What are two-way tables?

- In statistics we call categorical variables present in an experimental design: FACTORS
- Each possible value of the categorical variable (factor) is called a level of the factor.
- With this language a two-way table is a representation of an experiment that studies the relationship between two factors.



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

Marginal distributions

We can look at each categorical variable separately in a two-way table by studying the row totals and the column totals. They represent the marginal distributions, expressed in counts or percentages (They are written as if in a margin.)

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

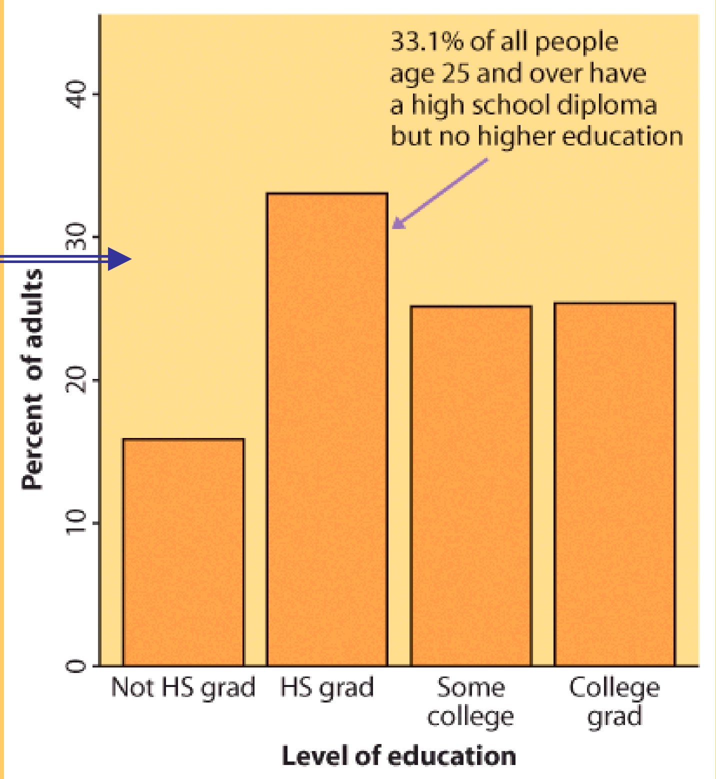
	Age group			
Education	25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

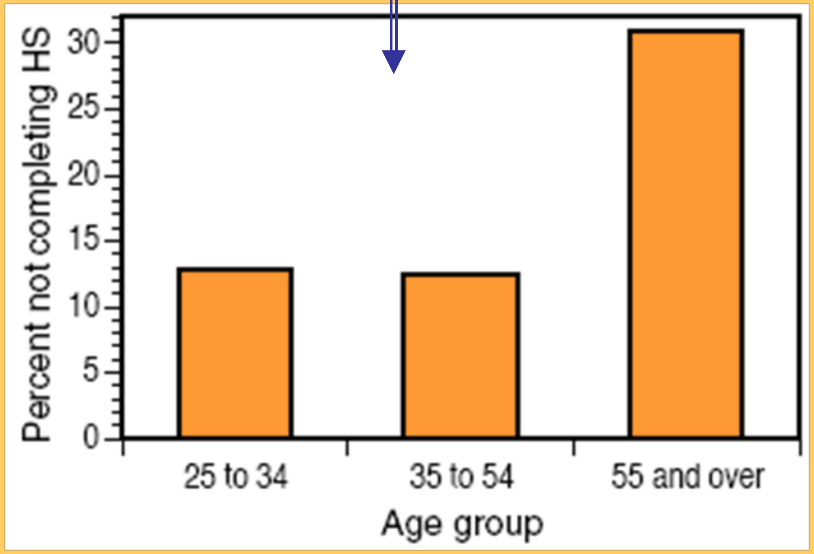
2000 U.S. census

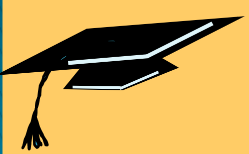
The marginal distributions can then be displayed on separate bar graphs, typically expressed as percents instead of raw counts. Each graph represents only one of the two variables, completely ignoring the second one.

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

Education	Age group: 25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230







Parental smoking

Does parental smoking influence the smoking habits of their high school children?

Summary two-way table: High school students were asked whether they smoke and whether their parents smoke.

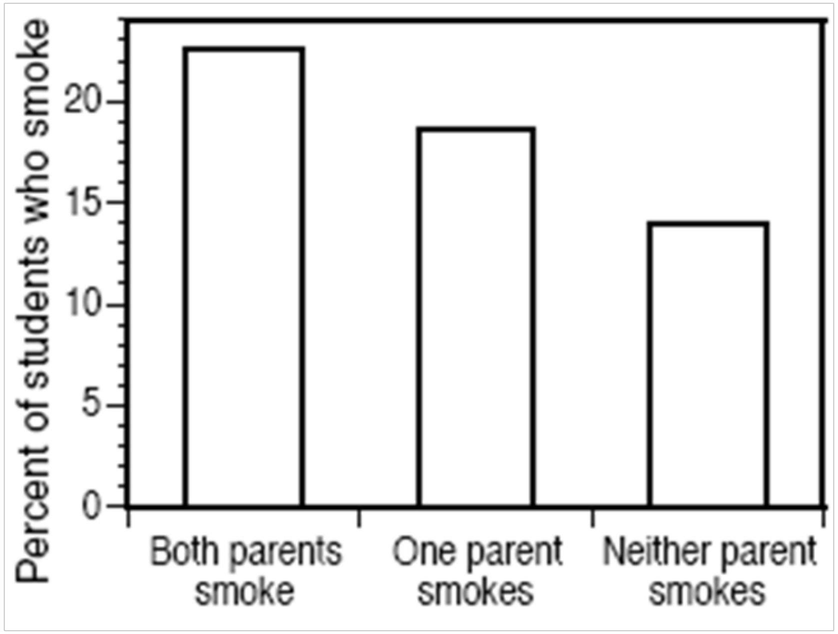
	Student smokes	Student does not smoke	*Total*
Both parents smoke	332.49	1447.51	1780
One parent smokes	418.22	1820.78	2239
Neither parent smokes	253.29	1102.71	1356
Total	1004	4371	5375

Marginal distribution for the categorical variable "parental smoking":

The row totals are used and re-expressed as percent of the grand total.

Neither parent smokes	One parent smokes	Both parents smoke
13.9%	18.6%	22.5%

The percents are then displayed in a bar graph.



Relationships between categorical variables

The marginal distributions summarize each categorical variable independently. But the two-way table actually describes the relationship between both categorical variables.

The cells of a two-way table represent the intersection of a given level of one categorical factor with a given level of the other categorical factor.

Because counts can be misleading (for instance, one level of one factor might be much less represented than the other levels), we prefer to calculate percents or proportions for the corresponding cells. These make up the conditional distributions.

Conditional distributions

The counts or percents within the table represent the **conditional distributions**. Comparing the conditional distributions allows you to describe the "relationship" between both categorical variables.

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

	Age group			
Education	25 to 34	35 to 54	55 and over	Total
Did not complete high school	4,459	9,174	14,226	27,859
Completed high school	11,562	26,455	20,060	58,077
College, 1 to 3 years	10,693	22,647	11,125	44,465
College, 4 or more years	11,071	23,160	10,597	44,828
Total	37,786	81,435	56,008	175,230

	25 to 34	35 to 54	55 up	All
1:NotHS	4459	9174	14226	27859
	11.80	11.27	25.40	15.90
2:HSgrad	11562	26455	20060	58077
	30.60	32.49	35.82	33.14
3:SomeCo	10693	22647	11125	44465
	28.30	27.81	19.86	25.38
4:CollGr	11071	23160	10597	44828
	29.30	28.44	18.92	25.58
All	37785	81436	56008	175229
	100.00	100.00	100.00	100.00

l Contents-
Count
% of Col

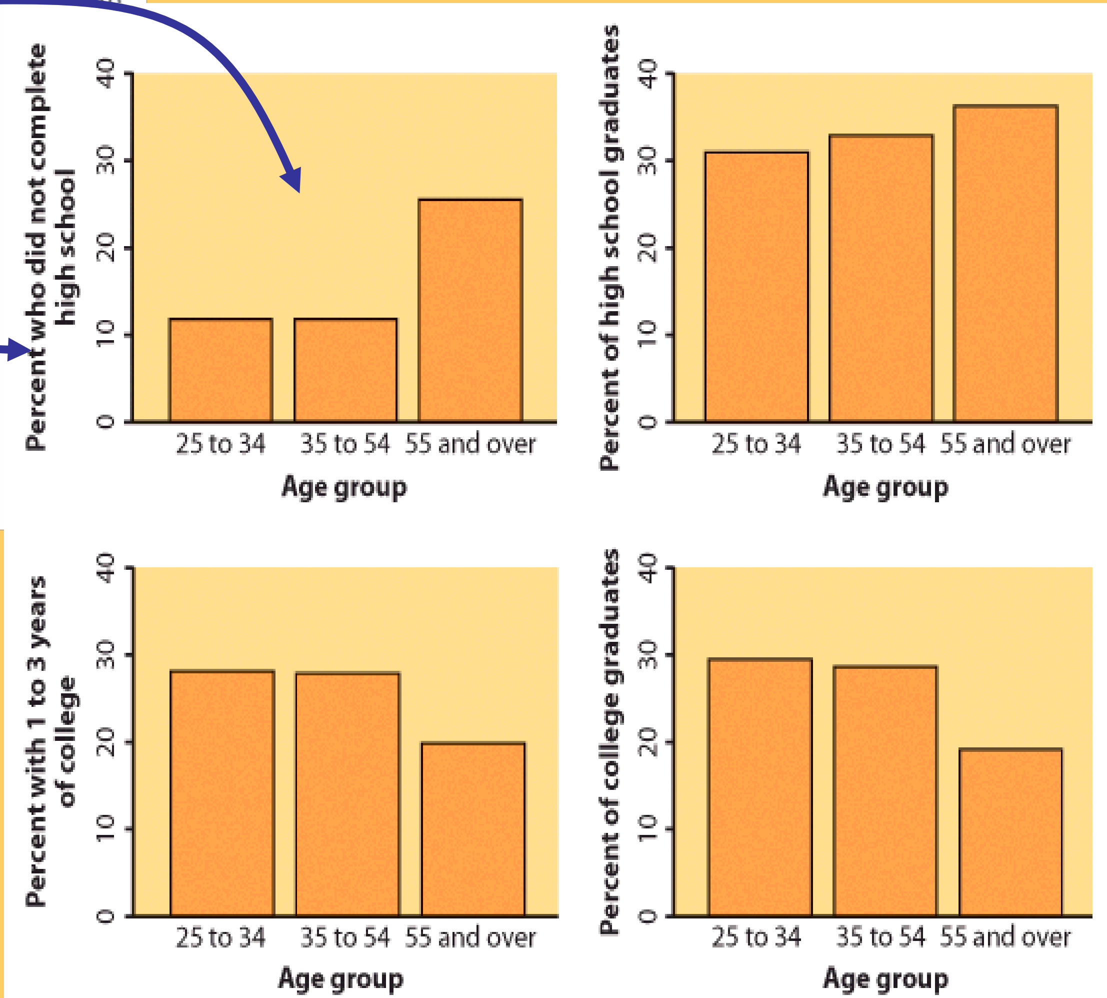
Here the percents are calculated by age range

$$29.30\% = \frac{11071}{37785} = \frac{\text{cell total}}{\text{column total}}$$

The conditional distributions can be graphically compared using side by side bar graphs of one variable for each value of the other variable.

	25 to 34	35 to 54	55 up	All
1:NotHS	4459	9174	14226	27859
	11.80	11.27	25.40	[illegible]
2:HSgrad	11562	26455	20060	
	30.60	32.49	35.82	
3:SomeCo	10693	22647	11125	
	28.30	27.81	19.86	
4:CollGr	11071	23160	10597	
	29.30	28.44	18.92	
All	37785	81436	56008	
	100.00	100.00	100.00	

Cell Contents–
Count
% of Col



Here the percents are calculated by age range (columns).

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.

$$\frac{30}{84} = 35.7\% = \frac{\text{cell total}}{\text{column total}}.$$

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.

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

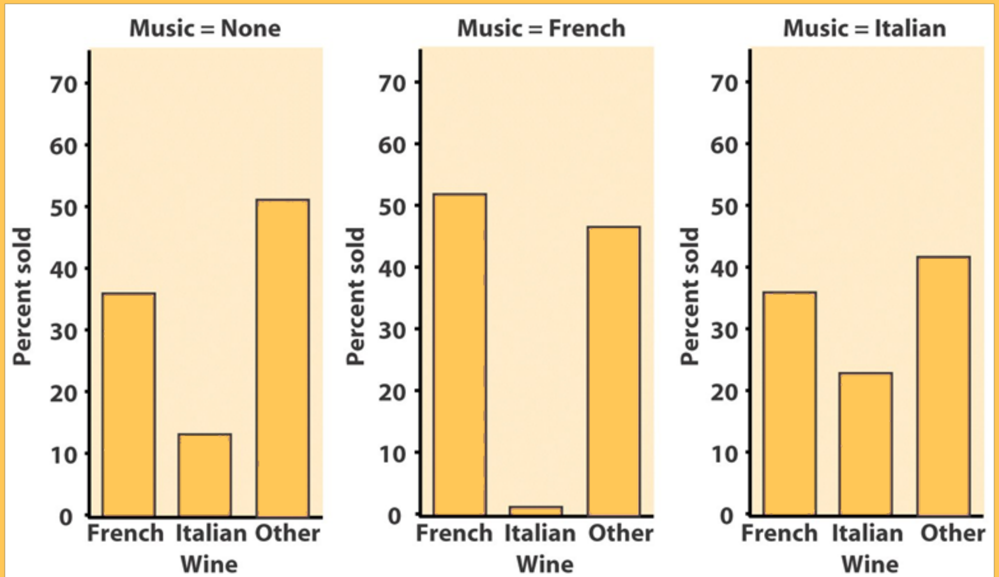
Column percents for wine and music

	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

For every two-way table, there are two sets of possible conditional distributions.

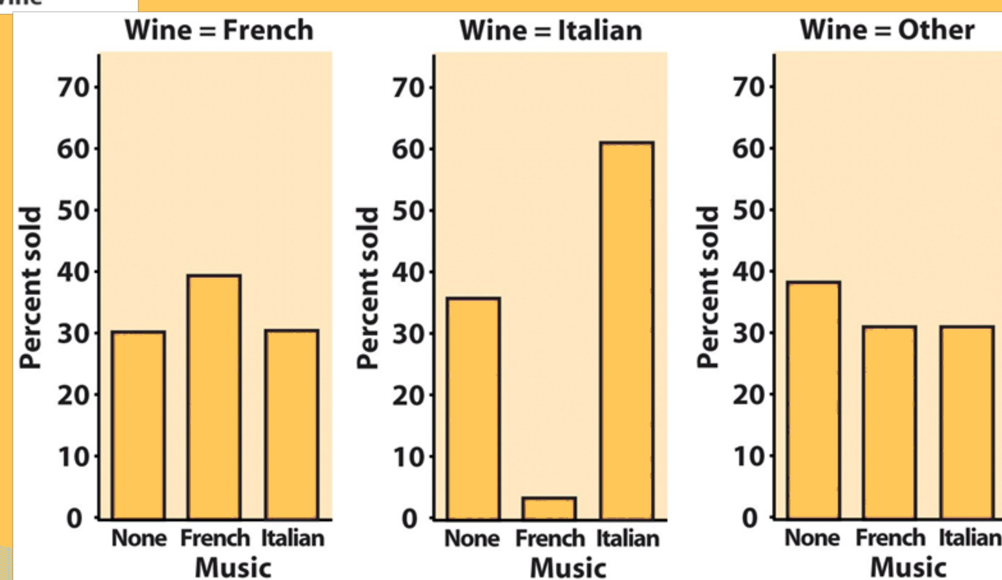
	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

Does background music in supermarkets influence customer purchasing decisions?



Wine purchased for each kind of music played (column percents)

Music played for each kind of wine purchased (row percents)



Simpson's paradox

An association or comparison that holds for all of several groups can reverse direction when the data are combined (aggregated) to form a single group. This reversal is called Simpson's paradox.

Example: Hospital death rates

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800
% surv.	97.0%	98.0%

On the surface, Hospital B would seem to have a better record.

Patients in good condition		
	Hospital A	Hospital B
Died	6	8
Survived	594	592
Total	600	600
% surv.	99.0%	98.7%

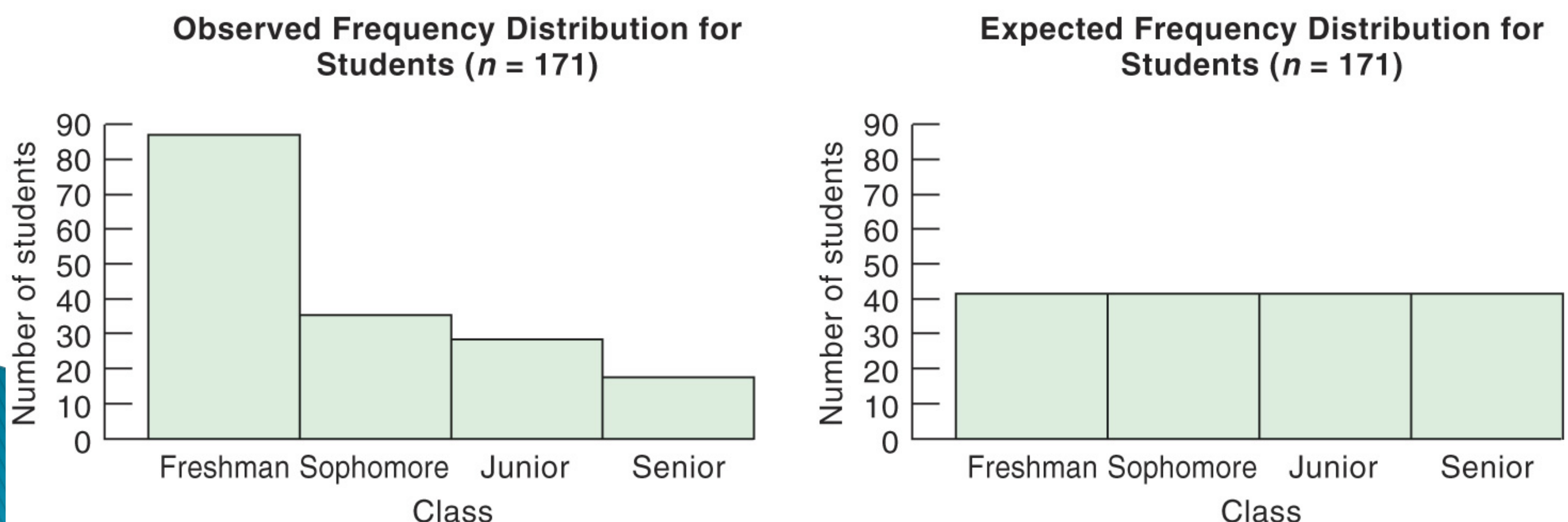
Patients in poor condition		
	Hospital A	Hospital B
Died	57	8
Survived	1443	192
Total	1500	200
% surv.	96.2%	96.0%

But once patient condition is taken into account, we see that hospital A has in fact a better record for both patient conditions (good and poor).

Here patient condition was the lurking variable.

Inference for Two-Way tables.

- The main test is to check whether or not the two factors are independent or if there is a relationship between them.
  - Put it differently we check if the differences in sample proportions that are observed are likely to have occurred by just chance because of the random sampling.
- To assess this we use a **chi-square (χ^2) test** to check the null hypothesis of no relationship between the two categorical variables of a two-way table.



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 (under independence hypothesis) is:

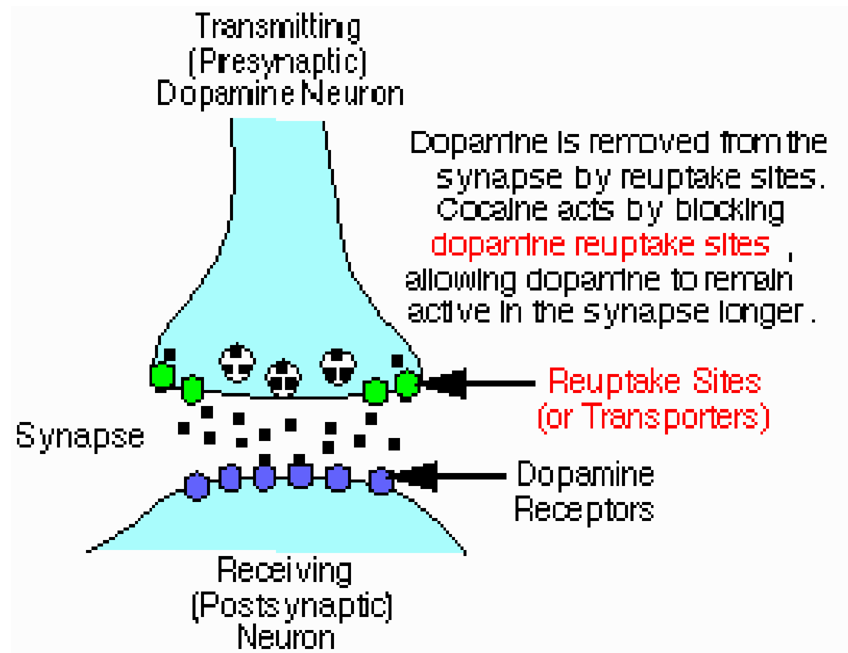
$$\boxed{\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}}$$

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.**

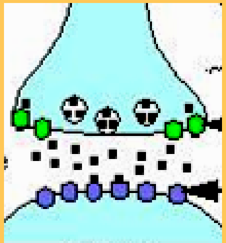
The pleasurable high followed by unpleasant after-effects encourage repeated compulsive use, which can easily lead to dependency.

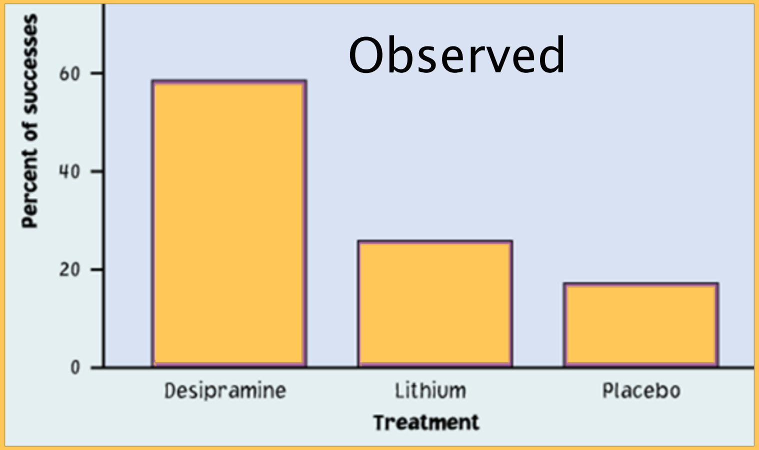
Desipramine is an **antidepressant** affecting the brain chemicals that may become unbalanced and cause depression. It was thus tested for recovery from cocaine addiction.



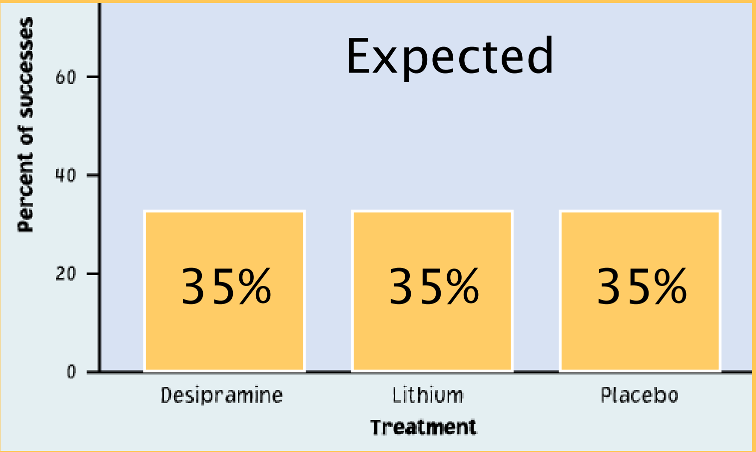
Treatment with desipramine was compared to a standard treatment (lithium, with strong anti-manic effects) and a placebo.

Cocaine addiction





	Relapse: 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 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 $r * c$ cells in the table)

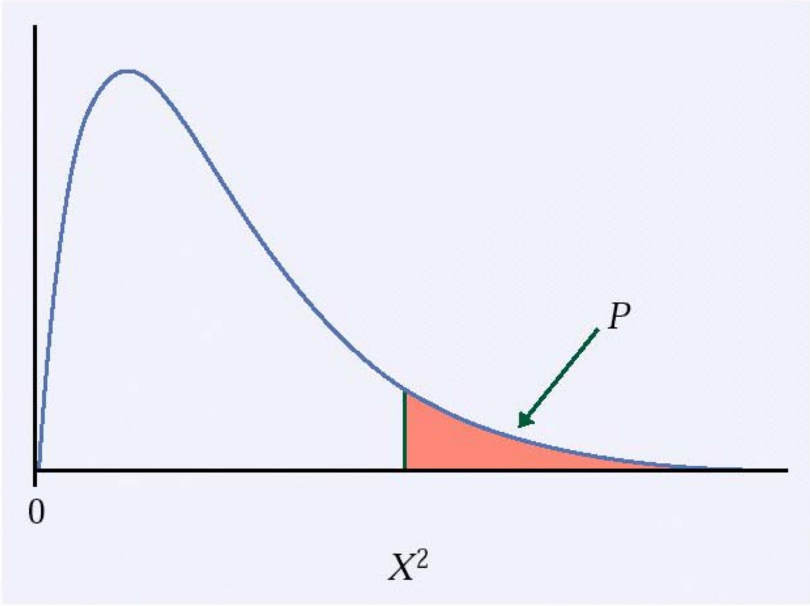
$$\boxed{\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}}$$

Large values for χ^2 represent strong deviations from the expected distribution under the H_0 and providing 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 χ^2 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 ransom samples (SRS).
- All individual expected counts are 1 or more
- No more than 20% of expected counts are less than 5

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

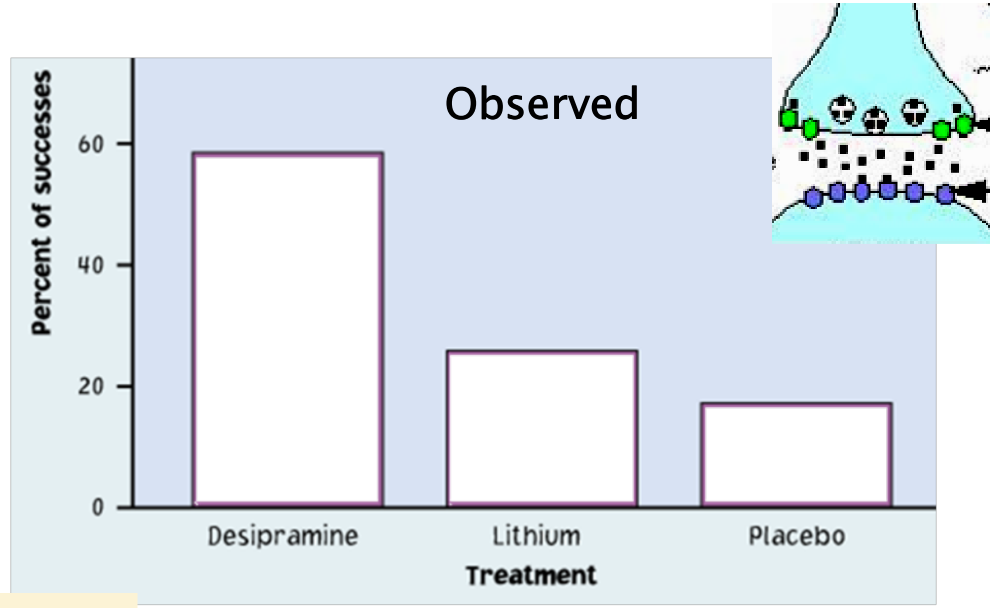
Chi-square test vs. z-test for two proportions

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: $X^2 = (z)^2$.

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 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).

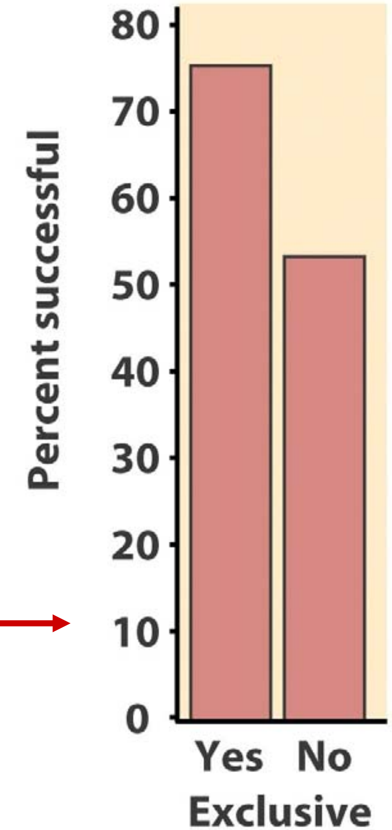
➔df = (2 − 1)(2 − 1) = 1

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 α 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

R code:

- In R you create a matrix with elements the counts.
- Then to perform the chi-square test use simply:

```
chisq.test()
```

- More details can be found on pages 136-137 of the R textbook.