Chi-Square worked examples:

1. Criminologists have long debated whether there is a relationship between weather conditions and the incidence of violent crime. The article "Is There a Season for Homocide?" classified 1361 homicides according to season, resulting in the following table (*Criminology*, 1988: 287-296):

Winter	Spring	Summer	Fall
328	334	372	327

Test to see if there is a relationship between the seasons and violent crime.

---- Solution ----

We have two possibilities:

(a) homicides do not change from season to season (homicides are independent of seasons)
(b) homicides do change from season to season (homicides are not evenly distributed across seasons)

Looking at the table, it's obvious that the homicides are **not** evenly distributed across seasons. Remember, however, that our data is only a sample of 1361 homicides. The differences among seasons that we've observed might simply be due to sampling error.

We can state our null and alternate hypotheses now. Our null hypothesis always says "nothing happens." In this case, our null hypothesis would say that homicides are independent of seasons. Our alternate hypothesis would say that homicides change across seasons. We can state these hypotheses in a variety of ways:

1. In terms of probabilities

H_0: P(winter) = P(spring) = P(summer) = P(fall)
H_A: Not H_0

2. In terms of independence

H_0: Homicides are independent of seasons
H_A: Not H_0

3. In terms of distributions

H_0: Homicides are evenly distributed across seasons
H_A: Not H_0

We can now run our chi-squared "badness-of-fit" test on this data. To do this, we need to know both the **observed** and **expected** number of homicides for each season. The observed numbers of homicides are obvious – they are stated in the table. The expected number of homicides in each season are unknown – we have no way of knowing what to expect each season. If we assume our null hypothesis is true, however, we can calculate our expectations.

Under our null hypothesis, the number of homicides are evenly distributed across seasons. Since we observed 1361 homicides across 4 seasons, we expect to see:

$$\frac{1361}{4} = 340.25 \text{ homicides each season}$$

We can display these expectations in our table:

Winter	Spring	Summer	Fall
Observed = 328 Expected = 340.25	Observed = 334 Expected = 340.25	Observed = 372 Expected = 340.25	Observed = 327 Expected = 340.25

When we have a table of categorical data and we want to compare observed numbers of observations to expected numbers of observations, we can use the chi-squared goodness-of-fit test.

Chi-Squared Goodness-of-Fit Test Statistic: $\chi^2_{c\text{-}1} = \sum_{\text{All Cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

We'll calculate this step-by-step in the table:

Winter	Spring	Summer	Fall
Observed = 328 Expected = 340.25 $\frac{(328 - 340.25)^2}{340.25} = 0.441$	Observed = 334 Expected = 340.25 $\frac{(334 - 340.25)^2}{340.25} = 0.115$	Observed = 372 Expected = 340.25 $\frac{(372 - 340.25)^2}{340.25} = 2.963$	Observed = 327 Expected = 340.25 $\frac{(327 - 340.25)^2}{340.25} = 0.516$

Notes: If we forget to square the numerator, our test statistic will always equal zero.
Dividing by the expected number of observations takes sample size into account

We now calculate our test statistic as: $0.441 + 0.115 + 2.963 + 0.516 = 4.035$. When we have a table with only one row (or only one column), the chi-squared test statistic has $c - 1$ degrees of freedom, where c = the number of columns (or rows) in our table.

So, for this table, we have $\chi^2_{4-1} = \chi^2_3 = 4.035$.

To see if this value is significant, we would compare it to a critical value from our chi-squared table. If our calculated test statistic is bigger than the critical value in the table, we would reject our null hypothesis.

Using a computer, I calculate the p-value for our chi-squared test to be $p = 0.2577$. This means we should retain our null hypothesis, so we conclude that there is no significant relationship between the number of homicides and the seasons.

2. Each individual in a random sample of high school and college students was cross-classified with respect to both political views and marijuana usage, resulting in the following data ("Attitudes about Marijuana and Political Views," *Psychological Reports*, 1973: 1051-1054):

		Marijuana Usage		
		Never	**Rarely**	**Frequently**
Political Views	**Liberal**	**479**	**173**	**119**
	Conservative	**214**	**47**	**15**
	Other	**172**	**45**	**85**

Does the data support the hypothesis that political views and marijuana usage level are independent within the population?

---- Solution ----

We have a 3x3 contingency table (categorical data), so we can run our chi-squared test for independence.

Chi-Squared Test for Independence: $\chi^2_{(r-1)(c-1)} = \sum_{\text{All Cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

Notes: Yep, this looks almost identical to the Chi-squared test for independence.
Only the degrees of freedom change to $(\text{rows} - 1)(\text{columns} - 1)$

We need to state our hypotheses, calculate our expected number of observations in each cell, calculate our test statistic, and state our conclusion.

We always state our hypotheses in terms of the two variables in our study (the row and column variables). Under a null hypothesis, we assume the variables are independent. We can, therefore, state our hypotheses as:

Hypotheses:
H_0: Marijuana use and political views are independent
H_A: Not H_0

To get the expected number of observations, we must recall what we know about independence. Last semester, we learned that if two events (A and B) are independent, then:

1. Knowledge of the outcome of one event does not influence the probability of the other event
2. $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$

The last result is the most important. When two events are independent, we can multiply their probabilities to calculate the probability that they both happen simultaneously. We can use this to obtain our expected number of observations.

Let's take another look at our data:

		Marijuana Usage			
		Never	**Rarely**	**Frequently**	**TOTAL**
Political Views	**Liberal**	479	173	119	**771**
	Conservative	214	47	15	**276**
	Other	172	45	85	**302**
	TOTAL	**865**	**265**	**219**	**1349**

Suppose we're interested in calculating our expected number of observations in the top-left cell (Liberals who never used marijuana).

We can first calculate the probability that someone randomly sampled from our data is a Liberal. Since we have a total of 771 Liberals out of our sample of 1349 individuals, the probability is calculated as:

$$\text{P(Liberal)} = \frac{771}{1349} = 0.572$$

We can then calculate the probability that someone from our study has never used marijuana:

$$\text{P(Never)} = \frac{865}{1349} = 0.641$$

Using what we know about independence: $\text{P(Liberal} \cap \text{Never)=P(Liberal)P(Never)} = (.5715)(.6412) = 0.3664$

This is the **expected probability** for our first cell. Since we want to know the expected number of observations, we multiply this probability by the total number of observations in our data.

Expected number of **Liberals** who **Never** used marijuana = $(.3664)(1349) = 494.3$

We can speed-up these calculations using the following:

Calculating Expectations: $\text{Expected} = \dfrac{(\text{row total})(\text{column total})}{\text{total sample size}} = \dfrac{rc}{N}$

Using this formula, we can quickly calculate the other expected numbers of observations. For example, to calculate the expected number of **Conservatives** who **Frequently** use marijuana:

$$\text{Expected} = \frac{(276)(219)}{1349} = 44.81$$

The observed and expected numbers of observations are reported in the table on the next page. I also use the chi-square calculation in each cell.

	Never	Rarely	Frequently
Liberal	O = 479 E = 494.38 $\chi^2 = 0.48$	O = 173 E = 151.46 $\chi^2 = 3.06$	O = 119 E = 125.17 $\chi^2 = 0.30$
Conservative	O = 214 E = 176.98 $\chi^2 = 7.74$	O = 47 E = 54.22 $\chi^2 = 0.96$	O = 15 E = 44.81 $\chi^2 = 19.83$
Other	O = 172 E = 193.65 $\chi^2 = 2.42$	O = 45 E = 59.33 $\chi^2 = 3.46$	O = 85 E = 49.03 $\chi^2 = 26.39$

Our overall test statistic is: $\chi^2_{(3-1)(3-1)} = \chi^2_4 = 0.48 + 7.74 + 2.42 + 3.06 + 0.96 + 0.30 + 19.83 + 26.39 = 64.64$

We compare this to a chi-square distribution with 4 degrees of freedom. A computer calculates the p-value of this test statistic to be $p < .0001$, so we would reject the null hypothesis.

We can, therefore, conclude that there is a relationship between political views and marijuana usage. In order to find the nature of that relationship, we can look at the chi-square values we calculated in each cell. The biggest chi-square values show us the biggest discrepancies between what we observed and what we expected.

For this data, our biggest chi-square values were 26.39 and 19.83. So the relationship between political views and marijuana usage is due to those two cells of the table. Now we must look at the data in those cells to see what is going on.

In the bottom-right cell, we see that we had many more observations that we expected. So we can conclude that individuals with **Other** political views are more likely to **Frequently** use marijuana.

In the cell right above that one, we see that we had fewer observations that we expected. Therefore, we can conclude that **Conservatives** are less likely to **frequently** use marijuana.

This is a clumsy way of finding our conclusions. We can use relative probabilities and odds rates to help our clarify our conclusions.

Relative probabilities:

I'm most interested in the individuals who frequently use marijuana, so I'm going to focus on that column of data.

	Never	Rarely	Frequently	Total
Liberal	479	173	119	771
Conservative	214	47	15	276
Other	172	45	85	302
Total	865	265	219	1349

First, I will calculate the probability that a Liberal frequently uses marijuana: $\frac{119}{771} = 0.154$

Now I will calculate the probability that a Conservative frequently uses marijuana: $\frac{15}{276} = 0.054$

We calculate the relative probability by taking the ratio of these probabilities: $\frac{.154}{.054} = 2.85$.

This relative probability has a simple interpretation. Liberals are 2.85 times more likely to frequently use marijuana than Conservatives.

We could also do the same thing with odds. Recall that odds are calculated: $\text{Odds} = \frac{P(A)}{1 - P(A)}$

The odds that a Liberal frequently uses marijuana are: $\frac{.154}{1 - .154} = 0.182$

The odds that a conservative frequently uses marijuana are: $\frac{.054}{1 - .054} = 0.057$

Therefore, the odds ratio is calculated as: $\frac{.182}{.057} = 3.193$.

The odds of a Liberal frequently using marijuana are more than 3 times higher than the odds of a Conservative frequently using marijuana.