How do we show that two qualitative variables are independent? We will also test if two or more populations experience side effects from a drug at the same rate.

Statistics
Class Notes
Tests for Independence and the Homogeneity of Proportions (Section 12.2)

Tests for Independence:
Earlier we found that employment status was associated with the level of education a person achieved. We explored the percentages of people employed for each level of education (did not finish high school, high school graduate, some college, or Bachelor's degree or higher) and found as the level of education increased, so did the percentage employed. But is this apparent association really **statistically significant**?

We will perform a test very similar to our last section using the **Chi-Square distribution**. What we are aiming to do is to test if the two variables are ***not* associated** or **independent** (where one does *not* affect the other).

Our test will be

H_0: The two variables are *not* associated or independent.
H_1: The two variables are associated or dependent.

How do we go about this? Do you remember the formula for finding the probability that *both* of two independent events happen? Write it here (and rejoice, *you* are the master of all you know).

So, what if we compare the observed values with what they should be, assuming independence? We can use that cool chi-square test statistic we just saw. Here's the plan.

The Test of Independence:
Step 1: Determine the null and alternative hypotheses.

This is the only option for step 1.

H_0: The two variables are *not* associated or independent.
H_1: The two variables are associated or dependent.

Step 2: Select a level of significance, α, depending on the seriousness of making a Type I error.

Step 3a: Calculate the **expected counts**, E_i for each cell in the contingency table. We will refer to the i^{th} cell in the table. This is found by assuming the variables are independent and therefore $P(E\ and\ F) = P(E)\cdot P(F)$. In practice, we will use the formula

$$E_i = \frac{(row\ total)(column\ total)}{table\ total} \text{ for the } i^{\text{th}} \text{ cell in the table.}$$

This is *not* needed if you use technology.

The Test of Independence (continued):

Step 3b: Verify that

1. all expected counts are greater than or equal to 1 (all $E_i \geq 1$), and
2. no more than 20% of the expected counts are less than 5.

If these are *not* met, we can combine two or more columns or rows, or increase the sample size.

Step 3c:

Compute the test statistic $\chi_0^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ where O_i is the *observed* count for the i^{th} cell.

Step 4 (Classical Approach):
Determine the critical value using Table VIII. **All tests for independence are right-tailed**, so the critical value is χ_α^2 with $(r-1)(c-1)$ degrees of freedom (Figure 2). Compare the test statistic to the critical value. If $\chi_0^2 > \chi_\alpha^2$, reject the null hypothesis.

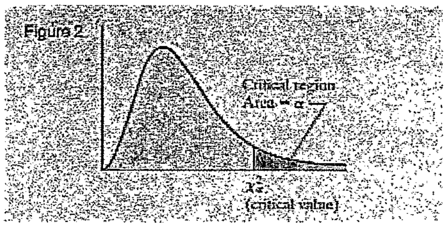


Figure 2

Here, *r* is the number of rows and *c* is the number of columns in the contingency table.

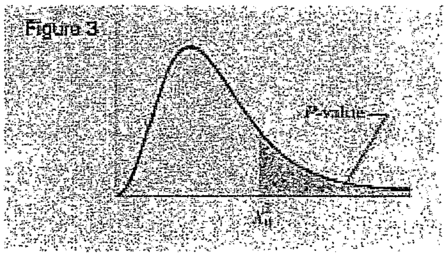


Figure 3

Step 4 (*P*-value Approach):
Use Table VIII or technology to approximate the *P*-value by determining the area under the chi-square distribution with $(r-1)(c-1)$ degrees of freedom to the **right** of the test statistic (Figure 3). If the *P*-value < α, reject the null hypothesis.

Instructions for technology are given on the next page.

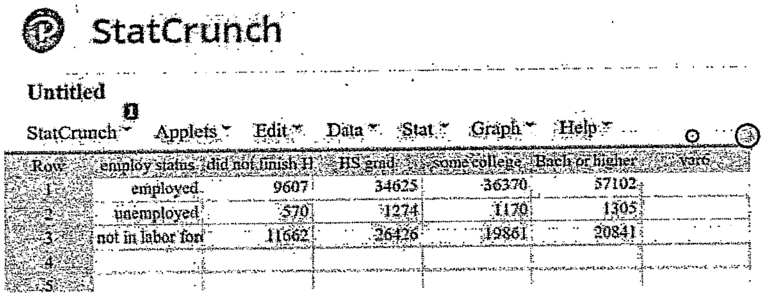
Step 5: State the conclusion.

Remember, if we do *not* reject the null hypothesis, we are *not* saying that it *must* be true. We simply state that we do *not* have evidence to conclude it is as stated in the alternative hypothesis.

If we do reject the null hypothesis, we say we found evidence to support the alternative hypothesis. We do *not* say the null hypothesis must be false.

Instructions for StatCrunch:

1. Enter the contingency table into the spreadsheet. The column marked **var1** is to be the row variable *categories*. Then use the subsequent columns for the data, labeling all columns. The result will look just like your contingency table. Here is a picture of how the data in the next example looks.

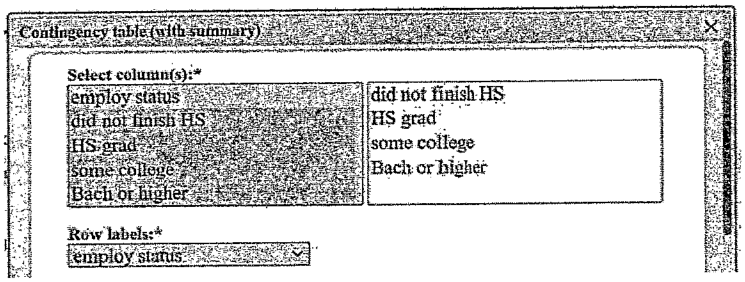


Row	employ status	did not finish H	HS grad	some college	Bach or higher	var6
1	employed	9607	34625	36370	57102	
2	unemployed	570	1274	1170	1305	
3	not in labor for	11662	26426	19861	20841	
4						
5						

I have renamed columns var1 – var5.

2. Select **Stat** > **Tables** > **Contingency**. Then highlight **With Summary**.

3. You will tell it which columns comprise the column variable. Next, tell it the **row** (variable) **labels**. Here is a picture.



Contingency table (with summary)

Select column(s):*

employ status
did not finish HS
HS grad
some college
Bach or higher

did not finish HS
HS grad
some college
Bach or higher

Row labels:*
employ status

4. Under **Display**, select **Expected count** and **Contributions to Chi-Square**. Under **Hypothesis tests**, the default is what we want, **Chi-Square test for independence**. The **Confidence level** does *not* apply to what we are doing; ignore it.

5. Under **Graph** at the bottom, you can show a **heatmap** or not. You do *not* need it. It is a display using varying color hues or intensities to show the magnitude of numbers in a data set.

6. Click **Compute!** (and exclaim Compute! at the top of your lungs as you do so).

7. You might have to scroll down to see the test results. It will be labeled **Chi-Square test**. The **Value** will be the test statistic and the ***P*-value** is also shown.

Instructions for the calculator are given in the book. A method using matrices is given but you can also find the expected counts by hand and enter those along with the observed counts and the degrees of freedom as done in the last section for goodness-of-fit tests. If you do so, calculate the degrees of freedom as described above in step 4.

expl 1: We looked at this data earlier and determined that level of education and employment status were associated. Let's be a bit more rigorous. Test the hypothesis that the variables are independent at the $\alpha = 0.05$ level of significance. Do this in StatCrunch. Be sure to state the conclusion.

Table 10

Employment Status	Level of Education Did Not Finish High School	High School Graduate	Some College	Bachelor's Degree or Higher	Totals
Employed	9607	34,625	36,370	57,102	137,704
Unemployed	570	1274	1170	1305	4319
Not in the Labor Force	11,662	26,426	19,861	20,841	78,790
Totals	21,839	62,325	57,401	79,248	**220,813**

Verify that in step 3b.

Ho: Employment status and level of education are independent.

H1: They are not indpt (dependent).

STATCRUNCH → pvalue < 0.0001 (Value ≈ 7623)

Since pvalue < α = 0.05, we reject Ho and conclude that we have sufficient evidence to say level of education and employment status are not indpt.

One affects the other.

Expected Counts Calculations by Hand:

Doing this in StatCrunch does *not* require us to calculate the expected counts. But some people like to do it themselves. We are told that the $E_i = \dfrac{(row\ total)(column\ total)}{table\ total}$, but why?

Let's use the last example's data to show this. Truly, assuming the null hypothesis,

$P(\text{employed and did not finish HS}) = P(\text{employed}) \cdot P(\text{did not finish HS}) = \dfrac{137,704}{220,813} \cdot \dfrac{21,839}{220,813}$.

For an expected count of these people, we would take this and multiply it by the total number of people, 220,813. Notice, this is equivalent to the formula for E_i given above. It is simply easier to do when we are working by hand.

$\left(\dfrac{\text{row total}}{\text{overall total}} \cdot \dfrac{\text{column total}}{\cancel{\text{overall}}\ \text{total}}\right) \cancel{\text{overall total}} = \dfrac{(\text{row total})(\text{col total})}{\text{overall total}}$ ✓

Tests for Homogeneity of Proportions:
We will compare the population proportions from two or more independent samples. If you have two or more populations in which you want to determine equality of proportions, use this test.

Besides the null and alternative hypotheses being different, the procedures for performing this test are the same as the test for independence. We will replace Step 1 with the following.

This is the only option for step 1.

Step 1: Determine the null and alternative hypotheses.

H_0: $p_1 = p_2 = \ldots = p_n$ (for n populations)
H_1: At least one of the population proportions is different from the others.

expl 2: Dizziness is a common side effect of medicine. The following data are from clinical studies of several medicines used to treat osteoarthritis and rheumatoid arthritis and the numbers of patients who experienced dizziness. We are interested to see if a particular drug causes this side effect more than other drugs. Test at the $\alpha = 0.01$ level to see if the proportions of patients (in the populations) on the different drugs would experience dizziness at the same rates.

Dizziness	**Drug**					
side effect	**Celebrex**	**Placebo**	**Naproxen**	**Diclofenac**	**Ibuprofen**	**Total**
yes	83 $\hat{p}_c \approx 0.0200$	32 $\hat{p}_p \approx 0.0172$	36 $\hat{p}_n \approx 0.0264$	5 $\hat{p}_d \approx 0.0129$	8 $\hat{p}_i \approx 0.0232$	164
no	4063	1832	1330	382	337	7944
Total	4146	1864	1366	387	345	8108

a.) What are the null and alternative hypotheses?

H_0: $p_c = p_p = p_n = p_d = p_i$ (from populations)
H_1: At least one proportion here is not the same as the others.

b.) Find a few expected counts to get a feel for it. You do *not* need to do them all. We will use technology to do the heavy lifting.

$$E_1(\text{celebrex}) = \frac{(\text{row tot})(\text{col tot})}{\text{overall tot}} = \frac{164 * 4146}{8108} \approx 83.86$$

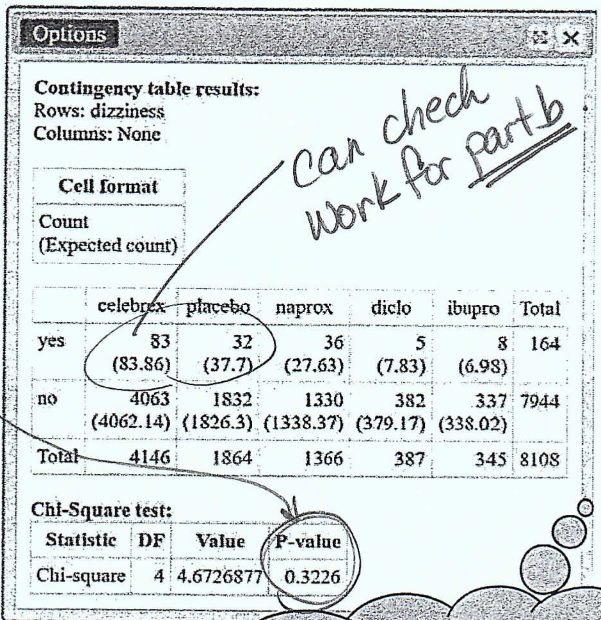
$$E_2(\text{placebo}) = \frac{(\text{row tot})(\text{col tot})}{\text{overall tot}} = \frac{164 * 1864}{8108} \approx 37.70$$

(We'll check on next page)

expl 2 (continued):

c.) Here is the output as given in StatCrunch. What is your conclusion? Do we have evidence that the percentages of patients who experience dizziness on each drug are *not* the same? Write your conclusion properly.

The pvalue ≈ 0.3226 which is not less than $\alpha = 0.01$. Hence, we do not reject Ho and conclude we do not have sufficient evidence to say that any one drug causes dizziness at a different rate than any other drug.



Contingency table results:
Rows: dizziness
Columns: None

Cell format
Count (Expected count)

	celebrex	placebo	naprox	diclo	ibupro	Total
yes	83 (83.86)	32 (37.7)	36 (27.63)	5 (7.83)	8 (6.98)	164
no	4063 (4062.14)	1832 (1826.3)	1330 (1338.37)	382 (379.17)	337 (338.02)	7944
Total	4146	1864	1366	387	345	8108

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	4	4.6726877	0.3226

can check work for part b

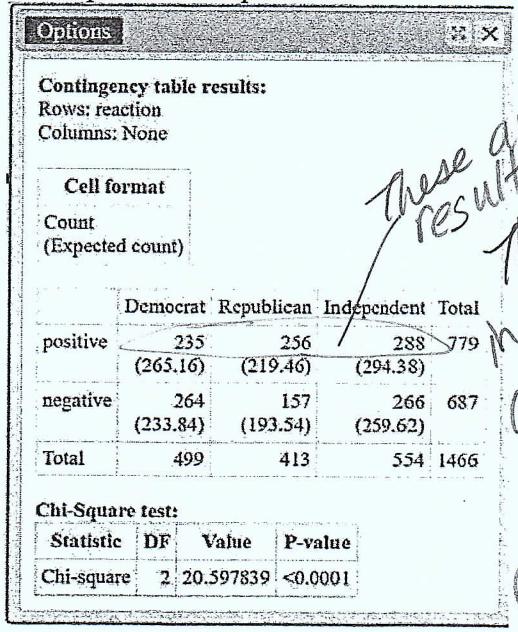
The numbers in parentheses are the expected counts. Check those you computed. ✓

expl 3: Do people on different ends of the political spectrum have different reactions to certain words? A recent Pew Research Center poll asked adult Democrats, Republicans, and Independents if they had a positive or negative view of the word *capitalism*. Below are the results along with the StatCrunch output for the test for homogeneity of proportions. The expected counts are provided in parentheses. What is your conclusion? Is there evidence to support the notion that at least one of the population proportions is different from the others?

$H_0: p_D = p_R = p_I$ (has positive view)

H_1: At least 1 prop is not the same.

Since p-value < 0.0001, we'll reject Ho at the $\alpha = 0.01$ (or more) level. We do have sufficient evidence to say at least one party has a different proportion.



Contingency table results:
Rows: reaction
Columns: None

Cell format
Count (Expected count)

	Democrat	Republican	Independent	Total
positive	235 (265.16)	256 (219.46)	288 (294.38)	779
negative	264 (233.84)	157 (193.54)	266 (259.62)	687
Total	499	413	554	1466

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	2	20.597839	<0.0001

These are poll results.

The numbers in parentheses are expected counts assuming independence

For each party, find the proportion who had a positive reaction to the word *capitalism*. Do these numbers support your conclusion above?

$\hat{p}_D = 235/499 \approx 0.471$

$\hat{p}_R = 256/413 \approx 0.620$

$\hat{p}_I = 288/554 \approx 0.520$

We can see these proportions are not very close to each other.