Chi-Square

Donna Dietz

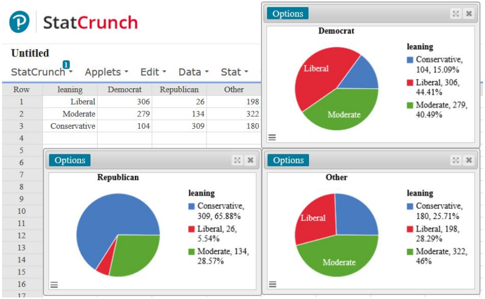
American University

dietz@american.edu

STAT 202 - Spring 2020

When would you use Chi-Square?

Recall earlier, we had sets of pie-charts like these ones:



Untitled

Row	leaning	Democrat	Republican	Other
1	Liberal	306	26	198
2	Moderate	279	134	322
3	Conservative	104	309	180

Although the data are fabricated, they represent the types of questions we want to address. If you line up several groups which are subdivided into the same sub-categories, are the proportional variations meaningful between the groups? Are Democrats more likely to be liberal? Are Republicans more likely to vote? Who is more likely to use mail-in ballots? All these things can be analysed with a Chi-Square test.

Roadmap

This discussion consists of three main parts, and they fit together in the shape of the letter

$$Y.$$

- What is the *Chi-Square* family of distributions?
- What is the *Chi-Square* statistic?
- What is the *Chi-Square* test?

You ought to know what this family of distributions is, and if you're going to even play around a little bit with it by hand, you need to know how to calculate the statistic. The test is a process by which you look up the value you just calculated, on a table which represents this distribution, and this generates a $p - value$ that tells you how unlikely it would be to get such an extreme result if your Null is true.

One tail

The Chi-Square family is a one-tail family, and it will soon make sense why. (But if you really think about why, you may only become more convinced as I am, that one-tail tests on the Normal curve are always nonsensical.)

What is χ and why did we square it?

In the study of statistics, there seems to be a lot of vocabulary. That's because every time you do something to a variable or a distribution, statisticans give it a new name!

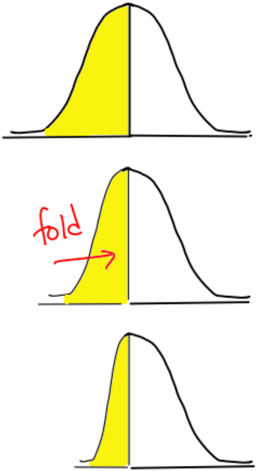
So what is χ?

"Chi", which rhymes with 'sky' and sounds like the name "Kai", χ is the Greek letter "ch" as in "choromatic" or "archeology" or "chaos". (Chai is a beverage and starts with the same sound as cheese, but that is a different 'ch' sound!)

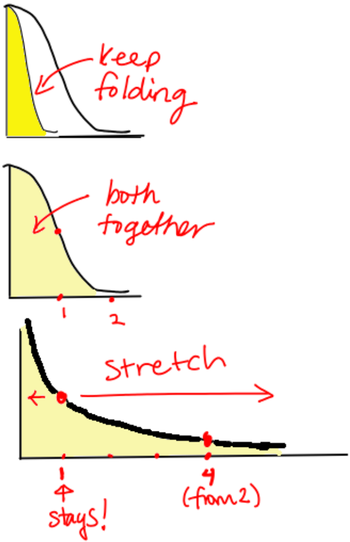
χ^2 revealed

χ^2 revealed

Folding the Normal Curve



Folding and Stretching the Normal Curve



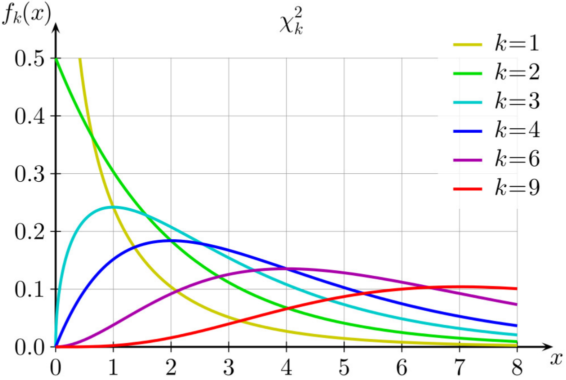
χ^2

The χ^2 family of distributions begins with this one, which is just found by squaring the Normal Distribution. It's as if you took every item in the Standardized Normal distribution and squared it. The height of the distribution we had at z=1 will be double that height on the χ^2 graph at $x = 1$. The height that was at z=2 will be doubled and go to 4. The height that was at z=0.5 doubles and goes to 0.25 and so forth.

What you know from the Empirical Rule is useful here! You know that 68% of the time, you are to the left of $x = 1$, so 32% of the time, you are in the tail to the right of $x = 1$.

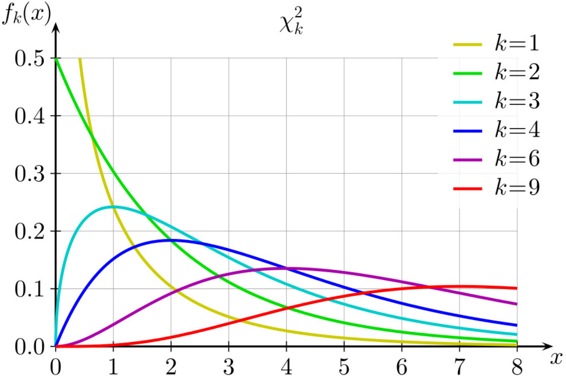
Notice that you are doing a one-tail test. However, this is a folded form of the Normal Distribution, so a one-tail test is automatically like a two-tail test!

The whole family!



χ^2_1 is the distribution for z_1^2 if z_1 is sampled randomly from N(0,1).

Adding in the whole family!

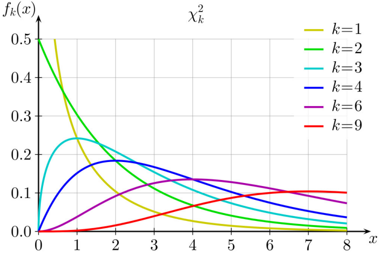


The $k = 2$ curve is

$$z_1^2 + z_2^2$$

as just explained, if z_1 and z_2 are from N(0,1).

And of course...



The $k = 3$ curve comes from $z_1^2 + z_2^2 + z_3^2$.

The $k = 4$ curve comes from $z_1^2 + z_2^2 + z_3^2 + z_4^2$.

Recall

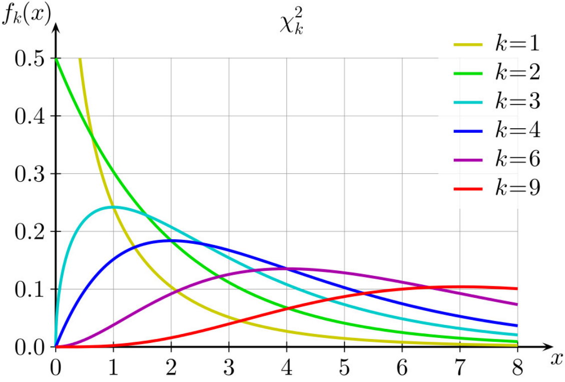
Recall how previously, we have done exercises where we added two or more independent random Normal variables:

$$N(a, b) + N(c, d) = N(a + c, \sqrt{b^2 + d^2})$$

This should feel similar in some way, because after the squaring step, you are just adding the pieces together, and each piece you are adding is just from χ_1^2. In our diagram, that is called $f_1(x)$.

Hard to estimate areas

Using this diagram, it's hard for you to estimate things like, "What is the expected value for $k = n$?"

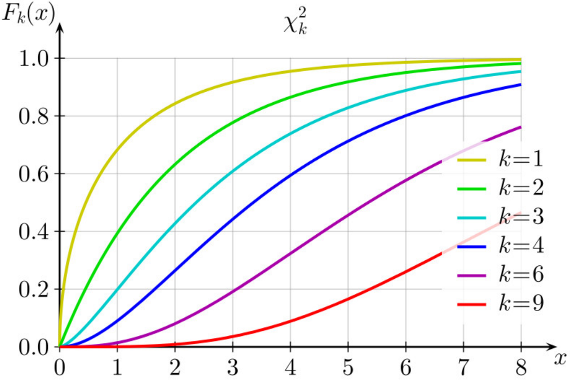


Chi-Square Distribution

Since it's hard to estimate areas...

Easier to estimate areas

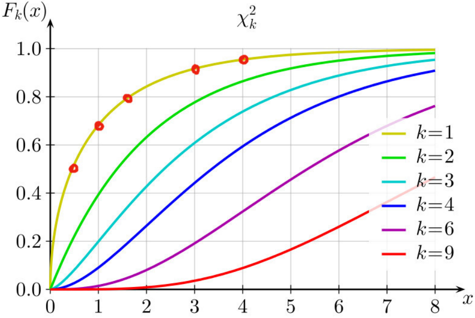
We have another way of expressing these distributions. It's a visual form of the χ^2 Tables.



Cumulative Distribution Function

Using this diagram, you can estimate things like,
"What is the expected value for $k = n$?"

Easier to estimate areas



Cumulative Distribution Function

First, verify that the $k = 1$ curve hits about 0.68 when $x = 1$. That agrees with the Empirical Rule. It also hits about 0.95 when $x = 4$.

It hits the 0.5 mark near $x = 0.5$. This means that roughly half population of N(0,1) when squared will be under one half! About 80% is under 1.6 and 90% is below 3.

CLT

Central Limit Theorem

Any distribution when sampled randomly and independently, will eventually limit to a Normal Distribution.

We saw this already when the Binomial distributions converged to the Normal. So, a distribution with just two bars (0 and 1, or success and failure) did this! All distributions have this property, and χ_1 is no exception!

Note: Mathematicans could probably come up with counterexamples, but they would not apply to real world problems!

Long term behavior

$N(k, \sqrt{2k})$

As k grows, the χ^2_k mean gets closer to k and the standard deviation gets closer to $\sqrt{2k}$.

Like our binomial distribution, this is similar to "counts" or "totals" rather than averages. You can, of course, divide by k to find the averages if you wish. We don't need to do that in this course.

Computers and Tables

Just as with our other distributions, you have paper tables as well as StatCrunch tools that give you progressively more accurate values for calculation purposes. But, there is nothing like a picture to help you understand what's happening!

Transition to the Statistic

You will get practice on the worksheet with looking at the graphic, and the table, and the StatCrunch tool. Now, we turn our discussion to the statistic we will be calculating, and associated concepts.

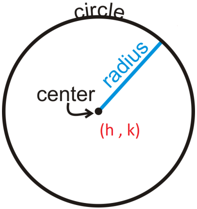
After that, we'll put those ideas together.

Degress of Freedom

df

We use the concept of *degrees of freedom* frequently enough in Statistics that we have an abbreviation for it: *df*. 'Degrees of freedom' means simply, "How many choices do we have?"

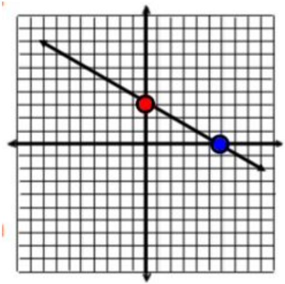
You can fully define a circle with a center (which has two coordinates) and a radius, for a total of three choices. So, a circle has three degrees of freedom.



Degrees of Freedom

A line on the x-y plane can be drawn by defining a y-intercept and a slope, which is two choices or two degrees of freedom.

But you may have instead drawn a line by placing a point on each of the two axes, but you still only made two choices, so this only confirms that a line has two degrees of freedom.



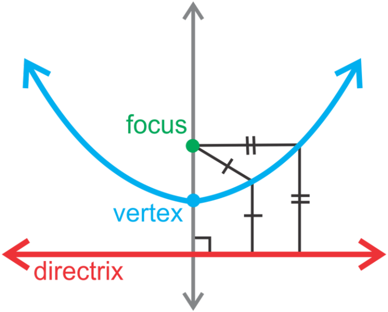
There are often multiple ways of describing your choices, but the total number of choices remains the same.

A parabola

A quadratic, which is really a parabola with its line of symmetry parallel to the y-axis, can be defined using a vertex (2 choices) and a directrix (one choice, since it has to be horizontal).

This same quadratic can be defined by choosing a, b, and c in the formula $y = ax^2 + bx + c$, but you still have just 3 choices to make.

3 choices



Focus and Directrix for Parabola

You have 3 decisions to make for a parabola of this type. You choose the focus, and your directrix will have the form $y = C$ where you get to choose C. (The vertex is always halfway between the focus and the directrix!)

What does *df* mean in this context?

	A	B	Total
P			80
Q			220
Total	100	200	300

What does *df* mean in this context?

	A	B	Total
P	**?**		80
Q			220
Total	100	200	300

You may choose the number of items which exhibit properties A and P, but once you do that, you have no more freedom. Likewise, you could have picked another cell to fill, but once you do, you're done. So, here, **you have one degree of freedom**.

What about this?

	A	B	C	D	E	F	G	Total
P								33
Q								30
R								27
Total	8	12	15	11	14	17	13	90

What about this?

	A	B	C	D	E	F	G	Total
P	?	?	?	?	?	?		33
Q	?	?	?	?	?	?		30
R								27
Total	8	12	15	11	14	17	13	90

One way to think of this is to ask how many cells you could fill in before the rest of the cells are determined. In this case, we have 12 cells that we can fill before the rest are determined. So we had **12 degrees of freedom**. We may have chosen different cells to fill, but we would have selected a total of 12. We do have to be careful which 12 we choose, because we could not, for example, choose all the cells in column A. But let's presume we only make selections which are non-conflicting. So we have 12 choices. Note: $(7-1)\cdot(3-1)=12$.

The statistic itself

In order to do a Chi-Square test, you need Chi-Square tables, we need a Chi-Square *statistic* that we will look up on a table or computer. How do we calculate this?

The χ^2 statistic

χ^2

Add together the χ^2 contributions for each cell:

$$\frac{(Observed - Expected)^2}{Expected}.$$

Example

Let us do this one together:

Color	Candies	Cookies	Total
Yellow	10	20	
Black	20	30	
Total			

We have a total of 80 colored food items.
There are $\frac{30}{80}$ Yellow items and $\frac{30}{80}$ Candies, so we would expect a ratio of $\frac{30\cdot 30}{80\cdot 80} = 0.140625$ or 11.25 Yellow Candies.
We instead see we have slightly fewer, or 10.

$$\frac{(11.25 - 10)^2}{11.25} = 0.1389.$$

Example

Color	Candies	Cookies	Total
Yellow	10	20	
Black	20	30	
Total			

Recall our Expected Values shortcut:
Row sum times column sum divided by overall sum.

Expected Yellow Candies = (30*30)/80 = 11.25
Expected Yellow Cookies = (30*50)/80 = 18.75
Expected Black Candies = (50*30)/80 = 18.75
Expected Black Cookies = (50*50)/80 = 31.25

Example

Now χ^2 contributions:

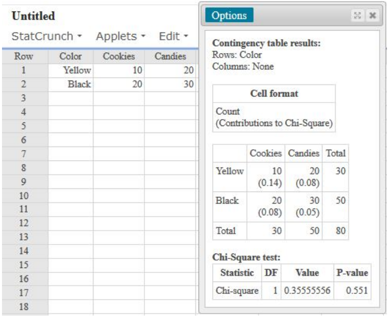
Color	Candies	Cookies	Total
Yellow	10	20	
Black	20	30	
Total			

Yellow Candies: $(11.25 - 10)^2/11.25 = 0.1389$
Yellow Cookies: $(18.75 - 20)^2/18.75 = 0.0833$
Black Candies: $(18.75 - 20)^2/18.75 = 0.0833$
Black Cookies: $(31.25 - 30)^2/31.25 = 0.05$

Add those: 0.3555

On StatCrunch

Stat > Tables > Contingency > With Summary
Select columns: Cookies, Candies
Row Variable: Color
Display: Contibutions to Chi-Square



Row	Color	Cookies	Candies
1	Yellow	10	20
2	Black	20	30
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			

Contingency table results:
Rows: Color
Columns: None

Cell format
Count (Contributions to Chi-Square)

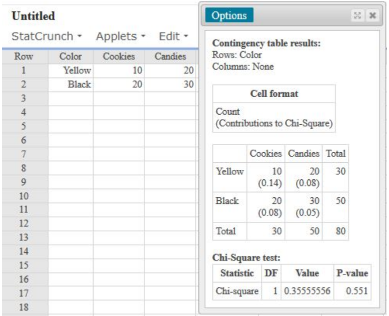
	Cookies	Candies	Total
Yellow	10 (0.14)	20 (0.08)	30
Black	20 (0.08)	30 (0.05)	50
Total	30	50	80

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	0.35555556	0.551

Consistent with intuition, this is not significant, as you can see from the p-value.

Verify the *df* and the Value.



Row	Color	Cookies	Candies
1	Yellow	10	20
2	Black	20	30
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			

Contingency table results:
Rows: Color
Columns: None

Cell format
Count (Contributions to Chi-Square)

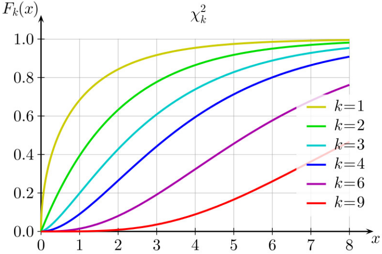
	Cookies	Candies	Total
Yellow	10 (0.14)	20 (0.08)	30
Black	20 (0.08)	30 (0.05)	50
Total	30	50	80

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	0.35555556	0.551

I think we agree, right? If we have four counts with four forced totals, we can only make one choice, like saying we have 10 yellow cookies. Once we make that choice, the other 3 categories are set. Notice they were all off by exactly the same count value of 1.25 (but not the same percent)!

Looking up the value.



Since *df* $= 1$ in this situation, look at that curve. But this curve is backwards from what we're going to want for a p-value. This gives the total area, not the remaining area! For our statistic value of 0.3555, we should estimate that the area captured thus far is near 0.4. You can't do much better than that on this graphic. That's close enough. That would give a remaining tail of roughly 0.6 which is not remotely significant. This graphic is quite accurate enough to allow you to reject the hypothesis that the color distribution between the cookies and candies is anything other than random.

Use the paper table!

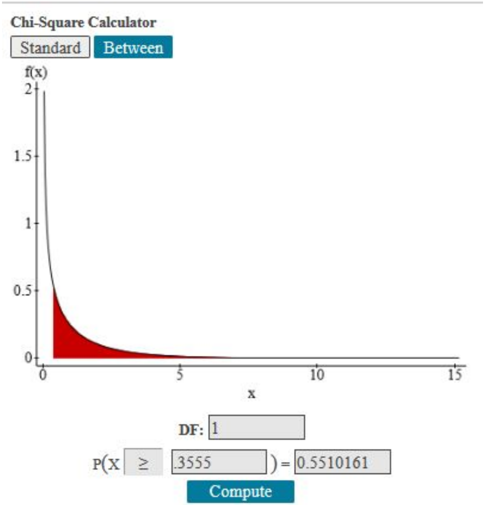
You will notice that a value of 0.3555 doesn't appear in the table anyhwere. The lowest value on the paper table anywhere at all is 1.32.

This is not inconclusive! On the contrary, we get the exact same result as we did before, which is to say that the tail probability is too high to do anything other than fail to reject the null. We have no evidence to make a claim of difference between these groups.

And StatCrunch?

We can, of course, use technology to generate uselessly accurate values.

Stat > Calculators > Chi-Square



DF: 1

P(X ≥ .3555) = 0.5510161

Why does that work?

It may not be obvious why this process that we use to calculate the chi-square statistic is equivalent to adding up n copies of z^2. However, if you are interested in seeing the algebraic steps in detail, Wikipedia has a great article on this and explains it nicely.

https://en.wikipedia.org/wiki/Chi-squared_distribution

One tail or two?

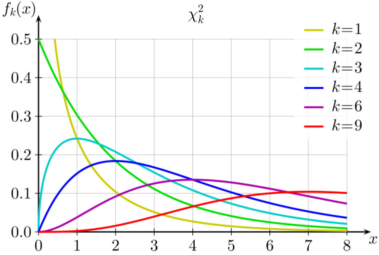
For the χ^2 family, when we cut off the tail on the high end, this is really like a two-tail test that we would do on a Normal distribution. Is it clear to you why that is?

When we squared the Normal distribution, we combined the two tails into one tail, on right side.

Example: If the average of a sample is 10, the contribution to the χ^2 statistic from 9 or 11 will be the same. The contribution from 5 or 15 will be the same. What we are looking at is based off the difference to the mean, so being on the right or left side doesn't matter. All the contributions get added up.

But they do have two tails!

Yes, many of the χ^2 curves have two tails, however, if you are anywhere to the left of that maximum "hump", your p-value is insignificant, unless for some reason you're trying to demonstrate that someone's data are *too perfect*. However, I have (strangely) never heard of any statistician being accused of fabricating data because they're too perfect.



Simpson's Paradox

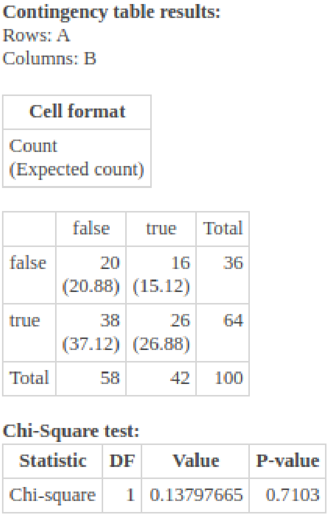
You now have the tools to fully explore the Simpson's Paradox exercises. In addition to the analysis you've done, you can test both the split data as well as the combined data to see if there is evidence that the effects you are claiming to see are even meaningful or not!

A progression

As before with LSR, let's investigate a progression where two things are not related in any way, but they progress until they're totally related.

The truth of A will remain constant, but the B truth values will slowly start to agree more and more with A's truth values.

Progression 1: No interaction



Contingency table results:
Rows: A
Columns: B

Cell format
Count (Expected count)

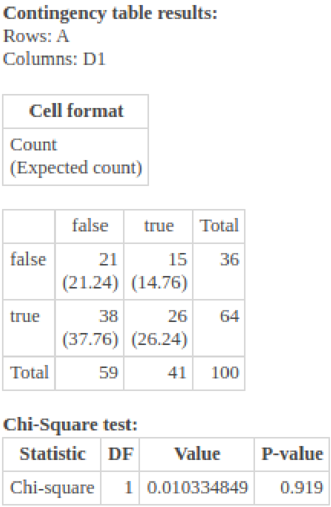
	false	true	Total
false	20 (20.88)	16 (15.12)	36
true	38 (37.12)	26 (26.88)	64
Total	58	42	100

Chi-Square test:

Statistic	**DF**	**Value**	**P-value**
Chi-square	1	0.13797665	0.7103

A can be true or false, and so can B. Currently there is no interaction.

Progression 2



Contingency table results:
Rows: A
Columns: D1

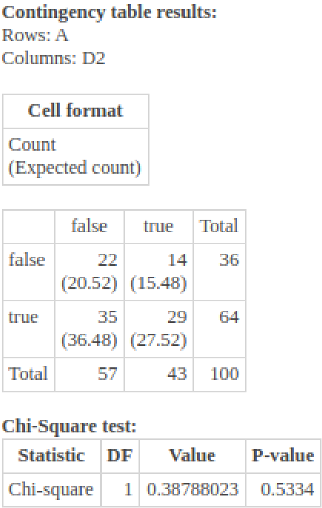
Cell format
Count (Expected count)

	false	true	Total
false	21 (21.24)	15 (14.76)	36
true	38 (37.76)	26 (26.24)	64
Total	59	41	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	0.010334849	0.919

Progression 3



Contingency table results:
Rows: A
Columns: D2

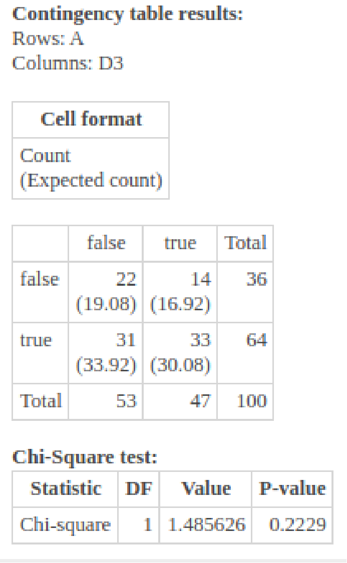
Cell format
Count (Expected count)

	false	true	Total
false	22 (20.52)	14 (15.48)	36
true	35 (36.48)	29 (27.52)	64
Total	57	43	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	0.38788023	0.5334

Progression 4



Contingency table results:
Rows: A
Columns: D3

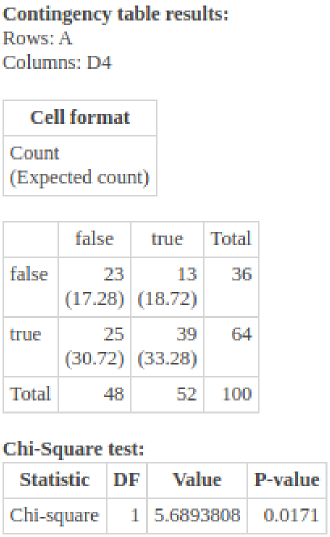
Cell format
Count (Expected count)

	false	true	Total
false	22 (19.08)	14 (16.92)	36
true	31 (33.92)	33 (30.08)	64
Total	53	47	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	1.485626	0.2229

Progression 5



Contingency table results:
Rows: A
Columns: D4

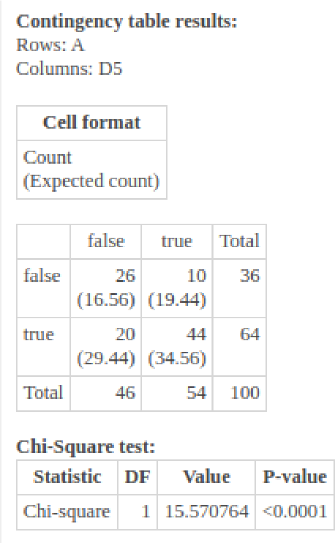
Cell format
Count (Expected count)

	false	true	Total
false	23 (17.28)	13 (18.72)	36
true	25 (30.72)	39 (33.28)	64
Total	48	52	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	5.6893808	0.0171

Progression 6



Contingency table results:
Rows: A
Columns: D5

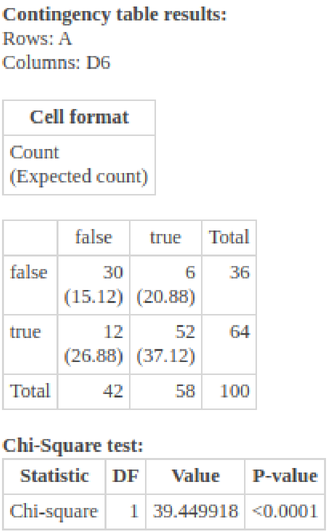
Cell format
Count (Expected count)

	false	true	Total
false	26 (16.56)	10 (19.44)	36
true	20 (29.44)	44 (34.56)	64
Total	46	54	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	15.570764	<0.0001

Progression 7



Contingency table results:

Rows: A

Columns: D6

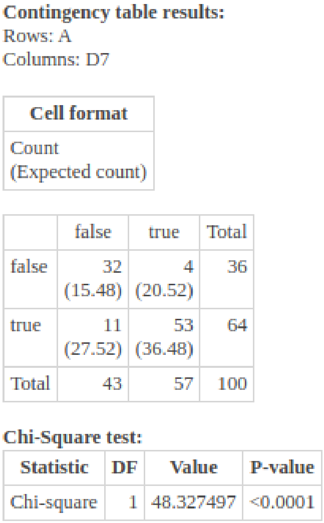
Cell format
Count (Expected count)

	false	true	Total
false	30 (15.12)	6 (20.88)	36
true	12 (26.88)	52 (37.12)	64
Total	42	58	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	39.449918	<0.0001

Progression 8



Contingency table results:
Rows: A
Columns: D7

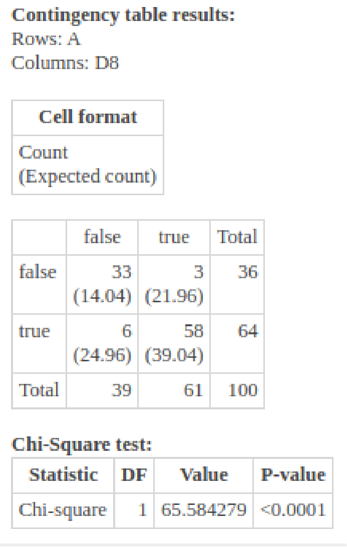
Cell format
Count (Expected count)

	false	true	Total
false	32 (15.48)	4 (20.52)	36
true	11 (27.52)	53 (36.48)	64
Total	43	57	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	48.327497	<0.0001

Progression 9



Contingency table results:

Rows: A

Columns: D8

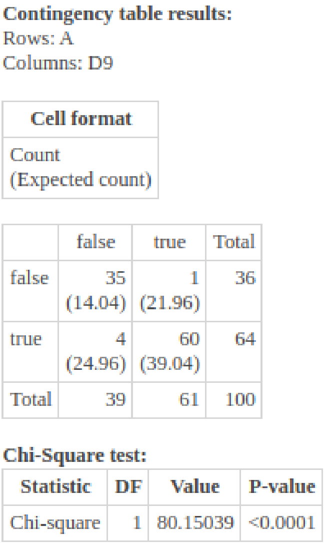
Cell format
Count (Expected count)

	false	true	Total
false	33 (14.04)	3 (21.96)	36
true	6 (24.96)	58 (39.04)	64
Total	39	61	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	65.584279	<0.0001

Progression 10



Contingency table results:

Rows: A

Columns: D9

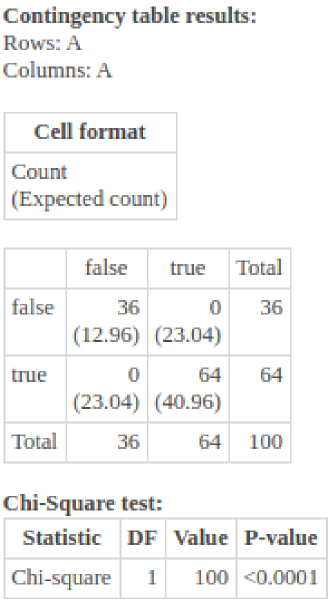
Cell format
Count (Expected count)

	false	true	Total
false	35 (14.04)	1 (21.96)	36
true	4 (24.96)	60 (39.04)	64
Total	39	61	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	80.15039	<0.0001

Progression 11



Contingency table results:
Rows: A
Columns: A

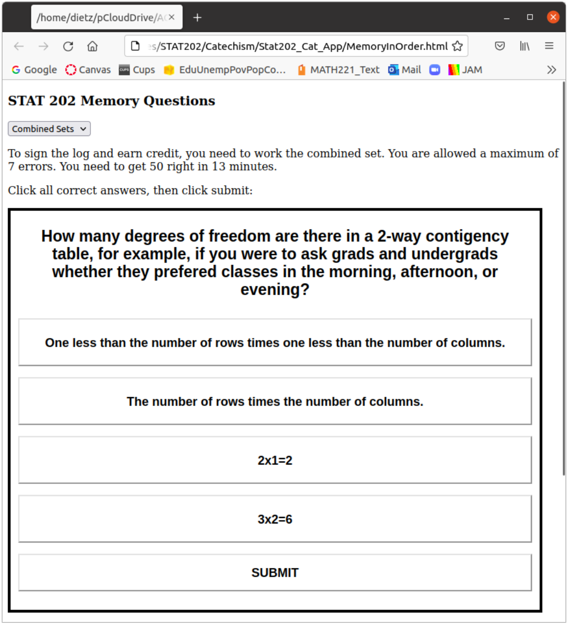
Cell format
Count (Expected count)

	false	true	Total
false	36 (12.96)	0 (23.04)	36
true	0 (23.04)	64 (40.96)	64
Total	36	64	100

Chi-Square test:

Statistic	**DF**	**Value**	**P-value**
Chi-square	1	100	<0.0001

MEMORY QUESTION


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
Combined Sets

STAT 202 Memory Questions

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

How many degrees of freedom are there in a 2-way contigency table, for example, if you were to ask grads and undergrads whether they prefered classes in the morning, afternoon, or evening?

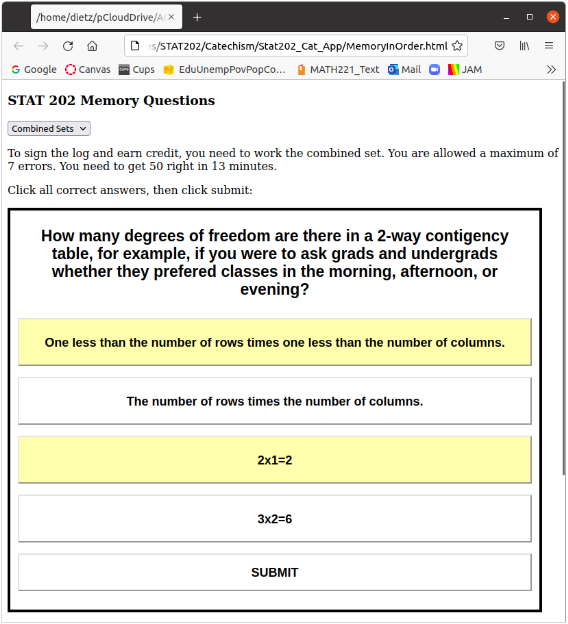
One less than the number of rows times one less than the number of columns.

The number of rows times the number of columns.

2x1=2

3x2=6

SUBMIT



/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

How many degrees of freedom are there in a 2-way contigency table, for example, if you were to ask grads and undergrads whether they prefered classes in the morning, afternoon, or evening?

One less than the number of rows times one less than the number of columns.

The number of rows times the number of columns.

2x1=2

3x2=6

SUBMIT