Only Two More Weeks!!

Unit 10 – Contingency Tables

Almost-Made-It-Through Professor Colton

Unit 10 - Outline

<u>Unit 10 – Contingency Tables</u>

Intro

Chi Square Test for Dependence

- Observed and Expected Matrices
- Hypotheses Statements
- Test Statistic: Chi-Square Test and p-value
- Examples

Review

<u>Contingency Tables</u>

- Contingency Tables helped us organize data on two variables!
- We used them to find probabilities such as: P(Statistics), P(Art and Poor Attendance), P(Good Attendance | Chemistry), etc.

	Statistics	**Art**	**Chemistry**	***Total***
Perfect	100	40	80	220
Good	20	50	70	140
Poor	30	15	30	75
Total	150	105	180	435

- We also learned about relationship between EVENTS, such as Perfect Attendance and Statistics.
  - Are these events mutually exclusive?? NO
  - How about independent??

<u>Independence of EVENTS</u>

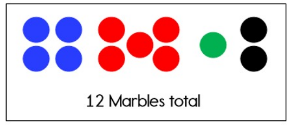
- Two EVENTS were independent if the prior EVENT had NO effect on the subsequent EVENT.
- If this is true, the first EVENT does NOT change the probability of the second EVENT!
  - We could write this in terms of conditional probability:

$$P(B|A) = P(B) \quad ==> \quad P(A \text{ and } B) = P(A) \times P(B|A), \text{ simplify now!}$$

$$= P(A) \times P(B)$$

LCQ: Multiplication Rule for Independent Events

Setup: If I have a jar of colored marbles and want to select <u>two</u> of them <u>with replacement</u>.



1) Are these events independent or dependent?

Independent → so now we can use the simpler multiplication rule!

2) What is the probability both are blue?

P(Success) = P(Blue) = 4/12 → P(Blue and Blue) = P(Blue) x P(Blue) = 4/12 x 4/12 = 0.11

New

Independence of VARIABLES

- Now we are going to study the **independence** of two entire VARIABLES, not just individual events

	Statistics	Art	Chemistry	*Total*
Perfect	100	4	80	220
Good	20	50	70	140
Poor	30	15	30	75
Total	150	105	180	435

Comparison

- Statistics OVERALL (globally) has P() = 150/435 ≈ 0.34
- Is this probability similar to Statistics JUST WITHIN (locally) Perfect Attendance???
- P(Statistics | Perfect) = 100/220 ≈ 0.45?? Is this close enough??

- For this example, our variables are Attendance and Major.
  - Are Attendance and Major related / associated??? Is there a dependence relationship here?? Or are they independent??
- To answer this, we need to actually look at the ALL of the relationships between events of each variable simultaneously (at the same time)!!
  - So that means analyzing: Perfect Attendance and Statistics, Perfect Attendance and Art,… Poor Attendance and Chemistry ALL AT THE SAME TIME!!
  - How to think about this → We are comparing the Global data (column / row totals) to the Local data (middle cells), are the patterns the same??
  - Sounds complex, but we actually have a nice Hypothesis Test that will do this!
- Answering the questions above could be useful and is a very common interest in practice!
  - Let's say the University was reviewing their attendance policy when all classes went virtual, should they have a department specific rule??
  - Or is a University-wide rule effective enough?

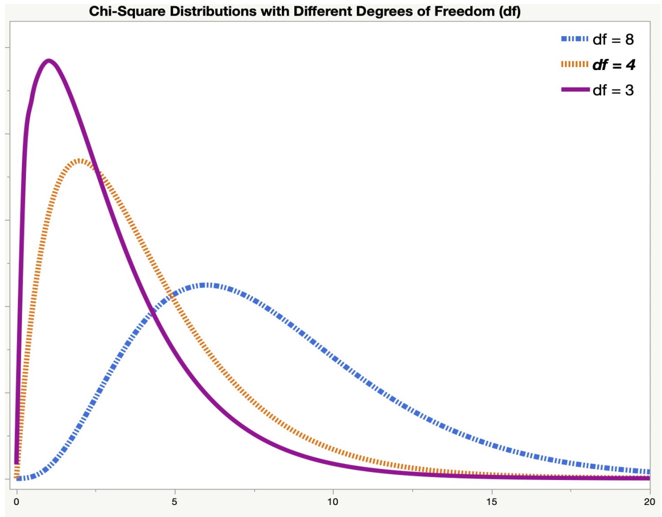
Chi-Square Test of Independence χ^2

<u>Chi-Square Test of Independence</u>

- The formal name of the test we will be doing is a **Chi-Square Test of Independence** (X^2 Test of Independence)
    - This test is based on the Chi-Square (X^2) distribution, hence the name
    - It is a right-skewed, continuous distribution that has a degrees of freedom parameter

- This test determines if two categorical variables are associated or not

*** We aren't going to be checking any assumptions, we just need two categorical variables!*



https://www.jmp.com/en_ca/statistics-knowledge-portal/chi-square-test/chi-square-distr

<u>Process</u>

- We start by assuming two variables are unrelated. Here are some examples:

    1) Movie genre (variable 1) and Snacks (variable 2)
        - Our idea is that the type of movie someone goes to see and whether or not they purchased snacks is unrelated
        - If true, easier to estimate how many snacks will be sold on any given night because what showings are available wouldn't have an impact
    2) Dog breed (variable 1) and Brand of food (variable 2)
        - We think the breed of dog a family has and the dog food brand they buy are unrelated.
        - If true, a store wouldn't have to market small dog or big dog food

- Then we look at the contingency table of the collected data and evaluate our assumption by <u>comparing what actually happened</u> to <u>what should have happened</u> *if they were indeed unrelated*!!

- Said another way, the independence test checks to see if the actual data is "close enough" to the expected counts that would occur if the two variables are independent → Let's demonstrate!

https://www.jmp.com/en_au/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html

Hypotheses

Logic

- This is pretty much the SAME thing that we have been doing with Hypothesis Tests for Proportions and Means
- Except now we are studying the relationship between two variables (not parameters), specifically testing for **independence** between the row and column variables of a contingency table

Hypotheses

- Null Hypothesis → We start by assuming the row and column variables are **independent** (i.e. **NOT related**)
- Alternative Hypothesis → Then we are trying to show the opposite, that the row are column variables are **NOT independent**, or **dependent** (i.e. **related**)
- So in general:

  H_0: The row and column variables are independent

  H_A: The row and column variables are dependent

- But we NEED to add CONTEXT for our problem!
- Examples:

  1) H_0: Movie genre and Snacks are independent

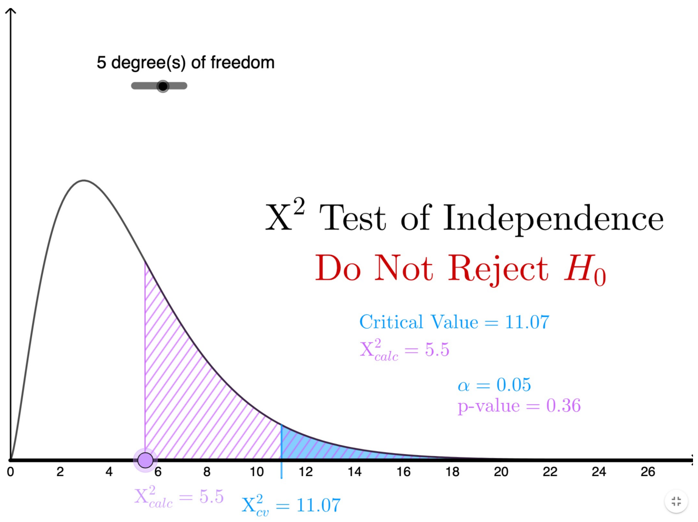
     H_A: Movie genre and Snacks or not are NOT independent (or dependent)

  2) H_0: Dog breed and Type of food are NOT related

     H_A: Dog breed and Type of food ARE related

Test Statistic and P-value

https://www.geogebra.org/m/smhy8cxz



Test Statistic

- The **Test Statistic** X^2_{stat} has the following formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \qquad \text{with } df = (r-1)(c-1)$$

where:
O = Observed count
E = Expected count

where:
r = number of rows
c = number of columns

- The Observed count is our sample data in the contingency table
- The Expected count is what should have happened *if our two variables were indeed unrelated*!! In other words, the counts under the Null Hypothesis!

P-Value

- Chi-Square Test for Independence is always RIGHT-tailed
- So the p-value is the probability of getting our Test Statistic or greater

Calculator

- We are going to use the calculator to calculate everything! Phew!
- But we need to understand what is happening behind the scenes first!

Decisions

- We will make our decisions to Reject or Fail to Reject using the p-value method so that we don't have to find the Critical Value for the X^2 distribution

Expected Counts

- Lets go through how **expected counts** are calculated very slowly with the following example
- Here are our hypotheses:

  H_0: Attendance and Major are unrelated

  H_A: Attendance and Major are related

	Statistics	Art	Chemistry	*Total*
Perfect	100	40	80	220
Good	20	50	70	140
Poor	30	15	30	75
Total	150	105	180	435

Process

- We are going to use probabilities to calculate the expected counts, so lets first think about the marginal probabilities

	Statistics	Art	Chemistry	*Total*	
Perfect				220/435 0.51	P(Perfect) = 0.51
Good				140/435 0.32	P(Good) = 0.32
Poor				75/435 0.17	P(Poor) = 0.17
Total	150/435 0.34	105/435 0.24	180/435 0.41	435/435 1	
	P(Statistics) = 0.34	P(Art) = 0.24	P(Chemistry) = 0.41		

- Now we need to fill in the probabilities in the middle (joint probabilities, example: P(Perfect AND Statistics)). Well under the Null hypothesis, Attendance and Major are unrelated
  - In probability, that means they are independent!!
  - How can we calculate the probability of independent events???? Just multiply the marginal probabilities of each!
    - P(Perfect and Statistics) = P(Perfect) x P(Stats)

Expected Counts

- So lets do that for each of the middle cells!
- P(Perfect AND Statistics) = P(Perfect) x P(Statistics), … , P(Poor AND Chemistry) = P(Poor) x P(Chemistry)

	Statistics	Art	Chemistry	*Total*	
Perfect	0.34 x 0.51 0.17	0.24 x 0.51 0.12	0.41 x 0.51 0.21	220/435 0.51	P(Perfect) = 0.51
Good	0.34 x 0.32 0.11	0.24 x 0.32 0.08	0.41 x 0.32 0.13	140/435 0.32	P(Good) = 0.32
Poor	0.34 x 0.17 0.06	0.24 x 0.17 0.04	0.41 x 0.17 0.07	75/435 0.17	P(Poor) = 0.17
Total	150/435 0.34	105/435 0.24	180/435 0.41	435/435 1	
	P(Statistics) = 0.34	P(Art) = 0.24	P(Chemistry) = 0.41		

- Now to get the counts, what can we do???
- We have the PERCENT of students with Perfect Attendance and Statistics Major

P(Perfect and Stats) = 0.17

- Now I want the actual NUMBER of students in this group! Logically, it would make sense to just multiply this PROBABILITY by the TOTAL number of students in the sample. Correct!!

Expected Count of Perfect and Stats = 0.17 x 435 = 73.95

Expected Counts

- So lets do that!!

	Statistics	Art	Chemistry	*Total*	
Perfect	0.17 x 435 73.95	0.12 x 435 52.2	0.21 x 435 91.35	220	*Observed row totals*
Good	0.11 x 435 47.85	0.08 x 435 34.8	0.13 x 435 56.55	140	
Poor	0.06 x 435 26.1	0.04 x 435 17.4	0.07 x 435 30.45	75	
Total	150	105	180	435	
	Observed column totals				

- We know have our table of **Expected Counts**
- These counts represent what we would expect to see if there truly is no relationship between Attendance and Major.
  - Restated: If their probabilities were independent, then we should have observed these new counts based on the original row and column totals
  - Ex) If there is not a relationship between Attendance and Major, we would expect 73.95 students to have Perfect attendance and be Statistics majors!

  - Now, our Global data matches the Local data, same proportions! → Our marginal probabilities are the same as our conditional probabilities!

    - Ex) P(Statistics) = P(Statistics | Perfect)

$$\frac{150}{435} = 35\% \quad \approx \quad 34\% = \frac{73.95}{220}$$

	Statistics	Art	Chemistry	*Total*
Perfect	100	40	80	220
Good	20	50	70	140
Poor	30	15	30	75
Total	150	105	180	435

Compared to originally (observed): P(Statistics | Perfect) $\frac{100}{220} = 0.45$

and P(Good) = P(Good | Chemistry)

$$\frac{140}{435} = 32\% \quad \approx \quad 31\% = \frac{56.55}{180}$$

*** Note that the numbers,*

- *So these co*

Expected Counts

- Another way to get the Expected counts is by using this formula:

$$\textbf{Expected Count} = \frac{(\textbf{row total})(\textbf{column total})}{\boldsymbol{n}}$$

- <u>This is the formula you will see in every resource</u>. We just went the long way first to show what is actually happening

- If we look back at our calculations, we will see that we were implicitly doing this formula!

<u>Demonstration</u>

- For Perfect and Statistics, the we found the probability by multiplying the two marginal:

$$P(\text{Perfect and Stats}) = \frac{220}{435}\left(\frac{150}{435}\right) = 0.17$$

	Statistics	Art	Chemistry	*Total*	
Perfect	0.34 x 0.51	0.24 x 0.51	0.41 x 0.51	220/435	P(Perfect) = 0.51
	0.17	0.12	0.21	0.51	
Good	0.34 x 0.32	0.24 x 0.32	0.41 x 0.32	140/435	P(Good) = 0.32
	0.11	0.08	0.13	0.32	
Poor	0.34 x 0.17	0.24 x 0.17	0.41 x 0.17	75/435	P(Poor) = 0.17
	0.06	0.04	0.07	0.17	
Total	150/435	105/435	180/435	435/435	
	0.34	0.24	0.41	1	
	P(Statistics) = 0.34	P(Art) = 0.24	P(Chemistry) = 0.41		

- Then we multiplied this by the sample size to get the expected count!

$$\text{Expected Count of Perfect and Stats} = \frac{220}{435}\left(\frac{150}{435}\right) \times 435 = 73.95$$

	Statistics	Art	Chemistry	*Total*
Perfect	0.17 x 435	0.12 x 435	0.21 x 435	220
	73.95	52.2	91.35	

- If we simplify the fraction in the calculation above, we see that the following happens:

$$\text{Expected Count of Perfect and Stats} = \frac{row\ total}{n}\left(\frac{column\ total}{n}\right)n = \frac{220}{435}\left(\frac{150}{\cancel{435}}\right)\cancel{435} = \frac{220(150)}{435} = \frac{(row\ total)(column\ total)}{n}$$

- Which brings us to the formula presented at the top!

	Statistics	Art	Chemistry	*Total*
Perfect	220 x 150 / 435	220 x 105 / 435	220 x 180 / 435	220
	75.86	53.10	91.03	
Good	140 x 150 / 435	140 x 105 / 435	140 x 180 / 435	140
	48.28	33.79	57.93	
Poor	75 x 150 / 435	75 x 105 / 435	75 x 180 / 435	75
	25.86	18.10	31.03	
Total	150	105	180	435

Comparison of Observed and Expected Counts

- Now that we have calculated **expected** counts, we can compare them to the **observed** counts

	Statistics	**Art**	**Chemistry**	***Total***
Perfect	100 (73.95)	40 (52.2)	80 (91.35)	220
Good	20 (47.85)	50 (34.8)	70 (56.55)	140
Poor	30 (26.1)	15 (17.4)	30 (30.45)	75
Total	150	105	180	435

Observed count
(Expected count)

- **If there is no relationship between Attendance and Major, the Observed (actual) counts and the Expected counts will be similar**! If there is a relationship, the Observed and Expected will be different.

- Finally, to calculate the **Test Statistic** X^2_{stat} we would use this formula to the right:
    - And eventually p-value and decide to reject or fail to reject and interpret!
    - For this test, only <u>reject</u> for <u>large values of the Test Statistic</u> (which have small p-values)!
        - This is because of the right-skew of the distribution

- Now let's see how to do this in the calc

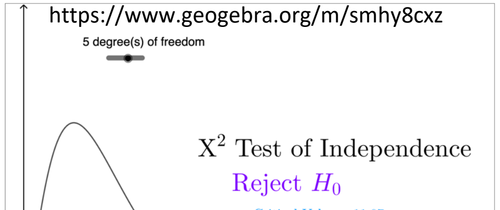
  (after another example)!

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

with $df = (r-1)(c-1)$

where:
O = Observed count
E = Expected count

where:
r = number of rows
c = number of columns



Comparison of Observed and Expected Counts

- Here's another example and visualization of this comparison of the Observed and Expected counts with the Movie and Snack context

Table 2: Contingency table for movie snacks data with row and column totals

Type of Movie	Snacks	No Snacks	**Row totals**
Action	50	75	**125**
Comedy	125	175	**300**
Family	90	30	**120**
Horror	45	10	**55**
Column totals	**310**	**290**	**GRAND TOTAL = 600**

Table 3: Contingency table for movie snacks data showing actual count vs. expected count

Type of Movie	Snacks	No Snacks	Row totals
Action	50 **65**	75 **60**	125
Comedy	125 **155**	175 **145**	300
Family	90 **62**	30 **58**	120
Horror	45 **28**	10 **27**	55
Column totals	310	290	GRAND TOTAL = 600

Observed count

Expected count

Here is an alternate dataset

- The row and column totals are the same
- But the Yes/No splits for each movie genre are different (zoom in ☺):

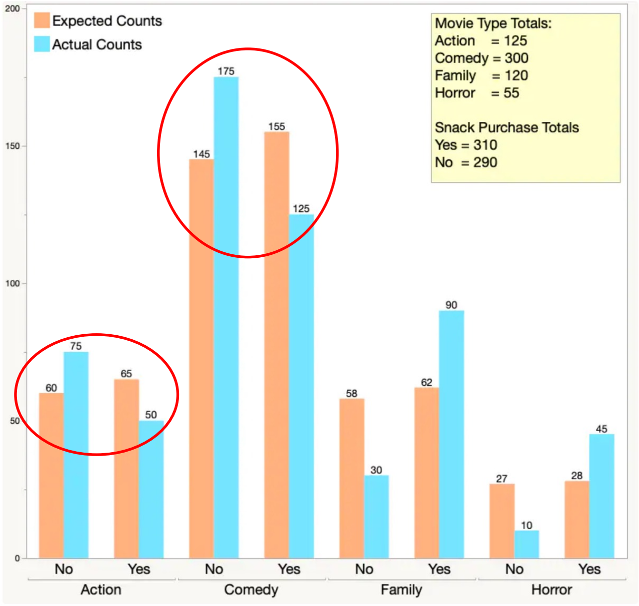


Figure 1: Bar chart showing the expected and actual counts for the different movie types

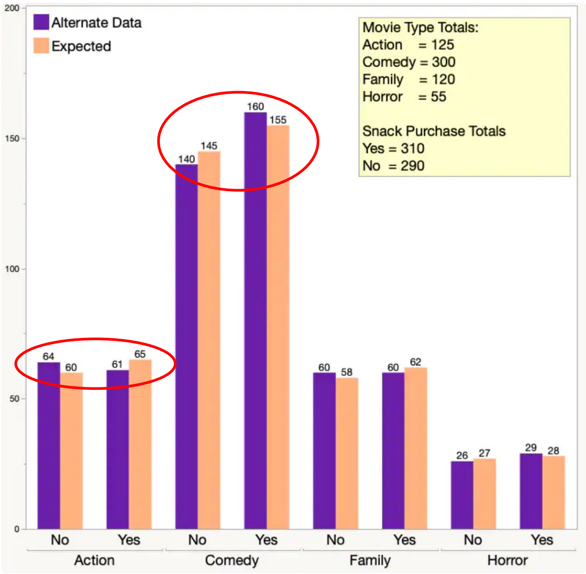


Figure 2: Bar chart showing the expected and actual counts using different sample data

Table 2: Contingency table for movie snacks data with row and column totals

Type of Movie	Snacks	No Snacks	Row totals
Action	50 61	75 64	125
Comedy	125 160	175 140	300
Family	90	30	120
Horror	45	10	55
Column totals	310	290	GRAND TOTAL = 600

- The expected counts are the SAME because they are based on the row and column totals
- But the observed have changed

source

Test Statistic X^2_{stat} = 65.03 > 7.815 = Critical Value → Reject!

Test Statistic X^2_{stat} = 0.903 < 7.815 = Critical Value → Fail to Reject

https://www.jmp.com/en_au/statistics-knowledge-portal/chi-square-test/chi-square-test-of-independence.html

Using Calc – X^2 Test of Independence

Setup
The University was reviewing their attendance policy when all classes went virtual. Is there enough evidence to conclude there is a significant relationship between Attendance and Major? Use α = 0.1

GOAL: Conduct a Hypothesis Test!

	Statistics	Art	Chemistry	*Total*
Perfect	100	40	80	220
Good	20	50	70	140
Poor	30	15	30	75
Total	150	105	180	435

1. Enter contingency table data
   a) Edit Matrix
   b) Enter correct dimensions (excluding row and column totals)
   c) Enter data
2. *X^2–Test*
   a) Observed = matrix of contingency table data
   b) Expected = Output of expected counts → Our calculator computes the Expected Counts for us! So this is where is stores the calculated Expected Counts

   Calculate or Draw

Hypotheses

H_0: Attendance and Major are unrelated

H_A: Attendance and Major are related

Using Calc – X^2 Test of Independence

Setup
The University was reviewing their attendance policy when all classes went virtual. Is there enough evidence to conclude there is a significant relationship between Attendance and Major? Use $\alpha = 0.1$

GOAL: Conduct a Hypothesis Test!

	Statistics	Art	Chemistry	*Total*
Perfect	100	40	80	220
Good	20	50	70	140
Poor	30	15	30	75
Total	150	105	180	435

1. Enter contingency table data
   a) Edit Matrix
   b) Enter correct dimensions (excluding row and column totals)
   c) Enter data

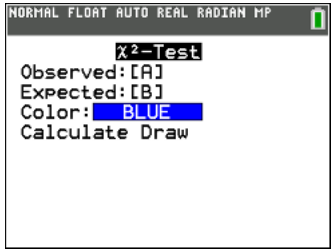
2. *X^2–Test*
   a) Observed = matrix of contingency table data
   b) Expected = Output of expected counts → Our calculator computes the Expected Counts for us! So this is where is stores the calculated Expected Counts

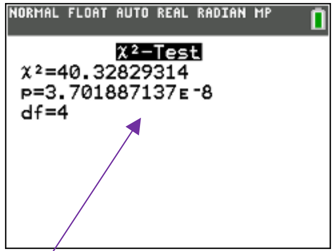
   Calculate or Draw

Hypotheses

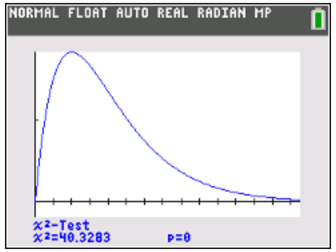
H_0: Attendance and Major are unrelated

H_A: Attendance and Major are related





Calculate Output
X^2 = Test Statistic
p = p-value
df = Degrees of Freedom

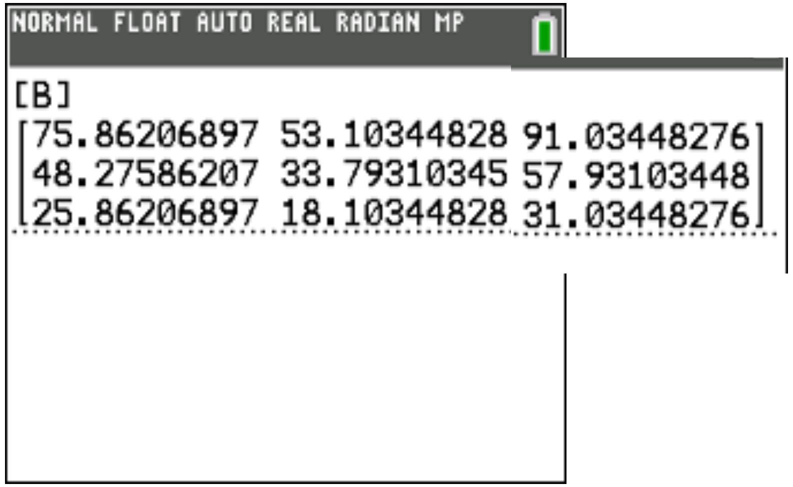


Draw Output
Plot (and displays values) of p = p-value and $X^2 = X^2_{stat}$ on the standard X^2 curve with df = (r – 1)(c – 1)

*** Calculator notation for result here: 3.7 E-8 = 3.7 x 10^{-8} = 3.7/100,000,000 (super small!) → can just say it's ≈ 0*

Comparing our Expected Counts to the Calculator's

	Statistics	Art	Chemistry
Perfect	0.17 x 435	0.12 x 435	0.21 x 435
	73.95	52.2	91.35
Good	0.11 x 435	0.08 x 435	0.13 x 435
	47.85	34.8	56.55
Poor	0.06 x 435	0.04 x 435	0.07 x 435
	26.1	17.4	30.45



- We did pretty good! Just roundoff error when we did it by hand (I rounded all the probabilities to 2 decimals)

LCQ - Interpretations

Problem: Write the conclusions and interpretations for our example.

Setup
The University was reviewing their attendance policy when all classes went virtual. Is there enough evidence t conclude there is a significant relationship between Attendance and Major? Use α = 0.1

Solution:

LCQ - Interpretations

Problem: Write the conclusions and interpretations for our example.

Setup
The University was reviewing their attendance policy when all classes went virtual. Is there enough evidence t conclude there is a significant relationship between Attendance and Major? Use α = 0.1

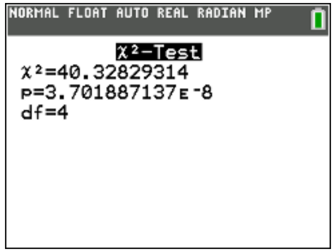
Solution:

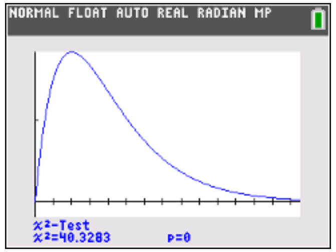
Need these:

Hypotheses

H_0: Attendance and Major are unrelated

H_A: Attendance and Major are related





Conclusion
Because the p-value ≈ zero is less than α = 0.1, we reject the null hypothesis!

Interpretation
There IS enough evidence to conclude the alternative → NOT full credit, CONTEXT!!!
There IS enough evidence to conclude that Attendance and Major are related!

It's important to note that we CAN NOT conclude that the Attendance CAUSES a MAJOR (or vice versa).
- *The independence test tells us ONLY whether there is a relationship or not*
- *Tt does NOT tell us that one variable causes the other*

LCQ – Whole Test

Problem:

The table below gives test results for drug use by 100 college students along with information about whether or not the student is actually a drug user. Conduct a hypothesis test to determine if the drug test result is dependent on whether the student actually uses drugs at the 5% level of significance.

	Positive Test	Negative Test
Drug User	26	9
Not a drug user	7	58

Solution:

LCQ – Whole Test

Problem:

The table below gives test results for drug use by 100 college students along with information about whether or not the student is actually a drug user. Conduct a hypothesis test to determine if the drug test result is dependent on whether the student actually uses drugs at the 5% level of significance.

	Positive Test	Negative Test
Drug User	26	9
Not a drug user	7	58

Solution: *** You should know all the steps required when conducting a full Hypothesis Test problem*

<u>Hypotheses</u>

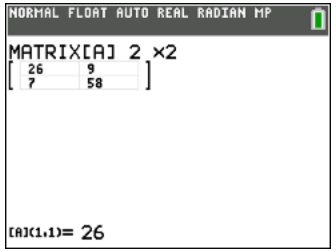
H_0: Positive test and drug use are unrelated → *NOT CORRECT! Because positive test result is just a single EVENT, we are talking about the entire VARIABLES*
H_0: Test result and drug use are unrelated
H_A: Test result and drug use are related

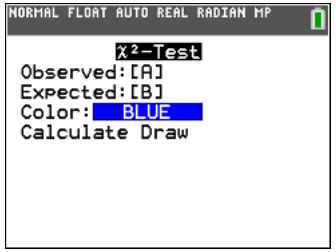
*** Even though the menu options won't ever change for this test, we still have to show our work by writing it out! This includes saying where we put the data*

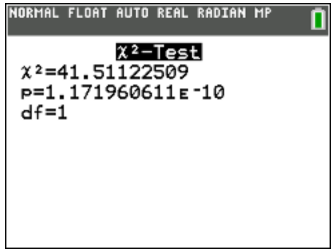
<u>P-Value (and Test Statistic)</u>

Entered contingency table into matrix [A]

P-value = X^2-Test(Observed = [A], Expected = [B]) ≈ 0
Test Statistic X^2 = 41.511, df = 1







<u>Conclusion and Interpretation</u>

Since our p-value ≈ 0 is less than alpha = 0.05, we reject the null hypothesis!
We have sufficient evidence to conclude that drug use and test result are related

<u>For demonstration purposes</u>

- *Let's say p-value = 0.10, what re the conclusion and interpretation now??*
- *Since our p-value 0.1 is greater than significance level = 0.05, we fail to reject the null hypothesis! There is NOT enough evidence to that test result and drug use are related*