1. Pick up the Project Introduction (brown packet). ← one per pair

2. Quietly read and score it using the scoring rubric but don't write on it.

Return the packet

1. Pick up the Project Introduction (brown packet).

2. Quietly read and score it using the scoring rubric but don't write on it.

HWCheck

Pick up the solutions. Do a quick check. Have them back up front in 5 minutes ! Ask questions as needed. LCQ tomorrow on these items.

1 Construct an expected frequency table for the following contingency tables:

a

	Likes chicken	*Dislikes chicken*	*sum*
Likes fish			60
Dislikes fish			40
sum	75	25	100

EXERCISE 11E.1

1 a

	Likes chicken	*Dislikes chicken*	*sum*
Likes fish	45	15	60
Dislikes fish	30	10	40
sum	75	25	100

b

	Drove to work	*Cycled to work*	*Public transport*	*sum*
Male				44
Female				36
sum	46	14	20	80

---

b

	Drove to work	*Cycled to work*	*Public transport*	*sum*
Male	25.3	7.7	11	44
Female	20.7	6.3	9	36
sum	46	14	20	80

c

	Junior school	*Middle school*	*High school*	*sum*
Plays sport	35	59	71	165
Does not play sport	23	27	35	85
sum	58	86	106	250

c

	Junior school	*Middle school*	*High school*	*sum*
Plays sport	38.28	56.76	69.96	165
Does not play sport	19.72	29.24	36.04	85
sum	58	86	106	250

d

	Wore hat and sunscreen	*Wore hat or sunscreen*	*Wore neither*	*sum*
Sunburnt	3	5	13	
Not sunburnt	36	17	1	
sum				

d

	Wore hat and sunscreen	*Wore hat or sunscreen*	*Wore neither*	*sum*
Sunburnt	10.92	6.16	3.92	21
Not sunburnt	28.08	15.84	10.08	54
sum	39	22	14	75

p. 333 ... #5

(b)

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \cdot \sum (y-\bar{y})^2}} = \frac{-152}{\sqrt{(71.5)\cdot(376)}} = -0.927$$

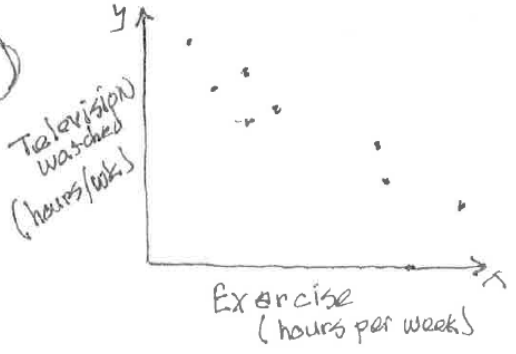
LCQ

$\bar{x} = 4.75$

$\bar{y} = 12$

(c) → see next sheet

(a)



c) Describe the correlation:

There is a strong, negative, linear correlation between exercising and hours watching television.

d) $\bar{x} = 4.75$, $\bar{y} = 12$, $S_x = 2.9896$

Covariance

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

(L_1 and L_n)

$$= \frac{-152}{8} = -19$$

LCQ tomorrow

LSRL

Equation of LSRL

$$y - y_1 = \frac{S_{xy}}{(S_x)^2}(x - x_1)$$

using the mean point (4.75, 12)

$$y - 12 = \frac{-19}{(2.9896)^2}(x - 4.75)$$

$$\boxed{y - 12 = -2.13(x - 4.75)}$$

can convert to $y = mx + b$ format to check with GDC

$y = -2.13x + 22.1$

e) 

For every additional hour of Exercise per week, they watch 2.13 hours less of TV per week.

f) $y = -2.13x + 22.1$

$= -2.13(5) + 22.1$

$= 11.45$ or 11.5 hours of TV watching at 5 hours of exercise.

Your calculator can also generate the expected values as we'll see later

d

	Wore hat and sunscreen	*Wore hat or sunscreen*	*Wore neither*	*sum*
Sunburnt	3	5	13	
Not sunburnt	36	17	1	
sum				

Schedule

Monday ---The full Chi-Square Test of Indep. Process

Tuesday- Special Situations + Evaluate other Criteria for project, **LCQ**

Wednesday- Get a list of Unit 2 Test items, continued practice,

Packet P3 (Info on selecting a project and Ideas for project)

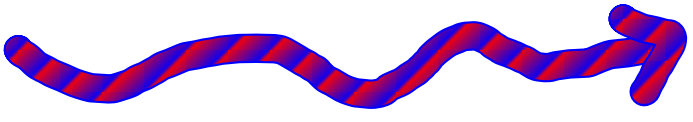
Thursday - Evaluate a past project (using the scoring guide) + Use

Computer spreadsheet to calculate r and LSRL

Friday - Review Questions + Start Numerical Trigonometry

Monday - Test on Unit 2 (Statistical Applications)

Recap from Part I of the Chi Square Test of Independence.



If you were not here last class, hopefully you already checked the blog!

The test is for categorical variables only.

- Set up a contingency table for two categorical variables.
- Assume independence to start
- Calculate expected values

We only use the term "correlation" with with numerical data.

Pick up the Class Notes

read the first 4 slides

X^2 is a statistic that measures the difference between observed values and expected values in a contingency table

Observed Frequencies

	Regular exercise	*No regular exercise*	
Male	112	104	216
Female	96	88	184
	208	192	400

EXPECTED frequencies

	Regular exercise	*No regular exercise*	*sum*
Male	$\frac{216\times208}{400} \doteq 112.3$	$\frac{216\times192}{400} \doteq 103.7$	216
Female	$\frac{184\times208}{400} \doteq 95.7$	$\frac{184\times192}{400} \doteq 88.3$	184
sum	208	192	400

If the chi square value that we calculate is big enough, then we can establish a:

linkage between two variables

association between to variables

relationship between the variables

If the variables in this example are, indeed, associated, then gender might have an effect on regular exercise but just being associated or linked does not prove causation.

What you can say is.....

Chi Square Statistic is :

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e}$$

and we compare it to.....

....the cutoff, or critical Chi-Square Value which is either given to you (or found in a resource table) .

..... which, in turn, will tell us whether to accept or reject the assumed independence between the two variables.

Independent ⟷ Not Independent

Dependent

	Early	late
M	20	27
F	42	6

Associated
Linked
Related
Connected

AIM:

Calculate the Chi-Square Statistic, 3 different ways

Carry out the entire Test of Independence

Before we go on to a new situation we need to practice calculating X^2 by using the formula itself.

For this we'll continue to use the same example from yesterday

handout

Once the expected cell frequencies are computed, it is convenient to enter them into the original table as shown below. The expected frequencies are in parentheses.

	Graduated	Failed to Graduate	Total
Experimental	73 (59.042)	12 (25.958)	85
Control	43 (56.958)	39 (25.042)	82
Total	116	51	167

Observed frequencies

	Graduated	Failed to Graduate	Total
Experimental	73	12	85
Control	43	39	82
Total	116	51	167

Expected frequencies

	Graduated	Failed to Graduate	Total
Experimental	59.042	25.958	85
Control	56.958	25.042	82
Total	116	51	167

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(73 - 59.042)^2}{59.042} + \frac{(12 - 25.958)^2}{25.958} + \cdots = 22.0$$

χ^2

alternative:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

L_1 f_o	L_2 f_e	L_3 $\frac{(L_1 - L_2)^2}{L_2}$		
73	59.042			
12	25.958			
43	56.958			
39	25.042			

22.0 χ^2.

and now with matrices

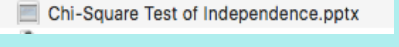
We find the Chi Squared value by putting the values from the table of observed values and the table of expected values into the calculator.

a. Choose **MATRIX** and go to **EDIT**
b. Make sure your matrix is the right size
c. Enter your **Observed** values in **Matrix A**
d. Choose **STAT** and go to **TESTS**
e. Scroll down to **χ^2-Test** and press **ENTER**
f. Choose **Calculate**.
g. Your **expected** values can now also be found in **Matrix B**

a. Choose **MATRIX** and go to **EDIT**
b. Make sure your matrix is the right size
c. Enter your **Observed** values in **Matrix A**
d. Choose **STAT** and go to **TESTS**
e. Scroll down to **χ^2-Test** and press **ENTER**
f. Choose **Calculate**.
g. Your **expected** values can now also be found in **Matrix B**

We'll step back and just observe example of the whole process

(ppt: Chi Square ppt)



B.B.

When finished, pick up the notes

Full example of a Chi-Square Problem

Is someone's favorite subject at school associated with gender?

The Chi-Squared Test of Independence

This test determines if two types of ***categorical*** variables are associated with each other. (if not associated then they are independent of one another). The test does not prove causation. *There are two ways to run the test:*

A. *using the* χ^2 statistic B. Using a probability value

Note: The term *correlation* is not appropriate to use with this test or with categorical data.

A. The Chi Squared Test of Independ. (using the χ^2 statistic)

Follow the steps below if performing the Chi-Square Test of Independence *for a project* or *for an assignment.*

1. **Make a Contingency Table and State the Null Hypothesis.** Place the data into a contingency table showing the respective observed frequencies. (must be integers)

Here is an example question:
A student was going to investigate whether favorite subject ***is associated*** with gender. So, the student gave his 150 classmates a questionnaire to fill out. The results for the question on the gender of the student and their favorite subject are given in the table below, which is a 2 × 3 contingency table of **observed** frequencies.

State a Null Hypothesis (H_0) and an Alternative Hypothesis (H_1).

H_0: Favorite subject is **independent** of gender.

H_1: Favorite subject and gender are **associated** (or you can say Favorite subject and gender are not independent). Note: Being associated does not mean one variable **causes** the other variable to happen.

Then, show a Contingency Table

	History	Biology	French	
Female	22	20	18	
Male	20	11	9	

Expected values:

	History	Biology	French
Female	n	18.6	16.2
Male	q	r	10.8

The expected values for each cell are calculated from row and column totals from the original contingency table as follows:

To find *n for example* : $\frac{60}{100} \cdot \frac{42}{100} \cdot 100 = 25.2$ *which is equal to* $\frac{\text{Row total} \cdot \text{Column total}}{n}$

On IB exams, you will want to know how to produce these values using the above method and from your GDC

3. **Check if the test will be valid** (results trustworthy?):
All **EXPECTED** values must be greater than 5. If not, the test will be *invalid*. Do not continue the test.

If your contingency table is larger than a 2 by 2, then you might be able to condense the number of columns or rows in a sensible way and re-start the test. If so, be sure to show your new contingency table and then re-start step 2. (On projects -- be sure to discuss the reasoning behind your new category divisions). To avoid this problem, always collect an abundance of data !

4. State **significance level.** This is usually 5%. *Values such as 1% or 10% are also common. On most problems this value will be given to you. If not, choose 5%.*

EXAMPLE STATEMENT: ***"This test will be run using a 5% significance level".***

5. Find the number of **degrees of freedom** and the **critical value**. This is simple:

d.f. = **(number of rows – 1) x (number of columns – 1)**, in our example, (2-1) x(3-1) = 1 x 2 = **2**

The critical value: Will be given to you on IB exams. On projects and problems from class you will have to cite and use a table.

a. Go down the left column until you reach the number of degrees of freedom (2 in this example)
b. Go across until you reach the column 0.95 and read off your value. It is 5.99
   (note: we use 0.95 because there is usually a 5% significance level and 0.95 is 100% – 5%)
   Note: If the test was being run at a 1% contingency level, the go across to the 0.01 column.

Critical values of the χ^2 distribution

degree of freedom → ν

significance level → p

ν	10%	5%		1%	
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267

degree of freedom → ν

significance level → p

ν	10%	5%		1%	
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188

6. State the ***rejection inequality***

"If $\chi^2 > 5.99$, then I will reject H_0 "

Translation: You will reject the null hypothesis if the chi-square statistic (that you calculate on the next step) is greater than the critical value from the table. (This means the differences between the expected and observed values are large enough to reject independence. If not, you will accept the null hypothesis if the chi-square statistic is not greater than the critical value.

7. Calculate the **Chi Squared Statistic**

You need to know how to calculate this statistic using the formula and by your GDC, depending on the question.

a) By "formula" (you may at times need to show a process with the formula)

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$ where $f_o = observed\ frequency$ (*a natural number*)

$f_e = expected\ frequency$ (*a natural number*)

1. For each cell in the table, subtract the ***expected value*** from the ***observed value*** and square the difference. Divide this by the expected value.
2. Find the sum, which is the Chi-Square statistic.
3. On a project this process needs to be shown clearly.

b) With your GDC:

a. Choose **MATRIX** and go to **EDIT**
b. Make sure your matrix is the right size
c. Enter your **Observed** values in **Matrix A**
d. Choose **STAT** and go to **TESTS**
e. Scroll down to **χ^2-Test** and press **ENTER**
f. Choose **Calculate**.
g. Your **expected** values can now also be found in **Matrix B**

.....*Ok, in our favorite subject/gender example:* $\chi^2 = 1.78$

Note when working on projects: If the degree of freedom is 1 (in other words for *all 2 by 2 tables*) the ***Yates Continuity Correction*** must be used. In this case, do not use the statistic from your calculator. It will be invalid. here is the revised formula.

If $df = 1$, we use

$$\chi^2_{calc} = \sum \frac{(|f_o - f_e| - 0.5)^2}{f_e}$$

where $|f_o - f_e|$ is the **absolute value** or **modulus** of $f_o - f_e$

8. **Make two summary statements** about the test.

A. First make a statement, in a complete sentence, using the following guidelines:

If the χ^2 value is **less than or equal to** the critical value, we **accept the null hypothesis** (or are unbable to reject the null hypothesis)

or if If the χ^2 value is **greater than** the critical value, we **reject the null hypothesis.**

B. Then, in a complete sentence, make a **general statement** of what the results tell you about the particular situation you are studying using the vocabulary of the situation.

In our Example:

"Since 1.78 is less than the critical value, 5.99, I accept H_0 which indicates independence between the gender and favorite subject . "

and the general statement could be something like:

"Based on these results, there is no apparent association between favorite subject at my school and one's gender."

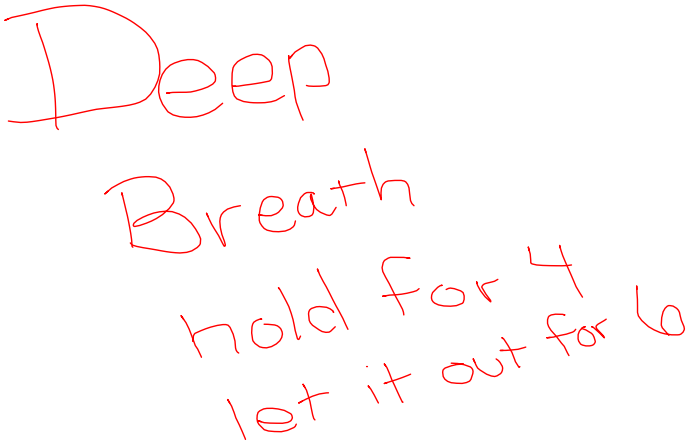
(do not use the word "***correlation***" or "***affect***" when using this test)

Note:

Be cautious of the wording of your conclusions during those times when you reject the Null Hypothesis , H_0. *The Chi Square Test of Independence does not prove anything. It supplies evidence to support if two categorical variables are associated with each other or not. If our example had had the opposite result, the general statement might read as:*

"Since 9.72 is greater than the critical value, 5.99, I reject H_0 which indicates that favorite subject and gender are not independent."

These results supply evidence to support the possibility that favorite subject and gender of the students I studied are associated.



B. The Chi Squared Test of Indep.
(using Probability)

You may be asked to perform the test using **probability** instead of the Chi-Square statistic. In that case, you will follow steps 1 through 4. But continue with the following steps instead:

5. State the ***rejection inequality***

from GDC

"If the p-value is **less than** the significance level " (5% or 0.05 for many cases), then I will reject H_0.

6. **Calculate the *p-value*.**
This step must be done on a calculator. Follow the same calculator steps as if you are calculating the Chi-square statistic.

"From my GDC, p = 0.41 "

7. Make two summary statements

A. First make a statement, in a complete sentence, using the following guidelines:

If the ***p value*** is **greater than or equal to** the signifance level (5% in most problems), we **accept the null hypothesis** (or are unbable to reject the null hypothesis)

or if If the ***p value*** is **LESS THAN** than the signifance level, we **reject the null hypothesis.**

B. **Then**, in a complete sentence, make a **general statement** of what the results tell you about the particular situation you are studying using the vocabulary of the situation.

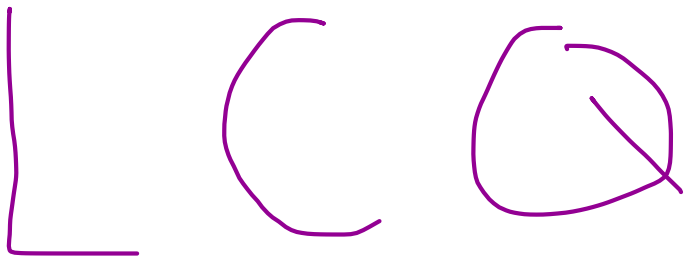
Example:
"Since my p-value of 0.41 is greater than the 0.05 significance level, I must accept H_0 which indicates independence between gender and favorite subject. "

and the general statement could be something like:
"Based on these results, there is no apparent association between favorite subject at my school and one's gender and favorite subject."

(do not use the word ***correlation*** or ***affect*** when using this test)

don't lose this packet.

~ Write your name on it.
~ Use it this fall....
~ Bring it to the review sessions in April.



Assignment: Ch.11 Packet

p. 337.... 2abc

p. 341.... #1 (use X^2 statistic)

.....#3 (use probability)

follow and label all steps

Optional Extra Practice with Correlation and LSRL by "hand": handout with Gross Domestic Product and Infant Mortality Rate

How do you tell one bathroom full of statisticians from another?

Check the p-value