CHI SQUARE TEST OF INDEPENDENCE/ASSOCIATION

RESEARCH PROBLEM

Sometimes we want to know about whether or not categories (attributes/groupings) of one variable "line up with" categories (attributes/groupings) of another.

- How the categories "overlap"
- How categories of one variable *vary by/are distributed along* categories of another.

RESEARCH PROBLEM

Are black individuals more likely to be Democrats?

- Relationship between race (2 or more groups) and political party affiliation (2 or more groups)

Do cities (compared to suburbs or rural areas) experience more robberies?

- Relationship between type of location (3 groups) and type of crimes (8 categories)

Are liberals or conservative more permissive in their childrearing?

- Relationship between party affiliation (2 groups) and permissiveness (2 categories)

RESEARCH PROBLEM

Categories (Attributes/Groups):

THE ANSWER IS CHI (NOT CHAI)

Chi Square

- Examines relationship between (categories of) <u>two</u> *<u>nominal/ordinal</u>* variables

LOGIC OF CHI-SQUARE

Examines *differences* between cell frequencies in two crosstabulations of the relationship between the two variables:

- one crosstabulation with actual observed frequencies
- one crosstabulation with frequencies expected if the variables are unrelated/independent of one another

LOGIC OF CHI-SQUARE

Crosstabulation (Crosstab)/Contingency Table:

- Matrix/table that depicts distribution or frequencies in categories of one variable across categories of another variable

Observed frequencies (f_o)

- Cell frequencies **actually observed** from real data in a bivariate table (crosstab)

Expected frequencies (f_e)

- Cell frequencies that are expected to occur, if the two variables were statistically independent (e.g. no association/relationship b/w variables or the null hypothesis were true)

LOGIC OF CHI-SQUARE

Observed frequencies (f_o)

- Actual data

First Gen Status	Public Affairs	Sociology	Total
Firsts	691	1245	1936
Nonfirsts	1259	1425	2684
Total	1950	2670	4620

Expected frequencies (f_e)

- What are the data if H_0 were true?
  - No association between categories of the variables

First Gen Status	Public Affairs	Sociology	Total
Firsts	?	?	1936
Nonfirsts	?	?	2684
Total	1950	2670	4620

LOGIC OF CHI-SQUARE

Expected frequencies (f_e)

- We calculate expected frequencies for each empty cell.
- Consider the overlapping categories (row and column) to which a given cell belongs
- For a cell, calculate the **proportion of all cases** that come from the total of just *one of those categories* (either row or column totals/marginals)
- Then, you adjust/weight the total number of cases that come from *the other category* by multiplying it by your calculated proportion. This would be expected value for that cell.
  - This uses multiplicative law for joint probabilities of independent events, b/c we're taking into account that H0 assumes that the variables are independent.

LOGIC OF CHI-SQUARE

Expected frequencies (f_e)

- Top left cell is for Public Affairs students who are first-generation.
- 1936 of all 4620 students are first-gen. 1936/4620 = .419 students were first-gen.
- If there was no relationship between the variables (independent), we would expect .419 of all Public Affairs students to be first-gen, and .419 of all Sociology students to be first-gen.
- For all Public Affairs students, we adjust PA total by the proportion, or, 1950 * .419 = 817.14. This would be expected value for that cell.

First Gen Status	Public Affairs	Sociology	Total
Firsts	?	?	1936
Nonfirsts	?	?	2684
Total	1950	2670	4620

$$f_e = \frac{(\text{Column marginal})(\text{Row marginal})}{N}$$

LOGIC OF CHI-SQUARE

Expected frequencies (f_e)

- (1936/4620) * 1950 = 817.14

Row Marginal

First Gen Status	Public Affairs	Sociology	Total
Firsts	817.14		1936
Nonfirsts			2684
Total	1950	2670	4620

Column Marginal

$$f_e = \frac{(\text{Column marginal})(\text{Row marginal})}{N}$$

N

LOGIC OF CHI-SQUARE

Expected frequencies (f_e)

- What are the data if H_0 were true?
  - No association between categories of the variables

First Gen Status	Men	Women	Total
Firsts	817.14	1118.86	1936
Nonfirsts	1132.86	1551.14	2684
Total	1950	2670	4620

$$f_e = \frac{(\text{Column marginal})(\text{Row marginal})}{N}$$

LOGIC OF CHI-SQUARE

Actual data: Observed frequencies (f_o)

First Gen Status	Public Affairs	Sociology	Total
Firsts	691	1245	1936
Nonfirsts	1259	1425	2684
Total	1950	2670	4620

H_0 true data: Expected frequencies (f_e)

First Gen Status	Public Affairs	Sociology	Total
Firsts	817.14	1118.86	1936
Nonfirsts	1132.86	1551.14	2684
Total	1950	2670	4620

CHI-SQUARE

Chi Square Test of Independence (X^2)

- Examines relationship between two nominal/ordinal variables
  - Tests the independence of (absence of association between categories of) two variables

CHI-SQUARE (RESEARCH QUESTION)

Is variation in X related to variation in Y?

- How is variation in the categories of X associated with variation in the categories of Y?

CHI SQUARE (VARIABLE TYPES)

IV: nominal, ordinal (e.g. categorical/discrete)

- Grouping variable

DV: nominal, ordinal (e.g. categorical/discrete)

- Grouping variable

CHI SQUARE (ASSUMPTIONS)

1. Independence of Observations
   - Groups are not related or dependent upon each other. Case can't be in more than one group. No ties between observations
2. Normality of Distribution
   - Distribution must be relatively normal
     - If 20% or more expected cell frequencies (f_e) are below n=5, you violate the assumption

CHI SQUARE (HYPOTHESES)

Null hypothesis (H_0)

- There *is no relationship/no association* between the two cross-tabulated variables, in the population, therefore the variables are statistically independent.
- There is no difference between the observed and expected frequencies for categories of the variables. Frequencies of dependent variable are expected to be the same across groups of the independent variable
  - H_0: $f_e = f_o$

Research hypothesis (H_1)

- There *is a relationship/an association* between the two variables, in the population
- There is a difference between the observed and expected frequencies for categories of the variables.
  - H_1: $f_e \neq f_o$

CALCULATING CHI SQUARE (X^2)

Measures *size of difference* between observed and the expected frequencies

- By calculating the difference for *each* cell

CALCULATING CHI SQUARE (X^2)

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

Where

f_o = observed frequencies

f_e = expected frequencies

$$f_e = \frac{(Column\ marginal)(Row\ marginal)}{N}$$

CALCULATING CHI SQUARE (X^2): MAJOR AND FIRST-GEN STATUS

	Public Affairs		Sociology		
First Gen Status	f_o	f_e	f_o	f_e	Total
Firsts	691	817.14	1245	1118.86	1936
Nonfirsts	1259	1132.86	1425	1551.14	2684
Total	1950		2670		4620

CALCULATING CHI SQUARE (X^2): MAJOR AND FIRST-GEN STATUS

Cell	f_o	f_e	f_o - f_e	$(f_o - f_e)^2$	$(f_o - f_e)^2$ / f_e
PA/Firsts	691	817.14	-126.14	15911.2996	19.47
PA/Non-Firsts	1259	1132.86	126.14	15911.2996	14.04
Soc/Firsts	1245	1118.86	126.14	15911.2996	14.22
Soc/Non-Firsts	1425	1551.14	-126.14	15911.2996	10.26
∑					X^2 = 57.99

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

HYPOTHESIS TESTING WITH CHI SQUARE

Is the difference between observed expected frequencies “extremely” different from what is expected by chance/the null hypothesis?

X^2 TEST AND THE X^2 DISTRIBUTION

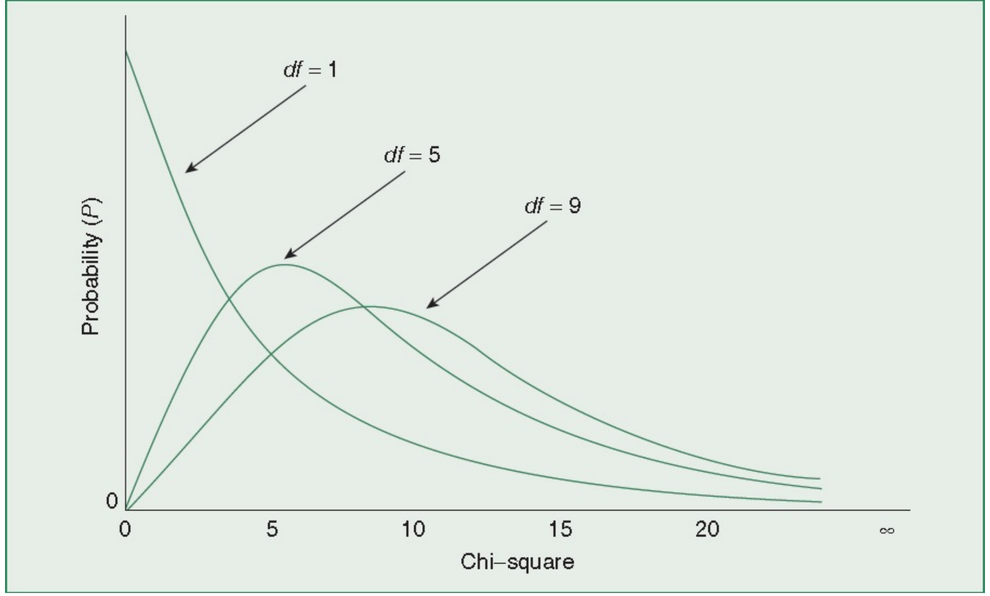
The X^2-distribution has multiple curves

- Each curve based on sample size or degrees of freedom

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

- Where *r* is the number of rows in the crosstab and *c* is the number of columns in the crosstab.

Figure 11.1 Chi-Square Distributions for 1, 5, and 9 Degrees of Freedom



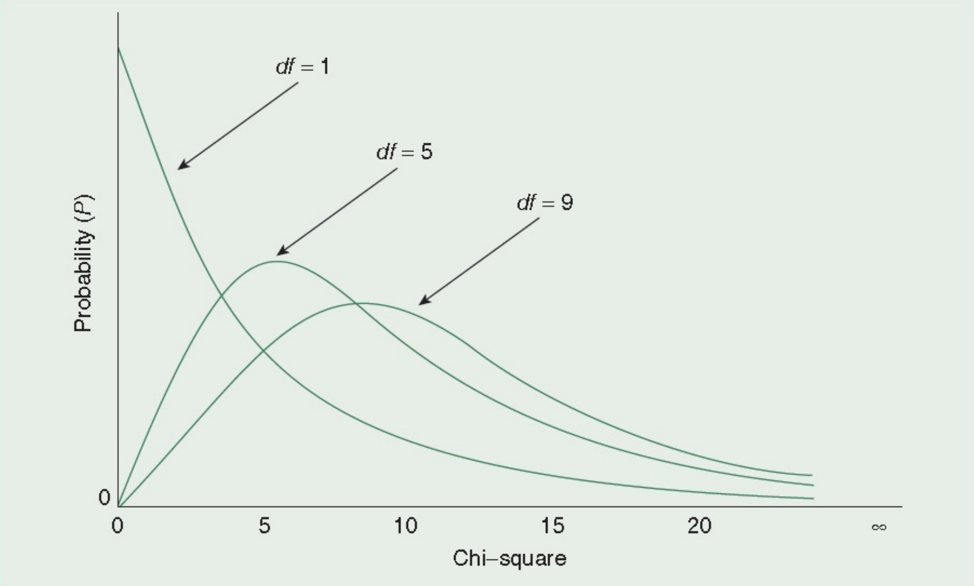
Distribution is one-sided (squaring gets us positive numbers), ranging from 0 to +∞

IS THE X^2 EXTREME?

If the overall *difference* between the expected and observed frequencies *is extreme enough*

But what counts as extreme enough?

Figure 11.1 Chi-Square Distributions for 1, 5, and 9 Degrees of Freedom



IS THE X^2 EXTREME?

Similar to the *t*-Test:

- If $|t_{obtained}| \geq |t_{critical}|$, then…
  - Difference between mean of group 1 and mean of group 2 is so extreme that we can't blame it on sampling error, therefore… H_0 is probably not true, so…
  - $t_{obtained}$ is in rejection region
  - Reject H_0
  - $p \leq \alpha$
  - Statistically significant difference

IS THE X^2 EXTREME?

…And similar to the ANOVA *F*-test

- If $F_{obtained} \geq F_{critical}$, then…
  - Overall differences between group means is so extreme that we can't blame it on sampling error, therefore… H_0 is probably not true, so…

  - $F_{obtained}$ is in rejection region
  - Reject H_0
  - $p \leq \alpha$
  - Statistically significant difference between group means

IS THE X^2 EXTREME?

For the X^2 test:

- If $X^2_{obtained} \geq X^2_{critical}$, then…
  - Differences between observed and expected frequencies are so extreme that we can't blame it on sampling error, therefore… H_0 is probably not true, so…
  - $X^2_{obtained}$ is in rejection region
  - Reject H_0
  - $p \leq \alpha$
  - Statistically significant relationship between the two variables

IS THE X^2 EXTREME?

X^2 has multiple distributions, based *df* and α:

- Because of squaring differences , F will always be positive
  - Thus, distribution only has one-tail

- We must refer to a table (X^2 Table) to figure out what $X^2_{critical}$ is.
  - Your α
    - Usually $\alpha = .05$
  - Your *df*
    - *df* = (r-1)(c-1)

- Then, evaluate to see if $X^2_{obtained} \geq X^2_{critical}$

IS THE X^2 EXTREME?

Hypothesis Testing:

- If $X^2_{obtained} \geq X^2_{critical}$, reject the null hypothesis
- If $X^2_{obtained} < X^2_{critical}$, fail to reject the null hypothesis

REPORTING CHI SQUARE

Report

- The test used
- If you reject or fail to reject the null hypothesis
- The variables used in the analysis
- The degrees of freedom, calculated value of the test, and p-value
  - X^2(df) = Chi-square$_{obtained}$, p-value

- "Using the Chi Square test of independence, I reject/fail to reject the null hypothesis that there is no relationship between one variable and the other variable, in the population, X^2(?) = ?, p ? .05"

CALCULATING CHI-SQUARE

In the example between major and first-gen status

- $X^2_{obtained}$ = 57.99
- df = (2-1)(2-1) = 1

In our X^2 Table, we follow the df = 1 row and the α = .050 column, to see where they intersect.

- $X^2_{critical}$ = 3.84

- Because $X^2_{obtained} \geq X^2_{critical}$ we reject the null

First Gen Status	Public Affairs	Sociology	Total
Firsts	691	1245	1936
Nonfirsts	1259	1425	2684
Total	1950	2670	4620

Cell	f_0	f_e	$f_0 - f_e$	$(f_0 - f_e)^2$	$(f_0 - f_e)^2 / f_e$
PA/Firsts	691	817.14	-126.14	15911.2996	19.47
PA/Non-Firsts	1259	1132.86	126.14	15911.2996	14.04
Soc/Firsts	1245	1118.86	126.14	15911.2996	14.22
Soc/Non-Firsts	1425	1551.14	-126.14	15911.2996	10.26
Σ					X^2 = 57.99

REPORTING CHI SQUARE

Report

- The test used
- If you reject or fail to reject the null hypothesis
- The variables used in the analysis
- The degrees of freedom, calculated value of the test, and p-value
  - X^2(df) = Chi-square$_{obtained}$, p-value

- "Using the Chi Square test of independence, I reject the null hypothesis that there is no relationship between gender and first-gen status, in the population, $X^2(1) = 57.99$, $p < .05$"

EXAMPLE: POLITICAL ORIENTATION AND VIEWS OF BLM

Using the following cross-tab, hand calculate:

- X^2
- degrees of freedom
- determine it's significance level
- Fully and completely report your findings (and whether you reject/fail to reject the null hypothesis

Views of BLM	Political Orientation		Total
	Liberals	Conservatives	
Positive	47	5	52
Negative	3	35	38
Total	50	40	90