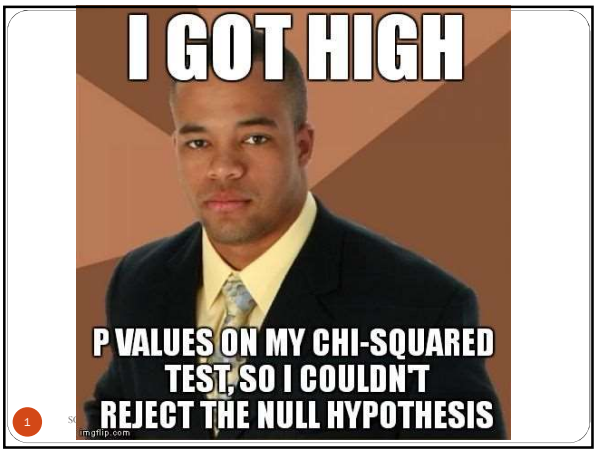


1

Warnings in the Syllabus

Work Alone

Do not work together when you have been instructed to work alone. Plan ahead as much as possible, and ask frequent questions early, but DO NOT complete assignments (other than group labs) together! Undertake study partners or groups with caution, and terminate them at the appropriate stage of your work. Discuss the contents of any assignment, and plan how to carry it out, with any other member of the class, the instructor, or anyone else. However, once you have actually begun the work on the assignment, only ask the instructor for help, no matter how minor. You may always consult written course materials, your own notes, or SPSS program "help" features, which we will review.

Careful Being Helpful

Academic dishonesty, as defined by the CSUN Schedule of Classes and Catalog Supplement, includes cheating, fabrication, plagiarism as well as facilitating academic dishonesty. Facilitating academic dishonesty includes allowing another student to see your work, which they then reproduce in whole or in part and submit as their own, whether you know it or not.

I exercise zero tolerance for academic dishonesty (including cheating and plagiarism), as outlined in the section on Academic Dishonesty in the *University Catalog* (https://catalog.csun.edu/policies/academic-dishonesty), and take any compromise of that policy very seriously. Ignoring this instruction will result in wasted time and investment, will damage your academic record, and will forego the opportunity to actually learn the material and benefit from the course.

is, give no credit for) that entire assignment. Do not, therefore, steal so much as a sentence fragment without proper attribution. And note that zeroes awarded for academic dishonesty will not be dropped, and *will* lower your semester grade.

5

SOC424 – Statistics w/ Dr. Ellis Godard

"Chi-square"

x^2

COLUMN percentages

- I used that phrase over 30 times in the last lecture
  - It's the *only* of the available percentages that I mentioned
  - It's also *all* over the notes, and in the handout @ Canvas
- I warned *not* to get all of them and triple the %s in the table
  - JUST the… ***column percents***
- But, on the last quiz…

Crosstab cells should have the cell count (observed frequency) and...

I haven't even mentioned this (yet)

All of these	6 respondents	30 %
The column percentage	13 respondents	65 % ✓
The expected count	1 respondent	5 %
The total percentage		0 %
The row percentage		0 %

Do NOT get these! They aren't COLUMN

6

Outline for Today…

- **Reminders**
  - Academic Dishonesty
  - COLUMN percents
- **Intro to "Chi-Square"**
  - Concept
  - Univariate Chi-square
  - Bivariate Chi-square
- **The Chi-square test**
  - Steps involved (same 5 as any test)
  - Two more examples
  - Observed and Expected Values
- **The Test in SPSS**
  - 1 example, 1 required lab, 1 extra credit lab (if you want it?)

4

"Chi^2" Measures Dependence

- **H_A: Values of 1 variable depend on those of another**
  - Percentages are *significantly* different across the rows
  - Highest percentages move row to row, maybe making a diagonal?
- **H_o: statistical independence, no association**
  - Null hypothesis imagines a distribution random
  - What would data look like if Ha is wrong
  - Independence – essentially the same %'s across each row
    - Or, at least, *not* significant differences across any of the rows

7

"Chi^2" Compares 2 Sets of #s

- **Chi-square compares *observed* data to *imaginary***
  - Observed frequencies are the data we have
  - Expected frequencies are what %s would be if null true
- **Want to find large difference between the two**
  - The bigger the difference (from the null) we observe (in the sample data),
  - The more comfortable we are that the null probably isn't "right", so
  - The less risk there is in rejecting the null, and saying there's probably a difference in the full population, too (not just in the sample)

Univariate* Chi-Square Test

* ***Not*** what you'll be using, but the idea *might* help understand what's next

- **Q: Do students watch reality shows?**
  - Of 48 students, 20 watch them and 28 do not
  - 28/48 is a majority (58.3%) in the sample
  - But is it a big enough majority to generalize to the population?
    - Is the difference *statistically significant?*
- **Test with "chi-square", which compares:**
  - Observed frequencies: 20 and 28 (*Fo*)
  - Expected frequencies: 24 and 24 (Fe)
    - If the null is "true", there's no difference at *all*

Univariate Chi-Square Test, cont'd

- **Calculated Chi-square**

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{16}{24} + \frac{16}{24} = \frac{32}{24} = 1.5$$

- **Compare to critical (aka tabular), like t & F**
  - Varies w/ degrees of freedom (c-1)
  - Here, for alpha of .05, critical value is 3.841
- **Calculated < Critical, so can't reject null**
  - Our sample data is not significantly different from null hypothesized values of independence
  - 20 and 28 are not different enough from 24

Chi-square for a 2x2 Table

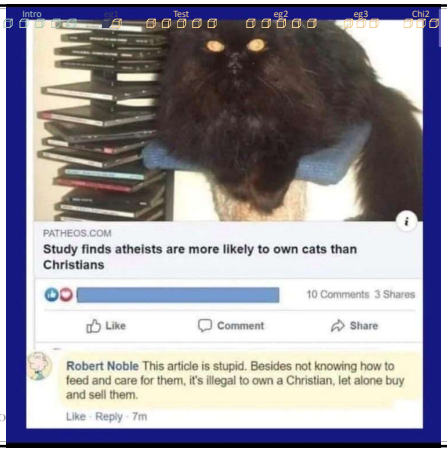
- Tests Dependence vs. Independence
  - Variables are *statistically independent* if the *population* conditional distributions equal the marginal distribution (%'s same across rows)
  - They are statistically *dependent* if the population conditional distributions differ (%'s do change)
- Is an Inference from Sample Data
  - We examine *sample* conditional distributions to look for evidence of differences from the null
  - There's always *some* sample difference, but is it large enough to think we'd also find one in the population? Is the difference *significant?*

Table of Independence

- Seek relationship: DV varies w/ values of IV
- Here, they don't – variables are independent:

	LibProt	ConsProt	Catholic	None
IRM ok	750 30%	1200 30%	900 30%	150 30%
Not ok	1750 70%	2800 70%	2100 70%	350 70%
Total	2500	4000	3000	500

- Distribution of DV the *same* for each IV value
  - %'s don't change across the rows at *all*



Atheists more likely to own cats?

- Hypothetical sample:
  - 100 atheists, of whom 30 own a cat
  - 100 Christians, of whom 20 own a cat
- Actual Crosstab

	Atheists	Christians	Totals
Cats	30 (30%)	20 (20%)	50 (25%)
Dogs	70 (70%)	80 (80%)	150 (75%)
Total	100	100	200

- Null table would look like this:

	Atheists	Christians	
Cats	25 (25%)	25 (25%)	50 (25%)
Dogs	75 (75%)	75 (75%)	150 (75%)
Neither	100	100	200

Steps in Chi-Square Test of Independence

3. *Test Statistic:* $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

AGAIN: You are NOT doing this by hand!!!!
Seriously. PLEASE. Just no. Use SPSS's #s!!

where $f_e = \frac{(\text{Row Total})(\text{Column Total})}{\text{Total Sample Size}}$

4. *P-value:* Use Table C. P is the right-hand tail beyond observed χ^2 value, for chi-square with degree of freedom (df) equal to (r-1)(c-1), where r is the number of rows and c is the number of columns.
5. *Conclusion:* Reject H_0 at α-level (a conventional level of significance) if $P < \alpha$.

Steps in Chi-Square Test of Independence

Same five steps as in any other test…

1. *Assumptions*
   1. Two categorical variables
   2. Random sample
   3. Minimum expected frequencies
      1. For 2x2, *expected* observations in *each cell* must be 5 or greater
      - For larger tables, at least 75% should have expected count of 5 or more & all cells must have an expected count greater than 1.
2. *Hypotheses*
   - H_0: Variables are statistically independent in population
   - H_a: Statistical dependence of the variables

Test Statistic: Similar Formula

If you try to calculate this by hand in lab, I will be shook! DON'T!

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Observed number in a given category

Expected number in a given category

1. Chi-Square values are always positive
2. Chi-Square increases as the difference between the observed and expected number increases

Null and Alternative Hypothesis

General Form

H_a : The variables are statistically dependent

H_o : The variables are statistically independent

Example:

H_a : Obedience is statistically dependent upon Region

H_o : Region &"Obedience" are statistically independent

Calculating (Bivariate) Chi^2

Not. By. Hand.

- Test statistic – as always, want it to be large so that the p-value will be small, so we can reject the null

$$x^2 = \sum \frac{(f_{observed} - f_{expected})^2}{f_{expected}}$$

- Frequencies:
  - The observed frequency for any cell is just the raw count in that cell.
  - The expected frequency for any cell is equal to the total number of cases in that row multiplied by the total number in that column, all divided by the total sample size.

$$f_{expected} = \frac{(rowtotal)x(columntotal)}{samplesize}$$

Dependence? (Observed data)

		REGION: North	REGION: South	Row Total
Obey Most Important	Yes	127 (17.6%)	68 (25.9%)	195 (19.9%)
	No	592 (82.4%)	195 (74.1%)	787 (80.1%)
Column Totals		719 (73.2%)	263 (26.8%)	982

- Differences suggest that respect for obedience varies by region
  - Both **Across rows, and**
  - **Modal percentages** (highest in each row)
- The null focuses on the **row marginals**

Expected & Observed Freqs

- Chi-square applies row % to column counts
- Estimates how many *per* cell if null true
  - e.g. 19.9% x 719 = 142.8

		REGION: North	REGION: South	Row Total
Obey Most Important	Yes	E=142.8 O=127	E=52.2 O=68	195 (19.9)
	No	E=576.2 O=592	E=210.8 O=195	787 (80.1)
Column Totals		719 (73.2)	263 (26.8)	982

(Null) Table of Independence

What would the numbers look like if there was not an association?

Same percentages across each row

		REGION: North	REGION: South	Row Total
Obey Most Important	Yes	f_e=19.9%	f_e=19.9%	195 (19.9)
	No	f_e=80.1%	f_e=80.1%	787 (80.1)
Column Totals		719 (73.2)	263 (26.8)	982

Computing a Chi-Square Statistic

Does the same thing (f_e and more) for every cell

Cell	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
Yes-N	127	142.8	-15.8	249.64	1.75
Yes-S	68	52.2	15.8	249.64	4.78
No-N	592	576.2	15.8	249.64	0.43
No-S	195	210.8	-15.8	249.64	1.18
Total	982	982	0	----	8.14

This is the Chi-Square

Computing Expected Frequencies

For any cell,

$$f_e = \frac{(\text{row total})(\text{column total})}{\text{total sample size}} = \text{row\% x col\#}$$

For those in the north who chose "obey":

$$f_e = \frac{(195)(719)}{(982)} = (0.199)(719) = 142.8$$

Another Example: Observed Data

- **Parental style (DV) by political affiliation (IV)**

	Conservative	Moderate	Liberal	Total
Permissive	7 (21.9%)	9 (30.0%)	14 (51.9%)	30
Moderate	10 (31.3%)	10 (33.3%)	8 (29.6%)	28
Authoritarian	15 (46.9%)	11 (36.7%)	5 (18.5%)	31
Col. Marginal	32	30	27	89

- **Modal %s (highest in each row) fall on "main diagonal"**
  - **Corner-to-corner pattern predicted by our H_A**

Example2, cont'd: Null Table

- **Table of calculated Expected Frequencies (f_e)**

$$f_{expected} = \frac{(rowtotal)x(columntotal)}{samplesize}$$

	Conservative	Moderate	Liberal
Permissive	10.79 = 30×32/89	10.11 = 30×30/89	9.10 = 30×27/89
Moderate	10.07 = 28×32/89	9.44 = 28×30/89	8.59 = 28×27/89
Authoritarian	11.15 = 31×32/89	10.45 = 31×30/89	9.40 = 31×27/89

- **This table illustrates statistical independence, to compare to the data we actually observed**

Example 2: Calculating Chi²

f_O	f_E	(f_O-f_E)	$(f_O-f_E)^2$	$(f_O-f_E)^2/f_E$
7	10.79	-3.79	14.36	1.33
9	10.11	-1.11	1.23	.12
14	9.10	4.90	24.01	2.64
10	10.07	-.07	.005	.00
10	9.44	.56	.31	.03
8	8.49	-.49	.24	.03
15	11.15	3.85	14.82	1.33
11	10.45	.55	.30	.03
5	9.40	- 4.40	19.36	2.06
				7.57

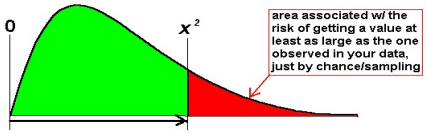
$$= \sum \frac{(f_o - f_e)^2}{f_e}$$

Bivariate Chi²: Interpretation

- Chi² itself has no interpretation (like variance or "F")
  - Summarizes how different the observed data is from the null table
  - Can only understand that summary by use of a p-value
  - Varies with the "degrees of freedom"
    - For a 1-variable chi-square test, df = categories minus 1
    - For a 2-variable chi-square, df = (# rows – 1) x (# columns – 1); here, df = 4
- Table is similar to t, F, and other tables
  - Computed value is 9.488 is how big chi² would have to be to leave a rejection region ("tail") of only 0.05 (5%)
  - Since our chi² is smaller than that, our p-value is bigger
    - The difference does not stretch to the right as far as we need it to
    - The "rejection region" is thus bigger than we wanted it to be
  - The risk of being wrong in rejecting the null is too great
  - Since we can't reject the null (of independence) we don't find support that there is a dependent relationship

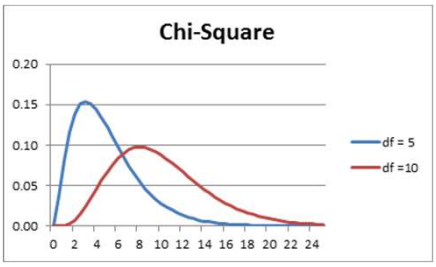
The Chi-Square Distribution

- **As in all statistical tests, some observed distance is associated with a p-value (probability level) of getting *more* than that distance – risk of being wrong if we reject the null**
- **Here, the "distance" is between observed frequencies (i.e. your research data) & expected frequencies (i.e. what the data would look like if there were *no* relationship, and therefore if the null hypothesis was "true")**
- **If & only if p value is smaller than 0.05 (5%), we reject the null.**



Shape changes w/ Degrees of Freedom

- For example, flatter w/ a larger table (more rows or columns):



athiestcats.sav

Atheists & Cats in SPSS....

- I've put the data above (30/20) into a dataset
- I'll show you xtabs & chisquare output from it
  - There ***is*** a diagonal pattern, but ***not*** a significant one
- I also created a 2nd Pet variable, showing what happens if *all* Athiests have cats & *no* Christians do
  - Very different p value, amirite?

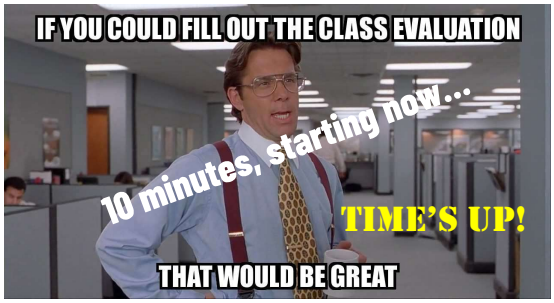
Required Lab: Revisit Your Crosstab

- Use whatever dataset and whatever two variables you used to make a crosstab in the previous lab – but MUST SUBMIT IT AS A SEPARATE LAB!
- Request, report, and interpret chi-square output (note: use Pearson's), this time including all five steps of a hypothesis test for your analysis
- Yes, if in once-weekly labs, this means use the data you just used minutes ago – but you may want to show me what you did, before you proceed, to ensure it made sense ☺

Extra Credit Lab: RSHORD x YEAR

- **Create a dichotomous measurement of whether someone watches *any* reality shows**
  - Look at a frequency distribution first
  - Be sure to declare for any missing values, if needed
  - Use RECODE…INTO DIFFERENT (not COMPUTE)
- **Request a crosstabulation that examines whether seniors are less likely to watch reality shows than are juniors**
  - 3 = junior, 4 = senior, 9 = missing (!)
- **Write a few sentences summarizing the cell percentages (pay special attention to *modal percentages*), and a sentence summarizing the chi-square results.**

http://www.csun.edu/sef



10 9 8 7 6 5 4 3 2 1