PS 0700
Basic Statistical Methods: Crosstabulation

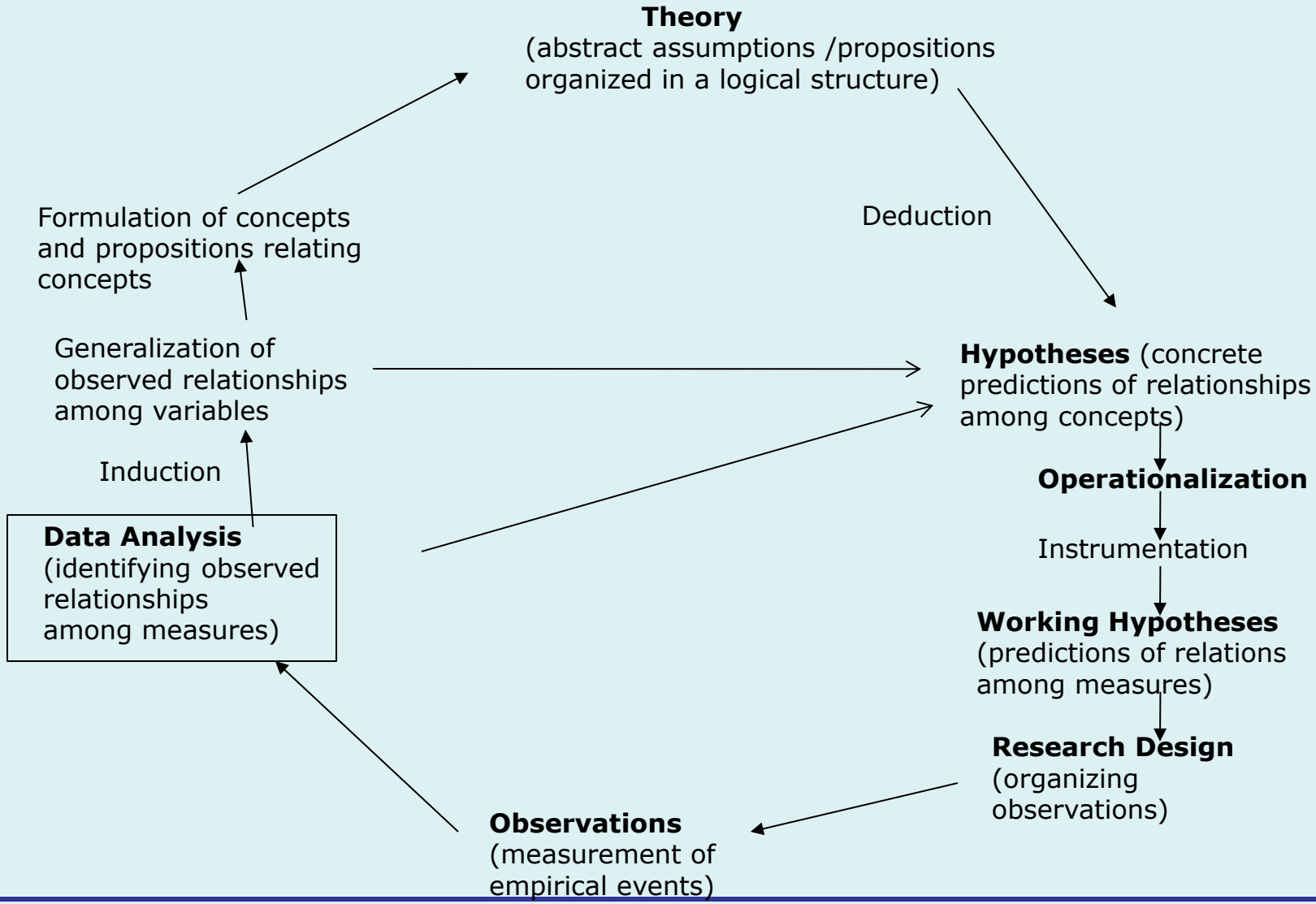
Political Science Research Methods

Professor Steven Finkel

Fall Semester 2022

Week 12

A Model of the Research Process



Crosstabulation

- When you want to examine the relationship between nominal or ordinal independent variables and nominal/ordinal dependent variables, the appropriate statistical technique is called "crosstabulation" or "crosstabs" or "contingency table analysis"
- A crosstab is simply a table that shows you the percentages (or proportion) of units from each category of the independent variable that are also in each category of the dependent variable
- We then compare the percentages (proportions) in some category of the DV for each category of the IV, and assess how strongly the two variables are related
- After that, we conduct a statistical test called "chi-square" (χ^2) to see whether the relationship is *statistically* significant, i.e., whether the results we see in our sample were so unlikely to have come about if there were *no relationship* in the population that we reject the null hypothesis of "no relationship"
- We follow the same steps in hypothesis testing as we did with "t-tests" in the previous lectures

The Relationship between Immigration Status (IV) and Big or Small Salary Increases (DV)

```
library(haven)
job <- read_dta(file = "job_training.teaching.dta")
```

```
library(gmodels)
CrossTable(job$bigchange, job$immigrant, expected = FALSE, prop.c=TRUE, prop.r = FALSE, prop.t = FALSE, prop.chis
q = FALSE, chisq = FALSE)
```

```
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  474
##
##
##                | job$immigrant
## job$bigchange |         0 |         1 | Row Total |
## --------------|-----------|-----------|-----------|
##             1 |       162 |        75 |       237 |
##               |     0.438 |     0.721 |           |
## --------------|-----------|-----------|-----------|
##             2 |       208 |        29 |       237 |
##               |     0.562 |     0.279 |           |
## --------------|-----------|-----------|-----------|
##  Column Total |       370 |       104 |       474 |
##               |     0.781 |     0.219 |           |
## --------------|-----------|-----------|-----------|
```

Note: The salary increase variable (salchange) was recoded earlier for pedagogical purposes into two categories (Small/Big) divided at the median ($1324). This is the new variable "bigchange".

Category 1 is "small change", and Category 2 is "big change"

```
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  474
##
##
##                | job$immigrant
## job$bigchange |         0 |         1 | Row Total |
## --------------|-----------|-----------|-----------|
##             1 |       162 |        75 |       237 |
##               |     0.438 |     0.721 |           |
## --------------|-----------|-----------|-----------|
##             2 |       208 |        29 |       237 |
##               |     0.562 |     0.279 |           |
## --------------|-----------|-----------|-----------|
##  Column Total |       370 |       104 |       474 |
##               |     0.781 |     0.219 |           |
## --------------|-----------|-----------|-----------|
##
```

Each "cell" contains the raw number of people who fall within it, and the proportion of immigrants or non-immigrants who fall in that particular category of the dependent variable (this is the COLUMN proportion or, multiplied by 100, the column percentage– VERY IMPORTANT!!!!)

So: 208 non-immigrants (.562 or 56.2% of all non-immigrants) showed a "BIG" salary increase, while 29 immigrants (.279 or 27.9%) of all immigrants showed a "BIG" salary increase

Things to Note in Any Crosstab Table

- What is the independent variable and what is the dependent variable? (Independent variable *should* be across the columns, dependent variable *should* be down the rows, but make sure this is the case)
- What proportion or percentage of cases in one category of the IV falls into a given category of the DV, and what proportion or percentage of cases in the other category or categories of the IV falls into that **same category of the DV**? This will tell you the *raw bivariate relationship* between the IV and the DV
- Here we have 56.2% of non-immigrants showing "big" salary increases, while only 27.9% of immigrants show "big" salary increases. This is indicative of a 28 percentage point difference, and this tells you that there is a *moderate* relationship between the two variables

Characterizing the Strength of a Relationship

- There are no hard and fast rules about how strong a relationship is based on the percentage difference that is shown in a crosstab table
- Informally, some rules of thumb:
  - Percentage differences of 10 or less are pretty weak
  - Percentage differences of 10-25 are "modest"
  - Percentage differences of 25-40 are "moderate"
  - Percentage differences greater than 40 are "strong"
- But:
  - These are not hard and fast rules
  - Even strong substantive relationships may not be statistically significant (and vice versa!)
  - There are lots of statistics that can be calculated to give added insight into the strength of the relationship. These are called "measures of association", but we don't have time to cover them

Establishing Statistical Significance

- Next step: Is the relationship "statistically significant?"
- As we have discussed, a "statistically significant" relationship is one that is very unlikely to have come about by chance, or by a random sampling process that could have produced the *appearance* of a relationship in a given sample ***even though no relationship really exists in the overall population***
- So we will specify the null hypothesis ("no relationship in the population"), construct a statistical test to see how different our sample relationship is from a situation of "no relationship", and then determine how likely it was to have observed our sample relationship *if* there truly was no relationship in the population
- If there was less than a 5% chance, we will **reject** the null hypothesis and claim that there is a *statistically significant relationship at the .05 level between the IV and the DV*
- The statistical test is the "chi-square test of independence" (χ^2)

Step 1: Specify the Null and Research Hypotheses

Null Hypothesis: No relationship in the population

- H_0: Immigration status and the size of salary increases are unrelated in the population; or (equivalently)
- H_0: Immigration status and the size of salary increases are *statistically independent* in the population

Research or Alternative Hypothesis

- H_A: Immigration status is related to the size of salary increases in the population; specifically
- H_A: Immigrants are less likely to register big salary increases than non-immigrants

Step 2: What would our Crosstab Table Have Looked Like if the Null Hypothesis were True?

- We need to calculate the *expected* number of people in each cell under the condition that the null hypothesis is true
- If there were no relationship between the IV and the DV in this case, then we would expect 50% of immigrants to have big salary increases and 50% of non-immigrants to have big salary increases. This is because 50% of the entire sample has big salary increases, so if there is no relationship between salary and immigration status, 50% of people in **every** category of the IV should have a big salary increases
- There is nothing special about 50% here that would necessarily apply to other tables – the procedure is based on the overall percentage of a sample in a given category of a dependent variable
- For example, if, say, 26% of the overall sample graduated from college, and we were testing the effect of immigration on college graduation rates, the prediction from the null hypothesis would be that 26% of immigrants graduated from college and 26% of non-immigrants graduated from college

Expected Cell Counts under H_0

- Lower Left Cell: There are 370 non-Immigrants, so if 50% of them would be expected to have big salary increases if the null hypothesis were true, that would mean 185 people in the lower left cell
- Lower Right Cell: There are 104 immigrants, so if 50% of them would be expected to have big salary increases if the null hypothesis were true, that would mean 52 people in the lower right cell
- You can do these calculations for all of the cells of any cross-tab table
- Shortcut: (Lower left): Total in that row (237)* Total in that column (370), Divided by Overall Total (474)=185
- Shortcut for the Expected Cell Count in a given cell: R*C/T
  - Row Total*Column Total, Divided by Total N

The "Expected" and "Observed" Crosstab Tables

```
##
##
##                | job$immigrant
## job$bigchange |         0 |         1 | Row Total |
## --------------|-----------|-----------|-----------|
##             1 |       162 |        75 |       237 |
##               |     0.438 |     0.721 |           |
## --------------|-----------|-----------|-----------|
##             2 |       208 |        29 |       237 |
##               |     0.562 |     0.279 |           |
## --------------|-----------|-----------|-----------|
##  Column Total |       370 |       104 |       474 |
##               |     0.781 |     0.219 |           |
## --------------|-----------|-----------|-----------|
```

Observed (O)

Expected (E)

```
#
#                | job$immigrant
# job$bigchange |         0 |         1 | Row Total |
# --------------|-----------|-----------|-----------|
#             1 |       162 |        75 |       237 |
#               |   185.000 |    52.000 |           |
# --------------|-----------|-----------|-----------|
#             2 |       208 |        29 |       237 |
#               |   185.000 |    52.000 |           |
# --------------|-----------|-----------|-----------|
#  Column Total |       370 |       104 |       474 |
# --------------|-----------|-----------|-----------|
#
```

```
CrossTable(job$bigchange, job$immigrant, expected = TRUE, prop.c=FALSE, prop.r = FALSE, prop.t = FALSE, prop.chis
q = FALSE, chisq = FALSE)
```

Calculating Chi-Square, the Test Statistic

- How different is the "observed" table from what would have been "expected" under H_0? We summarize the extent of the deviation between *O* and *E* with a statistic called "chi-square" (χ^2)

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

- So for each cell of the table, take the observed frequency, subtract the expected frequency, square it, and divide by the expected. Then sum these quantities for all the cells, and that is χ^2

Our Example

- $\chi^2 = (162\text{-}185)^2/185 + (208\text{-}185)^2/185)$

  $+ (75\text{-}52)^2/52 \quad + (29\text{-}52)^2/52$

  $= 2.86 + 2.86 + 10.17 + 10.17 = \mathbf{26.06}$

- Now, is this value large enough for us to say that it was unlikely to have come about by chance?
- It turns out that we know the exact sampling distribution of χ^2 under the H_0, that is, we know how likely it is to obtain χ^2 of different values under repeated random sampling from populations where H_0 is true
- R will produce a "p" –value for the χ^2 and if it is less than .05, we know that it *would have been very unlikely to have observed a* χ^2 of 26.06 or larger if the null hypothesis were true. We then say that we reject H_0 at the .05 level, or that the relationship between immigrant status and salary increases is statistically significant

```
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  474
##
##
##                | job$immigrant
## job$bigchange |         0 |         1 | Row Total |
## --------------|-----------|-----------|-----------|
##             1 |       162 |        75 |       237 |
##               |   185.000 |    52.000 |           |
##               |     0.438 |     0.721 |           |
## --------------|-----------|-----------|-----------|
##             2 |       208 |        29 |       237 |
##               |   185.000 |    52.000 |           |
##               |     0.562 |     0.279 |           |
## --------------|-----------|-----------|-----------|
##  Column Total |       370 |       104 |       474 |
##               |     0.781 |     0.219 |           |
## --------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  26.06507     d.f. =  1     p =  3.301017e-07
##
```

"Degrees of Freedom"

- The "critical value" to be able to say that χ^2 is statistically significant --- that is was unlikely to have come about by random sampling error --- depends on what is known as the "degrees of freedom" in the table
- Degrees of freedom (*df*) in crosstabulation refers to how many *cell frequencies* are free to take on *any value* before you know *all of the cell frequencies*, given the **total** number of people in each row and column of the table
- For example: in our case we have 370 non-immigrants, 104 immigrants, and 237 people with "small" increases and 237 people with "big" increases
- Given this information, how many cells in the table are free to vary? If we know, for example, that the upper left cell has 200 people, then the upper right *must* have 37 people in it, the lower left cell *must* have 170 people, and lower right *must* have 67 people
- ***Therefore a 2x2 table with 2 rows and 2 columns has 1 df!!!!***

- How about a 2x3 table, with 2 rows and 3 columns? How many cells are free to vary? That's right, 2!!
- How about a 3x3 table? That's right, 4 !!!!
- We can calculate the *df* of any table by the following formula: (r-1)*(c-1)
- With larger *df*, your χ^2 needs to be larger to reject H_0
- This has been worked out for all possible *df*, and R will provide the exact "p" of observing the magnitude of χ^2 that you observed, given the number of df in your table and given that H_0 is true.
- **Whenever this "p" value is less than .05, reject H_0 at the .05 significance level, and say that the relationship between X and Y is "statistically significant", or very unlikely to have come about by chance from a population where no relationship between X and Y exists**

Notes on Crosstabulation

- We could still be wrong. That is, we will make a mistake 5 out of 100 times -- we will say there is a relationship between the variables when there really isn't. All statistical analysis represents some risk of error, but when the chances are so small we are willing to take that risk
- Establishing a statistically significant relationship between variables does not mean that there is a *strong* relationship between the variables. Increasing sample size, for example, will make relationships "significant" in the sense of being different from zero, but they may not be substantively important. So do not rely on χ^2 to tell us how strong a relationship is substantively. Use Proportion or Percentage Differences for that (or the many measures of association we don't have time to talk about!)
- Establishing statistical significance does not establish a *causal* connection between variables either -- for that you need to "control" for all those outside Z variables that might contaminate the relationship