Day 0 Statistics for Day 1 Impostor Syndrome

Soren Jordan

August 23, 2019

Day 1 impostor syndrome

- What have I done? Why am I here?
  - You picked a program in Public Administration *on purpose*, because some part of you cares about this work and has a goal to accomplish
- How am I supposed to read this much?
  - Make a schedule that works for you. Become *way* more effective at budgeting your time (the one equally distributed resource!)
- What do all of these numbers mean? What do all of these words mean?
  - If you knew that already, you wouldn't need the classes! Just be willing to *learn* and *ask questions*
- I bet everyone else already gets it. I'm so behind
  - Literally everyone feels this way

Day 0 scientific vocabulary

- First priority: feel comfortable with the *scientific vocabulary*
- *Theory*: a general explanation of how the world works
    - What qualifies as a "theory"? Depends on the scientist
    - Universally believed that you cannot directly test a theory
    - Which is why we *never* say the word "prove" in science-land
- *Model*: a pretty general explanation of how a more specific part of the world works
    - Most importantly: is not a "theory"
- *Hypothesis*: a *specific, testable* proposition derived from a theory or a model
    - The standard-form hypothesis: as $X \uparrow / \downarrow$, $Y \uparrow / \downarrow$
- "All models are wrong, but some are useful." –George Box (–Michael Scott)

Day 0 measurement vocabulary

- Our theories are built around *concepts* that are related
- In the social sciences, those concepts are not really observed numerically, with few exceptions
    - I believe that autocratic governments affect economic growth differently than democracies for . . . IR reasons
    - I can record the number for GDP
    - How about "dictatorship" or "authoritarianism"?
- *Operationalization*: moving from *concept* to *measurement*
- *Validity*: how well our measure reflects the concept
    - The presence of free and fair elections
    - Self-reporting as a dictatorship
- Even in qualitative research, it's useful to have indicators to help us identify changes. This is often called *scoring*

Day 0 data vocabulary

- Our theories relate an *independent* variable to a *dependent* variable
- Independent variable: X. The thing doing the causing (explanatory variables)
- Dependent variable: Y. The thing being caused
- Often easiest to express as a "formula"
  - $Y = X_1 + X_2 + X_3$
- Are we equally interested in all of the X variables? Usually not
  - Whichever one you build your dissertation around
  - Like nursing home proximity
- *Control* variables versus independent variables "of interest" or "key"

Day 0 data vocabulary

- Independent: it's in the name
- Which is sometimes why we refer to X as being “fixed” in repeated samples (hold on)
- But imagine: $Y =$ Vote choice
- $X =$ Age, Gender, and Presidential feelings on a 0-100 scale (“feeling thermometer”)
- Do some of these feel more “fixed” than others?
- (Like most other problems, we usually just assume this away)
- But it's a unique challenge of the social world

Day 0 data vocabulary

- Dataset: your collection of X and Y for some selected group
- *Variables* are in *columns*
- *Observations* are in *rows*
  - We call them the “units of analysis” or “cases”: what kind of level are you observing?
- So one whole row is a single observation (you’re “observing” something on all of your variables and recording the measures)
- “Data” is technically the plural of “datum,” because we collect multiple measures for each observation
  - So you’ll see sentences like “The data were collected . . .”
  - Welcome to academics

Day 0 data vocabulary

- Sample observational data set

Time	Y	X	Z_1	$\dots$	Z_n
001	21	0	10	$\dots$	0
002	28	1	9	$\dots$	1
003	37	5	12	$\dots$	1
004	46	4	10	$\dots$	1
005	48	0	7	$\dots$	0
006	50	0	18	$\dots$	1
007	72	4	21	$\dots$	1

Day 0 data vocabulary

- Z?
- We often refer to control variables as "spurious" variables
- Failing to control for relevant independent variables will often lead to the wrong conclusions
- "Omitted variables *bias*"
- "All models are wrong, but some are useful."

Day 0 statistical vocabulary

- *Bias*: systematically away from the “true” answer
- *Unbiased*: approximately the “true” answer
- *Consistency*: in repeated procedures, an estimate (of something) converges to the “true” answer
- Truth?

Day 0 statistical vocabulary

- *Population*: the universe of cases we are interested in
- People = all people
- Dictatorships = all dictatorships (over all time? How "general" is that theory?)
- Do we ever collect data on all cases? No.
- *Sample*: a subset of the universe of cases we are interested in
  - Good ones are *random* and *representative*

Day 0 statistical vocabulary

- We use samples to test hypotheses about populations
- This is called statistical *inference*: using something we know (the sample) to talk about something we don't (the population)
- The most common framework (but not the only one) is "frequentist statistics"
- There is a true value for the relationship between X and Y in the population, and we can estimate it using the sample

Day 0 statistical vocabulary

- Using statistics, we can quantify how likely that the "true" population value is *not* some number, given our sample estimate
  - Because we can use samples to rule things *out*, but not rule things *in*
- Most often, we try to rule out the *null hypothesis*: the likelihood that X and Y are unrelated
- This is *statistical significance*: how confident can we be that X and Y are *not unrelated* (notice the double negative) in the population, using our sample
  - Gather sample data
  - Calculate the relationship
  - See how likely certain relationships are in the population
  - See how confident we can be to rule out no relationship
- We summarize this confidence at a completely arbitrary level: 95% (most of the time)

Day 0 statistical vocabulary

- If we can be 95% sure that X and Y, we reject the null hypothesis that *they are not unrelated* **in the population** and conclude that *they are related* **in the population**
- That is statistically significant: you'll often see $p < 0.05$ (where p summarizes our confidence in that null hypothesis rejection)
- Relationships can also be *substantively significant*: whether it is a relationship that "makes a difference"
  - Up to the scientist to interpret!

Day 0 statistical vocabulary

- More formally: really smart statisticians spend their time deriving distributions of statistics
- We use data to generate a test statistic
- That statistic follows a distribution
- We see how unlikely the null hypothesis is, given that distribution
- There are *lots* of distributions, but you'll see the normal, the t, and the χ^2 the most

Day 0 statistical methods

- Σ is the summation operator
  - So a mean is calculated as $\frac{\Sigma_{i=1}^{n} X_i}{n}$
- Median, mode
- Variance: sum of the squared deviations from the mean (a measure of spread)
  - So variance is calculated as $\Sigma_{i=1}^{n}(X_i - \bar{X})^2$
- Problem: variables have different means, so variance could "explode" because a variable has large values
- Solution: standard deviation, the square root of variance ($\sqrt{Var}$)
  - So standard deviation is calculated as $\sqrt{\Sigma_{i=1}^{n}(X_i - \bar{X})^2}$
- Central tendency versus spread ("moments" of a distribution)

Day 0 statistical methods

- Bivariate (two-variable) methods
- Tabular analysis:
  - If two variables can be summarized by a table, how different are the entries in that table from no relationship?
  - $\chi^2 = \Sigma \frac{(O-E)^2}{E}$, where O is observed and E is expected
- Difference of means:
  - $t = \frac{(\bar{X_1} - \bar{X_2})}{se(\bar{X_1} - \bar{X_2})}$
  - Where X_1 and X_2 are two groups differentiated by an outcome variable
- Pearson's r
  - Formula: $r = \frac{cov_{XY}}{\sqrt{var_X var_Y}}$
  - t statistic: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

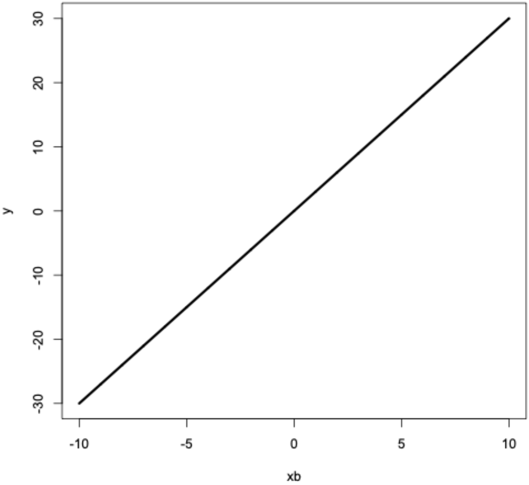
Day 0 statistical methods

- Most common: *regression*
- The logic: the best way to summarize a relationship is a line
- $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Where each of the β links the X to Y in the usual rise over run way

Day 0 statistical methods

- That's the "linear model"
- We can extend that linear model by using functions to transform the line into other shapes
- Probability curves, etc.
- This is the *generalized* linear model

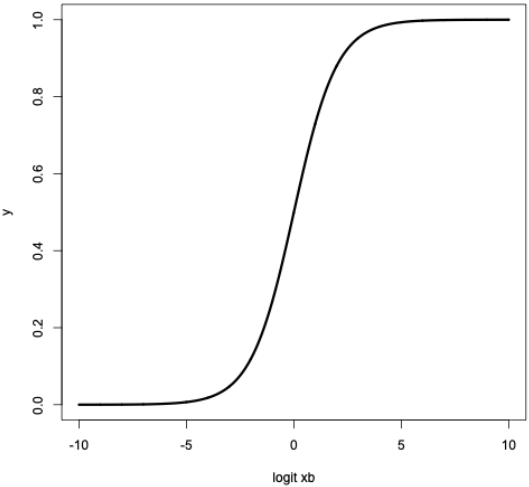
Day 0 statistical methods



Day 0 statistical methods

- $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Now imagine if we transform the right-hand side
- $\frac{\exp X}{1+\exp X}$
- The “logistic” transformation

Day 0 statistical methods



Day 0 statistical vocabulary

- *Random* and *stochastic*: without causes
- *Systematic*: has a cause!
- We usually think there is a *random* component to our models, as it's the social world and no one is right 100% of the time
- $Y = X_1 + X_2 + X_3 + \epsilon$
- Where ϵ is *error*

Practice

- Let's read and explore together
- Read `tinyurl.com/kynect-read`, making note of *everything* you do and don't understand
- Pay special attention to the analysis section
- Then we'll come back and discuss together!