Data, Causality, Stats Review

Tyler Ransom

Univ of Oklahoma

Jan 17, 2019

Today's plan

1. Review reading topics
   1.1 Types of data
   1.2 How type of data informs us about causality
   1.3 Statistics review
2. In-class activity: Working with data in R

Types of data

Experimental vs. Observational

- **Experimental data:**
  - Data from controlled experiments
  - e.g. a biologist experiments with two genetically identical plants
  - Relatively uncommon in the social sciences
- **Observational data:**
  - a.k.a. Nonexperimental data; Retrospective data
  - Data collected passively, after observing some outcomes
  - Collectors act as **observers** of what has happened
  - e.g. Census Bureau surveys households and asks about income & education

Cross-sectional, Time series, and Longitudinal Data

- **Cross-sectional:** Data on multiple units collected at a single point in time

- **Time series:** Data on one unit collected at a multiple points in time

- **Longitudinal:** Multiple units are followed over multiple time periods

    - “Longitudinal” a.k.a. “Panel”

© JosephJoseph

Cross-sectional, Time series, and Longitudinal Data

- **Cross-sectional:** Data on multiple units collected at a single point in time
  - e.g. Freshmen at OU in Fall 2018
- **Time series:** Data on one unit collected at a multiple points in time
- **Longitudinal:** Multiple units are followed over multiple time periods
  - "Longitudinal" a.k.a. "Panel"

© JosephJoseph

Cross-sectional, Time series, and Longitudinal Data

- **Cross-sectional:** Data on multiple units collected at a single point in time
  - e.g. Freshmen at OU in Fall 2018

- **Time series:** Data on one unit collected at a multiple points in time
  - e.g. Nominal GDP of USA

- **Longitudinal:** Multiple units are followed over multiple time periods
  - "Longitudinal" a.k.a. "Panel"

© JosephJoseph

Cross-sectional, Time series, and Longitudinal Data

- **Cross-sectional:** Data on multiple units collected at a single point in time
  - e.g. Freshmen at OU in Fall 2018
- **Time series:** Data on one unit collected at a multiple points in time
  - e.g. Nominal GDP of USA
- **Longitudinal:** Multiple units are followed over multiple time periods
  - e.g. OU Class of 2019, surveyed each academic year
  - "Longitudinal" a.k.a. "Panel"

© JosephJoseph

How the type of data informs us about causality

How the type of data informs us about causality

- Experimental data can be cross-sectional, time series, or longitudinal
- The two ideas are not related
- Experimental/Observational tells us how confident we can be that correlation $\Rightarrow$ causation
- Cross-Sec/Time-Series/Panel tells us how to correctly compute correlation

Stats Review

Sampling

- Cross-sectional data is a **random sample** from some population
- We look at data because we want to learn something about the population
  - e.g. Estimate a statistic
  - e.g. Test a hypothesis
- A random sample:
  - is representative of the population of interest
  - gives us the best chance of learning about the population

Estimators and Estimates

- **Estimator of θ:** Some rule that assigns each random sample a value of θ
  - e.g. Sample average, $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$
  - in this case, the *population parameter* is μ (the sample mean); $\overline{Y}$ estimates it
- **Estimate of θ:** the value the estimator spits out given a random sample
- Estimates depend on the sample $\Rightarrow$ estimates are random variables
  - **Sampling distribution:** The distribution of estimates (given all samples)
  - can characterize it by using summary stats (mean, variance, etc.)

Bias

- **Bias of an estimator W:** $\text{Bias}(W) = E(W) - \theta$
  - i.e. if we take a bunch of samples, the average of all estimates $= \theta$
  - $\overline{Y}$ is an unbiased estimator of μ
  - $\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$ is an unbiased estimator of σ^2
- But some poor estimators are unbiased
- And some great estimators are biased
- So bias isn't everything; but all else equal, we like unbiased estimators

Sampling Variance and Efficiency

- Bias focuses on the average of an estimator's sampling distribution
- It's also important to think about the *variance* of the sampling distribution
  - e.g. Variance$\left(\overline{Y}\right) = \frac{\sigma^2}{N}$
- An estimator is **efficient** if it has a lower sampling variance than all other estimators
- As econometricians, we should be using *unbiased* and *efficient* estimators!

Asymptotic Properties of Estimators

- **Asymptotic** means ”infinite-sample”
- A good way to evaluate an estimator is to look at its properties as $N \to \infty$
- An estimator W is **consistent** if its sampling distribution becomes more and more centered on θ as $N \to \infty$ (see: Law of Large Numbers)
- W is **asymptotically normal** if its sampling distribution increasingly resembles a Normal distribution as $N \to \infty$ (see: Central Limit Theorem)
- The best estimators are those that are CAN: Consistent and Asymptotically Normal

Hypothesis Testing

- **hypothesis testing:** A method to answer yes/no questions using a sample of data
  - e.g. Are Asian-Americans discriminated against in admissions to Harvard?

- Define null (H_0) and alternative (H_a or H_1) hypotheses
  - e.g. let $\theta = \mu_a - \mu_w$ be the difference in admissions rates of white and Asian Harvard applicants
  - if discrimination, then should have $\theta < 0$

$$H_0 : \theta = 0$$
$$H_a : \theta < 0$$

  - How low of T needed to conclude discrimination? (depends on $\text{Var}(T)$!)

Significance and Power

- **significance level:** our tolerance for making a Type I error (rejecting H_0 when we shouldn't)

- **power:** likelihood of *not* making a Type II error (failing to reject H_0 when we should)

- Set a significance level α (e.g. 5%)

- α quantifies our tolerance to make a Type I error

- Subject to α, want to maximize power

- There is a trade-off between power and significance

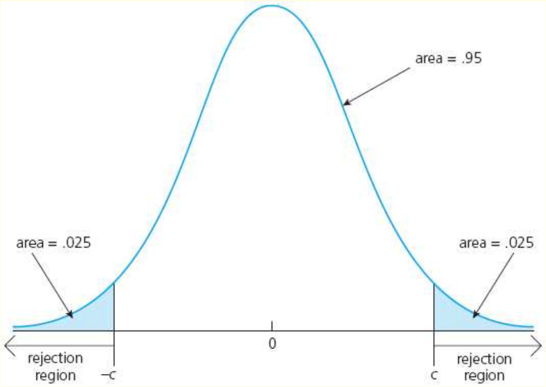
What goes into a hypothesis test

- Need two things to conduct a hypothesis test:

    1. **Test statistic (T):** Some function of the sample of data

    2. **Critical value (c):** Value of T such that we reject H_0 if, e.g. $|T| > c$

- c is implicitly a function of the significance level α



A two-sided test with $\alpha = 0.05$ (Wooldridge Fig. C.6)

Steps to performing a hypothesis test

- Declare a significance level (5% is most common)
- Determine if your H_a is one- or two-sided (two-sided most common)
- Determine what distribution your test statistic will have (t is most common)
- Given T and α, compute critical value
- Compute the value of T for your sample
- If $|T| > c$, reject H_0; otherwise, fail to reject H_0

Example: Discrimination in Harvard admissions

- Recall our hypothesis test introduced earlier:

$$H_0 : \theta = 0$$
$$H_a : \theta < 0$$

  where $\theta = \mu_a - \mu_w$ is the Asian-White difference in admissions rates

- Suppose our sample of data contains $N = 10,000$ applicants from each group.

- Let $\alpha = 0.05$. What is c?

  - Look up c in a t distribution table

  - For one-tailed test with $9,999$ degrees of freedom, $c = -1.65$

Example: Discrimination in Harvard admissions

- Suppose $\bar{y}_a - \bar{y}_w = -0.06$; and $\text{se}\left(\bar{y}_a - \bar{y}_w\right) = 0.01$

$$
\begin{aligned}
t &= \frac{\text{estimate} - \text{null}}{\text{std. err.}} \\
&= \frac{-0.06 - 0}{0.01} \\
&= -6
\end{aligned}
$$

- $-6 < -1.65$ so we reject H_0
- conclude that there is evidence of discrimination against Asian-Americans

p-values

- Often, the outcome of the hypothesis test will be summarized by the p-value instead of comparing T and c.

- **p-value:** The largest significance level at which we could conduct the test and still fail to reject H_0.

- Reject H_0 if $p < \alpha$

Hypothesis testing in R

- R will compute test statistics, critical values, and p-values automatically
- We'll practice this in more detail next time