Lecture 8: Comparing Means

Conduct 1 with Nathan Favero

November 10, 2022

AMERICAN UNIVERSITY
WASHINGTON, DC

Review: Hypothesis testing

Q: Identify which is a null hypothesis and which is an alternative hypothesis:

1. $Corr(x, y) \neq 0.2$
2. $Corr(x, y) = 0.2$

A:

1. Alternative
2. Null

Review: Hypothesis testing

Q: If I decide to increase the alpha level from .01 to .10 for all my research, which of the following is true?

(1) It is *harder* to reject a null hypothesis. Risk of false positive (type I error) $\Uparrow$; risk of type II error $\Downarrow$.

(2) It is *harder* to reject a null hypothesis. Risk of false positive (type I error) $\Downarrow$; risk of type II error $\Downarrow$.

(3) It is *easier* to reject a null hypothesis. Risk of false positive (type I error) $\Uparrow$; risk of type II error $\Downarrow$.

(4) It is *easier* to reject a null hypothesis. Risk of false positive (type I error) $\Downarrow$; risk of type II error $\Uparrow$.

A: (3) The alpha level indicates the probability of getting a type I error (if null is true), so increasing the alpha level increases risk of a false positive. A higher alpha level makes it easier to reject the null (easier for the p-value to be smaller than alpha), so risk of type II error decreases.

2-sample t-tests

- Example: Effect of a fuel additive
  - We have a sample of vehicles, some of which have received a fuel additive and some of which have not
  - Our dependent variable is fuel efficiency, measured in miles per gallon (mpg)
  - We want to consider whether the fuel additive has an effect on the mean gas mileage (in other words, does the population of treated cars have a different mean than the population of untreated cars?)
  - The mean for untreated cars is 21 mpg while the mean for treated cars is 22.75 mpg
  - We want to know whether this difference is statistically significant (in other words, how plausible is it that this difference is merely due to sampling error–the imprecision in our estimates caused by relying on a finite sample?)

2-sample t-tests

- There are multiple ways we can conduct hypothesis tests; for now, we'll do what's called a 2-sample t-test
- We could conduct an equivalent test using regression; we'll learn this alternative approach later (see Ch. 12 of the textbook)
- Conceptually, we can think of our two groups as separate populations (one where cars receive the fuel additive, one where they don't)
- We will also think of our sample as being divided into two separate samples–one for each population
- In a 2-sample t-test, we compute the difference in means between our two samples
- For this test, our degrees of freedom are computed as $n_1 + n_2 - 2$, where n_1 is the size of the first sample and n_2 is the size of the second sample

2-sample t-tests

- Example: Effect of a fuel additive
  - $H_0 : \mu_1 = \mu_2$
    - In plain English, the mean fuel efficiency is unaffected by the additive (mean without the additive is the same as mean with the additive)
  - $H_A : \mu_1 \neq \mu_2$ (two-sided)
    - The fuel additive makes a difference for fuel efficiency (could be a negative effect)
  - $\alpha = .05$
  - Type into Stata:

    ```
    use http://www.stata-press.com/data/r14/fuel3
    ttest mpg, by(treated)
    ```
  - $t = -1.43; p = 0.17$
  - Since $p > \alpha$, we fail to reject the null; the results are inconclusive
  - (If p were less than α, we would conclude that the additive made a difference for fuel efficiency; there'd be a statistically significant difference)

2-sample t-tests

- Example: Effect of a fuel additive
    - It can also be nice to create a graph comparing to two means, with confidence intervals (may need to install package first: `ssc install ciplot`):

      ```
      ciplot mpg, by(treated)
      ```

2-sample t-tests

- Assumptions for a 2-sample t-test:
    - Independence of observations (simple random sample/treatment satisfies this)
    - The variable of interest is normally distributed in the population (test performs reasonably well as long as sampling distributions are normally distributed, so we’re usually okay as long as both samples are reasonably large because of the central limit theorem)
    - The variances of the two populations are equal (this assumption is rather strict, but fortunately, there’s an easy correction for it; just use the “unequal” option in Stata)

Confidence intervals vs hypothesis testing

- Confidence intervals have several similarities to hypothesis tests
- Both rely on standard error estimates
- Both require us to pick an alpha level
- The results of the two approaches are interrelated
    - If the parameter value specified in our null hypothesis falls *within* the confidence interval, we fail to reject the null (in a 2-sided test, at least)
    - If the parameter value specified in our null hypothesis falls *outside* the confidence interval, we reject the null (in a 2-sided test)
    - Notice the Stata output for our means tests also provided confidence intervals
        - In the example we just discussed, the confidence interval for the difference between the means (of the two populations) is the one that will directly correspond to the hypothesis test

Confidence intervals vs hypothesis testing

- Hypothesis testing is probably the dominant interpretive tool in applied social science
- Nonetheless, confidence intervals are generally much more informative
- Hypothesis tests only allow you to confirm (or fail to confirm) that there's a non-zero effect; but this effect could be so tiny that we don't care
- In hypothesis testing, failing to reject the null tells us virtually nothing; with a confidence interval, it's possible to conclude that there is–at most–a tiny effect (if the confidence interval contains only a small range of values close to zero)
- These issues will come up prominently in the context of regression

Confidence intervals

- We've mainly focused on 95% confidence intervals, but Stata can compute confidence intervals for other alpha levels (e.g., a 90% confidence interval for an alpha level of .10)
- Stata finds a critical value (t^*) such that the probability of drawing a value from our (assumed) distribution that is further from the mean than the critical value is equal to our alpha-level
- Stata then creates the confidence interval by multiplying this critical value by the estimated standard error and adding/subtracting this from our point estimate

$$C.I. = \bar{x} \pm (t^* \times s_{\bar{x}})$$

Confidence intervals

- Example: Fuel efficiency without additive
  - We can get confidence intervals for two different alpha levels (.05 and .10):

```
use http://www.stata-press.com/data/r14/fuel3
keep if treated==0
mean mpg, level(95)
mean mpg, level(90)
```

One-way ANOVA

- A 2-sample t-test allows us to test for a difference in (population) means when we can divide our sample into 2 categories
- With ANOVA, we can do a difference in means test with more than 2 categories
- Two-sample t-test null hypothesis: $\mu_1 = \mu_2$
  - Alternative: $\mu_1 \neq \mu_2$
- ANOVA null hypothesis: $\mu_1 = \mu_2 = ... = \mu_k$ (where *k* is the number of categories)
  - Alternative: at least one category has a mean that is different from at least one of the others ($\mu_i \neq \mu_j$ for some *i* and some *j*)
  - If there are 4 categories ($k = 4$) and $\mu_1 = \mu_2 = \mu_3$ but $\mu_3 \neq \mu_4$, then null is false and alternative is true

One-way ANOVA

Example: Attitudes about whites

- Do attitudes about whites differ by racial identity?
- Data: survey results from 2016 ANES
  - V162314: feelings about whites
  - V161310X: race of respondent
- Null: $\mu_{white} = \mu_{black} = \mu_{asian} = \mu_{native} = \mu_{hispanic} = \mu_{other}$
- Alternative: At least one racial group has a different mean attitude than one of the others ($\mu_{white} \neq \mu_{black}$ or $\mu_{white} \neq \mu_{asian}$ or $\mu_{black} \neq \mu_{asian}$ or...)
- Type into Stata:

```
oneway V162314 V161310X if V162314 >=0 & V161310X>=0, tabulate
```

- $p = 0.000$, so we reject the null and conclude that there are differences among racial groups
- The results of the F-test don’t allow us conclude anything about which racial groups are different

One-way ANOVA

- Intuition of ANOVA: examine variance to see whether the differences across categories are bigger than one would expect from random noise (sampling error)
- We can think of each category as having its own subpopulation (e.g., a subpopulation of whites, a subpopulation of blacks, etc.)
- We want to know whether all the subpopulations have the same mean (e.g., whether average attitudes about whites are the same for people of all racial categories)

One-way ANOVA

- We can estimate the mean for each category, but we know our sample-based estimates will be imperfect. Even if all categories have the same population mean, our subsample means (e.g., sample mean attitude for each racial group) will probably not be perfectly equal because of sampling error.
- Thus, we need to create a baseline for how much we expect the subsample means to differ because of sampling error (even if all categories have the same population mean) and then see whether the differences are bigger than what we'd expect if all categories have the same population mean

One-way ANOVA

- As with any hypothesis test, if the p-value is less then alpha, we conclude that the null hypothesis is implausible (we reject it), and we accept the alternative hypothesis
    - If $p > \alpha$, we fail to reject the null and conclude nothing
- Assumptions (very similar to two-sample t-test):
    - Independence of observations
    - The variable of interest is normally distributed in each population
    - The variances of all populations (for each category) are equal (Stata automatically conducts a test of this)

One-way ANOVA vs t-tests

- If we only have 2 categories, one-way ANOVA and two-sample t-tests give equivalent results
- ANOVA allows us to move beyond looking at 2 categories at a time, but with more than 2 categories, the hypothesis test does not allow us to conclude *which* categories are different from each other (we just conclude that not all categories are the same)

One-way ANOVA vs t-tests

- Instead of or in addition to ANOVA, we may wish to conduct a series of two-sample t-tests to see if there is evidence that specific pairs of categories differ from one another
    - Example: we can see if blacks differ from Latinos, if whites differ from blacks, if whites differ from Latinos, etc.
- In some cases, we may wish to adjust the threshold for rejecting the null hypothesis when conducting a series of hypothesis tests
    - If we conduct three tests and each one has a 5% chance of producing a false positive (if each null hypothesis is true), the probability of producing *at least one* false positive will be greater than 5%

Multiple-comparison tests

- If we make an adjustment to the p-values (or critical values/alpha levels) when conducting several hypothesis tests, we can make the probability of getting at least one false positive (approximately) equal to the alpha level
- Several adjustment procedures (which make a variety of assumptions) have been developed
    - One of the simplest and most common adjustments is the Bonferroni adjustments
    - Our textbook (pg. 104) explains the Tukey HSD adjustment

Multiple-comparison tests

- Using the `oneway` command in Stata, we can select an option that will automatically compute adjusted p-values for multiple-comparison tests (e.g., `bonferroni`)
    - If we want to use multiple-comparison adjustment, we reject any null hypothesis for which the adjusted p-values is below our alpha level (otherwise, we fail to reject)
    - For any comparison, the adjusted p-value indicates the probability (assuming a true null) of any of the differences (from the multiple comparisons) being as (or more) extreme (relative to sample size and variance) as the one observed for this comparison

Multiple-comparison tests

- Should you use a multiple-comparison adjustment?
    - It depends on what you want to know
    - The adjusted and unadjusted p-values/hypothesis tests tell you different things about how unlikely a particular result is:
        - The unadjusted p-value tells you how likely it is to get a result this extreme for this particular comparison if the null is true
        - The adjusted (e.g., Bonferroni) p-value tells you how likely it is to get at least one result this extreme (out of all the comparisons) if the null is true
    - Different practices with regards to multiple-comparison adjustments are typical in different disciplines and with different types of analyses

Review questions: key points from this lecture

- What is the null hypothesis for a 2-sample t-test? What about for an ANOVA?
- If I conduct a 2-sample t-test and get a p-value that is less than my alpha-level (eg., .05), what do I conclude? What if p is greater than alpha?
- If I conduct an *ANOVA* and get a p-value that is less than my alpha-level (eg., .05), what do I conclude? What if p is greater than alpha?
- Why is it potentially problematic for me to run several hypothesis tests simultaneously? How can a multiple-comparison adjustment help me deal with this problem?