Lecture 7: Hypothesis Testing

Intro Stats with Nathan Favero

American University (Washington, DC)

August 2, 2024

Overview of hypothesis tests

1. Select a null hypothesis and an alternative hypothesis (statements about population parameters)
2. Select an alpha level (how often we're willing to be wrong with this process)
3. Calculate a p-value using a statistical model (which has assumptions) to evaluate a sample of data
4. Comparing the p-value to our alpha level:
    - If $p < \alpha$, we reject the null hypothesis and accept the alternative hypothesis
    - If $p > \alpha$, we fail to reject the null hypothesis and conclude nothing (inconclusive results)

Hypothesis tests

- Example: ADHD medication
  - Taken from textbook (pg. 93-95); full case study: `https://onlinestatbook.com/2/case_studies/adhd.html`
  - Subjects are children with ADHD, and they are asked to complete the delayed gratification multiple times on separate occasions, during which they were given different dosages of the drug
  - We want to compare children's performance when they got no drug (only a placebo) versus when they got a 0.6 mg dose of methylphenidate
  - Key variable is the difference in the delayed gratification score between the two conditions for each child (score with the drug - score without the drug)

Hypothesis tests

- Example: ADHD medication
  - If drug has no effect, the population mean (μ) of the difference (score with the drug - score without the drug) should be 0. So 0 is our baseline.
  - Step 1
    - Our null hypothesis is $H_0 : \mu = 0$
    - The logical alternative is our alternative hypothesis: $H_A : \mu \neq 0$
  - Step 2
    - We'll stick with the standard default alpha level of 0.05, which means we're willing to be wrong 5% of the time with our approach
    - $\alpha = 0.05$

Hypothesis tests

- Example: ADHD medication
  - Step 3 – Calculate p-value in Stata:

```
insheet using https://nathanfavero.com/class/adhd.csv
ttest diff=0
```

Hypothesis tests

```
. ttest diff=0

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    diff |      24    4.958333    1.538726    7.538188    1.775236    8.141431
------------------------------------------------------------------------------
    mean = mean(diff)                                             t =   3.2224
Ho: mean = 0                                     degrees of freedom =       23

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.9981         Pr(|T| > |t|) = 0.0038          Pr(T > t) = 0.0019
```

Hypothesis tests

- Example: ADHD medication
  - Step 3
    - According to Stata (focusing on the lower-middle portion of the output), $p = 0.0038$
  - Step 4
    - $0.004 < 0.05$, thus $p < \alpha$
    - We reject the null hypothesis and conclude that the mean difference between performance with and without the drug is *not* zero (as stated in our alternative hypothesis)
    - In other words, we conclude the medication has an effect on children's delayed gratification

Key elements of hypothesis tests

- Null hypothesis (H_0): a statement that some (population) parameter is equal to a precise value that will be used as a benchmark that the data will be compared to
  - Examples:
    - Population A and population B have the same mean
    - There is no correlation between variables X and Y (in the population)
    - My district's "true" average SAT score is 1000 (the national mean)
  - Usually, the null hypothesis indicates that there is no effect (or difference among groups)
  - We use the null hypothesis as an assumption, and then we test the feasibility of that assumption based on the data
    - We ask: is the null hypothesis believable given the values found in our sample?

Key elements of hypothesis tests

- Alternative hypothesis (H_A): the logical alternative to the null hypothesis
  - Examples:
    - Population A and population B have *different* means
    - There is a (non-zero) correlation between variables X and Y (in the population)
    - My district's "true" average SAT score is greater than 1000
  - While the null hypothesis specifies a precise parameter value/condition (e.g., correlation=0), the alternative hypothesis will specify a range of possible values (e.g., correlation$\neq$0 or correlation>0)
  - 2 types of alternative hypotheses:
    - Directional/1-tailed (e.g., population correlation is *greater* than zero)
    - Non-direction/2-tailed (e.g., population correlation is *different* from zero)

Key elements of hypothesis tests

- Alpha level: the threshold used to decide at what point a result is “unlikely” enough (under a null hypothesis) that the null hypothesis is rejected
    - In hypothesis testing, we’re thinking about whether the data we’ve collected (our sample) would be unusual to observe if the null hypothesis were true
    - Example: if the true population mean is 1000, we would expect that getting a sample mean of 998 is more likely than getting a sample mean of 800
    - But at what point do we decide that our sample data is so unlikely under the null hypothesis (e.g., our sample mean is so far from 1000) that we conclude the null hypothesis must be false?
    - The alpha level represents the threshold we choose

Key elements of hypothesis tests

- Alpha level
  - There's an inherent tradeoff in choosing a threshold
  - If we have very little tolerance for unusual data but the null hypothesis happens to be *true*, we have fairly high odds of falsely rejecting the (true) null with our test
    - A larger alpha increases the risk of a type I error or "false positive" (rejecting a true null)
  - If we have a lot of tolerance for unusual data but the null hypothesis happens to be *false*, we have fairly high odds of failing to reject the null hypothesis even though it's false
    - A smaller alpha increases the risk of a type II error
  - In real research, we never know for sure whether the null hypothesis is true, so we set the alpha level according to how conservative we want to be–how much tolerance we have for the risk of a type I error (versus a type II error)

Key elements of hypothesis tests

- Alpha level
  - The alpha level indicates the probability of getting a type I error *if the null hypothesis happens to be true* (assuming all assumptions of the statistical model are true)
  - An alpha level of .05 means that there is a 5% chance of getting a false positive (rejecting the null) if the null is true
  - The default alpha level is usually 0.05 ($\alpha = .05$)
  - Other common levels that are used are .10, .01, and .001

Key elements of hypothesis tests

- Statistical significance: refers to rejecting the null hypothesis
  - Example: the difference between two sample means is "statistically significant" if the null hypothesis of equal means has been rejected
- Type I error: rejection of a null hypothesis that is actually true (a "false positive")
- Type II error: failure to reject a null hypothesis that is false

Hypothesis tests

- Let's look at another example of a hypothesis test (also from the textbook): `https://onlinestatbook.com/2/tests_of_means/single_mean.html`
  - Question: do subjects prefer the images that are subliminally suggested?
  - Key variable is the number of subliminally suggested images that were chosen (out of 100 total pairs)
  - If subliminal suggestions make no difference, exactly half of the suggested images should be chosen just out of dumb luck, so we use 50 (out of 100) as our baseline
  - Step 1 – $H_0 : \mu = 50$
  - Step 2 – $\alpha = 0.05$

Hypothesis tests

- Example: `https://onlinestatbook.com/2/tests_of_means/single_mean.html`
  - Step 3 – Calculate p-value in Stata:

```
input score
45
48
49
49
51
52
53
55
57
end
ztest score == 50, sd(5)
```

Hypothesis tests

- Example: https://onlinestatbook.com/2/tests_of_means/single_mean.html
  - $p = 0.55$
  - Step 4 – $0.55 > 0.05$, thus $p > \alpha$. We fail to reject the null; the results are inconclusive
  - See Appendix 1 for further more details on z-tests

How Stata calculates p-values

- The only part of a hypothesis test that requires much math is step 3: computing the p-value
- I won't ever expect you to compute your own p-values (we'll let Stata handle that)
- But we'll go over one example of how they're calculated, using this data on subliminal suggestions
- In order to calculate the p-value, we first calculate our test statistic, in this case a z-score or z-statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$$

- We can use the formula we learned earlier to calculate the standard error ($\sigma_{\bar{x}}$) based on the (assumed) population variance (σ_x):

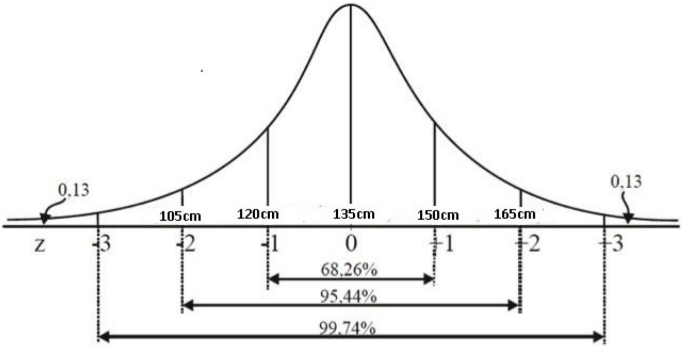
$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

How Stata calculates p-values

- Remember, an estimation process will have a sampling distribution, which will show all the possible values of the estimate (depending on who happens to get selected into our random sample)
- Based on the assumptions of a z-test (see the textbook), the sampling distribution of z (which is just a standardized version of our estimate $\bar{x}$) is equal to the standard normal distribution *if the null hypothesis is true*
- To decide how unlikely a given value of z (and the corresponding value of $\bar{x}$) is under the null hypothesis, we can see how far out in the tails z is in the standard normal distribution

How Stata calculates p-values

- Remember, we learned that 95% of the time, a value drawn from the standard normal distribution will fall between -2 and 2



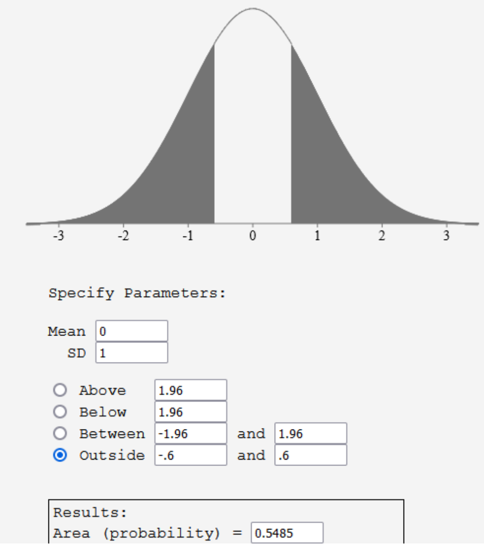
(Image: CC BY-SA 4.0, "Curva normal 2.jpg" by Matheus otoni)

How Stata calculates p-values

- We learned basic thresholds before (1, 2, 3), but Stata can convert any value of z to a p-value
- The p-value is the probability of obtaining a sample mean as far (or farther) from μ_0 as the one we obtained ($\bar{x}$), assuming that the null hypothesis is true
- In other words, the p-value answers the question "If μ_0 is the true population mean, how likely are we to get a sample mean as far from μ_0 as the one we got in our sample?"

How Stata calculates p-values

- Example: z=0.6 (two-tailed), p=.55



Other hypothesis tests

- There are lots of other hypothesis tests that can easily be conducted
- The key is to always know what the null hypothesis is
- For example, we might wish to compare means among more than two groups (more on this next week)
    - We could have a null hypothesis that all of the means are equal ($H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$)
    - Example: We want to know whether the number of ER visits varies by day of the week; the null hypothesis is that the mean number of visits is the same for all 7 days of the week

Review questions: key points from this lecture

- What are the steps of a hypothesis test?
- When do we reject versus fail to reject the null? (Or accept the alternative hypothesis versus conclude nothing?)
- How can you tell the difference between a null hypothesis and an alternative hypothesis?
- What is the typical alpha level?
- What's the difference between a 1-sided versus a 2-sided hypothesis test?
- What are type I and type II errors?
- When is something statistically significant?

Appendix 1: More about z-tests

Return to z-test example

- Remember, $\bar{x}$ is an imperfect estimate of μ
- We want to know whether we can reject the null hypothesis (conclude $\mu \neq \mu_0$) based on the value of our estimate $\bar{x}$ (as well as the precision of that estimate, measured as $\sigma_{\bar{x}}$)
- If $\bar{x}$ is a long way from μ_0 and our estimate is sufficiently precise, we'll conclude that $\mu \neq \mu_0$

Appendix 1: More about z-tests

- We can conduct a z-test (a type of hypothesis test) if we are willing to assume (1) there is a normal distribution and (2) some value we choose is the standard deviation of the distribution
    - This second assumption is often problematic because we rarely know the standard deviation of a population ahead of time (although with large samples, this becomes unimportant)
    - To avoid the second assumption, we often use a t-test instead, but we'll start with the z-test since it relies on a distribution we've already learned about (the normal distribution); see also Appendix 2
- The one-sample z-test is rarely used in practice, so it's not the most important test to be familiar with. But it's one of the simplest, so it provides a nice illustration of hypothesis tests and their assumptions.

Appendix 2: The central limit theorem

- So far, we've assumed that our underlying variable (x) is distributed according to a normal distribution, but violating this assumption usually doesn't create serious problems
- We usually turn to the normal distribution (or the t-distribution, which is similar) when it comes time to create a confidence interval or calculate a p-value because of the *central limit theorem*
- Central limit theorem: the sum (or average) of several independent random variables (with similar variances) will approach the shape of a normal distribution (even if the original variables don't have normal distributions) as the number of variables grows larger

Appendix 2: The central limit theorem

- If we have a sample of data for variable *x*, we can consider each value of *x* (in the sample) to be an independent random variable since each value depends on which observation is randomly selected into the sample
    - The values are independently drawn since the selection of one observation into the sample does not affect the probability of any other observation being selected into the sample (at least for an infinitely-large population)
    - The values have equal variance since they are all drawn from the same probability distribution (everyone in the population has an equal probability of being selected)

Appendix 2: The central limit theorem

- Therefore:
  - If we have a simple random sample,
  - AND if the sample size is reasonably large (say > 20),
  - the probability distribution for the sample mean of any variable x will look like a normal distribution (even if the distribution of x is not normal).