PHP 2510 Principles of Biostatistics & Data Analysis

Week 7: Hypothesis Testing

This Week's Plan

1

Exam #1

2

Hypothesis Testing 101

Paradigm

Core Components

3

Demonstrations

Tea Expertise

Applications to Data

4

Experimental Design

Core Components

Handout

OUTCOMES

After this week’s classes, along with the required readings (CHIHARA Chapter 3, 3.1-3.2; CHIHARA Chapter 8, 8.1-8.4.3), you should be able to:

- Define the core components and explain the logic of hypothesis testing
- Formulate appropriate null and alternative hypotheses for a given research question
- Evaluate the relationships between sample size, significance level, effect size, and power
- Carry out hypotheses tests in R and interpret results accurately

Inference Paradigm (last week)

1. Assume something about the true population
2. Assume something about your sampling/reporting process
3. Postulate a statistic; prove it is close to the population parameter you care about
4. Make claims about the population based on the value we actually observe in our dataset

Part 2

Hypothesis Testing Paradigm (this week)

1. Assume something is true about your population… null hypothesis
2. Create a (test) statistic
3. Determine how the test statistic behaves under (1) … null distribution
4. Determine if the realization of the test statistic in your dataset is unusual or not… p-value
    a. If rare: (1) was unreasonable… "statistically significant"
    b. If common: no evidence to disprove (1)… "can be explained by chance alone"

What are some things we can test?

1. Is the drug effective at all?
2. Is the drug as effective as the existing standard of care?
3. Is the drug as effective in subpopulation A as it is in subpopulation B?

What is the definition of “effective”?
What would be our null hypothesis for each?

ASIDE

The choose function

https://en.wikipedia.org/wiki/Binomial_coefficient

Part 3

The Tea Expert

- “Famous” experiment conducted by statistician Ronald Fisher in 1919
- A lady claims she can detect the order in which milk and tea have been added into a teacup (she preferred milk added into the cup first)
- Fisher presented the lady with 8 cups of tea
- He told her that there were 4 of each type before she had to guess
    - Why is this fact important?

Part 3

The Tea Expert

- What is the null hypothesis?
- What is the test statistic?
- What is the null distribution?
- What experimental outcomes allow us to claim expertise?

Note: Fisher's exact test is **non-parametric**, meaning we don't assume our data generation process was governed by a specific parameter; these are popular when datasets are small

we will first focus on parametric tests; please focus less on the exact calculations here, and more on the process / set-up / mental model

#1

Warm Up - Hypothesis Testing Framework

- What does it mean to “assume the null (hypothesis) is true”
- Why do we do such a thing?
- What do I mean by “data is evidence”?

Hypothesis Testing Paradigm (reminder)

1. Assume something is true about the world… null hypothesis
2. Create a test statistic
3. Determine how the test statistic behaves under (1) … null distribution
4. Determine if your realization of the test statistic in your dataset is unusual or not… p-value
    a. If rare: (1) was not reasonable… “statistically significant”
    b. If common: no evidence to disprove (1)… “can be explained by chance alone”

Part 3

The Chocolate Expert

- What is the null hypothesis?
- What is the test statistic?
- What is the null distribution?
- What experimental outcomes allow us to claim expertise?

#1

Recall from week 5 survey: Out of 26 respondents, 6 of you have “good” eyesight (never worn Rx glasses)

Do ivy-league intro-to-biostatistics students have worse eyesight than the general population, where 40% have never worn Rx glasses?

2

Part 3

Recall from week 5 survey: Out of 26 respondents, 6 of you have “good” eyesight (never worn Rx glasses)

Do ivy-league intro-to-biostatistics students have worse eyesight than the general population, where 40% have never worn Rx glasses?

- What is the null hypothesis? alternative?
- What is the test statistic?
- What is the distribution of the test statistic under the null?
- What is the p-value?
- What assumptions are we making?
- What would you change for a 2-sided test? (Draw a picture)

2

Hypothesis Testing Paradigm (reminder)

1. Assume something is true about the world… null hypothesis
2. Create a test statistic
3. Determine how the test statistic behaves under (1) … null distribution
4. Determine if your realization of the test statistic in your dataset is unusual or not… p-value
    a. If rare: (1) was not reasonable… "statistically significant"
    b. If common: no evidence to disprove (1)… "can be explained by chance alone"

Part 2

What are some common tests?

⭐ 1. **Group mean (note: a proportion is a mean too)**
⭐ 2. **Difference in group means**
3. Number of successful trials (exact test, Rx example)
4. Variance (Chi-sq test)
5. Ratio of variances (F-test)
6. Difference in counts (Chi-sq test)
7. …

SPEEGLE 5.5.4 Summary (pg. 143) - we are going to keep using these results!!

Part 3

The “standard” hypothesis test (t-test)

1. Population has some true mean μ and variance σ^2
   - We are interested in estimating μ
2. Take an iid sample
3. Per CLT (for large n), $\bar{X}$ follows a $N(\mu, \sigma^2/n)$
4. Because we don’t know σ^2, we have to estimate that as well

Step 4 introduces more uncertainty:

$(\bar{X}-\mu)/(\sigma/\text{sqrt}(n)) \sim N(0,1)$

$(\bar{X}-\mu)/(s/\text{sqrt}(n)) \sim$ t distribution w/ n-1 degrees of freedom

Part 3

The “standard” hypothesis test (t-test)

```
library(resampledata3)
data("IceCream")

t.test(IceCream$VanillaCalories - 200)
```

vs

Can we calculate the p-value “by hand” (without t.test function), for whether the true mean = 200?

The “standard” hypothesis test (t-test), 2 groups

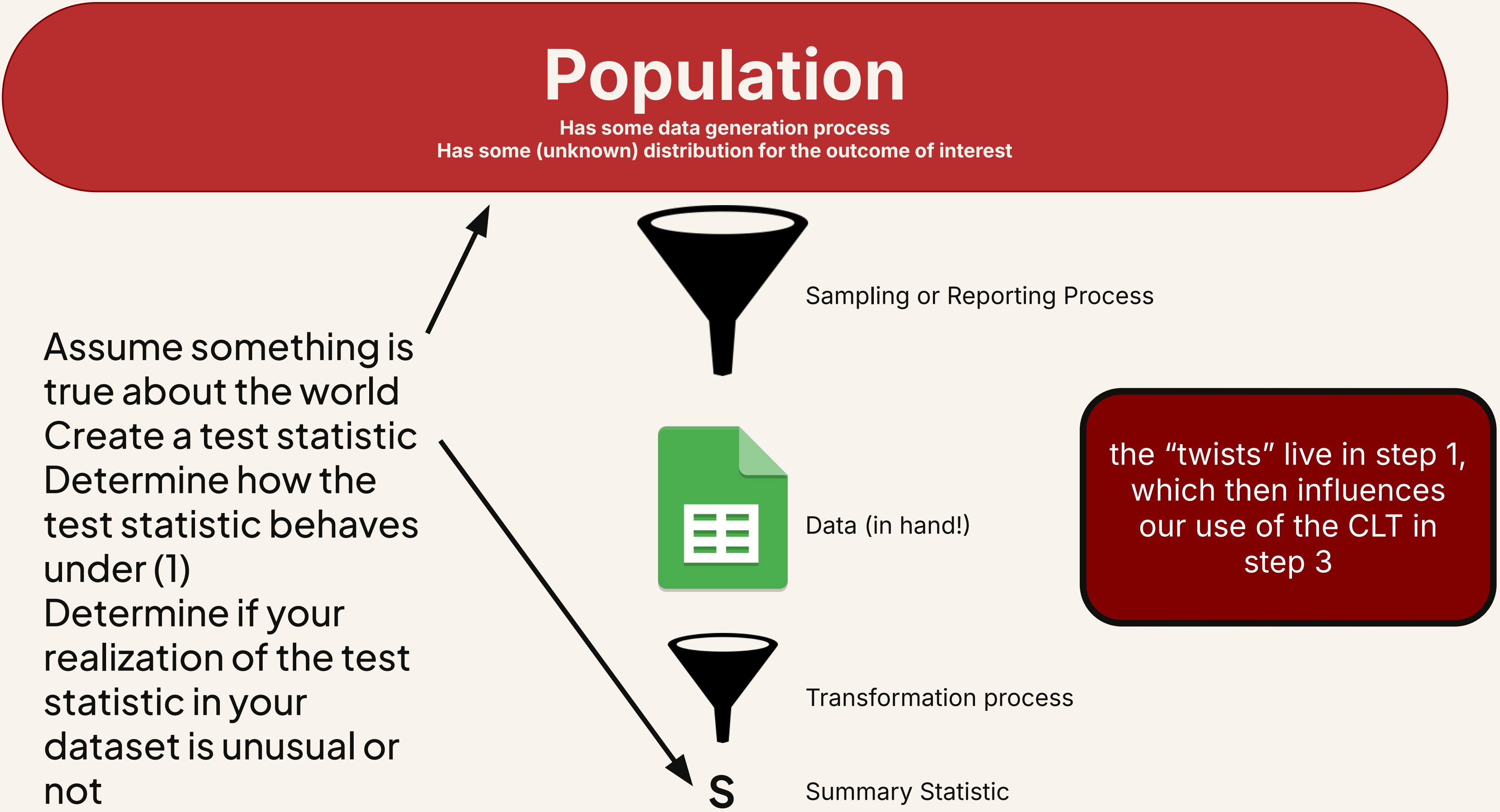
1. 2 independent populations; with true means μ_1 , μ_2 and variance σ_1^2, σ_2^2
   - We are interested in estimating $\mu_1 - \mu_2$
2. Take an iid sample
3. Per CLT (for large n), $\bar{X}_1$ follows a $N(\mu_1, \sigma_1^2/n_1)$, $\bar{X}_2$ follows a $N(\mu_2, \sigma_2^2/n_2)$
4. Because we don’t know σ_1^2, σ_2^2 we have to estimate them as well

Let’s figure out our test statistic:

But we might have a few “twists”

- Assume population distribution
  - Bernoulli (p, p(1–p) as our mean and variance)
  - Poisson (λ, λ as our population mean and variance)
  - ...
- Two groups
  - Paired (X and Y are not independent)
  - Equal population variances

How do we think about each of these “twists”?
Where do the book formulae actually come from?



Population

Has some data generation process
Has some (unknown) distribution for the outcome of interest

1. Assume something is true about the world
2. Create a test statistic
3. Determine how the test statistic behaves under (1)
4. Determine if your realization of the test statistic in your dataset is unusual or not

Sampling or Reporting Process

Data (in hand!)

Transformation process

S

Summary Statistic

the “twists” live in step 1, which then influences our use of the CLT in step 3

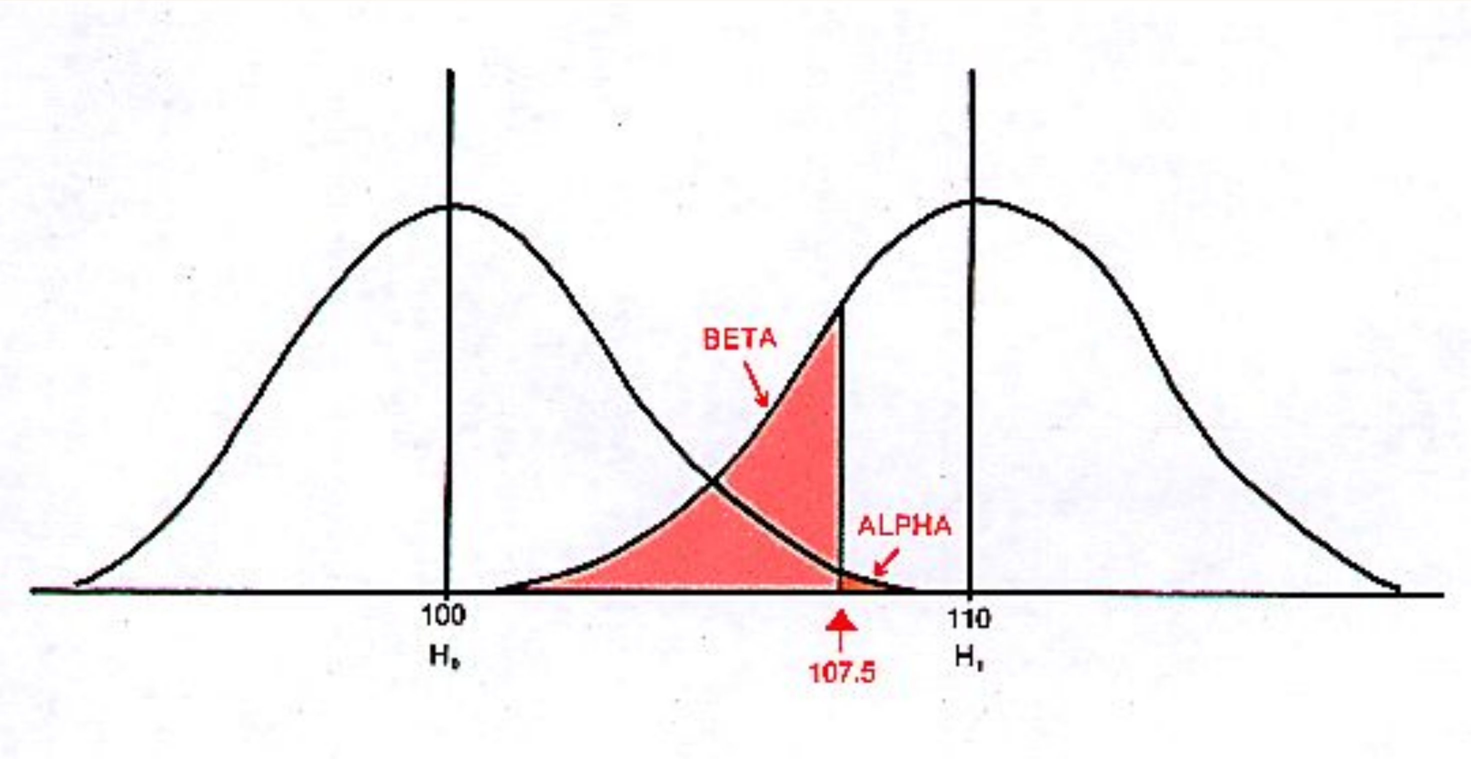
OUTCOMES

After this week’s classes, along with the required readings (CHIHARA Chapter 3, 3.1-3.2; CHIHARA Chapter 8, 8.1-8.4.3), you should be able to:

- Define the core components and explain the logic of hypothesis testing
- Formulate appropriate null and alternative hypotheses for a given research question
- Evaluate the relationships between sample size, significance level, effect size, and power
- Carry out hypotheses tests in R and interpret results accurately

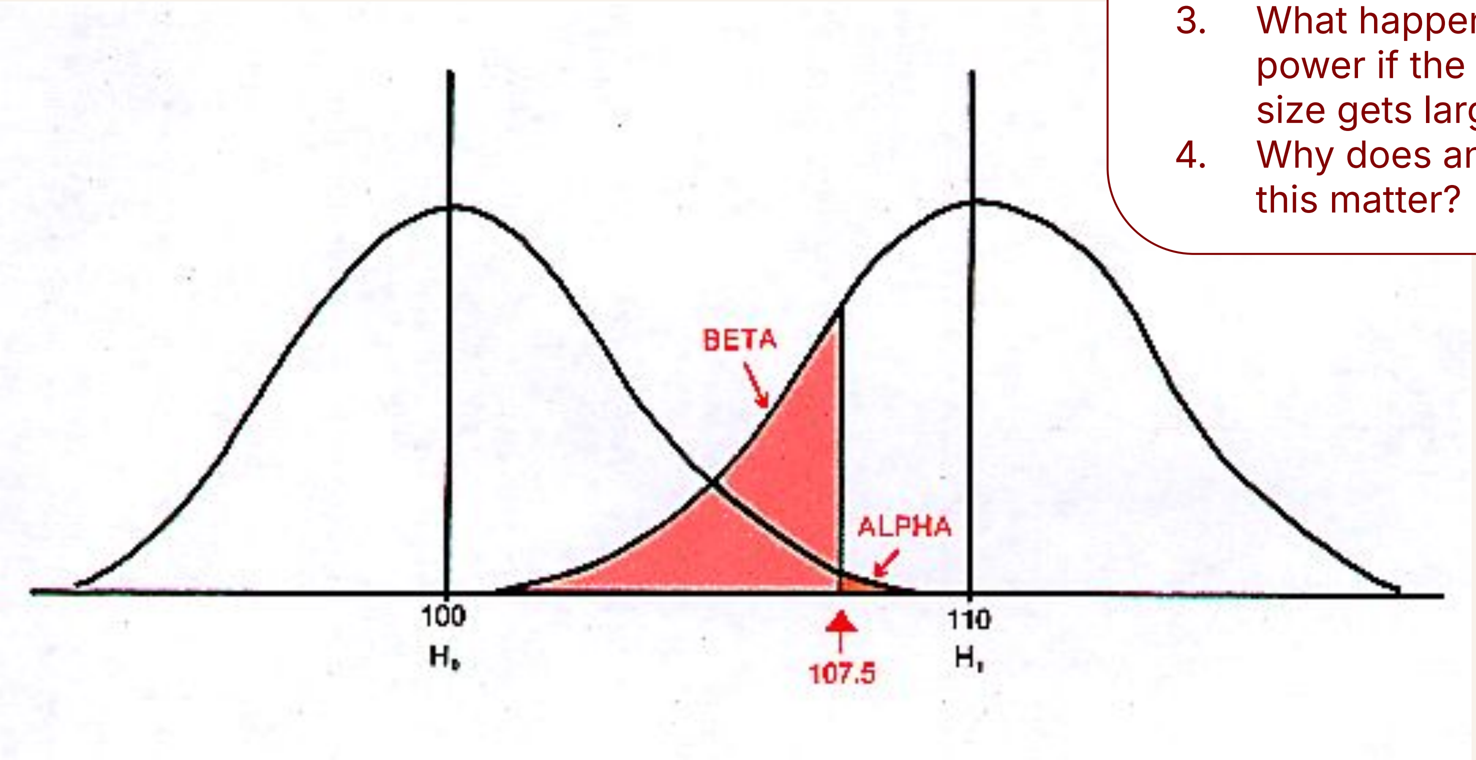
Part 4

Visualizing Hypothesis Testing



Part 4

Visualizing Hypothesis Testing



1. How do we visualize “power”?
2. What happens to these curves as n gets larger?
3. What happens to power if the effect size gets larger?
4. Why does any of this matter?

Part 4

Experimental Design Handout

- Let’s go over step 0 and step 1
- The rest is a take-home (coding) activity
  - Please submit to Canvas by 10/28 9am
  - Participation-only (ungraded for accuracy)
  - We will review answers next week (probably 10/28)

Part 4

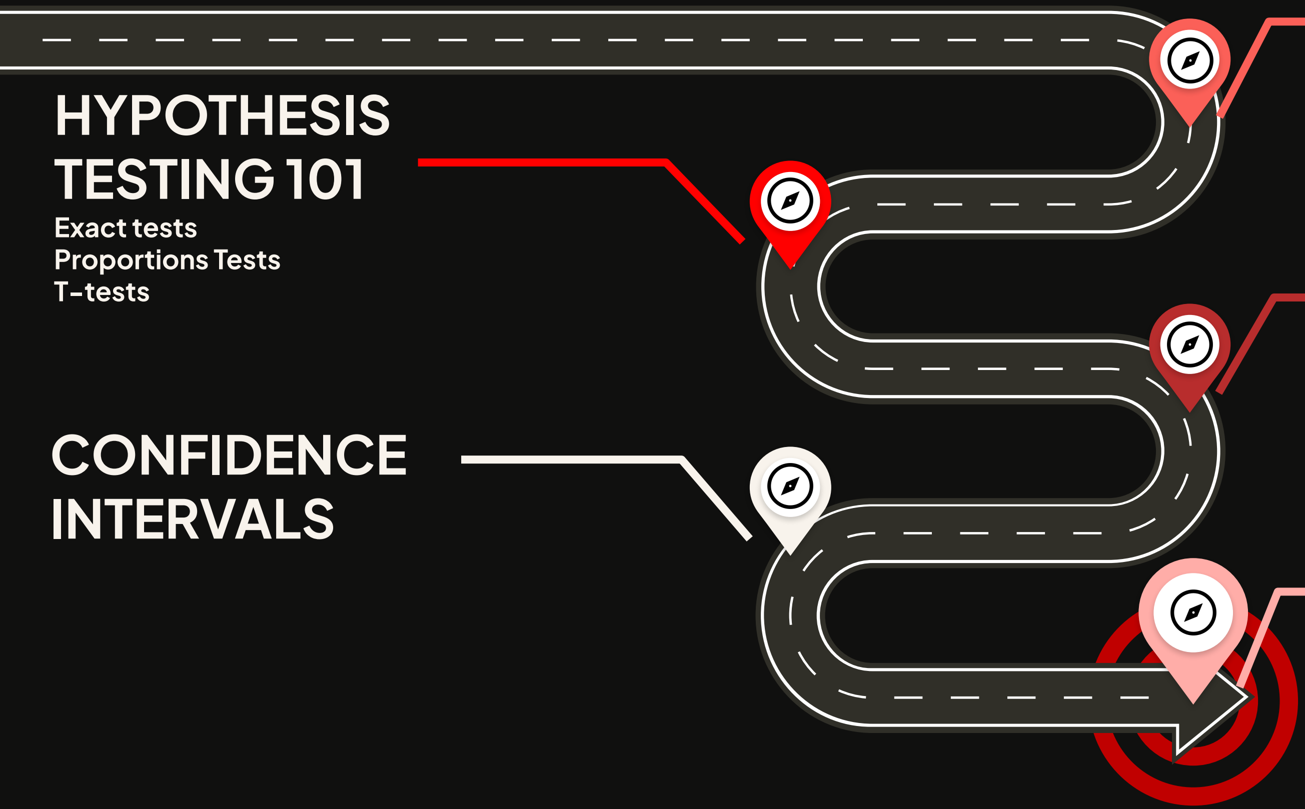
SPOILER ALERT: nothing is free

- To get more power*, you need more samples
- To detect a smaller effect size*, you need more samples
- To require a stricter significance level*, you need more samples
- Relationships are non-linear (e.g. 5% more samples does NOT get you 5% more power)

*holding all other parameters constant

ROADMAP

Free roadmap template



SAMPLING DISTRIBUTIONS & CLT

Exam Extra Credit

HYPOTHESIS TESTING 101

Exact tests
Proportions Tests
T-tests

EXPERIMENTAL DESIGN

Handout

CONFIDENCE INTERVALS

SYNTHESIS: THE INFERENTIAL PROCESS

Week 9 Lab
Assignment #2
Extra Practice Problems

OUTCOMES cont.

SPEEGLE Ch 8, 8.1–8.7 & 10.2 will be used to tie it all together (~1 week away)

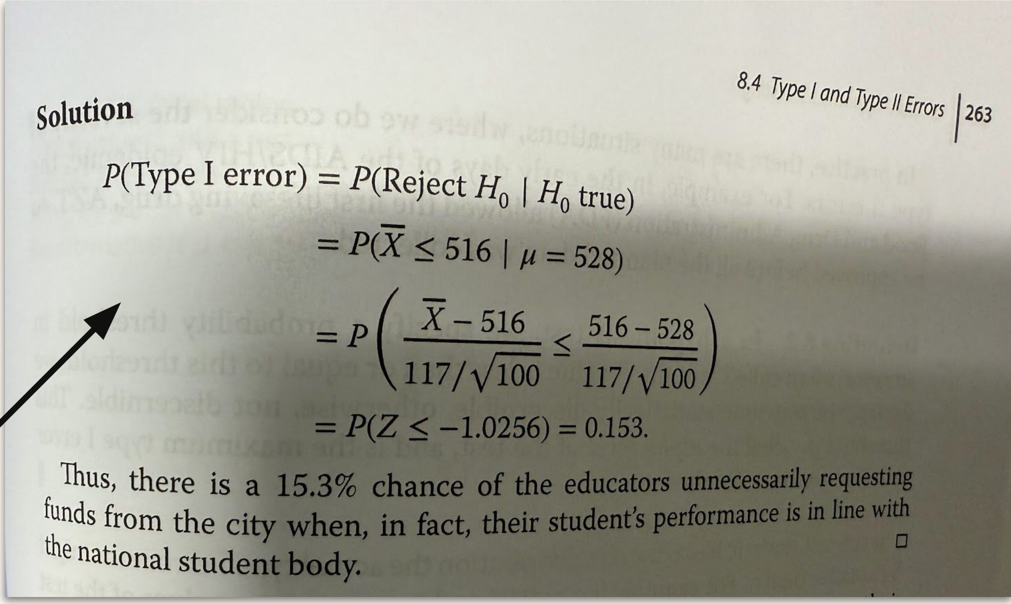
We will come back to these more advanced topics, but you should be familiar with them already

- Define the core components and explain the logic of hypothesis testing
- Formulate appropriate null and alternative hypotheses for a given research question
- Evaluate the relationships between sample size, significance level, effect size, and power
- Carry out hypotheses tests in R and interpret results accurately
- Identify the key assumptions underlying one- and two-sample t-tests
- Perform t-tests and interpret the output accurately
- Distinguish between designs that call for an independent t-test versus a paired t-test

Part 4

RECAP / LOOSE ENDS

- Hypothesis Testing Lab (questions; answers)
    - Prop.test vs t.test
    - “EXTRA” problem
- Extra Credit Exam Answer Key
    - Can you see how it is related to this Ed Discussion post?
- Did you find the CHIHARA 8.4.1 Type I Errors “typo”?
- Any other questions?



8.4 Type I and Type II Errors | 263

Solution

$$\begin{aligned} P(\text{Type I error}) &= P(\text{Reject } H_0 \mid H_0 \text{ true}) \\ &= P(\overline{X} \le 516 \mid \mu = 528) \\ &= P\left(\frac{\overline{X} - 516}{117/\sqrt{100}} \le \frac{516 - 528}{117/\sqrt{100}}\right) \\ &= P(Z \le -1.0256) = 0.153. \end{aligned}$$

Thus, there is a 15.3% chance of the educators unnecessarily requesting funds from the city when, in fact, their student’s performance is in line with the national student body. □

Part 4

PRACTICAL SUMMARY

- If I have 1s/0s, I use prop.test because that ensures my variance is the correct function of my underlying mean (remember, mean and variance can be related – they are for bernoulli, poisson, exponential, etc.). Under the hood, it is using the central limit theorem, applied to the bernoulli distribution. I can do this by hand if I wanted (I've learned enough!)
- If I don't want to assume a specifical random variable distribution for underlying population, I can use t.test. I can do inference for a single group mean or 2 groups means (their difference – paired or not). I am relying on the central limit theorem result, and doing a t-distribution look-up. I could do this by hand if I wanted (I've learned enough!)
- No matter what, I am asking the question: is my data aligned with a certain value (of the true mean) in the population? If my data is too strange for that value, I reject that value being true in the population.