BIOSTAT 702: Module 3

One Sample Inference; Part 1: Continuous Outcome

Dr. Marissa Ashner

Department of Biostatistics and Bioinformatics

Fall 2025

Duke University School of Medicine

Module Goals

- Understand the basics of null hypothesis statistical testing, as well as it's flaws
- Be able to run one-sample hypothesis tests for a continous outome and distinguish between when to use different tests

Resources for this Module

Textbooks

- ST21: Chapter 9
- ADLM: Chapter 2

Websites

- ASA Statement on p-values

Motivation

- A couple lectures ago, we talked about how research questions and hypotheses are part of the clinical research study life cycle
- In the last lecture, we talked about estimation of quantities from our data samples that we are interested in
- Now we will bring these ideas together

How do we test our research hypotheses using the quantities we have estimated from the collected data?

Null Hypothesis Statistical Testing (NHST)

- Set up statistical hypotheses
    - *null* and *alternative* hypotheses
    - mutually exclusive and exhaustive
    - We will primarily focus on two-sided tests
- Choose a *test statistic* and determine it's sampling distribution *under the null*
    - The magnitude of this statistic will help determine whether to reject the null hypothesis or not
    - Usually a function of the *estimates* we talked about in the previous lecture
- Calculate this statistic and corresponding *p-value*
    - Given the null is true, how likely is it that our realized sample (or something more "extreme") would be drawn from that distribution?
- Compare p-value to a pre-specified significance level to draw inference / interpret the results
    - We call the significance level α and it's usually equal to 0.05
        - more on that later…

One-Sample Test for a Population Mean μ

We think the average height of Duke students is 5’10’ ’ (70 inches). We know our sample mean will likely vary from the true average height, but we want to know if our sample mean is likely drawn from a distribution where the true average height is 5’10’ ’.

- *Statistical Hypotheses:* $H_0 : \mu = 70$, $H_A : \mu \neq 70$
- *Test Statistic:* $T = \frac{\bar{Y}-70}{s/\sqrt{n}} \sim t_{n-1}$.
  - Why did we choose this?
- ☆ *p-value:* $P(T \geq |t_{obs}||H_0 \text{ is true})$
- *interpret:*
  - If $p < \alpha$, reject the null, concluding that there is sufficient evidence to support the alternative
  - If $p \geq \alpha$, do not reject the null, concluding there is *insufficient* evidence to support the alternative

t {
样本量小
σ未知
正态分布

z: σ已知, n很大

Errors in Hypothesis Testing

confidence level $= 1 - \alpha$

- If the null hypothesis is actually true, but we *reject the null*, this is a **Type I Error**
  - This is equal to our significance level $\alpha = P(\text{reject } H_0 | H_0 \text{ true})$
- If the alternative hypothesis is actually true, but we *fail to reject the null*, this is a **Type II Error**
  - Known as $\beta = P(\text{fail to reject } H_0 | H_A \text{ true})$
  - $1 - \beta$ is known as the *power*, or the probability of correctly rejecting H_0 when H_A is true
- Want to balance these two errors, since as one increases, the other will decrease
  - Usually α is fixed and we choose a statistical test that minimizes the Type II Error (or maximizes the power)

Multiple Testing

5% reject Ho is wrong.

- Let's say we are running 10 hypothesis tests, each with a Type I error $\alpha = 0.05$.
- Assuming these tests are independent of one another, the probability of *at least 1* type I error occurring of all 10 would be $0.05 * 10 = 0.5$
  - This means we would likely see at least one false positive
- **Multiple Testing Corrections** are used to fix this problem
  - Will talk about this more later, but wanted to introduce the idea now

Inversion of a Hypothesis Test

- We talked about *Confidence Intervals (CIs)* in the last lecture, but they are directly related to hypothesis testing
- CIs are actually created by “inverting” the hypothesis test

Statistical vs Clinical/Practical Significance

CI 不包含 H_0. → 统计显著. H_0 reject

CI 包含 H_0 → 不显著 fail to reject

- These are *distinct* concepts
- **Statistical Significance:** Does the confidence interval contain values of the parameter under the null hypothesis?
- **Clinical Significance:** Does the confidence interval contain values of the parameter that are or are not substantively "large"?
    - "large" will depend on the context

Flaws of NHST / p-values

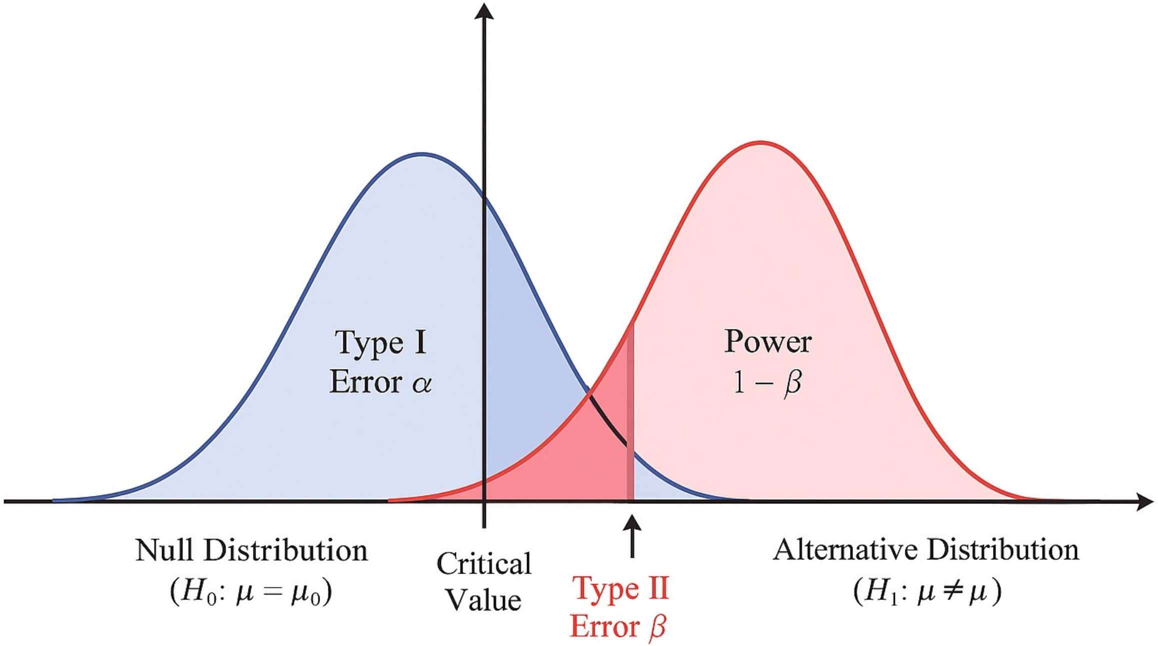
- p-values have in some ways, become a “gatekeeper” for whether work is publishable
  - Leads to “p-hacking”

ASA Statement on P-Values

- p-values can indicate incompatibility of data with the model used
- p-values do not measure the probability that the study hypothesis is true
- scientific conclusions should not be based only on p-value thresholds
- proper inference requires full reporting / transparency
- p-values do not measure effect size or result importance
- p-values do not provide a good measure of evidence

Non-Parametric Tests

- In the last lecture, we talked about how sampling distributions can be appromxiated using the t distribution when samples are large or small samples are approximately normal
  - What happens when we have small samples that are not normal-looking? Or largeish samples that are *really* not normal-looking?
  - We might consider applying *non-parametric* tests, which do not make *parametric distributional* assumptions on the data
  - Two common ones are the **sign test** and the **Wilcoxon signed rank test**
    - These test the *median*


Type I
Error α
Power
$1 - \beta$
Null Distribution
$(H_0\colon \mu = \mu_0)$
Critical
Value
Type II
Error β
Alternative Distribution
$(H_1\colon \mu \neq \mu\,)$

Sign Test

test median

- Center the data at the hypothesized median (e.g., 70 inches for our example)
  - Then discard all 0 values
- Count the number of positive values as the test statistic
- Under the null, the test statistic is distributed binomial with $\pi = 0.5$
- This test has low power (probability to correctly rejecting the null) because it doesn't take into account the magnitudes of differences from the hypothesized median, only the direction

Wilcoxon Signed Rank Test

[illegible] sign test

- This test is more powerful than the sign test, because it uses magnitudes of differences
  - However, this comes with the assumption that the data is symmetric about the hypothesized median
- Center the data at the hypothesized mean/median (e.g., 70 inches for our example)
  - Then discard all 0 values
- Transform the centered sample into ranks *ignoring the sign* (i.e., low to high)
  - If there are ties, average the ranks they would've uniquely received
- Sum the ranks associated with positive centered values (W^+) and those with negative centered values (W^-)
  - Choose the smallest as the test statistic
- The null distribution for W can be enumerated based on the fact that positive and negative values are equally likely under the null
  - In the event of ties, this distribution is approximated using a normal distribution

Q & A

Questions?