# Additional Topics in Hypothesis Testing

- One-Sided vs. Two-Sided Tests
- Error Types
- Choice of Level of Significance
- Cautions About Hypothesis Testing
- Importance of Results

University of Pittsburgh



## Review: Hypothesis Testing Terminology

- **Null hypothesis:** initial assumption made about a parameter
- **Alternative hypothesis:** conclusion that is made if the evidence suggests the null hypothesis is false
- **Test statistic:** a criterion used to decide if a hypothesized value is a plausible value for the parameter
- **P-value:** measure of how unusual the test statistic is
- **Level of significance:** the cutoff that signifies that a p-value is small enough to reject the null hypothesis
  - Denoted by $\alpha$ ("alpha") and usually set to either 0.01, 0.05, or 0.10



## Review: Hypothesis Testing Procedure

- We will cover 8 hypothesis tests this semester. Each is customized to the variable situation, but all follow the same procedure:

| Step | Description |
|---|---|
| | Determine the appropriate test to use |
| 1 | Determine the null and alternative hypotheses |
| | Collect the data |
| 2 | Ensure the conditions for performing the test hold |
| 3 | Calculate the test statistic |
| 4 | Calculate the p-value |
| 5 | Calculate a confidence interval that matches the level of significance |
| 6 | Write a conclusion using the above results |

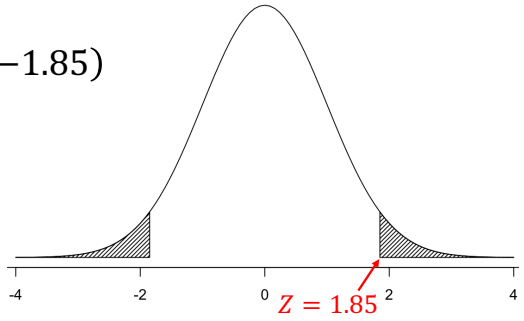
# Example: One-Sided vs. Two-Sided Test

- **Scenario:** You suspect a friend is using unfair dice in a game of Monopoly, so you keep track of his next 40 die rolls and observe a sample mean of 4.00. A fair die has $\mu = 3.50$ and $\sigma = 1.708$.
- **Question:** What hypotheses should we test if we simply suspect the dice are unfair?
- **Answer:** $H_0$: _____ vs. $H_A$: _____
- **Question:** Rolling higher numbers moves you around the board faster in Monopoly. What hypotheses should be tested knowing that it is often advantageous to have higher die rolls?
- **Answer:** $H_0$: _____ vs. $H_A$: _____
  - Knowing that higher die rolls are _____, we have a suspicion that the dice are deliberately weighted towards _____ die rolls.



# Example: One-Sided vs. Two-Sided Test

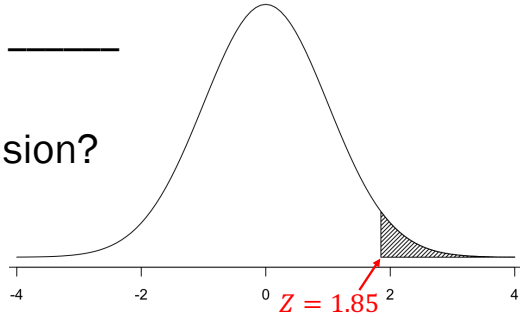
- **Scenario:** You observe a sample mean of 4.00 from the next 40 die rolls. A fair die has $\mu = 3.50$ and $\sigma = 1.708$.
- **Question:** Using the two-sided test ($H_0: \mu = 3.50$ vs. $H_A: \mu \neq 3.50$), what would the test reveal using a 5% level of significance?
- **Answer:** Mean die roll _____ from 3.50
  - **Test Statistic:** $Z = \frac{4.00 - 3.50}{1.708/\sqrt{40}} = 1.85$
  - **P-Value:** $p = P(|Z| > 1.85) = 2 \times P(Z < -1.85)$
    $= 2(0.0322)$
    $=$ _____
  - **Decision:** _____






# Example: One-Sided vs. Two-Sided Test

- **Statistics:** $\bar{X} = 4.00$, $\mu_0 = 3.50$, and $\sigma = 1.708$
- **Question:** Without redoing the test, what would the p-value be if using the upper one-sided test ($H_0: \mu = 3.50$ vs. $H_A: \mu > 3.50$)?
- **Answer:** _____
  - Test statistic is _____ – no _____ in the calculation changed
  - Only need the area in _____
  - P-value of a one-sided test will always be _____ the p-value of the _____
- **Question:** Does this change the conclusion?
- **Answer:** _____
  - _____and conclude that the mean die roll is _____



# Comparing One-Sided and Two-Sided Tests

- When deciding on the form of the alternative, consider the following:
  - Form of the alternative does not impact the _____
    - Test statistic measures _____ the sample mean is from the hypothesized mean
  - Form of the alternative impacts the _____
    - Two-sided alternative will always be _____ the value of the corresponding one-sided tests
  - Alternative hypothesis should be chosen based on what the researcher suspects is true _____ the data
    - Cannot collect the data, find the sample mean, and then choose the alternative
  - If a one-sided test is _____, hedge with a _____ test
    - More _____, but does not bring the risk of choosing the "_____" direction



# Error Types

- **Type I error:** occurs when the null hypothesis is true in reality, but the evidence tells us to reject $H_0$
  - **Common Reason:** A bad sample resulted in a small p-value
- **Type II error:** occurs when the null hypothesis is false in reality, but the evidence tells us to fail to reject $H_0$
  - **Common Reason:** Hypothesized value was wrong, but the sample did not pick up enough evidence to reject that value

| | | Reality of the Situation | |
|---|---|---|---|
| | | **$H_0$ True** | **$H_0$ False** |
| Your Decision | **Reject $H_0$** | Type I error | Correct decision |
| | **Fail to reject $H_0$** | Correct decision | Type II error |



# Example: Identifying Error Types

- **Scenario:** You are buying a two-person kayak for you and a friend to use on the ocean. The two of you together weigh 400 pounds, which is the claimed weight capacity of the kayak. You want to test the hypotheses $H_0: \mu = 400$ vs. $H_A: \mu < 400$ by randomly sampling a set of reviews from previous buyers.
- **Result:** Many people lied about their weight in their review. You find that the average weight the kayak can hold is less than 400 pounds when in fact the true weight capacity is 400 pounds.
- **Question:** What type of error was made?
- **Answer:** _____
  - **Decision:** _____
  - **Reality:** _____

## Example: Identifying Error Types

- **Scenario:** You are buying a two-person kayak for you and a friend to use on the ocean. The two of you together weigh 400 pounds, which is the claimed weight capacity of the kayak. You want to test the hypotheses $H_0: \mu = 400$ vs. $H_A: \mu < 400$ by randomly sampling a set of reviews from previous buyers.
- **Result:** Previous buyers were honest about their weights. You find that 400 is a plausible value for the weight based on the sample. After testing it out, you find that it cannot hold both of you.
- **Question:** What type of error was made?
- **Answer:** _____
  - **Decision:** _____
  - **Reality:** _____



## Example: Ramifications of Errors

- **Scenario:** You are buying a two-person kayak and want to test the weight capacity using the hypotheses $H_0: \mu = 400$ vs. $H_A: \mu < 400$.
- **Question:** What are the ramifications of each error type?
- **Answer:**
  - **Type I Error:** You _____ buying a kayak that would have _____ and instead look for a _____.
  - **Type II Error:** You spend money on a kayak that _____. Even if you can _____, you and your friend will _____.
- **Question:** Which type of error is worse?
- **Answer:** _____



## Example: Describing Error Types

- **Scenario:** Rubber pellets are often used on synthetic turf football fields. However, these pellets often contain lead. The federal limit for lead levels in these areas is 400 $\mu$g/g. Suppose a new football field is built. We want to test the hypotheses $H_0: \mu = 400$ vs. $H_A: \mu < 400$ by randomly sampling pellets from the field.
- **Question:** How could a Type I and Type II error occur?
- **Answer:**
  - **Type I Error:** Conclude that the field is _____ (i.e. _____) when in reality the field is _____ (i.e. _____)
  - **Type II Error:** Conclude that the field is _____ (i.e. _____) when in reality the field is _____ (i.e. _____)

# Example: Describing Error Types

- **Scenario:** Rubber pellets are often used on synthetic turf football fields. However, these pellets often contain lead. The federal limit for lead levels in these areas is 400 $\mu$g/g. Suppose a new football field is built. We want to test the hypotheses $H_0: \mu = 400$ vs. $H_A: \mu < 400$ by randomly sampling pellets from the field.
- **Question:** What are the ramifications of each error type?
- **Answer:**
  - **Type I Error:** Many athletes get _____.
  - **Type II Error:** The field is _____ and _____.

- **Question:** Which error type is more serious?
- **Answer:** _____



# Error Types and Level of Significance

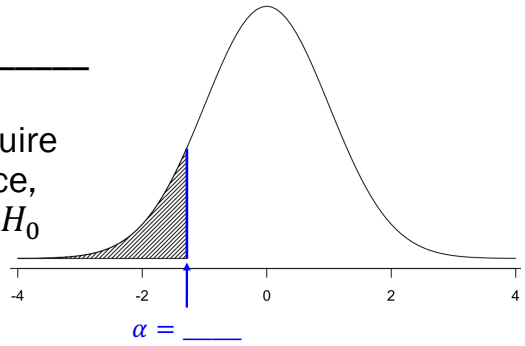
- The level of significance and the errors types are related.
  - 5% is a commonly used level of significance that will suffice in a majority of situations
  - Your choice of a level of significance may also be dictated by if a Type I or Type II error is worse.

- If a Type I error is worse, we should _____ the level of significance.
- If a Type II error is worse, we should _____ the level of significance.
- If neither error type is clearly worse, then stick with ____.



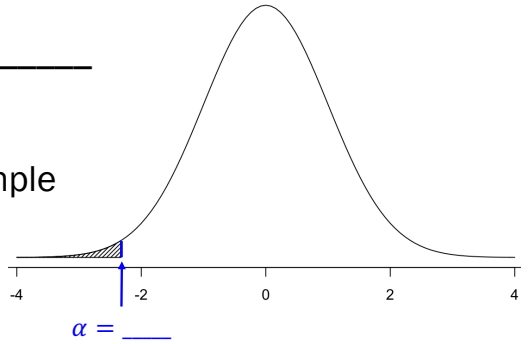
# Example: Appropriate Level of Significance

- **Scenario:** You are buying a two-person kayak and want to test the weight capacity using the hypotheses $H_0: \mu = 400$ vs. $H_A: \mu < 400$.
- **Question:** What level of significance would be appropriate to use?
- **Answer:** _____
  - _____ error is more serious – want to be confident that _____ _____
  - Type II error can only occur when _____
    - Fail to reject a false null hypothesis
  - _____ level of significance doesn't require quite as much _____ for significance, giving us a better chance of _____ $H_0$
    - Makes Type II error _____



# Example: Appropriate Level of Significance

- **Scenario:** You need to test the lead levels of a new football field for safety using the hypotheses $H_0: \mu = 400$ vs. $H_A: \mu < 400$.
- **Question:** What level of significance would be appropriate to use?
- **Answer:** _____
  - _____ error is more serious – want to be confident that _____ _____
  - Type I error can only occur when $H_0$ is _____
    - Reject a true null hypothesis
  - Choosing a _____ level of significance forces us to find a _____ sample mean to reject $H_0$
    - Makes Type I error _____






# Cautions About Hypothesis Testing

- Researchers have a decent amount of control over how significant their results are as they can often choose:
  - The population being studied
  - Form of the alternative hypothesis → Discussed earlier
  - Value of the hypothesized mean
  - Sample size
- It is also possible to run multiple analysis on different samples of the data or using different hypotheses, but only report the desired results.
  - This is called **p-hacking** and is extremely unethical



# Example: Choice of Population

- **Scenario:** A market researcher is doing a study on the largest amount that females would pay for a pair of jeans and believes the mean is greater than \$100. She surveys 30 random females who are shopping at a large, suburban mall. Assume $\sigma = \$25$.
  - **Results:** $\bar{X} = 116, Z = 3.51, p = 0.0002$, C.I.: $(107.05, 124.95)$
  - **Conclusion:** Researcher writes an article claiming the average amount a female would be willing to pay for jeans exceeds \$100.
- **Question:** What is wrong with the conclusion here?
- **Answer:**
  - Population that was studied was _____ – not all females _____
    - Could be able to generalize it to all females who shop at _____
  - Sample is not _____ of the _____

# Example: Choice of Hypothesized Mean

- **Scenario:** An SAT prep course claims to improve scores. The mean improvement between attempts for 40 students was 85 points. Assume the standard deviation of improvements is 80 points. The prep course needs to decide how to present the results.
- **Results:**

| Claim | Hypotheses | $Z$ | P-Value |
|---|---|---|---|
| 0+ Pt. Improvement | $H_0: \mu = 0$ vs. $H_A: \mu > 0$ | 6.72 | $9.08 \times 10^{-12}$ |
| 50+ Pt. Improvement | $H_0: \mu = 50$ vs. $H_A: \mu > 50$ | 2.77 | 0.0028 |

- **Question:** Which result makes a more compelling case for the effectiveness of the course?
- **Answer:** Improving scores by _____
  - Test was still _____ – more convinced the course actually _____ as opposed to scores increasing _____



# Example: Choice of Sample Size

- **Scenario:** Two analyses are performed using $\alpha = 0.05$:
  - Test if mean fastball speed in baseball exceeds 96 mph
  - Test if mean PPM of a chemical released from a car engine is less than 8
- **Results:**

| Analysis | Hypotheses | $\bar{x}$ | $n$ | $Z$ | P-Value |
|---|---|---|---|---|---|
| Fastball | $H_0: \mu = 96$ vs. $H_A: \mu > 96$ | 96.06 | 7059 | 1.68 | 0.0465 |
| PPM | $H_0: \mu = 8$ vs. $H_A: \mu < 8$ | 6 | 5 | 1.63 | 0.0515 |

- **Question:** Which result is more meaningful?
- **Answer:** _____
  - Even though it failed to achieve _____, it came very close despite only having a _____
  - Fastball speed analysis required over _____ to barely attain significance



# Importance of Results

- **Scenario:** Study of fastballs in baseball found an average speed of 96.06 mph. This result yielded a p-value of 0.0465 when testing $H_0: \mu = 96$ vs. $H_A: \mu > 96$ from a random sample of 7059 pitches.
- **Question:** Why does the p-value not indicate whether a result is important?
- **Answer:** We can _____ by simply taking a _____ enough sample
  - If we _____ the standard error $\left(\frac{\sigma}{\sqrt{n}}\right)$ by _____ $n$, then the test statistic $\left(Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)$ will _____
  - Smaller p-values result from _____ test statistics
  - Is the difference between 96 mph and 96.06 mph _____?

## Importance of Results

- **Scenario:** Study of a new car engine found that it releases only 6 ppm of a certain chemical into the air. This result yielded a p-value of 0.0515 when testing $H_0: \mu = 8$ vs. $H_A: \mu < 8$ from a random sample of 5 engines.
- **Question:** Why might a study with a non-significant result still be important?
- **Answer:** Studies with small sample sizes have _____ achieving significance, but are often used as _____ to show _____ _____
  - Almost attained significance with only 5 engines
  - Could convince an investor to _____ that could afford a _____ sample size



## Danger of Running Multiple Tests

- **Scenario:** A hospital studied the relationship between cell phone usage and 20 types of brain cancers (called gliomas). There was no relationship found between cell phones and 19 of the gliomas, but one glioma was found to be related to cell phone usage at the 5% level of significance. Only this one result was reported.
- **Question:** Why is this unethical?
- **Answer:**
  - _____ the 19 gliomas that were _____ – reporting only one of the results can cause _____ in the general population
  - Would expect ___ of the 20 tests to be significant _____
    - $\alpha = 0.05$ means extreme results occur 5% of the time even when $H_0$ is _____
    - Likely a _____
  - Relevant xkcd comic depicting a humorous play on p-hacking