Module : Hypothesis Testing

California State University - Northridge

Ovande Furtado, Jr., Ph.D.
March 11, 2022

```
xaringanExtra::use_webcam()
```

Example

Research question: Do Kin grad student at CSUN perform differently from the entire population when comes to reaction time?

Information to consider:

- Dependent variable: reaction time scores[1]
  - continuous (ratio)
- Independent variable: none

Data set

We will use the data set provided here.

[1] online test - Human Benchmark website

Short Review

- Inferential statistics: the process of estimating population parameters based on sample statistics
  - We rarely have access to the entire population, and even if we did, it would likely be impractical to test them all
  - We take a representative sample and estimate the parameter of interest based on the sample statistic.
- Parameter—a characteristic of a population
- Statistic—an estimate of a parameter based on sample value

Short Review, cont.

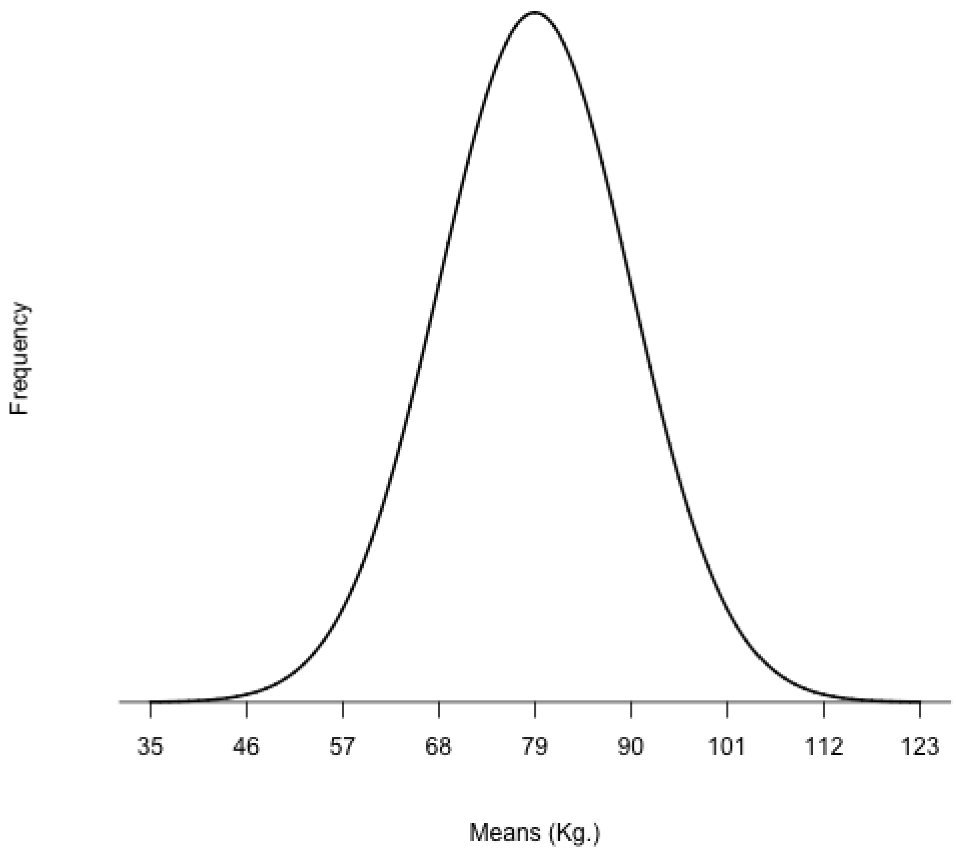
- Sampling error—the amount of error in the estimate of a population parameter that is based on a sample statistic
- Standard error of the mean (SEM)
  - It is an estimate of the amount of error when a sample mean is used to estimate the population mean.
  - The population mean cannot truly be known.
  - We do know the sample mean and an estimate of the error we should expect.

Short Review, cont.

Assume the following:

- The population = 175 lb (79 kg); SD = 25 lb (11 kg).
  - ~ 68% of all scores between 150 and 200 lb (68 and 91 kg)
  - ~ 95% of all scores between 125 and 225 lb (57 and 102 kg)
- Imagine you have access to all scores in the population.
- At random, pick 50 scores and calculate the mean.
- Return those 50 scores to the pot and repeat multiple times (i.e., 8000).
- Create a frequency distribution of all those sample means (sampling distribution of the mean).
- Distribution would be normal (even if the samples are not normal) due to the central limit theorem.
- The mean of the sample means is equal to the population mean.
- The standard deviation of the sample means is called the standard error of the mean.

Short Review, cont. SE_M



Again, assume that:

- Population mean (μ) = 175 lb (79 kg)
- Population SD (σ) = 25 lb (11 kg)
- If we take 8000 sample means
    - Mean of these sample means will be about 175 lb (79 kg)
    - SD of sample means = 1.6 kg which is the standard error of mean

Short Review, cont. SE_M

- We cannot sample the population multiple times, only once.
- The solution is to estimate the SE_M, and

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

- Use a Confidence Interval (CI) combined with a certain Level of Confidence (LOC)
  - CI: the range of values associated with a level of confidence
  - LOC: % that establishes the probability that a statement is correct

Statistical Hypothesis Testing

Statistical hypothesis testing—create two mutually exclusive and exhaustive mathematical statements about the outcome of the analysis

- Statistical hypotheses:
  - H_0 - the null hypothesis
  - H_a — the alternate hypothesis
  - Mutually exclusive — only one of the two can be true
  - Exhaustive — no other option can exist

We hypothesize that males and females differ on reaction time scores.

When H_a is a non-directional hypothesis

$$H_0 : \mu_m = \mu_f \text{ vs. } H_1 : \bar{x}_m \neq \bar{x}_f$$

Whem H_1 is a directional hypothesis

$$H_0 : \mu_m \leq \mu_f \text{ vs. } H_1 : \bar{x}_m > \bar{x}_f$$

Statistical Hypothesis Testing, cont.

- The interest is typically on H_a, but the H_0 is the one being tested; ***always!!!***
    - We either **reject** or **fail to reject**[1] H_0
- Recall the idea of the sampling distribution of the mean. Now extend this idea to mean differences.
    - Assume we have the reaction time scores of all males and females. Also assume that the H_0 is TRUE.
    - Take ONE sample of males and females and calculate the mean difference
    - plot the value in a distribution (sample mean differences)
    - now continue doing this for until you collected 10.000 sample mean difference scores.

[1] We never "**accept** H_0", since the actual state of the H_O is never known.

Statistical Hypothesis Testing, cont.

- If H_0 is true, then the mean difference ought to be ~0, but due to sampling error, the mean difference is unlikely to be exactly 0.
- Repeat over and over to create a sampling distribution of mean differences.
- The **mean of the mean differences** should be zero (since H_0 is true).
- The standard deviation of the distribution of mean differences is called the **standard error of mean differences**.

WAIT!!! I do not have access to the entire population...

TRUE, see next slide....

Statistical Hypothesis Testing, cont.

- Indeed, we do not have access to the entire population
- You take one sample from each group, calculate the mean difference, **and estimate the probability (*p*) that you could have gotten a mean difference this big or bigger if H_0 is true.**

- *p* (data$\mid$ H_0) == probability of the data, given H_0.
- If the *p*-value is small, we get suspicious about the truth of the null hypothesis. If the *p*-value is small enough, we get so suspicious that we reject the null hypothesis and accept the alternate hypothesis.
- How small is small? We set the criteria for small by setting α. The typical α is 0.05. This is equivalent to a 95% level of confidence (95% LOC).

Type I and Type II Error

- Decision about H_0 based on probabilities; may be wrong
- $\alpha \Rightarrow$ probability of committing a type I error (set by researcher)
- $\beta \Rightarrow$ probability of committing a type II error
- Power $\Rightarrow$ the probability of rejecting H_0 when H_0 is false
    - Power is equal to 1 − β.
    - Power is increased by decreasing noise in the data and increasing the sample size.

	H_0 **True**	H_0 **false**
Fail to reject	Correct decision	Type II error (β)
Reject H_0	Type I error (α)	Correct decision

Still, a theoretical example

- Center - sampling distribution of $\bar{x}$ ($H_0 : \mu = 100$)
  - This is the value being tested
- Distributions on either side - distributions that would be **true** if H_0 is **false** (alternative hypothesis)[1]
- To test a hypothesis, we take a sample from the population and determine if it could have come from the hypothesized distribution with an acceptable level of significance (α)[2] value - shaded areas in each tail of the H_0
- If the sample mean marked as $\bar{x}_1$ is in the tail of the distribution of H_0, we conclude that the probability that it could have come from the H_0 distribution is less than alpha - ***we reject*** H_0

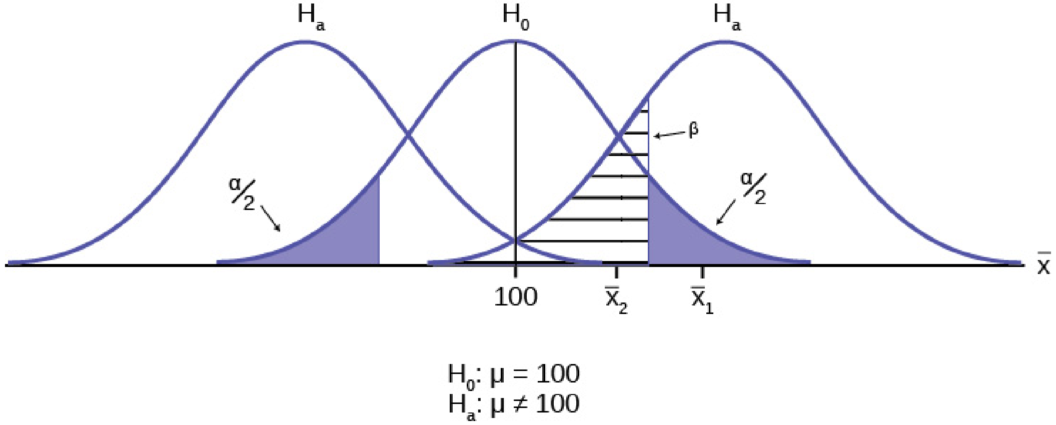


Figure representing a sampling distribution of $\bar{x}$ ($H_0 : \mu = 100$) and two alternative distributions that would be true if H_0 is false.[3]

[1]. We do not know which is true, and will never know. There are, in fact, an infinite number of distributions from which the data could have been drawn if Ha is true, but only two of them are on the illustration representing all of the others.

[2]. Each area is actually α/2 because the distribution is symmetrical and the alternative hypothesis allows for the possibility for the value to be either greater than or less than the hypothesized value--called a two-tailed test).

[3]. 9.2 Outcomes and the Type I and Type II Errors - Introductory Business Statistics | OpenStax. (2022, February 28). Retrieved from https://openstax.org/books/introductory-business-statistics/pages/9-2-outcomes-and-the-type-i-and-type-ii-errors

Still, a theoretical example

]

- The truth may be that this $\bar{x}_1$ did come from the H_0 distribution, but from out in the tail. If this is so, then we have falsely rejected a true null hypothesis and have made a Type I error.
- What statistics has done is provide an estimate about what we know, and what we control, and that is the probability of us being wrong, α.
- We can also see in the figure that the sample mean could be really from an H_a distribution, but within the boundary set by the alpha level.
- Such a case is marked as $\bar{x}_2$. There is a probability that $\bar{x}_2$ actually came from H_a but shows up in the range of H_0 between the two tails. This probability is the β error, the probability of **accepting a false null.**

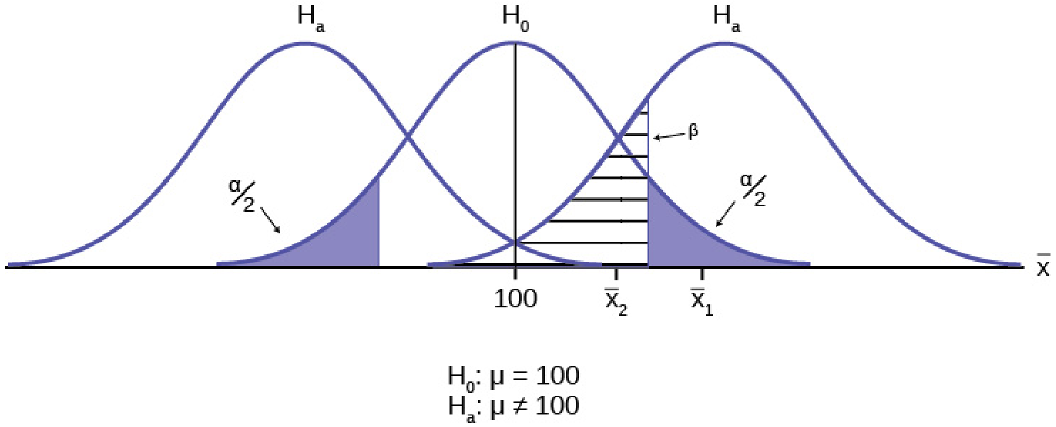


Figure representing a sampling distribution of $\bar{x}$ ($H_0 : \mu = 100$) and two alternative distributions that would be true if H_0 is false.[1]

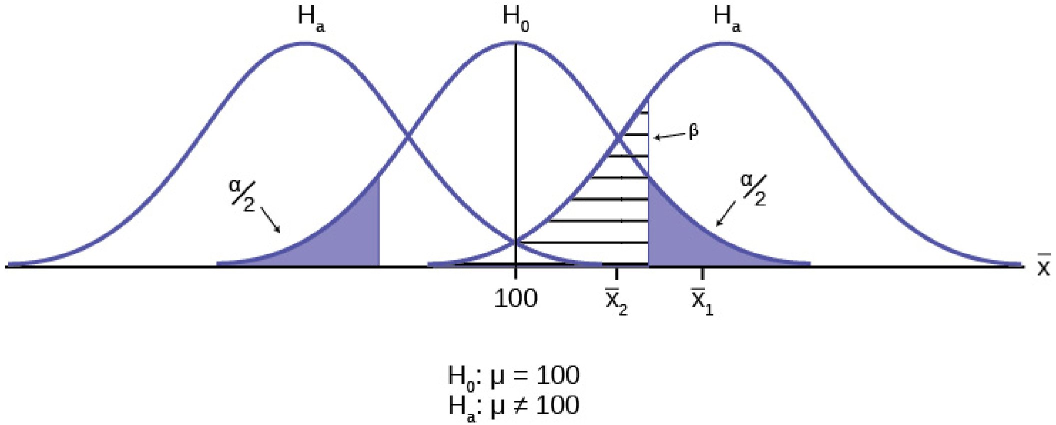
We will not reject a null hypothesis unless there is a greater than 90, or 95, or even 99 percent probability that the null is false.

[1]. 9.2 Outcomes and the Type I and Type II Errors - Introductory Business Statistics | OpenStax. (2022, February 28). Retrieved from https://openstax.org/books/introductory-business-statistics/pages/9-2-outcomes-and-the-type-i-and-type-ii-errors

Finally, an example!

Suppose the null hypothesis, H_0, is: Frank's rock climbing equipment is safe.[1]

- Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe.
- Type II error: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.
- Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks his rock climbing equipment is safe, he will go ahead and use it.)
- This is a situation described as "accepting a false null".



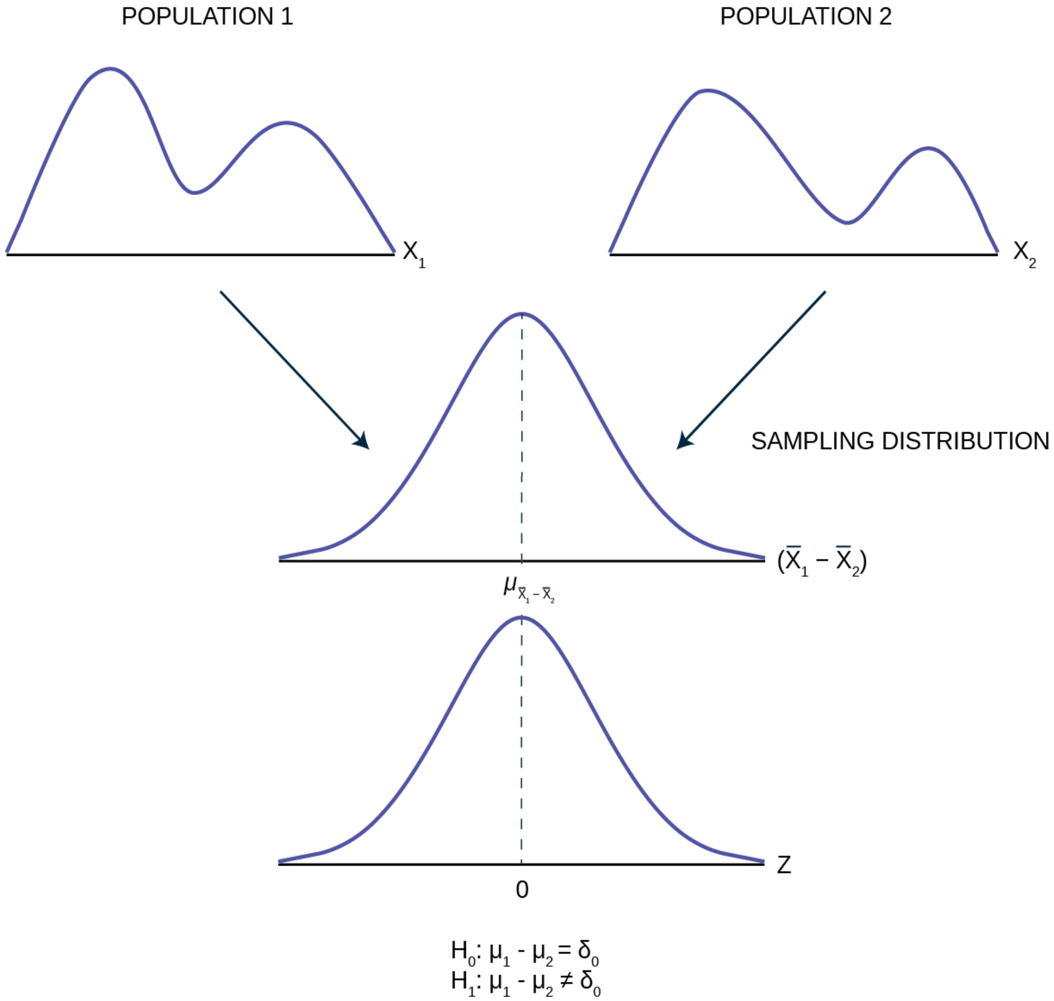
[1]. 9.2 Outcomes and the Type I and Type II Errors - Introductory Business Statistics | OpenStax. (2022, February 28). Retrieved from https://openstax.org/books/introductory-business-statistics/pages/9-2-outcomes-and-the-type-i-and-type-ii-errors

Overview

- Set α, say at .05 (95% LOC).
    - Accept a 5% risk of making a type I error, if H_0 is true
- Perform analyses and calculate *p*.
    - *p* $\Rightarrow$ probability of the getting data that you did, if H_0 is true
    - p (data $\mid$ H_0)
- If p ≤ α, reject H_0.
- If we reject H_0, **we say the result is statistically significant.**
    - Not necessarily important or meaningful[1]

[1] Not all significant tests are meaningful - We will address this soon.

Two ore more pulations


POPULATION 1
POPULATION 2
X_1
X_2
SAMPLING DISTRIBUTION
$(\bar{X}_1 - \bar{X}_2)$
$\mu_{\bar{X}_1 - \bar{X}_2}$
Z
0
$H_0: \mu_1 - \mu_2 = \delta_0$
$H_1: \mu_1 - \mu_2 \neq \delta_0$

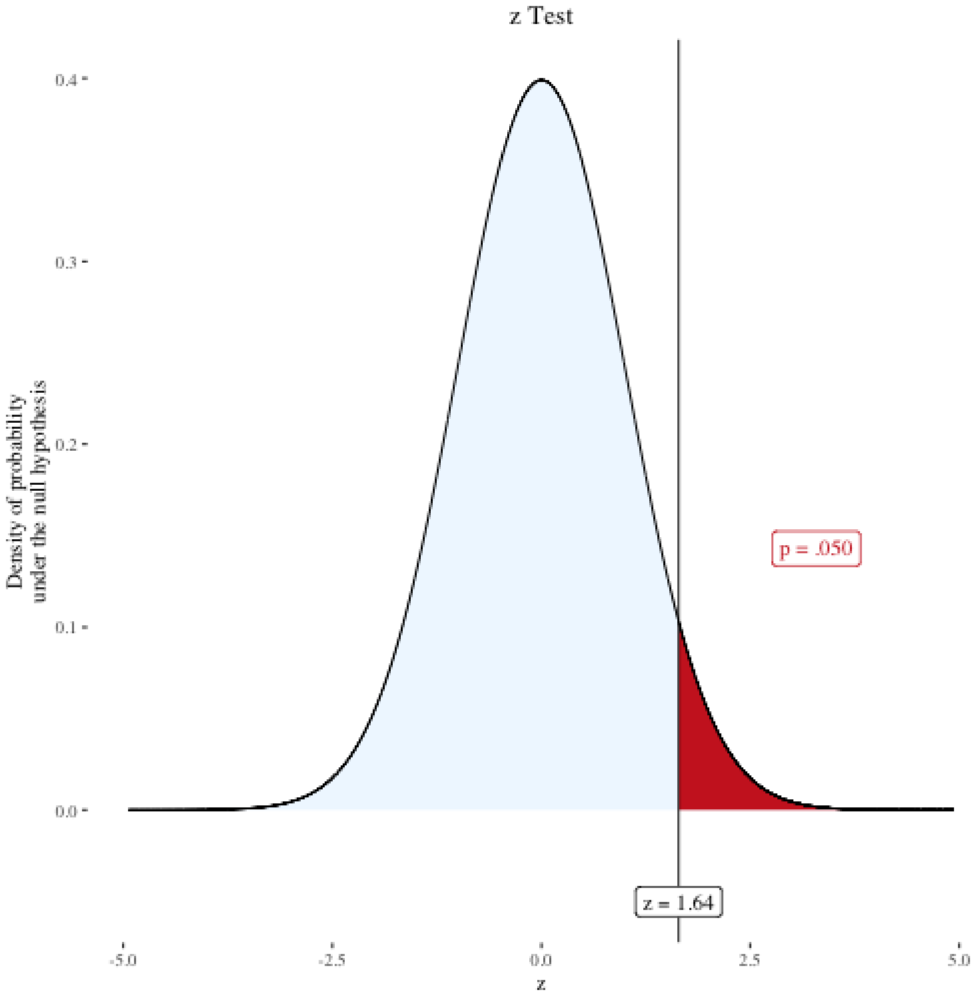
- Previous example was for ONE sample
- Do these two samples come from the same population distribution?
- The H_0 being tested now is mean differences

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs. } \mu_1 - \mu_2 \neq 0$$

One and Two-tailed tests

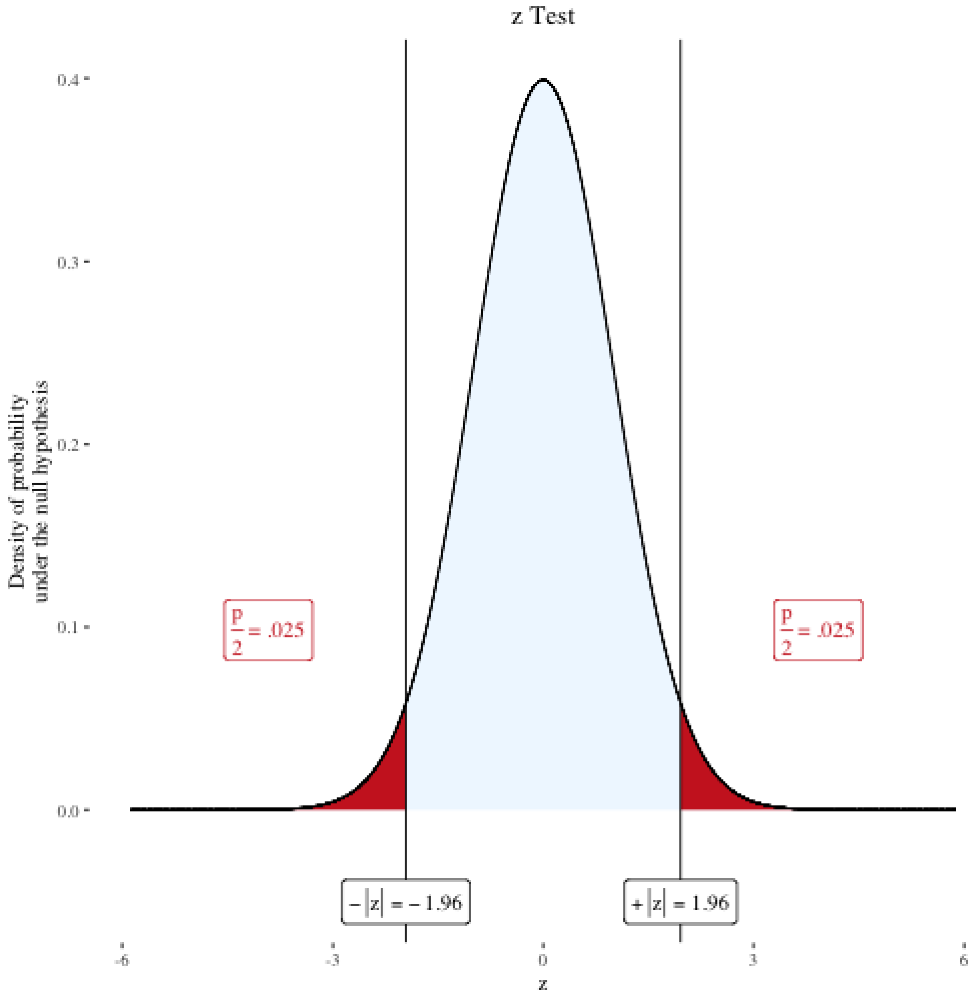
Two-tailed test: We state H_0 such that difference between means = 0. Or that the μ_1 is equal to the μ_2.

- $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \bar{x}_1 - \bar{x}_2 \neq 0$
- or $H_0 : \mu_1 = \mu_2$ vs. $H_a : \bar{x}_1 \neq \bar{x}_2$
- Half of rejection area divided between the two tails of the sampling distribution



One and Two-tailed tests

- **One-tailed test:** We hypothesize that one condition (or group mean) is larger than another.
    - $H_0 : \mu_1 \leq \mu_2$ vs. $H_a : \bar{x}_1 > \bar{x}_2$
    - All rejection area in one tail of the sampling distribution



Confidence Levels

The table below shows the uncorrected critical p-values and z-scores for different confidence levels.

z-score (Standard Deviations)	p-value (Probability)	Confidence level
< -1.65 or > +1.65	< 0.10	90%
< -1.96 or > +1.96	< 0.05	95%
< -2.58 or > +2.58	< 0.01	99%

Applying Confidence Intervals

- Confidence intervals are a**n alternative approach** to the null hypothesis statistical test.
- The approach is based on the same underlying statistical model as the null hypothesis statistical test, but instead of making a binary decision about the acceptability of H_0, the analyst simply calculates an interval around which it is estimated that the population value truly exists.

For 95% Confidence Interval:

95% CI = $\bar{x} \pm 1.96 * SE_M$

Example from normal curve

SEM = $\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$

n = 50, $\bar{x}$ = 35 cm (13.8 in), s = 10 cm (3.9 in)

SEM = $\sigma_{\bar{x}} = \frac{10}{\sqrt{50}}$ = 1.4 cm (0.6 in)

95% CI = 35 ±1.96 (1.4) = 32.3 to 37.7 cm (12.7 to 14.8 in)

Applying Confidence Intervals, cont.

Can also apply to mean differences

CI includes zero $\Rightarrow$ we fail to reject H_0

CI does not includes zero $\Rightarrow$ we reject H_0

For example, if BMI mean difference = 3 kg/m2, and SE of mean differences = 2 kg/m2.

A 95% CI = 3 ±1.96 (2) = −0.9 to 6.9 kg/m2.

Note the 95% CI includes zero so we do not reject the null.

Reporting significance

Usual notation	Signif. stars	English translation	The null is…
$p > 0.05$		The test wasn’t significant.	Retained
$p < 0.05$	*	The test was significant at α = 0.05; but not at α =.01 or α = 0.001.	Rejected
$p < 0.01$	**	The test was significant at α = 0.05 and α = 0.01; but not at α = 0.001.	Rejected
$p < 0.001$	*	The test was significant at all levels	Rejected

Running the hypothesis test in practice

Doing it with jamovi

- Click here to see an example from Navarro and Foxcroft, 2019
- iskdata
  - While in class, we test the following hypothesis

As a group, do Kinesiology CSUN students perform differently from the entire population[1]

[1] You should exercise caution when interpreting the results since you comparing a group of graduate students with the entire population.

Effect Size and Power

Cohen's d is a measure of "effect size" based on the differences between two means.

It measures the relative strength of the differences between the means of two populations based on sample data.

The calculated value of effect size is then compared to Cohen’s standards of small, medium, and large effect sizes.

Size of effect	d
Small	0.2
Medium	0.5
Large	0.8

Cohen's Standard Effect Sizes

Effect Size and Power

(1 − β) is called the Power of the Test.

The statistical power of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis when a specific alternative hypothesis is true.

Statistical power ranges from 0 to 1, and as the power of a test increases, the probability β of making a type II error by wrongly failing to reject the null hypothesis decreases.

Power analysis: can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size.

Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size.

There are several online tests, but G*Power is the best option:

- Download G*Power
- Examples using G*Power

Thanks!

CSUN® CALIFORNIA STATE UNIVERSITY NORTHRIDGE

ovandef@csun.edu

@ofurtado

@drfurtado