PS 211: Introduction to Experimental Design

Fall 2025 · Section C1

Lecture 9: Effect sizes, power, & single-sample t-tests

Updates and Reminders

- Homework 2 is due Friday.
  - We are going to try to return grades quickly so you can use feedback to prepare for Exam 2.
- Please submit as a PDF.
- If you would like to go over your exam, please come to office hours this week:
  - TODAY (10/7): 12:30 - 2:30 pm (Juneau, 111 Cummington Mall Room 242)
  - Wednesday (10/8): 10 am - 12 pm (Kate, 111 Cummington Mall Room 208)
- Coming up:
  - Thursday (10/9): Exam 2 Review
  - Tuesday (10/14): No class (Monday schedule)

Review: Z Scores & Z Tests

Z Scores are a form of *standardization*.

- They tell us where a score falls relative to other scores in a dataset.
- Useful for comparing scores *across* distributions.
- Can be computed for individual datapoints or individual means in a distribution of means.

Z Tests are a form of statistical hypothesis testing.

- We use *z* tests when we:
  - **know** information about a population distribution, including its standard deviation.
  - want to compare a **sample** mean to the **population** mean.

Review: Z Test Steps

Conducting a *Z* Test

1. Transform "raw" mean to *z* score using the formula for a distribution of means.

2. Use a *z* table or a computer program to determine the probability of obtaining this *z* score under the null hypothesis.

- *Given the population mean and standard error, what is the probability of obtaining a z score as extreme as the one calculated?*

3. Determine cutoffs or critical values based on your alpha level (usually the top or bottom 5% of the distribution for one-tailed tests, and the top *and* bottom 2.5% of the distribution for two-tailed tests).

4. Compare your probability level (*p* value) to your cuttoff values and determine if you should *reject* or *fail to reject* the null hypothesis.

Review: Confidence Intervals

- A common type of interval estimate is a confidence interval.
- A confidence interval is the range of values within which a population parameter is estimated to fall.
- A confidence interval is usually expressed in terms of a percentage, such as 95% or 99%. This percentage is called the confidence level.
- A 95% confidence interval means that if we were to take many samples and compute a 95% confidence interval for each sample, then approximately 95% of those intervals would contain the true population parameter.

Calculating confidence intervals with *z* distributions

You have:

- Sample mean (M)
- Sample size (n)
- Population standard deviation (SD)

Note: This is rare in practice! We will learn how to handle unknown population SDs later in the course.

You want:

- 95% Confidence interval (CI) around the sample mean.

Calculating confidence intervals with *z* distributions (Continued)

1. Assume your **sample** mean lies in the center of a normal distribution.
2. Determine the area under the normal curve that corresponds to your desired confidence level (e.g., 95%).

- For a 95% CI, this is .95. This leaves .05 in the tails, or .025 in each tail. From this, you can determine the bounds of the CI.
- For a 95% CI, the upper bound is .975. The lower bound is .025.

Calculating confidence intervals with *z* distributions (Continued)

3. Find the corresponding *z* scores for these bounds using a *z* table or computer programming.

- For the upper bound (.975), the *z* score is approximately 1.96.
- For the lower bound (.025), the *z* score is approximately -1.96.

4. Convert the z statistic to raw scores.

- Use the formula: $X = \mu + z \times SE$, where $SE = SD/\sqrt{n}$.

5. Now we have our 95% CI!

Interpretation: About 95% of CIs computed from repeated samples would include the true population mean.

Example: Screen Time of Teens vs. Adults

- Nielsen (2018): U.S. adults spend **11 hours/day** on screens ($\mu = 11, SD = 2$)
- We measure a sample of **20 teens** → $M = 9$

Do teens and adults differ in how much time they spend on screens?

- Compute the probability (*p* value) that the teen population distribution and the adult population distribution are the same.
- Compute a 95% CI for the *population mean* of teens.

For the *p* value (*z*-test)

1. Convert 9 to *z* score:

$$\frac{(9-11)}{2/\sqrt{(20)}} = -4.47$$

2. Use *z* table to convert to percentile / probability: $< .001$

For the 95% CI

1. Compute SE: $2/\sqrt{20} = 0.447$
2. Find $z_{crit} = 1.96$
3. Compute bounds: $9 \pm 1.96(0.447)$ → 8.12, 9.88

Confidence Intervals and Significance

- CI and hypothesis testing give consistent conclusions.
- If the **population mean** (e.g., 11 hours) **falls outside** the 95% CI, we reject the null hypothesis.
- If it **falls inside**, we fail to reject.

💡 **A CI provides more information** than a simple reject/fail-to-reject decision—it is an interval estimate that shows the range of plausible effects.

Effect Size

- Statistical significance depends on the sample size.
- The larger the sample size, the more likely it is that a statistically significant result will be found.

Based on what we know about *z* tests, why is this the case?

Answer: Larger sample sizes provide more "certain" estimates of the population parameters, reducing the standard error and increasing the likelihood of finding a statistically significant effect.

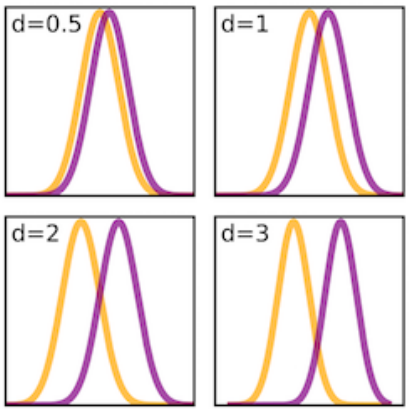
- The **effect size** tells us *how big* the effect actually is.
- It measures the magnitude of a difference in standardized units.

Example: Researchers examine standardized math test scores in over 500,000 students across the country. They find that, on average, girls score .3 points lower than boys. This effect is *statistically significant*. But is it meaningful?

What is effect size?

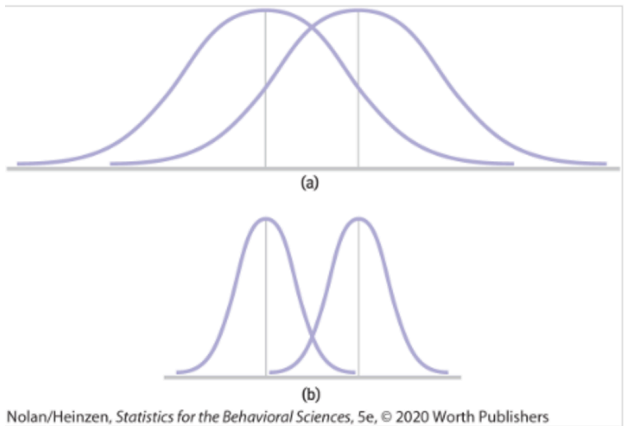
- Tells us the size of a difference that is *unaffected* by sample size.
- Allows standardization across studies
- Tells how much two populations do not overlap – the less overlap, the bigger the effect size
- Said in another way, the effect size is a quantitative measure of the magnitude of the experimental effect.
  - The larger the effect size, the stronger the relationship between two variables.

Effect size tells us how much two populations do not overlap



Overlap can be decreased in two ways:

1. When two population means are far apart, the overlap of the distributions is less and the effect size is bigger.



2. When variability within each distribution is smaller, overlap decreases and effect size increases.

Calculating effect size

- Cohen's d is a common statistical measure of effect size.
- We calculate Cohen's d by taking the difference between two means and dividing by the data's standard deviation.
- Cohen (1990) used the small but statistically significant correlation between height and IQ to explain the difference between statistical significance and practical importance.
  - His sample size was big: 14,000 children!
  - Cohen calculated that a person would have to grow by 3.5 feet to increase her IQ by 30 points (2 standard deviations)
  - Or, to increase her height by 4 inches, she would have to increase her IQ by 233 points!
  - Height may have been statistically significantly related to IQ, but there was no practical real-world application.

Calculating effect size (continued)

- To calculate Cohen's d, we use the same formula as for the *z* statistic, but we substitute the standard deviation for the standard error:

Cohen's *d*:

$$d = \frac{M_1 - M_2}{SD}$$

- This way, the effect size (Cohen's d) is based on the spread of the distribution of individual scores (standard deviation), not the distribution of means (standard error).

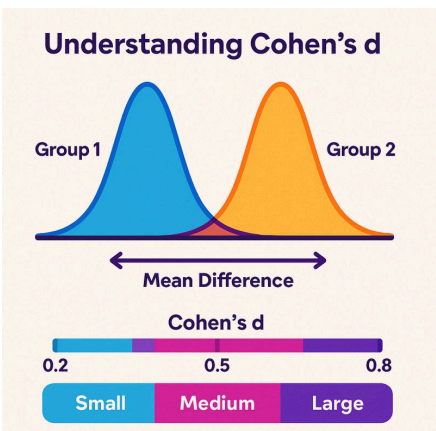
What is the effect size for the difference in teens' vs. adults' screen time?

$M1 = 11; M2 = 9; SD = 2$

Answer: d = 1 This tells us that on average, teens' screen time is 1 full standard deviation below that of adults'. This is a *large* effect.

Conventions for interpreting effect sizes

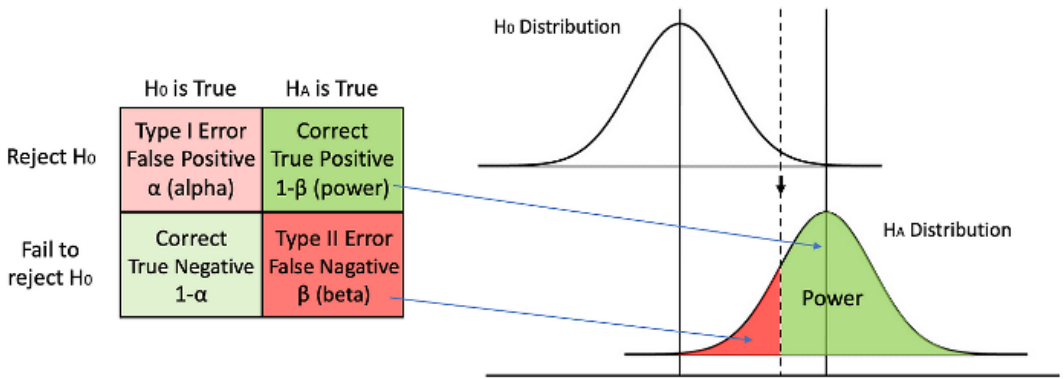
- Jacob Cohen published guidelines (or conventions) based on the overlap between two distributions to help researchers determine whether an effect is small, medium, or large.
- These numbers are not cutoffs; they are merely rough guidelines to help researchers interpret results.
- If the difference between two groups' means is less than 0.2, the difference is not important, even if it's significant.



Statistical Power

- **Statistical Power** = probability of correctly rejecting H_0 when it's false (avoiding a Type II error).
- In other words, power is the likelihood we will reject the null hypothesis *when we should*.
- Ranges from probability of 0.00 to probability of 1.00
- Probability of 0.80 (80%) is the conventional goal.
  - Many studies in psychology are underpowered!

Statistical Power (Continued)



- When testing hypotheses, there are two ways we can be correct and two ways we can be wrong:

1. Correctly rejecting the null hypothesis (true positive)
2. Correctly failing to reject the null hypothesis (true negative)
3. Incorrectly rejecting the null hypothesis (Type I error)

Five factors that influence power

Power increases when:

1. Alpha **increases**

- This is usually not a good idea: This is like changing the rules of a basketball game by shortening the basket height, or widening the goalposts in football or soccer
- Increasing the alpha level from 0.05 to 0.1 increases the probability of type I error from 5% to 10%!

2. Turn a **two-tailed test** into a **one-tailed test**

- This is only appropriate if you have a strong theoretical reason to predict the direction of the effect.

Five factors that influence power (continued)

Power increases when:

3. Sample size (n) **increases**

- This is a good idea: More data gives us a clearer picture of the population.
- Larger samples give us more precise estimates of population parameters, reducing standard error and increasing the likelihood of finding a statistically significant effect

4. Difference in means **increases**

- This is usually not under our control, but we can try to design studies that maximize effect size.
- For example, we can use extreme groups (e.g., comparing very high vs. very low anxiety individuals) to increase the difference between means.

Five factors that influence power (continued)

Power increases when:

5. Standard deviation **decreases**

- This is also a good idea: Populations with less variability make it easier to detect differences between groups.
- Often not under our control, but we can try to use reliable measures and reduce measurement error to decrease variability within groups.
- For example, if we are measuring anxiety, we can use a well-validated questionnaire rather than a single-item measure to reduce measurement error and variability within groups.

When and how do we use power?

We use power calculators in two ways:

1. Calculate power after conducting study from several pieces of information (*post hoc*).
2. Conduct power analyses before conducting study to determine sample size necessary to achieve given level of power given estimate of effect size (*a priori*).

A priori power calculations are especially useful because they help us determine the sample size needed to achieve 80% power with an alpha level of 0.05

We can use online calculators or packages for R to conduct power analyses.

Computing power is largely beyond the scope of this course, but it is very important you understand power at a conceptual level.

Parametric vs. Non-Parametric Statistics

A brief but important aside

- We have so far talked about one kind of statistical test: a *z* test.
- This is an example of a **parametric** statistical test, which depends on certain assumptions.
- **Parametric** statistical tests generally have more power than **non-parametric** statistical tests.
- However, they are more powerful in large part because they take advantage of *assumptions* about the distribution of the data.
- It is important to ensure that these assumptions are largely met before using *parametric* statistics.

Assumptions of parametric statistical tests

1. Normality

- Data within each group are (roughly) normally distributed.
- We can often assume this because of the central limit theorem!

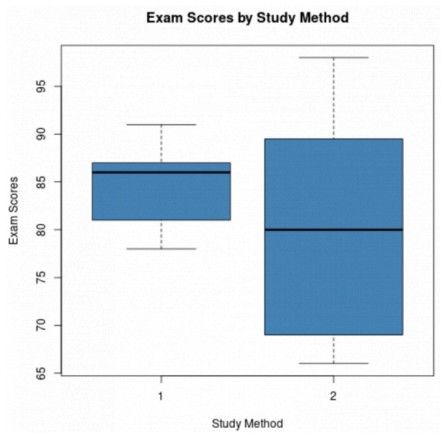
Why is the central limit theorem relevant here?

Answer: The central limit theorem states that the distribution of sample means will be approximately normal. This means that if we are working with sample means, we can assume our sample mean is drawn from a normal distribution.

Assumptions of parametric statistical tests (continued)

2. Equal variance

- Data within each group have approximately equal variance.



Assumptions of parametric statistical tests (continued)

3. Independence

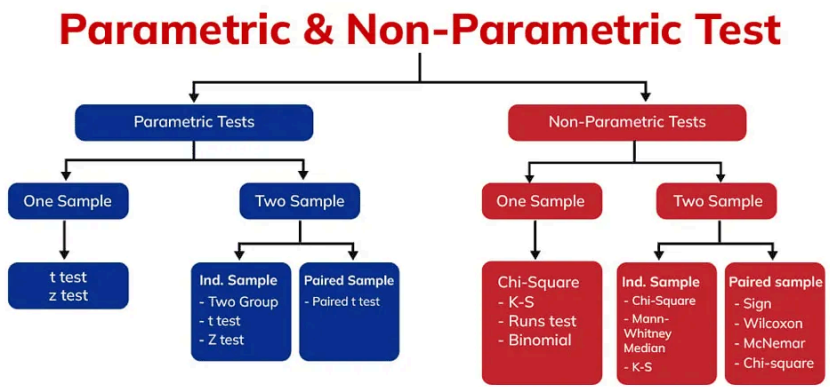
- Data in each group are randomly and *independently* sampled from a population.
- Sampling one data point does not influence the next data point that will be sampled.

4. No extreme outliers

- There should be no extreme outliers.

What if assumptions are not met?

- Often, you can still use parametric statistical tests even if assumptions are *a little* violated.
- However, there are also **non-parametric** statistical tests that do not rely on these assumptions and can be used instead.
- For every parametric statistical test, there is a non-parametric equivalent.



Moving from *z* to *t*

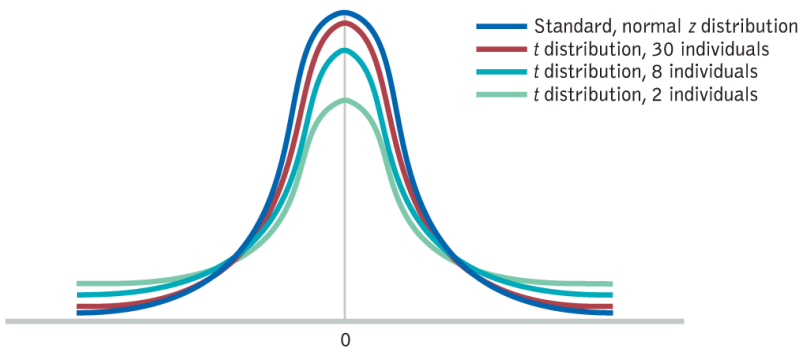
- *z* tests are one way we can compare two distributions to ask: Is the population from which this sample is drawn *different* from this other population?
- However, *z* tests assume we know the **population SD (σ)**.
- In reality, we rarely do!
- **More often, researchers use *t* tests**, which are more versatile than *z* tests.
- ***t* tests** are used:
  - when we do not know the population standard deviation (σ)
  - when we want to compare two samples to each other (vs. a sample to a known population)

Normal distributions vs. *t* distributions

- For *z* tests, we relied on normal distributions.
- For *t* tests, we will instead use *t* distributions.
- *t* distributions allow us to compare one sample to a population when we don't know all the parameters of the population.

The *t* Distribution

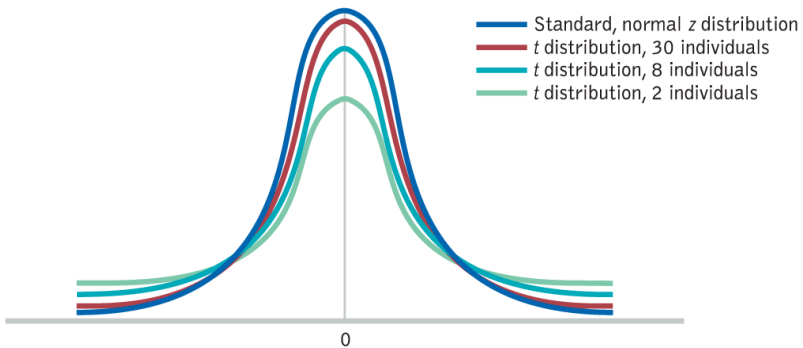
- There isn't just one *t* distribution — there's one for every sample size.
- As $n \to \infty$, the *t* distribution approaches the normal *z* distribution.
- Because we need to estimate *SD*, the *t* distribution is wider and has heavier tails than the *z* distribution. These heavier tails reflect the added uncertainty in our estimate of the population variance and, therefore, in the distribution of sample means.



Nolan/Heinzen, *Statistics for the Behavioral Sciences*, 5e, © 2020 Worth Publishers

The *t* Distribution

- For smaller samples (n = 2 or 8), *t* distributions are wider and flatter (more variability) than the *z* distribution
- As sample size increases (e.g., n = 30), the *t* distributions look more like the *z* distribution.
  - Why? A larger sample size is closer to the entire population. We expect many variables within the entire population to follow a normal distribution.



Nolan/Heinzen, *Statistics for the Behavioral Sciences*, 5e, © 2020 Worth Publishers

Types of *t* tests

- When we run *t* tests, we are asking if two means likely came from the same population distribution. Just like *z* tests!

1. Single-sample *t* test:

Used when comparing a sample mean to a population mean when the population standard deviation is not known.

2. Paired-sample *t* test:

Used when comparing two samples and every participant is in both samples. Common for *within-group* designs.

3. Independent-samples *t* test:

Used when comparing two samples and every participant is in only one sample. Common for *between-groups* designs.

How to compute a *t* statistic

- The first step to conducting a single-sample *t* test is to estimate the population standard deviation.
- The population standard deviation is based on the sample SD.
- This step is the only practical difference between conducting a z test with the *z* distribution vs. a *t* test with a *t* distribution!

The sample standard deviation formula is:

$$s = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$$

How to compute a *t* statistic (continued)

- Now that we have an estimate of the population standard deviation, we need an estimate of the spread of the distribution of means.
- We need this because:
  - We are comparing means, rather than individual scores.
  - The distribution of means is less variable than the distribution of scores.
  - Remember, when we did *z* tests, we used the standard error, not the standard deviation, which tells us the spread of sample means.
- We need to calculate standard error here too:

$$S_M = \frac{s}{\sqrt{N}}$$

How to compute a *t* statistic (continued)

- Now we can calculate our *t* statistic!

$$t = \frac{M - \mu_M}{S_M}$$

Look familiar?

This is very similar to the *z* statistic calculation:

$$z = \frac{X - \mu}{\sigma}$$

Conducting a single-sample *t* test

- The single-sample *t* test is a hypothesis test in which we compare a sample (data we collected) to a population.
- We know the mean of the population, but not the standard deviation.
- When we are using *t* distributions, we use a *t* table instead of the *z* table. The *t* table takes our sample size into account.

A *t* table

Degrees of freedom

- To use a *t* table and run a *t* test, we need to determine our **degrees of freedom**.

> Degrees of freedom = number of scores that are free to vary when we estimate a population parameter from a sample.

- Degrees of freedom reflect the amount of **independent information** available.

Degrees of freedom for a single-sample t test

df = n-1

- Here, we are estimating the population *standard deviation* from our sample data.
- We *know* the mean.
- Our degrees of freedom reflect the number of scores that could vary when a given parameter is known.
- Because our mean is known, that means all the scores in our dataset could vary, except one. Once we know the values of the first n-1 scores, the last score MUST take on a specific value.

Example: You know the mean of 3 exam scores is 90. The first two scores could be anything! But once you know them, there is only one possible score for the third exam to make the mean 90. If one score is 85 and the other score is 92, the third score **must** be 93.

Using a t table

- To use the *t* table, we look up the df to find the critical value of the *t* statistic at a given alpha level
- If the *t* statistic from our *t* test calculation is greater than the critical value, we reject the null hypothesis.

Degrees of freedom and critical *t* values

- The more degrees of freedom, the lower the critical *t* value

Comparing the *z* table and the *t* table

- As the degrees of freedom in the *t* table get very large (larger sample sizes), the *t* statistic for the 95th percentile approaches 1.96
- This is the same as the *z* statistic for the 95th percentile
- Why? The central limit theorem!
- The larger the sample, the more the *t* distribution looks like the *z* distribution (the more 'normal' it becomes)
- The *t* statistic merges with the *z* statistic as the sample size increases.

Recap: Conducting single-sample *t* tests

Used when comparing a **sample mean** to a **population mean** when population SD is unknown.

Six Steps:

1. Identify populations & assumptions.
2. State H_0 and H_1.
3. Determine characteristics of comparison distribution.
4. Find critical *t* values (using df = n – 1).
5. Calculate test statistic.
6. Make a decision.

Example: Counseling Session Contracts

- We are studying counseling sessions attended by students at a university. We want to know if signing a contract to attend counseling improves attendance.
- We ask students to sign contracts to attend a set number of counseling sessions (10)
  - Sample: Students at this counseling center who sign the contract to attend at least 10 sessions
  - Population: All students who attended counseling sessions at this university and did not sign the contract
- We sample 5 students who sign a contract. They attended 6, 6, 12, 7, and 8 counseling sessions
- The university average of students who did not sign contracts is 4.6 counseling sessions attended.

Did students who sign the contract attend a different number of sessions than those who did not?

That’s all for today!

See you next time for the Exam 2 review session!