

Lesson 14: Power and sample size calculations for means

TB sections 5.4

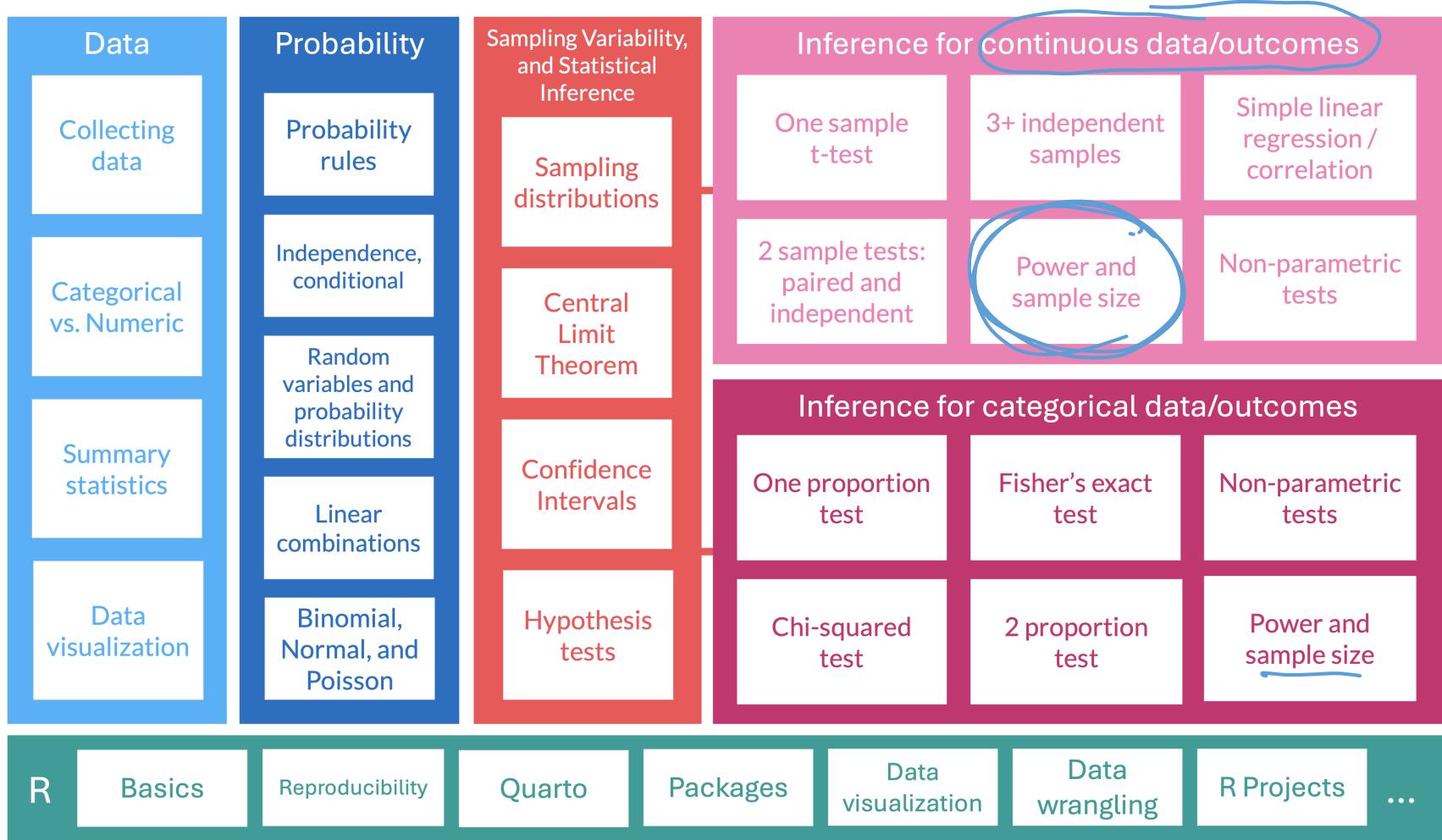
Meike Niederhausen and Nicky Wakim

2024-11-20

Learning Objectives

1. Understand the four components in equilibrium in a hypothesis test.
2. Define the significance level, critical value, and rejection region.
3. Define power and understand its role in a hypothesis test.
4. Understand how to calculate power for two independent samples.
5. Using R, calculate power and sample size for a single mean t-test and two independent mean t-test.

Where are we?



Before we get into power and sample size

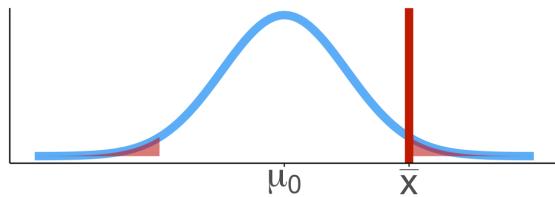
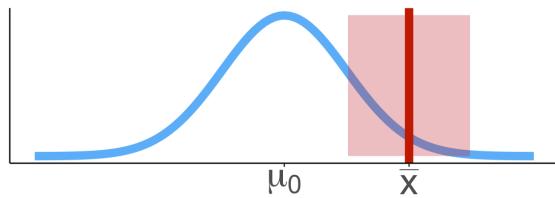
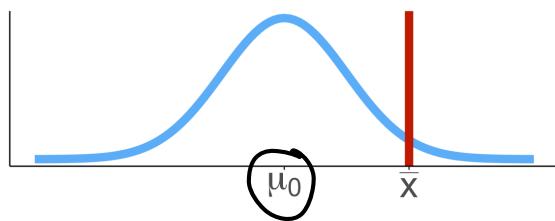
Let's watch this youtube video to jog our memory (remind us of what we learned):

One sample t-test vs Independent t-test vs Paired t-test



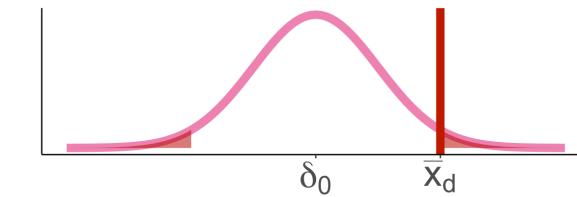
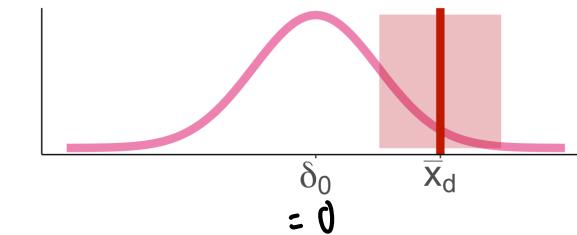
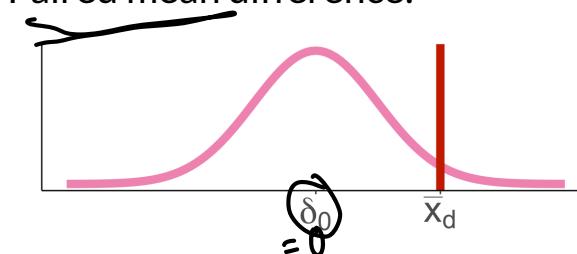
What would the distribution look like for 2 independent samples?

Single-sample mean:



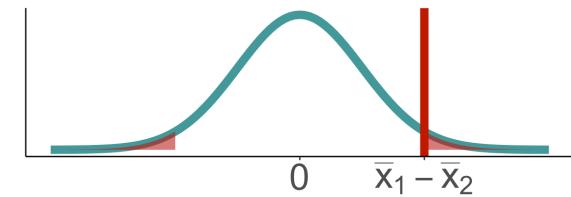
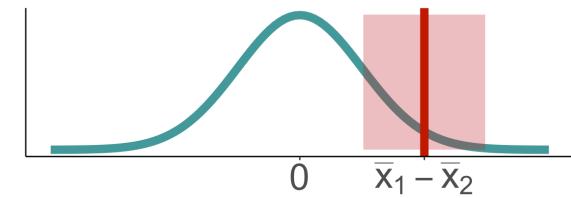
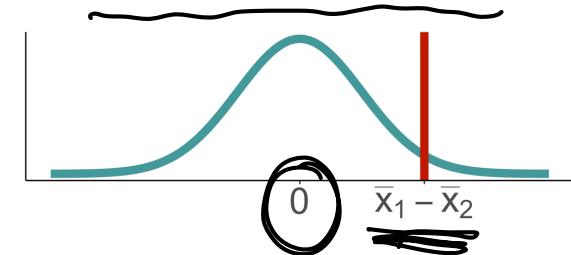
μ_0 : standard accepted
for pop

Paired mean difference:



diff or $\stackrel{=}{\text{not}}$?

Diff in means of 2 ind samples:



grp 1 same as
grp 2?

| | one sample t-test | independent 2 samp t-test | paired t-test |
|----------------------|--|---|--|
| example | <ul style="list-style-type: none"> body temp 98.25°F sample mean vs. 98.6°F (pop val) | <ul style="list-style-type: none"> caffeine, taps \bar{x}_1 caff vs. \bar{x}_2 decat | <ul style="list-style-type: none"> vegetarian diet Chol before & after for each person $x_{\text{diff}} = x_{\text{after}} - x_{\text{bef}}$ |
| hypothesis test | | | |
| test statistic | | | |
| confidence intervals | | | |

Learning Objectives

1. Understand the four components in equilibrium in a hypothesis test.
2. Define the significance level, critical value, and rejection region.
3. Define power and understand its role in a hypothesis test.
4. Understand how to calculate power for two independent samples.
5. Using R, calculate power and sample size for a single mean t-test and two independent mean t-test.

From Lesson 13: Does caffeine increase finger taps/min (on average)?

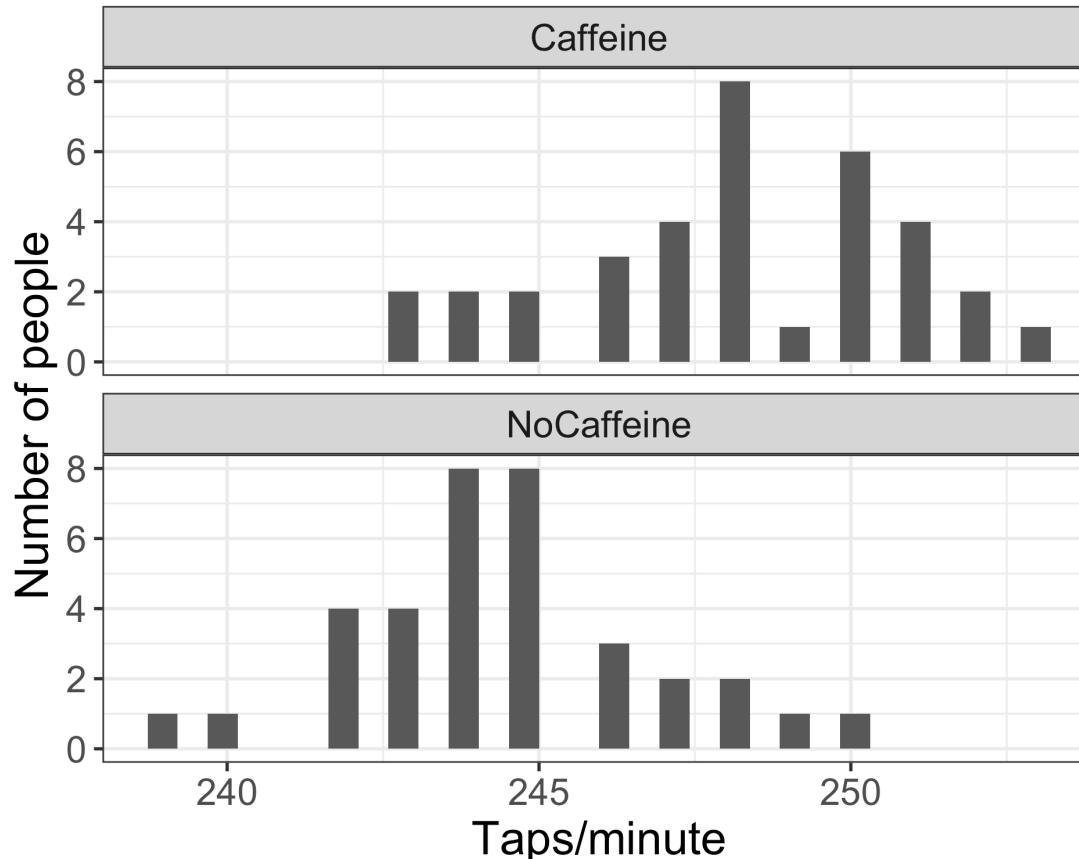
- Use this example to illustrate how to calculate a confidence interval and perform a hypothesis test for two independent samples

Study Design:¹

- 70 college students students were trained to tap their fingers at a rapid rate
- Each then drank 2 cups of coffee (double-blind)
 - Control group: decaf
 - Caffeine group: ~ 200 mg caffeine
- After 2 hours, students were tested.
- Taps/minute recorded

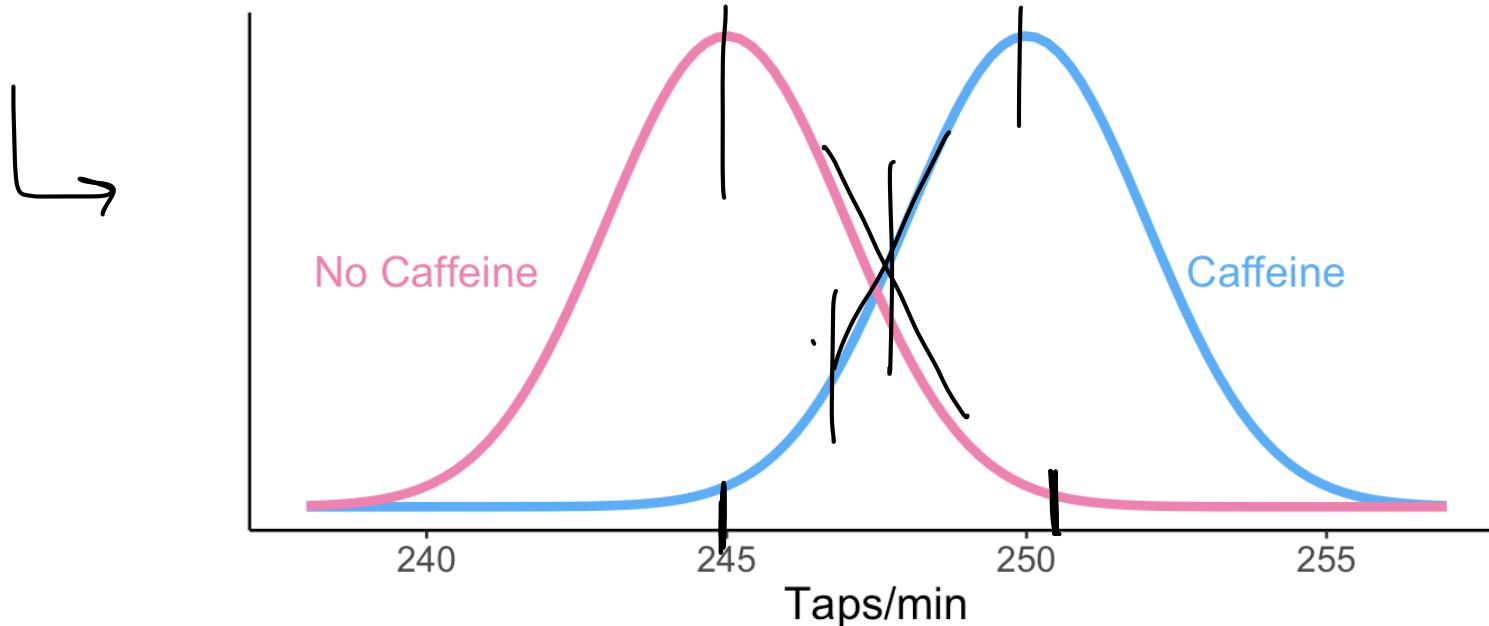
We started looking at the taps/min for each group

- ▶ Code to make these histograms



What if the following were the true population distributions? Case 1

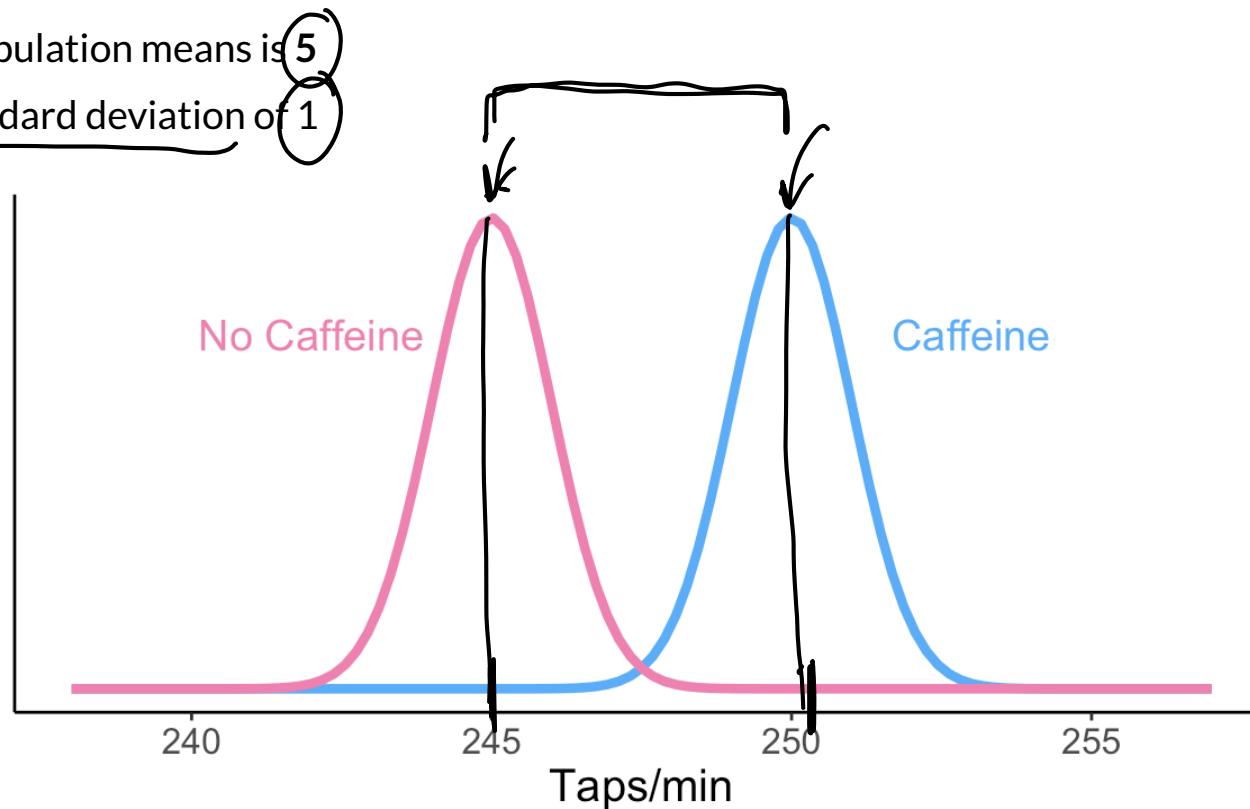
- Difference in population means is 5 ✓ caffeine has 5 taps/min more in avg
- Both have a standard deviation of 2



- When we take two samples from these groups, do you think it would be easy to distinguish between the mean taps/min?
 - Depends on the number of samples we get: we might need a lot ✓

What if the following were the true population distributions? Case 2

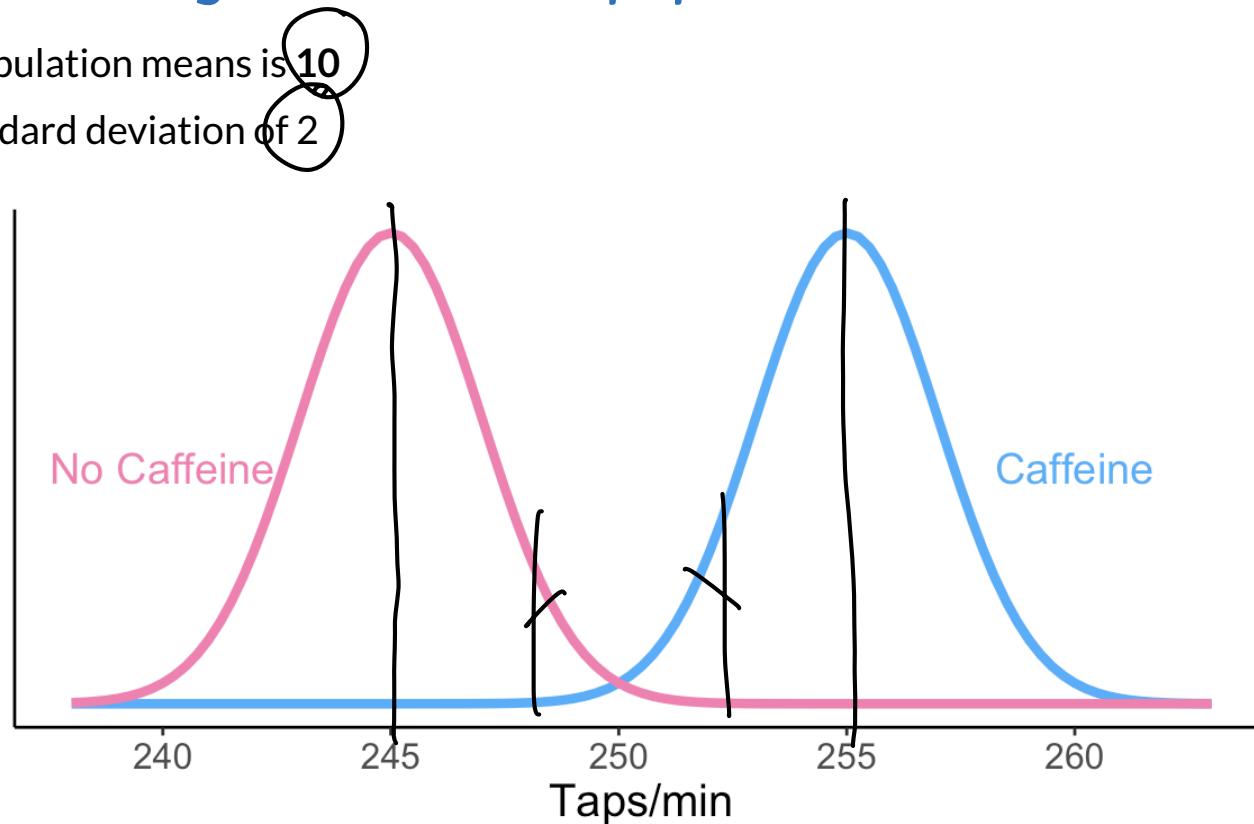
- Difference in population means is 5
- Both have a standard deviation of 1



- When we take two samples from these groups, do you think it would be easy to distinguish between the mean taps/min?
 - Seems easier to distinguish here. How did the standard deviation decrease?

What if the following were the true population distributions? Case 3

- Difference in population means is 10
- Both have a standard deviation of 2



- When we take two samples from these groups, do you think it would be easy to distinguish between the mean taps/min?
 - Also seems easier to distinguish here

There are a few things at play here

- There are several measurements that affect how easy it is to distinguish between two populations
- “Distinguish between two populations” = correctly reject the null hypothesis that they are the same

- What elements are at play?

1. Difference in population means

$\uparrow \text{diff}$ - more easily disting.

2. Number of samples from each population

$\uparrow n \downarrow SE$ - more easily dist,

$$SE = \frac{S}{\sqrt{n}}$$

3. The significance level that we use for a cut off

4. The power of our test

- More familiar with first two, but let's define #3 and #4 more

Learning Objectives

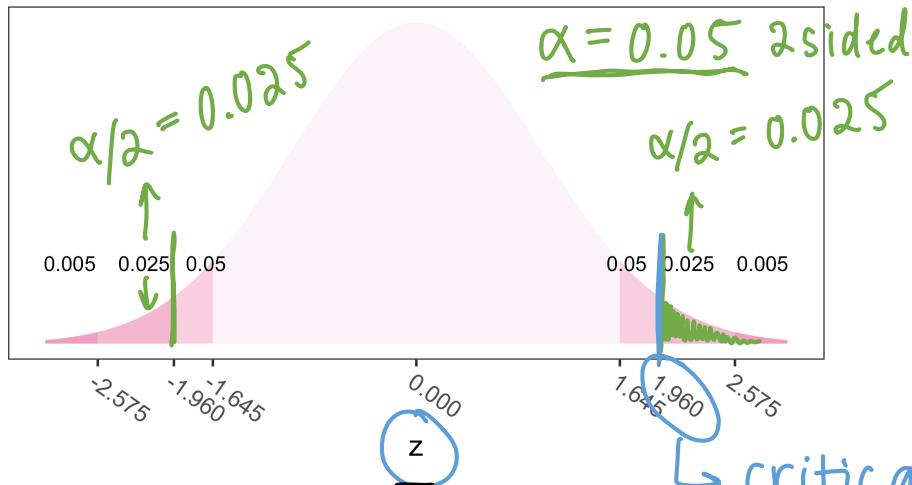
1. Understand the four components in equilibrium in a hypothesis test.
2. Define the significance level, critical value, and rejection region.
3. Define power and understand its role in a hypothesis test.
4. Understand how to calculate power for two independent samples.
5. Using R, calculate power and sample size for a single mean t-test and two independent mean t-test.

Significance levels and critical values

t^* 95% CI

- **Critical values** are the cutoff values that determine whether a test statistic is statistically significant or not
 - Determined by the **significance level**
- If a test statistic is greater in absolute value than the critical value, we reject H_0

Critical Values for a Normal Distribution



- Critical values are determined by
 - the significance level α , ✓
 - whether a test is 1- or 2-sided, & ✓
 - the probability distribution being used to calculate the p-value (such as normal or t-distribution) ↗ df.

- We have been referring to critical values from the t-distribution as t^*
 - See how we calculate a specific confidence interval in Lesson 10

critical value of 1.96 is :
 $P(Z \geq 1.96) = 0.025 = \frac{\alpha}{2}$

Poll Everywhere Question 1

$$P(Z \leq z^*) = 0.95$$

13:39 Wed Nov 20

Join by Web PollEv.com/nickywakim275

If we wanted to find the two-sided critical value for the $\alpha = 0.10$ under the standard normal distribution, what R function would we use?

$qnorm(p = 0.05, lower.tail = F)$

qnorm($p = 0.9$) 47%

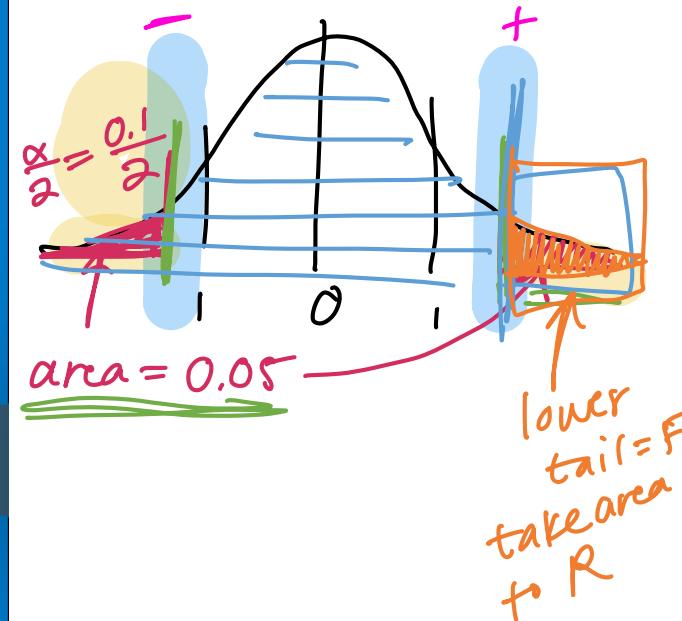
qnorm($p = 0.95$) ✓ 24%

0.05 qnorm($p = 0.10$) 24%

pnorm($p = 0.10$) 6%

Powered by  Poll Everywhere

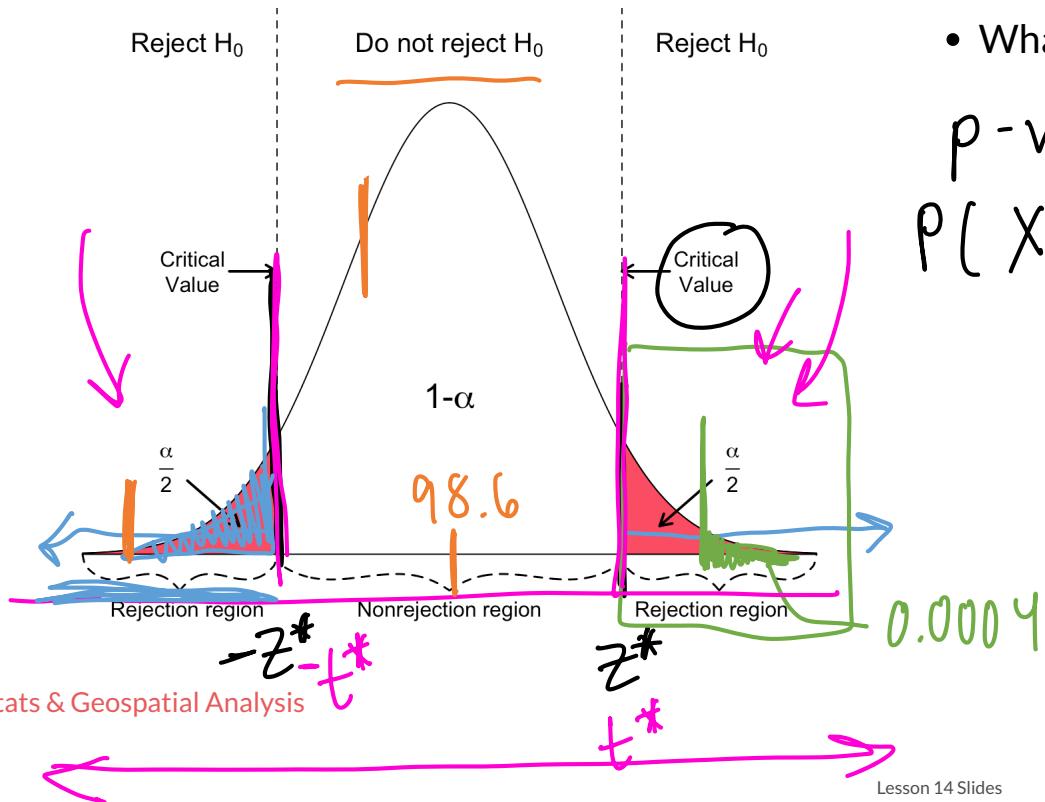
qnorm gives us the z-value for a prob.



Rejection region, significance levels, and critical values

- If the absolute value of the test statistic is greater than the critical value, we reject H_0
 - In this case the test statistic is in the **rejection region**.
 - Otherwise it's in the non-rejection region.

if we have $z \geq |z^*|$
 $z \leq -z^* \quad z \geq z^*$



- What do rejection regions look like for 1-sided tests?

$p\text{-value} < \alpha$
 $P(X \geq 98.25) = 0.0004 < \alpha$
 $t = \frac{98.25 - 98.6}{S/\sqrt{n}} > t^*$
or $t < -t^*$

Learning Objectives

1. Understand the four components in equilibrium in a hypothesis test.
2. Define the significance level, critical value, and rejection region.
3. Define power and understand its role in a hypothesis test.
4. Understand how to calculate power for two independent samples.
5. Using R, calculate power and sample size for a single mean t-test and two independent mean t-test.

Let's start with some important definitions in words

- Type I error (α): Probability of rejecting the null hypothesis given that the null is true incorrect
- Type II error (β): Probability of failing to reject the null hypothesis given that the null hypothesis is false
- Power (or sensitivity) ($1 - \beta$): Probability of rejecting the null hypothesis given that the null is false (correct)
- Specificity ($1 - \alpha$): Probability of failing to reject the null hypothesis given that the null is true (correct)

our conclusion

| | | <u>Fail to reject null hypothesis</u> | <u>Reject null hypothesis</u> |
|-------------------------|-----------------------------|--|--|
| <i>underlying truth</i> | Null hypothesis is true | Correct! ✓ (true negative) | Type I error (false positive) probability = α |
| | Null hypothesis is false | Type II error (false negative) probability = β | Correct! (true positive) power |

What does that look like with our two populations?

null: both equal to mean of no caffeine

| | Fail to reject null hypothesis | Reject null hypothesis |
|--------------------------|--|--|
| Null hypothesis is true | Correct! (true negative) | Type I error (false positive) probability = α |
| Null hypothesis is false | Type II error (false negative) probability = β | Correct! (true positive) |

• α = probability of making a Type I error

- This is the significance level (usually 0.05)

- Set before study starts

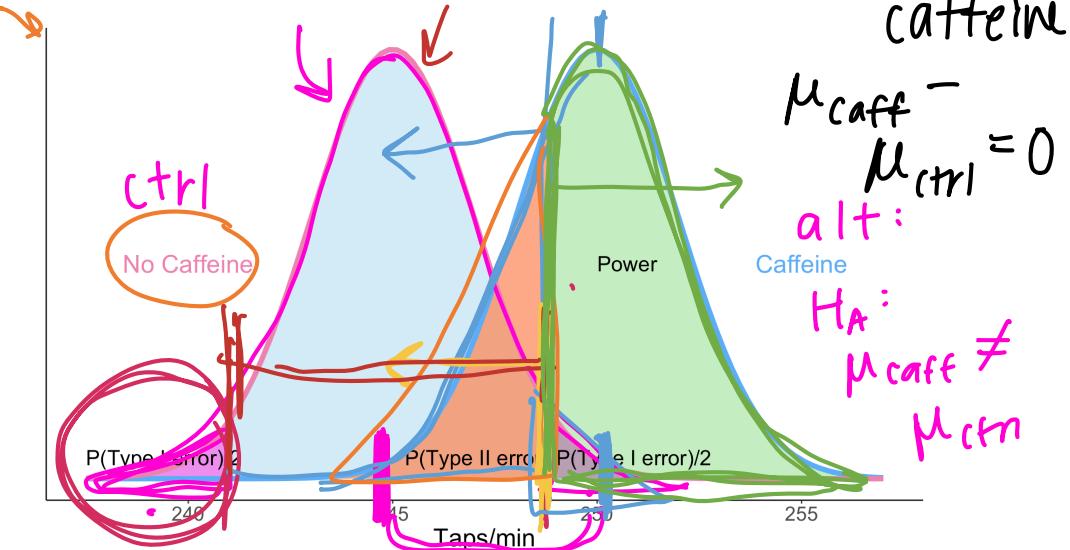
• β = probability of making a Type II error

- Ideally we want

- small Type I & II errors and

- big power

detect diff when diff exists



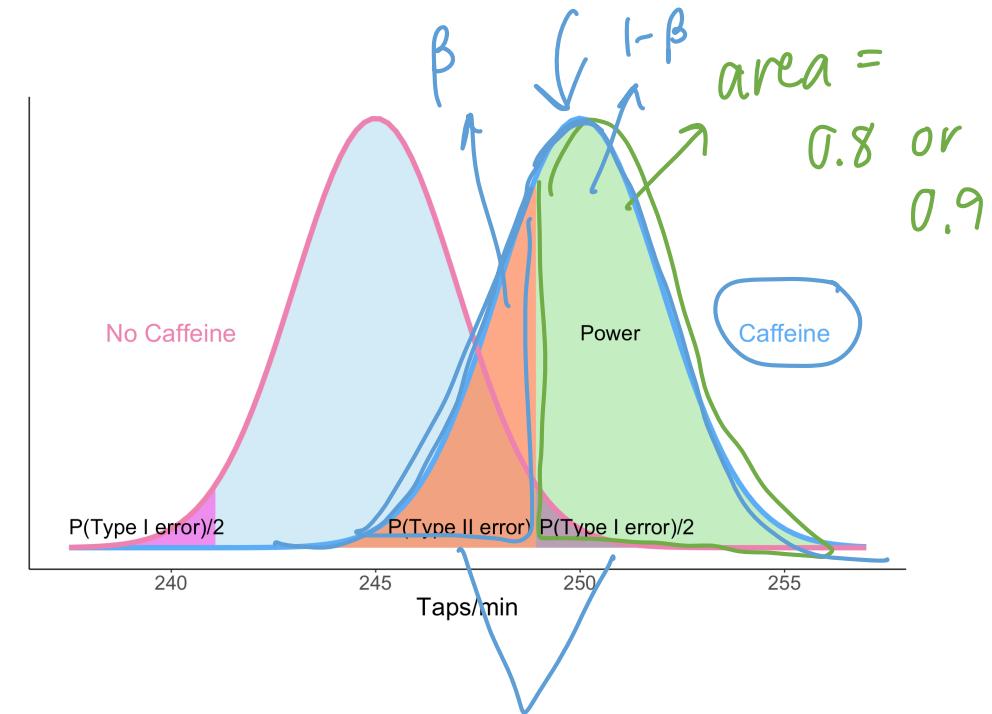
Power (or sensitivity) $(1 - \beta)$: Probability of rejecting the null hypothesis given that the null is false (correct)

- Power is the correct region that is usually **in line with our study design**: studies are often seeing if there is a distinction between two populations

Power

- Power (or sensitivity) ($1 - \beta$): Probability of rejecting the null hypothesis given that the null is false (correct)

- Power is also called the
 - true positive rate,
 - probability of detection, or
 - the *sensitivity* of a test
- Typically, we aim for 80% or 90% power



$$P(\text{type 2 error}) + \text{power} = 1$$

Let's demonstrate the relationship between error and power

From the applet at <https://rpsychologist.com/d3/NHST/>

Let's look at the following scenarios:

1. Solve for power: decreasing type 1 error (α)

2. Solve for power: increasing type 1 error (α)

3. Solve for power: decrease sample size

4. Solve for power: increase sample size

5. Solve for power: increase difference of means

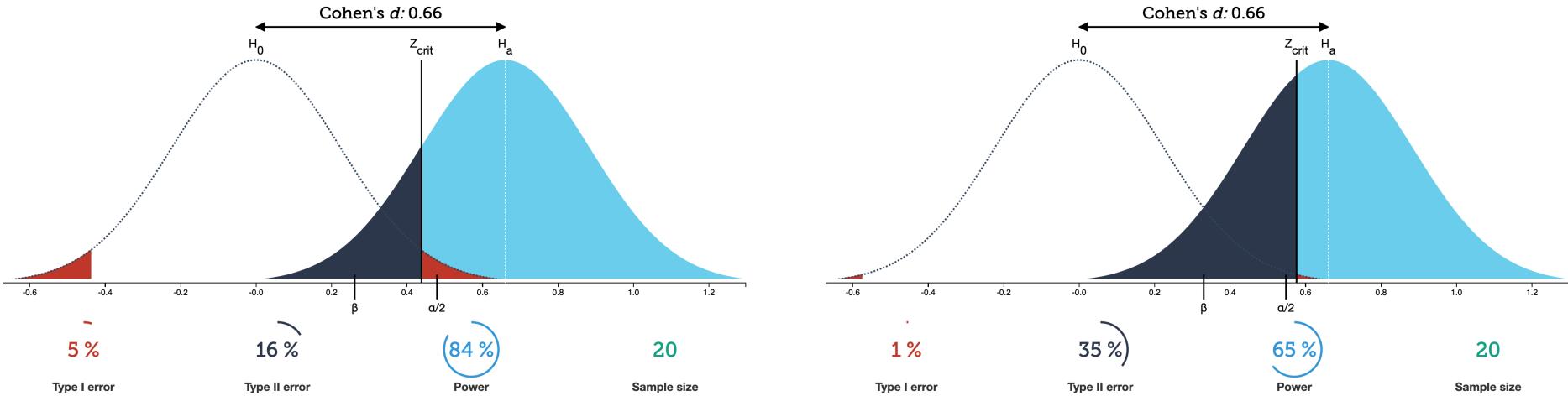
6. Solve for power: decrease difference of means

- Takeaway: cannot minimize both type 1 and 2 error

- Takeaway: increasing sample size increases power

If you want to keep revisiting these concepts!

From the applet at <https://rpsychologist.com/d3/NHST/>



- Cohen's d is just a standardized value to represent the difference in means:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

Learning Objectives

1. Understand the four components in equilibrium in a hypothesis test.
2. Define the significance level, critical value, and rejection region.
3. Define power and understand its role in a hypothesis test.
4. Understand how to calculate power for two independent samples.
5. Using R, calculate power and sample size for a single mean t-test and two independent mean t-test.

Calculating power or sample size

- Typically, before we set up a research study, we try to find the needed sample size to achieve 80% or 90% power
- If we have already have data, then we typically calculate the power based on the sample we have

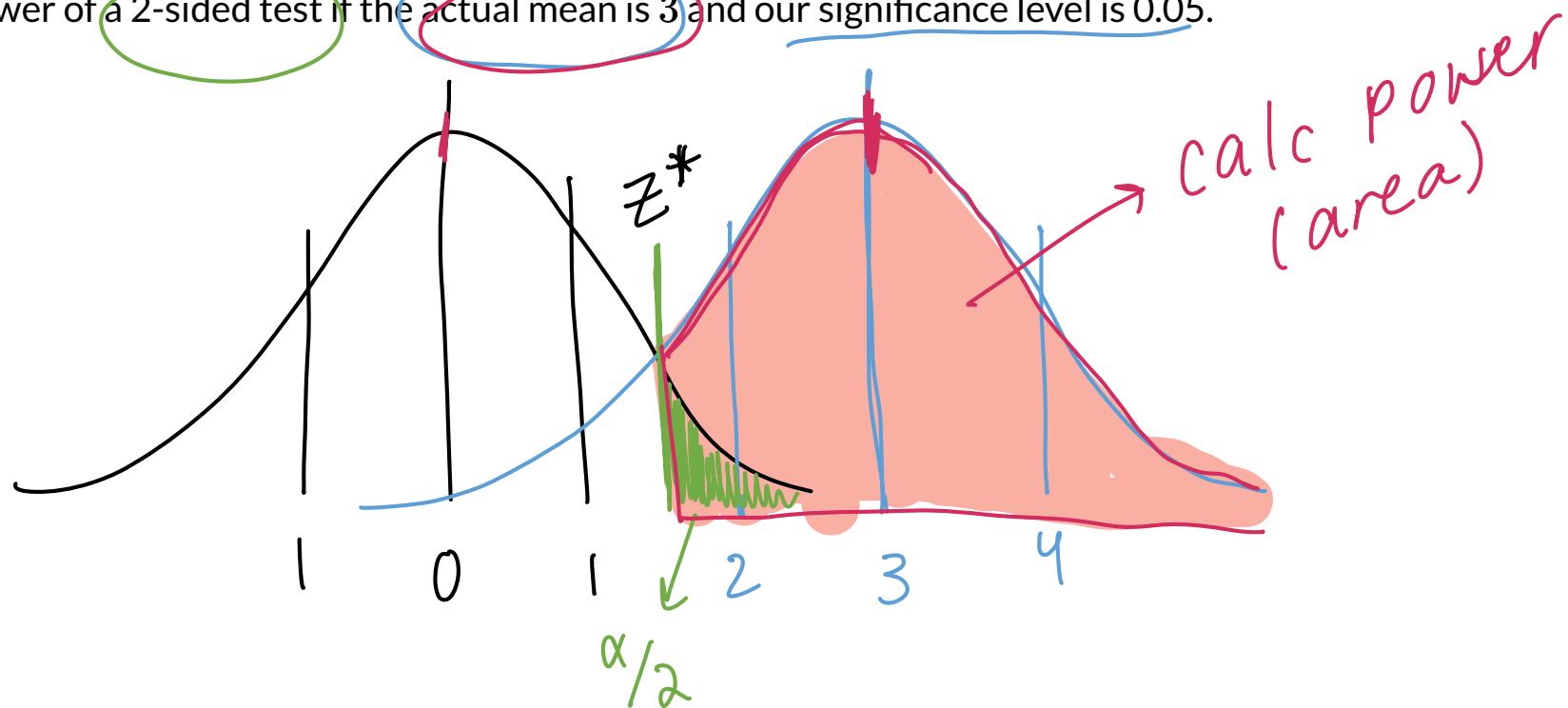
20% power – cut it off

Example calculating power (1/3)

Let's say we have:

- a null population with a normal distribution, centered at 0 with a standard deviation of 1 ($X_0 \sim Norm(0, 1)$)
- an alternative population, centered at 3 with a standard deviation of 1 ($X_A \sim Norm(3, 1)$)

Find the power of a 2-sided test if the actual mean is 3 and our significance level is 0.05.



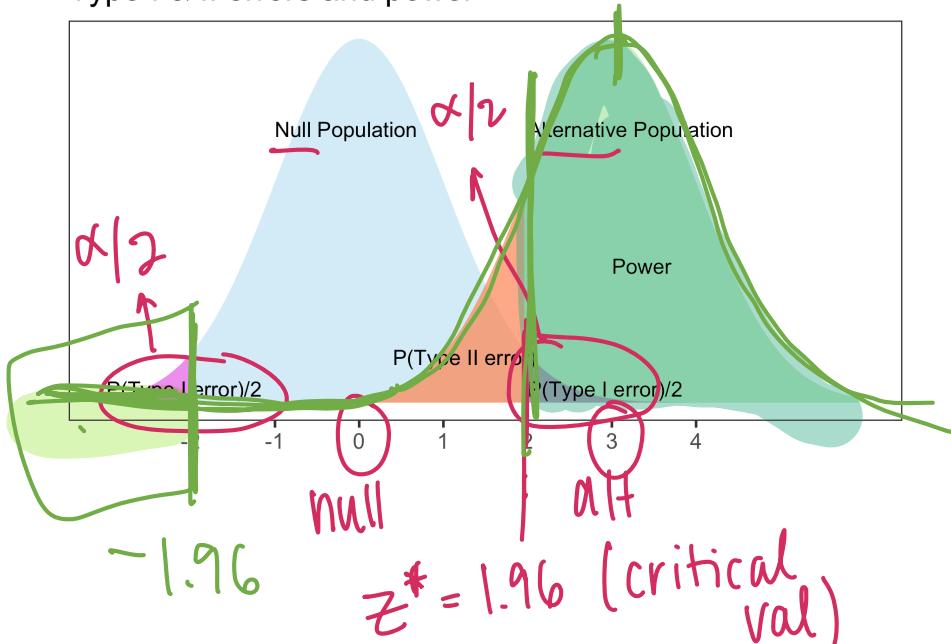
Example calculating power (2/3)

Let's say we have:

- a null population with a normal distribution, centered at 0 with a standard error of 1 ($X_0 \sim Norm(0, 1)$)
 - an alternative population, centered at 3 with a standard error of 1 ($X_A \sim Norm(3, 1)$)

Find the power of a 2-sided test if the actual mean is 3 and our significance level is 0.05.

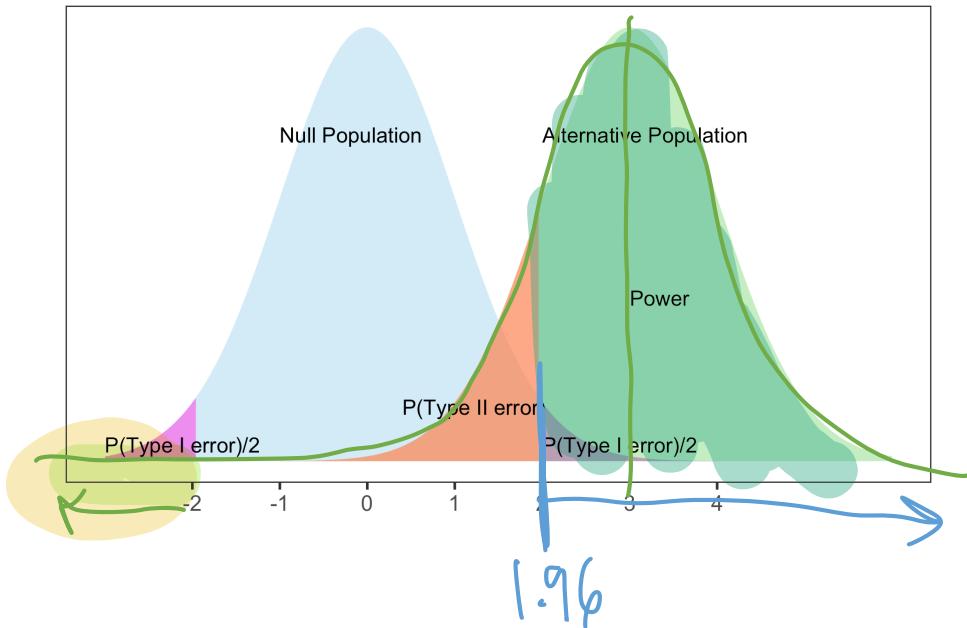
Type I & II errors and power



- Power = $P(\text{Reject } H_0 \text{ when alternative pop is true})$
 - Correctly reject null
 - When $\alpha = 0.05$, we reject H_0 when the test statistic z is at least 1.96 (critical value is 1.96 under the null distribution)
 - Then we need to calculate the probability that we are in the rejection regions given we are **actually in the alternative population**
 - Thus under the alternative population, we need to calculate $P(X_A \leq -1.96) + P(X_A \geq 1.96)$

220

Example calculating power (3/3)



- Thus under the alternative population, we need to calculate $P(X_A \leq -1.96) + P(X_A \geq 1.96)$
- Under the alternative population we have $X_A \sim Norm(3, 1)$

```
1 # left tail + right tail:  
2 pnorm(-1.96, mean=3, sd=1,  
3 lower.tail=TRUE) +  
4 pnorm(1.96, mean=3, sd=1,  
5 lower.tail=FALSE)  
[1] 0.8508304
```

right side

Answer: The power is 85%

- The left tail probability `pnorm(-1.96, mean=3, sd=1, lower.tail=TRUE)` is essentially 0 in this case.
- Note that this power calculation specified the value of the SE instead of the standard deviation and sample size n individually.

Learning Objectives

1. Understand the four components in equilibrium in a hypothesis test.
2. Define the significance level, critical value, and rejection region.
3. Define power and understand its role in a hypothesis test.
4. Understand how to calculate power for two independent samples.
5. Using R, calculate power and sample size for a single mean t-test and two independent mean t-test.

R package `pwr` for power analyses¹

- Use `pwr.t.test` for both one- and two-sample t-tests *for means!*
- Specify all parameters except for the one being solved for

```
1 pwr.t.test(n = NULL, — Sample size
2             d = NULL, → Cohen's d (effect size) / diff in means
3             sig.level = 0.05, →  $\alpha$ 
4             power = NULL, →  $1 - \beta$ 
5             type = c("two.sample", "one.sample", "paired"),
6             alternative = c("two.sided", "less", "greater"))
```

- Leave out:

- n: returns sample size
- d: returns Cohen's d/effect size (next slide)
- sig.level: get significance level (not typical)
- power: returns power

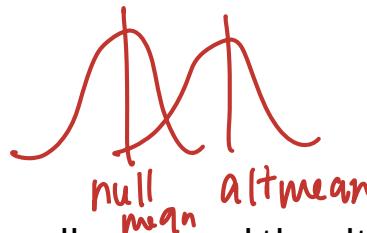
$$\begin{aligned} H_A: \quad & \mu_1 \neq \mu_2 && \text{two sided} \\ & \mu_1 > \mu_2 && \text{greater} \\ & \mu_1 < \mu_2 && \text{less} \end{aligned}$$

What is Cohen's d ?

- d is Cohen's d effect size
 - Just a standardized way to measure the distance between the null mean and the alternative mean
- Examples of values: small = 0.2, medium = 0.5, large = 0.8

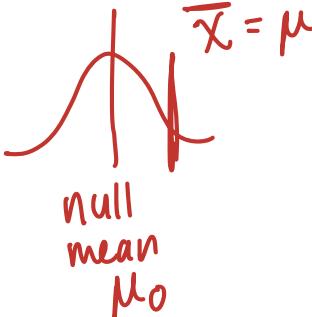
One-sample test (or paired t-test):

$$d = \frac{\bar{x} - \mu_0}{s}$$



Two-sample test (independent):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$$



- $\bar{x}_1 - \bar{x}_2$ is the difference in means between the two groups that one would want to be able to detect as being significant,
- s_{pooled} is the pooled SD between the two groups - often assume have same sd in each group

Power calculation for testing one mean

Conversely, we can calculate how much power we had in our body temperature one-sample test, given the sample size of 130.

- Calculate power,

- given α , n , “true” alternative mean μ , and null μ_0 ,
- assuming the test statistic is normal (instead of t-distribution)

$$1 - \beta = P(Z \leq z - z_{1-\alpha/2}) + P(Z \leq -z - z_{1-\alpha/2}) , \quad \text{where } z = \frac{\mu - \mu_0}{s/\sqrt{n}}$$

Φ is the probability for a standard normal distribution

```
1 mu <- 98.25; mu0 <- 98.6; sd <- 0.73; alpha <- 0.05; n <- 130  
2 (z <- (mu-mu0) / (sd/sqrt(n)) )
```

```
[1] -5.466595
```

```
1 (Power <- pnorm(z-qnorm(1-alpha/2)) + pnorm(-z-qnorm(1-alpha/2)))
```

```
[1] 0.9997731
```

If the population mean is 98.2 instead of 98.6, we have a 99.98% chance of correctly rejecting H_0 when the sample size is 130.

Sample size calculation for testing one mean

- Recall in our body temperature example that $\mu_0 = 98.6^\circ\text{F}$ and $\bar{x} = 98.25^\circ\text{F}$.

sample mean

- The p -value from the hypothesis test was highly significant (very small).
- What would the sample size n need to be for 80% power?

130 ppl

Calculate n

- given α , power ($1 - \beta$), "true" alternative mean μ , and null μ_0
- Calculate d : $d = \frac{\mu - \mu_0}{s}$

$$\frac{\bar{x} - \mu_0}{s} = \frac{98.25 - 98.6}{0.73}$$



Sd of
the sample

pwr: sample size for one mean test

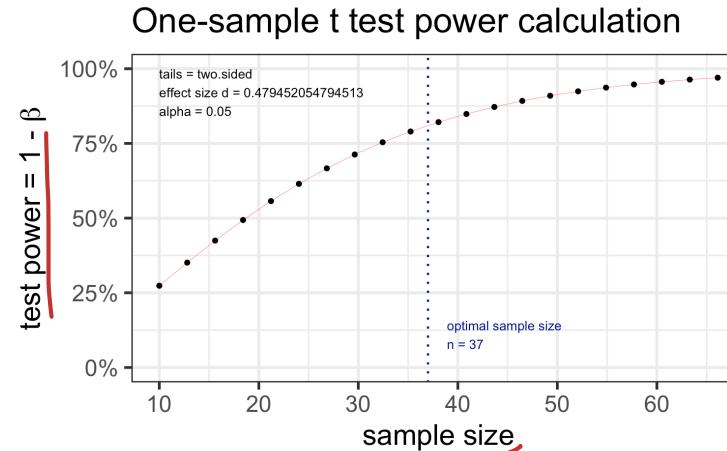
Specify all parameters except for the sample size:

```
1 library(pwr)
2 t.n <- pwr.t.test(
3   d = (98.6 - 98.25) / 0.73, J
4   sig.level = 0.05,
5   power = 0.80,
6   type = "one.sample")
7 alternative = "two.sided"
8 t.n
```

One-sample t test power calculation

n = 36.11196
d = 0.4794521
sig.level = 0.05
power = 0.8
alternative = two.sided

```
1 plot(t.n)
```



We need 37 individuals to detect this difference with 80% power.

pwr: power for one mean test

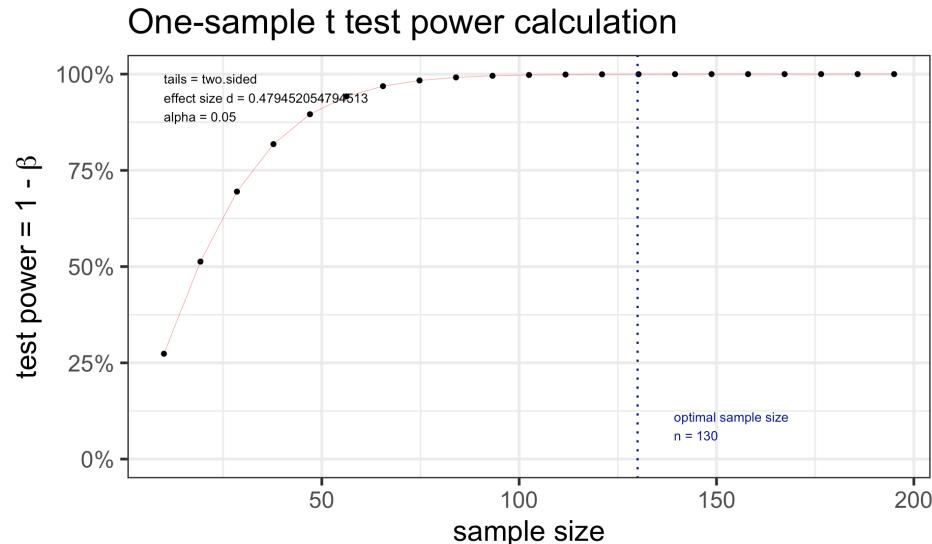
Specify all parameters *except* for the power:

```
1 t.power <- pwr.t.test(  
2   d = (98.6-98.25)/0.73,  
3   sig.level = 0.05,  
4   # power = 0.80,  
5   n = 130,  
6   type = "one.sample")  
7  
8 t.power
```

One-sample t test power calculation

```
n = 130  
d = 0.4794521  
sig.level = 0.05  
power = 0.9997354  
alternative = two.sided
```

```
1 plot(t.power)
```



We have **99.97% power** to detect this difference with 130 individuals.

pwr: Two-sample t-test: sample size

Example: Let's revisit our caffeine taps study. Investigators want to know what sample size they would need to detect a 2 point difference between the two groups. Assume the SD in both group samples is 2.6.

Specify all parameters *except* for the sample size:

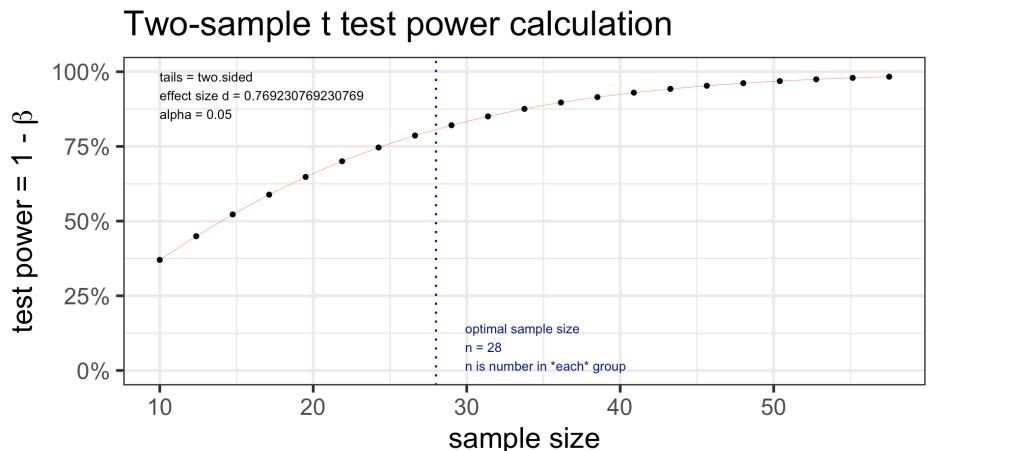
```
1 t2.n <- pwr.t.test(  
2   d = 2/2.6,  
3   sig.level = 0.05,  
4   power = 0.80,  
5   type = "two.sample")  
6  
7 t2.n
```

Two-sample t test power calculation

```
n = 27.52331  
d = 0.7692308  
sig.level = 0.05  
power = 0.8  
alternative = two.sided
```

NOTE: n is number in *each* group

```
1 plot(t2.n)
```



We **need 28 individuals** to detect this difference with 80% power.

:::::

pwr: Two-sample t-test: power

Example: Let's revisit our caffeine taps study. Investigators want to know what power they have to detect a 2 point difference between the two groups. The two groups are both size 35 (like in our previous example). Assume the SD in both group samples is 2.6.

Specify all parameters *except* for the power:

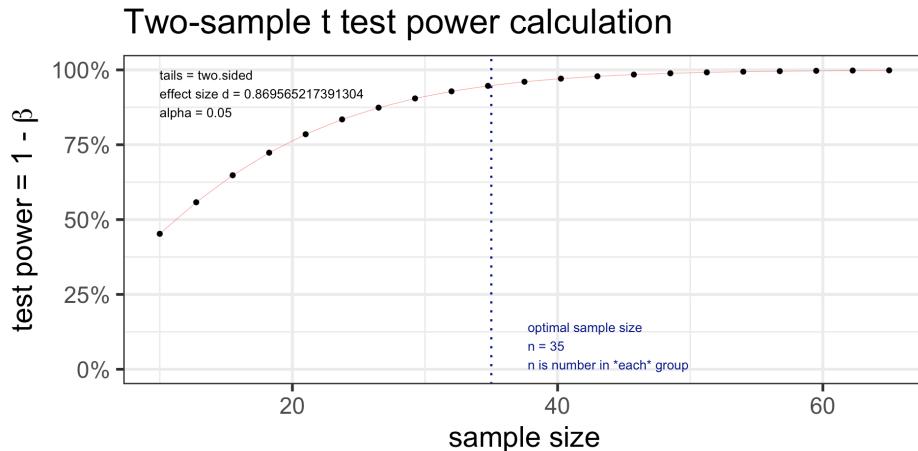
```
1 t2.power <- pwr.t.test(  
2   d = 2/2.3,  
3   sig.level = 0.05,  
4   n = 35,  
5   type = "two.sample")  
6  
7 t2.power
```

```
Two-sample t test power calculation

n = 35
d = 0.8695652
sig.level = 0.05
power = 0.9480091
alternative = two.sided
```

NOTE: n is number in *each* group

```
1 plot(t2.power)
```



We have **94.8%** power to detect this difference with 35 individuals in each group.

Resources for power and sample size calculations

More software for power and sample size calculations: PASS

- PASS is a very powerful (& expensive) software that does power and sample size calculations for many advanced statistical modeling techniques.
 - Even if you don't have access to PASS, their **documentation** is very good and free online.
 - Documentation includes formulas and references.
 - PASS documentation for powering **means**
 - One mean, paired means, two independent means
- One-sample t-test documentation: https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/PASS/One-Sample_T-Tests.pdf

OCTRI-BERD power & sample size presentations

- **Power and Sample Size 101**

- Presented by Meike Niederhausen; April 13, 2023
 - Slides: <http://bit.ly/PSS101-BERD-April2023>
 - [Recording](#)

- **Power and Sample Size for Clinical Trials: An Introduction**

- Presented by Yiyi Chen; Feb 18, 2021
 - Slides: <http://bit.ly/PSS-ClinicalTrials>
 - [Recording](#)

- **Planning a Study with Power and Sample Size Considerations in Mind**

- Presented by David Yanez; May 29, 2019
 - [Slides](#)
 - [Recording](#)

- **Power and Sample Size Simulations in R**

- Presented by Robin Baudier; Sept 21, 2023
 - [Slides](#)
 - [Recording](#)

