Standard Error

- Standard deviation of the sampling distribution.
- How spread out is the sampling distribution?
- How large are the typical chance differences?
- Later we will examine statistical power.
    - The spread of the sampling distribution is the standard error.
    - In what kinds of experiments are large and small differences likely to arise by chance?

Sampling Distributions and Randomization Inference

- First, read GG 3.0, 3.1, 3.4.
- Groups may differ by chance, even if the treatment has no effect.
  - How much would the groups differ if the treatment had no effect?
  - How large of an "effect estimate" would we reach by chance?
- Distribution of estimates one would reach if treatment had no effect.
  - How likely is this estimate to have just arisen by chance?
- Similar to observational studies, but:
  - Intuition easy to see in experiments.
  - Testing a hypothesis about our sample, not a population.
- Example code to walk through intuition on slides to follow.

Example: An Experiment With No Effect

- Does eating soybeans affect estrogen levels?
- 40 individuals: 20 men, 20 women.
- Simulate the potential outcomes of the control group.
- Simulate the potential outcomes of the treatment group.
- A simulated experiment with no effect.

```
group <- c(rep("Man",20),rep("Woman",20))
po.control <- c(seq(from = 1, to = 20), seq(from = 51, to = 70))
po.treatment <- po.control #no effect because potential outcomes in
treatment are the same
po.control:
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70
po.treatment:
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70
```

Random Assignment

- Define function to randomly assign units to treatment and control.
- Randomly pick 20 for treatment and 20 for control.
- Concatenate the two vectors.
- Get a different vector when you run it again.

```
randomize <- function() sample(c(rep(0,20),rep(1,20)))
randomize(): [1] 0 0 1 0 1 1 1 0 0 0 1 0 0 0 1 1 0 1 0 0 1 0 0 1 0 1 0 1
1 1 1 0 0 1 1 1 1 10 0
randomize(): [1] 0 0 1 1 0 0 0 1 1 1 0 0 0 0 0 1 1 0 1 0 1 0 1 1 0 1 0 1
1 1 1 0 1 0 0 1 0 0 1 1
treatment <- randomize() #Conduct randomization for this experiment
treatment: [1] 0 0 0 1 1 1 0 0 1 0 1 1 1 1 0 1 1 1 0 0 0 1 0 0 0 1 0 0 0
1 1 0 1 0 1 0 1 1 1 0
```

Realized Outcomes

- Treatment outcome for those randomized to treatment and control outcome for those randomized to control.
- Assign for each person in the vector.
- Same because we had an experiment with no effect.
- R code is often written in a compact manner; could also have been done separately for each group.
- Why are we doing this when there is no treatment effect?
  - Because it should also work when there is one.
- We're looking at what happens when we randomly assign people to control and treatment groups.

```
outcomes <- po.treatment * treatment + po.control*(1-treatment)
outcomes: [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 51 52
53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70
```

Function to Estimate the Average Treatment Effect

- Subtract the mean outcome for the control group from the mean outcome of the treatment group.
- How much higher is the average in the treatment group versus the control group?
- We may have randomly selected someone with a higher or lower level of estrogen.
- Even though we know the effect is 0, we see chance differences.

```
est.ate <- function(outcome, treat) mean(outcome[treat==1]) -
mean(outcome[treat==0])
ate <- est.ate(outcomes, treatment) #Compute the average treatment
effect for this experiment
ate: [1] 1.3 #Difference, despite there being no effect!
```

Rhetorical Posture of the Null Hypothesis

- You want to argue against a skeptic that a treatment has an effect.
- Assume the skeptic is right.
  - Treatment has no effect.
- What is the chance that we would see this estimate by chance in that scenario?
  - This is p-value.
  - We'll see where it comes from visually.

Average Size of the Difference

- Simulate this a few times to get a sense of how much our treatment effect estimate would vary by chance.
- We created an estimate function with the outcomes and the treatment group.
- Outcome vector will look the same regardless of the treatment vector.

```
est.ate(outcomes, randomize()) #Outcomes are randomized
est.ate(outcomes, randomize()): [1] -6.1
est.ate(outcomes, randomize()): [1] -14.7
est.ate(outcomes, randomize()): [1] 1.7
```

Outcome With Different Assignments

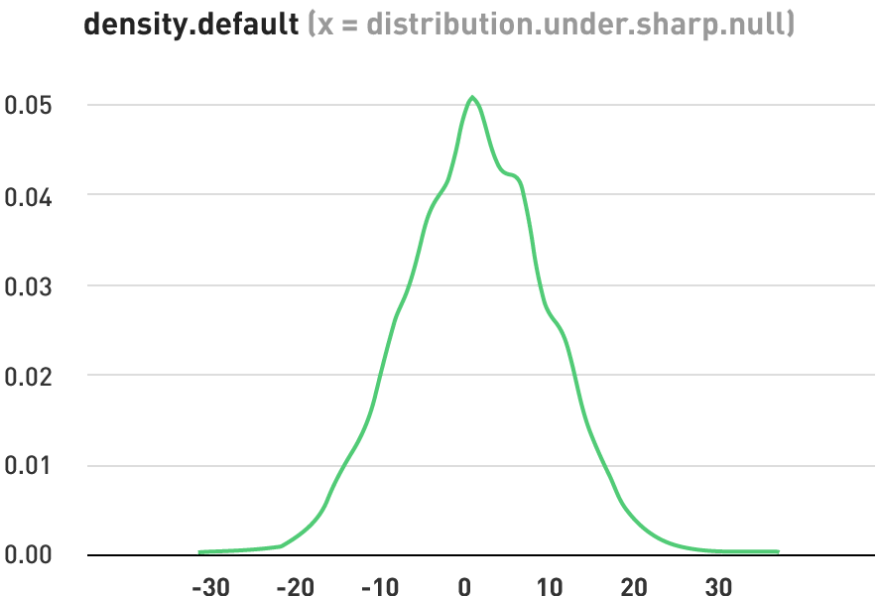
- Similar to resampling from a population.
- Rerandomizing from within the original population.
- Testing the null hypothesis from within the sample we already have.
- Reshuffle the 40 people between treatment and control.
- Assuming the treatment effect for everyone is zero.
- Sharp null hypothesis: For every unit, there is no effect.

Repeat to See Distribution of Effects

- Do this 5,000 times to get a sense of the distribution of effects.
- Replicate 5,000 times, and save to a vector.

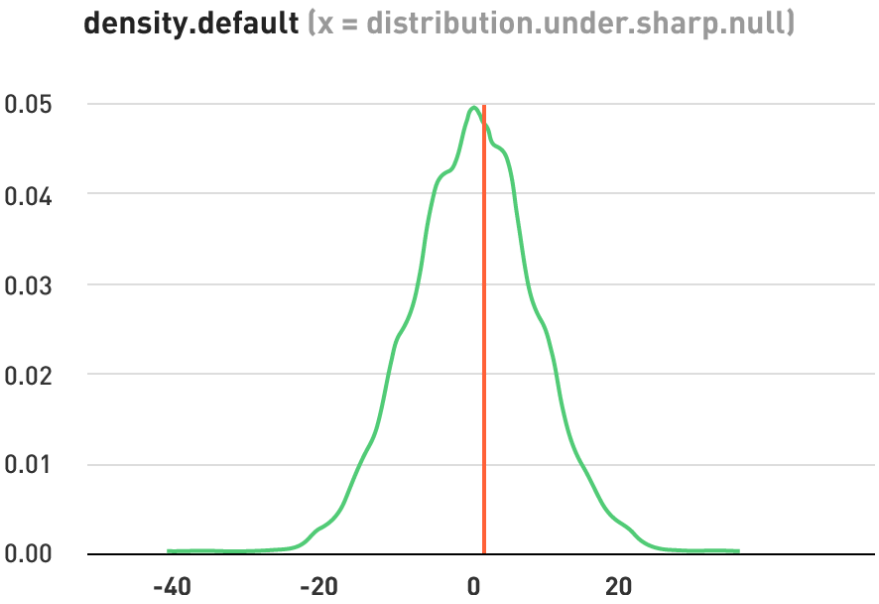
```
est.ate(outcomes, randomize())
distribution.under.sharp.null <- replicate(5000, est.ate(outcomes,
randomize()))
plot(density(distribution.under.sharp.null))
```

Distribution Under the Sharp Null




```
est.ate(outcomes, randomize())
distribution.under.sharp.null <- replicate(5000, est.ate(outcomes,
randomize()))
plot(density(distribution.under.sharp.null))
```

Distribution Under the Sharp Null

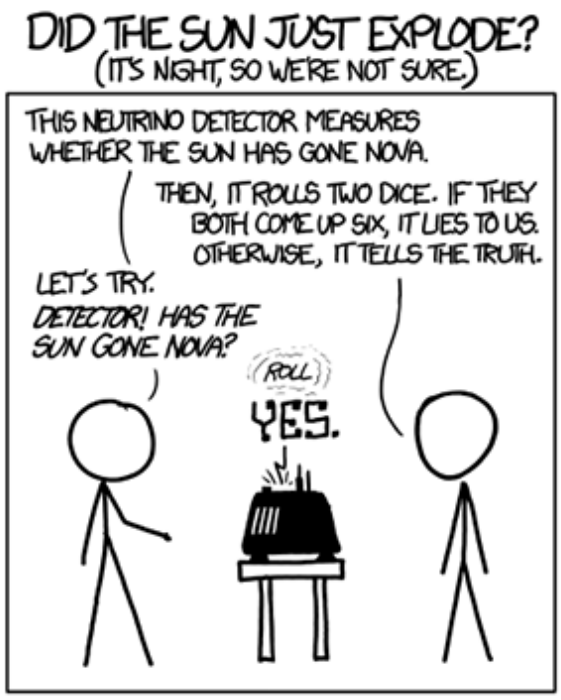



```
est.ate(outcomes, randomize())
distribution.under.sharp.null <- replicate(5000, est.ate(outcomes,
randomize()))
plot(density(distribution.under.sharp.null))
```

Size of the Observed Difference

- The p-value.
- How often did I get randomizations under the sharp null where the estimate was larger than my actual estimate?
- For each, is it larger than the average treatment effect estimate?
- This is a sampling distribution.
- How big is my estimate relative to the distribution of estimates?
- 41.2% chance we'd get a treatment effect, even if there isn't one.
- Later we'll show how to derive p-values with regression.

```
plot(density(distribution.under.sharp.null))
abline(v=ate) #Add a vertical line at our estimate
mean(ate < distribution.under.sharp.null): [1] 0.412
```



p-Values

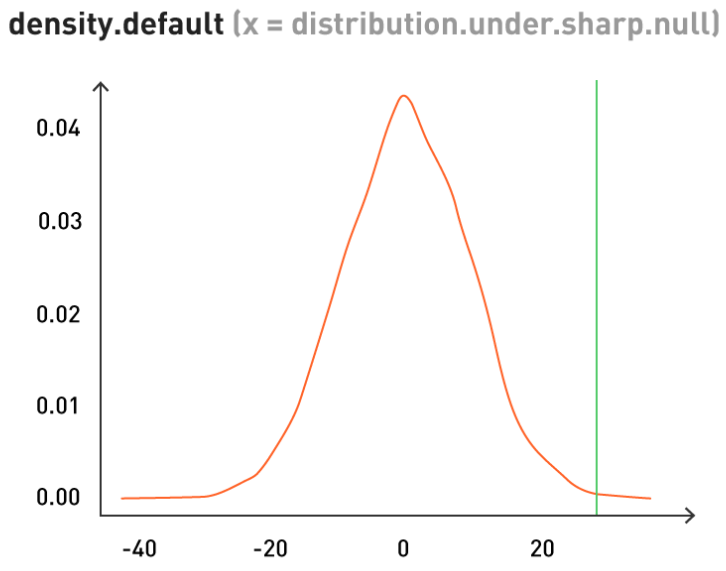
- If the treatment had no effect, how likely is it we would find a difference this extreme by chance?
    - What is the difference between the mean in the control and treatment groups?
    - Different from how likely it is the treatment has an effect
- Convention is to reject the null with p-value under 0.05.
- p-values don't tell you for sure that the treatment has an effect.
- They just tell you how likely it is you would have gotten that result by chance.
- The sampling distribution tells us how large the differences are we find by chance.
- Can find p-values < 0.05 even when the null hypothesis is correct.

Simulating an Experiment With a Large Effect

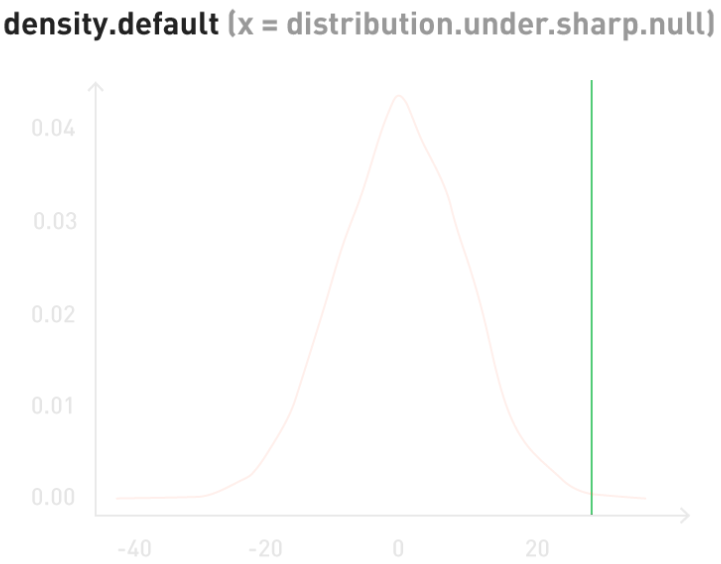
- Vector of outcomes and control
- 40-row table with potential outcomes in control and treatment, with a difference of 25

```
po.treatment <- po.control + 25
po.control: [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 51 52
53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70
po.treatment: [1] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
44 45 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95
treatment <- randomize()
outcomes <- po.treatment * treatment + po.control*(1-treatment)
outcomes: [1] 1 27 28 29 30 31 7 8 9 10 11 12 13 39 40 16 17 18 19 45 76
52 78 54 80 56 82 83 84 85 86 62 63 89 90 66 67 68 94 95
```

Estimating Average Treatment Effect



Estimating Average Treatment Effect



Estimating Average Treatment Effect

- Much bigger than experiment with no treatment effect.
- Sharp null hypothesis: For every unit, there is no effect.
- How likely to get 28.5 even with no treatment effect?
- p-value.
- 1 in 1000 chance we would reach estimate of this size if there is no effect.

```
ate <- est.ate(outcomes, treatment)
ate: [1] 28.5
distribution.under.sharp.null <- replicate( 5000, est.ate(outcomes,
randomize()) )
distribution.under.sharp.null <- replicate( 5000, est.ate(outcomes,
randomize()) )
plot(density(distribution.under.sharp.null))
abline(v=ate)
mean(ate < distribution.under.sharp.null): [1] 0.001
```

Statistical Power

DATASCI W241

Experiments and Causality

Detecting Nonzero Treatment Effects: Treatment Effect of 10

Detecting Nonzero Treatment Effects: Treatment Effect of 10

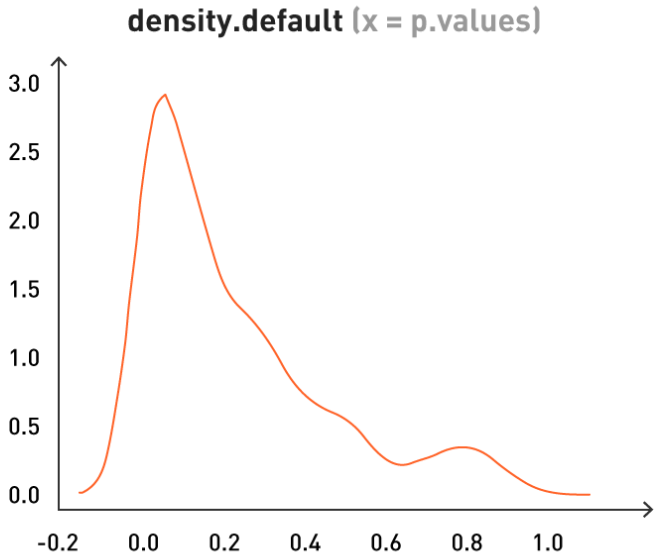
```
#assume treatment effect of 10
```

Detecting Nonzero Treatment Effects: Treatment Effect of 10

```
#assume treatment effect of 10
> p.values <- replicate(500, simulate.study(10))
> plot(density(p.values))
```

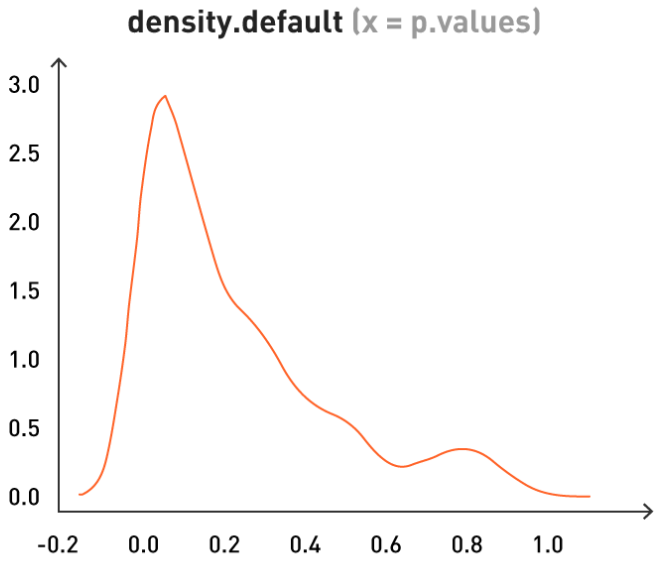
Detecting Nonzero Treatment Effects: Treatment Effect of 10

```
#assume treatment effect of 10
> p.values <- replicate(500, simulate.study(10))
> plot(density(p.values))
> mean(p.values < 0.05) #somewhat likely to
detect this effect
```



Detecting Nonzero Treatment Effects: Treatment Effect of 10

```
#assume treatment effect of 10
> p.values <- replicate(500, simulate.study(10))
> plot(density(p.values))
> mean(p.values < 0.05) #somewhat likely to
detect this effect
[1] 0.288
```



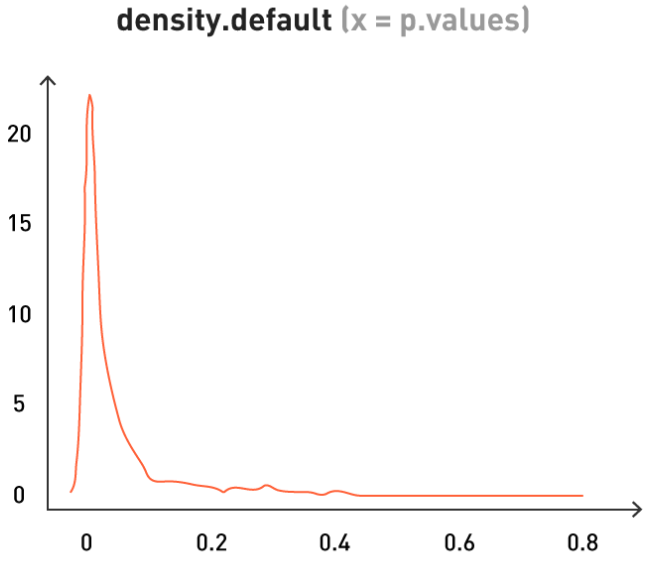
Detecting Nonzero Treatment Effects: Treatment Effect of 20

Detecting Nonzero Treatment Effects: Treatment Effect of 20

```
> p.values <- replicate(500, simulate.study(20))
```

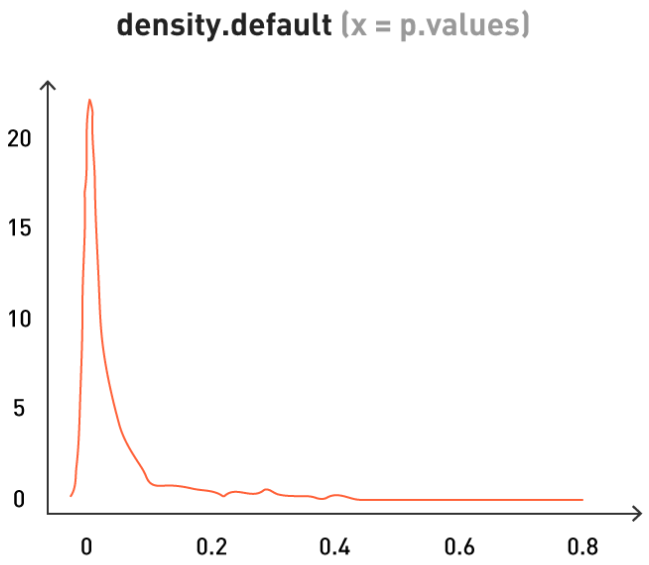
Detecting Nonzero Treatment Effects: Treatment Effect of 20

```
> p.values <- replicate(500, simulate.study(20))
> plot(density(p.values))
> mean(p.values < 0.05) #very likely to detect
this effect
```



Detecting Nonzero Treatment Effects: Treatment Effect of 20

```
> p.values <- replicate(500, simulate.study(20))
> plot(density(p.values))
> mean(p.values < 0.05) #very likely to detect
this eff ect
[1] 0.764
```



Increasing Statistical Power

Increasing Statistical Power

- Power increases with

Increasing Statistical Power

- Power increases with
  - Size of effect (larger effects easier to detect)

Increasing Statistical Power

- Power increases with
  - Size of effect (larger effects easier to detect)
  - Square root of sample size

Increasing Statistical Power

- Power increases with
  - Size of effect (larger effects easier to detect)
  - Square root of sample size
    - To detect an effect twice as small, sample size four times larger

Increasing Statistical Power

- Power increases with
  - Size of effect (larger effects easier to detect)
  - Square root of sample size
    - To detect an effect twice as small, sample size four times larger
- Statistical power of a test: probability that it correctly rejects the null hypothesis

Increasing Statistical Power

- Power increases with
  - Size of effect (larger effects easier to detect)
  - Square root of sample size
    - To detect an effect twice as small, sample size four times larger
- Statistical power of a test: probability that it correctly rejects the null hypothesis
  - Baseline statistical power: 0.05

Concentrated Tests

Concentrated Tests

- Suppose the FDA is testing effect of soybeans on estrogen levels.

Concentrated Tests

- Suppose the FDA is testing effect of soybeans on estrogen levels.
  - Study 1: Give one soybean each to a million people.

Concentrated Tests

- Suppose the FDA is testing effect of soybeans on estrogen levels.
    - Study 1: Give one soybean each to a million people.
    - Study 2: Give 10 soybeans each to 10,000 people.

Concentrated Tests

- Suppose the FDA is testing effect of soybeans on estrogen levels.
    - Study 1: Give one soybean each to a million people.
    - Study 2: Give 10 soybeans each to 10,000 people.
        - These two studies have the same statistical power.

Concentrated Tests

- Suppose the FDA is testing effect of soybeans on estrogen levels.
    - Study 1: Give one soybean each to a million people.
    - Study 2: Give 10 soybeans each to 10,000 people.
        - These two studies have the same statistical power.
- It is usually better to decrease the sample size and give a higher "dosage" to the treatment group.

Concentrated Tests

- Suppose the FDA is testing effect of soybeans on estrogen levels.
    - Study 1: Give one soybean each to a million people.
    - Study 2: Give 10 soybeans each to 10,000 people.
        - These two studies have the same statistical power.
- It is usually better to decrease the sample size and give a higher "dosage" to the treatment group.
- Concentrated tests increase statistical power by exposing a smaller treatment group to a more potent intervention.

Decreasing Statistical Power

Decreasing Statistical Power

- Power decreases with

Decreasing Statistical Power

- Power decreases with
  - Variation in outcome (larger differences by chance)

Decreasing Statistical Power

- Power decreases with
  - Variation in outcome (larger differences by chance)
- Standard deviation of distribution of outcome (crucial relationship to statistical power)

Decreasing Statistical Power

- Power decreases with
  - Variation in outcome (larger differences by chance)
- Standard deviation of distribution of outcome (crucial relationship to statistical power)
- Key statistic: ratio of true treatment effect to standard error of estimated effect

Decreasing Statistical Power

- Power decreases with
  - Variation in outcome (larger differences by chance)
- Standard deviation of distribution of outcome (crucial relationship to statistical power)
- Key statistic: ratio of true treatment effect to standard error of estimated effect
  - Standard error of estimated effect determined by square root of sample size and variation of outcome.

Confidence Intervals

- Confidence interval: "95% chance the true effect is in this range."
- Obtain using regression: estimate ±1.96 × standard error.

Confidence Interval Example

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	38.300	5.851	6.546	1.2e-07	***
treatment	**19.400**	**8.274**	2.345	0.0244	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- summary(lm(outcomes ~ treatment)) #Regression output. Allows us to compute confidence interval.
- In order to say there is a 95% chance we're in this range...
  - 19.4 +/- 1.96 * 8.274

Function: Estimate in Confidence Interval

```
estimate.in.confidence.interval <- function() {
+ true.effect <- 25
+ #Simulate outcomes
+ po.control <- c(seq(from = 1, to = 20), seq(from = 51, to = 70))
+ po.treatment <- po.control + true.effect
+ treatment <- randomize()
+ outcomes <- po.treatment * treatment + po.control*(1-treatment)
+ #Run regression
+ regression <- summary(lm(outcomes ~ treatment))
+ estimate <- regression$coefficients[2,1]
+ standard.error <- regression$coefficients[2,2]
+ lower.bound <- estimate - standard.error * 1.96
+ upper.bound <- estimate + standard.error * 1.96
+ #Is estimate in CI?
+ estimate.in.ci <- lower.bound < true.effect & upper.bound >
true.effect
+ #Expect estimate to be within range 95% of the time
+ return(estimate.in.ci)
}
mean(replicate(10000, estimate.in.confidence.interval()))
[1] 0.9427
```

We get closer to 95% with more permutations.

Regression vs. Permutation Inference

- In samples > 50, regression gives same answers as permutation inference, much more quickly.
- Permutation inference is useful for:
    - Understanding intuition behind uncertainty estimates
    - Blocking and clustering (next week)

Recap for the Week

- A sampling distribution is a distribution of estimates we would receive by chance were there no effect.
- This tells us how likely we are to get estimates as large as the estimates we got by chance. This is a p-value.
- p-values tell us P(data | we're wrong), not P(we're right | data).
- p-values can often reflect statistical power or precision.
- Improve precision by increasing sample size and decreasing noise.
- Confidence intervals represent our uncertainty around the size of an effect. Our "best guess" is the point estimate.