Gov 50: 22. More Hypothesis testing

Matthew Blackwell

Harvard University

Roadmap

1. Reviewing hypothesis testing
2. Issues with hypothesis testing
3. Power Analyses

1/ Reviewing hypothesis testing

Difference-in-means

```
library(gov50data)
trains <- trains |>
  mutate(treated = if_else(treatment == 1, "Treated", "Untreated"))
trains
```

```
## # A tibble: 115 x 15
##      age  male income white college usborn treatment ideol~1
##    <dbl> <dbl>  <dbl> <dbl>   <dbl>  <dbl>     <dbl>   <dbl>
##  1    31     0 135000     1       1      1         1       3
##  2    34     0 105000     1       1      0         1       4
##  3    63     1 135000     1       1      1         1       2
##  4    45     1 300000     1       1      1         1       4
##  5    55     1 135000     1       1      1         0       2
##  6    37     0  87500     1       1      1         1       5
##  7    53     0  87500     1       0      1         0       5
##  8    36     1 135000     1       1      1         1       4
##  9    54     0 105000     1       0      1         0       3
## 10    42     1 135000     1       1      1         1       4
## # ... with 105 more rows, 7 more variables:
## #   numberim.pre <dbl>, numberim.post <dbl>,
## #   remain.pre <dbl>, remain.post <dbl>, english.pre <dbl>,
## #   english.post <dbl>, treated <chr>, and abbreviated
```

Calculating the ATE

```
library(infer)
ate <- trains |>
  specify(numberim.post ~ treated) |>
  calculate(stat = "diff in means",
            order = c("Treated", "Untreated"))
ate
```

```
## Response: numberim.post (numeric)
## Explanatory: treated (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.383
```

Difference in means hypotheses

Hypotheses:

$$H_0 : \mu_T - \mu_C = 0$$
$$H_1 : \mu_T - \mu_c \neq 0$$

Observed difference in means:

$$\widehat{ATE} = \overline{Y}_T - \overline{Y}_C$$

Difference in means hypotheses

Hypotheses:

$$H_0 : \mu_T - \mu_C = 0$$
$$H_1 : \mu_T - \mu_c \neq 0$$

Observed difference in means:

$$\widehat{ATE} = \overline{Y}_T - \overline{Y}_C$$

How can we approximate the **null distribution**? **Permute** the outcome/treatment variables.

Permuting the treatment

Let's do 2 permutations to see how things vary:

```
set.seed(02138)
perm <- trains |>
  specify(numberim.post ~ treated) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000,
           type = "permute")
```

`generate(type = "permute")` shuffles to the outcomes, keeping treatment the same:

```
perm |> filter(replicate == 1)
```

```
## # A tibble: 115 x 3
## # Groups:   replicate [1]
##   numberim.post treated   replicate
##           <dbl> <fct>         <int>
##  1            3 Treated           1
##  2            2 Treated           1
##  3            5 Treated           1
##  4            3 Treated           1
##  5            3 Untreated         1
##  6            3 Treated           1
##  7            2 Untreated         1
##  8            2 Treated           1
##  9            3 Untreated         1
## 10            3 Treated           1
## # ... with 105 more rows
```

```
perm |> filter(replicate == 2)
```

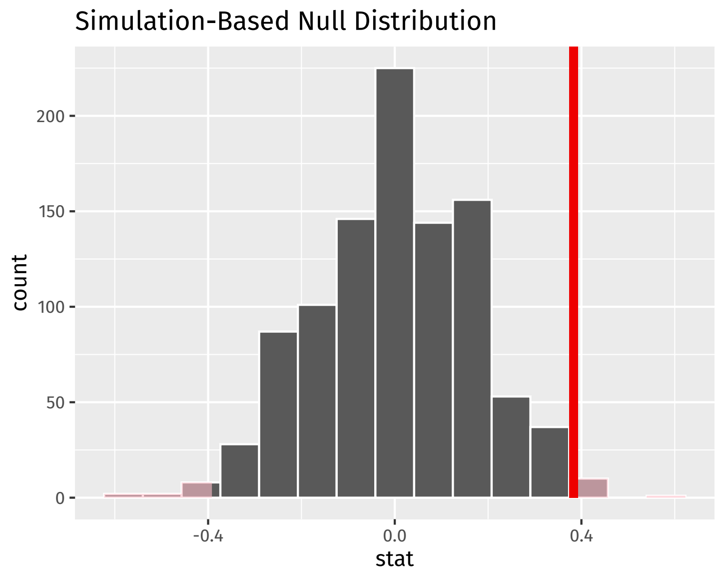
```
## # A tibble: 115 x 3
## # Groups:   replicate [1]
##   numberim.post treated   replicate
##           <dbl> <fct>         <int>
##  1            2 Treated           2
##  2            3 Treated           2
##  3            3 Treated           2
##  4            3 Treated           2
##  5            3 Untreated         2
##  6            4 Treated           2
##  7            2 Untreated         2
##  8            3 Treated           2
##  9            3 Untreated         2
## 10            2 Treated           2
## # ... with 105 more rows
```

Null distribution

The distribution of the differences-in-means under permutation will be mean 0 because shuffling the outcomes means that the outcomes in each permutation's treated and control group are coming from the same distribution.

```
null_dist <- trains |>
  specify(numberim.post ~ treated) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000,
           type = "permute") |>
  calculate(stat = "diff in means", order = c("Treated", "Untreated"))
```

```r
null_dist |>
  visualize() +
  shade_p_value(obs_stat = ate, direction = "both")
```



Interpreting p-values

```
get_p_value(null_dist, obs_stat = ate, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.022
```

Interpreting p-values

```
get_p_value(null_dist, obs_stat = ate, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.022
```

Hypotheses:

$$H_0 : \mu_T - \mu_C = 0$$
$$H_1 : \mu_T - \mu_c \neq 0$$

Observed difference in means:

$$\widehat{ATE} = \overline{Y}_T - \overline{Y}_C$$

Interpreting p-values

```r
get_p_value(null_dist, obs_stat = ate, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.022
```

Hypotheses:

$$H_0 : \mu_T - \mu_C = 0$$
$$H_1 : \mu_T - \mu_c \neq 0$$

Observed difference in means:

$$\widehat{ATE} = \overline{Y}_T - \overline{Y}_C$$

p-value: probability of an estimated ATE as big as $|\widehat{ATE}|$ by random chance if there is no treatment effect.

Rejecting the null

Decision rule: “reject the null if the p-value is below the **test level** α”

Rejecting the null

Decision rule: "reject the null if the p-value is below the **test level** α"

Rejecting the null in two-sample tests: there is a true difference in means.

Rejecting the null

Decision rule: “reject the null if the p-value is below the **test level** α”

Rejecting the null in two-sample tests: there is a true difference in means.

Test level α controls the amount of false positives:

	Null False (True difference)	Null True (No true difference)
Reject Null	True Positive	False Positive (Type I error)
Retain Null	False Negative (Type II error)	True Negative

Tests and confidence intervals

- There is a deep connection between confidence intervals and tests.

Tests and confidence intervals

- There is a deep connection between confidence intervals and tests.
- Any value outside of a $100 \times (1 - \alpha)\%$ confidence interval would have a p-value less than α if we tested it as the null hypothesis.

Tests and confidence intervals

- There is a deep connection between confidence intervals and tests.
- Any value outside of a $100 \times (1 - \alpha)\%$ confidence interval would have a p-value less than α if we tested it as the null hypothesis.
    - 95% CI for social pressure experiment: $[0.016, 0.124]$

Tests and confidence intervals

- There is a deep connection between confidence intervals and tests.
- Any value outside of a $100 \times (1 - \alpha)\%$ confidence interval would have a p-value less than α if we tested it as the null hypothesis.
    - 95% CI for social pressure experiment: $[0.016, 0.124]$
    - $\rightsquigarrow$ p-value for $H_0 : \mu_T - \mu_C = 0$ less than 0.05.

Tests and confidence intervals

- There is a deep connection between confidence intervals and tests.
- Any value outside of a $100 \times (1 - \alpha)\%$ confidence interval would have a p-value less than α if we tested it as the null hypothesis.
    - 95% CI for social pressure experiment: $[0.016, 0.124]$
    - $\rightsquigarrow$ p-value for $H_0 : \mu_T - \mu_C = 0$ less than 0.05.
- Confidence intervals are all of the null hypotheses we **can't reject** with a test.

CI in the trains example

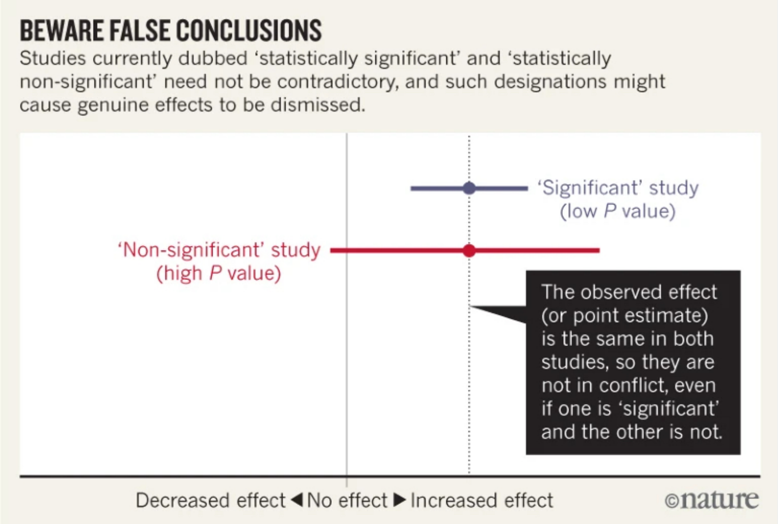
```
trains |>
  specify(numberim.post ~ treated) |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "diff in means",
            order = c("Treated", "Untreated")) |>
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0893    0.698
```

2/ Issues with hypothesis testing

Significant vs not significant

The difference between statistically significant and not statistically significant is itself not statistically significant:



What kind of significance

There are different types of significance that don't all have to be true together:

1. **Statistical significance:** we can reject the null of no effect.

What kind of significance

There are different types of significance that don't all have to be true together:

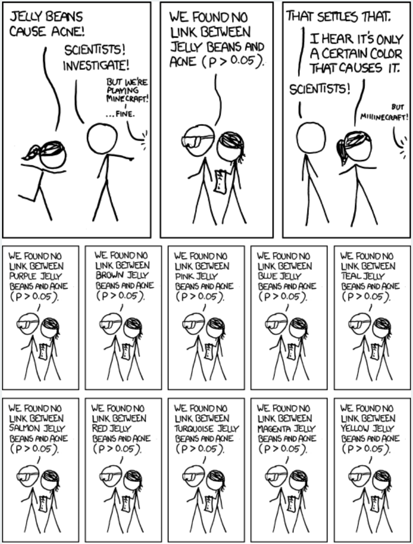
1. **Statistical significance:** we can reject the null of no effect.
2. **Causal significance**: we can interpret our estimated difference in means as a causal effect.

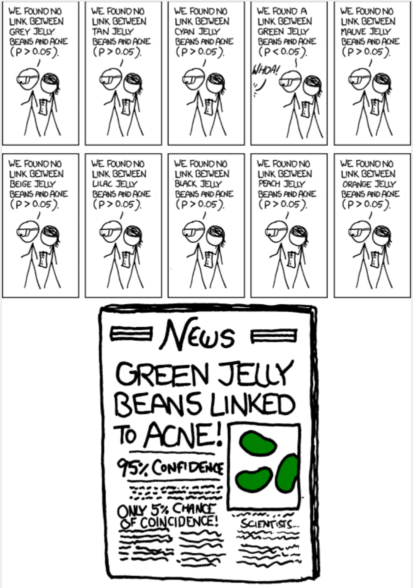
What kind of significance

There are different types of significance that don't all have to be true together:

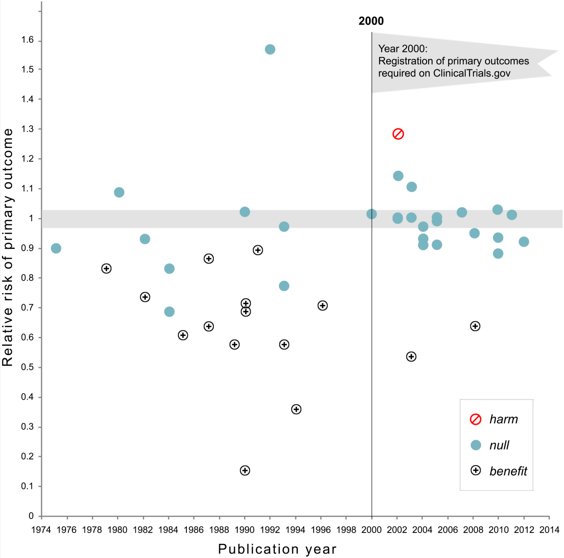
1. **Statistical significance:** we can reject the null of no effect.
2. **Causal significance**: we can interpret our estimated difference in means as a causal effect.
3. **Practical significance**: the estimated effect is meaningfully large.

p-hacking


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

p-hacking



2000
Year 2000:
Registration of primary outcomes
required on ClinicalTrials.gov
Relative risk of primary outcome
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014
Publication year
harm
null
benefit

3/ Power Analyses

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why did Gerber, Green, and Larimer use sample sizes of 38,000 for each treatment condition?

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why did Gerber, Green, and Larimer use sample sizes of 38,000 for each treatment condition?
- Choose the sample size to ensure that you can *detect* what you think might be the true treatment effect:

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why did Gerber, Green, and Larimer use sample sizes of 38,000 for each treatment condition?
- Choose the sample size to ensure that you can *detect* what you think might be the true treatment effect:
  - Small effect sizes (half percentage point) will require huge n

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why did Gerber, Green, and Larimer use sample sizes of 38,000 for each treatment condition?
- Choose the sample size to ensure that you can *detect* what you think might be the true treatment effect:
  - Small effect sizes (half percentage point) will require huge n
  - Large effect sizes (10 percentage points) will require smaller n

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why did Gerber, Green, and Larimer use sample sizes of 38,000 for each treatment condition?
- Choose the sample size to ensure that you can *detect* what you think might be the true treatment effect:
    - Small effect sizes (half percentage point) will require huge *n*
    - Large effect sizes (10 percentage points) will require smaller *n*
- **Detect** here means "reject the null of no effect"

Power of a test

- **Definition** The **power** of a test is the probability that a test rejects the null.

Power of a test

- **Definition** The **power** of a test is the probability that a test rejects the null.
    - Probability that we reject given some specific value of the parameter

Power of a test

- **Definition** The **power** of a test is the probability that a test rejects the null.
    - Probability that we reject given some specific value of the parameter
    - Power = $1 - \mathbb{P}(\text{Type II error})$

Power of a test

- **Definition** The **power** of a test is the probability that a test rejects the null.
  - Probability that we reject given some specific value of the parameter
  - Power = $1 - \mathbb{P}(\text{Type II error})$
  - Better tests = higher power.

Power of a test

- **Definition** The **power** of a test is the probability that a test rejects the null.
    - Probability that we reject given some specific value of the parameter
    - Power = $1 - \mathbb{P}(\text{Type II error})$
    - Better tests = higher power.
- If we fail to reject a null hypothesis, two possible states of the world:

Power of a test

- **Definition** The **power** of a test is the probability that a test rejects the null.
    - Probability that we reject given some specific value of the parameter
    - Power = $1 - \mathbb{P}(\text{Type II error})$
    - Better tests = higher power.
- If we fail to reject a null hypothesis, two possible states of the world:
    - Null is true (no treatment effect)

Power of a test

- **Definition** The **power** of a test is the probability that a test rejects the null.
    - Probability that we reject given some specific value of the parameter
    - Power = $1 - \mathbb{P}(\text{Type II error})$
    - Better tests = higher power.
- If we fail to reject a null hypothesis, two possible states of the world:
    - Null is true (no treatment effect)
    - Null is false (there is a treatment effect), but test had low power.

Why care about power?

- Imagine you are a company being sued for racial discrimination in hiring.

Why care about power?

- Imagine you are a company being sued for racial discrimination in hiring.
- Judge forces you to conduct hypothesis test:

Why care about power?

- Imagine you are a company being sued for racial discrimination in hiring.
- Judge forces you to conduct hypothesis test:
    - Null hypothesis is that hiring rates for white and black people are equal, $H_0 : \mu_w - \mu_b = 0$

Why care about power?

- Imagine you are a company being sued for racial discrimination in hiring.
- Judge forces you to conduct hypothesis test:
  - Null hypothesis is that hiring rates for white and black people are equal, $H_0 : \mu_w - \mu_b = 0$
  - You sample 10 hiring records of each race, conduct hypothesis test and fail to reject null.

Why care about power?

- Imagine you are a company being sued for racial discrimination in hiring.
- Judge forces you to conduct hypothesis test:
    - Null hypothesis is that hiring rates for white and black people are equal, $H_0 : \mu_w - \mu_b = 0$
    - You sample 10 hiring records of each race, conduct hypothesis test and fail to reject null.
- Say to judge, “look we don’t have any racial discrimination”! What’s the problem?

Power analysis procedure

- Power can help guide the choice of sample size through a **power analysis**.

Power analysis procedure

- Power can help guide the choice of sample size through a **power analysis**.
    - Calculate how likely we are to reject different possible treatment effects at different sample sizes.

Power analysis procedure

- Power can help guide the choice of sample size through a **power analysis**.
    - Calculate how likely we are to reject different possible treatment effects at different sample sizes.
    - **Can be done before the experiment**: which effects will I be able to detect with high probability at my n?

Power analysis procedure

- Power can help guide the choice of sample size through a **power analysis**.
    - Calculate how likely we are to reject different possible treatment effects at different sample sizes.
    - **Can be done before the experiment**: which effects will I be able to detect with high probability at my n?
- Steps to a power analysis:

Power analysis procedure

- Power can help guide the choice of sample size through a **power analysis**.
    - Calculate how likely we are to reject different possible treatment effects at different sample sizes.
    - **Can be done before the experiment**: which effects will I be able to detect with high probability at my n?
- Steps to a power analysis:
    - Pick some hypothetical effect size, $\mu_T - \mu_C = 0.05$

Power analysis procedure

- Power can help guide the choice of sample size through a **power analysis**.
    - Calculate how likely we are to reject different possible treatment effects at different sample sizes.
    - **Can be done before the experiment**: which effects will I be able to detect with high probability at my n?
- Steps to a power analysis:
    - Pick some hypothetical effect size, $\mu_T - \mu_C = 0.05$
    - Calculate the distribution of T under that effect size.

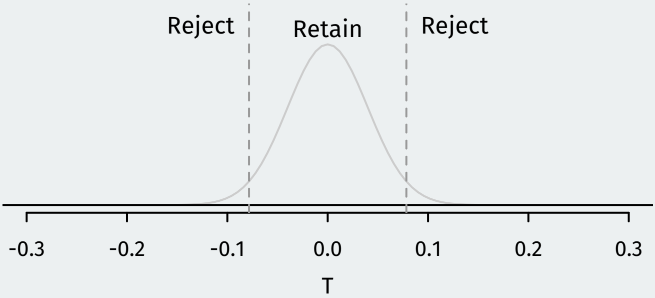
Power analysis procedure

- Power can help guide the choice of sample size through a **power analysis**.
  - Calculate how likely we are to reject different possible treatment effects at different sample sizes.
  - **Can be done before the experiment**: which effects will I be able to detect with high probability at my n?
- Steps to a power analysis:
  - Pick some hypothetical effect size, $\mu_T - \mu_C = 0.05$
  - Calculate the distribution of T under that effect size.
  - Calculate the probability of rejecting the null under that distribution.

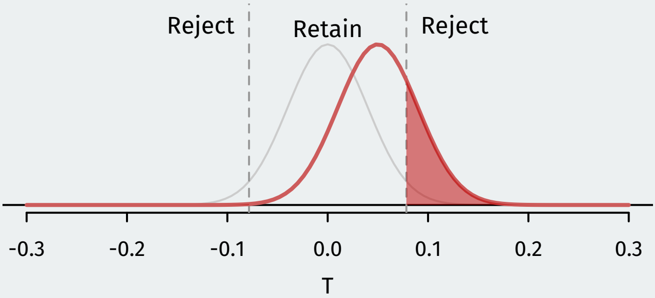
Power analysis procedure

- Power can help guide the choice of sample size through a **power analysis**.
  - Calculate how likely we are to reject different possible treatment effects at different sample sizes.
  - **Can be done before the experiment**: which effects will I be able to detect with high probability at my n?
- Steps to a power analysis:
  - Pick some hypothetical effect size, $\mu_T - \mu_C = 0.05$
  - Calculate the distribution of T under that effect size.
  - Calculate the probability of rejecting the null under that distribution.
  - Repeat for different effect sizes.

Power graph

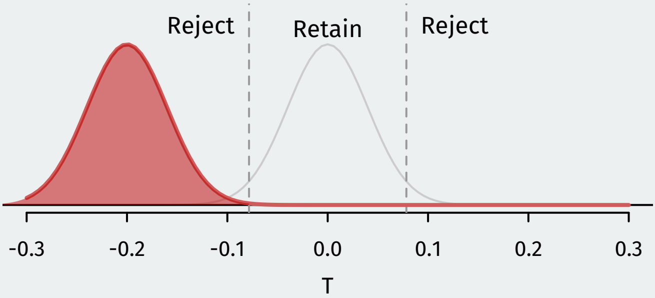


Power graph



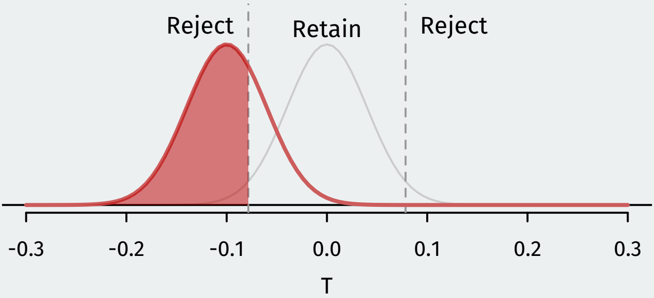
Assumed treatment effect = 0.05 and power = 0.24.

Power graph



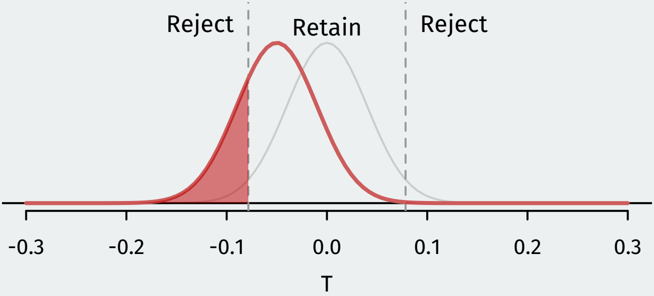
Assumed treatment effect = -0.2 and power = 0.999.

Power graph



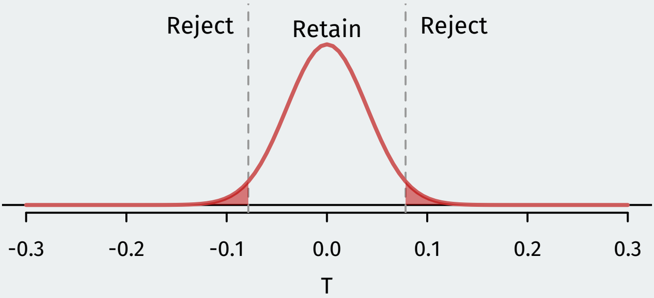
Assumed treatment effect = -0.1 and power = 0.705.

Power graph



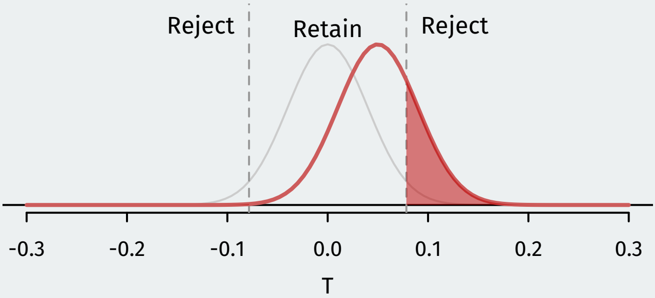
Assumed treatment effect = -0.05 and power = 0.24.

Power graph



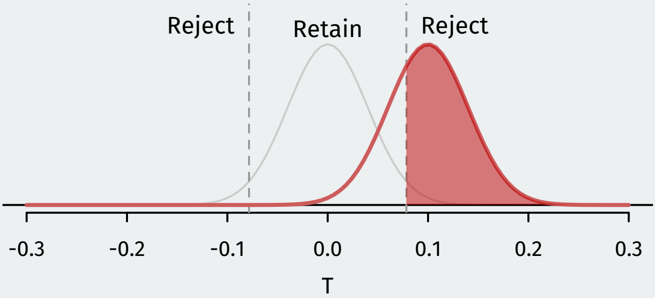
Assumed treatment effect $= 0$ and power $= 0.05$.

Power graph



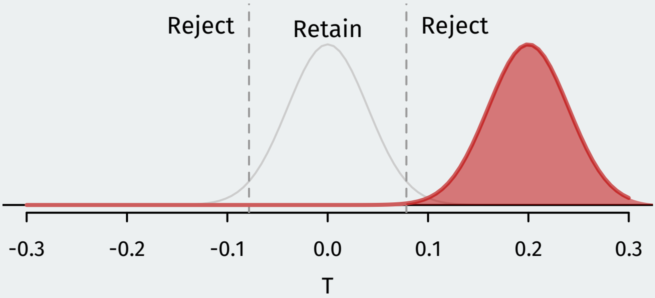
Assumed treatment effect = 0.05 and power = 0.24.

Power graph



Assumed treatment effect = 0.1 and power = 0.705.

Power graph



Assumed treatment effect = 0.2 and power = 0.999.

A power analysis

- We can calculate the power for every possible effect size and plot the resulting **power curve**:

A power analysis

- We can calculate the power for every possible effect size and plot the resulting **power curve**:
  - $n = 500$ (blue), 1000 (red), 10000 (black)

A power analysis

- We can calculate the power for every possible effect size and plot the resulting **power curve**:
  - $n = 500$ (blue), 1000 (red), 10000 (black)