Inference: Hypothesis Testing & Regression Uncertainty

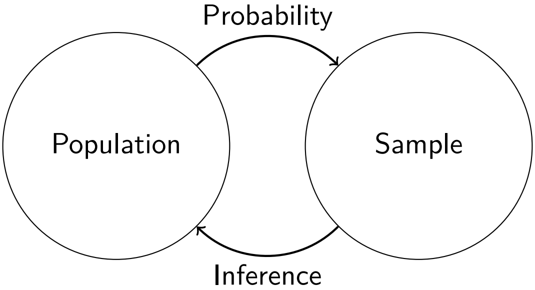
Week 12 (or 13? 14? I dunno)

Prof. Weldzius

Villanova University

Slides Updated: 2025-04-17

Remember our goal



- Psych! We're done with this.

The lady tasting tea

Your friend asks you to grab a tea with milk for her before meeting up and she says that she prefers tea poured before the milk. You stop by a local tea shop and ask for a tea with milk. When you bring it to her, she complains that it was prepared milk-first.

- You're skeptical that she can tell the difference, so you devise a test:
- Prepare 8 cups of tea, 4 milk-first, 4 tea-first
- Present cups to friend in a **random** order
- Ask friend to pick which 4 of the 8 were milk-first.

Assuming we know the truth

- Friend picks out all 4 milk-first cups correctly!
- Statistical thought experiment: how often would she get all 4 correct **if she were guessing randomly**?
  - Only one way to choose all 4 correct cups.
  - But 70 ways of choosing 4 cups among 8.
  - Choosing at random $\approx$ picking each of these 70 with equal probability.
- Chances of guessing all 4 correct is $\frac{1}{70} \approx 0.014$ or 1.4%.
- $\rightsquigarrow$ the guessing hypothesis might be implausible.

Statistical hypothesis testing

- Statistical hypothesis testing is a **thought experiment**.
  - Could our results just be due to random chance?
- What would the world look like **if we knew the truth**?
- *Example 1*:
  - An analyst claims that 20% of Philadelphia households are in poverty.
  - You take a sample of 900 households and find that 23% of the sample is under the poverty line.
  - Should you conclude that the analyst is wrong?
- *Example 2*:
  - Trump won 47.5% of the vote in the 2020 election.
  - Last YouGov poll of 1,363 likely voters said 44% planned to vote for Trump.
  - Could the difference between the poll and the outcome be just due to random chance?

Null and alternative hypothesis

- **Null hypothesis**: Some statement about the population parameters.
  - "Devil's advocate" position $\rightsquigarrow$ assumes what you seek to prove wrong.
  - Usually that an observed difference is due to chance.
  - Ex: poll drawn from the same population as all voters.
  - Denoted H_0
- **Alternative hypothesis**: The statement we hope or suspect is true instead of H_0.
  - It is the opposite of the null hypothesis.
  - An observed difference is real, not just due to chance.
  - Ex: polling for Trump is systematically wrong.
  - Denoted H_1 or H_a
- **Probabilistic** proof by contradiction: try to "disprove" the null.

Hypothesis testing example

- Are we polling the same population as the actual voters?
  - If so, how likely are we to see polling error this big by chance?
- What is the parameter we want to learn about?
  - True population mean of the surveys, p.
  - Null hypothesis: $H_0 : p = 0.475$ (surveys drawing from same population)
  - Alternative hypothesis: $H_1 : p \neq 0.475$
- Data: poll has $\bar{X} = 0.44$ with $n = 1363$.

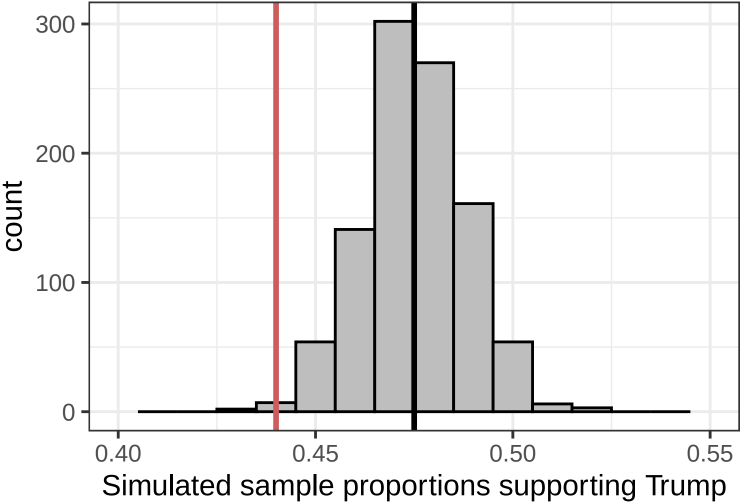
Statistical thought experiment

- If the null were true, what should the distribution of the data be?
    - X_i is 1 if respondents i will vote for Trump.
    - Under null, X_i is Bernoulli with $p = 0.475$.
    - $\sum_{i=1}^{n} X_i$ is the number in sample that will vote for Trump.
    - This sum will be Binomial with $n = 1363$ and $p = 0.475$.
- We can simulate draws from this distribution!
- Compare the distribution of proportions under the null to the observed proportion.

```
trump_voters <- rbinom(n = 1000, size = 1363, prob = 0.475)
trump_shares <- trump_voters / 1363

data.frame(trump_shares) %>%
ggplot(aes(x = trump_shares)) +
  geom_histogram(binwidth = 0.01, fill = "grey", color = "black") +
  xlim(0.4, 0.55) +
  xlab("Simulated sample proportions supporting Trump") +
  geom_vline(xintercept = 0.44, color = "indianred", linetype = "solid", si
  geom_vline(xintercept = 0.475, color = "black", linetype = "solid", size
```

Simulations of the null distribution



p-value

Definition (p-value)

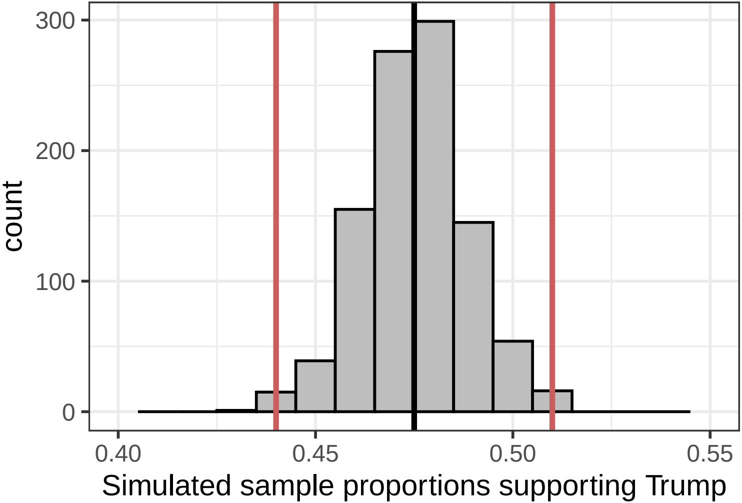
The **p-value** is the probability of observing data as or more extreme as our data under the null.

- If the null is true, how often would we expect polling errors this big?
  - Smaller p-value $\rightsquigarrow$ stronger evidence against the null
  - **NOT** the probability that the null is true!
- p-values are usually **two-sided**:
  - Observed error of 0.44 - 0.475 = -0.035 under the null.
  - p-value is probability of sample proportions being less than 0.44 **plus**
  - probability of sample proportions being greater than 0.475 + 0.035 = 0.51.

```
mean(trump_shares < 0.44) + mean(trump_shares > 0.51)
```

```
## [1] 0.008
```

The two-sided p-value



One-sided tests

- Sometimes our hypothesis is directional.
  - We only consider evidence against the null from one direction.
- Null: our polls are from the same population as actual voters
  - $H_0 : p = 0.475$
- **One-sided alternative**: polls underestimate Trump support.
  - $H_1 : p < 0.475$
- Makes the p-value one-sided:
  - What's the probability of a random sample underestimating Trump support by as much as we see in the sample?
  - Always smaller than a two-sided p-value.

```
mean(trump_shares < 0.44)
```

```
## [1] 0.005
```

Rejecting the null

- Tests usually end with a decision to reject the null or not.
- Choose a threshold below which you'll reject the null.
  - **Test level** α: the threshold for a test.
  - Decision rule: "reject the null if the p-value is below α"
  - Otherwise "fail to reject" or "retain", not "accept the null"
- Common (arbitrary) thresholds:
- $p \geq 0.1$ "not statistically significant"
- $p < 0.05$ "statistically significant"
- $p < 0.01$ "highly significant"

Testing errors

- A p-value of 0.05 says that data this extreme would only happen in 5% of repeated samples if the null were true.
  - $\rightsquigarrow$ 5% of the time we'll reject the null when it is actually true.
- Test errors:

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- Type I error because it's the worst
  - "Convicting" an innocent null hypothesis
- Type II error less serious
  - Missed out on an awesome finding

General sample means

- Earlier: hypothesis testing for a sample proportion.
  - Binary data $\rightsquigarrow$ easy setting.
  - Distribution of samples just depends on population proportion.
- This time: hypothesis testing for means of any variable.

Hypothesis testing procedure

Conducted with several steps:

1. Specify your **null** and **alternative hypotheses**
2. Choose an appropriate **test statistic** and level of test α
3. Derive the **reference distribution** of the test statistic under the null.
4. Use this distribution to calculate the **p-value**.
5. Use p-value to decide whether to reject the null hypothesis or not

- This procedure is general, but we'll focus on tests of a single population mean today.

Test statistic

Definition (Test statistic)

A **test statistic** is a function of data and possibly the null hypothesis used to adjudicate between the null and alternative hypotheses.

- Most common form for sample means is the z-statistic:

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- Put differently:

$$z = \frac{\text{observed} - \text{null guess}}{\text{SE}}$$

  - How many SEs away from the null guess is the sample mean?

- Usually replace population SD σ with sample SD $\hat{\sigma}$

Example: thermometer scores

- Social scientists often use **thermometer scores** to assess views toward groups.
    - 0-100 scale, where higher is "warmer" feeling toward group.
- You work at an advocacy group who got a survey with FT scores for transgender people.
    - $\bar{X} = 52.5$ and $\hat{\sigma} = 29.3$
    - Sample size, $n = 912$
- Co-worker Nully is weirdly insistent that these results are consistent with a population mean FT score of 50.
- Hypothesis tests to the rescue!

Calculating the test statistic

- Hypotheses:
  - $H_0 : \mu = 50$, population average is 50.
  - $H_A : \mu \neq 50$
- Test statistic:

$$Z_{\text{obs}} = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{52.3 - 50}{29.3/\sqrt{912}} \approx 2.35$$

- Observed average is 2.35 SEs away from the null!
  - Exactly how unlikely is this?

Determining the reference distribution

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- What is the distribution of Z?
- With sample proportions, we relied on the binomial distribution.
  - Doesn't work if variable is non-binary (age, income, etc)
- Central limit theorem to the rescue! In large samples and under the null:
  - $\bar{X}$ is normal with mean μ and standard deviation $\sigma/\sqrt{n}$.
  - Z will be standard normal (mean 0, SD 1)
- Large samples also justify using sample SD ($\hat{\sigma}$) in place of population SD (σ).

Finding the p-value

- **Step 4**: determine the p-value.
  - The **p-value** is the probability of observing a test statistic as extreme as Z_{obs}, if the null hypothesis is true.
  - Smaller p-values $\rightsquigarrow$ data less likely under the null $\rightsquigarrow$ null less plausible
- How to calculate?
  - We know Z is distributed standard normal $\rightsquigarrow$ use R!

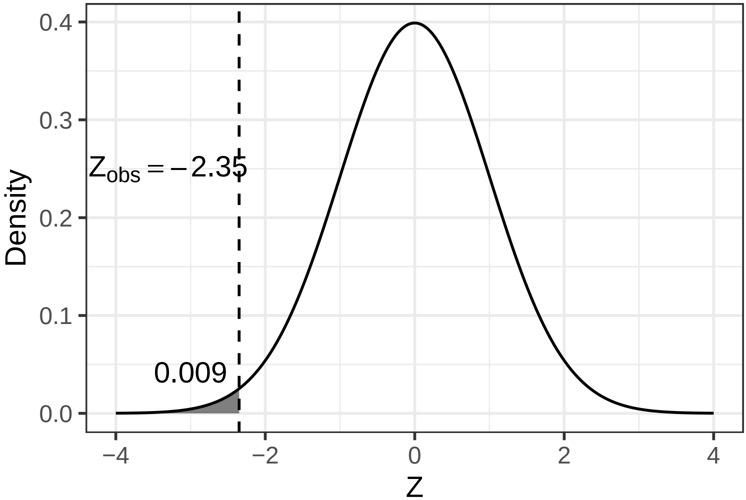
Standard normal probabilities in R

- The `pnorm(x)` function will give the probability of being less than `x` in a standard normal:

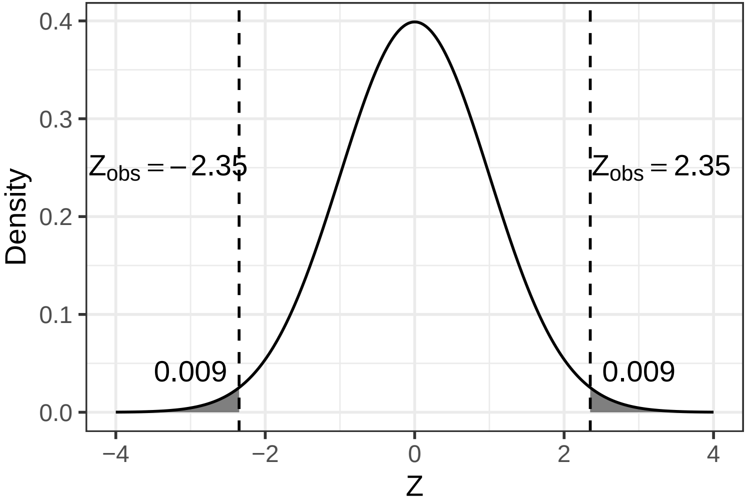
```
pnorm(-2.35)
```

```
## [1] 0.009386706
```

One-sided vs. two-sided tests



One-sided vs. two-sided tests



- two-sided p-value: 0.018.

Problem of small samples

- Central limit theorem justifies the z-test we've been doing.
  - "Sums and means of random variables tend to be normally distributed as sample sizes get big."
- What if our sample sizes are low?
  - Distribution of $\bar{X}$ will be unknown
  - $\rightsquigarrow$ can't determine p-values
  - $\rightsquigarrow$ can't get z values for confidence intervals
- Very difficult to get around this problem without more information.

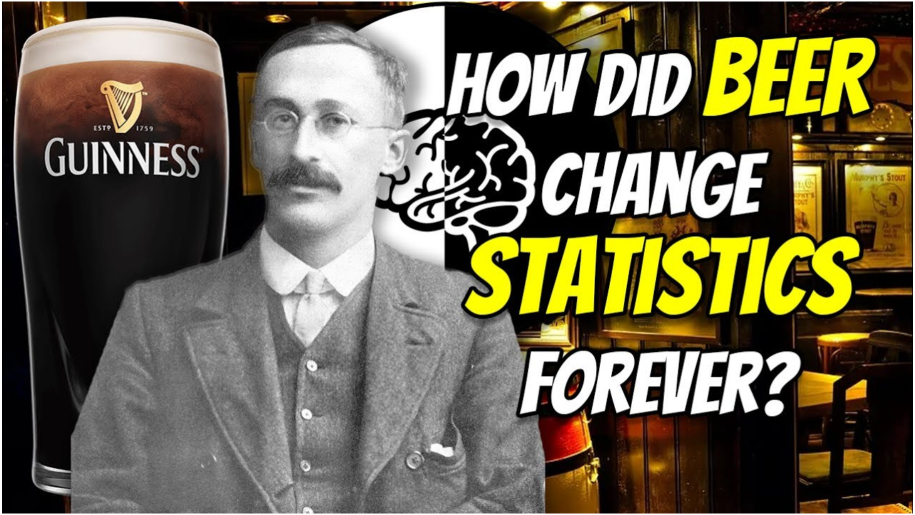
Solution to small samples?

- Common approach: assume data X_i are **normally distributed**
  - THIS IS AN ASSUMPTION, PROBABLY IS WRONG.
  - For instance, if X_i is binary, then it is very wrong.
- If true, then we can determine the distribution of the following test statistic:

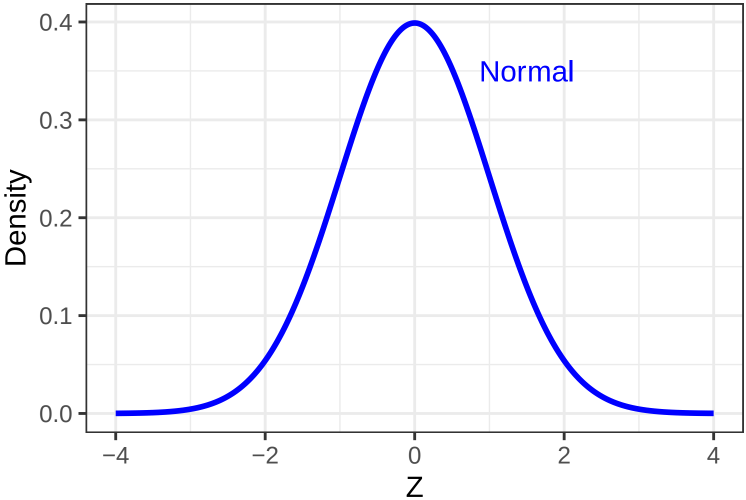
$$T = \frac{\bar{X} - \mu}{\widehat{SE}} \sim t_{n-1}$$

- T follows a Student's t distribution with $n - 1$ degrees of freedom.
  - Degrees of freedom determines the spread of the distribution.
  - Centered around 0
  - Similar to normal with fatter tails $\rightsquigarrow$ higher likelihood of extreme events.

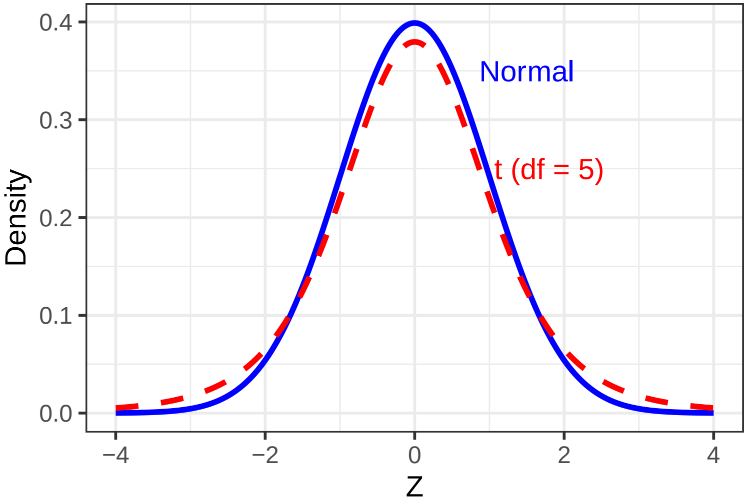
Who was Student?



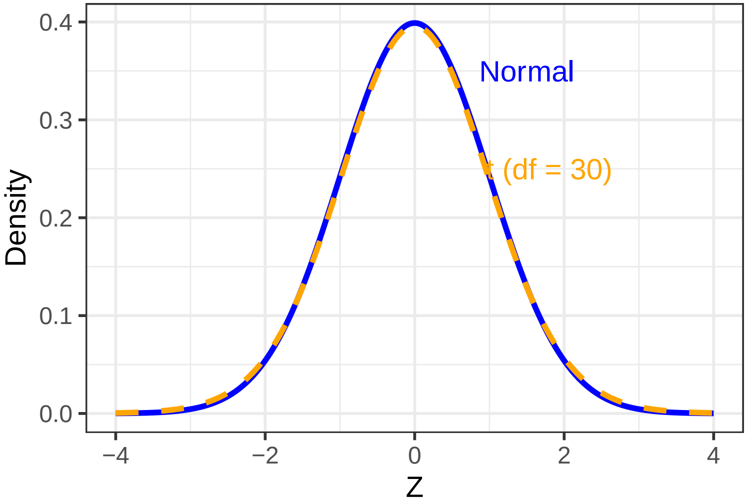
Student's t distribution



Student's t distribution



Student's t distribution



z-test vs. t-test

- z-tests are what we have seen: relies on the normal distribution.
  - Justified in large samples (roughly $n > 30$) by CLT
- t-tests rely on the t-distribution for calculating p-values.
  - Justified in small samples if data is normally distributed.
- Common practice is to use t-tests all the time because t is "conservative"
  - $\rightsquigarrow$ p-values will always be larger under t-test
  - $\rightsquigarrow$ always less likely to reject null under t
  - t-distribution converges to standard normal as $n \to \infty$
- R will almost always calculate p-values for you, so details of t-distribution aren't massively important.

Two-sample tests

- Statistical hypothesis testing is a **thought experiment**.
- What would the world look like **if we knew the truth**?
- Conducted with several steps:1. Specify your **null** and **alternative hypotheses**2. Choose an appropriate **test statistic** and level of test α3. Derive the **reference distribution** of the test statistic under the null.4. Use this distribution to calculate the **p-value**.5. Use p-value to decide whether to reject the null hypothesis or not

Recall from earlier

- We looked at hypothesis tests for means.
  - Tested null that true population mean was some value: $H_0 : \mu = \mu_0$
- Under the null hypothesis, we can determine the (approximate) distribution of the test statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

- Calculated p-values of this test statistic
- Today: generalizing to differences in means.

Social pressure example

- Back to the Social Pressure Mailer GOTV example.
  - Treatment group: mailers showing voting history of them and neighbors.
  - Control group: received nothing.
- Samples are **independent**
  - Example of dependent comparisons: **paired comparisons**

Two-sample hypotheses

- Parameter: **population ATE** $\mu_T - \mu_C$
  - μ_T: Turnout rate in the population if everyone received treatment.
  - μ_C: Turnout rate in the population if everyone received control.
- Goal: learn about the population difference in means
- Usual null hypothesis: no difference in population means (ATE $= 0$)
  - Null: $H_0 : \mu_T - \mu_C = 0$
  - Two-sided alternative: $H_1 : \mu_T - \mu_C \neq 0$
- In words: are the differences in sample means just due to chance?

Difference-in-means review

- Sample turnout rates: $\bar{X}_T = 0.37$, $\bar{X}_C = 0.30$
- Sample sizes: $n_T = 360$, $n_C = 1890$
- Estimator is the sample **difference-in-means**:

$$\widehat{ATE} = \bar{X}_T - \bar{X}_C = 0.07$$

- Standard error of difference in means of independent samples:

$$\text{SE}_{diff} = \sqrt{\text{SE}_T^2 + \text{SE}_C^2}$$

- Since turnout is binary, we can use the special proportions rule for the SEs:

$$\widehat{\text{SE}}_{\text{diff}} = \sqrt{\frac{\bar{X}_T(1-\bar{X}_T)}{n_T} + \frac{\bar{X}_C(1-\bar{X}_C)}{n_C}} = 0.028$$

CLT again and again

- $\bar{X}_T$ is a sample mean and so tends toward normal as $n_T \to \infty$
- $\bar{X}_C$ is a sample mean and so tends toward normal as $n_C \to \infty$
- $\rightsquigarrow \bar{X}_T - \bar{X}_C$ will tend toward normal as sample sizes get big.
- Using the z-transformation/standardization:

$$Z = \frac{(\bar{X}_T - \bar{X}_C) - (\mu_T - \mu_C)}{\text{SE}_{\text{diff}}} \sim N(0, 1)$$

- Same general form of the test statistic as with one sample mean:

$$\frac{\text{observed} - \text{null guess}}{\text{SE}}$$

The usual null of no difference

- Null hypothesis: $H_0 : \mu_T - \mu_C = 0$
- Test statistic:

$$Z = \frac{(\bar{X}_T - \bar{X}_C) - (\mu_T - \mu_C)}{\mathrm{SE}_{\mathrm{diff}}} = \frac{(\bar{X}_T - \bar{X}_C) - 0}{\mathrm{SE}_{\mathrm{diff}}}$$

- In large samples, we can replace true SE with an estimate:

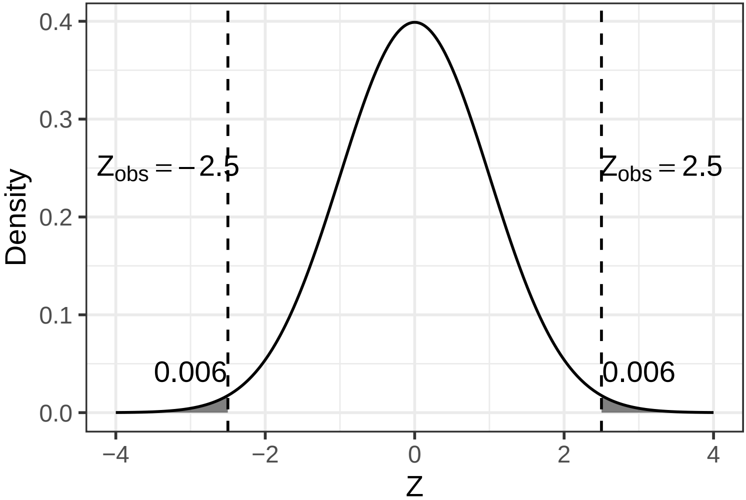
$$\widehat{SE}_{diff} = \sqrt{\widehat{\mathrm{SE}}_T^2 + \widehat{\mathrm{SE}}_C^2}$$

Calculating p-values

- Finally! Our test statistic in this sample:

$$Z = \frac{\bar{X}_T - \bar{X}_C}{\widehat{\mathsf{SE}}_{\mathsf{diff}}} = \frac{0.07}{0.028} = 2.5$$

- p-value based on a two-sided test: probability of getting a difference in means this big (or bigger) if the null hypothesis were true
- Lower p-values $\rightsquigarrow$ stronger evidence against the null.




```r
2 * pnorm(2.5, lower.tail = FALSE)
```

```
## [1] 0.01241933
```

Tests and confidence intervals

- Deep connection between confidence intervals and tests.
- A 95% CI contains all null hypotheses with p-values greater than 0.05.
  - All the nulls we couldn't reject if $\alpha = 0.05$
  - 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.