Lecture 16: Introduction to Hypothesis Testing

Announcements and reminders

- HW 3 posted! And due Monday 18 March (2 weeks)
- Feedback survey on Canvas (closes 18 March)

    https://canvas.colorado.edu/courses/24706/quizzes/53010

Previously, on CSCI 3022…

Proposition: If X is a normally distributed random variable with mean μ and standard deviation σ, then Z follows a standard normal distribution if we define:

$$Z = \frac{X - \mu}{\sigma} \qquad \text{and} \qquad X = \sigma Z + \mu$$

Fun fact: If Z is a standard normal random variable, then we can compute probabilities using the standard normal cdf

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(x)\ dx$$

A 100·(1-α)% confidence interval for the mean μ when the value of σ is known is given by

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \qquad \text{or} \qquad \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

A Thought Experiment

Example: After the introduction of the Euro, Polish mathematicians claimed that the new Belgian 1 Euro coin is not a fair coin. S'pose I hand you a Belgian 1 Euro coin. How could you decide whether or not it is fair? $p \neq 0.5$

Flip it 100 time

60? 65?

70?

$\boxed{\approx 62}$

Statistical hypotheses

Definition: A **statistical hypothesis** is a claim about the value of a parameter of a population characteristic.

Parameters: μ pop mean

σ^2 pop variance

p pop proportion

Examples:

- S'pose the recovery time of a person suffering from a disease is normally distributed with mean μ_1 and standard deviation σ_1
  Hypothesis: $\mu_1 > 10$ days

- S'pose μ_2 is the recovery time of a person suffering from the same disease, but also given some kind of new treatment.
  Hypothesis: $\mu_2 < \mu_1$

- S'pose μ_1 is the mean internet speed for Comcast and μ_2 is the mean internet speed for Century Link.
  Hypothesis: $\mu_1 \neq \mu_2$

Null vs Alternative hypotheses

In any hypothesis-testing problem, there are always **two competing hypotheses** under consideration:

1) H_0 : null hypothesis - default, our belief before we collect our data

2) H_1 or H_A : Alternative hypothesis

The objective of **hypothesis testing** is to choose, based on sampled data, between two competing hypotheses about the value of a population parameter

Classic Jury Analogy

Consider a jury in a criminal trial.

When a defendant is accused of a crime, the jury **presumes** that he or she is **not guilty**

→ Null hypothesis: **H_0 = not guilty**

The jury is then presented with **evidence**. If the evidence seems implausible under the assumption of non-guilt, we might **reject the null hypothesis** of non-guilt, and claim that the defendant is (likely) **guilty**

→ Evidence supported the Alternative hypothesis: **H_1 = guilty**

→ To do: Find P(evidence | H_0 true)
& data)

Null vs Alternative hypotheses

Is there strong evidence for the alternative hypothesis?

- The burden of proof is placed on those that believe the alternative claim

- The initially-favored claim (H_0) will not be rejected in favor of the alternative claim (H_1) unless the sample evidence provides **enough** support for the alternative

- Two possible conclusions:

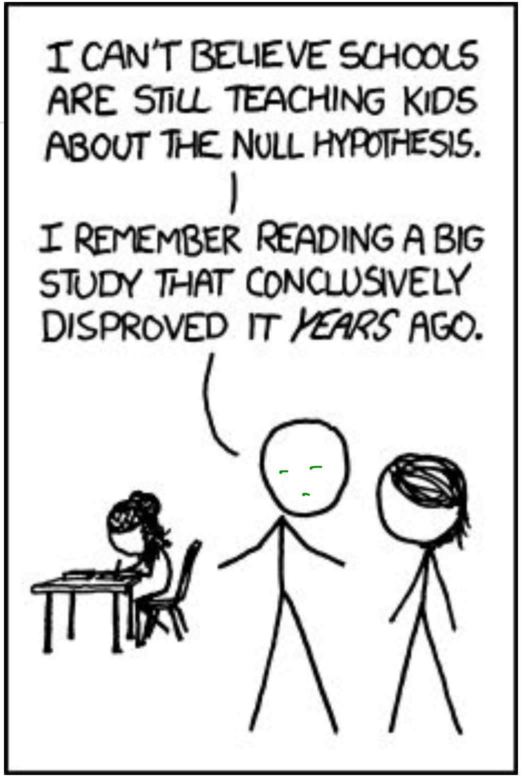
    1) Reject H_0 (in favor of H_1) if $P(\text{evidence} \mid H_0 = \text{true})$ is "low enough"

    2) Fail to reject the null hypothesis H_0

we don't say "we accept H_0"

Never prove H_0 is true/prove H_0

Null vs Alternative hypotheses



Why assume the Null Hypothesis?

- Sometimes we don't want to accept a particular assertion unless (or until) data can be shown to strongly support it

- Reluctance (cost, time, effort) to change

H_A or H_1: always what we want to find evidence in favor of

Example: S'pose a company is considering hiring a new outside advertising company to help generate traffic to their website. Under their current advertising, they get 200,000 hits/day on average. With μ denoting the true average number of hits/day they'd get using the new company's advertising, they would not want to switch companies (because it would be costly) unless evidence strongly suggested that μ exceed 200,000.

Null vs Alternative hypotheses

Example: S'pose a company is considering hiring a new outside advertising company to help generate traffic to their website. Under their current advertising, they get 200,000 hits/day on average. With μ denoting the true average number of hits/day they'd get using the new company's advertising, they would not want to switch companies (because it would be costly) unless evidence strongly suggested that μ exceed 200,000.

An appropriate problem formulation for hypothesis testing would be:

$H_0: \mu = 200{,}000$ hits/day under new ad compaign

$H_1: \mu > 200{,}000$

The conclusion that action is justified is identified with the alternative hypothesis, and it would take conclusive evidence to justify rejecting H_0 and switching to the new company.

Null vs Alternative hypotheses

$\theta = \mu_1$ or r^2 or p

The alternative to the null hypothesis H_0: $\theta = \theta_0$ will look like one of the following assertions (and variations of these)

Example

1) $\theta > \theta_0$ $\quad \mu > 200{,}000$
2) $\theta < \theta_0$ $\quad \mu < 200{,}000$
3) $\theta \neq \theta_0$ $\quad p \neq 0.5$

population is not proportion to 0.5

$H_0: p = 0.5$

$H_1: p \neq 0.5$

B/c we want to calculate:

$P(\text{data} \mid H_0 = \text{true})$

$P(\text{data} \mid p = 0.5)$

- $\boxed{\text{The equals sign is } \textbf{always} \text{ in the null hypothesis}}$
- The alternative hypothesis is the one for which we are seeking statistical evidence

some books: $H_0: p \leq .5$ $\quad$ us: $H_0: p = 0.5$

$H_1: p > .5$

Null vs Alternative hypotheses

The alternative to the null hypothesis H_0: $\theta = \theta_0$ will look like one of the following assertions (and variations of these)

1) $\theta > \theta_0$
2) $\theta < \theta_0$
3) $\theta \neq \theta_0$

- The equals sign is **always** in the null hypothesis
- The alternative hypothesis is the one for which we are seeking statistical evidence

Test statistics and evidence

$P(\text{data} \mid H_0 = \text{true})$

$p = 0.5$

Definition: A **test statistic** is a quantity derived from the sample data and calculated assuming that the null hypothesis is true. It is used in the decision about whether or not to reject the null hypothesis.

Intuition:

- We can think of the test statistics as our evidence about the competing hypotheses
- We consider the test statistic under the assumption that the null hypothesis is true by asking questions like

  How likely would we be to obtain this evidence if the null hyp were true?

$P(\text{data} \mid H_0 = \text{true})$

Test statistics and evidence

How likely would we be to obtain this evidence if the null hyp were true?

Example: To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

Test statistic : X = # flips that come up heads $\rightarrow \hat{p} = \frac{X}{n}$

Null Hypothesis : $H_0 : p = 0.5$

Alt Hyp : $H_1 : p \neq 0.5$

Test statistics and evidence

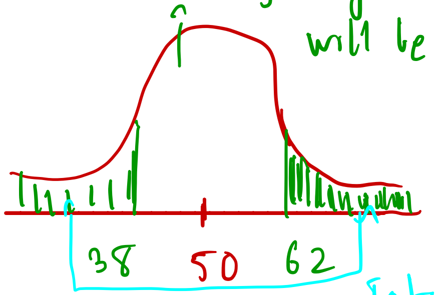
How likely would we be to obtain this evidence if the null hyp were true?

Example: To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

Question: What would it take to convince **you** that the coin is not fair?

$X = 62$ or 38 flips of heads are enough to convince us the coin is biased $(p \neq 0.5)$

anything outside this will be biased coin



Total area on the outside is significant level

Test statistics and evidence

How likely would we be to obtain this evidence if the null hyp were true?

Example: To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

Question: What would it take to convince **you** that the coin is not fair?

Let's say we wanted a 95% CI → 5% area in the tails

→ $z_{\alpha/2} = \text{ppf}(.975) = 1.96$

with a lot of flips

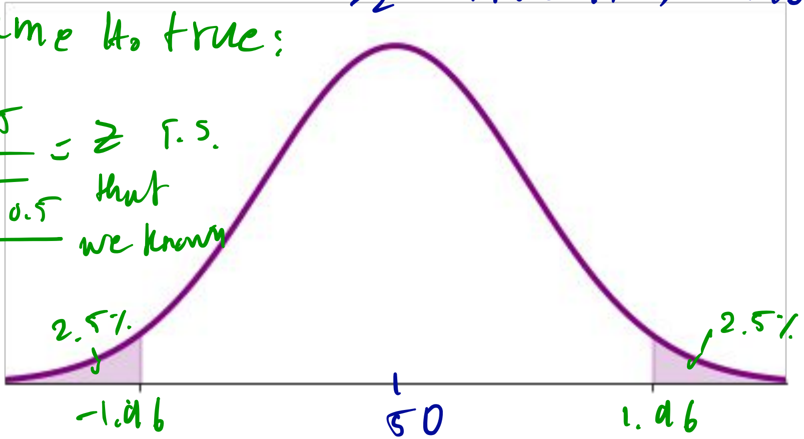
$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

↓ CLT

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Assume H_0 true:

$$\frac{\hat{p} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = z \text{ s.t. that we know}$$



Rejection regions and significance level

Example: To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

Definition: The **rejection region** is a range of values of the test statistic that would lead you to **reject** the null hypothesis.

$\alpha = 1 -$ [confidence level]

Definition: The **significance level** α indicates the largest probability of the test statistic occurring under the null hypothesis that would lead you to reject the null hypothesis.

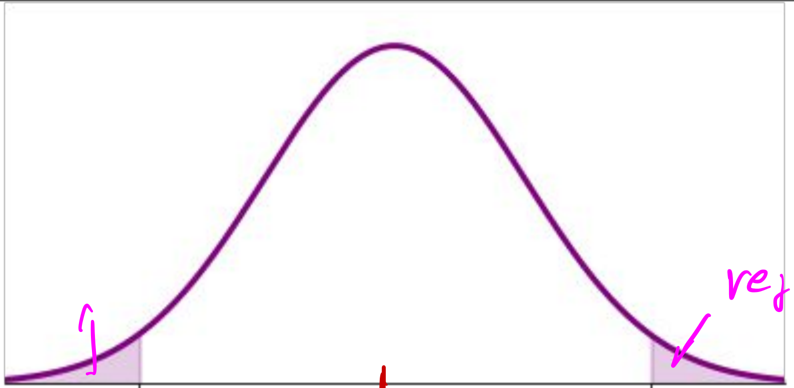
① write H_0 & H_1

② compute T.S (z)

③ compute to z_{crit}

(critical values are edge of reject region

rejection region



−1.96 0 1.96 → z

Rejection regions and significance level

[article about this example here]

Example: To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at the 0.1 significance level or not?

$H_0 : p = 0.5$

$H_1 : p \neq 0.5$

$\alpha = 0.1$ total area in tail $\rightarrow z_{TS} = 1.771$

Get our z_{crit} critical value by figuring out what z puts α probability in the tails;

calculate test statistic; Assume H_0 is true

$$\hat{p} = \frac{139}{250} \overset{CLT}{\sim} N\left(p = 0.5, \sigma^2 = \frac{p(1-p)}{250}\right)$$

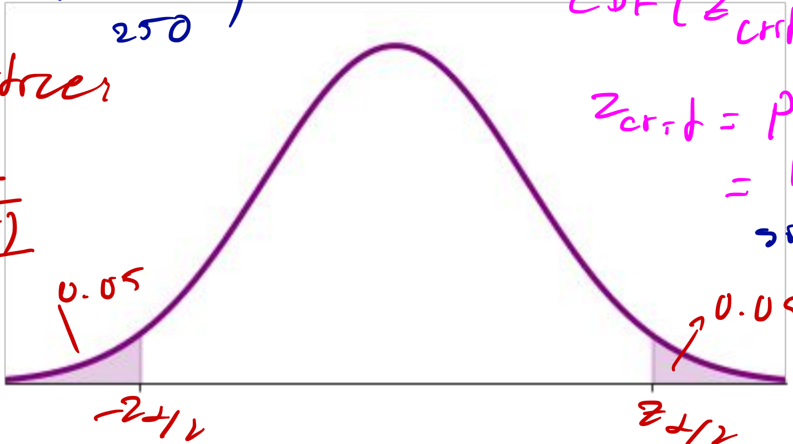
$$CDF(z_{crit} \text{ or } z_{\alpha/2}) = 1 - \frac{\alpha}{2} = .95$$

$$z_{crit} = ppf(.95) = 1.645$$

Box-Muller to standardize:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{139}{250} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{250}}}$$

test statistic



since $z_{TS} = 1.774 > 1.645$ we reject H_0 & conclude for 10% sig. level coin is

Rejection regions and significance level

[article about this example here]

Example: To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at the 0.1 significance level or not?

H_0: p = 0.5
H_1: p ≠ 0.5

$\left(1 - \frac{0.1}{2}\right) = 0.95$

$\alpha = 0.1 \rightarrow z_{\alpha/2}$ = stats.norm.ppf(0.95) = 1.645

→ reject H_0 if we see anything **more extreme** than $z_{crit} = \pm 1.645$

$$\hat{p} \sim N\left(0.5, \frac{0.5(1-0.5)}{250}\right)$$

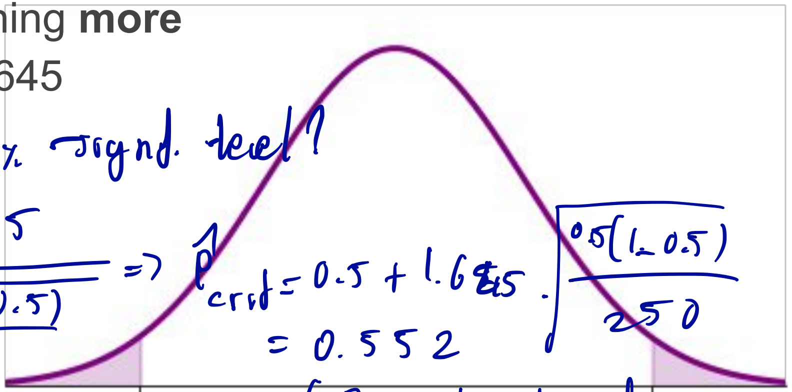
$$\Rightarrow \quad z = \frac{\frac{139}{250} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{250}}}$$

$$= \ldots \approx 1.771$$

How many flips is 10% signif. level?

$$z_{\alpha/2} = 1.645 = \frac{\hat{p}_{crit} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{250}}} \Rightarrow \hat{p}_{crit} = 0.5 + 1.645 \cdot \sqrt{\frac{0.5(1-0.5)}{250}}$$

$$= 0.552$$



↳ This number depends on sig. level 1.645

$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

$\rightarrow z = 1.771 > z_{\alpha/2} = 1.645$

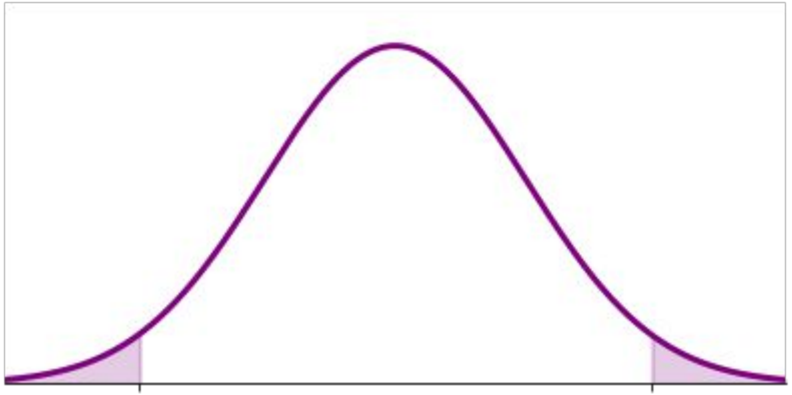
→ **Conclusion:** evidence supports rejecting null H_0

Rejection regions and significance level

[article about this example here]

Example: To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at the 0.05 significance level or not?

5% sig. level



Rejection regions and significance level

article about this example here

Example: To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at the 0.05 significance level or not?

We still find z = 1.77

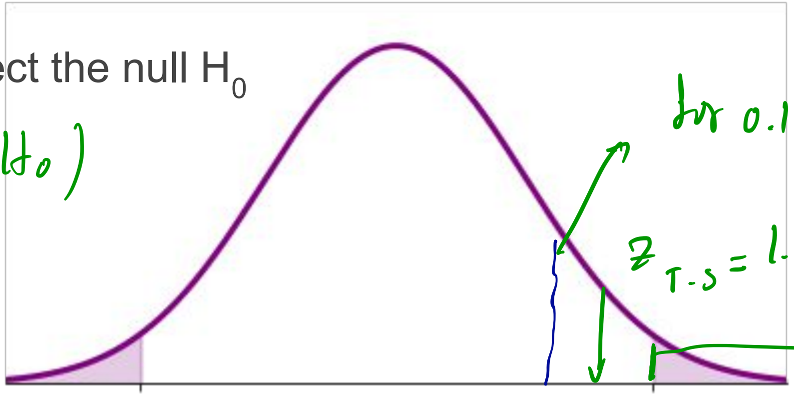
$\left(1 - \frac{\alpha}{2}\right) = 0.976$

What changes? Now z_{crit} is given by: $\alpha = 0.05 \rightarrow z_{\alpha/2}$ = stats.norm.ppf(0.975) = 1.96

→ z = 1.77 is **not** in the rejection region

→ **Conclusion:** fail to reject the null H_0

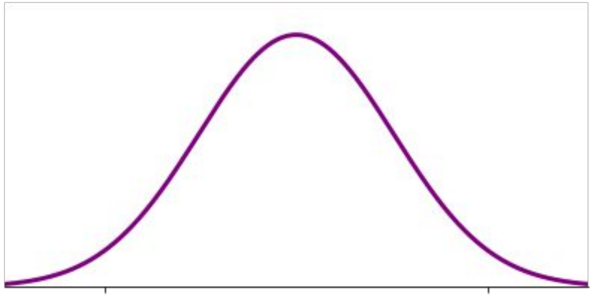
(we do not accept H_0)

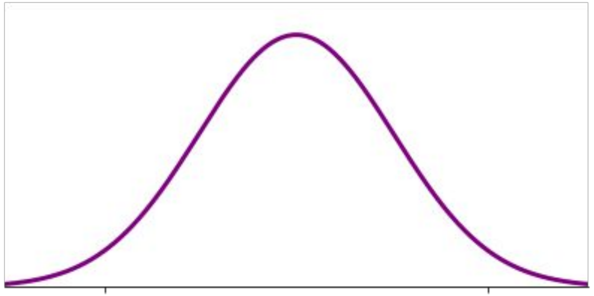


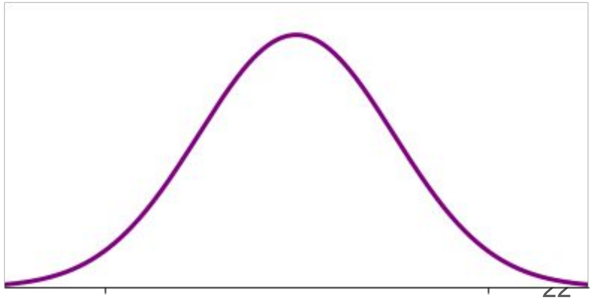
Different tests for different hypotheses

The coin example was an example of a **two-tailed hypothesis test**, because we would have rejected the null hypothesis if the coin had been biased towards heads **or** towards tails

Alternative hypothesis	Rejection region for level α test
H_1: $\theta > \theta_0$	$z \geq z_\alpha$
H_1: $\theta < \theta_0$	$z \leq -z_\alpha$
H_1: $\theta \neq \theta_0$	$(z \geq z_{\alpha/2})$ or $(z \leq -z_{\alpha/2})$

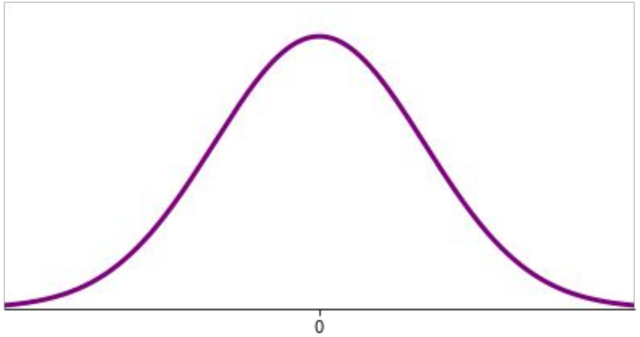






Switching advertising strategies

Example: S'pose a company is considering hiring a new outside advertising company to help generate traffic to their website. They currently get 200 thousand hits/day, with a standard deviation of 50 thousand hits per day. Suppose they hire the new ad agency for a 30-day trial. During those 30 days, their website gets 210 thousand hits/day. Perform a statistical hypothesis test to determine if the new ad campaign outperforms the old one at the 0.05 significance level.

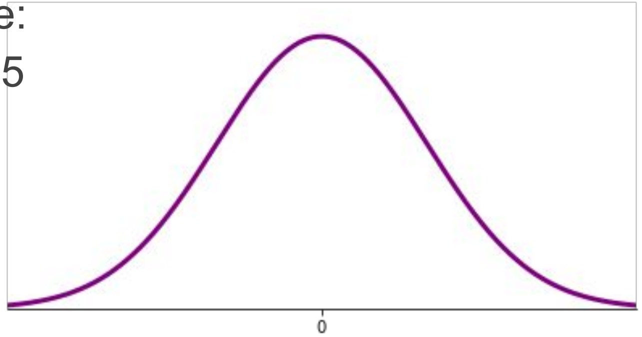


Switching advertising strategies

Example: S'pose a company is considering hiring a new outside advertising company to help generate traffic to their website. They currently get 200 thousand hits/day, with a standard deviation of 50 thousand hits per day. Suppose they hire the new ad agency for a 30-day trial. During those 30 days, their website gets 210 thousand hits/day. Perform a statistical hypothesis test to determine if the new ad campaign outperforms the old one at the 0.05 significance level.

Testing H_0: $\mu \leq 200$ against H_1: $\mu > 200$

$\alpha = 0.05$ and one-tail
→ rejection region critical value:
z_{crit} = norm.ppf(0.95) = 1.645



Test statistic:

$$X \sim N\left(200, \frac{50^2}{30}\right)$$

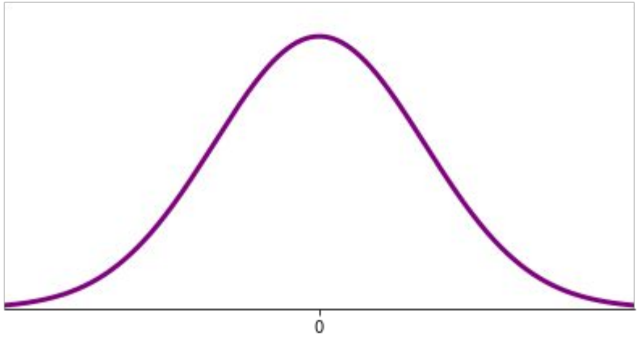
$$\Rightarrow \quad z = \frac{210 - 200}{\frac{50}{\sqrt{30}}}$$

$$\approx 1.095$$

⇒ 1.095 < 1.645, so **fail to reject H_0**

Important assumptions

Question: What assumptions did we make in the previous example?

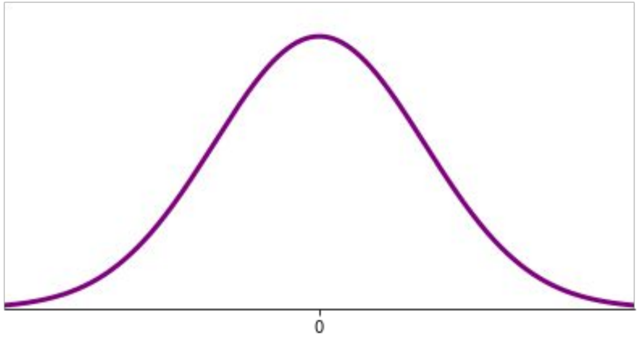


Important assumptions

Question: What assumptions did we make in the previous example?

1) Assumed that the CLT would hold -- n=30 samples (days)

2) Assumed that we can represent the distributions involved as Normal



Errors in hypothesis testing

	H_0 true	H_0 false
Rej. H_0	Type I Error	:)
Fail to reject	:)	Type 2 error

Definitions:

A **Type I Error** occurs when the null hypothesis is incorrectly rejected (it was, in fact, true)

→ **false positive**

A **Type II Error** occurs when the null hypothesis is incorrectly not rejected (it was false)

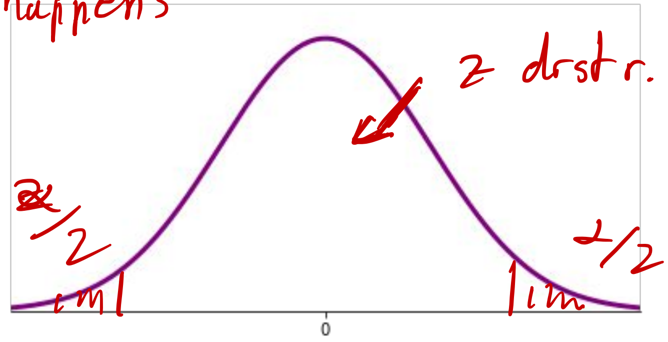
→ **false negative**

Question: What is the probability we commit a **Type I Error**?

α = P($\hat{p}$ (or other T.S) just happens by chance to land in the rejection region)

α = P(

z distr. look if H_0 is true



Errors in hypothesis testing

Definitions:

A **<u>Type I Error</u>** occurs when the null hypothesis is incorrectly rejected (it was, in fact, true)

→ **false positive**

A **<u>Type II Error</u>** occurs when the null hypothesis is incorrectly not rejected (it was false)

→ **false negative**

Question: What is the probability we commit a **Type I Error**?

Answer: This is exactly the significance level α

Why we care: Choose α by considering how willing you are to risk a Type I Error

H_0 = airplane safe

H_1 = airplane unsafe

What just happened?

- **Hypothesis testing** happened!
  - A way to formally ask questions like:

$$\mu_A \neq \mu_B \quad \text{or} \quad \mu_A < \mu_B$$

- **Significance level** -- how much evidence do you need in order to reject the null hypothesis?

- **Rejection regions** -- if your test statistic falls in here, you have evidence to reject the null hypothesis

- **Type I** and **Type II Errors**
  (false positives and negatives, respectively)