Welcome to STA 101!

Hypothesis testing

Example 1: is a mystery coin fair?

Setting: A carnival barker digs an unfamiliar coin out of his pocket and invites you to flip it as many times as you want.

Question: is the coin fair?

Hypotheses: two competing claims...

- H_0 : Prob(flipping heads) $= 0.5$;
- H_A : Prob(flipping heads) $\neq 0.5$

Data:

- You flip it 10 times, and get 60% heads. Is it fair?
- You flip it 50 times, and get 56% heads. Is it fair?
- You flip it 1,000,000 times and get 56% heads. Is it fair?

Example 2: is a medical consultant better than average?

Setting: To avoid complications, some prospective organ donors hire a medical consultant to advise on aspects of the surgery. The average complication rate for liver donor surgeries in the US is about 10%.

Question: does the consultant I am interviewing have a different complication rate than the US average?

Hypotheses: two competing claims...

- H_0: Prob(complications with this consultant) $= 0.1$;
- H_A: Prob(complications with this consultant) $\neq 0.1$.

Data: she has advised 62 liver donors, and 3 of them (4.8%) have had complications. Is she better than average?

Example 3: is yawning contagious?

Setting: the Mythbusters randomly split people into two groups:

- (control) didn't have a yawner near them;
- (treatment) had a yawner near them.

Question: are you more likely to yawn if someone yawns near you?

Hypotheses: two competing claims...

- H_0: Prob(yawning near a yawner) $=$ Prob(yawning alone);
- H_A: Prob(yawning near a yawner) $>$ Prob(yawning alone).

Data:

- proportion of yawners in the treatment group: $10/34 \approx 0.29$;
- proportion of yawners in the control group: $4/16 = 0.25$;
- difference: $0.2941 - 0.25 \approx 0.04$.

Hypothesis testing

Two competing claims *about the population*...

- **Null** (or **baseline**) **hypothesis**: "there's nothing going on;"
- **Alternative hypothesis**: "there's *something* going on."

In each example...

- we have evidence (data) in the form of a random sample;
- we have a best guess (point estimate);
- but there is uncertainty (eg. do I have enough data?);
- So what's the answer?

Which claim are the data most consistent with?

Do we have enough information to tell?

Could it be that our results are just due to chance?

The main idea

Setting: you have data and a best guess;

Hypothetical: assume the null is true;

Question: in a hypothetical world where the null is true, how crazy would it be to observe the data you observed?

Decision:

- if the data would be crazy, then the null must be bogus. Reject the null in favor the alternative.
- if the data would not be out of the ordinary, then you cannot rule the null out. You fail to reject the null.

Hypotheses

$$\mathrm{H}_0 : p = p_0$$
$$\mathrm{H}_A : p \neq p_0$$

- p is the true but unknown population parameter. You're trying to guess its value with data;
- In all the examples today, p is an unknown probability (a proportion or percentage), hence the notation p;
- p_0 is the *null* or *hypothesized* value. It's the "baseline" value you are testing for;
- H_0 is the status quo. "nothing special is going on;"
- H_A is the alternative. "something is going on;"
- "Innocent versus guilty."

Types of alternatives

Two-sided alternatives:

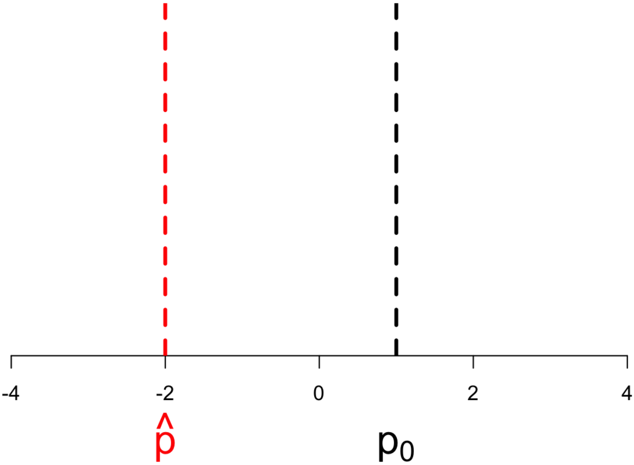
$$\mathrm{H}_0 : p = p_0$$
$$\mathrm{H}_A : p \neq p_0$$

One-sided alternatives:

$$\mathrm{H}_0 : p = p_0$$
$$\mathrm{H}_A : p > p_0$$

$$\mathrm{H}_0 : p = p_0$$
$$\mathrm{H}_A : p < p_0$$

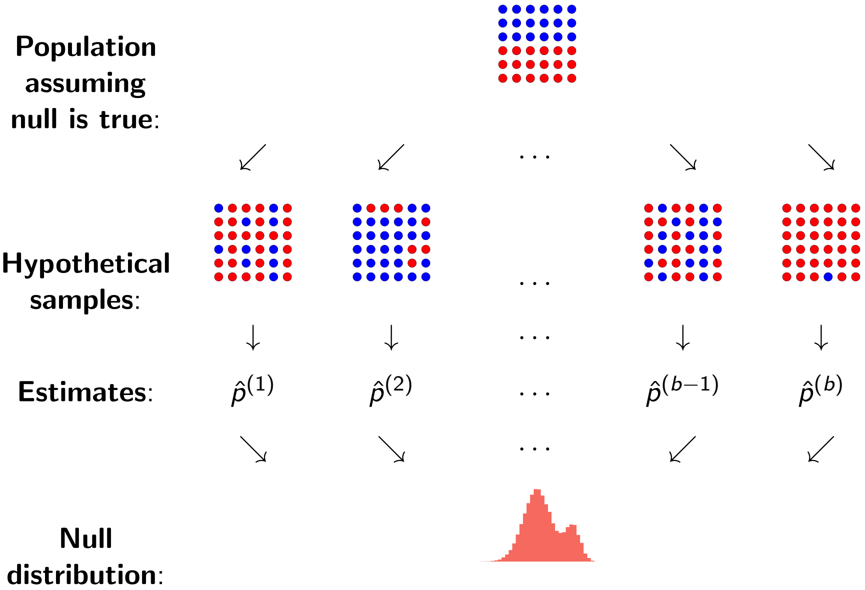
Sooo...what's the conclusion?



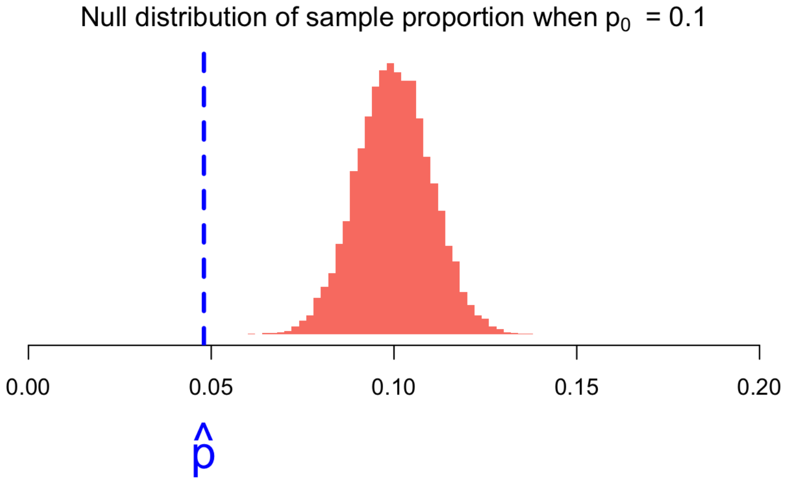
The null distribution

- hypothetical sampling distribution of the estimate *assuming the null were true*;
- visualizes the "menu of options" for the estimate in a world where the null is true.
- if the estimate you actually got would be "off the menu", the null was probably silly to begin with.
- if the estimate you actually got could be "on the menu," then the null is still in play.

The null distribution

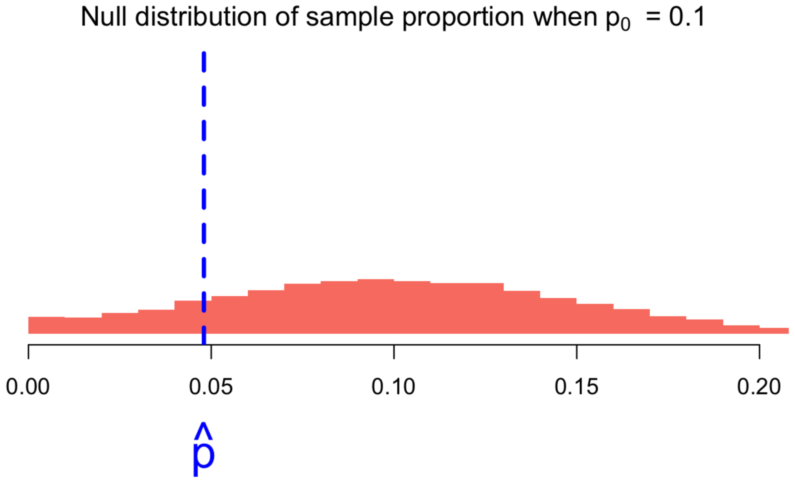


What if the null distribution looked like this?



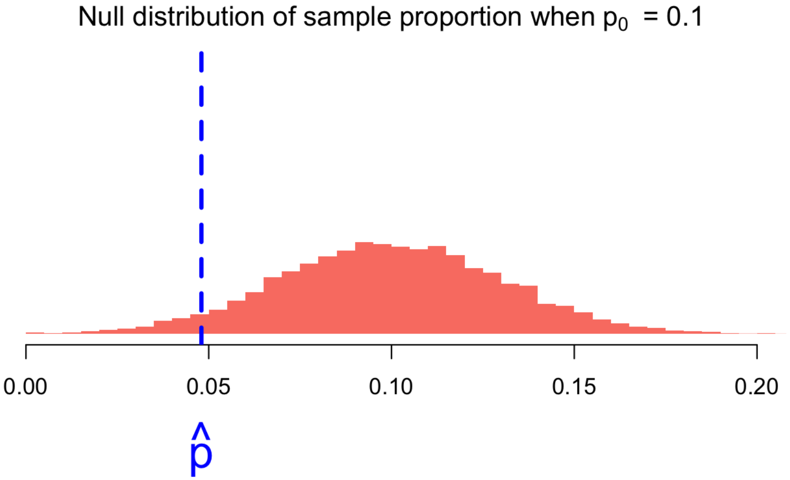
Reality (the estimate) and the hypothetical (the null distribution) look incompatible. Reject the null hypothesis.

What if the null distribution looked like this?



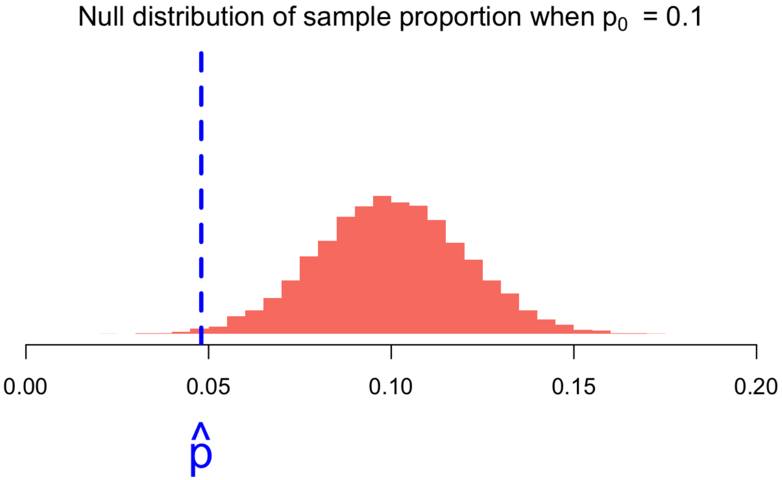
What would you conclude here?

What if the null distribution looked like this?



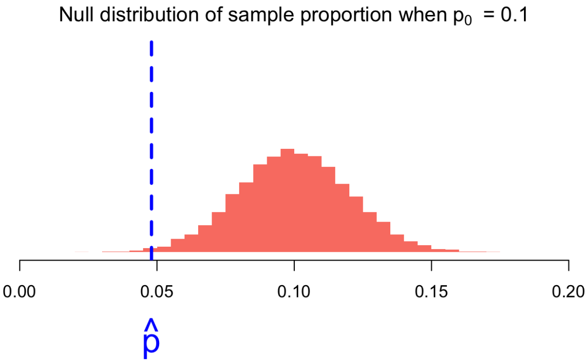
What would you conclude here?

What if the null distribution looked like this?



What would you conclude here?

Sooo...what's the conclusion?



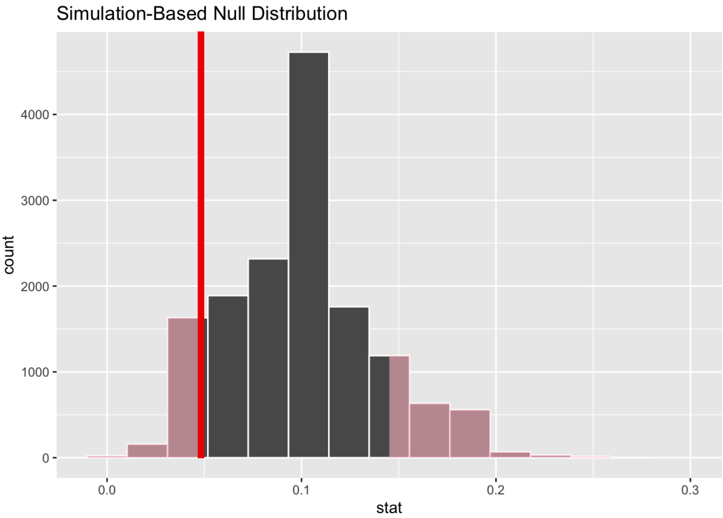
If our estimate is far out in the tails of the null distribution, this suggests the null was a bunch of malarkey from the start.

What do we mean by "far out in the tails"? Compare p-value to a threshold α called the discernability level:

- if p-value $< \alpha$, Reject H_0;
- if p-value $>= \alpha$, Fail to reject H_0.

p-value

"If the null were in fact true, what's the chance I would get results even crazier than what I actually got."



p-value

- Assuming the null is true, the p-value is the probability of get a result as extreme or more extreme than the one you actually got;
- If this probability is "large", your estimate feels right at home with the null. Fail to reject;
- If this probability is "small," the estimate and the null are incompatible. Reject the null in favor of the alternative.

Question: we've converted the question of "how close is close" to a question of "how small is small." Is this progress?

Follow-up: how do you decide the cut-off?

Recall: picking the confidence level of an interval estimate

Task: Choosing 75% vs 90% vs 95% vs 99% confidence?

Trade-off: We want an interval that is...

- ...wide enough to capture the truth with high confidence;
- ...narrow enough to teach us something meaningful about where the truth actually lives.

Silly example: The interval $(-\infty, \infty)$ is guaranteed to capture the truth every time. But it teaches us nothing.

Related: picking the discernability level of a test

The choice of cut-off α can be domain and application dependent, but the overall goal is to balance the risk of two types of errors:

		Your decision	
		Reject H_0	Fail to reject H_0
The	H_0 true	Type 1 error	Correct!
truth	H_0 false	Correct!	Type 2 error

- Type 1 error = false positive;
- Type 2 error = false negative.

Example: a judge sentencing defendants

Hypotheses:

$$H_0 : \text{person is innocent}$$
$$H_A : \text{person is guilty}$$

Outcomes:

		Your decision	
		Reject H_0	Fail to reject H_0
The truth	H_0 true	Jail innocent person	Free innocent person
	H_0 false	Jail guilty person	Free guilty person

- Aspects of the American trial system regard a Type 1 error as worse than a Type 2 error (reasonable doubt standard, unanimous juries, presumption of innocence, etc).

Example: a doctor treating patients

Hypotheses:

$$\begin{aligned} H_0 &: \text{ person is well} \\ H_A &: \text{ person is ill} \end{aligned}$$

Outcomes:

		Your decision: Reject H_0	Your decision: Fail to reject H_0
The truth	H_0 true	Treat healthy person	Ignore healthy person
	H_0 false	Treat sick person	Ignore sick person

- Doctors tend to prefer treating too much than too little.

Example: the boy who cried wolf

Hypotheses:

$$H_0 : \text{no wolf}$$
$$H_A : \text{Run! A wolf!}$$

Outcomes:

		Your decision	
		Reject H_0	Fail to reject H_0
The truth	H_0 true	Panic over nothing	Go about your day
	H_0 false	Run from wolf	Get eaten

- In Part 1 of the story, townspeople commit Type 1 error;
- In Part 2 of the story, townspeople commit Type 2 error.

Picking the discernibility level of a test

The choice of cut-off α can be domain and application dependent, but the overall goal is to balance the risk of two types of errors:

		Your decision: Reject H_0	Your decision: Fail to reject H_0
The truth	H_0 true	Type 1 error	Correct!
	H_0 false	Correct!	Type 2 error

- $\alpha \uparrow \Longrightarrow$ easier to reject H_0 $\Longrightarrow$ Type 1 $\uparrow$ Type 2 $\downarrow$
- $\alpha \downarrow \Longrightarrow$ harder to reject H_0 $\Longrightarrow$ Type 1 $\downarrow$ Type 2 $\uparrow$
- Typical choices: $\alpha = 0.01, 0.05, 0.10, 0.15$.

Cardinal Sins in Statistics, Part 2 of 91

- The *p*-value *is not* the probability that the null hypothesis is true. Would that it were so simple;
- The *p*-value is the probability of a crazier result than the one you got *assuming the null is true*;
- The null is either true or it is not, with probability zero or probability one.

Cardinal Sins in Statistics, Part 3 of 107

- If you've taken a statistics course before, or read papers that use hypothesis testing for drawing conclusions, you might have encountered the term "statistically significant" or "significance level".

- We will use the term "statistically discernable" or "discernability level", because "significant" has a different meaning in everyday language and this often causes confusion about what "statistically significant" means. It doesn't necessarily mean a notable or important event has happened, it just means the data are unlikely to have come from the null model.

Example

Setting: a coin flip comes up heads with probability 0.499.

Hypotheses:

$$\begin{aligned} \mathrm{H}_0 &: \text{Prob(heads)} = 0.5 \\ \mathrm{H}_A &: \text{Prob(heads)} \neq 0.5 \end{aligned}$$

Result: you flip the coin 10,000,000 times and get a p-value that's practically zero, and *correctly* conclude that the null is literally false. So what?

Punchline: the machinery of statistics *cannot* tell you if your results are "meaningful" and "important". It can only tell you if the results are likely or not under random sampling.

statistical significance $\neq$ importance

Hypothesis testing: an avalanche of itchy jargon

- **null hypothesis**
- **alternative hypothesis**
- **null distribution**
- p-**value**
- **discernability level**
- **Type 1 error**
- **Type 2 error**
- ...
- ...