Two Competing Claims

"There is nothing going on." (Null Hypothesis)	"There is something going on" (Alternative Hypothesis)
Promotion and gender are independent.	Promotion and gender are dependent.
There is no gender discrimination.	There is gender discrimination.
The observed difference in proportions is simply due to chance.	The observed difference in proportions is not due to chance.

Court Trial

Person is innocent until proven guilty

Burden of proof is on the prosecution

Collect evidence; determine if there's enough to convict

Verdict of not guilty does not guarantee innocence

Hypothesis Test

Variables are independent until proven dependent

Burden of proof is on the alternative hypothesis

Collect evidence; determine if there's enough to decide

Inability to reject null hypothesis does not make alternative hypothesis false

Example: Gender Discrimination

		Promoted		
		No	Yes	P(Yes|Sex)
Sex	Female	10	14	0.58
	Male	3	21	0.88

H_0: Promotion and gender are independent; difference in proportions is simply due to chance.

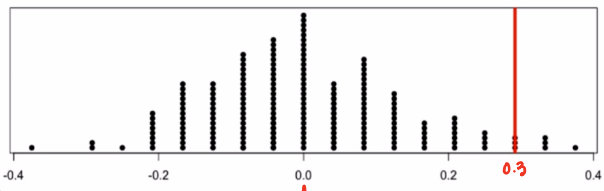
H_A: Promotion and gender are dependent; difference in proportions is not due to chance.

Simulation approach

- Consider an imaginary version of this experiment where the null is known to be true
- Simulate the results of this experiment
- Compare results of simulation to results of real-life experiment
- How unusual is the real-life result in the null hypothesis world?

Recall: Gender Discrimination

How likely is a 0.88 - 0.58 = 0.30 difference?



Why reject H_0?

Why is this considered evidence against the null hypothesis?

In our simulation, this level of disparity appeared only 5% of the time

Is it more plausible that we observed an event that should happen 1 out of 25 times, or that the null hypothesis is true?

Hypothesis Testing Framework

1. Start with a null hypothesis (H_0): a skeptic's position, the status quo
2. Identify an alternative hypothesis (H_A), our research question, the thing we want to test
3. Assuming H_0 is true, calculate probability of observing data
4. If this probability is small enough, conclude H_0 is unlikely and reject it

Statistical significance

If a hypothesis test results in rejecting the null hypothesis, the result of the test is called "statistically significant."

≠

Statistical vs Practical Significance

Suppose:

H_0: $\mu = 10$ H_A: $\mu > 10$ $\sigma = 2$

	$\bar{x} = 10.05$	$\bar{x} = 10.1$	$\bar{x} = 10.2$
n = 30	p≈0.45	0.39	0.29
n = 5000	p≈0.039	0.0002	0

```
> pnorm(10.1, 10, 2/sqrt(30), lower.tail=FALSE)
[1] 0.3920956
```

Decision Errors

	Fail to reject H_0	Reject H_0
H_0 is true	✓	Type 1 Error
H_A is true	Type 2 Error	✓

In US judicial system, H_0 = innocent

- Type 1 error
  - Innocent person convicted
- Type 2 error
  - Guilty person not convicted

Decision Error Probabilities

	Fail to reject H_0	Reject H_0
H_0 is true	1 - α	α
H_A is true	β	1 - β

α: probability of rejecting H_0 when you shouldn't, = the significance level

β: probability of failing to reject H_0 when you should

power (1 - β): probability of correctly rejecting H_0 of the hypothesis test

Power

Before carrying out a study, we should estimate the power of any tests we plan to do. (This is similar to planning how large a sample size we need for a given margin of error.)

Calculating power is more subtle, though, because it depends on several factors.

In particular, we need to decide what is a practically significant difference.

Example

Average systolic blood pressure in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg. We are investigating whether a medication increases blood pressure. Consider hypotheses:

H_0: μ = 130
H_A: μ > 130

Suppose the medicine increases average blood pressure to 132 mmHg. What is the power of a study with N = 100 to detect this difference?

Example

Breaking this down into two steps:

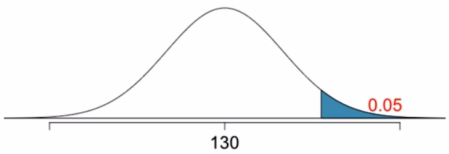
H_0: μ = 130
H_A: μ > 130

σ = 25, N = 100, α = 0.05

1. What is the minimum value of $\bar{x}$ that would lead you to reject H_0? (Hint: use qnorm in the distribution N(130, SE); work back from p = 0.05 to a value of $\bar{x}$)
2. What is the probability that we would see $\bar{x}$ greater than or equal to that if sample means actually come from the distribution: N(mean = 132, SE = 25 / √100)?

Calculating Power: Problem 1

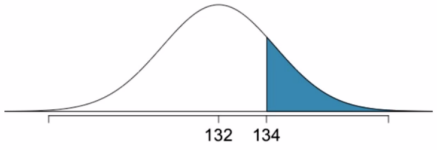
Which values of $\bar{x}$ represent sufficient evidence to reject H_0? (H_0: μ = 130, H_A: μ > 130)




```
> qnorm(0.05, 130, 2.5, lower.tail = FALSE)
[1] 134.1121
```

Calculating Power: Problem 2

What is the probability that we would reject H_0 if $\bar{x}$ came from N(mean = 132, SE = 2.5)?




```
> pnorm(134.1121, 132, 2.5, lower.tail = FALSE)
[1] 0.1991001
```

β = 0.8008999

Why worry about power?

Underpowered studies pose multiple problems:

- The obvious problem: by definition, an underpowered study is likely to fail to detect an effect
- If we do observe a statistically significant result, quantitative estimates are likely to be inflated

Why worry about power?

Notice a side effect: in order to detect the difference with a true mean of 132 mmHg, we had to *observe* a sample mean of 134 mmHg

If someone asks “How much does the drug raise blood pressure?”

- True answer: “It raises average BP by 2 mmHg”
- Our point estimate: “It raises average BP by 4 mmHg”

We overestimate by a factor of 2!

Why worry about power

It’s not just the fact that we may not detect an effect

As we see, underpowered studies tend to produce oversized estimates

This can bias future research, especially if statistical significance is used to decide whether to publish a result

Replication crisis

Today, there is a crisis in several fields of science--esp. psychology, medicine

Many published results cannot be replicated (suggesting they are not really true)

A 2018 study repeated experiments from 21 social science articles at higher power; only 13 results stood up

Decision errors

To reduce type 1 errors:

- Reduce α (this will raise β; more type 2 errors)

To reduce type 2 errors:

- Increase α (not usually a good idea)
- Increase the effect size desired by H_A (often not a good idea--be careful!)
- Reduce measurement variability
- Increase sample size