Business Experimentation and Causal Methods

Prof. Fradkin
Topic: Measuring Uncertainty

Last Time

1. Random Variables, Distributions
2. Expectations and Standard Deviation
3. Law of Large Numbers
4. Randomized Assignment and Selection Bias
5. Important Assumptions

This Time

Basic question:

We’ve run an experiment and get an estimate X. How far away from X is the true ATE likely to be?

This Time

Basic question:

We’ve run an experiment and get an estimate X. How far away from X is the true ATE likely to be?

Why this matters:

We could have gotten X just by chance. If there is a lot of uncertainty, true effect may be 0 or even -X.

Concepts We Will Learn

1. **Sampling Distribution**
2. Standard Error and Confidence Interval
3. Null Hypothesis and P-Value
4. Statistical Power

This is not intuitive!

FiveThirtyEight

Politics Sports **Science & Health** Economics Culture

NOV. 24, 2015 AT 12:12 PM

Not Even Scientists Can Easily Explain P-values

P-values have taken quite a beating lately. These widely used and commonly misapplied statistics have been blamed for giving a veneer of legitimacy to dodgy study results, encouraging bad research practices and promoting false-positive study results.

But after writing about p-values again and again, and recently issuing a correction on a nearly year-old story over some erroneous information regarding a study's p-value (which I'd taken from the scientists themselves and their report), I've come to think that the most fundamental problem with p-values is that no one can really say what they are.

http://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/

Where does uncertainty in experiments come from?

- **Who gets the treatment is random:**

  Many randomizations that could have occurred —> but we only observe one.

  Our randomization, just by chance, could have resulted in something closer or further away from the true average treatment effect.

- **Other one we won't talk about today:**

  Sometimes we randomly choose who is eligible to participate in a study or experiment.

  For example, a political survey may randomly sample 1000 people in Boston but the total population of Boston is much larger. Uncertainty because of who is sampled.

Two randomizations: Different Estimates

Person	In-person	Zoom	True Effect
John	1	1	0
Mary	0	1	-1
Suraj	0	0	0
Katerina	1	1	0
Molly	0	1	-1
Leroy	0	0	0

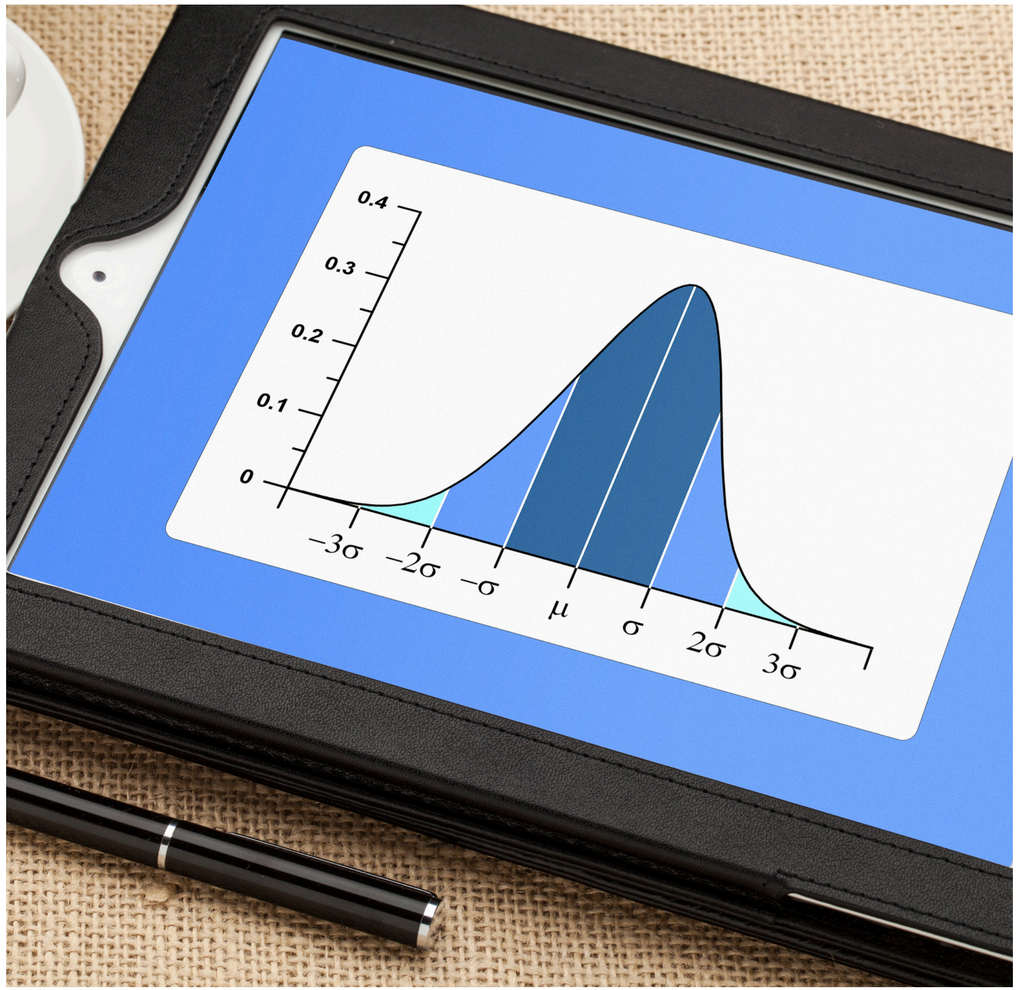
1/3 - 1/3 = 0

Person	In-person	Zoom	True Effect
John	1	1	0
Mary	0	1	-1
Suraj	0	0	0
Katerina	1	1	0
Molly	0	1	-1
Leroy	0	0	0

0 - 1 = -1

Sampling Distribution

- The distribution of estimates we get because of different randomizations.



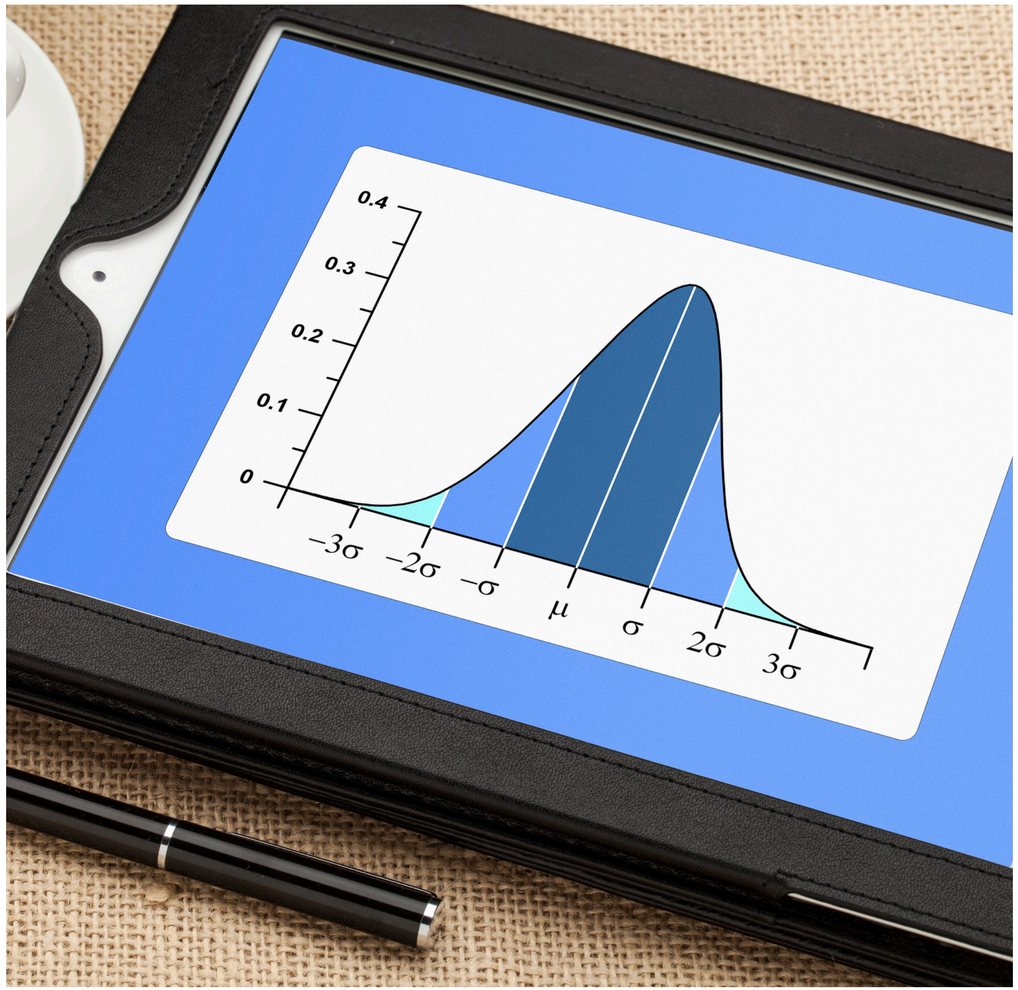
Python interlude

Concepts We Will Learn

1. Sampling Distribution
2. **Standard Error and Confidence Interval**
3. Null Hypothesis and P-Value
4. Statistical Power

Standard Error

- Definition: Standard deviation of the sampling distribution.
- It measures our uncertainty about the true value.
- The larger the standard error, the more uncertain we are about where the truth lies.
- Why might standard errors be large?
  - Small number of observations.
  - Large variance of outcomes.
  - Large variance in individual treatment effects.

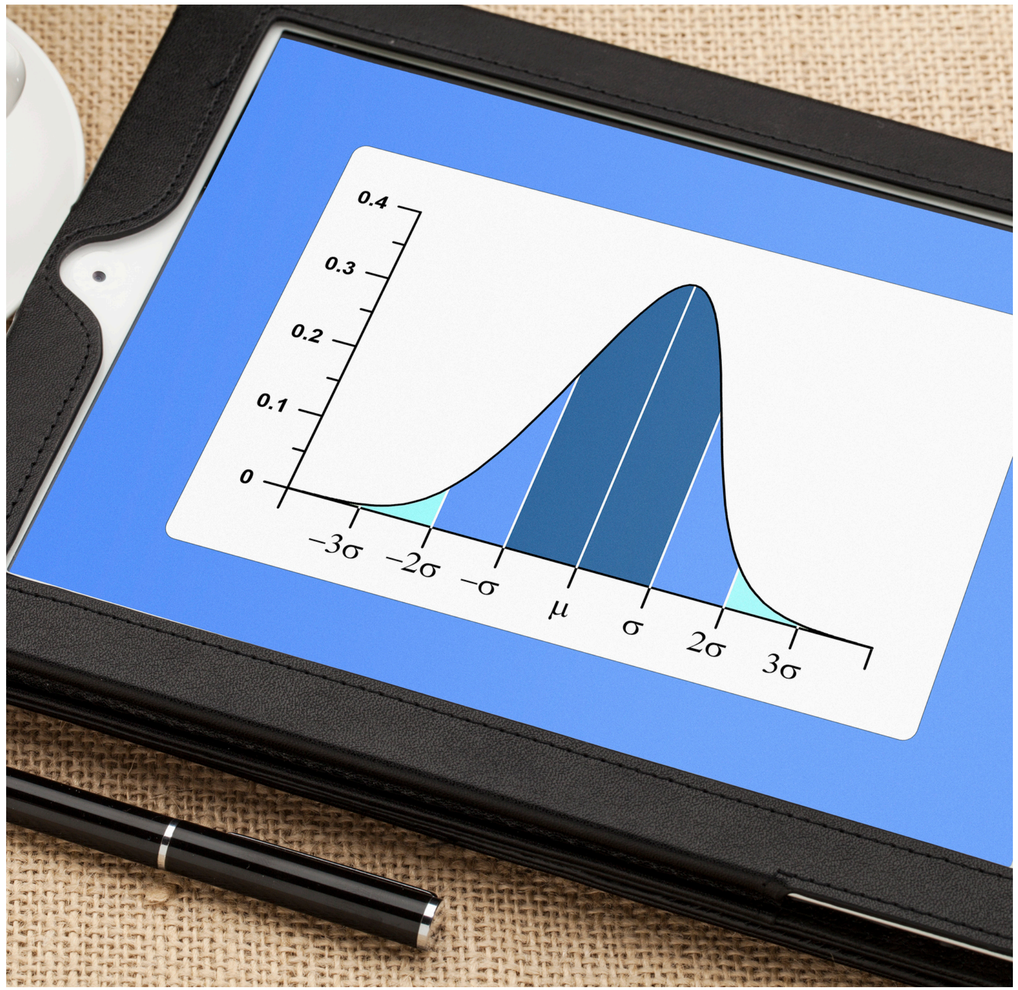


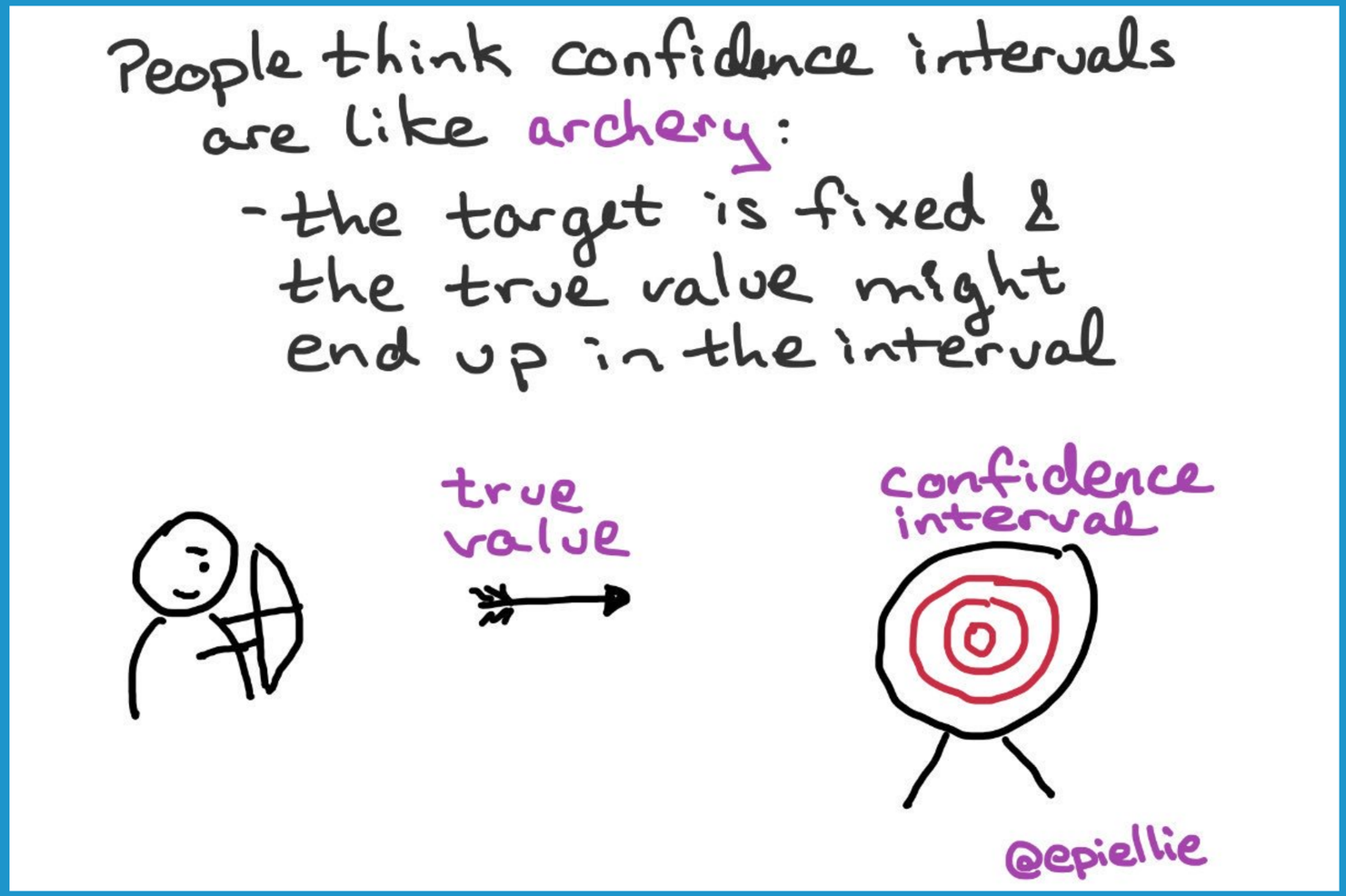
95% Confidence Interval

- It is an estimate such that:
  - 95% percent of the time we compute this interval, the true value is in the interval.
- We think of the confidence interval as a range that typically contains the truth.
- Compute 95% confidence interval as follows:

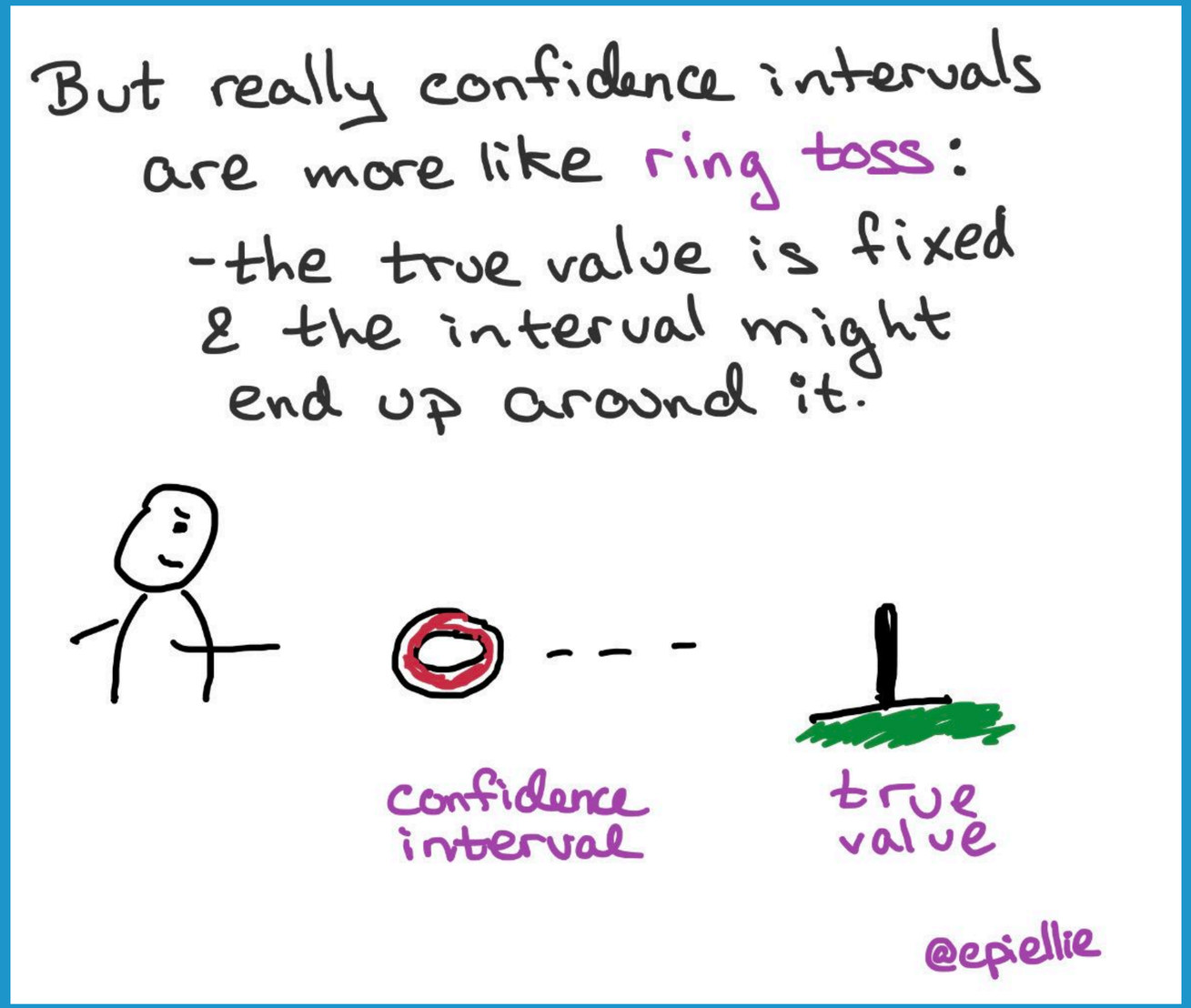
$$\hat{ATE} - 1.96 * \hat{SE}, \hat{ATE} + 1.96 * \hat{SE}$$

- Where $\hat{SE}$ is the standard error of the $\hat{ATE}$.
- 1.96 is the number for the 95% confidence interval.




People think confidence intervals are like archery:
- the target is fixed & the true value might end up in the interval
true value
confidence interval
@epiellie

Not correct!



Larger confidence interval (e.g. 99% vs 95%) means higher chance of hitting the toss.

Getting the confidence interval manually

```python
mean_diff = np.mean(treated_sample) - np.mean(control_sample)
std_err_diff = np.sqrt(np.var(treated_sample, ddof=1)/len(treated_sample) + np.var(control_sample, ddof=1)/len(control_sample))

# Confidence level and critical t-value
confidence_level = 0.95
t_crit = 1.96

# Confidence interval
CI_lower = mean_diff - t_crit * std_err_diff
CI_upper = mean_diff + t_crit * std_err_diff

print(f"Confidence Interval: {CI_lower}, {CI_upper}")
```

✓ 0.0s Python

```
Confidence Interval: -0.5906195274170887, 1.2572861940837554
```

Concepts We Will Learn

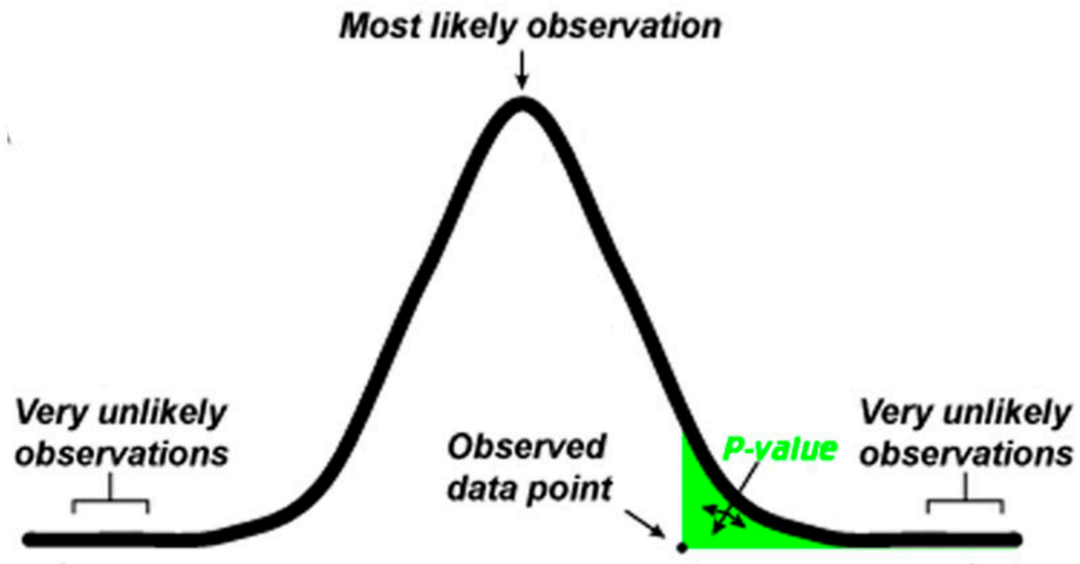
1. Sampling Distribution
2. Standard Error and Confidence Interval
3. **Null Hypothesis and P-Value**
4. Statistical Power

How many observations do we need?

- Skeptic: Suppose that the treatment had a 0 effect.
- Null Hypothesis: True effect of the experiment is 0.
- Assume that the skeptic is right. We still have a sampling distribution due to the fact that any randomization will have some differences between treatment and control.

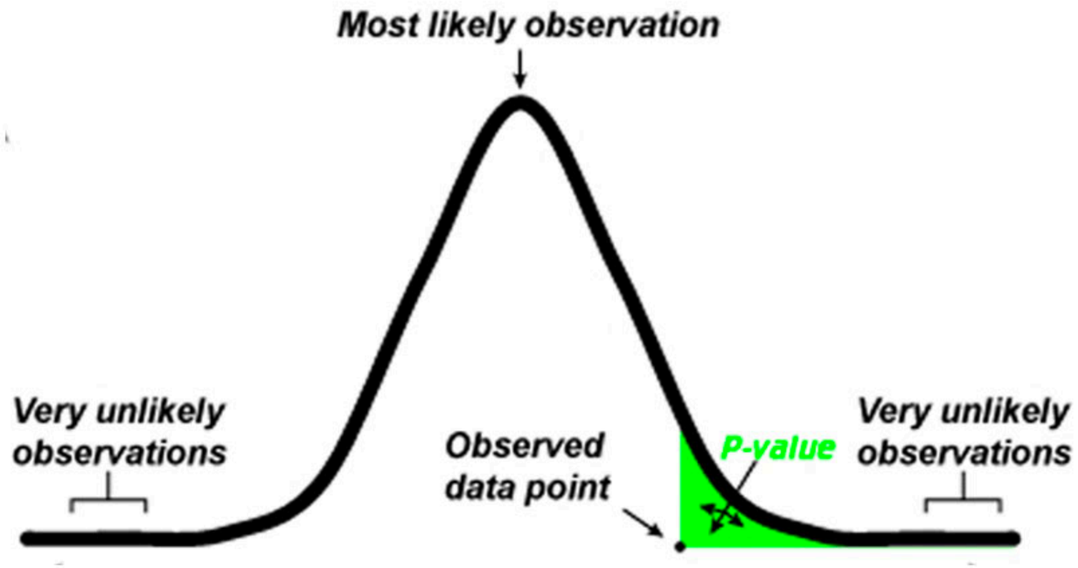
P-value!

- Probability under the null hypothesis that:
  - Just by chance we get an estimate at least as big as the one that we did get.
- For example, if we get a treatment effect of 1, what is proportion of the sampling distribution greater than 1.



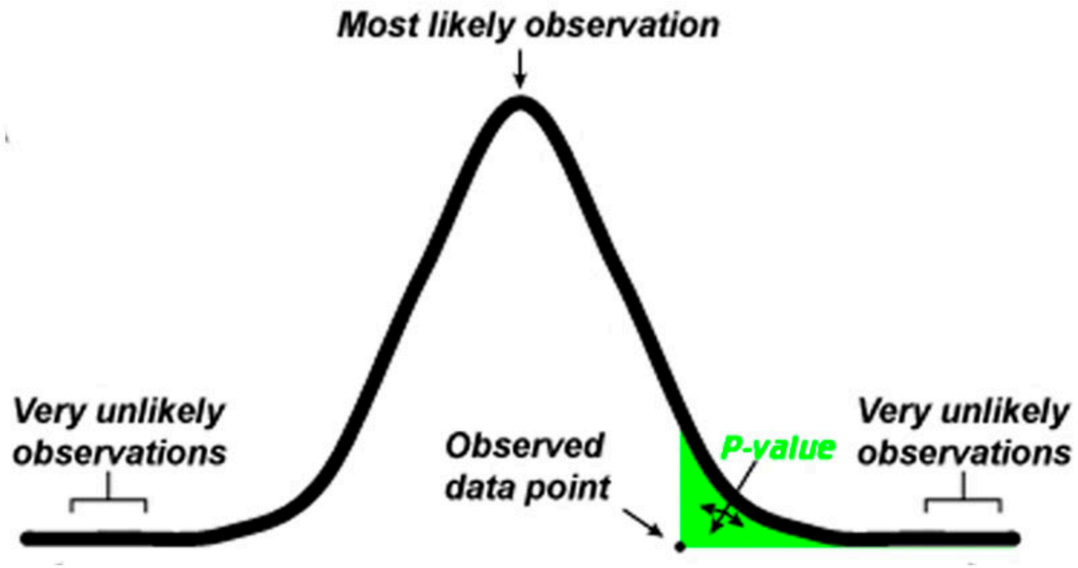
Rejecting the null hypothesis.

- It is a convention that we reject the null hypothesis with the p-value < .05 (5%).
- When we reject the null hypothesis, we can also say that we ‘detected’ an effect.



What does a high p-value mean?

- Suppose we get a p-value of 75%. We don’t reject the null hypothesis.
- If we don’t reject the null hypothesis, there may still be an effect. But we can’t confidently say so.
- We never ‘accept’ the null hypothesis. We fail to reject it.



Using ttest_ind to get the p-value

Two arguments:

Treated Outcomes Control Outcomes

```python
from statsmodels.stats.weightstats import ttest_ind, _tconfint_generic

treated_sample = data_po[data_po['T'] == 1]['Y']
control_sample = data_po[data_po['T'] == 0]['Y']
tstat, pvalue, df = ttest_ind(x1 = treated_sample, x2 = control_sample, alternative = 'two-sided', usevar = 'pooled', value = 0) #
df is the degree of freedom
reject = pvalue < alpha  # alpha = 0.05

print(f"t-stat (t): {tstat}")
print(f"P-value (p): {pvalue}")
print(f"Do we reject H_0? {reject}")
```

[11] ✓ 0.0s Python

```
t-stat (t): 0.7071067811865475
P-value (p): 0.5185185185185183
Do we reject H_0? False
```

Python interlude

Concepts We Will Learn

1. Sampling Distribution
2. Standard Error and Confidence Interval
3. Null Hypothesis and P-Value
4. **Statistical Power**

How many people do we need in our experiment?

- **Statistical Power:**
  - Probability of rejecting the null when there is a true treatment effect of some size.
- Suppose we have two treatments: A and B.
  - A has a true ATE of 10
  - B has a true ATE of 1
- It is easier to detect an effect with treatment A since it has a bigger effect.
- It is more likely that we reject the null with treatment A.

What do we need for a power analysis?

- **Minimum detectable effect.** How big of an effect do we want to detect with high probability?
- **Standard error of our estimate of the average treatment effect.** Intuition: the more uncertain our estimate is, the less likely it is we reject the null.
- How do we know the standard error before running the experiment? We don't, but we know something about what determines it.

What determines the standard error of $\widehat{ATE}$

- The variance of the outcomes.

  For example, if our outcome is revenue and we are selling cars, variance will be higher than if we are selling pens.

- The number of observations. The more observations, the lower the standard error.

Summarizing

Power increases with:

- The size of the effect
- The sample size.

Power decreases with:

- Variation in the outcome.

Cohen's D: Common Measure of Effect Size

$$\frac{\textit{Effect Size}}{\textit{Standard Deviation of Y}}$$

How to do power analysis in Python?

```
# Statistical Power
from statsmodels.stats.power import TTestPower

true_effect = 1
sd_outcome = np.std(outcome)
n = TTestPower().solve_power(effect_size = true_effect/sd_outcome,
                             nobs = None, alpha = .05, power = 0.8, alternative='two-sided')

print(f"Necessary sample size: {np.ceil(n)}")
```

```
Necessary sample size: 6041.0
```

```
TTestPower().power(effect_size = true_effect / sd_outcome,
                   alpha = .5, nobs = 300, alternative='two-sided')
```

```
0.5769761145963959
```

How many observations do we need?

```python
# Statistical Power
from statsmodels.stats.power import TTestPower

true_effect = 1
sd_outcome = np.std(outcome)
n = TTestPower().solve_power(effect_size = true_effect/sd_outcome,
                             nobs = None, alpha = .05, power = 0.8, alternative='two-sided')

print(f"Necessary sample size: {np.ceil(n)}")
```

```
Necessary sample size: 6041.0
```

Observations Needed

What is the power?

```
]  TTestPower().power(effect_size = true_effect / sd_outcome,
                      alpha = .5, nobs = 300, alternative='two-sided')
```

```
0.5769761145963959
```

Minimum Detectable Effect for 200 Observations

```
TTestPower().solve_power(effect_size = None,
                         power = 0.8,
                         alpha = 0.05,
                         nobs = 200)
```

✓ 0.0s

```
0.19906466118162786
```

Cohen's D of 20%

Python interlude

Summarizing

1. Sampling Distribution
2. Standard Error and Confidence Interval
3. Null Hypothesis and P-Value
4. Statistical Power