Hypothesis Testing with SciPy

Hillary Green-Lerman
@codecademy

codecademy

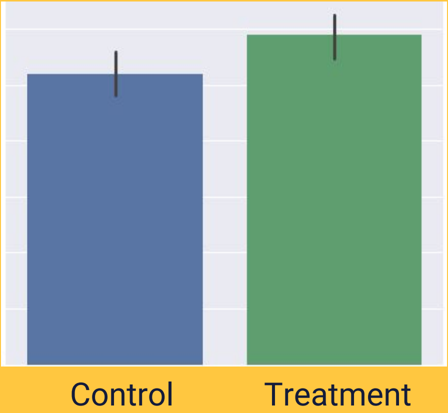
Why Test?

Example Problems

Scenario 1:

We're testing a new drug to raise good cholesterol levels. After six weeks, the mean cholesterol level in the treatment group is slightly higher than in the control group.

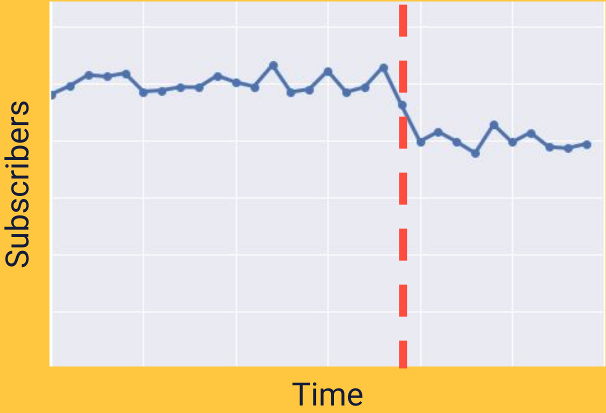
Does our new drug work?



Scenario 2:

Monthly subscriptions to our streaming service dipped after we released a new version.

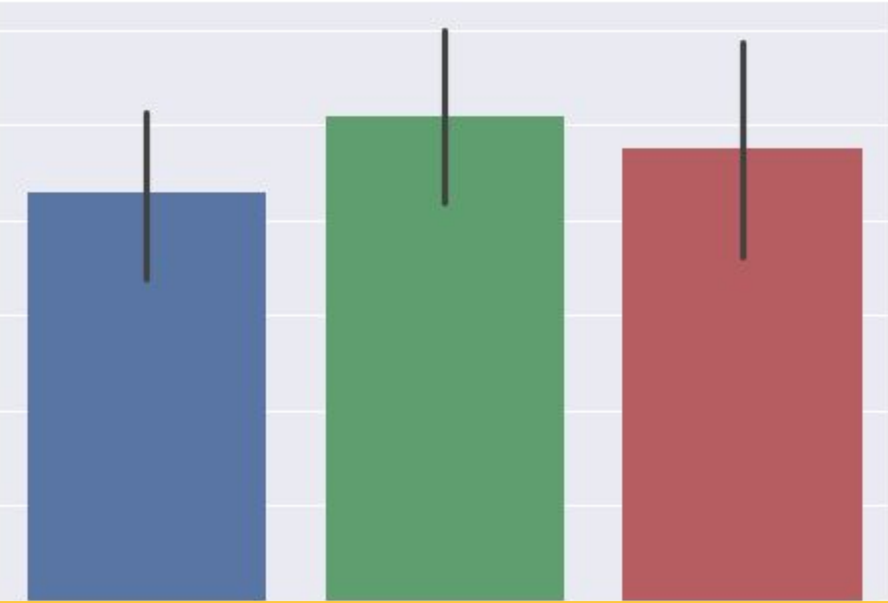
Is there a real change in subscription rates?



Not Every Difference is Significant

These three samples were generated from the same command:

```
np.random.normal(loc=50, scale=25, size=30)
```



What does it mean to be significant?

An observed difference between two quantities is **probably *not* due to chance**.

What does it mean to be *not* significant?

There is **not enough evidence** to say that an observed difference between two quantities is not random.

OR

If there is a difference, it is **smaller than we care about**.

Why do use want to learn hypothesis testing?

Hypothesis Testing

Hypothesis Testing

- Reframes our qualitative question (*"Is this difference real?"*) into a mathematical question (*"What is the **probability** that the difference I am observing is due to chance?"*)

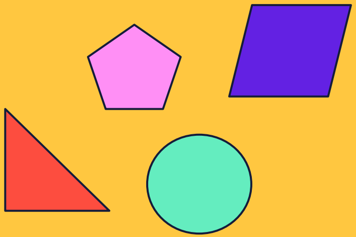
- Goal: reject the **null hypothesis**: *"The two populations I am comparing are identical and the differences I observe are due to chance."*

- We reject the null hypothesis by proving that the proving that it unlikely. We do that by calculating the **p-value** (using a hypothesis test). We generally want $p < 0.05$ (i.e., there's only a 5% chance that two identical distributions could have produced these results)

Question 1:
Is my data Categorical or Numerical?

- A professor expects an exam average to be roughly 75%, and wants to know if the actual scores line up with this expectation. Was the test actually too easy or too hard?
- A PM for a website wants to compare the time spent on different versions of a homepage. Does one version make users stay on the page significantly longer?

- A pollster wants to know if men and women have significantly different yogurt flavor preferences. Does a result where men more often answer "chocolate" as their favorite reflect a significant difference in the population?
- Do different age groups have significantly different emotional reactions to different ads?



Question 2:
How many samples am I comparing?

- 1 Sample (i.e., comparing to an ideal target)
  i.e., comparing an actual result against a desired target or KPI

- 2 Sample
  i.e., comparing a control and treatment group or an A/B test

- More than 2 Sample
  i.e., comparing three different variants of a landing page

Hypothesis Testing Options

		What type of data do you have?	
		Numerical	**Categorical**
What type of comparison are you making?	**Sample vs. Known Quantity or Target**	1 Sample T-Test	Binomial Test
	2 Samples	2 Sample T-Test	Chi Square
	More than 2 Samples	ANOVA and/or Tukey	

Let's Practice!

Long Link:

https://www.codecademy.com/courses/learn-scipy-hypothesis-testing/lessons/hypothesis-testing/

Short Link:

https://bit.ly/2H0ykMx

1 Sample T-Test

<table>
<tr><td><b>When to Use</b></td><td>Compares a sample mean to a hypothetical population mean. It answers the question "What is the probability that the sample came from a distribution with the desired mean?"
Use this when you are comparing against a known target (like a statistic from a paper or a target metric).</td></tr>
<tr><td><b>Usage</b></td><td>ttest_1samp requires two inputs, a distribution of values and an expected mean:<pre>
tstat, pval = ttest_1samp(example_distribution,
                          expected_mean)
print(pval)
</pre></td></tr>
</table>

2 Sample T-Test

<table>
<tr><td><b>When to Use</b></td><td>A 2 Sample T-Test compares two sets of data, which are both approximately normally distributed.

The null hypothesis, in this case, is that the two distributions have the same mean. Use this when you are comparing two different numerical samples.</td></tr>
<tr><td><b>Usage</b></td><td><code>ttest_ind</code> requires two distributions of values:<pre>
tstat, pval = ttest_ind(example_distribution1,
                                    example_distribution2)
print(pval)
</pre></td></tr>
</table>

ANOVA

<table>
<tr>
<td><b>When to Use</b></td>
<td>ANOVA compares more than 2 numerical datasets without increasing the probability of a false positive.
In order to use ANOVA,
1. The samples should be normally distributed (ish)
2. The standard deviations of the data should be similar (ish)
3. The samples should be independent</td>
</tr>
<tr>
<td><b>Usage</b></td>
<td>ttest_ind requires two distributions of values:
<pre>
fstat, pval = f_oneway(sample1,
                       sample2,
                       sample3)
</pre>
</td>
</tr>
</table>

Tukey

When to Use

Tukey's Range Test compares more than 2 numerical datasets without increasing the probability of a false positive. Unlike ANOVA, Tukey tells us ***which*** datasets are significantly different. Many statisticians use Tukey instead of Anova.

Note: pairwise_tukeyhsd is from StatsModels, not SciPy!

Usage

pairwise_tukeyhsd requires three arguments:

- A vector of all data (concatenated using np.concatenate)
- A vector of labels for the data
- A level of significance (usually 0.05)

```
v = np.concatenate([a, b, c])
labels = ['a'] * len(a) + ['b'] * len(b) + ['c'] * len(c)
tukey_results = pairwise_tukeyhsd(v, labels, 0.05)
```

Binomial Test

When to Use

Compares an observed proportion to a theoretical ideal.

Examples:

- Comparing the actual percent of emails that were opened to the quarterly goals
- Comparing the actual percentage of respondents who gave a certain survey response to the expected survey response

Usage

`binom_test` requires three arguments:

- The number of successes (the numerator of your proportion)
- `n` - the number of trials (the denominator of your proportion)
- `p` - the proportion you are comparing to

```
pval = binom test(numerator,
                  n=denominator,
                  p=proportion)
print(pval)
```

Chi Squared Test

When to Use

If we have two or more categorical datasets that we want to compare, we should use a Chi Square test. It is useful in situations like:

- An A/B test where half of users were shown a green submit button and the other half were shown a purple submit button. Was one group more likely to click the submit button?
- Men and women were both given a survey asking "Which of the following three products is your favorite?" Did the men and women have significantly different preferences?

Usage

chi2_contingency requires a contingency table of all results:

```
chi2, pval, dof, expected = \
    chi2 contenigency([[cat1 yes,cat1 no],
                        [ cat2 yes,cat2 no])
print(pval)
```

Experimental Design

How to Experiment

1. State your hypothesis
2. Pick a metric to track
3. Calculate your metric for the control group
4. Select the smallest difference that you want to be able to detect
5. Select your split
   Will equal numbers of people see control and variant?
6. Calculate desired sample size
7. Run your experiment
8. Perform a hypothesis test

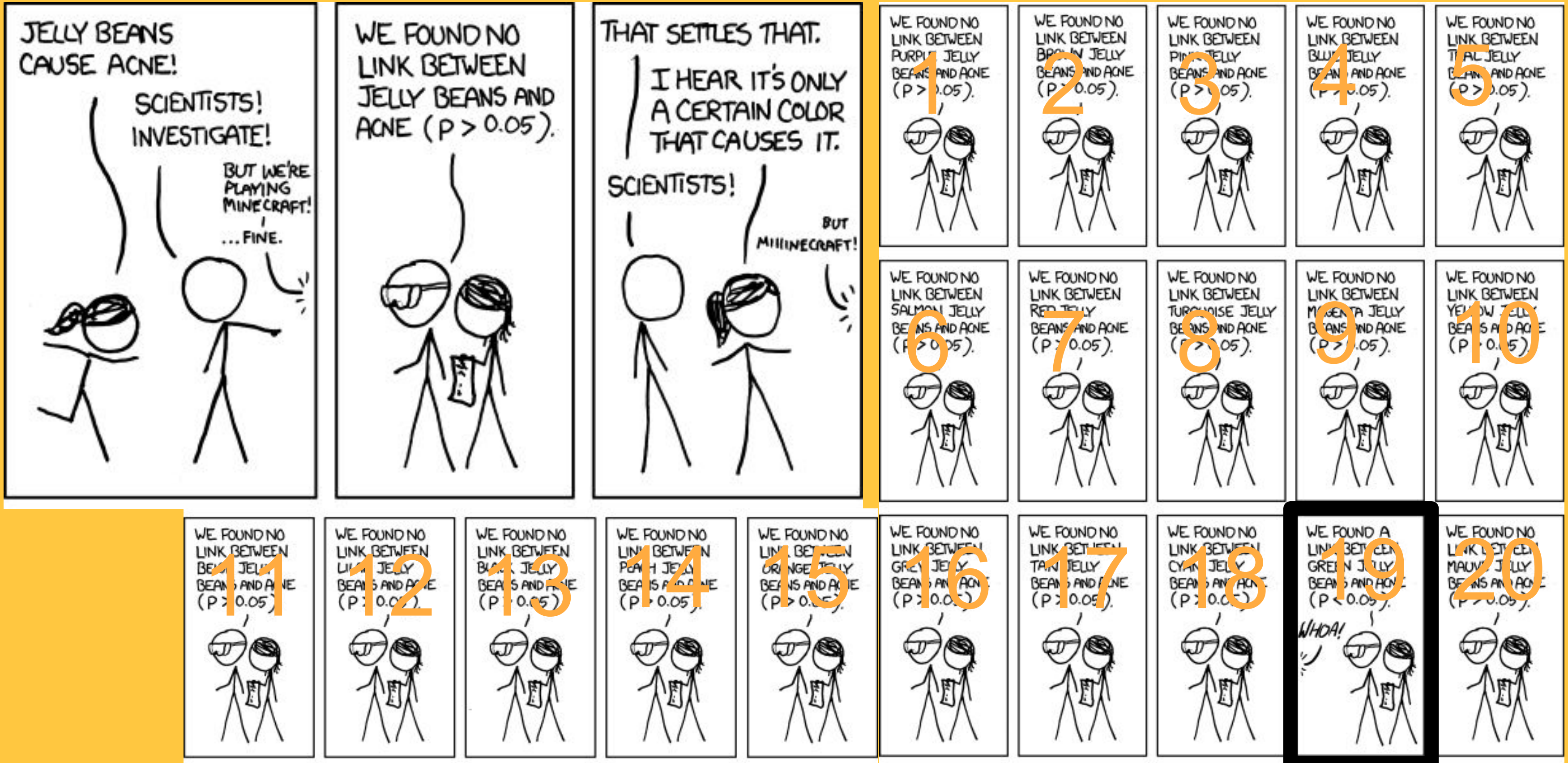
What if my results aren’t significant?

- Effectively, it means:
  “We’re pretty sure that ***if*** there is a difference between A and B, it’s smaller than X”
  - Q: How do we know X?
  - A: Sample size determination!
- Does not mean:
  - We need to run longer
  - There is **definitely** no difference between A and B

What Not to Do

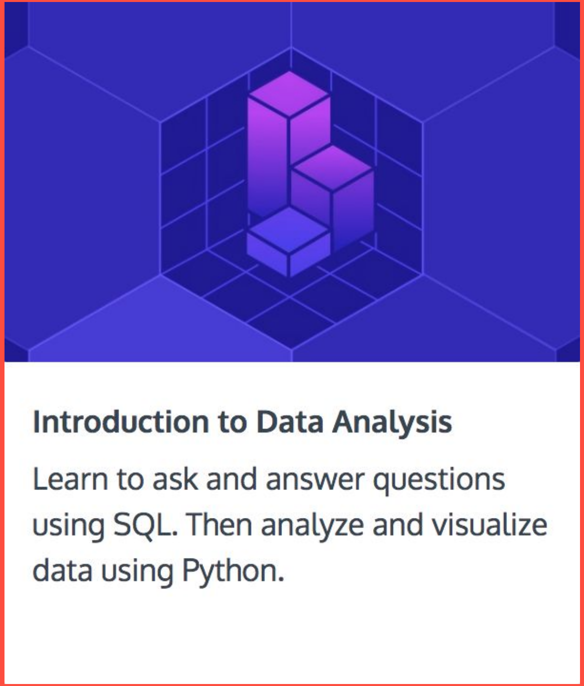
- Oh no! My experiment isn’t significant. Let’s run it longer!
- Yay! My experiment is already significant! Let’s kill it.

P-Hacking or Peaking


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
1
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
2
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
3
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
4
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
5
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
6
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
7
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
8
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
9
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
10
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
11
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
12
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
13
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
14
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
15
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
16
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
17
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
18
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
19
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
20
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).

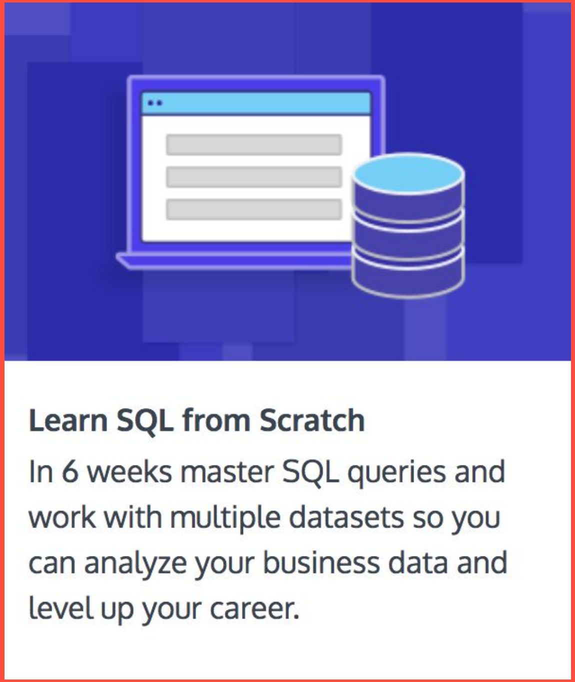
Want to Learn More?

codecademy



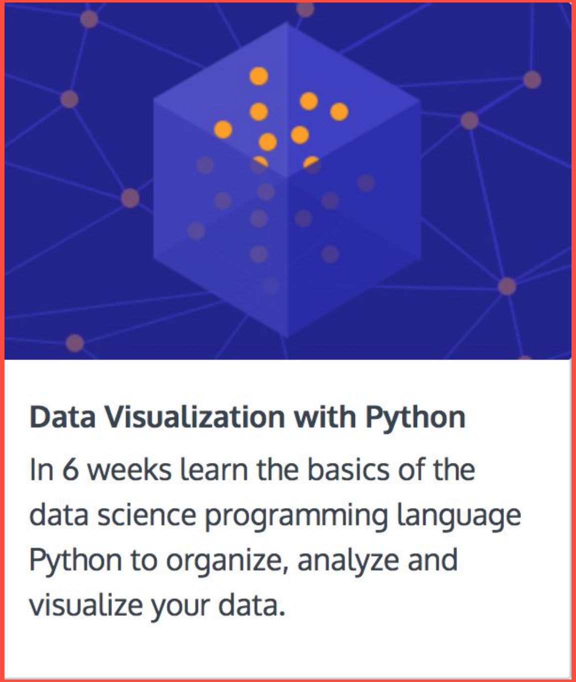
Introduction to Data Analysis

Learn to ask and answer questions using SQL. Then analyze and visualize data using Python.



Learn SQL from Scratch

In 6 weeks master SQL queries and work with multiple datasets so you can analyze your business data and level up your career.



Data Visualization with Python

In 6 weeks learn the basics of the data science programming language Python to organize, analyze and visualize your data.