

# Wrapping Up Hypothesis Tests: Guidelines, Errors and Caveats

## LECTURE 21

**CSCI 3022**

Maribeth Oscamou

Content credit: [Acknowledgments](#)

# Course Logistics: 8th Week At A Glance

Mon 3/4	Tues 3/5	Wed 3/6	Thurs 3/7	Fri 3/8
Attend & Participate in Class  Lesson: Hypothesis Tests with a Single Sample	(Optional): Attend Notebook Discussion with our TA (5-6pm Zoom)	Attend & Participate in Class  Lesson: Hypothesis Tests with Multiple Samples: A/B Tests, Randomized Controlled Trials and Causality	HW 7 Due 11:59pm	Attend & Participate in Class Lesson: Errors in Hypothesis Tests  QUIZ 5: Scope: L13-L16, nb 7, HW 6  HW 8 Released

### ⋮ ▼ Lesson Materials Week 8

⋮  [Lesson 20: AB Testing and Randomized Controlled Tests](#) 

⋮  [Lesson 20 Jupyter Demo: AB Testing, RCT](#) 

⋮  [Hypothesis Testing Guide.pdf](#)

# Today's Roadmap

---

CSCI 3022

- Recap/Review of Hypothesis Tests
- Finish Lesson 20:
  - Randomized Controlled Trials
- Wrapping Up Hypothesis Tests:
  - Errors
  - Statistical Power

## Recap: Steps in Hypothesis Testing

---

- **Define the null hypothesis and the alternative hypothesis**
- **Choose a significance level** (cutoff tail probability after which you will decide the null hypothesis is inconsistent with the observed data)
- **Choose a test statistic** to measure “discrepancy” between null hypothesis and data
- **Simulate the test statistic (or calculate directly when possible)** under the null assumptions
- **Gather observed data and compare** to the null hypothesis predictions:
  - Draw a histogram of (simulated) values of the test statistic
  - Compute the **observed test statistic and the p-value** from the real sample
- If the **p-value is less than (or equal to)** the significance level: Reject Null and Accept Alternative. Otherwise Fail to Reject Null.

## Hypothesis Test Concerns

---

The outcome of a hypothesis test can be affected by:

- **The hypotheses you investigate:**  
*How do you define your null distribution?*
- **The test statistic you choose:**  
*How do you measure a difference between samples?*
- **The empirical distribution of the statistic under the null:**  
*How many times do you simulate under the null distribution?*
- **The data you collected:**  
*Did you happen to collect a sample that is similar to the population?*
- **The truth:**  
*If the alternative hypothesis is true, how extreme is the difference?*

### Null Hypothesis:

- Null is meant to describe lack of an interesting pattern
  - Difference in sample results from null are **due to chance**
- Need to be able to either calculate null distribution theoretically or simulate data under the null hyp.

### Alternative Hypothesis:

- Should align with the question of interest

**Null** and **alternative** hypothesis can't be true at the **same time**. (Reject the Null → Alternative)

*"It is convenient to take this point [5%] as a limit in judging whether a deviation is to be considered significant or not." [Fisher 1925]*

## Guidelines for Choosing the Significance Level (i.e. p-value cutoff)

- Decide on it **before** seeing the results
  - Don't change it!
- Common values at 5% and 1%
  - follow conventions in your area

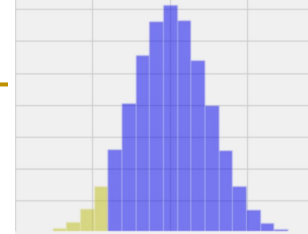
*"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the author [Fisher] prefers to set a low standard of significance at the 5 percent point ..." [Fisher 1926]*



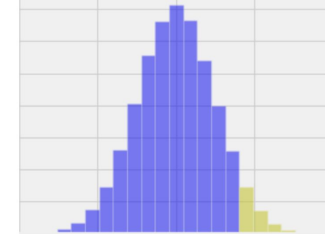
**Sir Ronald Aylmer Fisher [1890-1962]**  
**Pioneer of Modern Statistics**



## Review: Definition of the $p$ -value



Simulated values when **LOW** test statistics support the alternative hypothesis



Simulated values when **HIGH** test statistics support the alternative hypothesis

The  $p$ -value is the chance (probability),

- under the null hypothesis (i.e. given the null)
  - that the test statistic
  - is equal to the value that was observed in the data
  - or is even further in the direction of the alternative.
- Yellow area denotes the  $p$ -value
  - Red dot denotes the observed statistic.

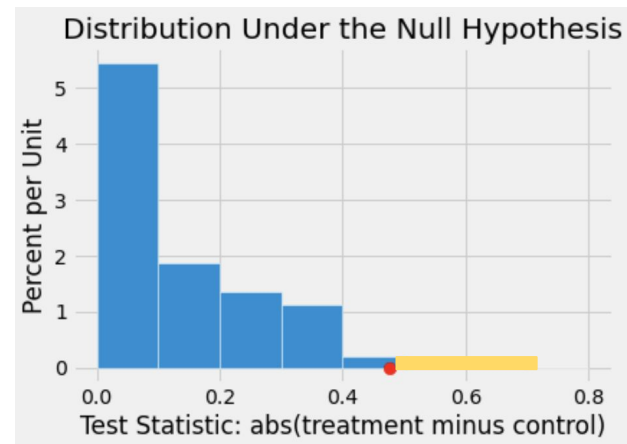
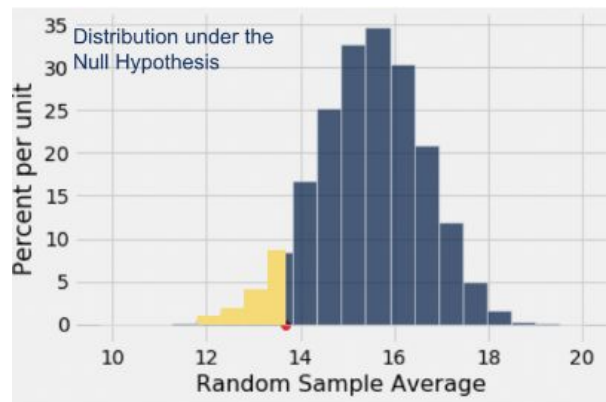
Formal name: **observed significance level**

Notice: The  $p$ -value is actually a Conditional Probability!

$$p\text{-value} = P(\text{observed data or more extreme} \mid \text{null hypothesis})$$

## Review: The p-Value as an Area

- Empirical distribution of the test statistic **under the null hypothesis**.
- Red dot denotes the observed statistic.
- Yellow area denotes the p-value



# Theoretical p-value vs Empirical p-value

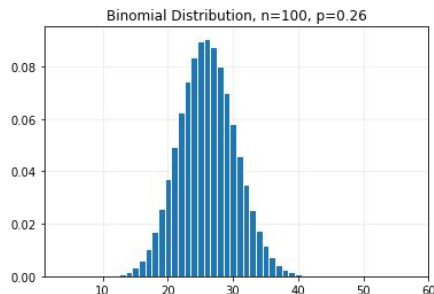
Recall: any random variable has a distribution:

## Probability (aka Population or Theoretical)

Distribution These are the distributions (pmf or pdf) of random variables or the distribution of some feature of some population.

```
k = np.arange(101)
p = special.comb(100, k)*(0.26**k)*(0.74**(100-k))

fig, ax = plt.subplots()
ax.bar(k, p, width=1, ec='white');
ax.set_axisbelow(True)
ax.grid(alpha=0.25)
plt.xlim(1,60)
plt.title("Binomial Distribution, n=100, p=0.26");
```



The p-value calculated using the **theoretical distribution** under the null hypothesis is called the **theoretical p-value (or just the p-value)**.

As your number of simulations increases, the empirical p-value will converge to the theoretical p-value

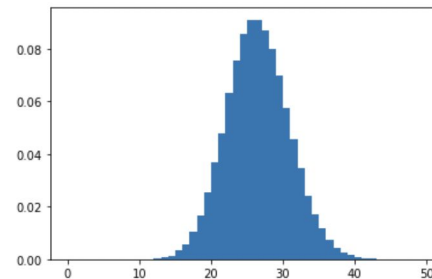
- Empirical (aka Simulated or Sample ) Distribution: based on random samples (or simulations)
- Observations can be from **repetitions of an experiment or random samples from a population**
  - All observed values
  - The proportion of times each value appears

```
#Simulate one experiment
def heads_in_n_tosses(n=100):
    return sum(np.random.choice(["H","T"],size=n,p=[.26, .74]) == 'H')

# Repeat the experiment m times:
num_simulations = 50000;
outcomes=[]

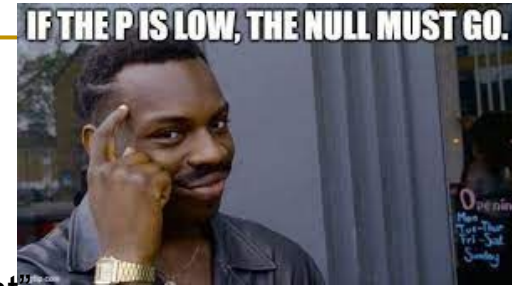
for i in np.arange(num_simulations):
    outcomes = np.append(outcomes, heads_in_n_tosses())

plt.hist(outcomes,bins=np.arange(0,50), density=True);
```



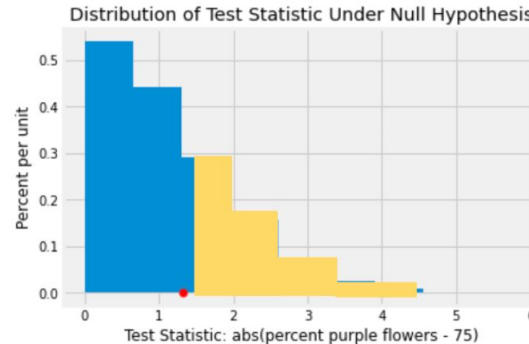
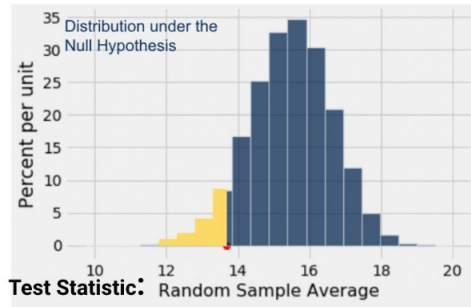
The p-value calculated using the **empirical distribution** under the null hypothesis is called the **empirical p-value**.

# Recap: Making a Concluding Decision in a Hypothesis Test



## Conclusion of Test:

- If  $p\text{-value} \leq \text{your predetermined significance level}$ 
  - Reject null and accept alternative
    - If  $p\text{-value} \leq 5\%$  it is called “statistically significant”
    - If  $p\text{-value} \leq 1\%$ , it is called “highly statistically significant”
- Else
  - Fail to reject null hypothesis



Red dot denotes the observed test statistic.

Yellow area denotes the p-value

# Back to Lesson 20

---

CSCI 3022

- Recap/Review of Hypothesis Tests
- **Finish Lesson 20:**
  - Randomized Controlled Trials

# Errors in Hypothesis Testing

---

CSCI 3022

- Finish Lesson 20:
  - Randomized Controlled Trials
- Wrapping Up Hypothesis Tests:
  - **Errors**
  - Statistical Power

## Discussion Question

---

Suppose there are 500 students enrolled in CSCI 3022. We give each student a separate coin and have them toss it 160 times to test whether or not the coin is fair.

**Null:** The coin is \_\_\_\_\_

**Alternative:** The coin is \_\_\_\_\_

- Test Statistic: \_\_\_\_\_
- Significance level (cutoff for the P-value): 5%

**Suppose in reality all the coins are fair.**

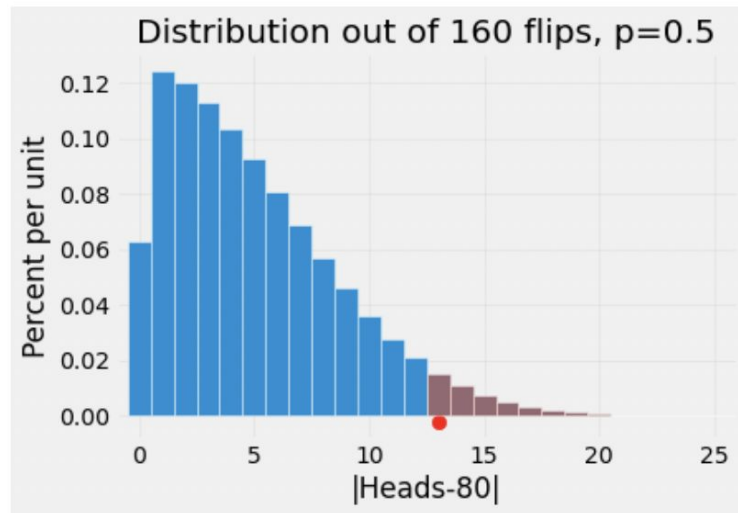
**About how many students will conclude that their coins are unfair using this hypothesis test?**

- A). 5                      B). 25                      C). 50                      D). 120                      E). 160

(Demo)

## Significance Level as an Error Probability

- If:
  - your **significance level (i.e. p-value cutoff)** is 5%
  - and the **null hypothesis happens to be true**
- Then there is a **5% chance** that **the test will INCORRECTLY reject the null hypothesis**.

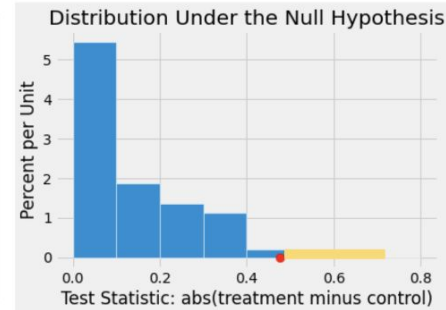


When null is true,  
5% of the time you  
will get an observed  
test statistic in tail  
shaded pink even  
when the coin is fair  
**JUST BY CHANCE!**



## Recap: Making a Concluding Decision in a Hypothesis Test



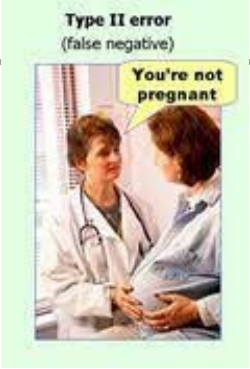

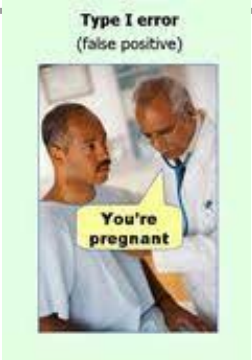

- P-value (You Compute It)
  - Depends on the observed data and simulation
  - Probability under the null hypothesis that the test statistic is the observed value or more extreme
  - $P(\text{data you observed or more extreme} \mid \text{null hypothesis})$
- Significance level (i.e. P-value cutoff ): You Pick It Before You Start
  - Does not depend on observed data or simulation
  - “Acceptable” probability of rejecting the null hypothesis when it is true. Common conventions: 5% or 1%









- Red dot denotes the observed statistic.
- Yellow area denotes the tail probability (p-value).

# Can the Conclusion be Wrong?

Yes.

	Null is true	Null is False
Test fails to reject null		 
Test rejects null	 	

Probability of this type of error is equal to the significance level.  
So choose a small significance level to control this error

	Null is true	Null is False
Test fails to reject null		 
Test rejects null	 	

Choose a small significance level to control this error

How do we minimize this type of error?

# Statistical Power





---

CSCI 3022

- Finish Lesson 20:
  - Randomized Controlled Trials
- Wrapping Up Hypothesis Tests:
  - Errors
  - **Statistical Power**

**Definition:** The **Statistical Power** of a hypothesis test is the probability of correctly rejecting the null when it is false.

Goal: Typical conventions are to set-up your test such that this is at least 80%.

	Null is true	Null is False
Test favors null		
Test rejects null		

**The Statistical Power** of a test is the probability of correctly rejecting the null when it is false.

## Back to our example

---

Suppose there are 500 students enrolled in CSCI 3022. We give each student a separate coin and have them toss it 160 times to test whether or not the coin is fair.

**Null:** The coin is fair

**Alternative:** The coin is unfair

- Test Statistic:  $|\text{num of heads} - 80|$
- Significance level (cutoff for the P-value): 5%

**Suppose in reality all the coins are UNFAIR, with  $P(H) = 45\%$**

**About how many students will CORRECTLY conclude that their coins are UNFAIR using this hypothesis test?**

- A). 5                      B). 25                      C). 50                      D). 120                      E). 160**

- **Definition:** The power of a hypothesis test is the probability of correctly rejecting the null hypothesis when the alternative is true.
  - $\text{Prob}(\text{rejecting null hypothesis} \mid \text{null is false})$

**Convention:** We usually try to design hypothesis tests so the Power is at least 80%.

**For calculating power or required sample size, there are four moving parts:**

- 1). Sample Size
- 2). Significance level (the p-value cutoff you chose)
- 3). Effect size (the minimal size of the effect you hope to be able to detect in a statistical test, such as a 5% difference in probability of heads or a 20% improvement in click rates on a website.
- 4). Power

**Specify any 3 of the above and the 4th is completely determined.**

**Most commonly, you would want to calculate sample size, so you must specify the other three**

## Steps for Estimating the Power of a Hypothesis Test

---

1. Start with some hypothetical data that represents your best guess about the data that will result (perhaps based on prior data)—for example, a box with 20 ones and 80 zeros to represent a .200 hitter, or a box with some observations of “time spent on website.”
2. Create a second sample simply by adding the desired effect size to the first sample—for example, a second box with 33 ones and 67 zeros, or a second box with 25 seconds added to each initial “time spent on website.”
3. Draw a bootstrap sample of size  $n$  from each box.
4. Conduct a permutation (or formula-based) hypothesis test on the two bootstrap samples and record whether the difference between them is statistically significant.
5. Repeat the preceding two steps many times and determine how often the difference was significant—that’s the estimated power.



If aren't using theoretical distribution, then need to decide number of simulations:

- large as possible: empirical distribution  $\rightarrow$  true distribution
- No new data needs to be collected (yay!)

**Number of observations:**

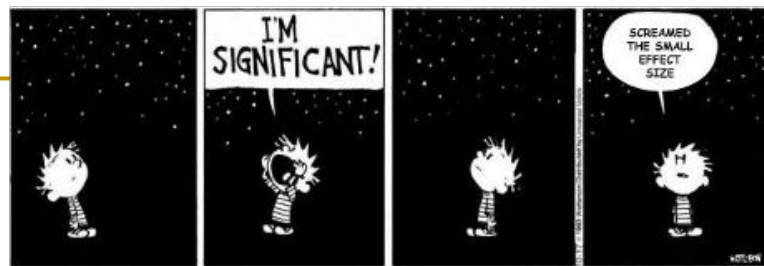
- A larger sample will lead you to reject the null more reliably if the alternative is in fact true (higher “statistical power”).

**Difference from the null:**

- If truth is similar to the null hypothesis (“small effect size”), then even a large sample may not provide enough evidence to reject the null.

# Statistically Significant vs “Practically” Significant

## Effect Size vs Statistical Significance:



- Statistical significance: After accounting for random sampling error, your sample suggests that a non-zero effect exists in the population.
- Effect sizes: The magnitude of the effect. It answers questions about how much or how well the treatment works. Are the relationships strong or weak?

No statistical test can tell you **whether the effect is large enough to be important** in your field of study. Instead, you need to apply your subject area knowledge and expertise to determine whether the effect is big enough to be meaningful in the real world. In other words, is it large enough to care about?

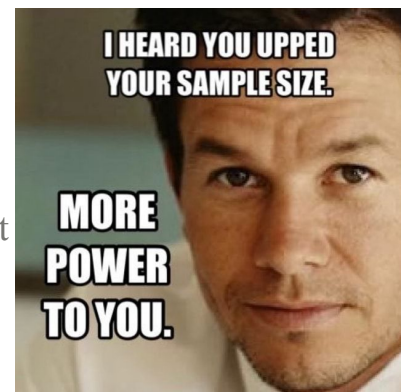
# Statistically Significant vs “Practically” Significant

Not all statistically significant differences are interesting!

- Here's how **small effect sizes** can still produce **tiny p-values**:
  - **You have a very large sample size.** As the sample size increases, the hypothesis test gains greater statistical power to detect small effects. With a large enough sample size, the hypothesis test can detect an effect that is so minuscule that it is meaningless in a practical sense.
  - **The sample variability is very low.** When your sample data have low variability, hypothesis tests can produce more precise estimates of the population's effect. This precision allows the test to detect tiny effects.

We need a method to determine whether the estimated effect (i.e. the difference between the treatment group and the control group) is still practically significant when you factor in the margin of error from sampling.

**Solution: Up Next - Confidence Intervals!**



# P-Hacking

---

CSCI 3022

- Finish Lesson 20:
  - Randomized Controlled Trials
- Wrapping Up Hypothesis Tests:
  - Errors
  - Statistical Power
  - **P-Hacking**

Ex: Suppose you do 20 different hypothesis tests (testing the relationship between jelly beans and acne) with a null hypothesis that there's no relationship. Assume you conduct each test at a significance level of 0.05.

If in reality *jelly beans aren't actually linked with acne*, what's the probability that NONE of our 20 tests are significant (i.e. that all of our tests correctly fail to reject the null)?

Poll:

A). ~95%

C). ~36%

E). ~20%

B). ~50%

D). ~5%

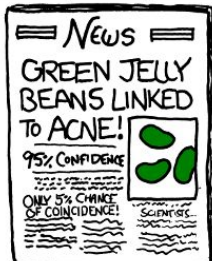
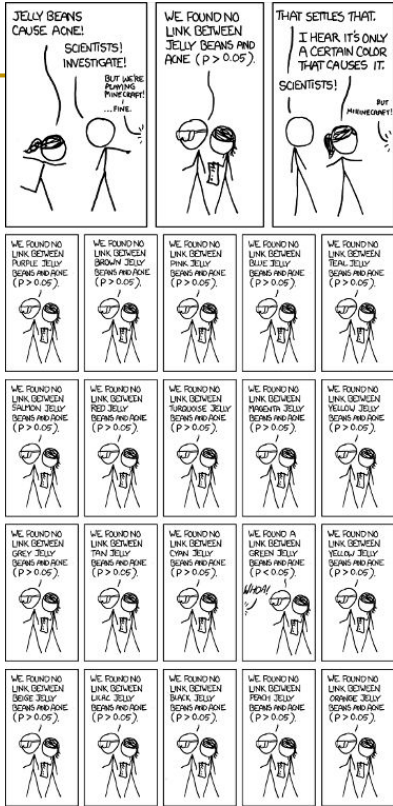
# Beware of P-Hacking

Ex: Suppose you do 20 different hypothesis tests (testing the relationship between jelly beans and acne) with a null hypothesis that there's no relationship. Assume you conduct each test at a significance level of 0.05.

If in reality jellybeans aren't actually linked with acne, what's the probability that NONE of our 20 tests are significant (i.e. that all of our tests correctly don't reject the null)?

$$0.95^{20} = 0.3584859224$$

THAT MEANS THAT ABOUT 64% OF THE TIME, ONE OR MORE OF THESE TESTS WILL BE SIGNIFICANT, JUST BY CHANCE, EVEN THOUGH JELLY BEANS HAVE NO EFFECT ON ACNE.



# Practice

---

CSCI 3022

## Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

**Question:** Do fewer people prefer Super Soda than its rival, or is this just chance?

**Null hypothesis:**

**Alternative hypothesis:**

**Test statistic:**

**p-value: Start at the observed statistic and look which way?**



## Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

**Question:** Do fewer people prefer Super Soda than its rival, or is this just chance?

**Null hypothesis:** The same proportion of people prefer Super as Rival

**Alternative hypothesis:**

**Test statistic:**

**p-value:** Start at the observed statistic and look which way?

## Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

**Question:** Do fewer people prefer Super Soda than its rival, or is this just chance?

**Null hypothesis:** The same proportion of people prefer Super as Rival

**Alternative hypothesis:** A smaller proportion of people prefer Super

**Test statistic:**

**p-value:** Start at the observed statistic and look which way?

## Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

**Question:** Do fewer people prefer Super Soda than its rival, or is this just chance?

**Null hypothesis:** The same proportion of people prefer Super as Rival

**Alternative hypothesis:** A smaller proportion of people prefer Super

**Test statistic:** Number of people (out of 200) who prefer Super

**p-value:** Start at the observed statistic and look which way?

## Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

**Question:** Do fewer people prefer Super Soda than its rival, or is this just chance?

**Null hypothesis:** The same proportion of people prefer Super as Rival

**Alternative hypothesis:** A smaller proportion of people prefer Super

**Test statistic:** Number of people (out of 200) who prefer Super

**p-value:** Start at the observed statistic and look which way? **LEFT**

**Conduct the test (DEMO)**

(Demo)

## Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

**Question:** Do fewer people prefer Super Soda than its rival, or is this just chance?

**Null hypothesis:** The same proportion of people prefer Super as Rival

**Alternative hypothesis:** A smaller proportion of people prefer Super

**Test statistic:** Number of people (out of 200) who prefer Super

**p-value:** Start at the observed statistic and look which way? LEFT

**Conduct the test (See NB 21)**

**What types of errors might result from this hypothesis test and how can we minimize them?**

(see nb 21)