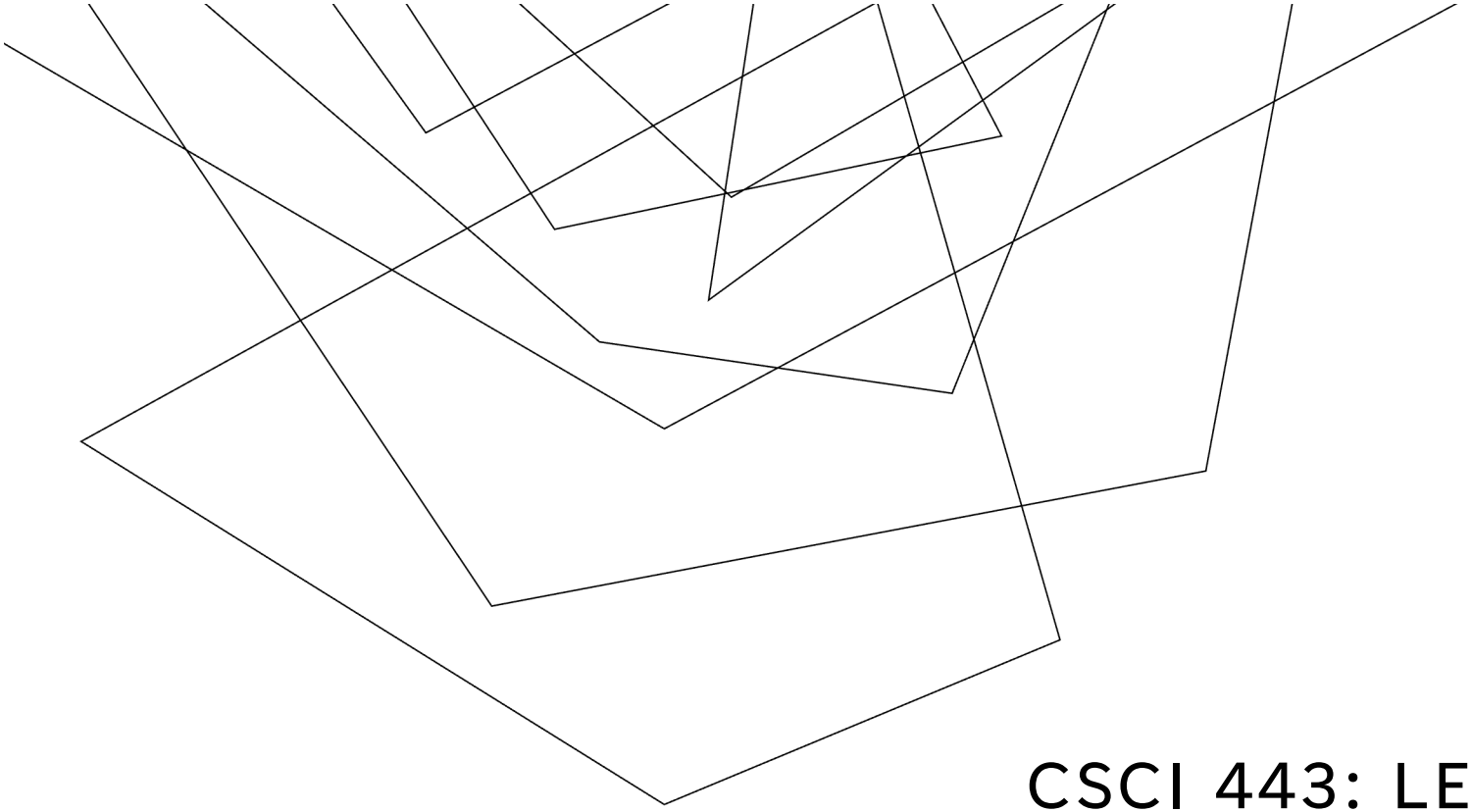


# CSCI 443: LECTURE 18 PERMUTATION TESTS.

Professor David Harrison

# OFFICE HOURS

| | |
|---|---|
| Tuesday | 4:00-5:00 PM |
| Wednesday | 12:30-2:30 PM |

# BLACKBOARD & GITHUB

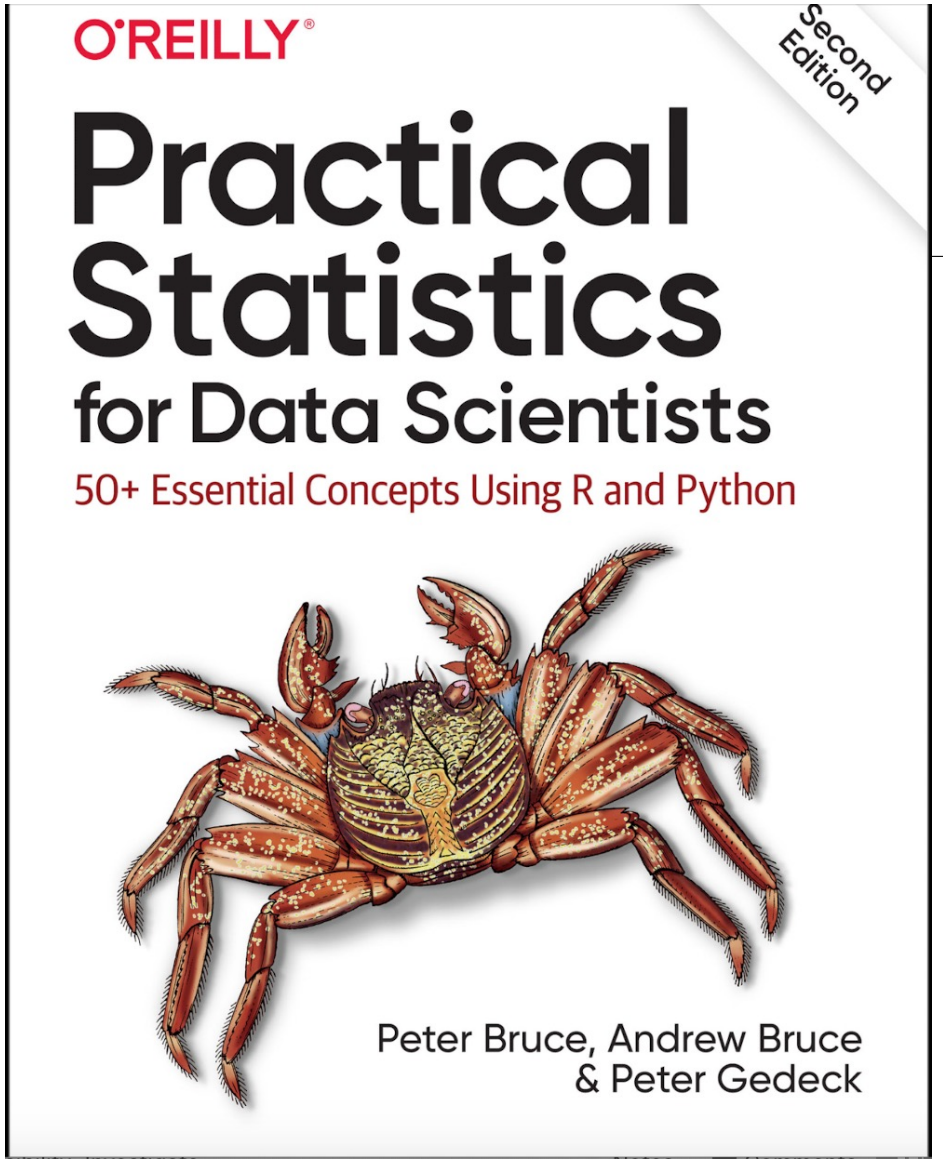
Slides up, handwritten notes AND a jupyter notebook for lecture 16 are on blackboard and in GitHub.

The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience

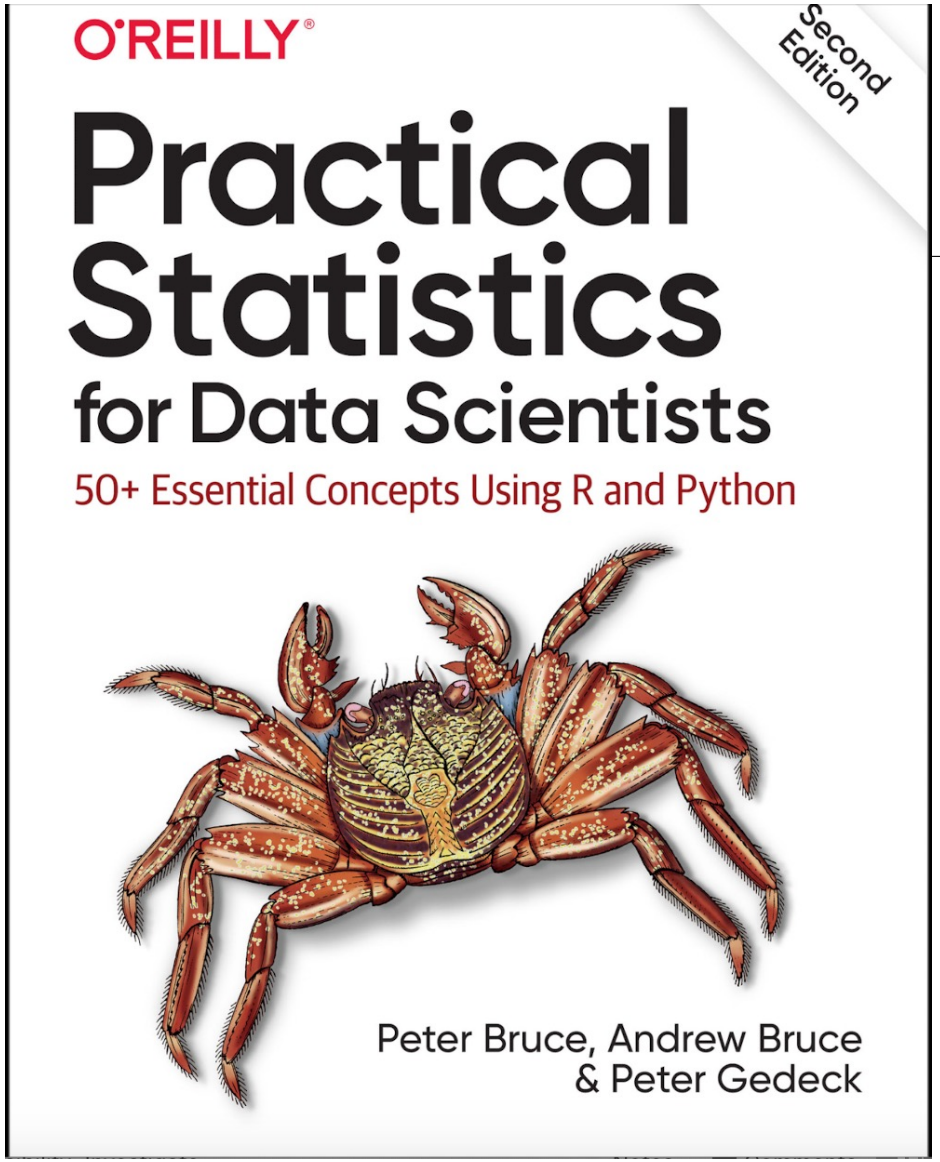
## READ ABOUT

- chapter 3: experiments, hypothesis testing
  - Permutation Tests
  - T-tests
  - ANOVA
  - Chi-square



## THINGS I WANT TO COVER TODAY

- Permutation Tests
- F-Statistic



# PREVIOUS LECTURE: AD COMPARISON

A:


B:

1000 views of each.

0.1% click through rate for A.

0.5% click through rate for B.

This is a hypothesis test between binomial random variables.

# MODIFIED FROM PREVIOUS LECTURE: AD COMPARISON

$\hat{p}_A$ = click through rate for A
= Bin($n_A$, $p_A$) / $n_A$

$\hat{p}_B$ = click through rate for B
Bin($n_B$, $p_B$) / $n_B$

A:


B:

# NON-BINOMIAL HYPOTHESIS TESTS

What if hypothesis test for random variables that are NOT binomial.

Example: trial for Novo Lilly's new life-extension drug Zombivia.

*X* = systolic blood pressure for patient in control group.

*Y* = systolic blood pressure for patient in exposed group.

*X* and *Y* are continuous numerical random variables

$$H_0 : \mu_x = \mu_y$$

$$H_A : \mu_x \neq \mu_y$$

# ZOMBIVIA TRIAL: STEP 1

**Step 0**: Choose alpha

$$\alpha = 0.05$$

**Step 1: Collect Data**

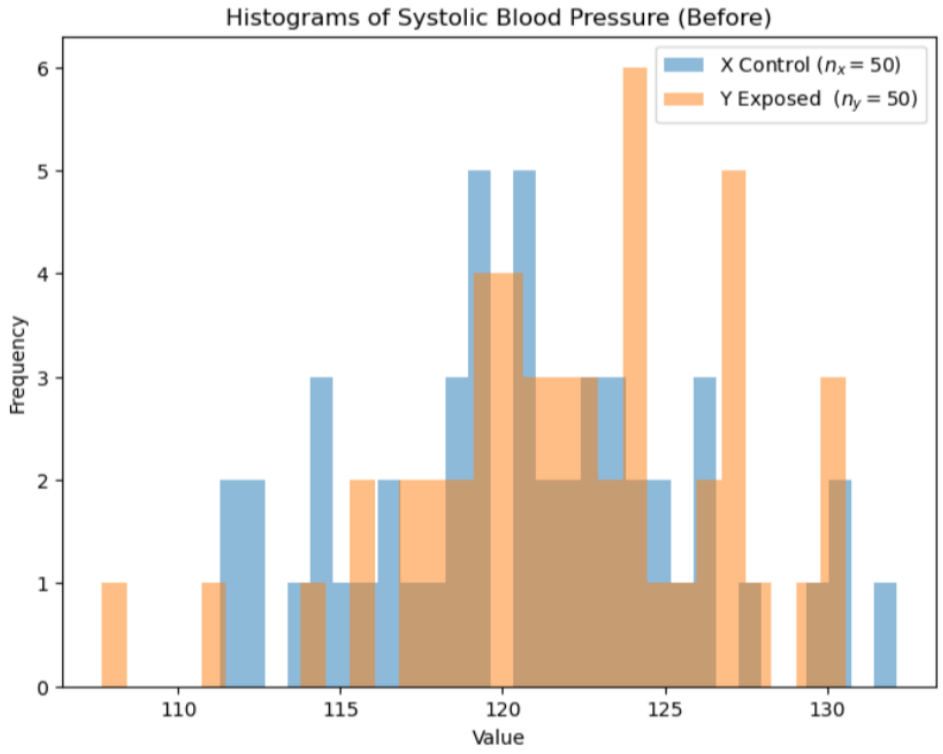
Gather systolic blood pressure measurements from patients in both the control group and the exposed group.

Ensure that the sample sizes ($n_x$ for the control group and $n_y$ for the exposed group) are sufficient to detect a significant difference

# ZOMBIVIA TRIAL: STEP 1 COLLECT DATA

**Measures blood pressure before.**

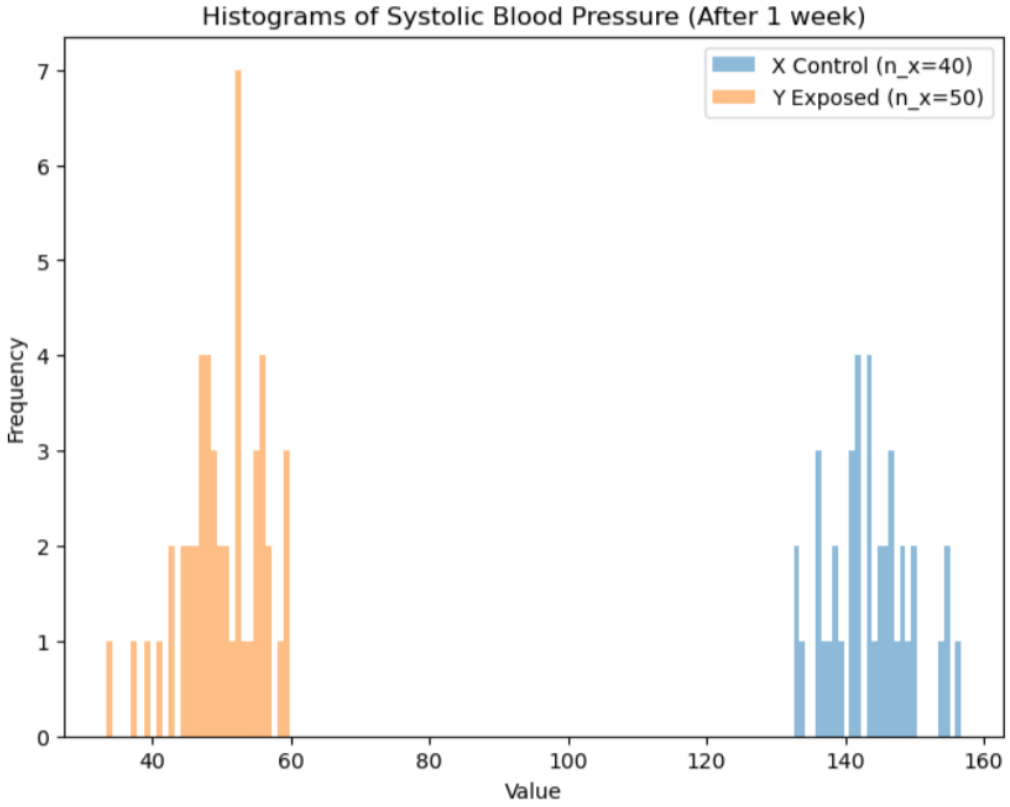
- Just to detect obvious sampling bias.




$$\bar{x} = 120.8, \quad \bar{y} = 122.1$$

# ZOMBIVIA TRIAL: STEP 1 COLLECT DATA

**Systolic blood pressure after 1 week.**

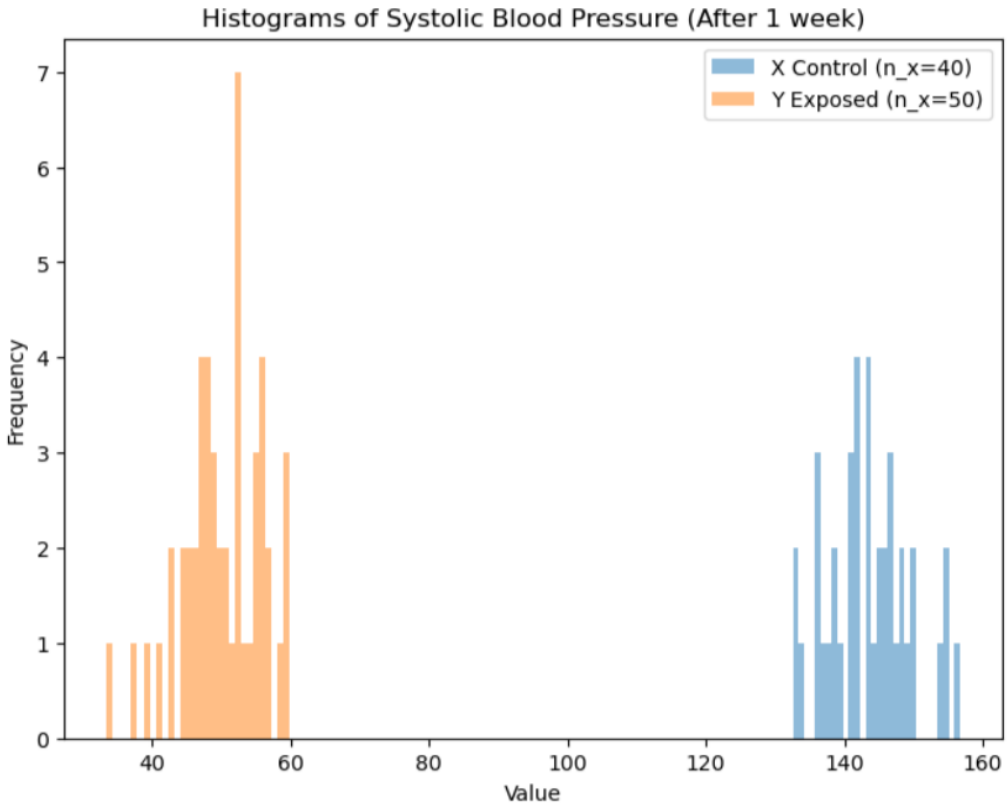



# ZOMBIVIA TRIAL: STEP 3

**Step 3: Test Assumptions**

Assumptions of a two-sample t-test.

- **Independence:** Systolic blood pressure measurements of *X* and *Y* should be independent.
  - Assume independence because of process.
- **Normality:** Distributions of X and Y should be approximately normal. Look for skew. Look for heavy tails.
- **Equal variances**: The sample variances in both groups should be roughly equal.

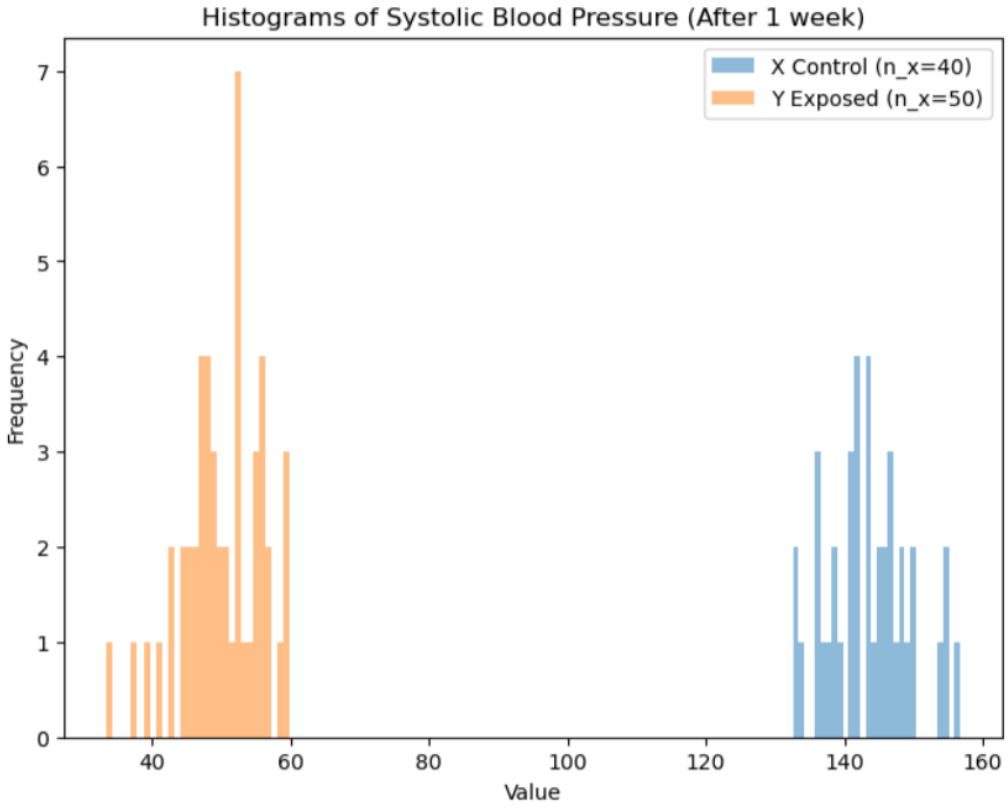


# ZOMBIVIA TRIAL: STEP 3

**Step 3: Test Assumptions (cont.)**

- **Normality:** Distributions of X and Y should be approximately normal. Look for skew. Look for heavy tails This assumption is true because trial is double blinded.
  - I don't see significant skew.
  - No heavy tails.
  - Assume normal.
- **Equal variances**: The sample variances in both groups should be roughly equal.

$$s_x = 6.0, \quad s_y = 5.8$$



# ZOMBIVIA TRIAL: STEP 4

**Step 4: Calculate the test statistic.**

**First calculate the pooled variance**

Let $n_x$ = number of patients in the exposed group.

Let $n_y$ = number of patients in the control group.

$n_x$ and $n_y$ started equal, but for some reason, members of the control group seemed to have disappeared.

Now $n_x < n_y$.

When $n_x \neq n_y$ then we use a weighted average weightd by the degrees of freedom in each sample set.

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

$$s_x^2 = 36.3, \quad s_y^2 = 33.8, \quad s_p^2 = 34.9$$

# ZOMBIVIA TRIAL: STEP 4

$$s_x^2 = 36.3, \quad s_y^2 = 33.8, \quad s_p^2 = 34.9$$

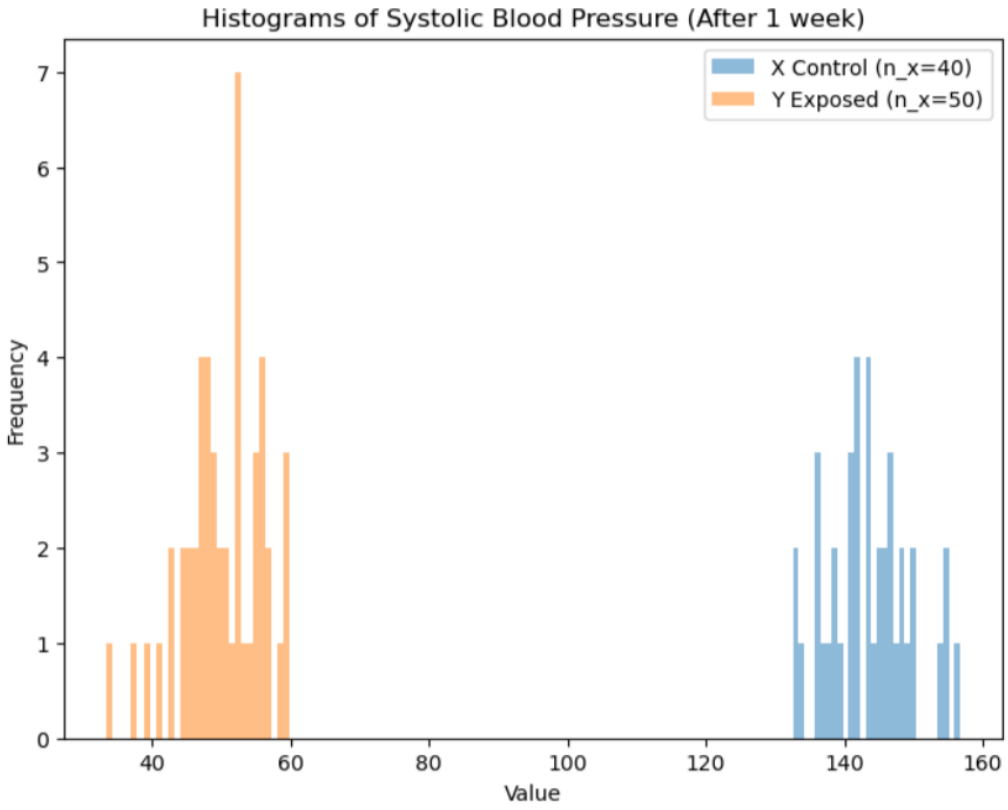
$$s_p = \sqrt{s_p^2}$$

$$s_x = 6.0, \quad s_y = 5.8, \quad s_p = 5.9$$

Standard Error of the difference between X and Y is

$$SE_{\bar{x}-\bar{y}} = \sqrt{s_p^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$$

t-statistic: $$t = \frac{\bar{x} - \bar{y}}{SE_{\bar{x}-\bar{y}}}$$



# ZOMBIVIA TRIAL: STEP 4

Test statistic:

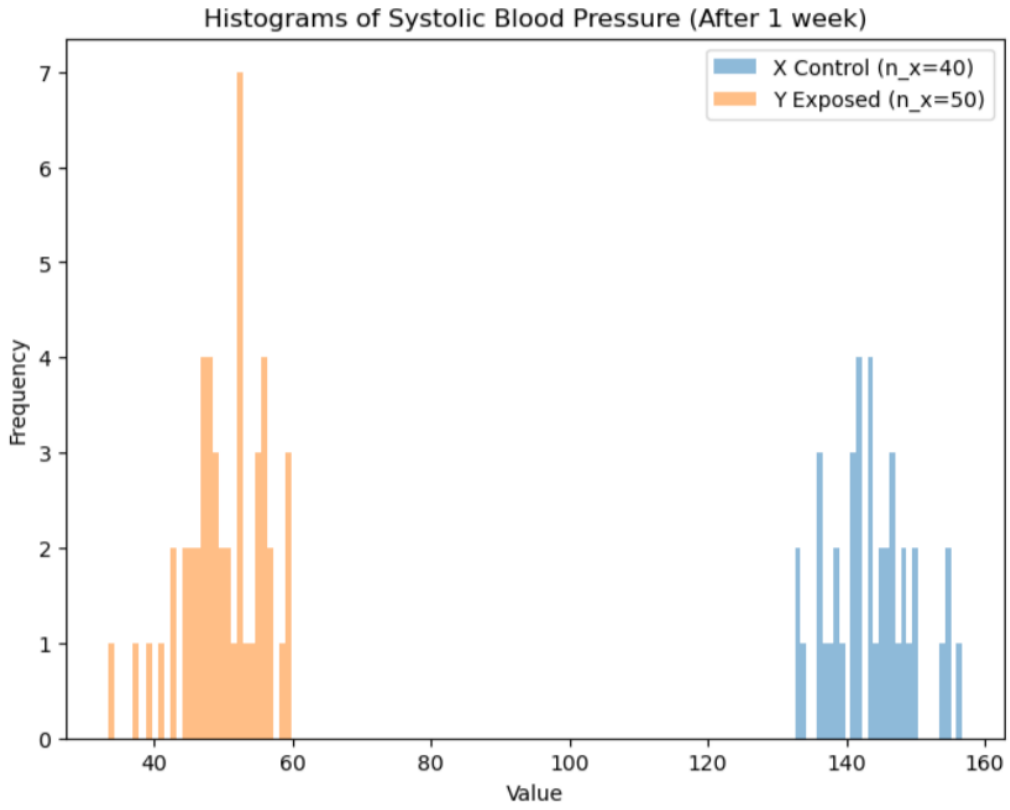
$$t = \frac{\bar{x} - \bar{y}}{SE_{\bar{x}-\bar{y}}}$$

$$\bar{x} = 143.6, \quad \bar{y} = 50.0$$

$$SE_{\bar{x}-\bar{y}} = 7.40$$

.

$$t = 12.65$$



# ZOMBIVIA TRIAL: STEP 4

$$t = 12.65$$

Compute degrees of freedom:

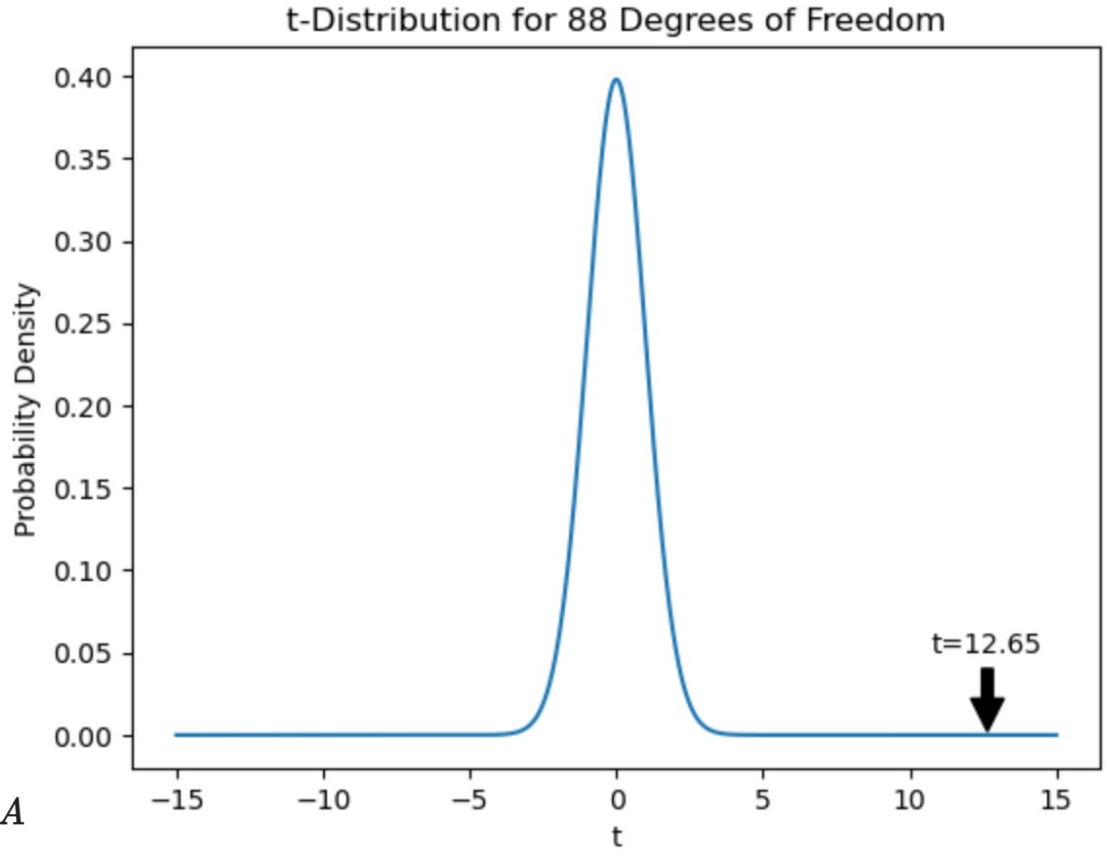
$$df = n_X + n_Y - 2$$

$$df = 88$$

Compute p-value.

$$p = 1.659307456321507e - 21$$

$p << \alpha$, so we reject $H_0$ in favor of $H_A$



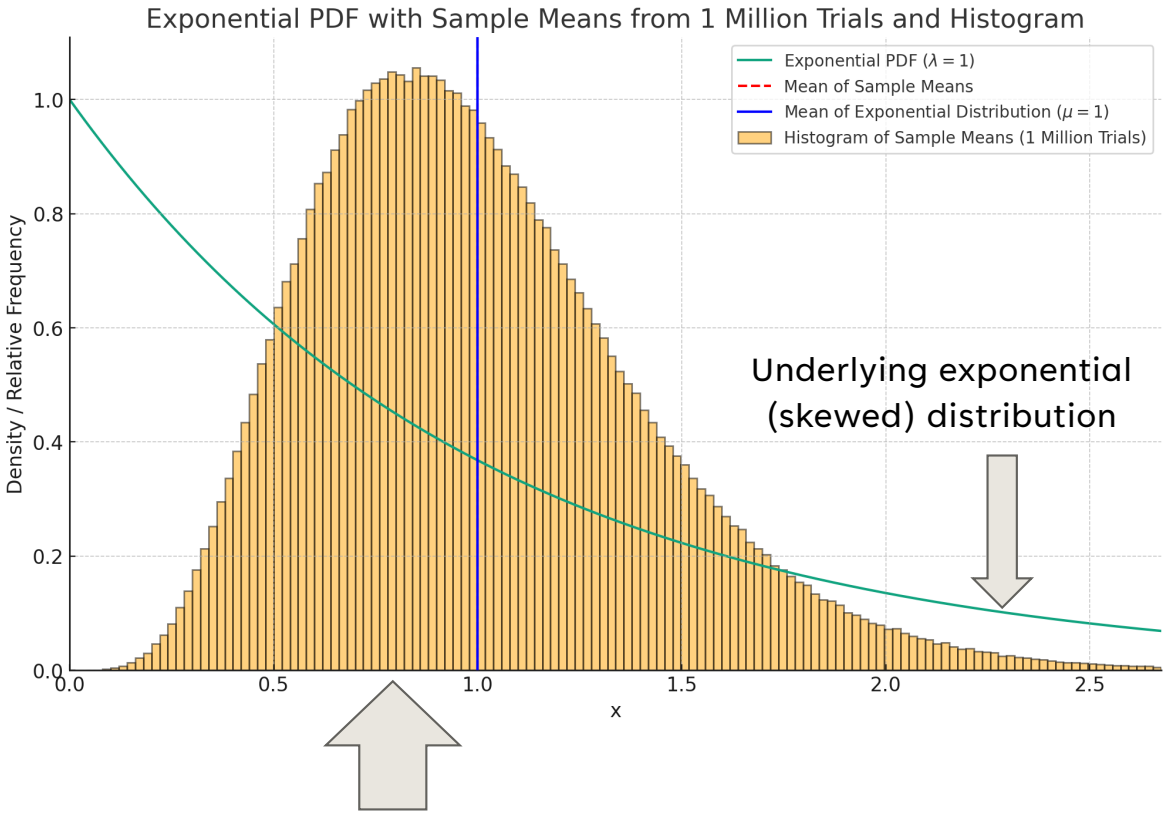

$H_A$ states that $\bar{x}$ and $\bar{y}$ are significantly different.

# PERMUTATION TEST

A *permutation test* is an alternative to hypothesis testing using either a t-distribution or Gaussian distribution as an approximation of the sampling distribution.

Used when

1) have small sample sizes
   - **CLT doesn't apply.**
2) the sampling distribution doesn't look normal.
3) uncertain of homoscedasticity (uncertain of equal variances)
4) have complex or uncommon statistical models
5) desire simplicity and robustness




Non-normal sampling distribution of the mean. Exhibits right skew.

# PERMUTATION TEST

A *permutation test* is as close to a one-size fits all we will get for hypothesis testing.

- Not historically used because of computation cost.
- Not a problem with computers except with vary large sample sizes
  - but in the large sample case CLT probably applies.

# HOW PERMUTATION TESTS WORK

Used with null hypothesis testing for A/B.

1. Combine samples from different groups into a single data set.
2. Shuffle the combined data set and randomly draw without replacement same size as group A.
3. Draw without replacement same size as group B
4. Measure test statistic.
5. Repeat until R times to build a permutation distribution.

The permutation distribution is an estimate of the sampling distribution.

# HOW PERMUTATION TESTS WORK

We can combine into a single dataset since we are proceeding from the assumption that the null hypothesis is true.

If it is true then the A and B at least have the same population mean.

1. Combine samples from different groups into a single data set.
2. Shuffle the combined data set and randomly draw without replacement same size as group A.
3. Draw without replacement same size as group B
4. Measure test statistic.
5. Repeat until R times to build a permutation distribution.

The permutation distribution is an estimate of the sampling distribution.

# EXAMPLE: INCANDESCENT VS. LED LIGHTS

We have been tasked with confirming that LED lights last longer than incandescent lights.

We gathered data from 100 light bulbs of each kind under identical simulated use patterns.

We continued the trial until 30% of the light bulbs fail for each kind.

We therefore have 30 failures of each kind in our sample sets.

# EXAMPLE: INCANDESCENT VS. LED LIGHTS

let $X$ = lifespan of an incadescent light bulb (in years)

let $Y$ = lifespan of an led light bulb (in years)

let $H_0 : \mu_x = \mu_y$

let $H_A : \mu_x \neq \mu_y$

# EXAMPLE: INCANDESCENT VS. LED LIGHTS

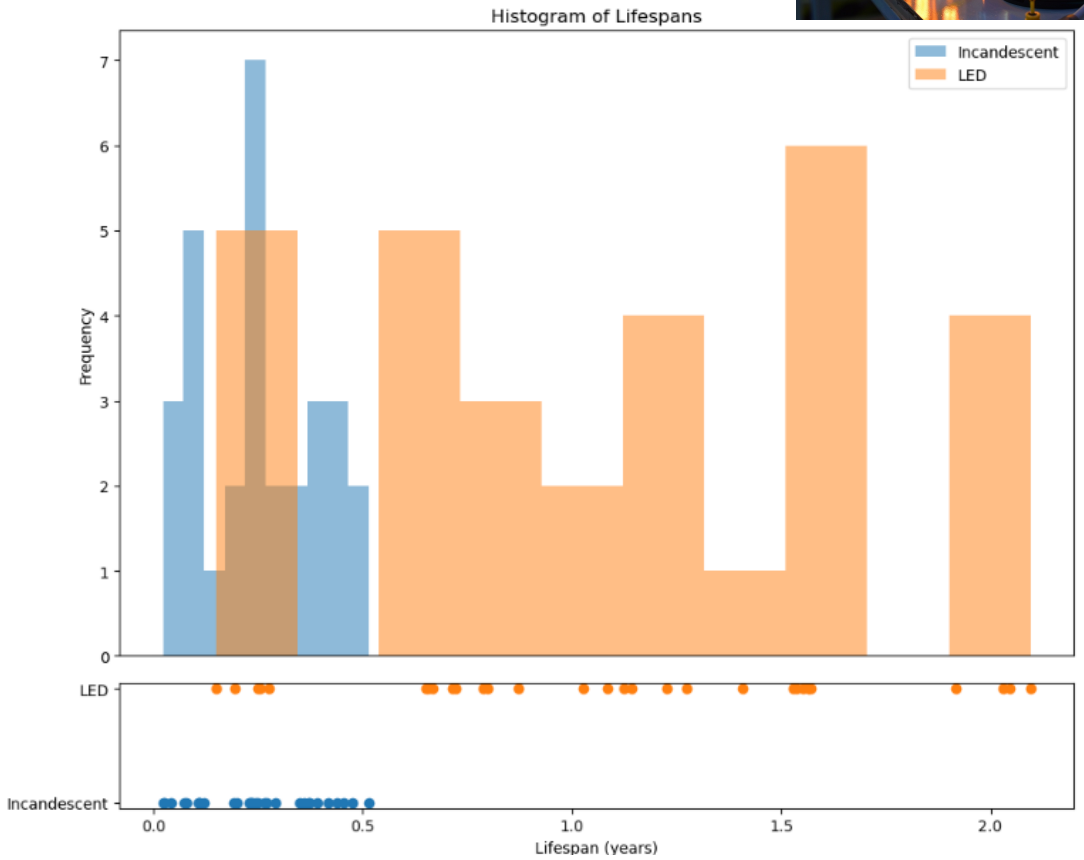
Neither distribution looks Gaussian.

- Seems to few samples for CLT to apply.

T-distribution is based on a Gaussian assumption

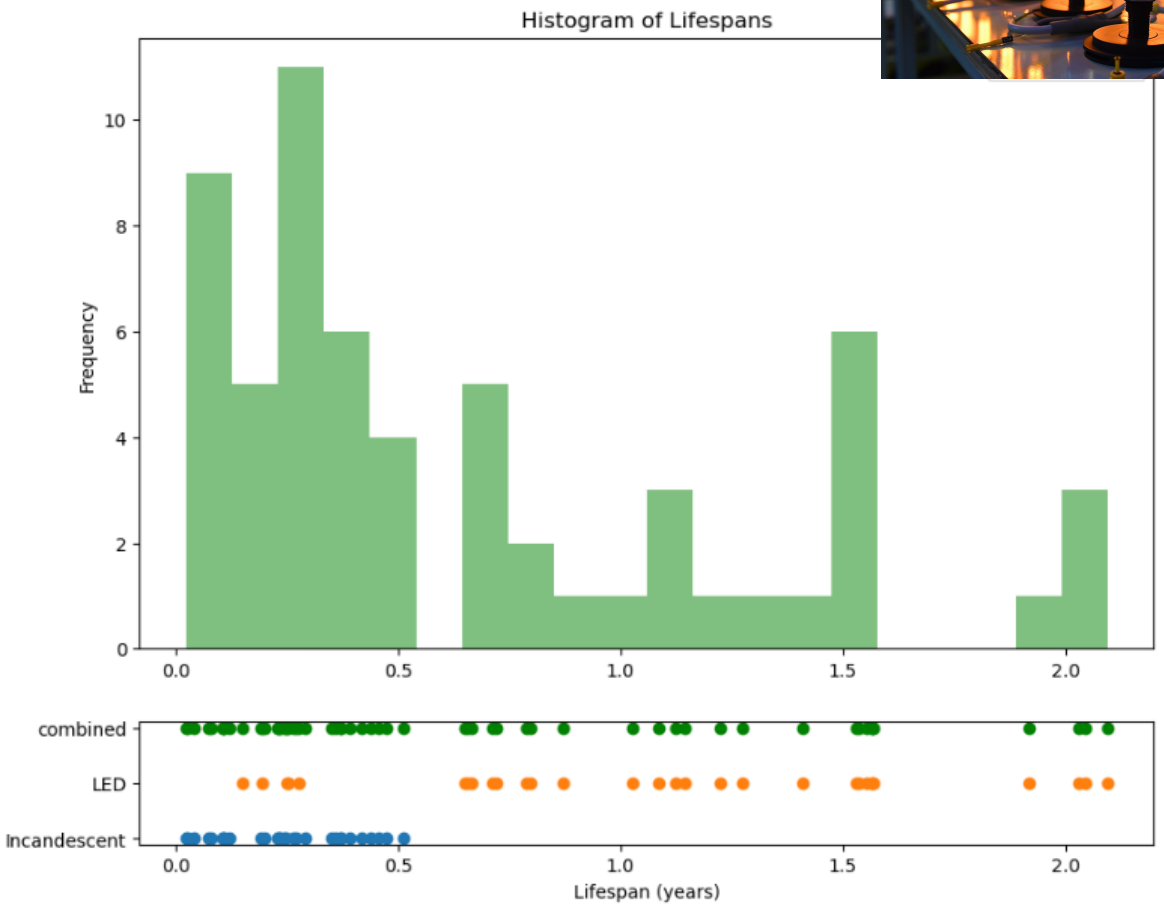
- Used when sample mean and sample variance are computed from the same samples.
- So no t-test.

When this happens, permutation tests make sense.



# EXAMPLE: INCANDESCENT VS. LED LIGHTS

Step 1. Combine samples from different groups into a single data set.



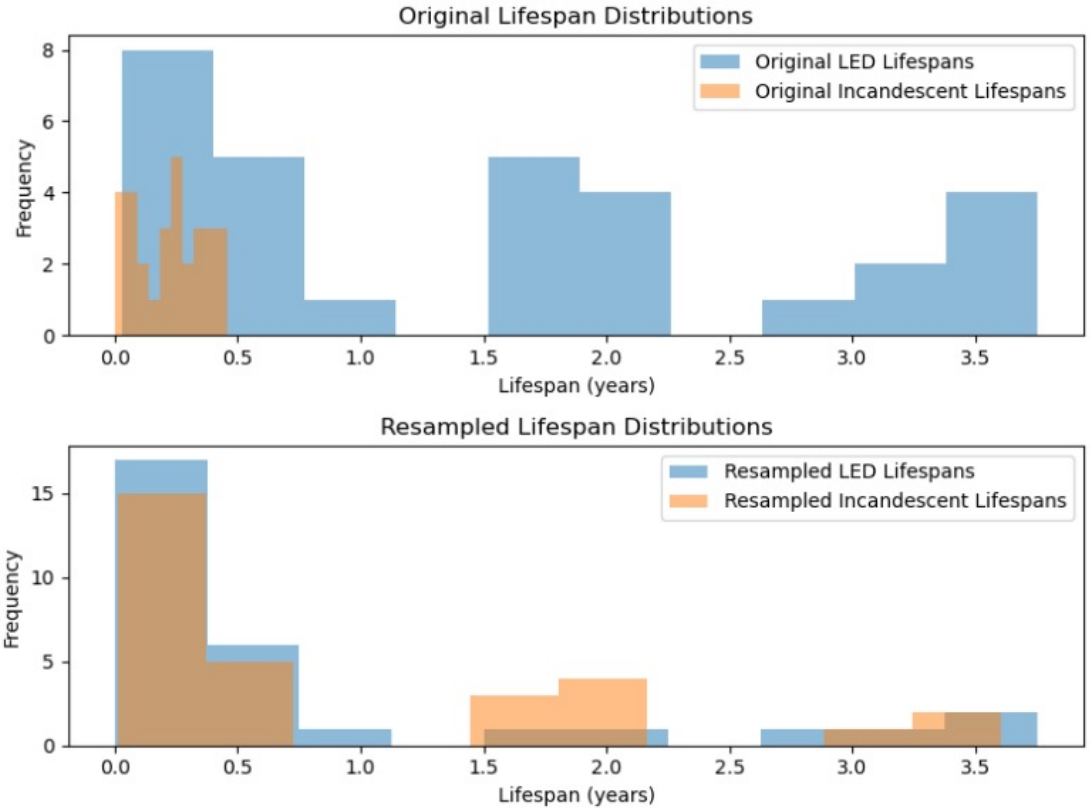
# EXAMPLE: INCANDESCENT VS. LED LIGHTS

Step 2. Shuffle the combined data set and randomly draw without replacement same size as group A.

- Here group A is the “Resampled LED lifespan” dataset.

Step 3: Draw without replacement same size as group B

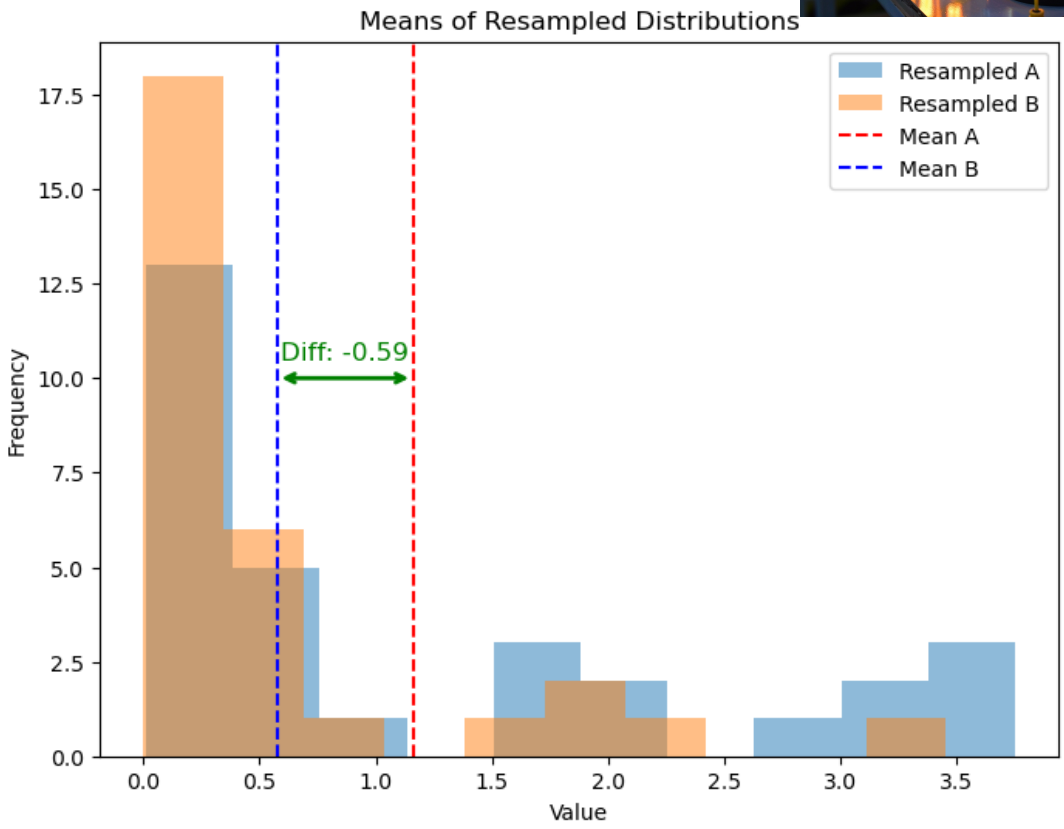
- Here group B is the “Resampled Incandescent Lifespans” dataset.



# EXAMPLE: INCANDESCENT VS. LED LIGHTS

Step 4. Measure test statistic.

- In this case we are measuring the difference in the means.



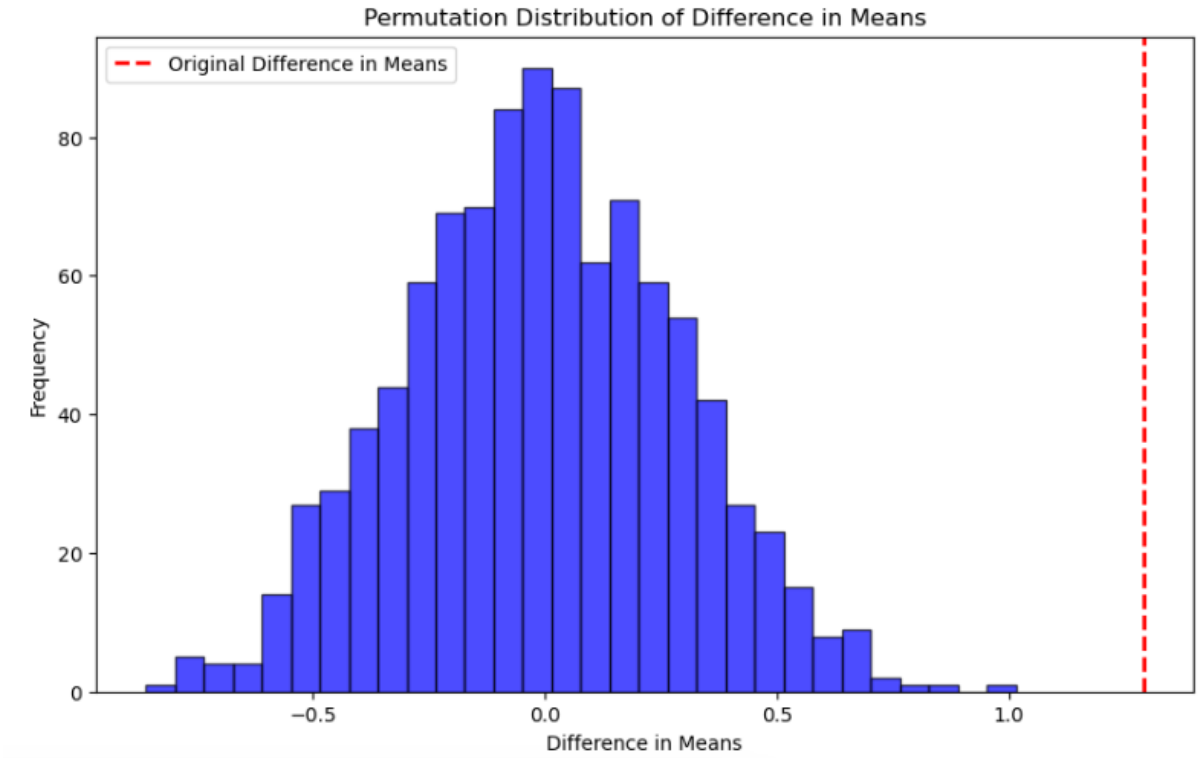
# EXAMPLE: INCANDESCENT VS. LED LIGHTS

Step 5.  Repeat until R times to build a permutation distribution.

R=100

A permutation distribution is analogous to a sampling distribution.

It estimates how much our sample means would vary if the null hypothesis is true.



# EXAMPLE: INCANDESCENT VS. LED LIGHTS

We can use the permutation distribution directly to estimate the p-value.

```python
def compute_p_value(permutation_diffs, original_diff):
    # Two-tailed test p-value
    extreme_values = np.abs(permutation_diffs) >= np.abs(original_diff)
    p_value = np.mean(extreme_values)

    return p_value


original_diff = np.mean(led_lifespans_sorted) - np.mean(incandescent_lifespans_sorted)

# Assuming permutation_diffs and original_diff are already defined
p_value = compute_p_value(permutation_diffs, original_diff)

print(f"P-value: {p_value}")
```

```
P-value: 0.0
```

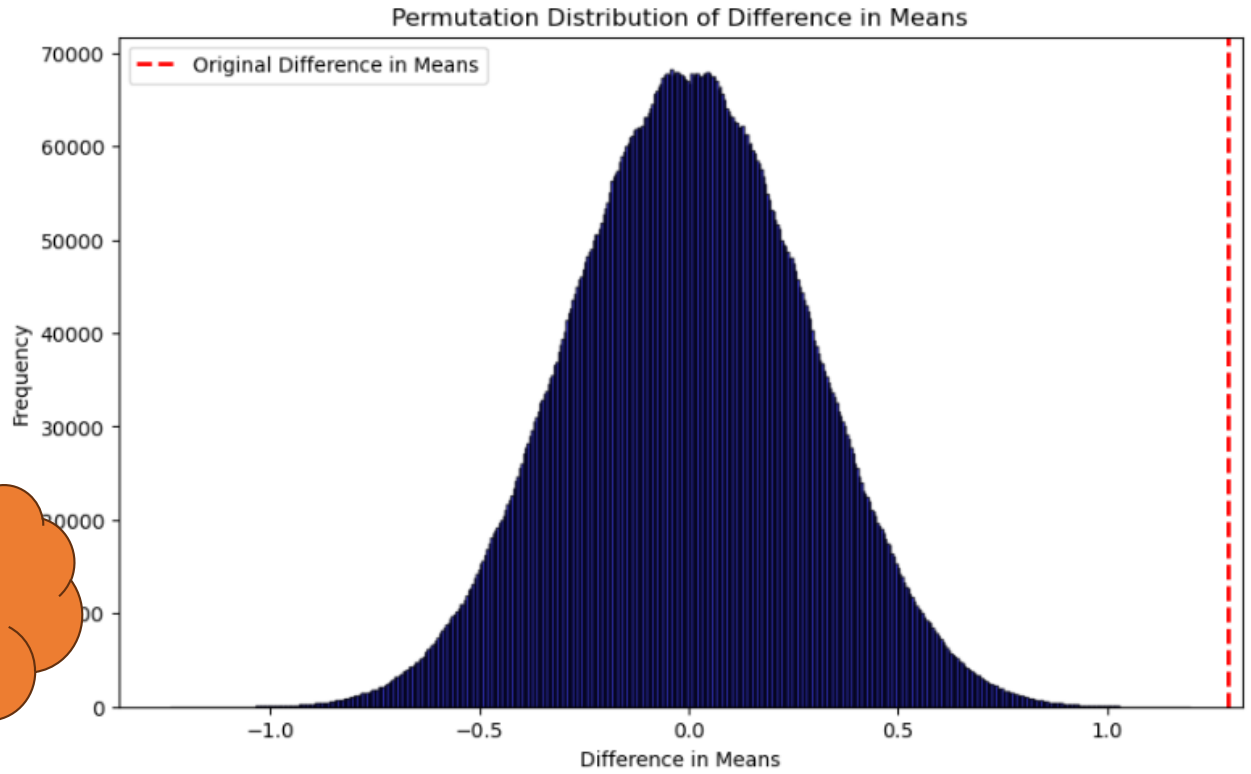
# EXAMPLE 2: INCANDESCENT VS. LED LIGHTS

Given that the original difference is greater than the differences in all 1000 resamples. We get a p–value of 0.

I reran with 10,000,000 resamples.

```
P-value: 0.0
```

P-value still ZERO!

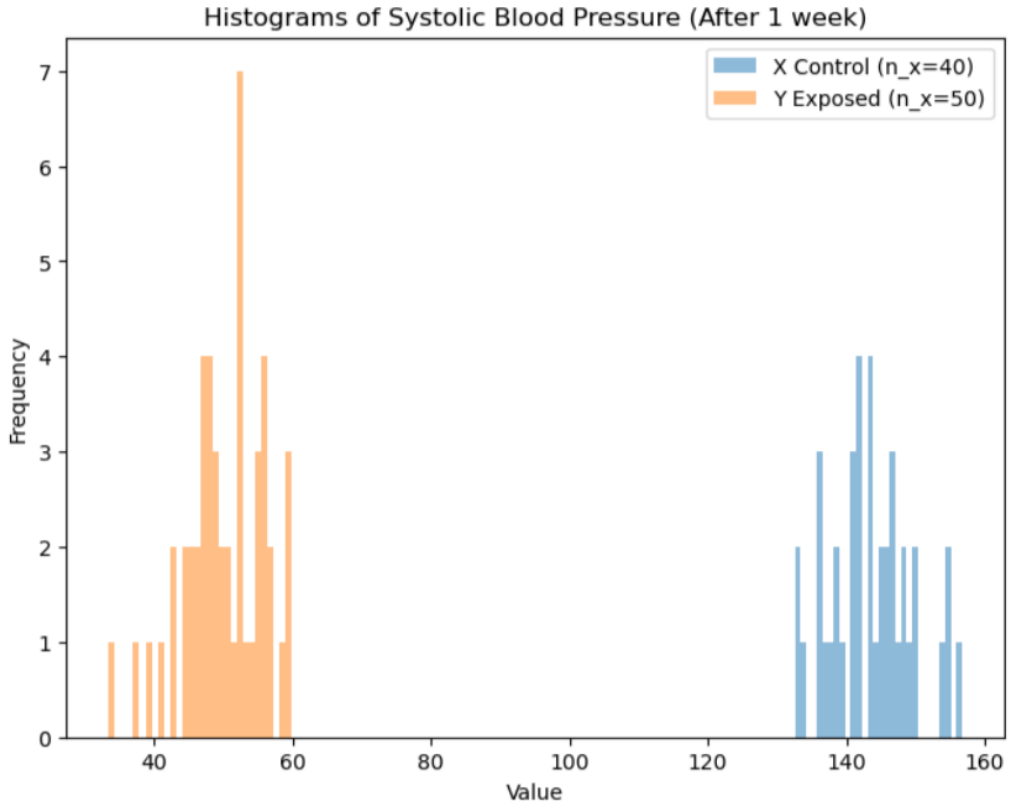


# EXAMPLE PERMUTATION TEST WITH ZOMBIVIA

Let’s say we looked at the distributions of blood pressure and think that the two groups don’t appear to have sufficiently Gaussian shape, so we decide to use a permutation test instead of a t-test.

# CREATE PERMUTATION DISTRIBUTION

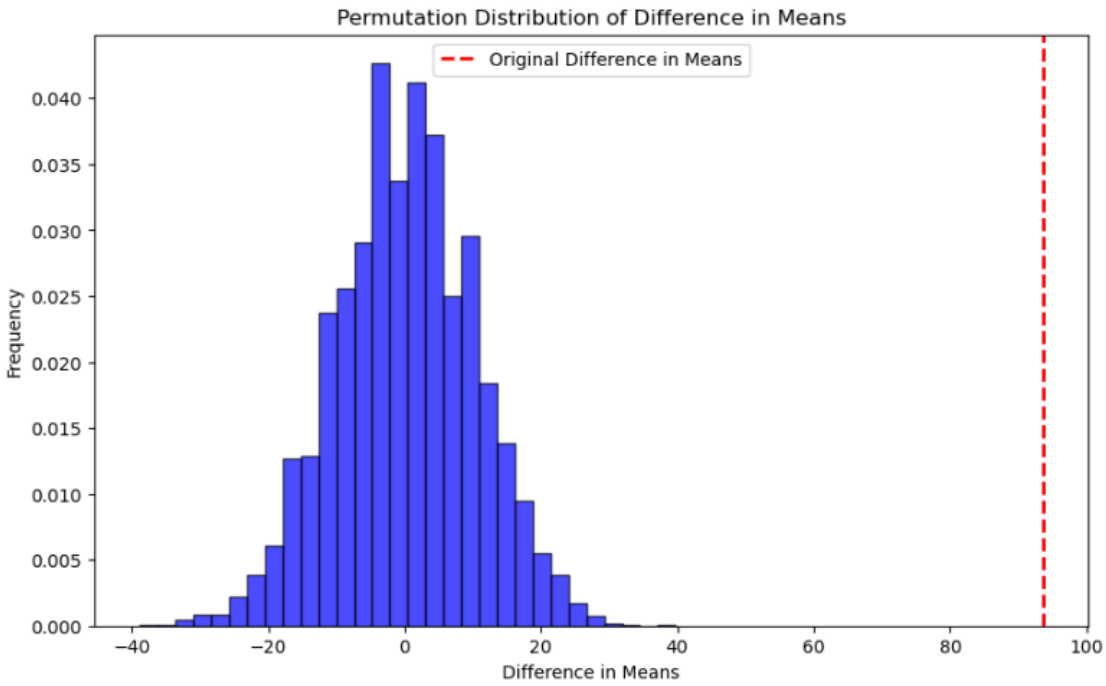
We will do resampling of the distribution of the blood pressures to create a permutation distribution from groups X and Y.



# COMPUTE P-VALUE FROM PERMUTATION DISTRIBUTION

We will do resampling of the distribution of the blood pressures to create a permutation distribution from groups X and Y.

```
P-value: 0.0
```



# COMPUTE P-VALUE FROM PERMUTATION DISTRIBUTION

Interesting point though is that the permutation distribution is much broader than the standard error.

$$s_x = 6.0, \quad s_y = 5.8, \quad s_p = 5.9$$

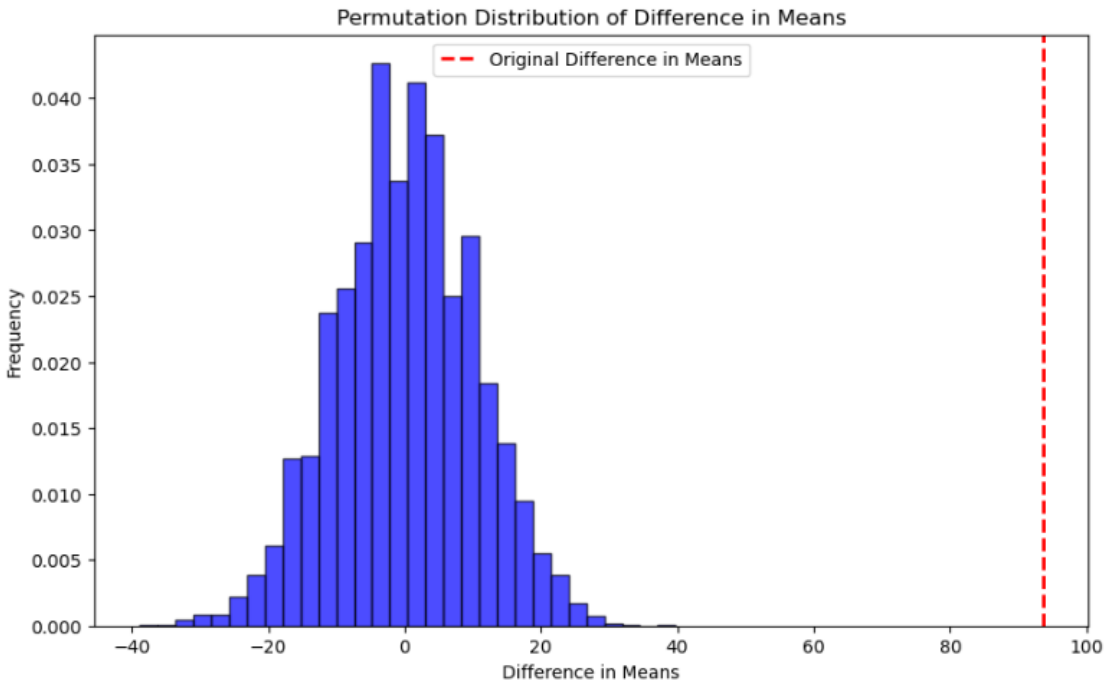
$$\bar{x} = 143.62, \quad \bar{y} = 49.98$$

$$SE_{\bar{x}-\bar{y}} = 7.40$$

$$t = 12.65$$

$$s_{permute} = 9.974$$

$$\frac{\bar{x} - \bar{y}}{s_{permute}} = 9.388$$



# ANOVA

When we have multiple (possibly many) groups and we want to know if any standout as having a significantly different mean.

We could perform t-tests for each pair, but the number of t-tests grows large.

For 7 groups, 7 choose 2 is 21. If we have alpha=0.05, we would expect 0.05 * 21 approximate 1 comparison to be a false discovery.

To deal with this, don't use pairwise t-tests to determine if there are standouts.

Instead state it as a single test.

ANOVA = ANalysis Of Variance

## KEY TERMS FOR ANOVA

***Pairwise comparison***

A hypothesis test (e.g., of means) between two groups among multiple groups.

***Omnibus test***

A single hypothesis test of the overall variance among multiple group means.

***Decomposition of variance***

Separation of components contributing to an individual value (e.g., from the overall average, from a treatment mean, and from a residual error).

***F-statistic***

A standardized statistic that measures the extent to which differences among group means exceed what might be expected in a chance model.

***SS***

"Sum of squares," referring to deviations from some average value.

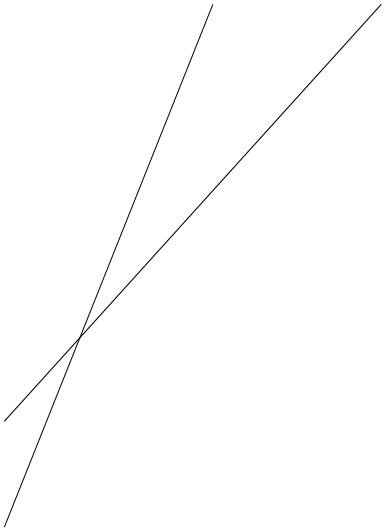
# ANOVA

ANOVA = ANalysis Of Variance

- Given multiple groups (presumed to be independent).
- Null hypothesis ($H_0$): "The group means are NOT different."
- Alternative hypothesis ($H_a$): "At least one group has different mean."

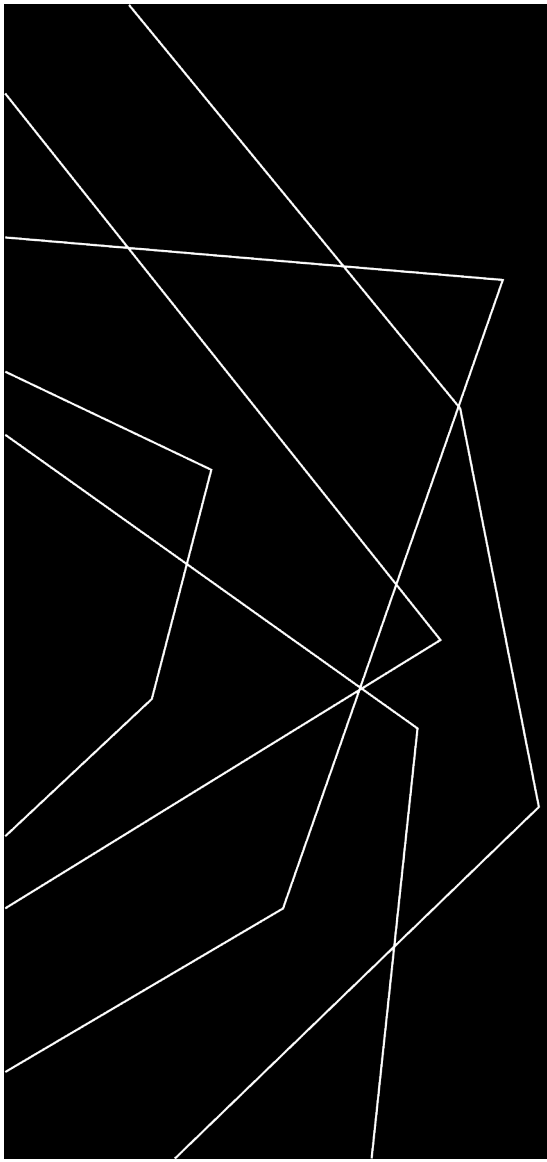
- Assumes  distributions for each group are Gaussian.
- Assumes the groups have equal variances (homoscedasticity).

- If sample size is large enough Gaussian may be satisfied due to CLT.
- If sample variances are similar, then homoscedasticity is satisfied.



# ANOVA

We will go into ANOVA in detail in the next lecture.



# THANK YOU

David Harrison

Harrison@cs.olemiss.edu