Session 2: Statistics Critical to Experimentation I

Shan Huang, HKU

Outline

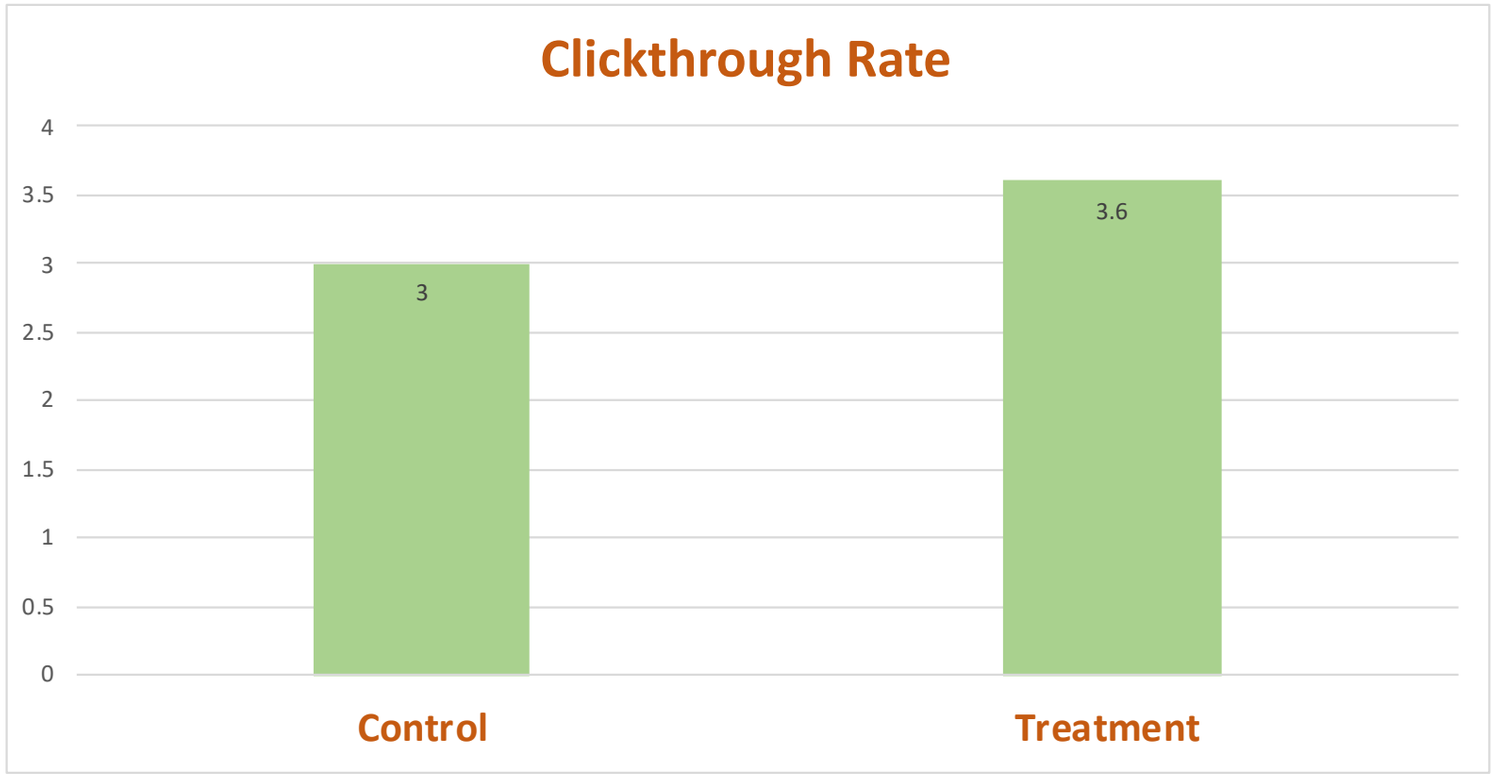
- Hypothesis testing
- t-test, z-test
- p-value
- Type I error
- Multiple Testing
- Confidence Interval (CI)
- Bootstrap CIs

Why You Care

- Statistics are fundamental to designing and analyzing experiments.
- I won't overwhelm you with maths.
- It's not a coding class.
- My objectives are to make you:
    1. Accurately understand the concepts and theories in the statistics critical to A/B testing.
    2. Apply them to solve problems in practice.
    3. Creatively deal with the new problems in practice.

Population or *Sample*?

Can you conclude the treatment increases ads clicks?



Treatment: a new ranking algorithm to target users for ads on WeChat Moments

Population vs. Sample

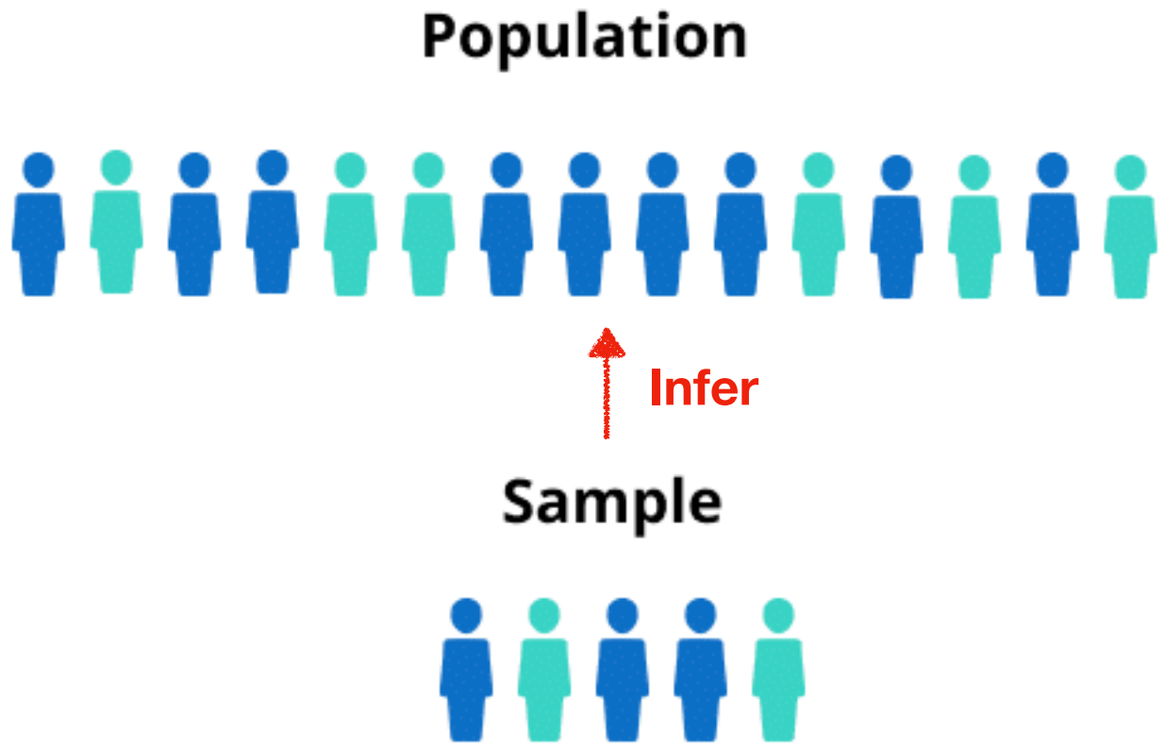
- Population
  - the entire group that you want to draw conclusions about.
    - e.g., all the WeChat Moments users (abstract and infinitely large)
- Sample
  - the specific group that you collect data from.
    - e.g., the random 10% users who logged in the WeChat Moments during the experiment (2 weeks)
  - Different samples can be drawn from the same population.
    - e.g., we run an experiment - draw samples of different groups from the population.

What do we want to learn from experiments?

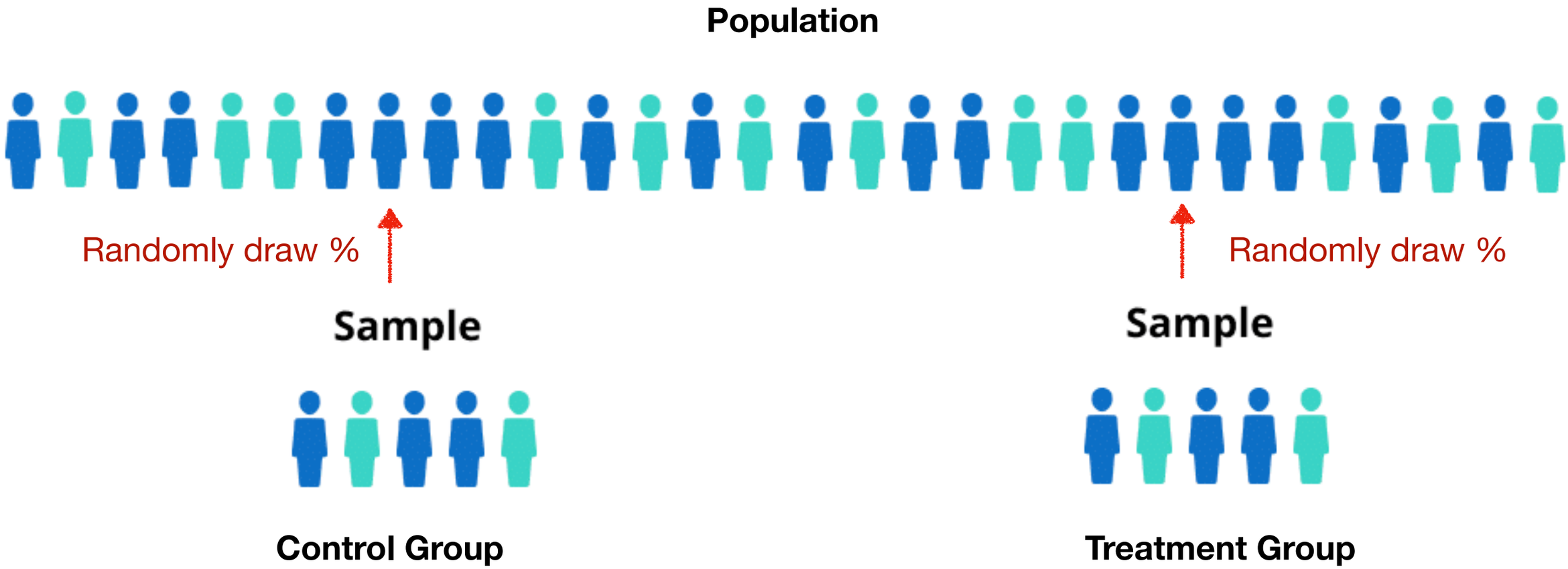
Sample Characteristics or
Population Characteristics

Statistical Inference

- Our focus is Population Characteristics
- However, we do not observe the entire population, just a sample.
- We infer the population characteristics (e.g., mean, median) based on a sample (or samples) drawn from that population.



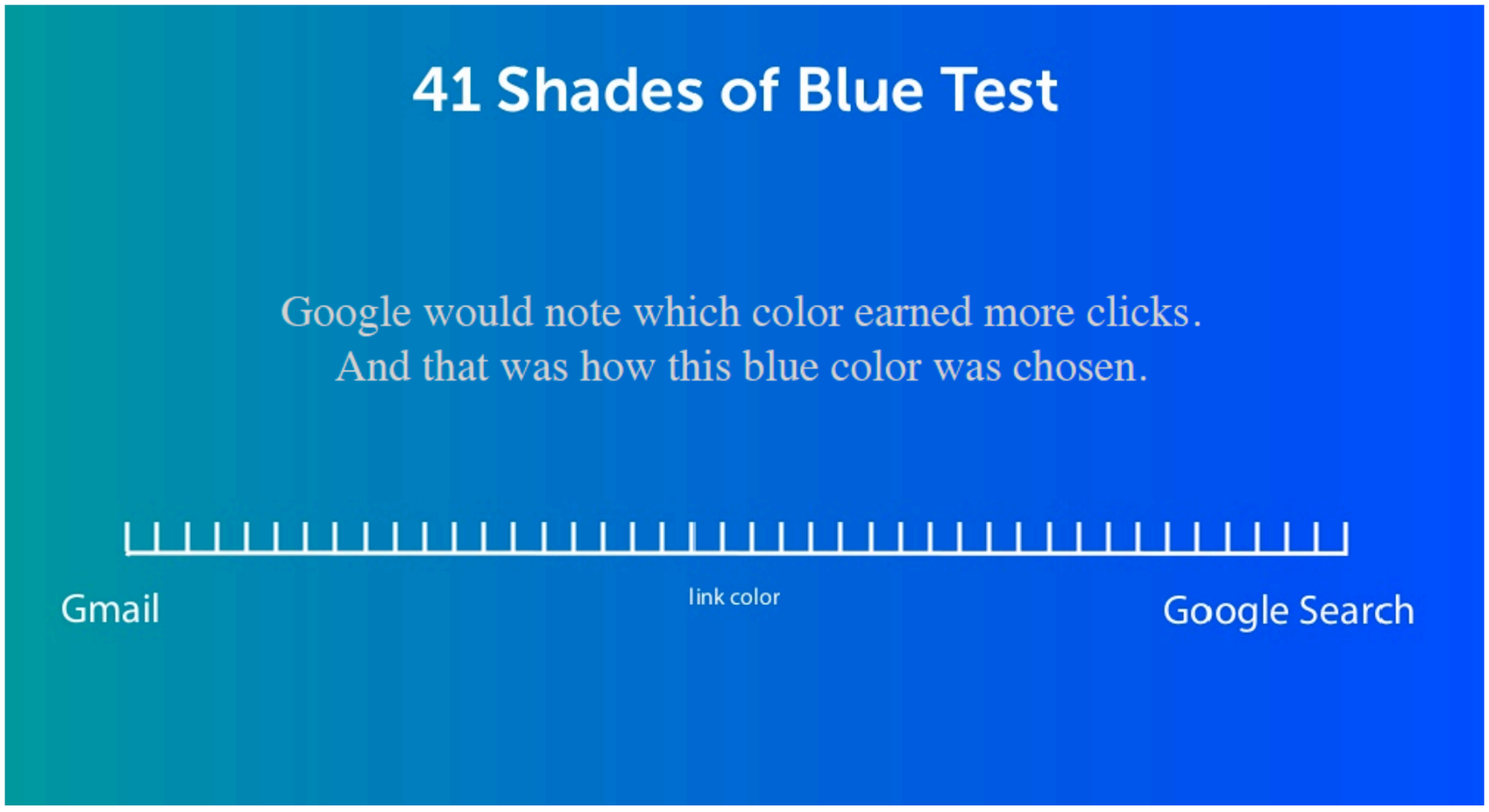
Population vs. Sample in Experiments



- Both the control and treatment groups are randomly drawn from the same population.
- We want to conclude whether there is a change in OEC when the population is treated differently (e.g., with a new feature or not).
- We use the sample difference between two variants to infer whether there is a change.

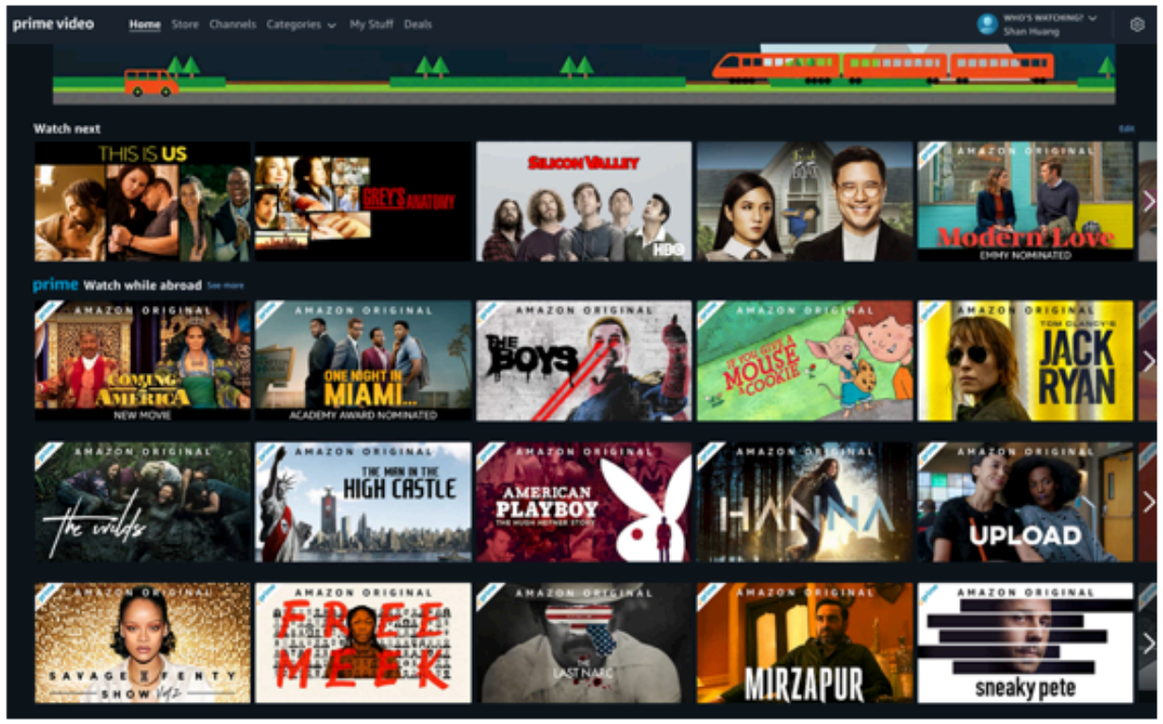
Class Exercise

- What are the population and sample for the experiment - 41 shades of blue?



Class Exercise

- What are the population and sample for the experiment on Amazon Prime?
  - As an engineer, you want to find the best parameters for your algorithm to optimize the user clicks on the movies recommended on Amazon Prime.
  - Assume that there are 8 combinations of the parameters *(k, b)*.
  - Could you design an experiment to find the *(k, b)* which optimizes user clicks?



Treatment Effects by potential outcomes framework

Setup:

- A random sample of units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$, randomly assigned
- Potential outcomes: $Y_i(0),\ Y_i(1)$
- Observed outcome: $Y_i = Y_i(D_i)$
- Number of treated/untreated units: $n_1 = \sum_{i=1}^{n} D_i$ and $n_0 = n - n_1$

Potential Outcome: the **outcome** for an individual under a **potential** treatment.

We cannot observe Yi (1) if unit_i was assigned to the control group, vice versa.

Two causal estimands, corresponding to where interest lies:

- Sample average treatment effect (SATE):

Random sampling

$$SATE = \frac{1}{n} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0)\}$$

- Population average treatment effect (PATE):

$$PATE = \mathbb{E}[Y_i(1) - Y_i(0)] \quad \text{(note what } \mathbb{E} \text{ here represents!)}$$

The assignment of variants is uncorrelated with all the factors impacting the potential outcomes.

Random assignment of treatment still implies: $\{Y_i(1),\ Y_i(0)\} \perp\!\!\!\perp D_i$

Estimation of Average Treatment Effect

Estimator = Observed difference in means between groups in sample:

- $\Delta = \frac{1}{n_1}\sum D_i \,.\, Y_i - \frac{1}{n_0}\sum (1 - D_i) \,.\, Y_i = m_1 - m_0$

- Δ is an unbiased estimator for SATE & PATE:
  - $E(\Delta \mid O) =$ SATE, O = the current sample.
  - $E(\Delta) = E[E(\Delta \mid O)] = E(SATE) = PATE$

Repeated random treatment assignments on a sample

Repeated ***random*** sampling (& random assignment)

We infer PATE (e.g. whether = 0) based on a realization of Δ observed from an experiment (sample).

Class Exercise

Di	Yi (0)	Yi (1)	Yi (1) - Yi (0)	Yi	DiYi	(1-Di)Yi
0	0	1	1	0	0	0
0	0	1	1	0	0	0
1	0	0	0	0	0	0
1	0	1	1	1	1	0
1	1	1	0	1	1	0
0	0	1	1	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	1	1	0	1	0	1
1	0	0	0	0	0	0
0	1	1	0	1	0	1
0	1	1	0	1	0	1
1	1	0	-1	0	0	0
1	1	0	-1	0	0	0
1	1	1	0	1	1	0
1	0	0	0	0	0	0
1	0	1	1	1	1	0
0	0	0	0	0	0	0
1	1	0	-1	0	0	0
0	0	1	1	0	0	0
SATE=			0.15		0.4	0.3
Delta=						0.1
N=10						

Users are randomized into control (Di=0) and treatment (Di=1) groups. Yi indicates whether user_i clicked the ads or not during the experiment.

Open the file: potential_outcome, and get the values of all the unfinished fields and calculate STAE and Δ for the experiment.

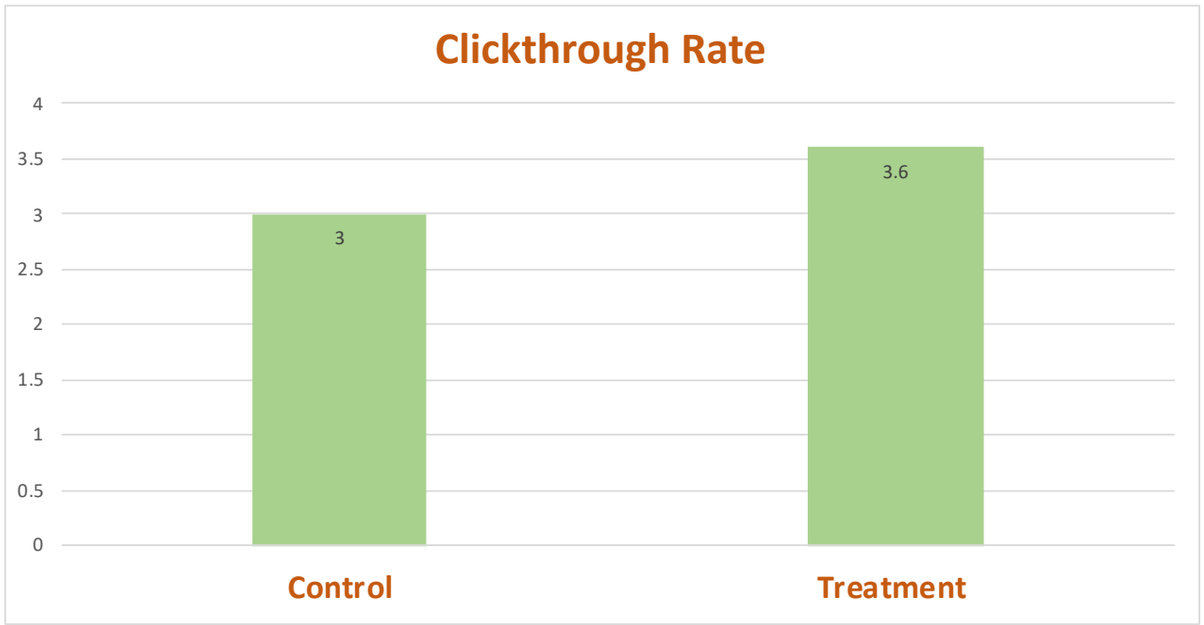
If we randomized users into two groups for 1000 times and get Δ for each randomization ,

Will the average Δ get closer to SATE?

If we randomly sample users from the population 1000 times and get Δ over 1000 randomizations for each sample for the experiment,

Will the average Δ get closer to PATE?

Does the new feature increase CTR?



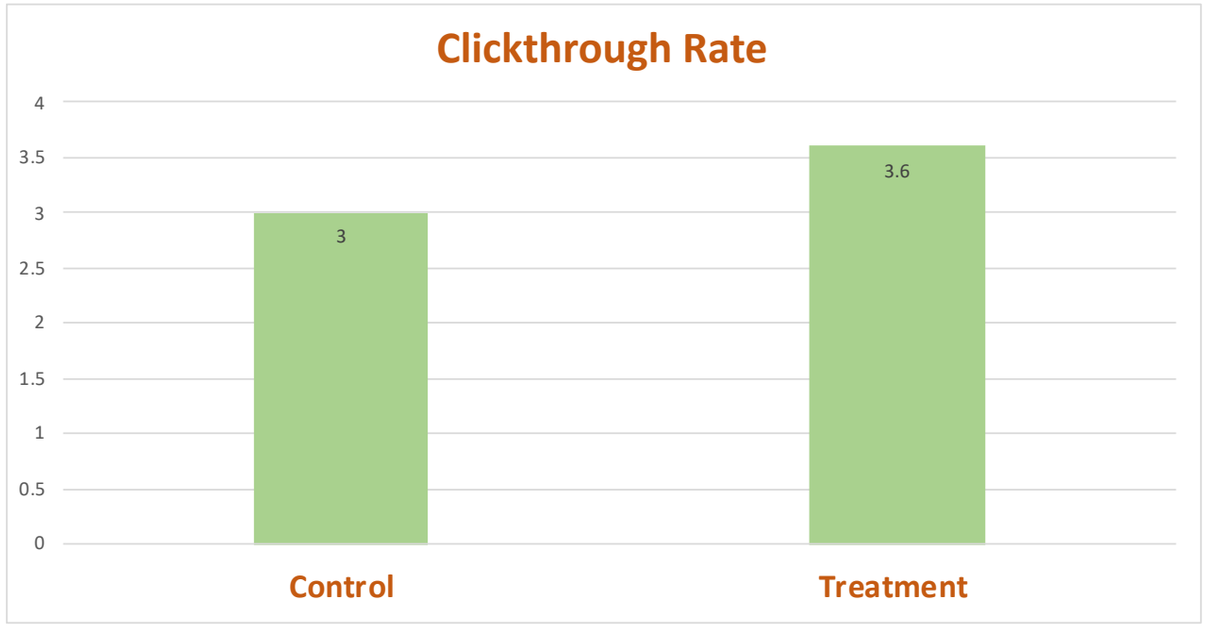
New Feature: a new ranking algorithm to target users for ads

Can we conclude that the treatment increases the population mean of CTR?

$$\Delta = m_1 - m_0 = 3.6 - 3 = 0.6$$

Can we conclude that the treatment increases the population mean of CTR?

Does the new feature increase CTR?



New Feature: a new ranking algorithm to target users for ads

If you find that there is a difference in sample means between Treatment and Control groups:

- If we involve many more users or collect another sample for the experiment, would we expect to continue to see the improvement?
- Is it due to a systematic factor (new algorithm makes ads targeting better) or due to chance?

Chance vs. Systematic Factors

- Systematic Factor
    - e.g., the algorithm targets more relevant users for ads and users tend to click more
    - The improvement is there when you observe an entire population.
- Chance
    - e.g., the improvement in CTR happens randomly to each observed sample
    - If you collect another sample or more users for the experiment, the improvement may disappear (i.e., specific to a sample).
- ***Statistically* Significant Difference -** the difference is this big, which is unlikely to have occurred by chance.

t-test

- The most common statistical test to compare the means of *independent* groups.
- The most common way to determine whether the difference we see between Treatment and Control groups is systematic or just by chance.
- *t* looks at how many standard errors the observed size of difference is away from the hypothesized difference, such as 0 (no difference).
- If *t* value indicates that the difference is too many standard errors (e.g., $t > 2$) away from the hypothesized value,
    - The difference is regarded as extreme compared to the differences under the *null distribution*.
    - We would then reject the Null, e.g., two groups are thus ***statistically*** significantly different from each other.

Hypothesis Testing

Ho (Null Hypothesis): $\delta = \mu_1 - \mu_0 = 0$ (population mean difference = o)

H1 (Alternative Hypothesis): $\delta \neq 0$ ($\mu_1 \neq \mu_2$)

- Y is the metric of interest (OEC, e.g., click)

$\Delta = m_1 - m_0$ (sample mean difference)

- m1, mo are sample means for control and treatment groups.

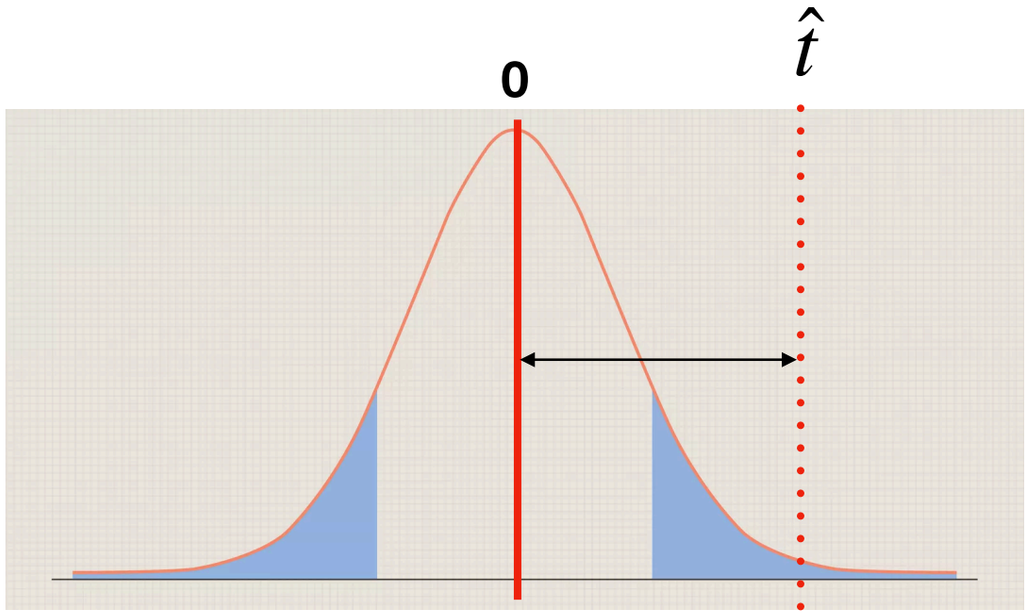
$$\hat{t} = \frac{\hat{\Delta} - \delta}{se(\Delta)} = \frac{\hat{\Delta}}{se(\Delta)}$$

- Δ is an estimator of δ
- t looks at how many standard errors away Δ from δ

User	Click	Treatment
111	1	1
112	0	0
113	1	1
114	0	0
115	0	0
116	1	0
117	0	1
118	0	0
119	0	1
120	1	1
121	1	1
122	1	0
123	1	1
124	0	0
125	1	1
126	0	0
127	0	1
128	0	0

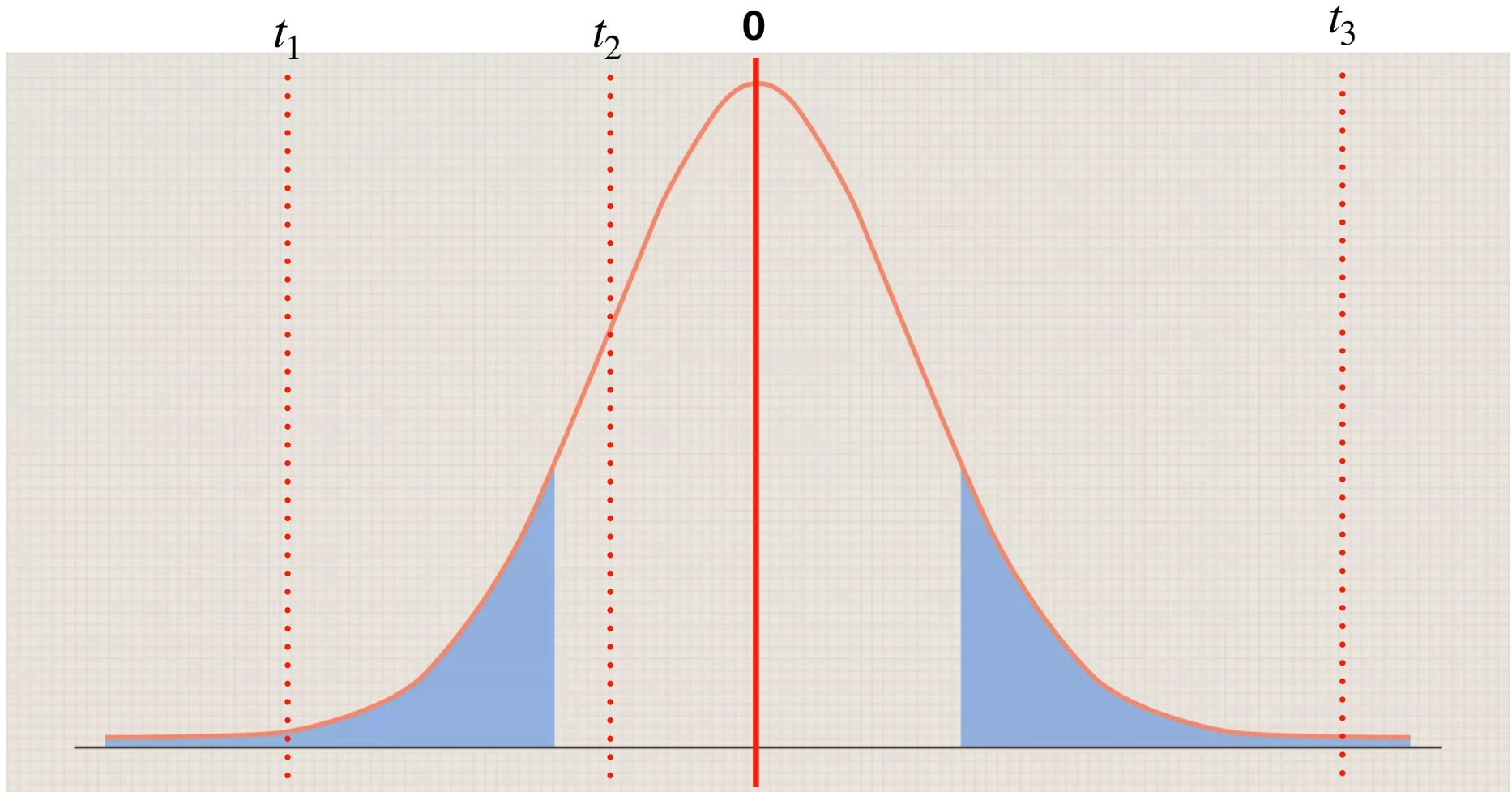
t-test

- Obtain the null distribution for t-stat, $t = \frac{\Delta}{se(\Delta)}$
  - When null is true (e.g., no difference), t would follow a student distribution with mean = 0
  - We draw a sample and calculate t based on our sample.
- Compare t to the null distribution and see how "extreme" it is.
  - If this $\hat{t}$ is extreme to happen (e.g., t>2), we would say this $\hat{t}$ is unlikely to be obtained from the null distribution
- Our observation from the sample is thus inconsistent with the null hypothesis
- Therefore, we reject the null hypothesis.



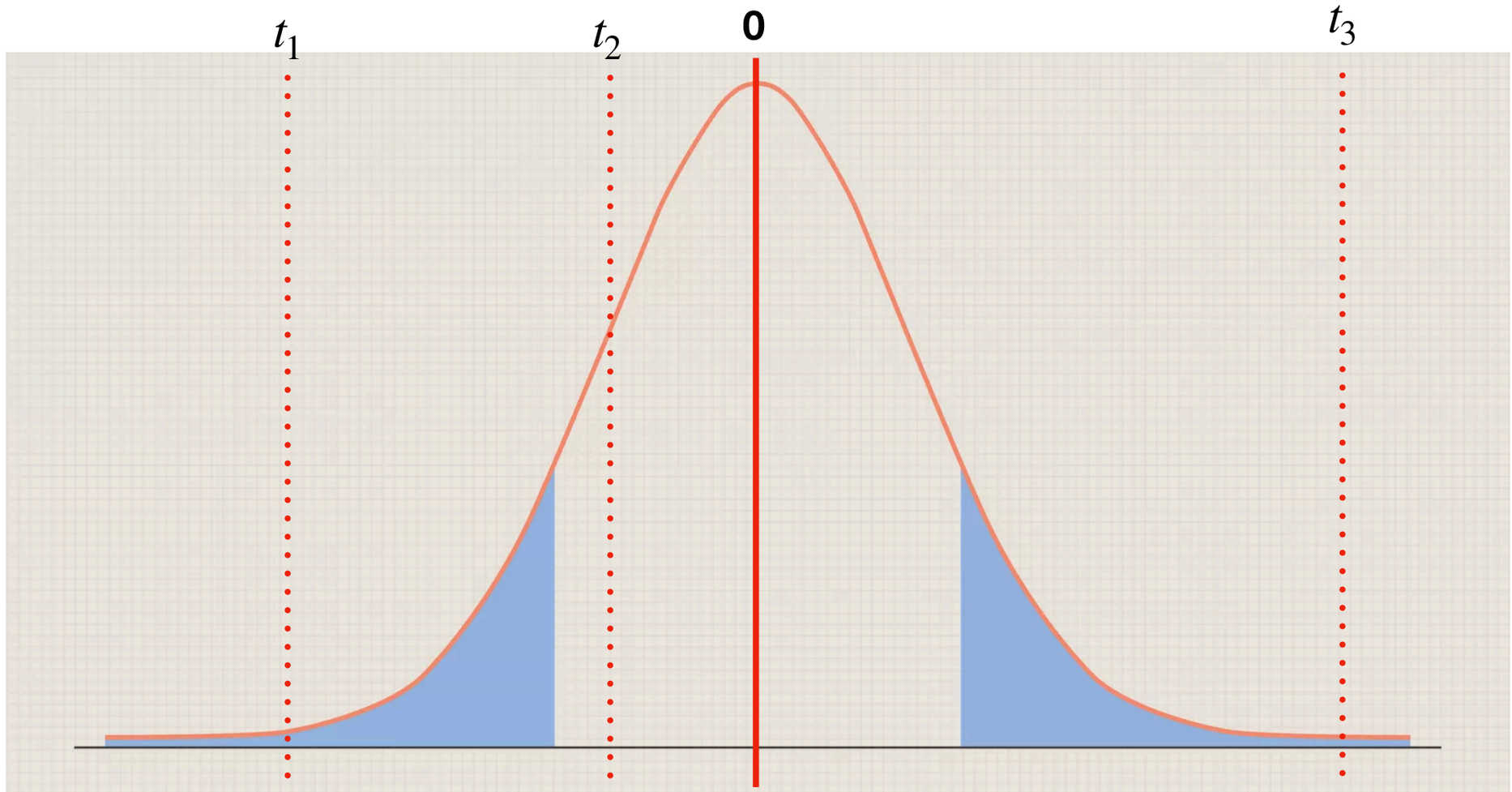
Class Exercise

- Which is more extreme to happen, under the null distribution?



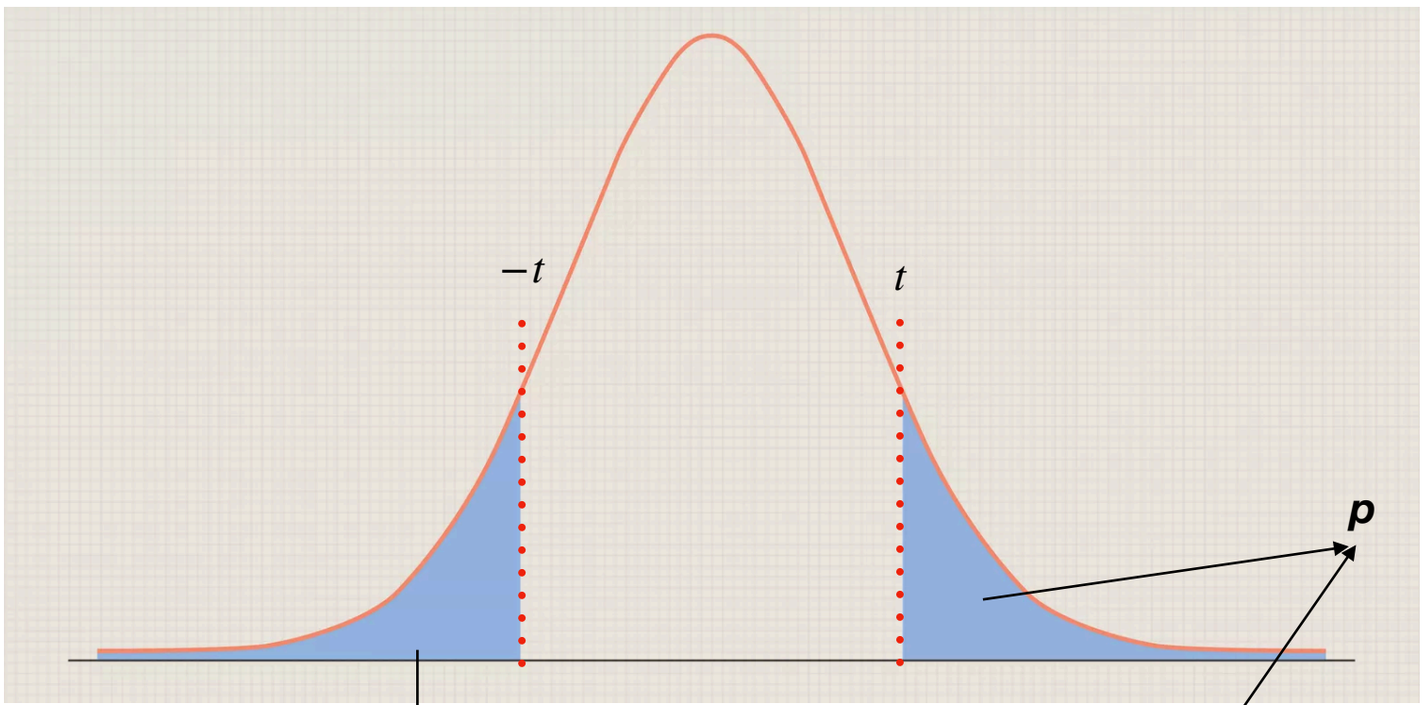
Class Exercise

- Could you describe/define how extreme they happen under the null distribution?



p-value

- p-value evaluates how extreme t (or Δ) occurs under the null distribution.
- p-value is the probability of observing t or more extreme assuming the Null is true.
  - $p = Pr(|T| \geq t \mid H_0)$
  - e.g., the probability of observing such difference or more extreme difference if there really is no difference between Treatment and Control groups (null is true).



Assume t is positive

p-value

- Any difference with a p-value smaller than α, e.g., 0.05 or 0.01, is considered to be extreme enough to reject the null & is called statistically significant (unlikely happen by chance)
- p-value cannot tell you about the effect size.
- It can only tell you whether the difference is statistically significantly different from 0, the hypothesized value.
- e.g., an 0.0001% increase can be statistically significantly larger than 0, but may not be practically significant.

Three Cases

1. The two samples are independent, and the variances of the populations are equal.
2. The two samples are independent, and the variances of the populations are not equal.
3. The two samples are not independent, e.g., the same users experience two conditions

Three Cases

1. The two samples are independent, and the variances of the populations are equal.
2. The two samples are independent, and the variances of the populations are not equal.
3. The two samples are not independent, e.g., the same users experience two conditions

It depends on whether the treatment changes the variances.

A Step-by-Step Example

	Treatment	Control
1	1	2
2	2	3
3	1	2
4	2	4
5	1	3
6	4	5
7	3	4
8	2	3
9	4	5
10	3	5
11	2	4
12	3	3
n1=n0=12		
Mean	2.33333333	3.58333333
VAR(Y)	1.15151515	1.17424242
VAR(m)	0.0959596	0.09785354
VAR(m1-m0)	0.19381313	
se(m1-m0)	0.44024213	
t	-2.8393466	

$$\text{Var}(Y) = \frac{1}{n-1}\sum (Y_i - \bar{Y})^2$$

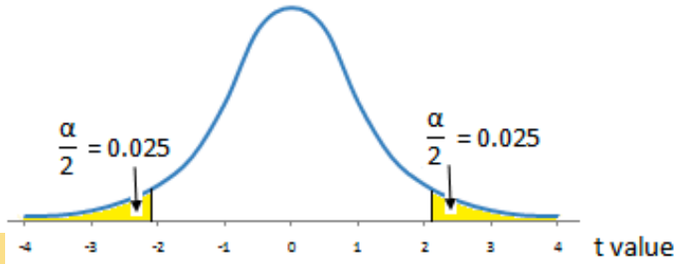
$$\text{Var}(\bar{Y}) = \frac{1}{n}\text{Var(Y)}$$

$$\text{Var}(\Delta) = \text{Var}(\bar{Y}_1 - \bar{Y}_0) = \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_0)$$

$$\text{se}(\Delta) = \sqrt{Var(\Delta)}$$

Student's t Distribution Table

For example, the t value for 18 degrees of freedom is 2.101 for 95% confidence interval (**2-Tail** $\alpha = 0.05$).



	90%	95%	97.5%	99%	99.5%	99.95%	1-Tail Confidence Level
	80%	90%	95%	98%	99%	99.9%	2-Tail Confidence Level
	0.100	0.050	0.025	0.010	0.005	0.0005	1-Tail Alpha
df	0.20	0.10	0.05	0.02	0.01	0.001	2-Tail Alpha
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192	
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991	
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240	
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103	
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688	
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588	
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079	
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413	
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809	
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869	
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370	
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178	
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208	
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405	
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728	
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150	
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651	
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216	
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834	
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495	
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193	
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921	
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676	
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454	
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251	
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066	
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896	
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739	
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594	
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460	

df= n1+ n2 -2

Reject the Null, when t-stat > 2.0739 or t-stat < -2.0739

t-statistic

- Sample sizes of different groups can be different.
- You can allocate different amounts of traffic to control and treatments.
    - e.g., you may assign less traffic to an “unsafe” feature.
    - N_1 unnecessarily equals N_0.
- Treatment may also change the variance.
    - $Var(Y_1)$ can different from $Var(Y_0)$
    - Use *welsh* tests with unequal variances

A Simple Example with python

```
1. import numpy as np, statsmodels.stats.api as sms
2. test0_s = [1,2,1,2,1,4,3,2,4,3,2,3]
3. ctrl0_s = [2,3,2,4,3,5,4,3,5,5,4,3]
4. cm0 = sms.CompareMeans(sms.DescrStatsW(test0_s),
   sms.DescrStatsW(ctrl0_s))
5. print(cm0.ttest_ind(alternative='two-sided',
   usevar='pooled'))
```

Equal Variance

```
(-2.8393466239285283, 0.009541766673068929, 22.0)
```

t-statistic

p-value

df

A Simple Example with python

```
1. import numpy as np, statsmodels.stats.api as sms
2. test0_s = [1,2,1,2,1,4,3,2,4,3,2,3]
3. ctrl0_s = [2,3,2,4,3,5,4,3,5,5,4,3]
4. cm0 = sms.CompareMeans(sms.DescrStatsW(test0_s),
   sms.DescrStatsW(ctrl0_s))
5. print(cm0.ttest_ind(alternative='two-sided',
   usevar='unequal')) Unequal Variance

0.4402421280535648

(-2.839346623928528 3, 0.009542324891921597,21.997899382545775)
```

t-statistic

p-value

df

Class Exercise

- Considering WeChat wants to use algorithms to rank the feeds on WeChat Moments instead of showing the organic feeds chronologically.
- Control Group: show feeds chronologically
- Treatment Group: Rank feeds with algorithms
- OEC: the number of days that a user clicks any feeds on WeChat Moments during the recent 30 days/during experiment.
  What distribution?
- Variants: Control vs. Treatment

Common Distributions of Metrics

- **Bernoulli Distribution** is a discrete probability distribution that represents the outcomes of a single trial that can result in one of two possible outcomes, often labeled as “success” (1) or “failure” (0).
    - These outcomes are typically labeled 1 (success) and 0 (failure), although they can represent any two distinct outcomes.
    - Parameter: p is the probability of success in each trial.
    - Examples: click, purchase, initiate chat (0,1)
    - mean=p, variance=p(1-p)

$$P(X = x) = p^x(1-p)^{(1-x)}$$
for $x = 0, 1$, where p is the probability of success.

- **Binomial Distribution** is a discrete probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success.
    - Two parameters: n is the number of trials and p
    - mean=np, variance = np(1-p)
    - Example: the number of days with click/visit/purchase during experiments

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$
for $k = 0, 1, 2, \ldots, n$, where:

Lift

$$lift = m_1/m_0$$

$$\Delta = m_1 - m_0 = (lift - 1)\,.\,m_0$$

- Lift is also called relative risk, and percent change is (lift -1)
- In practice, we often use lift to interpret the changes.
    - For example, we want this new feature improve 10% CTR.
- We can use difference Δ in statistical tests and then transform/interpret them to/as lifts.

Simulate Data of the Experiment

Population (we draw sample from but cannot observe)

- Lift : CTR(treatment)/CTR(control)
- Lift = 1.1,
- Control Group ~ B(n,p) = B(30,0.5)

We assume there is a lift in population after using the algorithm to rank the feeds. & CTR for now is p=0.5/day

```
lift = 1.1
ctr0=0.5
delta_p = 30*ctr0*(lift-1)
se_p_0 = np.sqrt(30*ctr0*(1-ctr0))
se_p_1 = np.sqrt(30*ctr0*lift*(1-ctr0*lift))
print(delta_p,se_p_0,se_p_1)
```

Binomial distribution
- Mean = np
- Variance = np(1-p) — Changed by Treatment

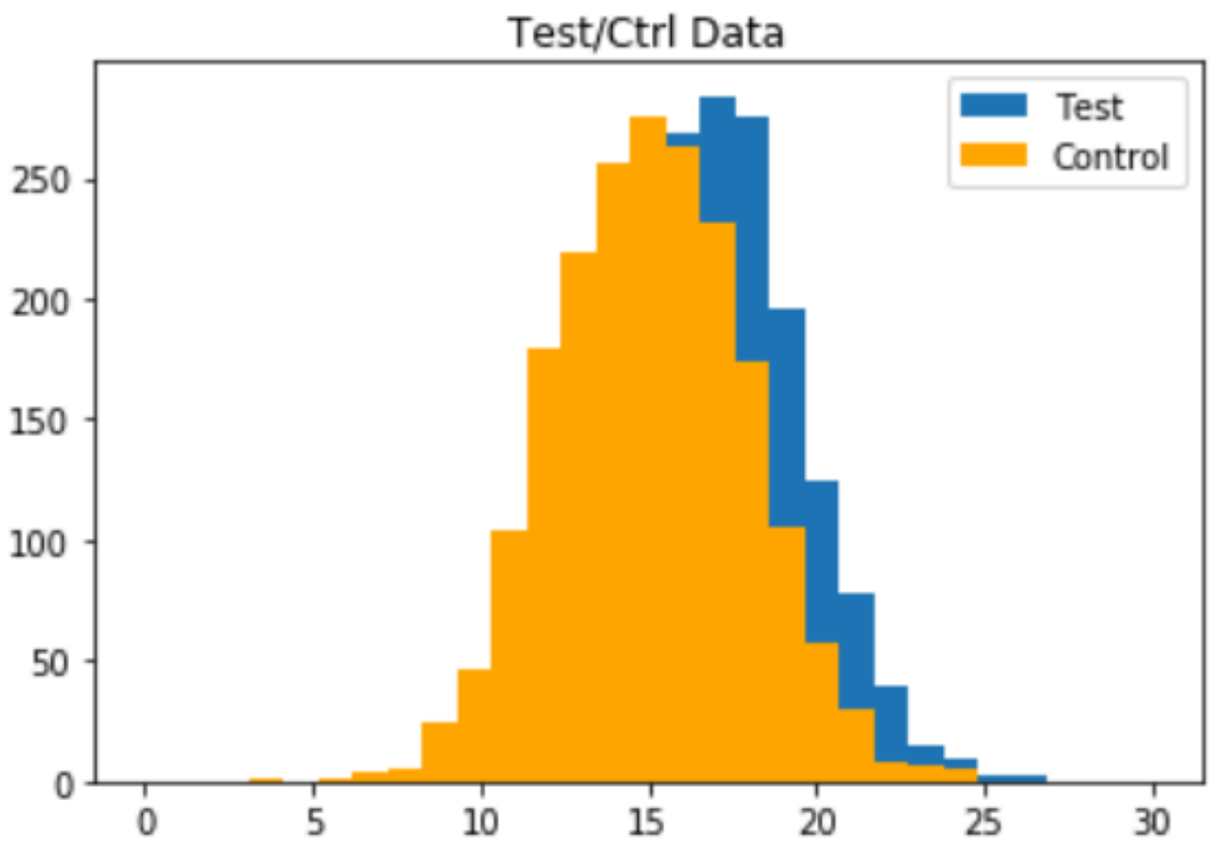
- We collect one sample from each the groups = randomly draw a sample from the population of control and treatment groups.

Sample - We draw samples from population for an experiment to infer δ, k = 1000

```
ctrl = np.random.binomial(30, p=ctr0, size=1000) * 1.0
test = np.random.binomial(30, p=ctr0*lift, size=1000) * 1.0
delta_s=np.mean(test)-np.mean(ctrl)
se0 = np.std(ctrl)
se1 = np.std(test)
print(delta_s,se0,se1)
```

Visualize the Data

```
import matplotlib.pyplot as plt
bins = np.linspace(0, 15, 15)
plt.hist(test, bins=bins, label='Test')
plt.hist(ctrl, bins=bins, label='Control', color='orange')
plt.title('Test/Ctrl Data')
plt.legend()
```



Can you decide whether the treatment is significantly larger than the control?

t-test unequal variance

```
cm = sms.CompareMeans(sms.DescrStatsW(test),sms.DescrStatsW(ctrl))
delta_s = np.mean(test)-np.mean(ctrl)
print(delta_s, cm.std_meandiff_separatevar)
print(cm.ttest_ind(alternative='two-sided', usevar='unequal'))

1.4969999999999999 0.12244436950559884
(12.225960295638982, 3.3707729243977326e-33, 1995.4365382578906)
```

Welsh ttest

How about a larger sample size, k=2000?

```
1.4329999999999998 0.08730332001042354
(16.414037860517876, 1.2156877260544895e-58, 3992.3761345226744)
```

SE becomes smaller

t-stat becomes more extreme

p-value become smaller indicating the difference is more unlikely to happen under NULL; (the difference is likely to happen)

Does it give more confidence to reject the null?

How about a smaller sample, k=20?

Will you trust the result? Why?

https://www.statsmodels.org/stable/generated/statsmodels.stats.weightstats.ttest_ind.html

Review

- Compare means between two groups with hypothesis testing.

H0 (Null Hypothesis): $\delta = \mu_1 - \mu_2 = 0$

H1 (Alternative Hypothesis): $\delta \neq 0$

The two samples are independent, and the variances of the populations are equal/unequal

- What is t statistics and p value?

Hypothesis Testing

Ho (Null Hypothesis): $\delta = \mu_1 - \mu_0 = 0$ (population mean difference = o)

H1 (Alternative Hypothesis): $\delta \neq 0$ ($\mu_1 \neq \mu_2$)

- Y is the metric of interest (OEC, e.g., click)

$\Delta = m_1 - m_0$ (sample mean difference)

- m1, mo are sample means for control and treatment groups.

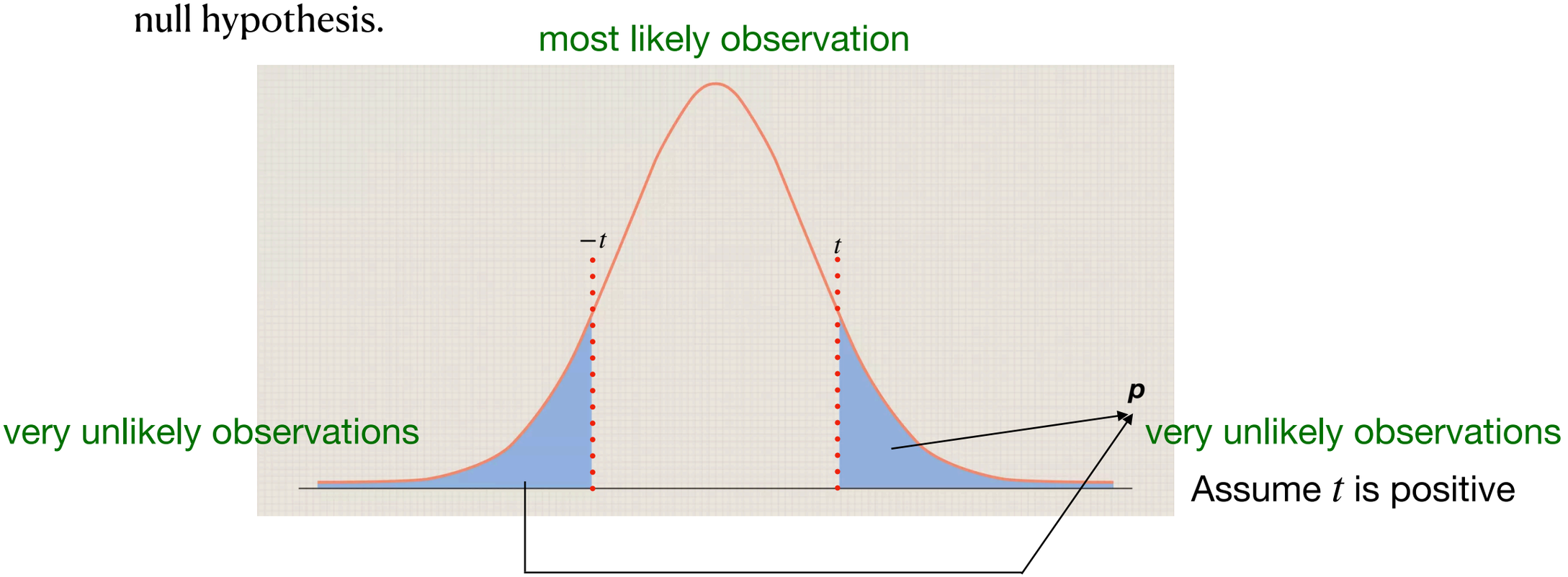
$$t_{stat} = \frac{\Delta - \delta}{se(\Delta)} = \frac{\Delta}{se(\Delta)}$$

- Δ is an estimator of δ
- t looks at how many standard errors away Δ from δ

User	Click	Treatment	
111	1	1	
112	0	0	
113	1	1	
114	0	0	
115	0	0	
116	1	0	
117	0	1	
118	0	0	
119	0	1	
120	1	1	
121	1	1	
122	1	0	
123	1	1	
124	0	0	
125	1	1	
126	0	0	
127	0	1	
128	0	0	

Review: p-value

- p-value evaluates how extreme *t* occurs under the null distribution.
- p-value is the probability of observing *t* or more extreme assuming the Null is true.
  - $p = Pr(|T| \geq t \mid H_0)$
    - the probability of observing such difference or more extreme difference if there really is no difference between Treatment and Control groups (null is true).
  - A very small *p*-value means that such an extreme observed t would be very unlikely under the null hypothesis.

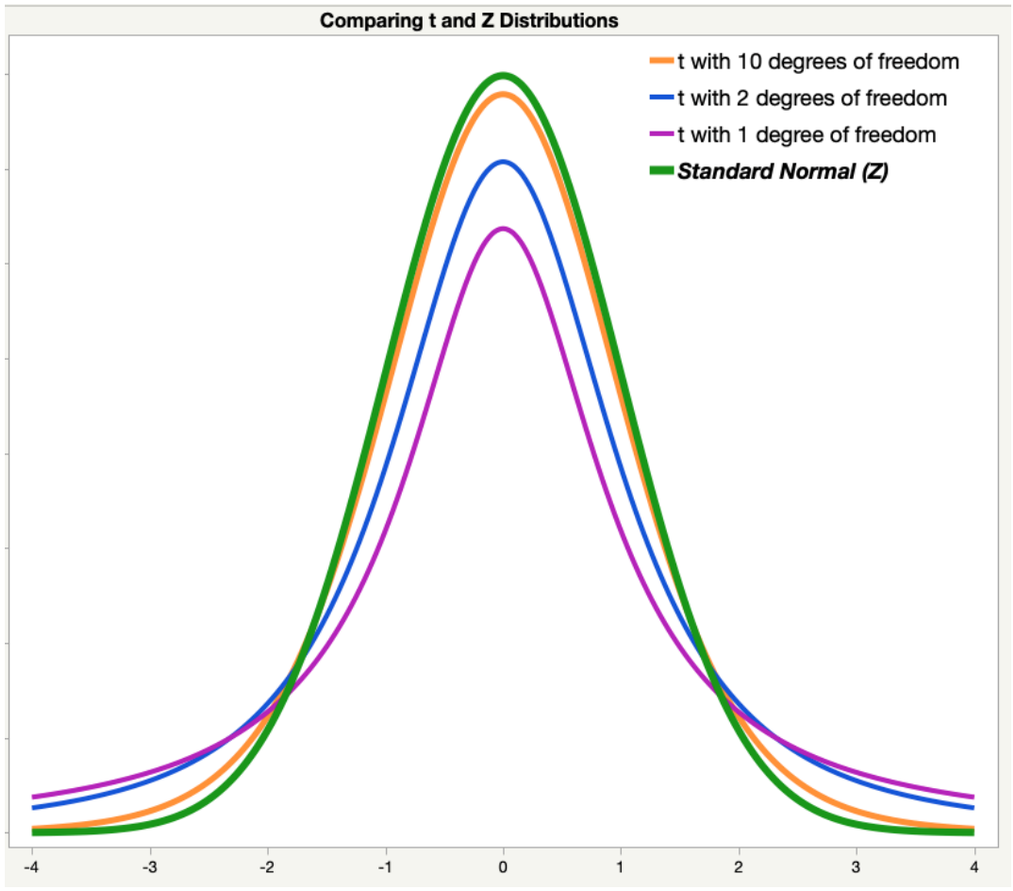


Are they correct or misinterpretations of p-value?

1. If the p-value = 0.06, the null hypothesis has a 6% chance of being true.
2. The p-value is calculated assuming that the Null hypothesis is true.
3. A non-significant difference (e.g., p-value > 0.05) means there is no difference between groups.
4. p-value = 0.05 means that we observed data that would occur only 5% of the time under the Null.
5. You are more confident to reject the Null, if p value is smaller.

t and z Distributions

- t-test is for small samples.
- t-distribution's shape depends on degrees of freedom
    - $df = n_1 + n_2 - 2$
- When N (df) is large enough, t-distribution approximates z-distribution, a standard normal distribution N(0,1).



Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.043	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∝	1.282	1.645	1.960	2.326	2.576	3.291

z-tests

- t-statistic follow student distribution at df=n1+no-2, T ~ t(df)

$$t = \frac{\Delta}{se(\Delta)}$$

- When sample size is sufficiently large, which is often the case in A/B tests
    - t distribution converge to z distribution, which is the standard normal distribution N(0,1), because of the *Central Limit Theorem*.
- z distribution does not depend on sample size.
- When sample is large, we reject the Null if z-statistics are larger than 1.96 or smaller than -1.96. (about 2 standard errors away)
- This also makes the calculation of Confidence Interval easier.

Central Limit Theorem

The distribution of sample means (*m*) approximates a normal distribution as the sample size gets larger, regardless of the population's (*Y*) distribution,

even if the original variables (Y) themselves are not normally distributed.

z-test

Ho (Null Hypothesis): $\delta = \mu_1 - \mu_0 = 0$ (population mean difference)

H1 (Alternative Hypothesis): $\delta \neq 0$ ($\mu_1 \neq \mu_2$)

$\Delta = m_1 - m_0$ (sample mean difference)

- m1, mo are sample means for control and treatment groups.

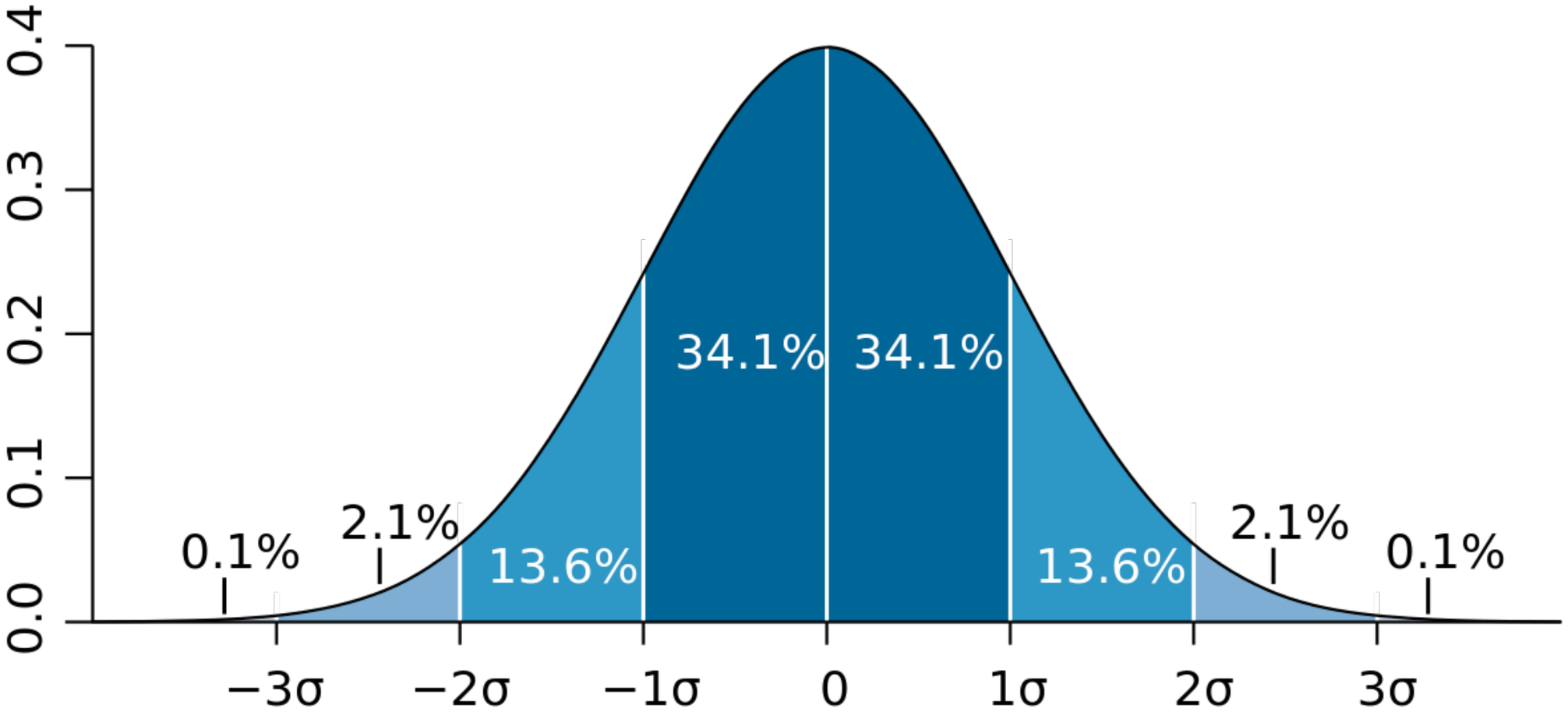
$$z_{stat} = \frac{\Delta - \delta}{se(\Delta)} = \frac{\Delta}{se(\Delta)}$$

Same with t tests!

- Δ is an estimator of δ
- z looks at how many standard errors away from δ

- Null distributions: Z distribution, Z ~ N(0,1)
    - different from (T ~ t(df)), when sample size is small

z distribution



$z = 2, \quad p = 2.2\% \times 2 = 0.044$

$z = 1.96, \quad p = 2.5\% \times 2 = 0.05$

If z > 1.96, we reject the null at $\alpha = 0.05$

Class Exercise

- N = 1000, z-test

```
print(cm.ztest_ind(alternative='two-sided',
usevar='unequal'))
```

(12.844694125583048, 9.211529158169 26e-38)

z-stat **p-value**

Please compare the z test results with those using t tests. What will you conclude?

What would you expect when N= 100?

- N = 1000, t-test

```
print(cm.ttest_ind(alternative='two-sided',
usevar='unequal'))
```

(12.844694125583048, 2.438448215186 5567e-36, 1993.3911532892944)

Multiple Testing

- You may compare many outcome variables (metrics) between control and treatments in one experiment.
- If you find some irrelevant metrics significantly different between control and treatments, with no theory suggesting the significant difference
    - the metrics unaffected by the new feature show significant differences, e.g., users' age, gender, and the behaviors irrelevant to the new feature.
- Will you conclude something wrong with the experiment?

In other words, if you find a metric significant, can you be 100% sure about there is a treatment effect?

at what chance?

Type I error occurs when concluding a significant difference between Treatment and Control when there is no real difference.

Reject the Null, when Null is true.

p-value < 0.05: The observed difference can still happen when Null is true, although the difference is unlikely to happen.

Type I Error

- We reject the Null if the observed difference is unlikely under the Null.
- However, it still can happen with low probability and commit Type I error.
    - For a specific test, what will be the chance of committing a Type I error?
- t-tests with a significance level (α) of 5%
    - 5% chance you will commit a Type I error when you run multiple tests.
- There are ongoing debates calling for a smaller significance level to decrease type I error.
    - Decrease the chance of concluding significant results, especially when there is unlikely a difference in theory.

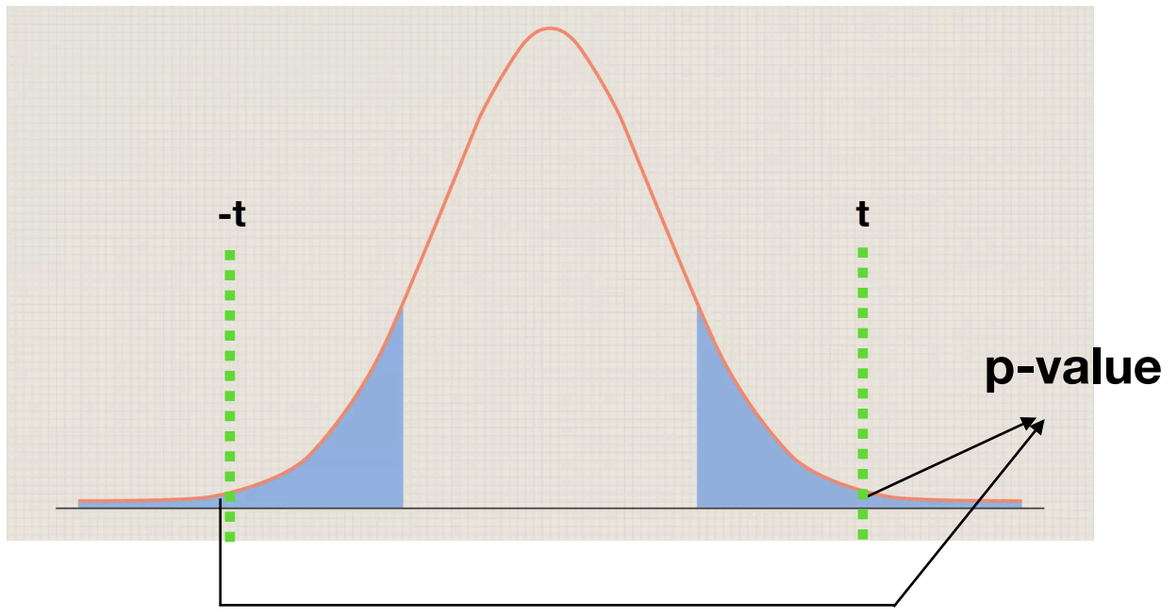
α is set ***before*** experiments/tests;
α is a characteristic of tests rather than for ***a specific*** test.

Type I Error: Right or Wrong?

α=5%

1. There is 5% probability that you incorrectly reject the null for a single t-test.
2. There are 5% of the t tests that may incorrectly reject the null.
3. For a single test, it's either correctly reject the null or incorrectly reject the null.
4. p-value = 0.05 means that if you reject the Null, the probability of a false positive for the test is 5%.

If you compute 20 (*independent)* metrics for your experiment, with the significance level = 5%, around 1 (1/20=5%) metric can still show significant results, even if the new feature does nothing to them.



Type I Error:

When you run 100 tests, you would wrongly reject 5% of the tests (5 test), if $\alpha = 5\,\%$.

How many tests would you wrongly reject, if $\alpha = 1\,\%$.

If you create 100 metrics for your experiment, you find 4 significant results?

Should you conclude the treatment has effects and new feature works?

Class Exercise

- The data (exp_data.csv) is from an experiment run by a company to test a new feature.
- The users were assigned to 3 different groups (conditions, variants).
- Please use t-tests at 5% significance level to compare users' characteristics (the first 14 users' variables).
- Would these users' characteristics be affected by the treatment?
- Will you expect significant differences in users' characteristics between any of the two groups?
- Will you assume equal/unequal variances between groups?

Class Exercise

How many significant differences in 42 tests are allowed by Type I Error , α=5% ?

```
gender
(1.338773515763321, 0.18068950041259244, 6750.0)
(0.9053695111401618, 0.36530258109856395, 6626.0)
(-0.41833581824238186, 0.6757151790218867, 6618.0)
age
(-0.2258780702654209, 0.8213030931691397, 6750.0)
(0.9497330944934246, 0.3422825277496915, 6626.0)
(1.1687121493699624, 0.24256172111339383, 6618.0)
device
(0.10889071847020237, 0.9132924051645153, 6750.0)
(0.5134641434648958, 0.6076438065780314, 6626.0)
(0.41388040743233595, 0.6789751051271262, 6618.0)
has_interest_online
(-1.2845094396946113, 0.19900784380497932, 6750.0)
(-1.6808531965006905, 0.0928385650413649, 6626.0)
(-0.407926212774742 84, 0.6833410435536809, 6618.0)
interestss_TVShows
(-0.5234272052816967, 0.6006941789213049, 6750.0)
(0.23350129033532566, 0.8153793808681323, 6626.0)
(0.6736860794427328, 0.5005344408716177, 6618.0)
interests_Travel
(0.06482087845276627, 0.9483185373224183, 6750.0)
(-0.5013152001110853, 0.6161659636359065, 6626.0)
(-0.5394472710402294, 0.5895964190860086, 6618.0)
interests_Society
(-0.6847791455098242, 0.4935068089879331, 6750.0)
(1.2351380997665236, 0.2168229057175556, 6626.0)
(1.9446141119994387, 0.0518637481483557, 6618.0)
interests_Pets
(0.9707276306009271, 0.3317187109438976, 6750.0)
(0.3147506430846481, 0.7529608999543934, 6626.0)
(-0.6519002384228214, 0.5144881696069725, 6618.0)
interests_Natural
(-1.0612678410854939, 0.2886061658249812, 6750.0)
(-1.8427709795172764, 0.06540709922408983, 6626.0)
(-1.3663540716083276, 0.17187427850386666, 6618.0)
interests_Cars
(0.6187802423559168, 0.5360820204981972, 6750.0)
(-1.3893444622272728, 0.1647747146564894, 6626.0)
(-1.7725664974397572, 0.07634656053156527, 6618.0)
interests_Foods
(1.0184637185219496, 0.30849413623581146, 6750.0)
(2.0767837483140696, 0.0378599681773576, 6626.0)
(1.016626487214549, 0.3093683127187254, 6618.0)
interests_Music
(-0.3262239225314997, 0.7442650563623125, 6750.0)
(0.715621661229616, 0.4742502108804256, 6626.0)
(0.9673603471637802, 0.3333992746546236, 6618.0)
interests_Digital
(-1.7427156587400414, 0.08142883367459593, 6750.0)
(-2.227313101078629, 0.025959817302320727, 6626.0)
(-0.45167712090086365, 0.6515164041179344, 6618.0)
interests_Life
(0.27387057610143695, 0.7841924640867347, 6750.0)
(1.7190916772349614, 0.08564438610483052, 6626.0)
(1.4489976215244424, 0.14738562626501506, 6618.0)
```

If p-value of one of the tests is very small, 0.0001,
will you have more concerns?

Decrease α, and see whether it still holds.

Class Exercise

How many significant differences in 42 tests are allowed by Type I Error α=1%?

gender
(1.338773515763321, 0.18068950041259244, 6750.0)
(0.9053695111401618, 0.36530258109856395, 6626.0)
(-0.41833581824238186, 0.6757151790218867, 6618.0)
age
(-0.2258780702654209, 0.8213030931691397, 6750.0)
(0.9497330944934246, 0.3422825277496915, 6626.0)
(1.1687121493699624, 0.24256172111339383, 6618.0)
device
(0.10889071847020237, 0.9132924051645153, 6750.0)
(0.5134641434648958, 0.6076438065780314, 6626.0)
(0.41388040743233595, 0.6789751051271262, 6618.0)
has_interest_online
(-1.2845094396946113, 0.19900784380497932, 6750.0)
(-1.6808531965006905, 0.0928385650413649, 6626.0)
(-0.407926212774742 84, 0.6833410435536809, 6618.0)
interestss_TVShows
(-0.5234272052816967, 0.6006941789213049, 6750.0)
(0.23350129033532566, 0.8153793808681323, 6626.0)
(0.6736860794427328, 0.5005344408716177, 6618.0)
interests_Travel
(0.06482087845276627, 0.9483185373224183, 6750.0)
(-0.5013152001110853, 0.6161659636359065, 6626.0)
(-0.5394472710402294, 0.5895964190860086, 6618.0)
interests_Society
(-0.6847791455098242, 0.4935068089879331, 6750.0)
(1.2351380997665236, 0.2168229057175556, 6626.0)
(1.9446141119994387, 0.0518637481483557, 6618.0)

interests_Pets
(0.9707276306009271, 0.3317187109438976, 6750.0)
(0.3147506430846481, 0.7529608999543934, 6626.0)
(-0.6519002384228214, 0.5144881696069725, 6618.0)
interests_Natural
(-1.0612678410854939, 0.2886061658249812, 6750.0)
(-1.8427709795172764, 0.06540709922408983, 6626.0)
(-1.3663540716083276, 0.17187427850386666, 6618.0)
interests_Cars
(0.6187802423559168, 0.5360820204981972, 6750.0)
(-1.3893444622272728, 0.1647747146564894, 6626.0)
(-1.7725664974397572, 0.07634656053156527, 6618.0)
interests_Foods
(1.0184637185219496, 0.30849413623581146, 6750.0)
(2.0767837483140696, 0.0001 6626.0)
(1.016626487214549, 0.3093683127187254, 6618.0)
interests_Music
(-0.3262239225314997, 0.7442650563623125, 6750.0)
(0.715621661229616, 0.4742502108804256, 6626.0)
(0.9673603471637802, 0.3333992746546236, 6618.0)
interests_Digital
(-1.7427156587400414, 0.08142883367459593, 6750.0)
(-2.227313101078629, 0.025959817302320727, 6626.0)
(-0.45167712090086365, 0.6515164041179344, 6618.0)
interests_Life
(0.27387057610143695, 0.7841924640867347, 6750.0)
(1.7190916772349614, 0.08564438610483052, 6626.0)
(1.4489976215244424, 0.14738562626501506, 6618.0)

Class Exercise

How many significant differences in 42 tests are allowed by Type I Error α=1%?

```
gender
(1.338773515763321, 0.18068950041259244, 6750.0)
(0.9053695111401618, 0.36530258109856395, 6626.0)
(-0.41833581824238186, 0.6757151790218867, 6618.0)
age
(-0.2258780702654209, 0.8213030931691397, 6750.0)
(0.9497330944934246, 0.3422825277496915, 6626.0)
(1.1687121493699624, 0.24256172111339383, 6618.0)
device
(0.10889071847020237, 0.9132924051645153, 6750.0)
(0.5134641434648958, 0.6076438065780314, 6626.0)
(0.41388040743233595, 0.6789751051271262, 6618.0)
has_interest_online
(-1.2845094396946113, 0.19900784380497932, 6750.0)
(-1.6808531965006905, 0.0928385650413649, 6626.0)
(-0.407926212774742 84, 0.6833410435536809, 6618.0)
interestss_TVShows
(-0.5234272052816967, 0.6006941789213049, 6750.0)
(0.23350129033532566, 0.8153793808681323, 6626.0)
(0.6736860794427328, 0.5005344408716177, 6618.0)
interests_Travel
(0.06482087845276627, 0.9483185373224183, 6750.0)
(-0.5013152001110853, 0.6161659636359065, 6626.0)
(-0.5394472710402294, 0.5895964190860086, 6618.0)
interests_Society
(-0.6847791455098242, 0.4935068089879331, 6750.0)
(1.2351380997665236, 0.2168229057175556, 6626.0)
(1.9446141119994387, 0.0518637481483557, 6618.0)
interests_Pets
(0.9707276306009271, 0.3317187109438976, 6750.0)
(0.3147506430846481, 0.7529608999543934, 6626.0)
(-0.6519002384228214, 0.5144881696069725, 6618.0)
interests_Natural
(-1.0612678410854939, 0.2886061658249812, 6750.0)
(-1.8427709795172764, 0.06540709922408983, 6626.0)
(-1.3663540716083276, 0.17187427850386666, 6618.0)
interests_Cars
(0.6187802423559168, 0.5360820204981972, 6750.0)
(-1.3893444622272728, 0.1647747146564894, 6626.0)
(-1.7725664974397572, 0.07634656053156527, 6618.0)
interests_Foods
(1.0184637185219496, 0.30849413623581146, 6750.0)
(2.0767837483140696, 0.0378599681773576, 6626.0)
(1.016626487214549, 0.3093683127187254, 6618.0)
interests_Music
(-0.3262239225314997, 0.7442650563623125, 6750.0)
(0.715621661229616, 0.4742502108804256, 6626.0)
(0.9673603471637802, 0.3333992746546236, 6618.0)
interests_Digital
(-1.7427156587400414, 0.08142883367459593, 6750.0)
(-2.227313101078629, 0.025959817302320727, 6626.0)
(-0.45167712090086365, 0.6515164041179344, 6618.0)
interests_Life
(0.27387057610143695, 0.7841924640867347, 6750.0)
(1.7190916772349614, 0.08564438610483052, 6626.0)
(1.4489976215244424, 0.14738562626501506, 6618.0)
```

What should you do when a metric is unexpectedly significant?

- Bonferroni Correction - Too Conservative !
  - Use smaller p-value (p-value/# tests)
- A **practical** simple two-step rule-of-thumb:

1. Separate all metrics into three groups:
   - First-order metrics: those you expect to be impacted by the new feature **5%**
   - Second-order metrics: those potentially to be impacted (e.g., through cannibalization) **1%**
   - Third-order metrics: those not to be impacted **0.1%**
2. Apply tiered significance levels to each group