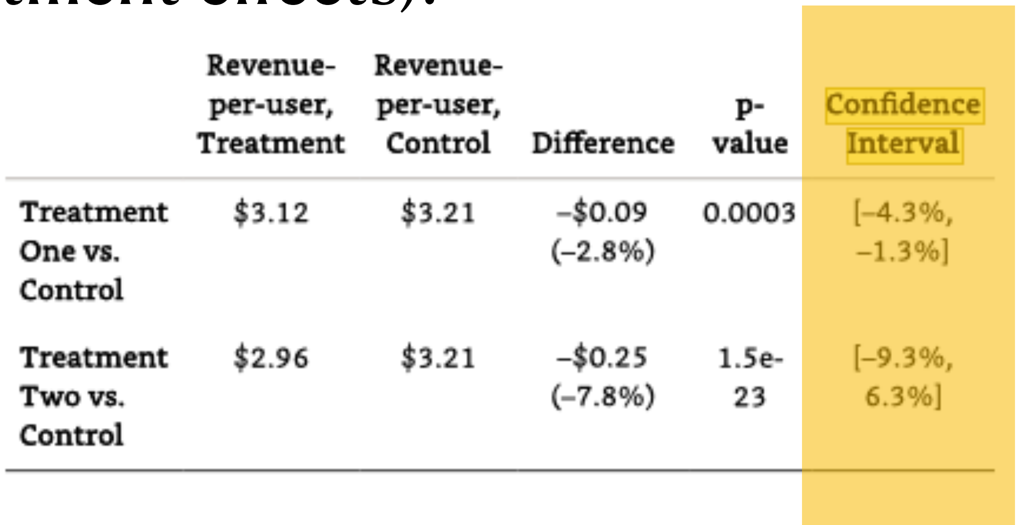
S3 Statistics Critical to Experimentation II

Shan Huang, HKU

Confidence Interval (CI)

- t-tests or z-tests can only conclude whether δ is significantly different from the hypothesized value (e.g., 0).
- t-tests or z-tests cannot tell you about the *effect size and the uncertainty of the effects*, say how large and volatile the differences (treatment effects) are. — based on our sample (observation)
- Confidence Intervals can inform you about both.
- The confidence level represents how often the confidence interval should contain the δ (e.g., the true/population treatment effects).



	Revenue-per-user, Treatment	Revenue-per-user, Control	Difference	p-value	Confidence Interval
Treatment One vs. Control	$3.12	$3.21	–$0.09 (–2.8%)	0.0003	[–4.3%, –1.3%]
Treatment Two vs. Control	$2.96	$3.21	–$0.25 (–7.8%)	1.5e-23	[–9.3%, 6.3%]

CI and Hypothesis Testing

- $(1-\alpha)$ CI represents that the CI should contain δ (e.g., true treatment effects, $\mu_1 - \mu_2$) for $(1-\alpha)$ of the times

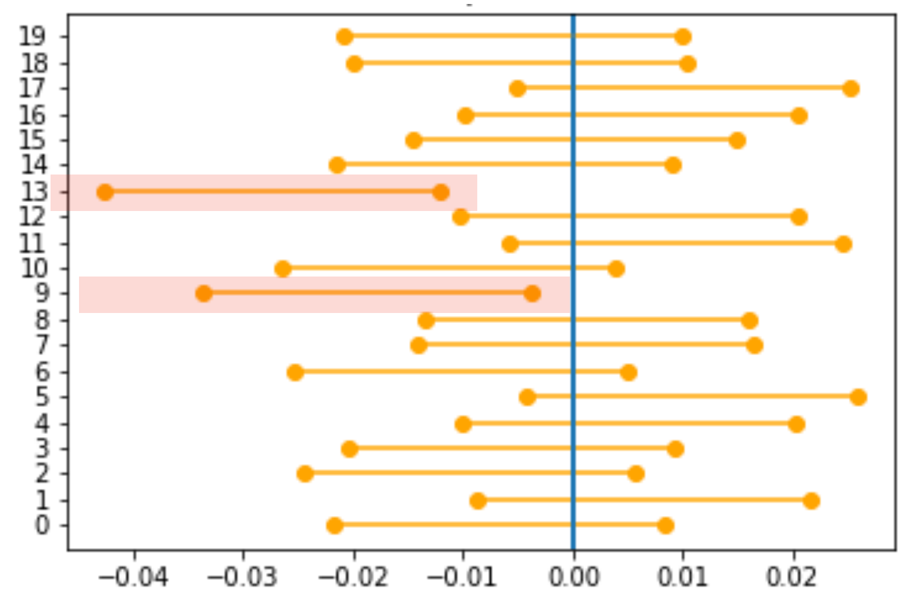
e.g., 95% CI should contain δ for 95% of the times.

- Duality between p-values and CIs:
  - Null Hypothesis: $\delta = \mu_1 - \mu_2 = x$ **= 0**
  - 95% CI does not contain x, implies $p < 0.05$
  - $(1-\alpha)$ CI does not contain x, implies $p < \alpha$

Reject the Null, δ is significantly different from x

Which CIs show significantly *negative* treatment effects?

Confidence Intervals of Treatment Effects

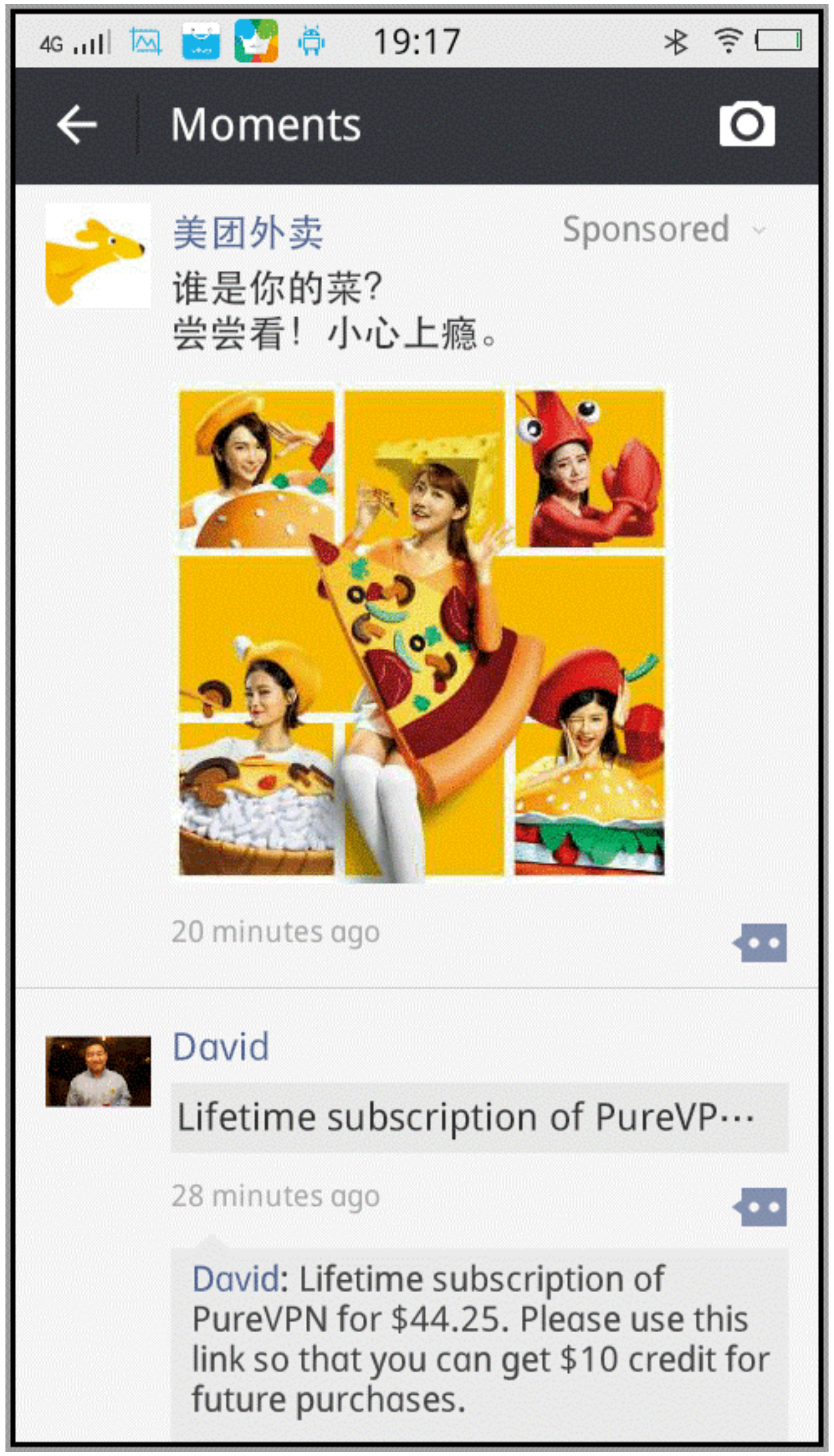


Confidence Interval (CI)

- $1 - \alpha$ Confidence Intervals of $\delta = \mu_1 - \mu_2$ have the form: $[\Delta - se\,.\,t_{\alpha/2}, \Delta + se\,.\,t_{\alpha/2}]$ (CIs)
- $\alpha = 5\,\%$, when n is sufficiently large, $t_{\alpha/2} = z_{\alpha/2}$ = 1.96, the form becomes: $[\Delta - se.1.96, \Delta + se.1.96]$ (z CIs)
- $\hat{\Delta} = m_1 - m_0$ is the point estimate of $\delta = \mu_1 - \mu_2$
- The length of CI is 3.92 (2*1.96) se.
- The smaller se is, the less uncertainties of CI indicate about δ .

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.
- Variants: Control vs. Treatment



美团外卖

谁是你的菜?
尝尝看！ 小心上瘾。

20 minutes ago

David

Lifetime subscription of PureVP…

28 minutes ago

David: Lifetime subscription of PureVPN for $44.25. Please use this link so that you can get $10 credit for future purchases.

Class Exercise

Calculate the CI of the treatment effects and decide how the treatment significantly improves OEC.

Create and save it as CI.ipynb

```
lift = 1.1
ctr0=0.5
ctrl = np.random.binomial(30, p=ctr0, size=1000) * 1.0
test = np.random.binomial(30, p=ctr0*lift, size=1000) * 1.0

delta_p = 30*ctr0*(lift-1)
delta_s = np.mean(test)-np.mean(ctrl)
print(delta_s)
se0 = np.std(ctrl)
se1 = np.std(test)
print(se0,se1)
```

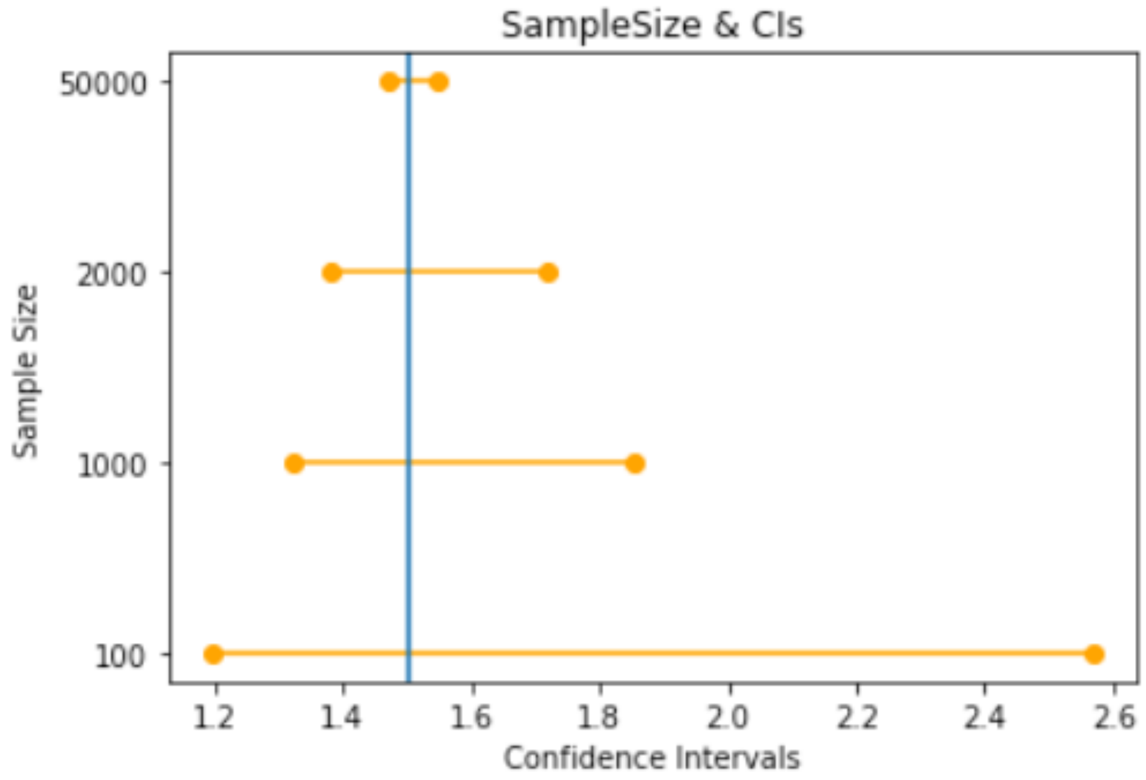
Class Exercise

Based on t_z_tests.ipynb, calculate the t and z confidence intervals, when k = 1000

```
print(cm.tconfint_diff(alpha=0.05, alternative='two-sided',
usevar='unequal'))
```

```
print(cm.zconfint_diff(alpha=0.05, alternative='two-sided',
usevar=‘unequal’))
```

Calculate the t confidence intervals, when k = 100, 1000, 2000, 5000. Which CI is the most accurate, telling you the most about δ ?



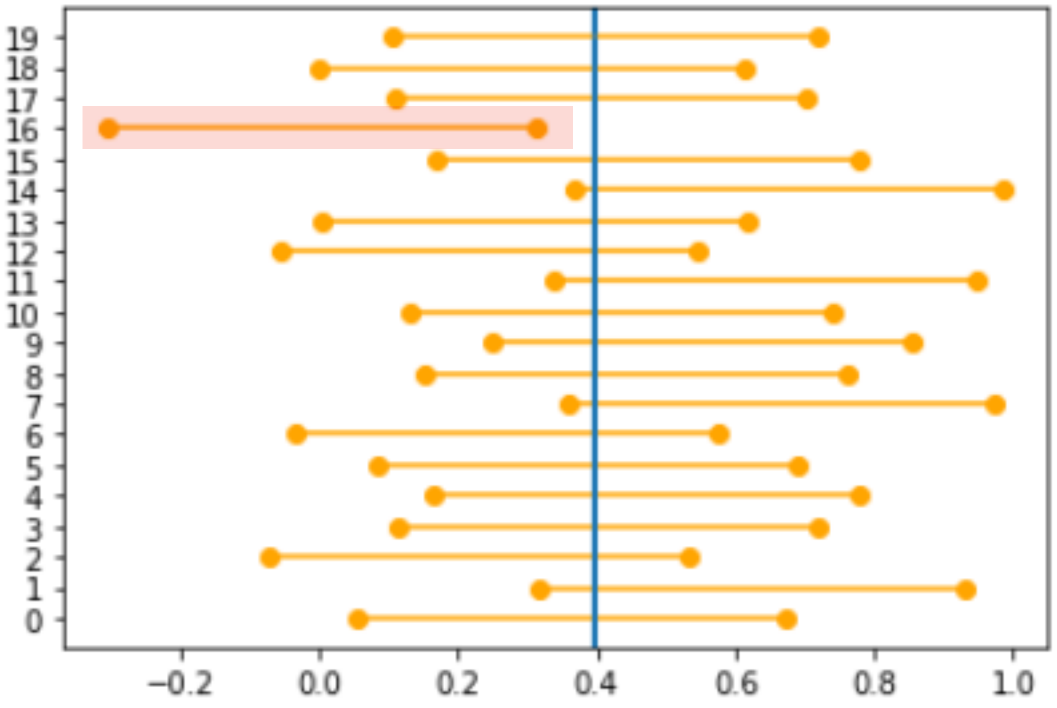
CI and Sample Sizes

Loosely speaking, the tighter CIs indicate a greater statistical power

Right or Wrong about CI?

- *For one sample (test),* there is 95% chance that a CI contain δ
- *For many samples (tests),* 95% of the CIs would be expected to contain δ
- For a given CI based on a sample, it either contains δ or does not contain δ

We repeatedly draw 20 random samples from the population with δ = 0.4. The confidence intervals computed by these 20 samples are as follow:



For a given CI based on one sample, it either contains δ or not.

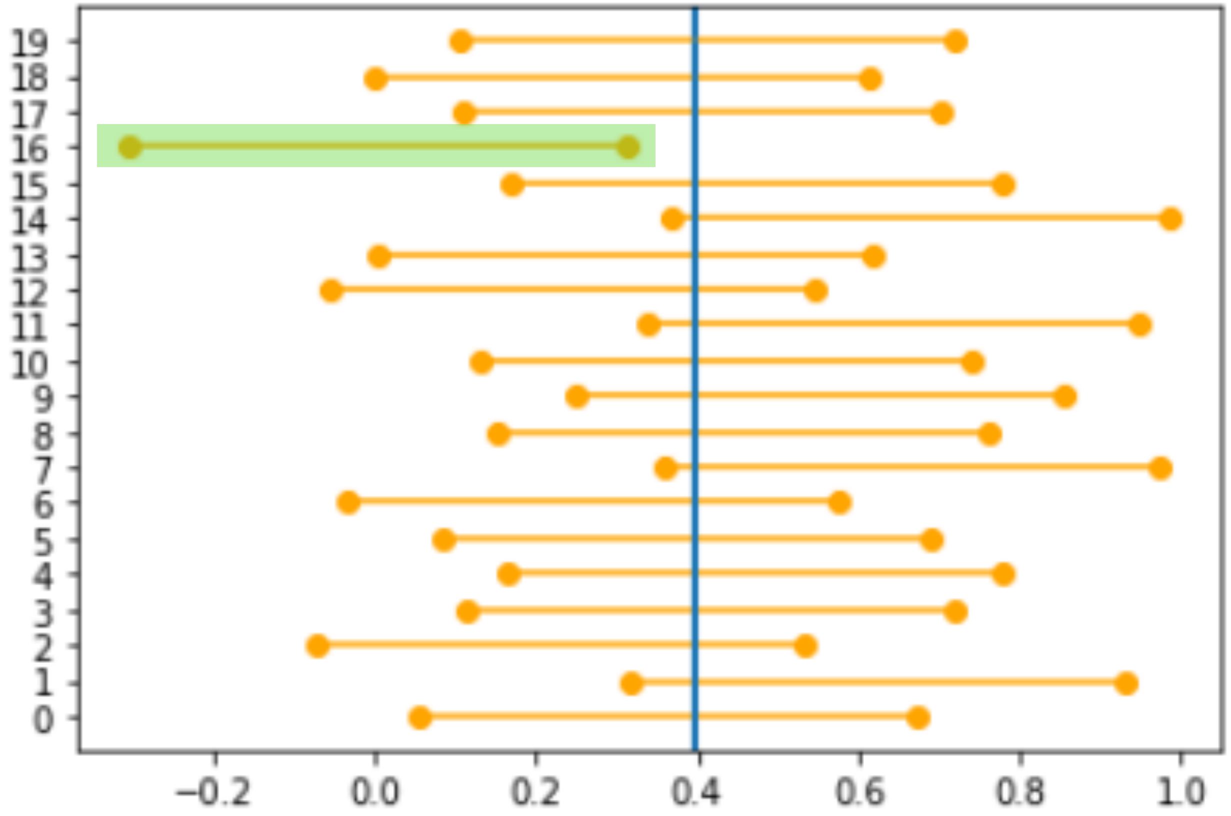
1 of the 20 CIs does not contain $\delta =$ 0.4.

CI and Type I Error

$(1-\alpha)$ CI represents that the CI should contain δ (e.g., true treatment effects, $\mu_1 - \mu_2$) for $(1-\alpha)$ of the times.

Type I Error is α:

- There is α of the times (samples) that $(1-\alpha)$ CIs *do not* contain δ**.**
- **For example,** Type I Error is 5% meaning that 95%-CIs do not contain δ for 5% of the times (samples).



1 out of 20 (5%*20) CIs do not contain δ = 0.4.

Class Exercise

CIs for testing the same variable (δ) based on different samples:

- Assume :

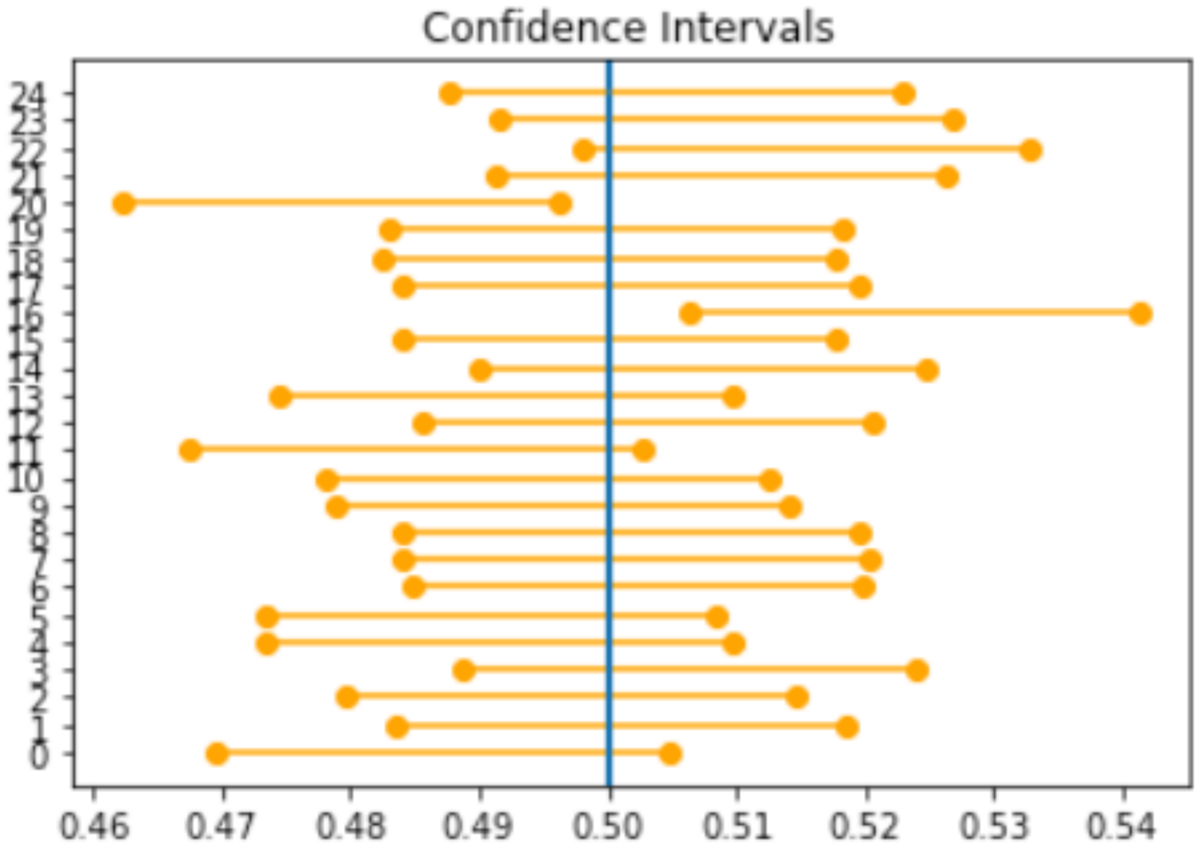
```
lift = 1.1
std=0.2
delta_p=5*(lift-1)
ctrl = np.random.normal(5,std,size = 1000)
test = np.random.normal(5*lift,std, size=1000)
```

- Draw 25 samples from the population.
- Show the CIs and interpret Type I error by the results.

Class Exercise

- Visualize CIs Example (you may use different approaches).

```
data_dict = {}
data_dict['Number'] = [x for x in range(25)]
data_dict['lower'] = [x[0] for x in ci]
data_dict['upper'] = [x[1] for x in ci]
dataset = pd.DataFrame(data_dict)
print(dataset)

for lower,upper,y in zip(dataset['lower'],dataset['upper'],range(len(dataset))):
    plt.plot((lower,upper),(y,y),'ro-',color='orange')
    plt.yticks(range(len(dataset)),list(dataset['Number']))
    plt.axvline(delta_p2)
plt.title('Confidence Intervals')
plt.savefig("CIMean0.png")
```



Draw less samples (10) and Show their CIs?

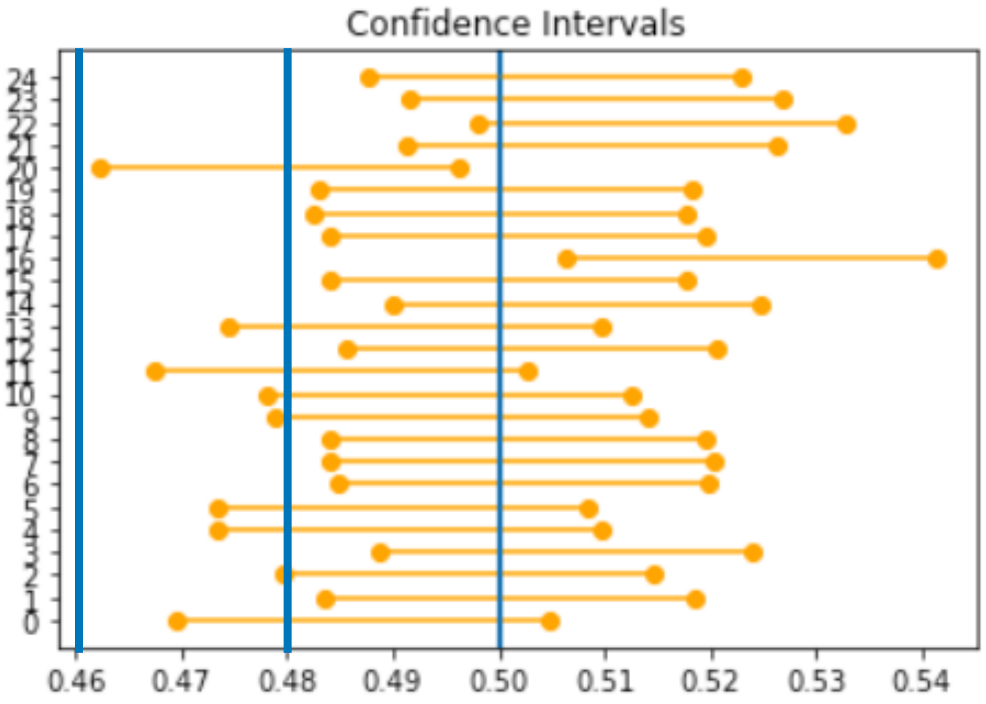
=< 2 CIs do not contain 0.5

Class Exercise

CIs for Multiple Testing

- Use the data (exp_data.csv), from an experiment that tests a new feature *(based on Mutiple_Testing.ipynb)*.
- The users were assigned to 3 different groups (variants).
- Please use 95% CIs to find whether users' characteristics (the first 14 users' variables) are statistically significantly different from one another.
- Will you assume equal/unequal variances between the variants?

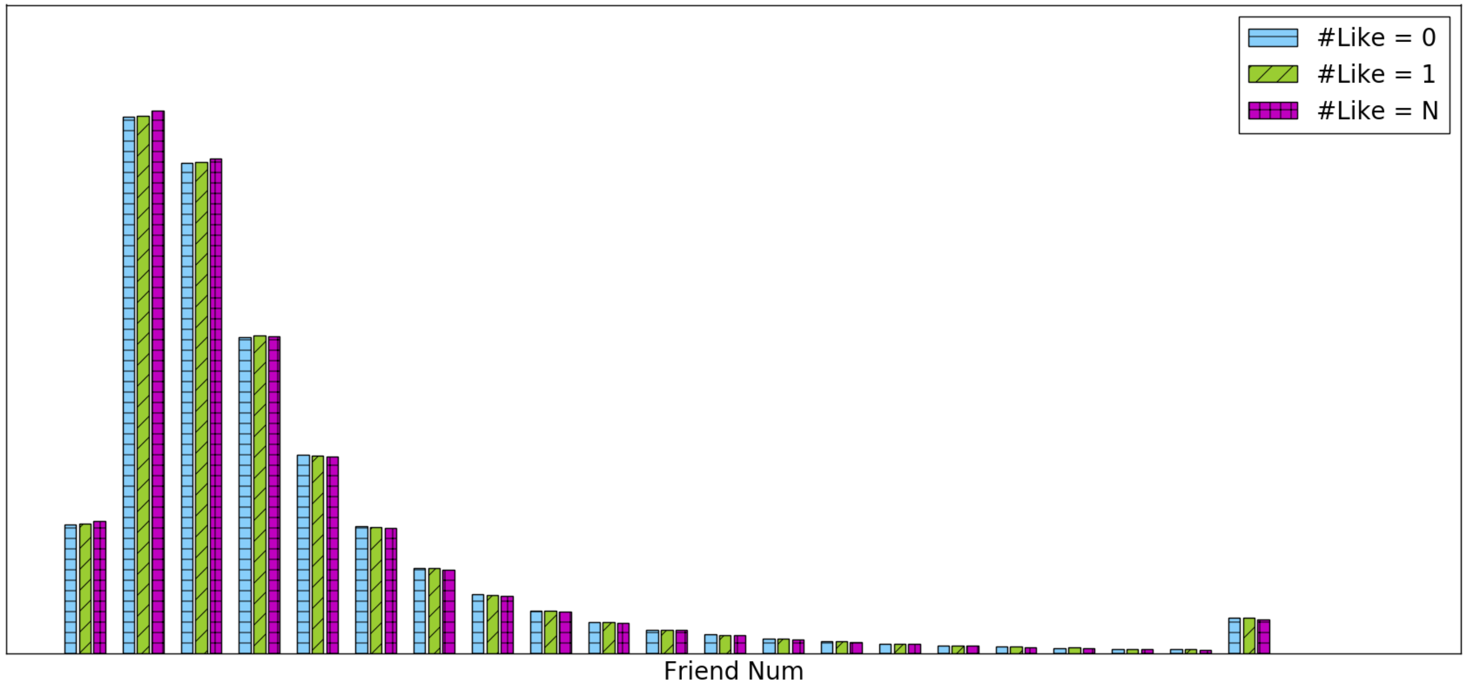
Confidence Intervals (CIs) vs. p-value



- CIs: 95% CI should contain δ for 95% of the times.
- Duality between p-values and CIs:
    - Null Hypothesis: $\delta = \mu_1 - \mu_2 = x$
    - 95% CI does not contain x, implies p < 0.05
    - $(1 - \alpha)$ CI does not contain x, which implies p < α
- Differences:
    - Hypothesis testing, the p-value, shows how extreme the observed value is compared to the hypothesized value. — infer whether δ is statistically significantly different than x (a certain value, 0)
        - When p value is smaller, it is more convincing to reject the null.
    - CIs show the effect size (how large) and the uncertainty (how volatile) of δ (population characteristic & mean difference).
        - Can tell you whether δ is statistically significantly larger (smaller) than x (x could vary)
        - Cannot tell you the confidence of rejecting the Null - δ is different from x

What will you do?

Should we stop using t/z tests, when metrics (Y) do not follow normal distributions?



Normality Assumption

Misunderstanding: Normality assumption on the sample distribution of metrics (e.g., revenue, sales, CTR, network degree, and etc) for t/z tests.

Central Limit Theorem

- The mean (e.g., $\bar{y}_1, \bar{y}_0$) of a large number of *independent* observations of a random variable (e.g., Y_1, Y_0) is approximately a normal random variable.
- The sum of two random variables $\Delta = \bar{y}_1 - \bar{y}_0 = m_1 - m_0$ that follow normal distributions would also follow a normal distribution
- Y_1, Y_0 do not have to follow normal distributions.

Normality Assumption

When:

1. There is a *large* number of units (e.g., users, user-ads)
2. The observations of different units are not correlated with one another
   - Different users' behaviors tend to be independent.
     - Social behaviors between friends can be correlated.
   - The behaviours of the same users on different pages/products/ads tend to be correlated.

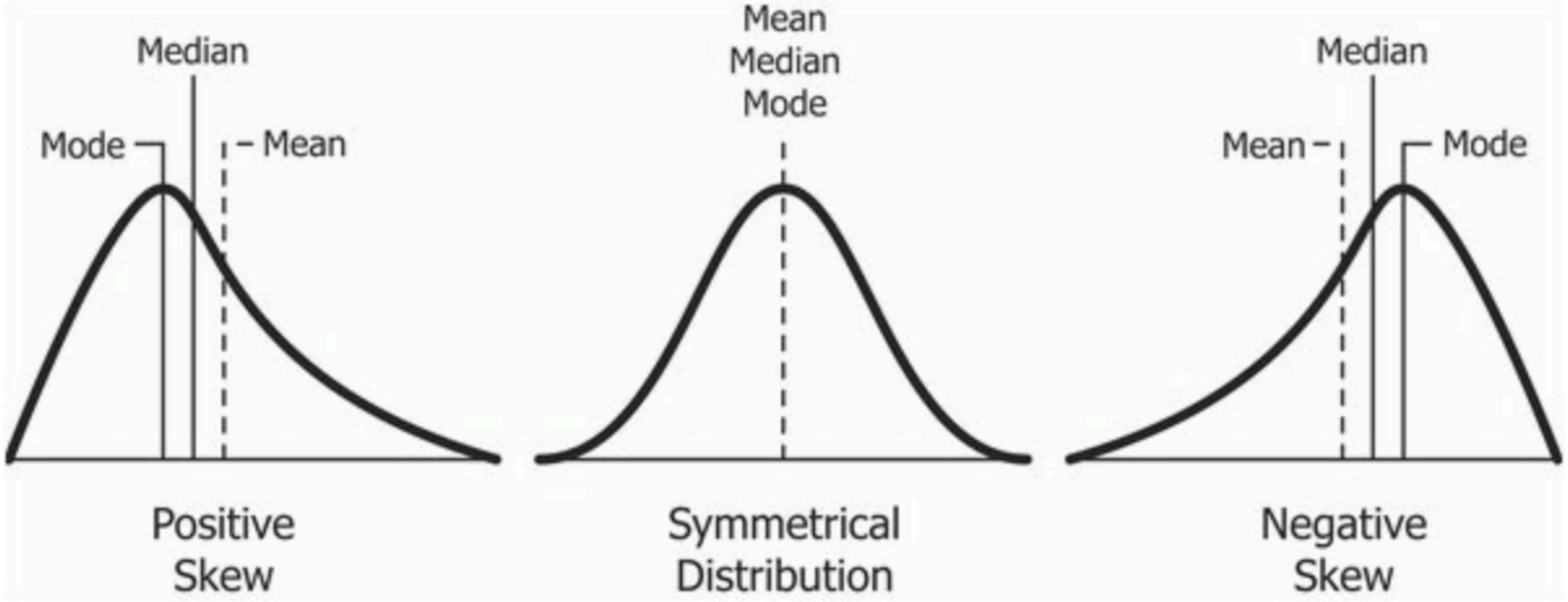
we can safely use t tests (or z tests).

- How *large* should the sample size be to assume normality on Δ?

Rule-of-Thumb: Sample Size and Skewness

- The minimum number of samples needed for the Δ to have normal distribution is **355**$\cdot s^2$ for each group (variant)
- s is the skewness coefficient of the sample distribution of the metric Y.
    - $s = E[Y - E(Y)]^3 / [Var(Y)]^{3/2}$
- You would need a larger sample size (n) for a more skewed metric (Y).

Skewness

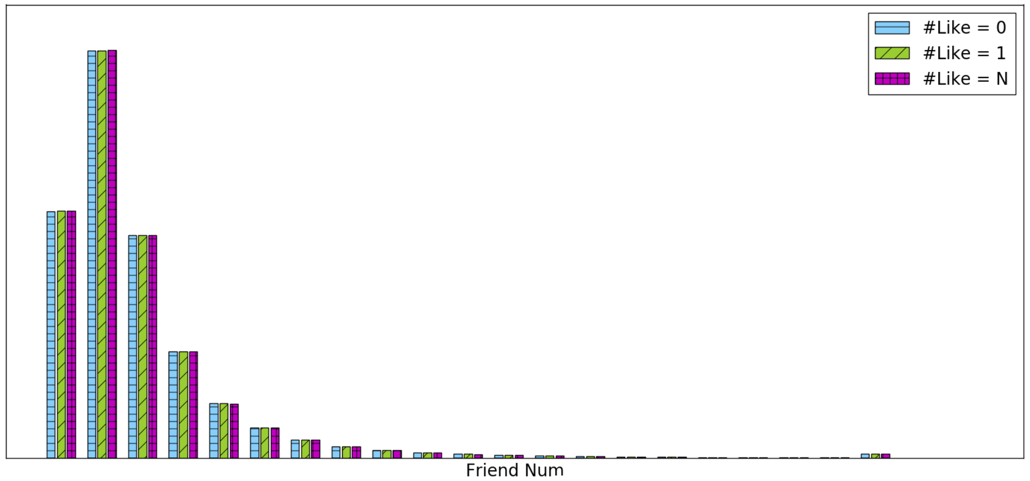


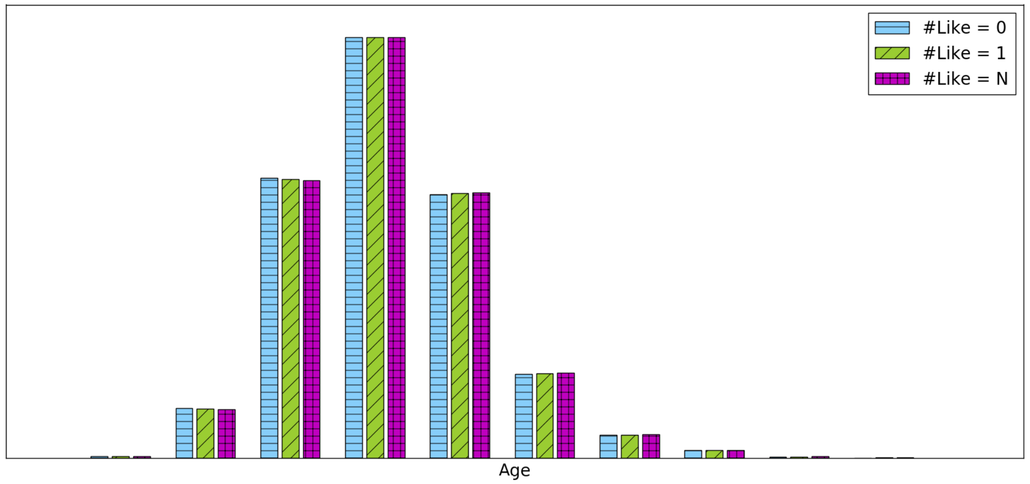
- "Skewness is **a measurement of the distortion of symmetrical distribution or asymmetry in a data set**. Skewness is demonstrated on a bell curve when data points are not distributed symmetrically to the left and right sides of the median on a bell curve."

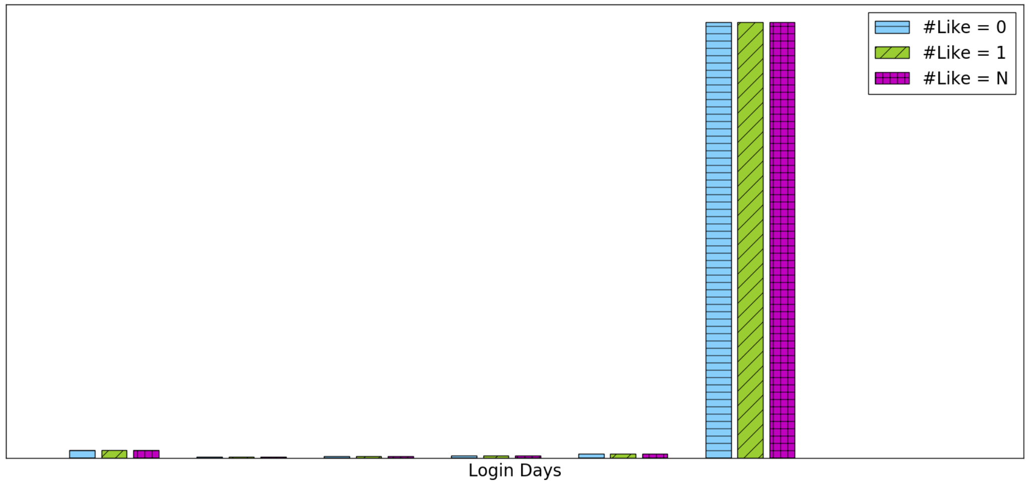
What metrics tend to be skewed?

Skewness

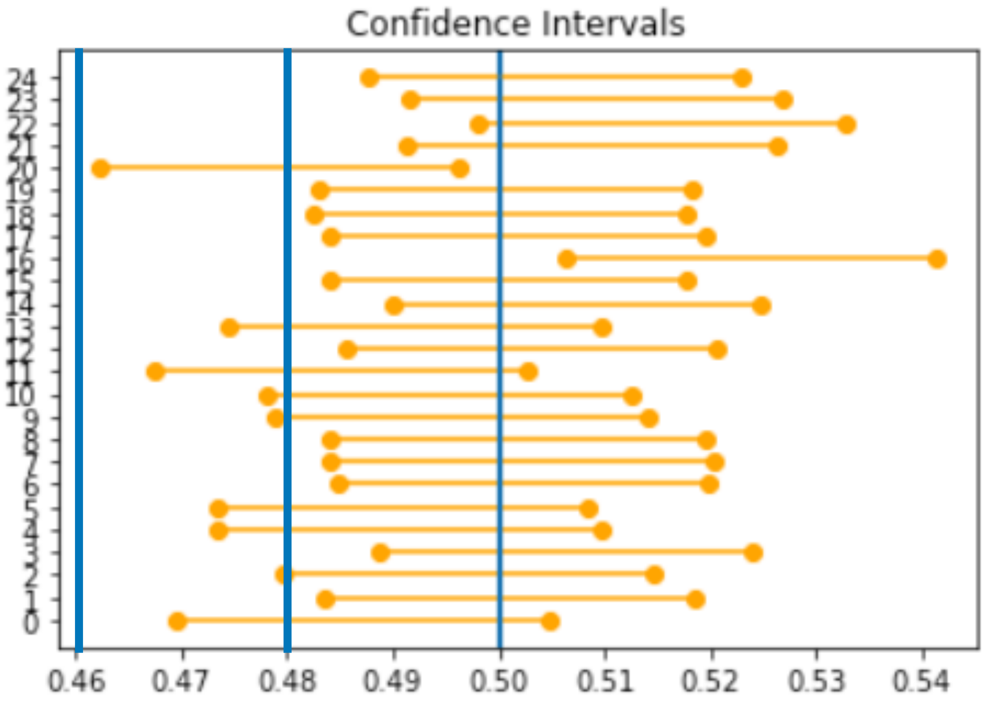
- Some metrics, such as product revenue and downloads, video views, number of friends, tend to have a high skewness
  - e.g., only a very small amount of products, videos and users have a very large amount of revenue, views, and friends.







Review: Confidence Intervals (CIs) vs. p-value



- CIs: 95% CI should contain δ for 95% of the times.
- Duality between p-values and CIs:
  - Null Hypothesis: $\delta = \mu_1 - \mu_2 = x$
  - 95% CI does not contain x, implies p < 0.05
  - $(1-\alpha)$ CI does not contain x, which implies p < α
- Differences:
  - Hypothesis testing, the p-value, shows how extreme the observed value is compared to the hypothesized value. — infer whether δ is statistically significantly different than x (a certain value, 0)
    - When p value is smaller, it is more convincing to reject the null.
  - CIs show the effect size (how large) and the uncertainty (how volatile) of δ (population characteristic & mean difference).
    - Can tell you whether δ is statistically significantly larger (smaller) than x (x could vary)
    - Cannot tell you the confidence of rejecting the Null - δ is different from x

If the sample size is too small to use t-tests, what will you do?

A. Increase the sample size

B. Reduce the skewness of Y

Reduce Skewness of Data

- Some metrics, such as product revenue and downloads, video views, number of friends, tend to have a high skewness
    - e.g., only a very small amount of products, videos and users have a very large amount of revenue, views, and friends.
- Transform the metric or cap the values.
    - Log transform the number of friends (long-tail distribution)
    - Cap the sales
        - e.g., after Bing capped Revenue/User to $10 per user per week, they saw skewness drop from 18 to 5.

Class Exercise: Skewness and Sample Size

Use the data (exp_data_2.csv), from an experiment that tests a new feature on newsfeed ads.

1. Visualize the variables (*age, gender, friend_cnt, sns_like_cnt, sns_comment_cnt*)
2. Observe their distribution - positive/negative skewed
3. Calculate their skewness
4. Find the minimum sample size for them to use t/z tests properly.

Class Exercise: Skewness and Sample Size

```
from scipy.stats import skew
>>> skew([1, 2, 3, 4, 5])
```

```
def skew(x):
    cprint(x,'red', 'on_yellow')
    s= df[x].skew()
    print('skew', s )
    df[x].plot.hist(bins=500,edgecolor='black')
    size = 355*s*s
    print('Sample Size',size)
```

When can't we use t/z tests (CIs)?

Central Limit Theorem

- The **mean** (e.g., $\bar{y}_1, \bar{y}_0$) of a large number of *independent* observations of a random variable (e.g., Y_1, Y_0) is approximately a normal random variable.

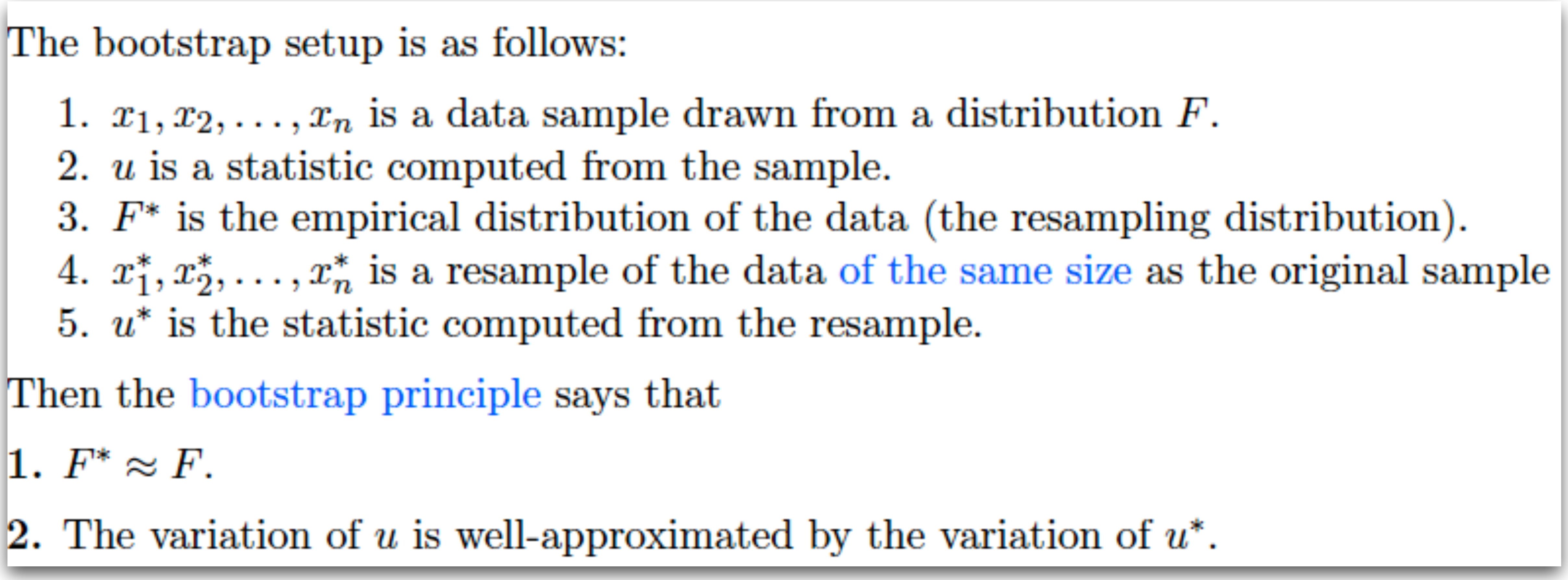
- Y_1, Y_0 do not have to follow normal distributions.

Bootstrap CIs

- If we also care about other summary statistics:
  - median, for example, the median of the revenue
  - %, percent change (lift change)
  - Other percentiles (e.g., quartiles)
  - standard errors (see whether the treatment changes it)
- We cannot use t/z CIs for the statistics other than mean. WHY?
- Bootstrap CIs are flexible, giving not only the CIs of means but also other summary statistics.
- It is a nonparametric method for constructing confidence intervals.
- *It requires a large computational power.*

Bootstrap CIs

- "Perform computations on the data itself to estimate the variation of statistics that are themselves computed from the same data."
- The data is "pulling itself up by its own bootstrap"



The bootstrap setup is as follows:

1. $x_1, x_2, \ldots, x_n$ is a data sample drawn from a distribution F.
2. u is a statistic computed from the sample.
3. F^* is the empirical distribution of the data (the resampling distribution).
4. $x_1^*, x_2^*, \ldots, x_n^*$ is a resample of the data of the same size as the original sample
5. u^* is the statistic computed from the resample.

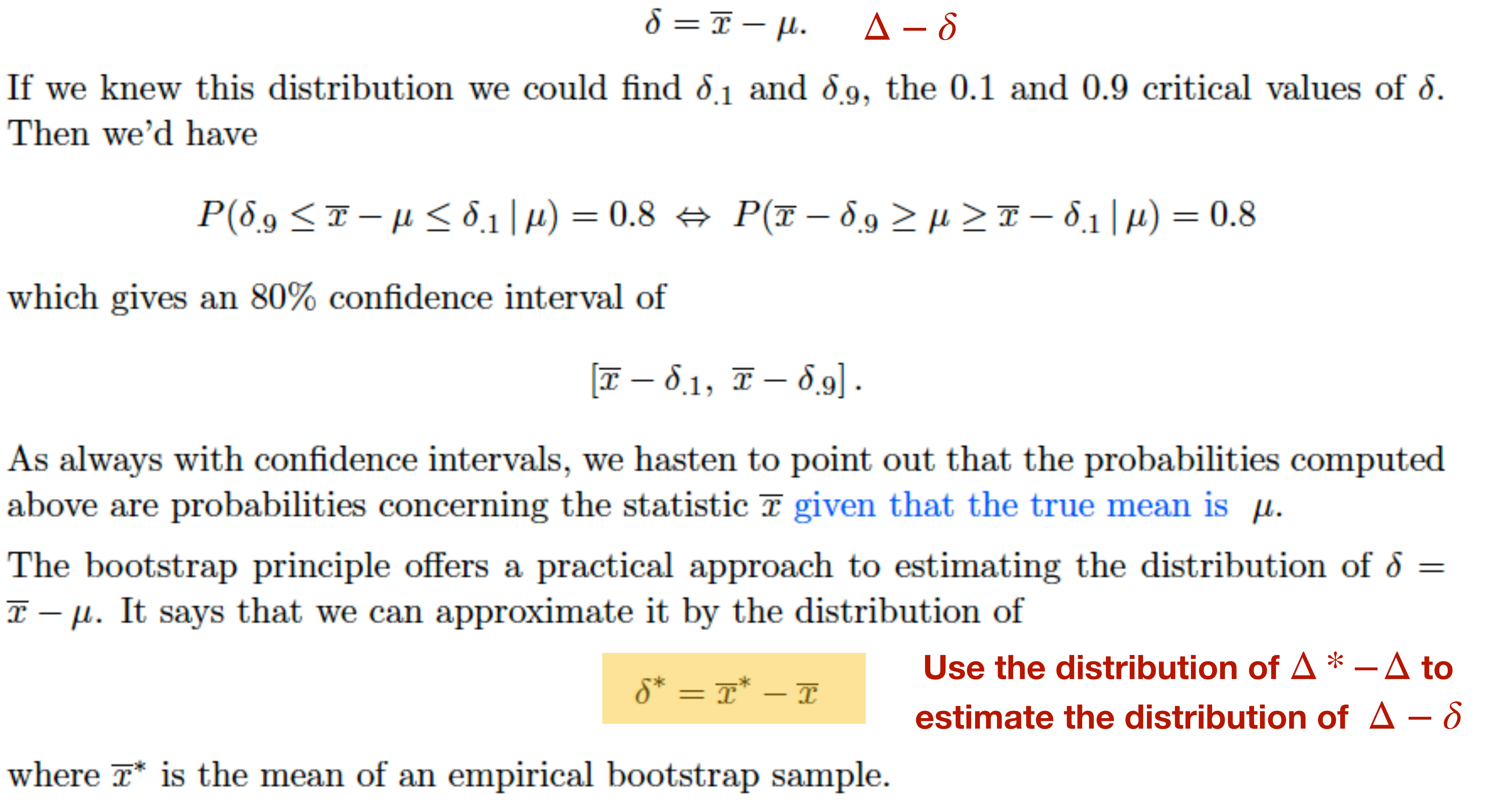
Then the bootstrap principle says that

1. $F^* \approx F$.

2. The variation of u is well-approximated by the variation of u^*.

https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading24.pdf

Bootstrap CIs



$$\delta = \overline{x} - \mu. \qquad \Delta - \delta$$

If we knew this distribution we could find $\delta_{.1}$ and $\delta_{.9}$, the 0.1 and 0.9 critical values of δ. Then we'd have

$$P(\delta_{.9} \le \overline{x} - \mu \le \delta_{.1} \,|\, \mu) = 0.8 \;\Leftrightarrow\; P(\overline{x} - \delta_{.9} \ge \mu \ge \overline{x} - \delta_{.1} \,|\, \mu) = 0.8$$

which gives an 80% confidence interval of

$$[\overline{x} - \delta_{.1},\ \overline{x} - \delta_{.9}]\,.$$

As always with confidence intervals, we hasten to point out that the probabilities computed above are probabilities concerning the statistic $\overline{x}$ given that the true mean is μ.

The bootstrap principle offers a practical approach to estimating the distribution of $\delta = \overline{x} - \mu$. It says that we can approximate it by the distribution of

$$\delta^* = \overline{x}^* - \overline{x}$$

Use the distribution of $\Delta * - \Delta$ to estimate the distribution of $\Delta - \delta$

where $\overline{x}^*$ is the mean of an empirical bootstrap sample.

https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading24.pdf

Class Exercise

- Compute Bootstrap CIs of the *difference and percent change of the means, medians, standard errors, and sums* between control and treatment groups.
- Use the code CI.ipynb

```
pip install bootstrapped
import bootstrapped.bootstrap as bs
import bootstrapped.compare_functions as bs_compare
import bootstrapped.stats_functions as bs_stats
ctrl = np.array(ctrl)
test = np.array(test)
print(bs.bootstrap_ab(
    test,
    ctrl,
    stat_func=bs_stats.mean,
    compare_func=bs_compare.difference,
    alpha=0.05))
```

stat_func=bs_stats.median

compare_func=bs_compare.percent_change,

https://github.com/facebookarchive/bootstrapped

A Common Mistake

Conclude there is no Treatment Effect, when the test does not show *statistically* significant result, $p > 0.05$ (or α).

- It's very likely that the experiment is underpowered to detect the effect size we are seeing.

Statistical Power

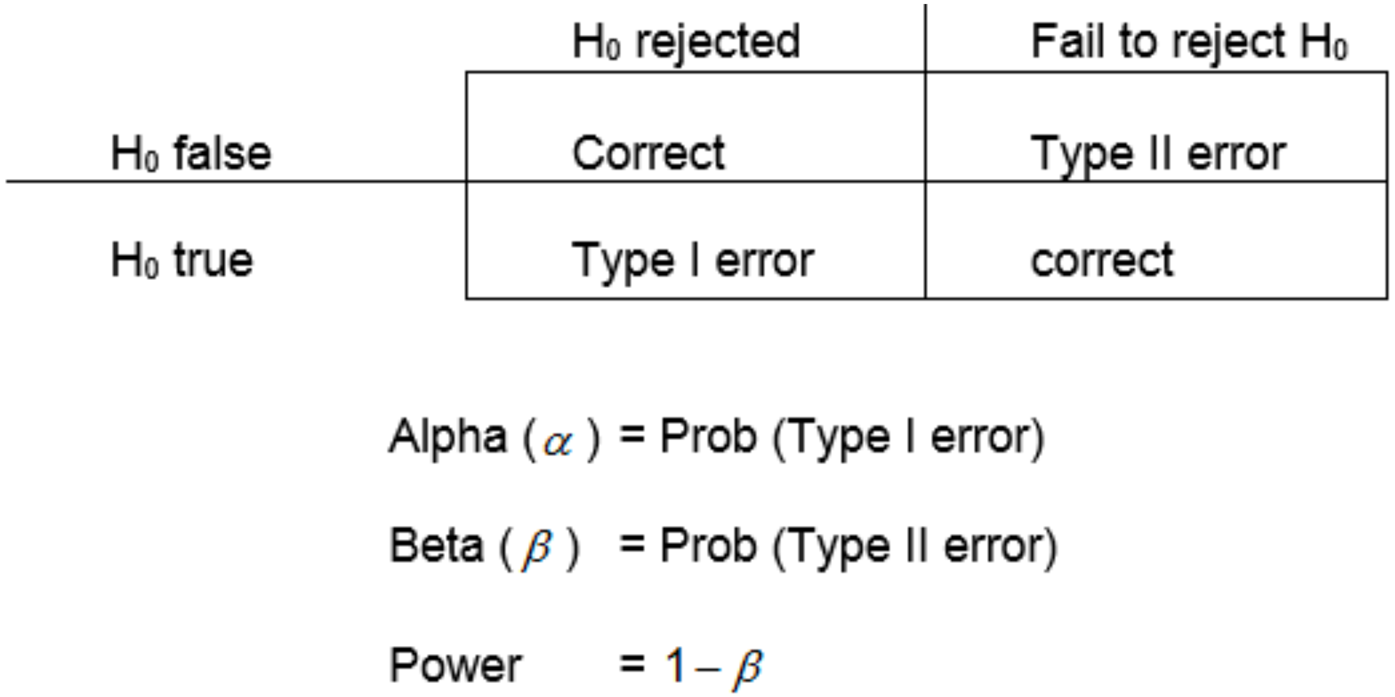
- Rule of Thumb: You need to have a large enough sample size to get at least *80% statistical power, the industry standard,* for your tests to compare the means (identify treatment effects).
- *Power:* the probability of detecting a difference (statistically significant difference) between the variants, when there is a difference $(\delta \neq 0)$
  - When the new feature makes the difference *that you care about*, the tests will conclude a significant difference (reject the Null).
  - Power = 1- Type II error
- *Type II error:* the probability of failing to reject the Null when there are significant differences

Type II Error

- Non-rejection of a false null hypothesis ("false negative”)
    - e.g., “a guilty person is not convicted”

If we draw 100 same-sized samples from the population with a difference between the treatment and control.

- Type II = % of the statistical tests accept the Null Hypothesis (there is no difference, $\delta = 0$)



	H_0 rejected	Fail to reject H_0
H_0 false	Correct	Type II error
H_0 true	Type I error	correct

Alpha (α) = Prob (Type I error)

Beta (β) = Prob (Type II error)

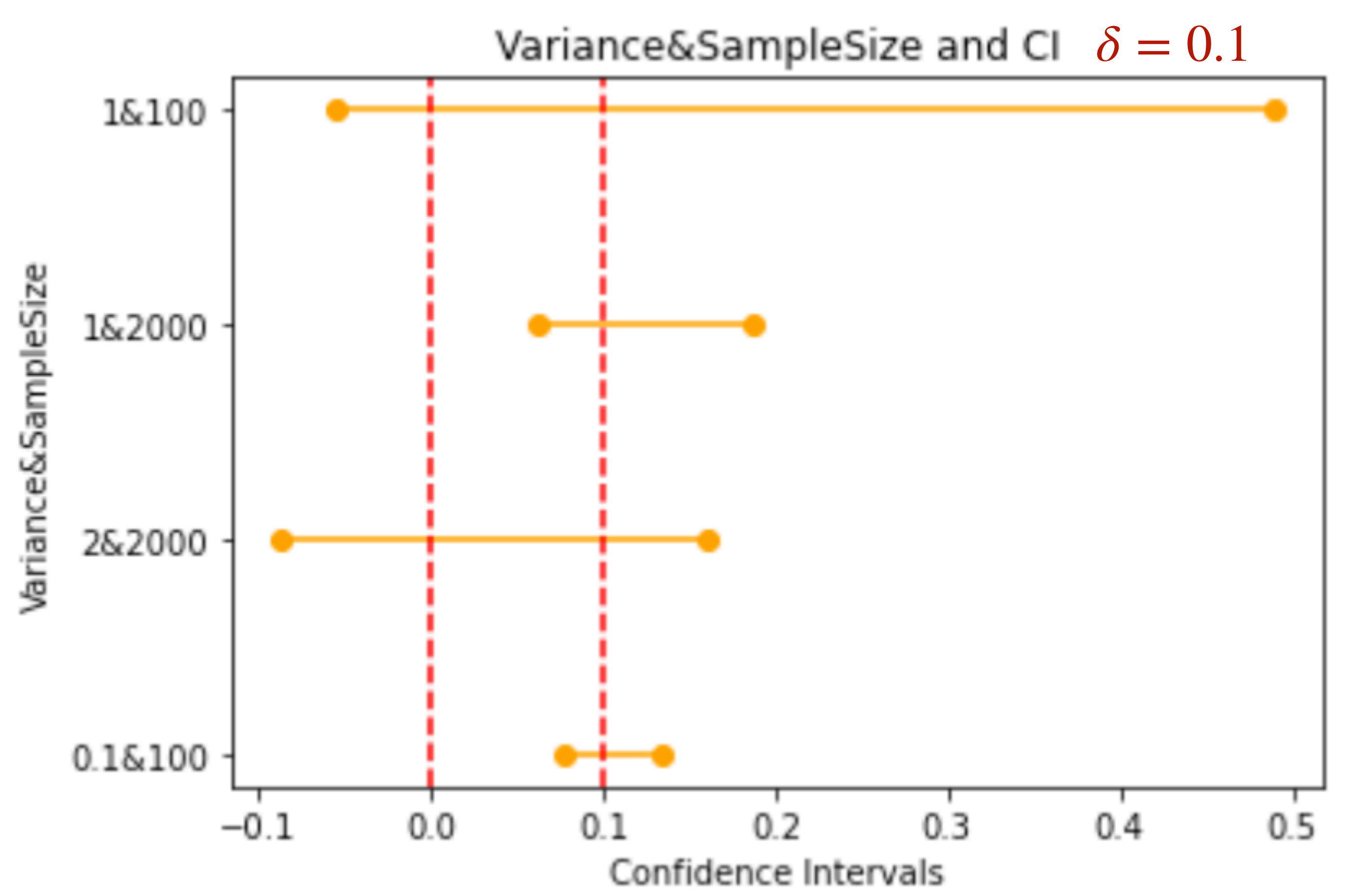
Power = $1 - \beta$

Statistical Power & CI

- Loosely speaking, a tighter CI indicates a larger statistical power
- When population variance of metric, σ, is small and n is large, the CI is tight.

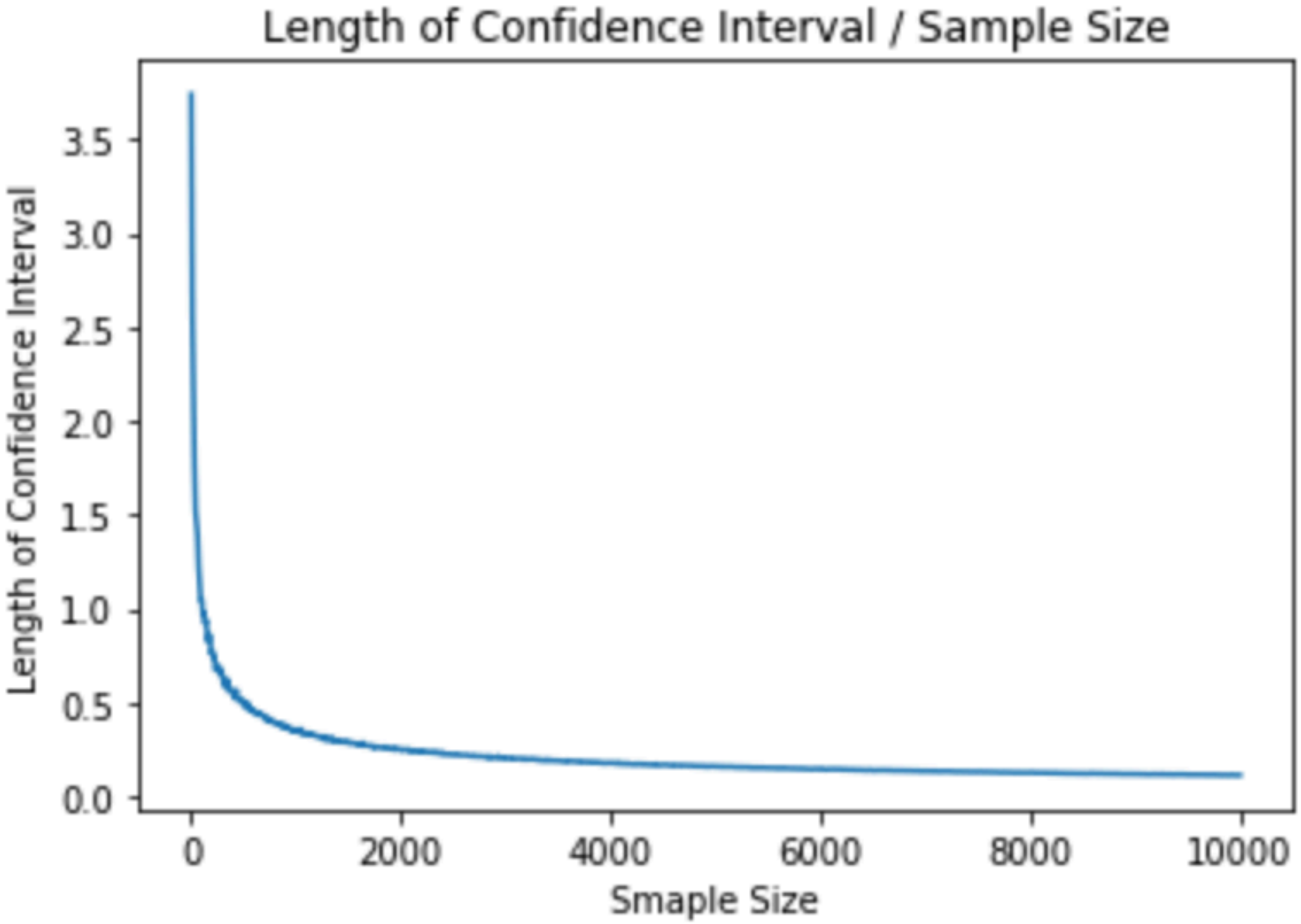
$$[\Delta - se\,.\,t_{\alpha/2}, \Delta + se\,.\,t_{\alpha/2}]$$

$$se = \sigma/\sqrt{n}$$

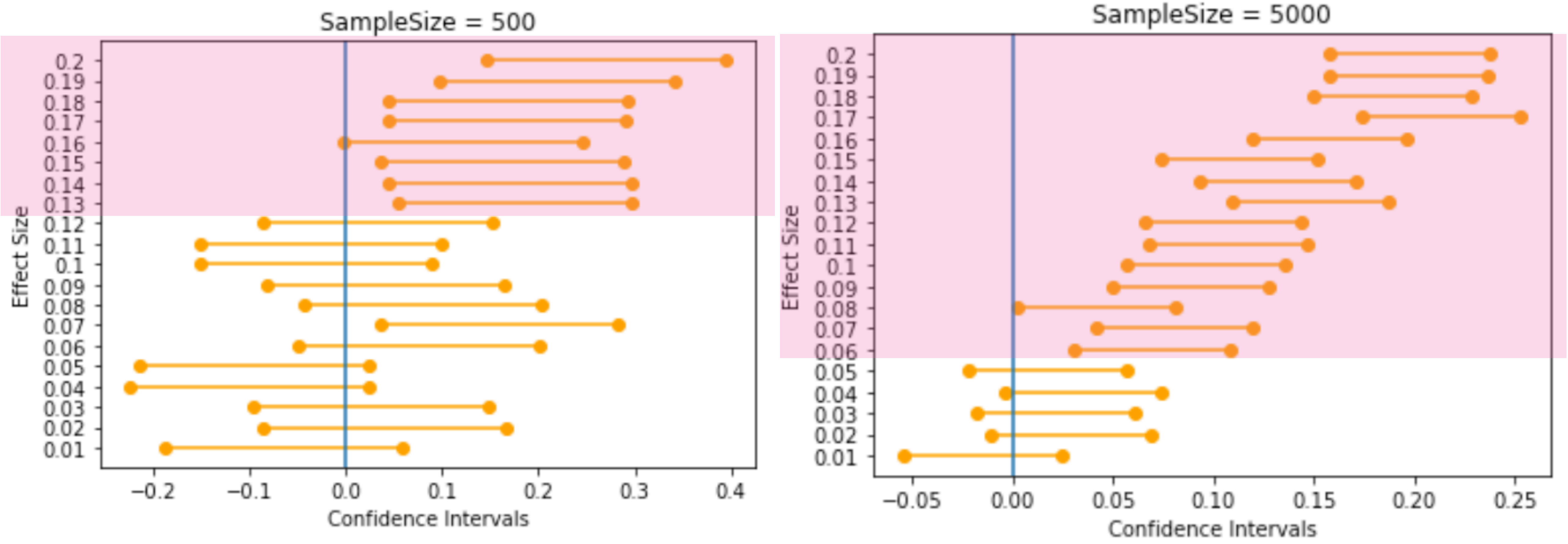


The practical significance is 0.1, meaning that you want H0 would be rejected when there is 0.1 or larger difference in the metric (y).

Sample Size Matters



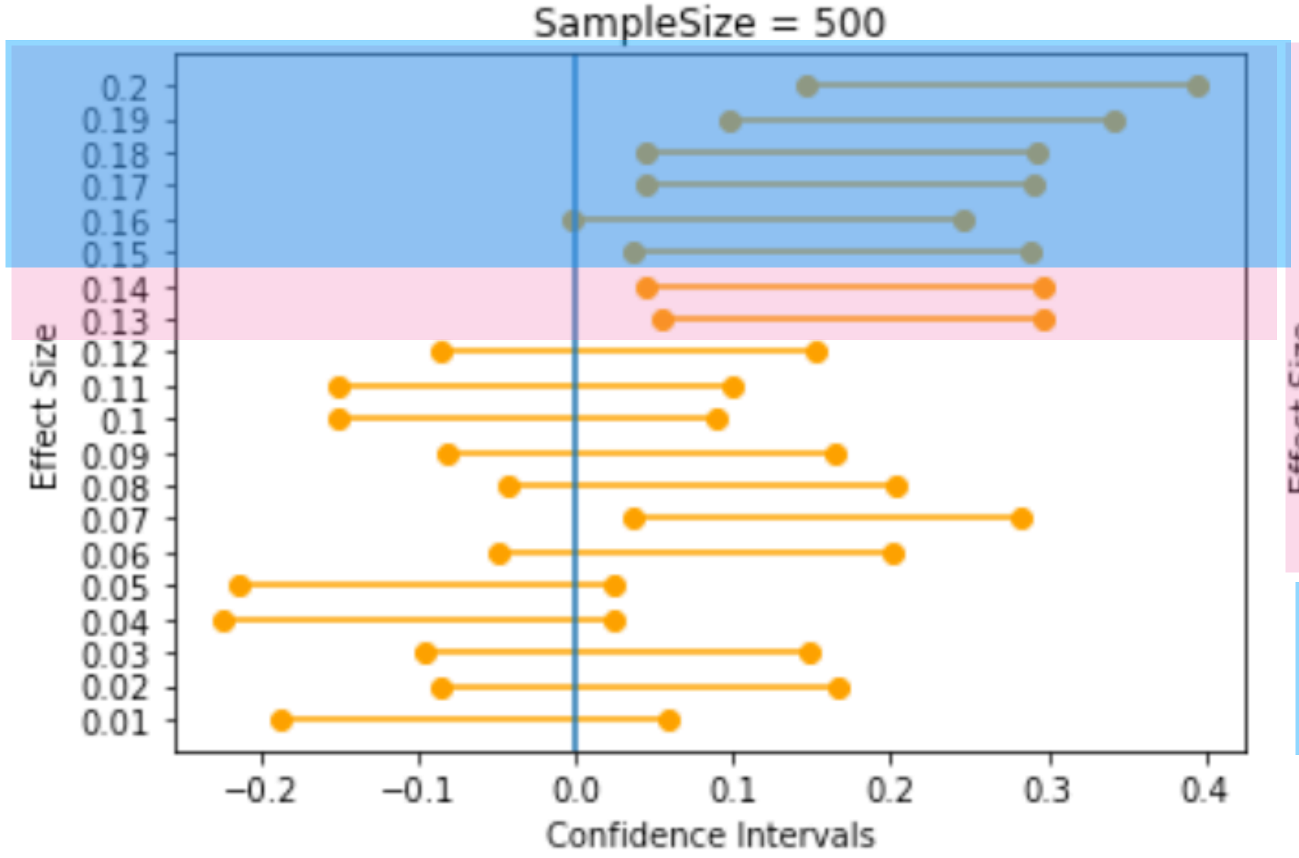
Sample Size and Effect Size Matter

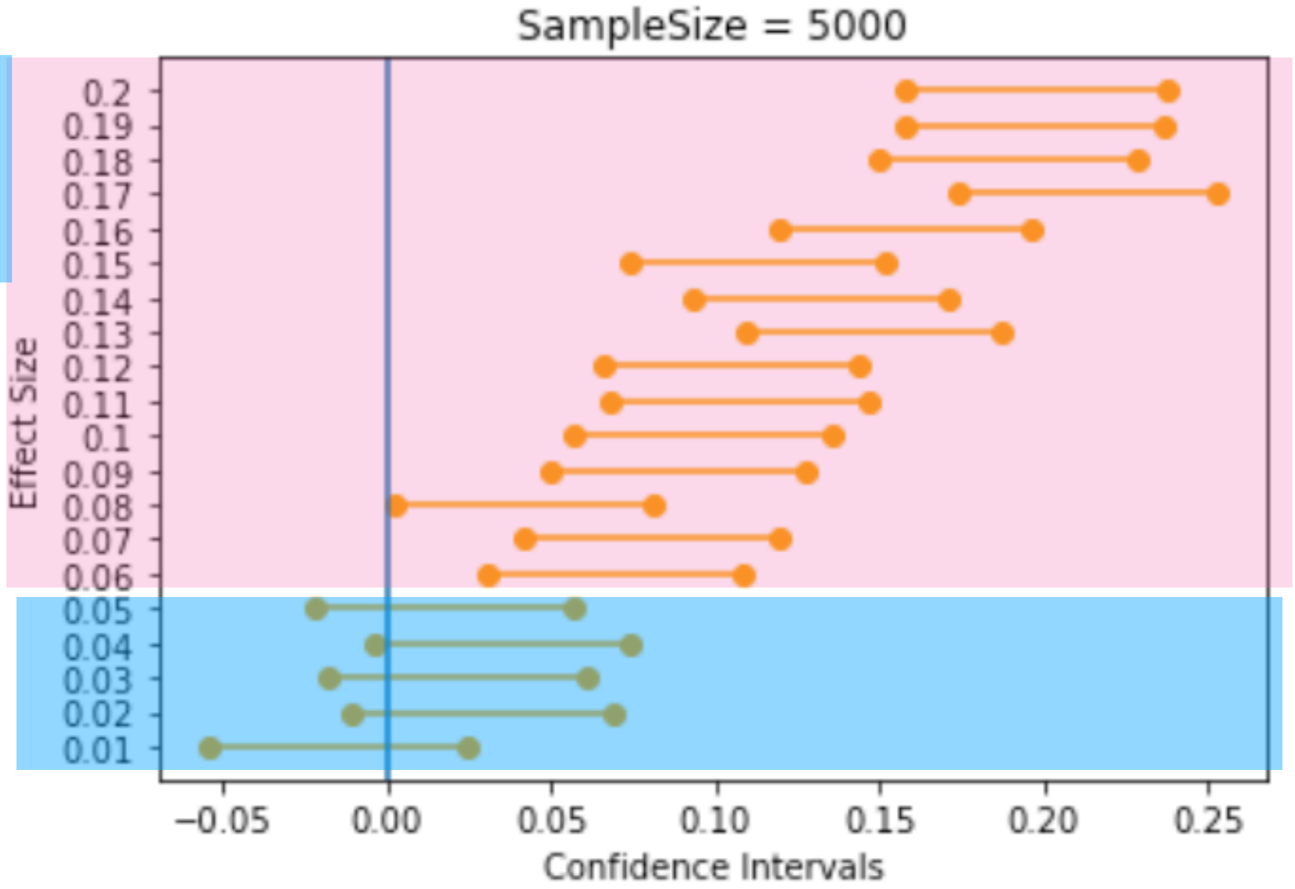


If sample size = 500, how large is the effect size that can be detected?

If sample size = 5000, how large is the effect size that can be detected?

Sample Size and Effect Size Matter

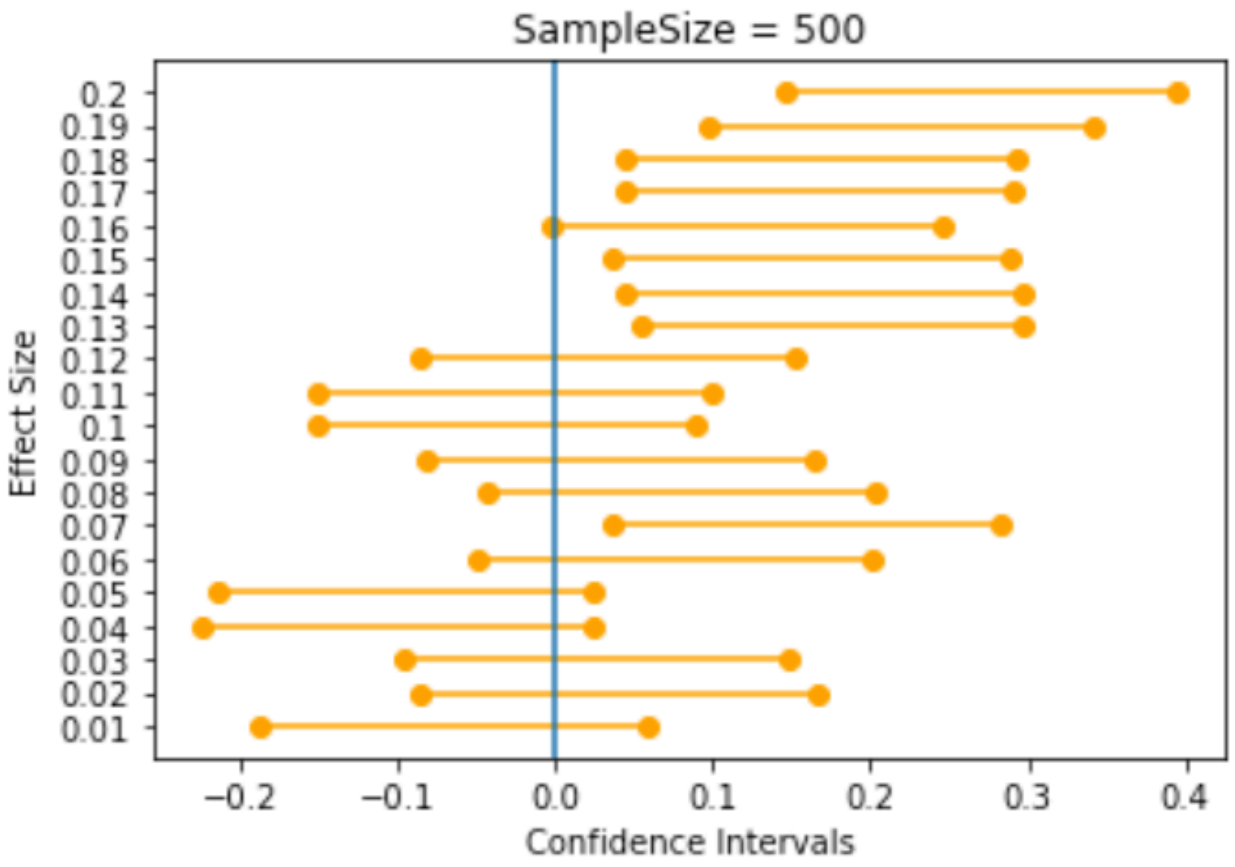


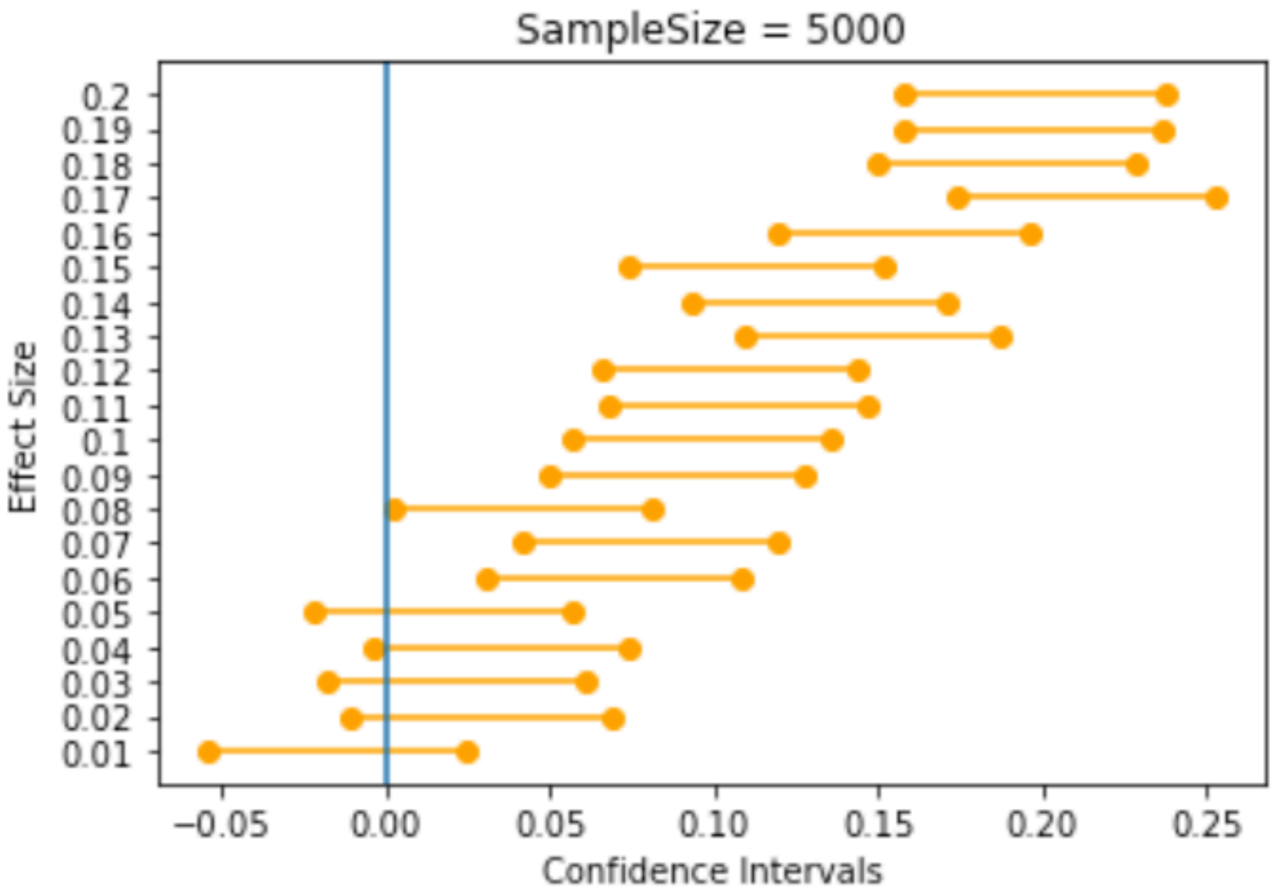


If you care the treatment effects greater than 15%, 500 is large enough.

If you care the treatment effects greater than 1%, 5000 is not large enough.

Class Exercise: Find the Sample Size





1. Practical significance is 13%
2. Practical significance is 6%
3. Practical significance is 1%

Statistical Power & Significance Level
Type I and Type II Errors

$$CI = [\Delta - \hat{se} \, . \, t_{\alpha/2}, \Delta + \hat{se} \, . \, t_{\alpha/2}]$$

- When α is larger, $t_{\alpha/2}$ is smaller and CI is tighter.
- A larger α means a higher Type I error but a smaller chance of missing a real difference and a lower Type II error.
- When you use a smaller α, you may need a larger N.

Class Exercise

- Calculate the statistical power
- Given the population distribution & δ
- Find the statistical power (1- Type II error) for k = 200, and 2000
- Save it as power_analysis.ipynb

```
lift = 1.05  – population mean difference
std=5 – population variance of Y
mean_d = 5*(lift-1)
k = [sample size]
ctrl = np.random.normal(5,std,k)
test = np.random.normal(5*lift,std,k)
```

Try different lifts and std

Class Exercise

```
ci=[]
#Sample Size = k
k = 200
for i in range(20):
    ctrl = np.random.normal(5,std,k)
    test = np.random.normal(5*lift,std,k)
    cm = sms.CompareMeans(sms.DescrStatsW(test), sms.DescrStatsW(ctrl))
    x,y = cm.tconfint_diff(alpha=0.05, alternative='two-sided', 
usevar='pooled')
    ci.append((x,y))

data_dict = {}
data_dict['Number'] = [x for x in range(20)]
data_dict['lower'] = [x[0] for x in ci]
data_dict['upper'] = [x[1] for x in ci]
dataset = pd.DataFrame(data_dict)
print(dataset)
```

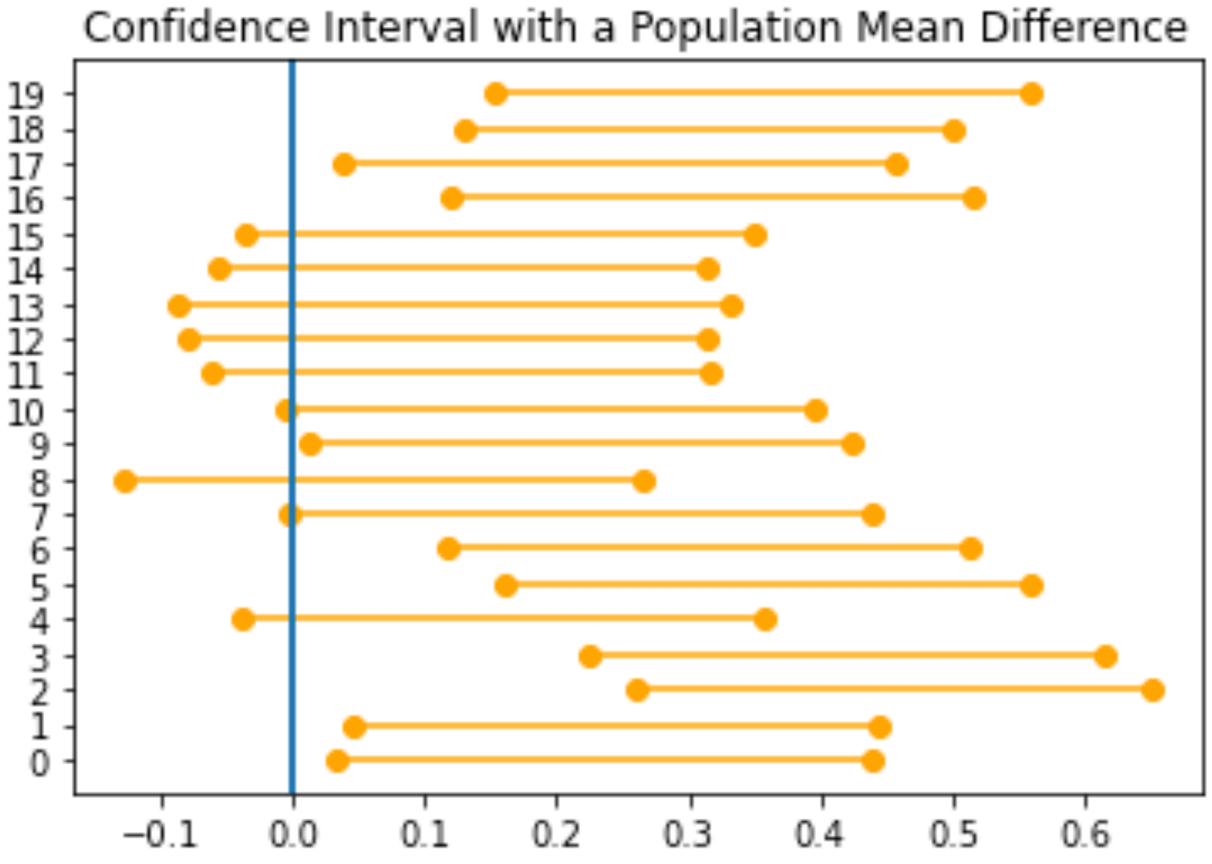
Class Exercise

```
for lower,upper,y in
zip(dataset['lower'],dataset['upper'],range(len(dataset))):
    plt.plot((lower,upper),(y,y),'ro-',color='orange')
    plt.yticks(range(len(dataset)),list(dataset['Number']))
    plt.axvline(x=0)
plt.title('Confidence Interval with a Small Population Mean
Difference')
plt.savefig("CIMeanSmall.png")
```

```
typetwo=sum((x[0]<=0 and x[1]>=0)
for x in ci)/20
power=1-typetwo
```

```
typesign=sum((x[1]<=0) for x in ci)/
20
power_s=1-typesign
```

```
print(typetwo,power,power_s)
```



Statistical Power

- Statistical Power is jointly decided by
  - Sample size (n) (how large the experiment is)
  - Population variance of metric i.e., Var(Y)
  - Effect size (how large the treatment effects that you care about, the practical significance; assume the population mean difference is this large)
  - Significance level α

Green: You can choose

Orange: More difficult to manipulate

How will you choose the sample size for your experiments?

Power Analysis

How to Decide the Sample Size

- To achieve 80% statistical power and 5% significance level, assuming Treatment and Control are of equal size, the same size for each variant is:

$$n = 16\sigma^2/\delta^2$$

σ^2 = population variance of the metric (Y)

δ = the effect size that practical significance requires (the difference between treatment and control)

How would you know δ before the experiment?

- We don’t know the size of the treatment effects before the experiment.
- But we know the size of δ that would matter in practice, with practical significance.
  - The current purchase rate is 5%, and we care about at least 5% increase.
  - δ = 5% * 5% = 0.25%
  - To be able to detect the difference, if population δ is this large.

Class Exercise - Find Sample Size

Let’s take an e-commerce site with 5% of users who visit during the experiment period ending up making a purchase. The conversion event is a Bernoulli trial with $p = 0.05$.

The variance, $\sigma^2 = p\,(1-p) = 0.05*0.95 = 0.0475$. You care about at least 5% increase on the purchase rate resulted from a new feature. Therefore, the sample size for each variant will be at least:

$$n = 16\sigma^2/\delta^2 = 16 \times 0.0475/(0.05 * 0.05)^2$$

=121,600 users

Class Exercise - Find Sample Size

You made a change to the checkout process, and you care about at least 5% change to the purchase rate. Purchase rate is 5% site visitors. Therefore, the sample size of site visitors for each variant will be at least

$$\sigma^2 = p(1-p) = 0.05 \times 0.95 = 0.0475.$$

$$n = 16\sigma^2/\delta^2 = 16 \times 0.0475/(0.05 \times 0.05)^2 = \textbf{121,600}$$

users

Class Exercise - Find Sample Size

You made a change to the checkout process, and you cared about at least 5% change to the purchase rate. You only include users who started the checkout process in your experiment. Assume that 10% of users initiate checkout, so that given the 5% purchase rate, half of them complete checkout. Therefore, the sample size for each variant will be at least:

$$p(purchase \mid checkout) = 5\,\% / 10\,\% = 50\,\% \;,$$

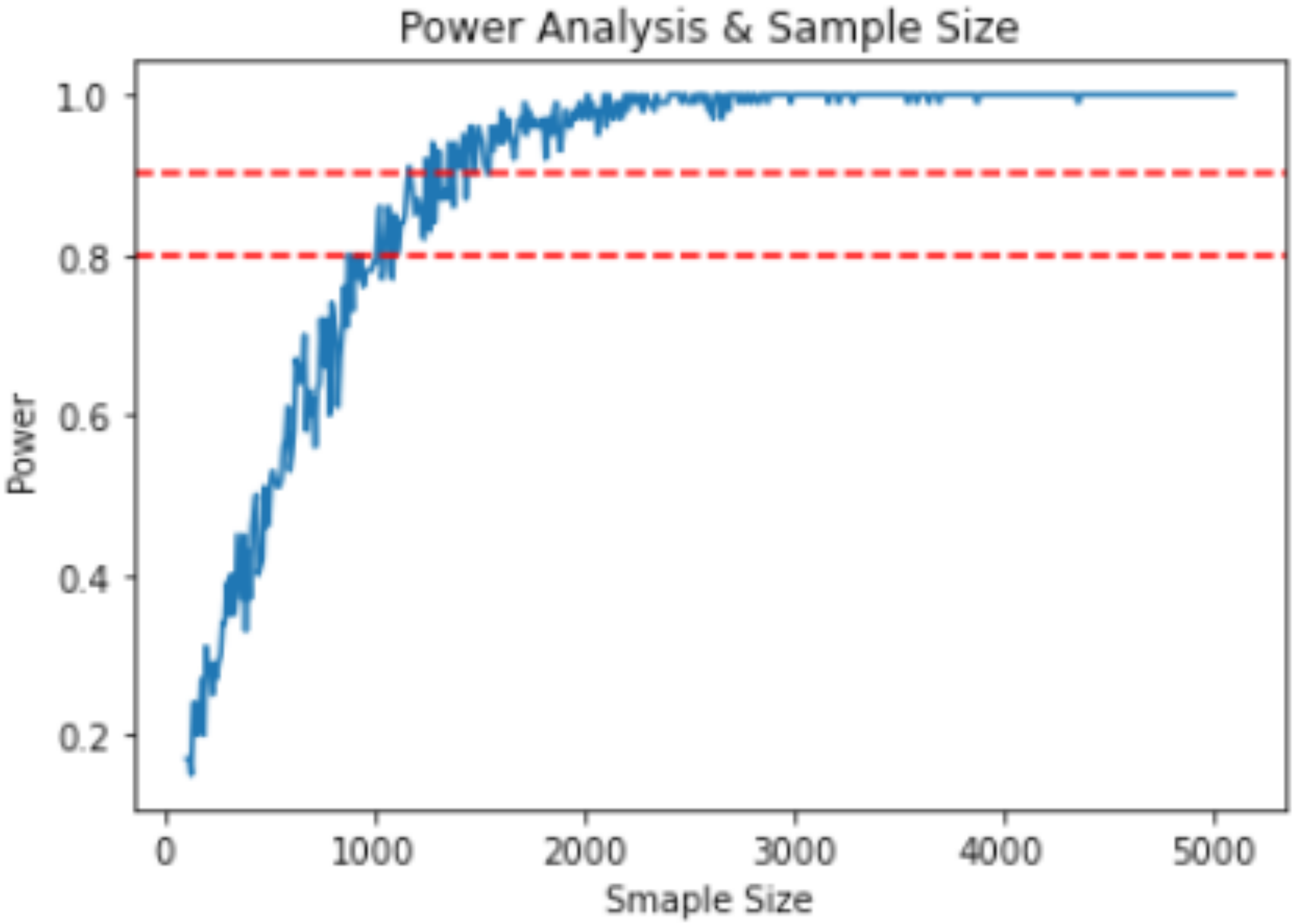
$$\sigma^2 = p(1-p) = 0.5 \times 0.5 = 0.25.$$

$$n = 16\sigma^2/\delta^2 = 16 \times 0.25/(0.5 \times 0.05)^2 = 6{,}400$$

A much smaller sample size !

Class Exercise - Find Sample Size

- Based on the code power_analysis.ipynb
- Find the **sample size** for the statistical power >= 0.8
- How about larger σ, smaller δ, and a larger required statistical power



Class Exercise - Find Sample Size

```
power=[]
n=100
m=500
for s in range(m):
    k =100+10*(s+1)
    ci=[]
    for i in range(n):
        ctrl = np.random.normal(5,std,k)
        test = np.random.normal(5*lift,std,k)
        cm = sms.CompareMeans(sms.DescrStatsW(test), 
sms.DescrStatsW(ctrl))
        a,b = cm.tconfint_diff(alpha=0.05, alternative='two-sided', 
usevar='pooled')
        ci.append((a,b))
    t2=sum((x[0]<=0 and x[1]>=0) for x in ci)/n
    pw=1-t2
    power.append((k,pw))
```

Class Exercise - Find Sample Size

```
l_y=[x[1] for x in power]
s_x=[x[0] for x in power]
plt.plot(s_x,l_y)
plt.title('Power Analysis' + ' & Sample Size')
plt.xlabel('Smaple Size')
plt.ylabel('Power')
plt.axhline(y=0.8, color='r', linestyle='--')
plt.axhline(y=0.9, color='r', linestyle=‘--')


index_=[power.index(x) for x in power if x[1] >= 0.7999
and x[1] <= 0.81]

size_exp=[power[i][0] for i in index_]

print(size_exp)
```

You made a change to the checkout process, and you cared about at least 5% change to the purchase rate. You only include users who started the checkout process in your experiment. Assume that 10% of users initiate checkout, so that given the 5% purchase rate, half of them complete checkout. Therefore, the sample size for each variant will be at least:

Use the simulation approach to estimate the sample size

Class Exercise - Find Sample Size

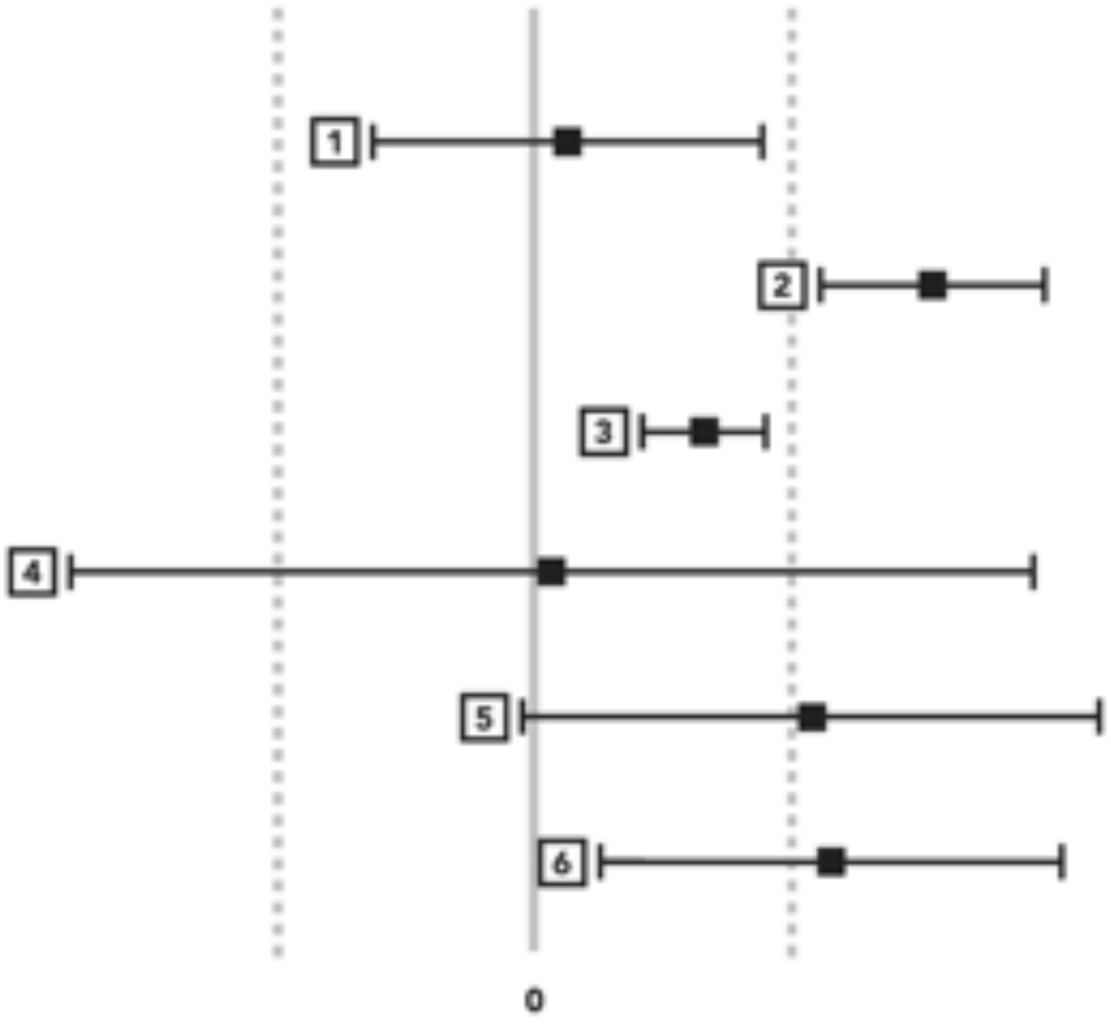
- Use the formula to calculate the sample size.

```
from statsmodels.stats.power import
zt_ind_solve_power

size_80=zt_ind_solve_power(effect_size=(mea
n*(lift-1)/std), alpha=0.05, power=0.8,
ratio=1,alternative="two-sided")

size_95=zt_ind_solve_power(effect_size=(mea
n*(lift-1)/std), alpha=0.05, power=0.95,
ratio=1,alternative="two-sided")
```

Would you launch the new feature?

1. The effect is significantly less than 0.5.

DON'T LAUNCH!

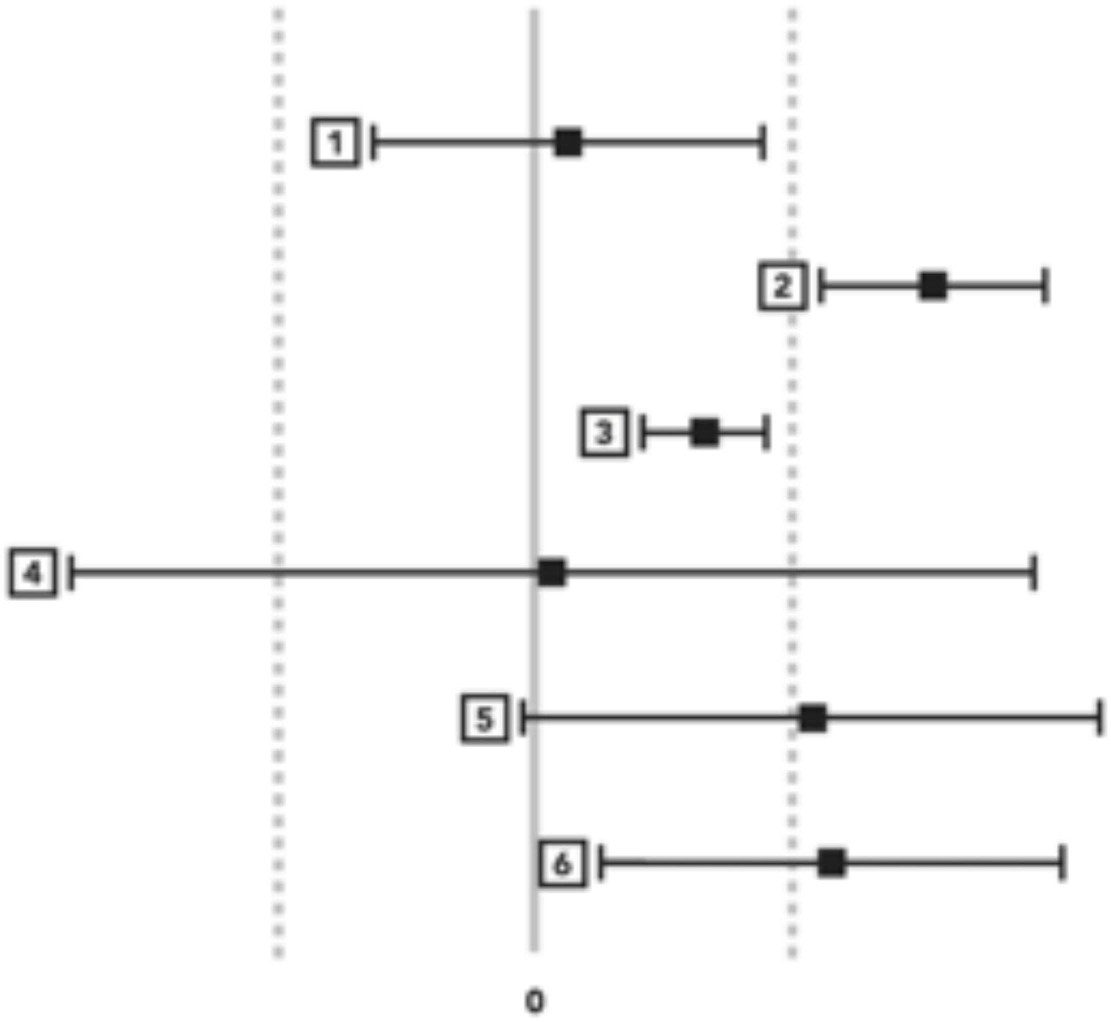


0.5
practical
significance

Would you launch the new feature?

2. The effect is significantly more than 0.5.

LAUNCH!

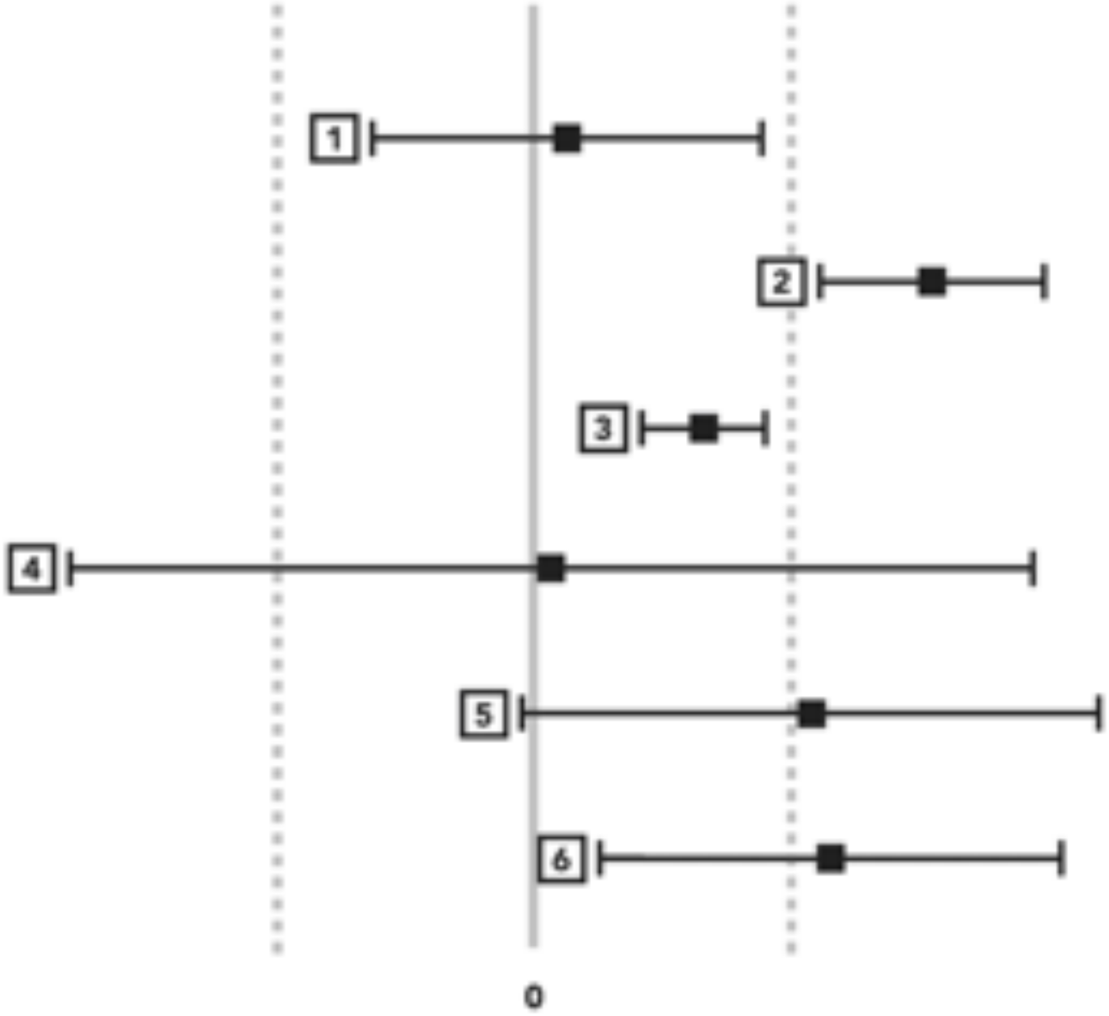


0.5 practical significance

Would you launch the new feature?

3. The effect is significantly more than 0, but less than 0.5.

DON’T LAUNCH!



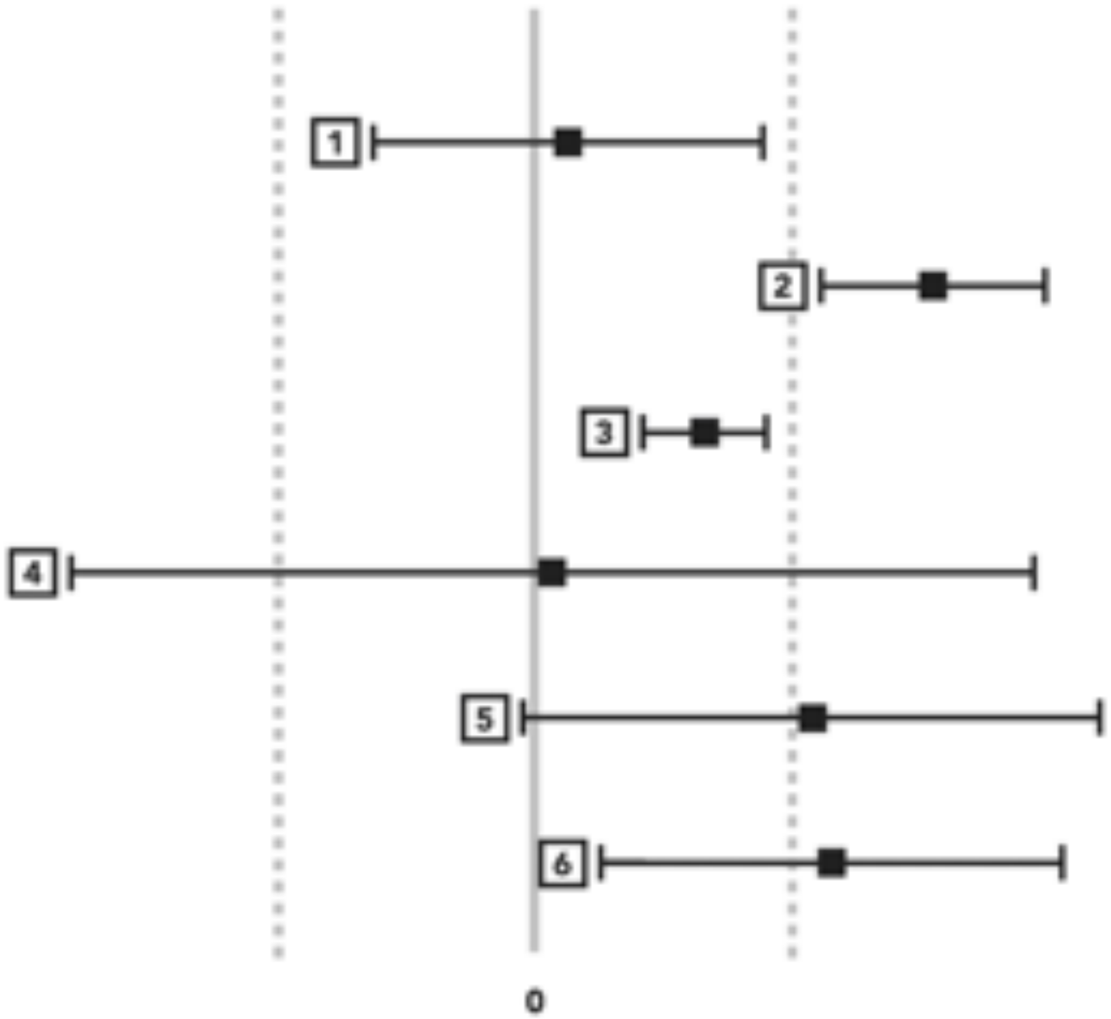
0.5 practical significance

Would you launch the new feature?

4. The effect is not significant and mean < 0.5.

Uncertain!

The tests may not have enough statistical power.



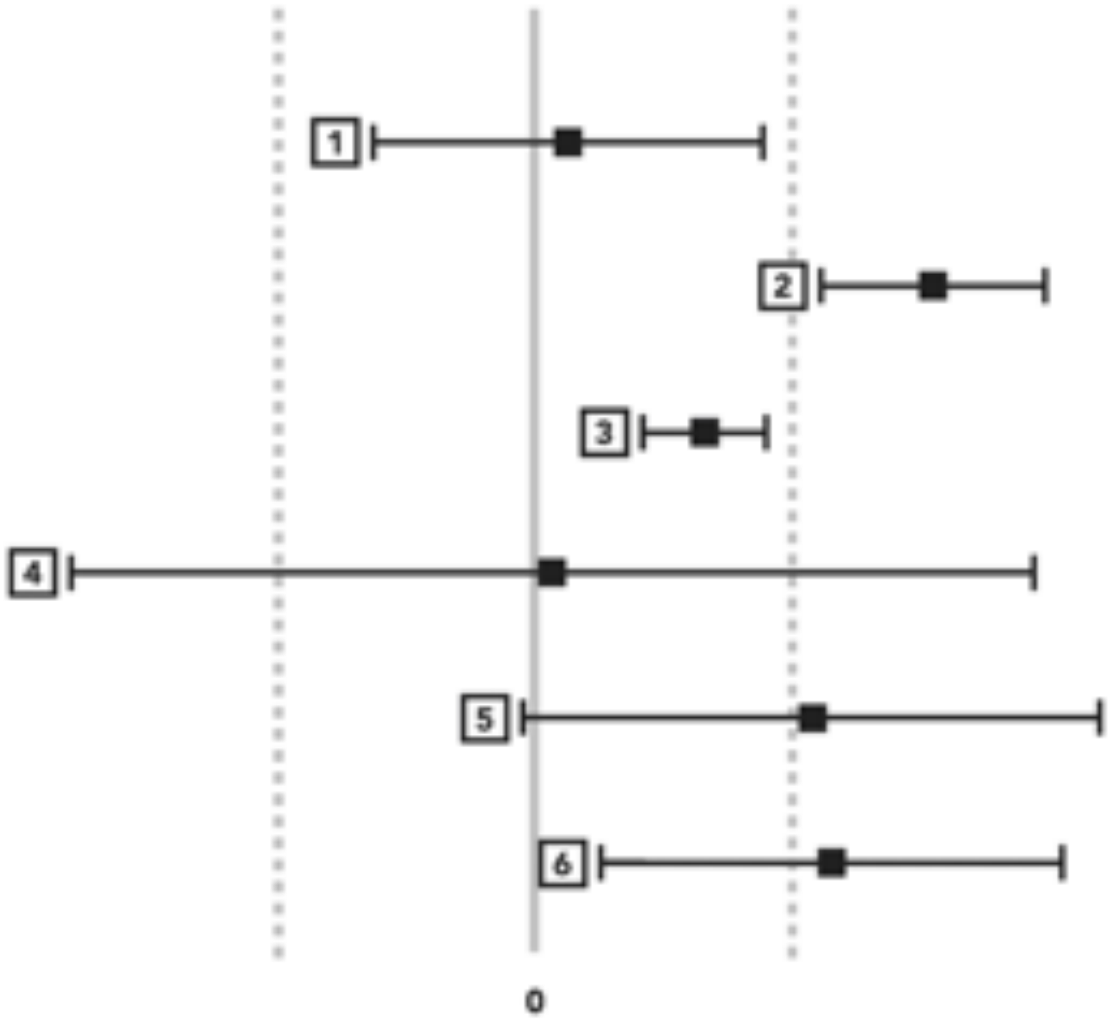
0.5 practical significance

Would you launch the new feature?

5. The effect is not significant and mean > 0.5.

Uncertain!

The tests may not have enough statistical power.

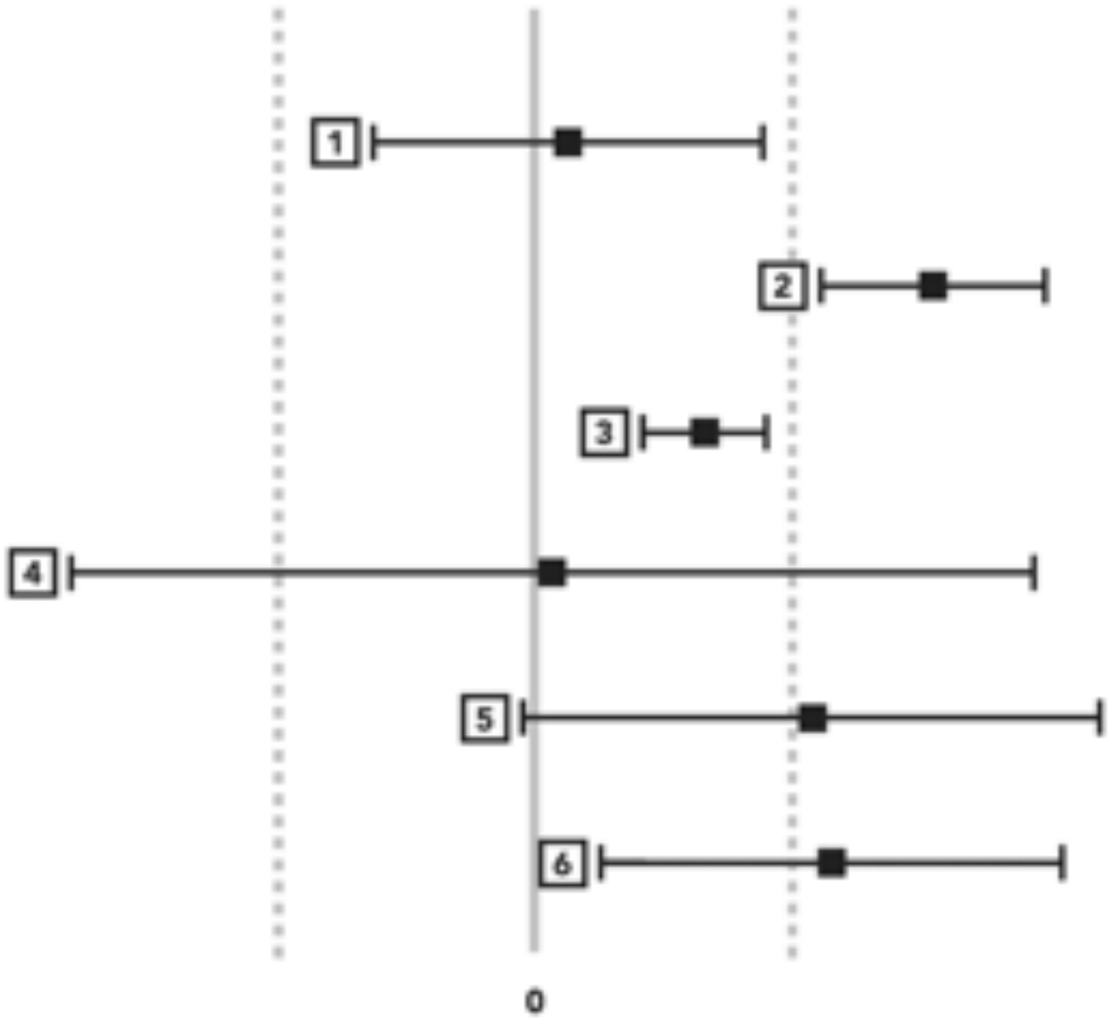


0.5 practical significance

Would you launch the new feature?

6. The effect is significantly larger than 0, but not significantly larger than 0.5.

It's ok to launch.
If it is possible, we should further increase the sample size.



0.5 practical significance

What will you do?

You are not allowed to get a large enough sample for your experiment, which is underpowered.

- The new feature has big uncertainties and may hurt user experience.
- In the early stage, there are not enough users.

How to Increase Statistical Power

- Increase the duration of the experiment.
  - The early users are different from the later ones.
  - In the later stage, users can be repeated users.
- Fisher’s Meta-analysis
  - Combining results from multiple small experiments
    - Instead of running one large experiment, run multiple small experiments.
    - To get verify the surprising results.

Fisher’s Meta-analysis

- *Intuition:* You replicate your experiment on another group of users (or with another treatment assignment), independent from the users participating in the old experiment (or the old treatment assignment). If the two experiments both show significant results, you will be more confident about your results.

Fisher’s Meta-analysis

- Combine the results from several independent tests bearing upon the same overall hypothesis (Ho).
- A technique for data fusion or “meta analysis" (analysis of analyses).
- Ho: All the separate null hypotheses are true (e.g., $\delta = 0$).
- H1: at least one of the separate *alternative* hypotheses is true.

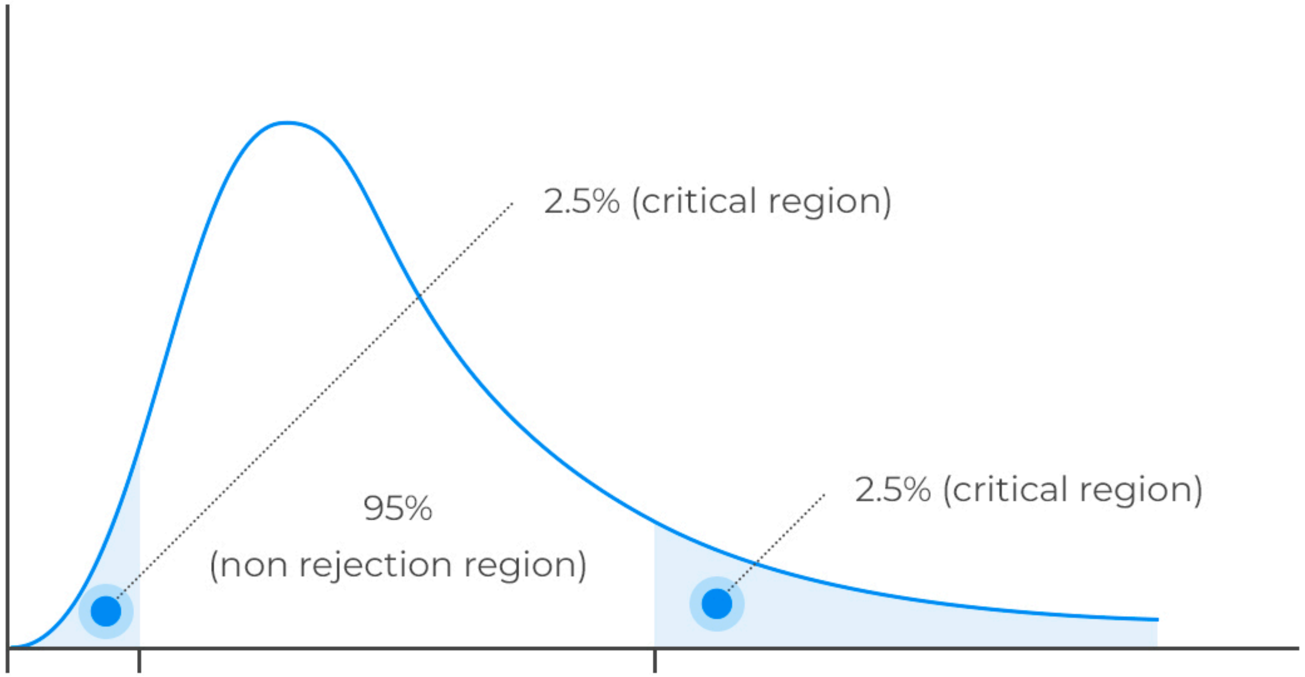
Fisher's Meta-analysis

- Combine p-values from k independent tests into one test statistic:

$$X^2_{2k} = -2\sum_{1}^{k} ln(p_i) \sim \chi^2(2k)$$

- p_i is for *i-th* experiment
- Compute the p-value for this new statistics.
- When p is small, reject the NULL Hypothesis.

Two-tailed Chi-Square test (5% significance)

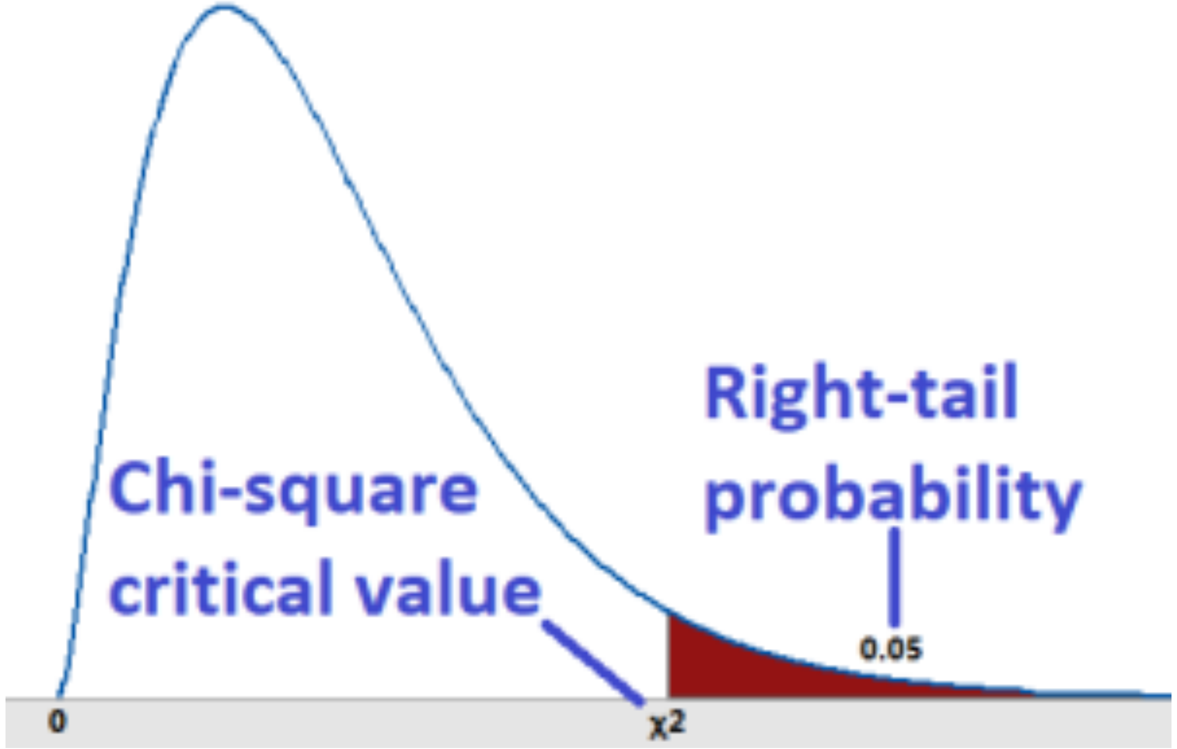


Fisher’s Meta-analysis

- If all k Null are true, this test statistic follows a chi-square distribution:
    - Combine p-values from k independent tests into one test statistic:

$$X^2_{2k} = -2\sum_{1}^{k} ln(p_i) \sim \chi^2(2k)$$

    - p_i is for *i-th* experiment
- Compute the p-value for this new statistic.
- When p-value is small, reject the NULL Hypothesis.



Class Exercise - Fisher’s Meta-analysis

pi	ln(pi)
0.04	-3.2188758
0.07	-2.65926
0.5	-0.6931472
0.08	-2.5257286
0.3	-1.2039728
0.05	-2.9957323
0.2	-1.6094379
0.03	-3.5065579
0.05	-2.9957323
0.06	-2.8134107
SUM	-24.221856
Chi-square	48.4437111
k=2*10	20
new p value	0.000368469