Do problem 1, 2, 3, 9, 10

1. Appropriate task to perform A/B test

Which of the following use cases can you reliably conduct an A/B test? (True/False)

1.1. Frontend person wants to change color of the 'Go' button on a search bar. Will it increase conversion rate?

1.2. The data team created four versions of machine learning model for product recommendations to new users of an app. Which one is the best?

1.3. Two managers from different factions have Layout A and Layout B for a physical convenience store. Which one should we use?

1.4. Mr. Rabbito thinks offline stores are the best channel to distribute our products, whereas Ms. Rakko thinks online websites are the way to go. Who is right?

1.5. Your boss wants to add a premium version to your freemium service. Is it a good idea?

1.6. The backend team came up with a new setup that they think will speed up the website load time. Should we implement this change?

1.7. Kuruma Inc., a car dealer, wants to change the banner on their homepage to see if it will attract more repeated customers. Average time between purchase of the car company is 5 years. How do you know if the banner change has an effect?

1.8. Your company undergoes a total revamp of its corporate identity. Is it the right call?

1.9. Elastic ninja at your company wants to show 15 products on the first page of search results instead of 20 products. Should you allow them?

1.10. Marketing person wants to know who respond better to our ads campaigns between iOS users and Android users. How to tell?

Solution:

1.1. Frontend person wants to change color of the 'Go' button on a search bar. Will it increase conversion rate?

: True

1.2. The data team created four versions of machine learning model for product recommendations to new users of an app. Which one is the best?

: True

1.3. Two managers from different factions have Layout A and Layout B for a physical convenience store. Which one should we use?

: False, it is extremely difficult to isolate confounding factors in a physical store setting (e.g., customer demographics, weather, or store location).

1.4. Mr. Rabbito thinks offline stores are the best channel to distribute our products, whereas Ms. Rakko thinks online websites are the way to go. Who is right?

: False, it fundamentally different channels with multiple variables.

1.5. Your boss wants to add a premium version to your freemium service. Is it a good idea?

: False, behavioral change and multiple variables.

1.6. The backend team came up with a new setup that they think will speed up the website load time. Should we implement this change?

: True

1.7. Kuruma Inc., a car dealer, wants to change the banner on their homepage to see if it will attract more repeated customers. Average time between purchase of the car company is 5 years. How do you know if the banner change has an effect?

: False, too long feedback loop.

1.8. Your company undergoes a total revamp of its corporate identity. Is it the right call?

: False, holistic changes affecting multiple variables at once.

1.9. Elastic ninja at your company wants to show 15 products on the first page of search results instead of 20 products. Should you allow them?

: True

1.10. Marketing person wants to know who respond better to our ads campaigns between iOS users and Android users. How to tell?

: True

2. Choose the method

What are the metrics you should use for the following A/B tests? Assume that the granularities are: page views and unique visitors.

2.1. Which button colors will make customers find it more easily? clicks / __

2.2. Which sets of products on a landing page will make customers more likely to buy? purchases / __

2.3. Which types of promotion coupons will be more effective? purchases / __

2.4. Which website layouts will attract more customers to click on sign up button? clicks / __

Solution:

2.1. Which button colors will make customers find it more easily?

: clicks / page views

2.2. Which sets of products on a landing page will make customers more likely to buy?

: purchases / unique visitors

2.3. Which types of promotion coupons will be more effective?

: purchases / unique visitors

2.4. Which website layouts will attract more customers to click on sign up button?

: clicks / page views

3. Choose the period

Based on the transaction table below,

3.1. what are the event-based conversion rate of 2020-11?

3.2. what are cohort-based conversion rate of 2020-11?

Assume 7-day attribution period. Conversion rate is calculated by purchases / unique users.

date	user	event
2020-11-01	A	visit
2020-11-01	A	purchase
2020-11-05	B	visit
2020-11-13	B	visit
2020-11-30	C	visit
2020-12-05	C	purchase

Solution:

3.1. what are the event-based conversion rate of 2020-11?

$$\text{Conversion rate of November} = \frac{\text{conversions within November}}{\text{number of users that visited in November}}$$
$$= \frac{A\ purchase}{UniqueUsers\ A, B, C} = \frac{1}{3} = 33.33\%$$

3.2. what are cohort-based conversion rate of 2020-11? Assume 7-day attribution period.

$$\text{Conversion rate of November Cohort}$$
$$= \frac{\text{conversions within November}+7}{\text{number of users that visited in November}} = \frac{A, C\ purchase}{UniqueUsers\ A, B, C} = \frac{2}{3}$$
$$= 66.67\%$$

4. Familarity with the incoming data

Give 3 examples of values that are usually distributed in the following manner (do not use examples from class):

4.1. Bernoulli/Binomial distributions: __, __, __

4.2. Normal/Student t's distribution: __, __, __

4.3. Exponential distribution: __, __, __

4.4. Poisson distribution: __, __, __

5. Design experiments

Which variables should you control for in an A/B test of the following cases?

5.1. We want to test if SMOKING -> CANCER (Smoking causes cancer) and we know that AGE -> SMOKING and AGE -> CANCER. We should control for __

5.2. We want to test if GUN OWNERSHIP -> CRIMES and we know that GUN OWNERSHIP -> GUN SALES and CRIMES -> GUN SALES. We should control for __

5.3. We want to test if CROP BURNING -> LUNG DISEASES and we know that CROP BURNING -> PM2.5 and PM2.5 -> LUNG DISEASES. We should control for __

6. LLN

The Law of Large Numbers (LLN) says that sample mean will converge to expectation as sample size grows. Assuming that this is true, prove that sample variance will converge to variance as sample size grows.

7. P-value

What is p-value? (Choose one or more)

7.1. Assuming that the null hypothesis is true, what is the probability of observing the current or more extreme data.

7.2. Based on the observed data, what is the probability of the null hypothesis being true.

7.3. Based on the observed data, what is the probability of the null hypothesis being false.

7.4. Assuming that our hypothesis is true, what is the chance that we reject the null hypothesis.

8. False positive

If we conduct a frequentist statistical test at 5% significance level repeatedly for 4,000 times, how many times can we expect to have statistically significant results even if group A and B are exactly the same?

9. Hamster Inc. and His Color Package

Hamster Inc. once again wants to test the conversion rates between package colors of its sunflower seeds; this time it is Red Package vs Gold Package. The Red Package is the existing group with average conversion rate of 11%. If they think the minimum detectable effect is 1% and want to make a 80/20 control/test split, how many unique users should see each package color before we decide which one performs better? Assume that they are testing at significance level of 15%. Show your work.

In [1]:
```
# significance level
import scipy.stats
scipy.stats.norm.ppf(0.85)
```

Out[1]:
```
1.0364333894937898
```

the problem did not give power, so we use power=0.5 same as lecture slide

Q9.

$$n = \frac{m+1}{m}\left(\frac{(z_\alpha + z_\beta)\sigma}{MDE}\right)^2$$

} Cntrl / Tst Split: 80/20

$\therefore m = 4$

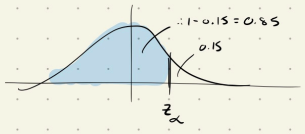
} MDE = 1% = 0.01

$$= \frac{5}{4} \times \frac{1}{0.01^2} \times ((z_\alpha + z_\beta)\sigma)^2$$

} Power = 0.5 = 50%

$\therefore z_\beta = 0$

} Significant = 15% = 0.15



scipy.stats.norm.ppf(0.85) = 1.03643

$\therefore z_\alpha = 1.03643$

$$= \frac{5}{4} \times \frac{1}{0.01^2} \times (1.03643 + 0)^2 \sigma^2$$

} Conversion Rate: 11%

└ Bernoulli: $\therefore p = 0.11$

$\sigma^2 = p \cdot (1-p)$: variance of Bernoulli

$= 0.11\,(1-0.11)$

$\therefore \sigma^2 = 0.0979$

$$= \frac{5}{4} \times \frac{1}{0.01^2} \times 1.03643^2 \times 0.0979$$

Cntrl - Red Tst - Gold

80 : 20

$\therefore n = 1314.5365 \approx 1315$

4 × 1315 : 1315 ∴ 5260 unique users for cntrl(red) group.

5260 : 1315 1315 unique users for test(gold) group.

Conclusion:

$5,260$ unique users for control(red) group.
$1,315$ unique users for test(gold) group.

10. Hamster Inc. and His A/B Testing Experiment

Let us say Hamster Inc. ran the experiment and got the following results.

10.1. At significance level of 7%, which variation should be chosen to run at 100% traffic? Show your work.

10.2. What are the confidence intervals at 7% significance of conversion rates for Red and Gold? Show your work.

campaign_id	clicks	conv_cnt	conv_per
Red	59504	5901	0.099170
Gold	58944	6012	0.101995

Solution:

10.1. At significance level of 7%, which variation should be chosen to run at 100% traffic? Show your work.

Two-Proportion Z-Test

ref. https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/z-test/

- The groups must be independent.
- The data must be selected randomly and independently from a homogenous population
- The population should be at least ten times bigger than the sample size.

Null Hypothesis (one-tail test/directional test):

$$H_0 : p_{\text{Gold}} \leq p_{\text{Red}}$$

$$H_1 : p_{\text{Gold}} > p_{\text{Red}}$$

pooled proportion

$$\hat{p} = \frac{5901 + 6012}{59504 + 58944} = 0.10058$$

In [2]:
```
p_pooled = (5901+6012)/(59504+58944)
print(p_pooled)
```

```
0.10057578008915305
```

standard error

$$SE = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{59504} + \frac{1}{58944})} = 0.00175$$

In [3]:
```
import numpy as np
se = np.sqrt(p_pooled*(1-p_pooled)*(1/59504+1/58944))
print(se)
```

```
0.001747833171389562
```

z-score

$$z = \frac{p_{\text{Gold}} - p_{\text{Red}}}{SE} = 1.61646$$

In [4]:
```
z = (6012/58944 - 5901/59504) / se
print(z)
```

```
1.616464512796479
```

significance level of 7%

In [5]:
```
alpha = 0.07
z_sig = scipy.stats.norm.ppf(1 - alpha)
print(z_sig)
```

```
1.47579102817917
```

In [6]:
```
if np.abs(z) > z_sig:
    print("Reject H_0")
else:
    print("Fail to reject H_0")
```

```
Reject H_0
```

Reject H_0(Gold<=Red) in one-tail test.

$\therefore$ We can say that H_1 Gold is better than Red.

Conclusion:

Gold should be chosen to run at 100% traffic.

10.2. What are the confidence intervals at 7% significance of conversion rates for Red and Gold? Show your work.

Confidence Intervals for a One-Tailed Test each color:

$$CI = p \pm z_\alpha \times SE \quad ; SE = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

In [7]:
```
p_red = 5901/59504
ci_red_upper = p_red + z_sig*se
ci_red_lower = p_red - z_sig*se
print(f"CI for red: ({ci_red_lower}, {ci_red_upper})")
```

```
CI for red: (0.09659036719758425, 0.10174924022376557)
```

In [8]:
```
p_gold = 6012/58944
ci_gold_upper = p_gold + z_sig*se
ci_gold_lower = p_gold - z_sig*se
print(f"CI for gold: ({ci_gold_lower}, {ci_gold_upper})")
```

```
CI for gold: (0.099415677493424, 0.10457455051960532)
```

Conclusion:

CI_{red}: (0.09659036719758425, 0.10174924022376557)

CI_{gold}: (0.099415677493424, 0.10457455051960532)

11. Understanding of A/B Testing

Which of the following are true about frequentist A/B tests? (True/False)

11.1. It does not tell us the magnitude of the difference between control and test groups.

11.2. We can never know when to stop the experiments.

11.3. We can never determine if the null hypothesis being true.

11.4. We can run one or as many experiments as we want using the same significance level.

11.5. If we have too many samples in each group, the validity of the test can be jeopardized.

11.6. If you have set up the experiment based on desired minimum detectable effect and significance level, statististical significance is the only factor in determining which group is the better one.

11.7. We can only test difference between two proportions.

11.8. More samples in control and test groups are always better.