Homework Stats - AB Testing Exercise

Week 4

6733172621 Patthadon Phengpinij

Collaborators. ChatGPT (for LaTeX styling and grammar checking)

1 AB Testing

Problem 1. Select The Problem:
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?

Solution. True
Because, we can randomly assign users to see either version of the button (A: original color, B: new color) and measure the conversion rates for both groups. By comparing the conversion rates, we can determine if the color change has a statistically significant effect on user behavior.

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?

Solution. True
Because, we can randomly assign new users to one of the four versions of the recommendation model (A, B, C, D) and track key performance metrics such as click-through rate, conversion rate, or user engagement for each group. By analyzing the results using statistical methods, we can identify which model performs best and make data-driven decisions on which one to implement.

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

Solution. False
Because, conducting an A/B test in a physical store setting can be challenging due to factors such as customer traffic patterns, store location, and external influences that may affect customer behavior. Additionally, it may be difficult to randomly assign customers to different layouts without introducing bias.

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?

Solution. False
Because, conducting an A/B test to compare offline stores and online websites can be complex due to differences in customer demographics, purchasing behavior, and external factors that may influence sales.

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

Solution. False
Because, introducing a premium version to a freemium service involves multiple factors that may not be easily isolated in an A/B test. These factors include pricing strategy, feature differentiation, and user willingness to pay, which can all impact the success of the premium offering.

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?

Solution. True
Because, we can randomly assign users to either the current website setup (A) or the new backend setup (B) and measure the website load times for both groups. By comparing the load times using statistical analysis, we can determine if the new setup significantly improves performance.

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?

Solution. False
Because, the long purchase cycle makes it difficult to measure the immediate impact of the banner change on customer behavior. A/B testing may not capture the delayed effects of the banner change on repeat purchases over such a long timeframe.

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

Solution. False
Because, a total revamp of corporate identity is a significant change that may not be easily tested through A/B testing. The impact of such a change on brand perception, customer loyalty, and overall business performance may take time to manifest and may be influenced by various external factors.

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?

Solution. True
Because, showing fewer products on the first page may impact user engagement and conversion rates. A/B testing can help determine if the change leads to better performance or if it negatively affects user behavior.

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

Solution. **True**

Because, we can segment users based on their operating system (iOS vs. Android) and randomly assign them to receive different ad campaigns. By measuring key performance metrics such as click-through rates, conversion rates, and return on ad spend for each group, we can determine which platform responds better to the ads.

Problem 2. Choose The Methods:

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 /** ______

Solution. **clicks / page views**

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

Solution. **purchases / unique visitors**

2.3 Which types of promotion coupons will be more effective? **purchases /** ______

Solution. **purchases / unique visitors**

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

Solution. **clicks / page views**

Problem 3. Concern The Period:
Based on the transaction table below,

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

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

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

Solution. 1 purchase (A) / 3 unique visitors (A, B, C)
Because, event-based conversion rate considers all events (visits and purchases) that occurred within the attribution period. In November 2020, there were a total of 3 unique visitors (A, B, C) who visited the site. However, only user A made a purchase within the 7-day attribution period. Thus, the event-based conversion rate for November 2020 is calculated as 1 purchase divided by 3 unique visitors, resulting in a conversion rate of approximately 33.33%.

3.2 What are the cohort-based conversion rate of 2020-11?

Solution. 2 purchases (A, C) / 3 unique visitors (A, B, C)
Because, cohort-based conversion rate considers users who made a purchase within the attribution period, regardless of when they first visited the site. In November 2020, there were a total of 3 unique visitors (A, B, C). Users A and C made purchases within the 7-day attribution period following their visits. Thus, the cohort-based conversion rate for November 2020 is calculated as 2 purchases divided by 3 unique visitors, resulting in a conversion rate of approximately 66.67%.

Problem 4. Familiar With 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

Solution.

1. coin tosses
2. chance of a lottery ticket to be a winning one
3. chance of a product being defective

4.2 Normal/Student t's distribution

Solution.

1. heights
2. weights
3. standardized test scores

4.3 Exponential distribution

Solution.

1. wait times between buses
2. customer spending
3. number of days between machine breakdowns

4.4 Poisson distribution

Solution.

1. number of bus arrivals at a stop per hour
2. number of site visitors per day
3. number of road accidents per day during the holiday season

Problem 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.

Solution. We should control for **AGE**.

5.2 We want to test if GUN OWNERSHIP ⇒ CRIMES and we know that GUN OWNERSHIP ⇒ GUN SALES and CRIMES ⇒ GUN SALES.

Solution. We should control for **NOTHING**.

5.3 We want to test if CROP BURNING ⇒ LUNG DISEASES and we know that CROP BURNING ⇒ PM2.5 and PM2.5 ⇒ LUNG DISEASES.

Solution. We should control for **NOTHING**.

Problem 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.

Solution. From the definition of sample variance,

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}^2), \text{ where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Apply LLN, we have,

$$\text{As } n \to \infty,\ \bar{X} \to \mu$$

Therefore,

$$\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \\
&= \frac{1}{n}(\sum_{i=1}^{n} {X_i}^2 - 2\mu\sum_{i=1}^{n} X_i + n\mu^2) \\
&= \frac{\sum_{i=1}^{n} {X_i}^2}{n} - \frac{2\mu\sum_{i=1}^{n} X_i}{n} + \mu^2 \\
&= \frac{\sum_{i=1}^{n} {X_i}^2}{n} - 2\mu\bar{X} + \mu^2 \\
&= \frac{\sum_{i=1}^{n} {X_i}^2}{n} - 2\mu^2 + \mu^2 \\
&= \frac{\sum_{i=1}^{n} {X_i}^2}{n} - \mu^2 \\
&= E[{X_i}^2] - E[X_i]^2 \\
&= Var(X_i) \\
s^2 &= \sigma^2
\end{aligned}$$

Problem 7. P-values:
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.

Solution.

- 7.1. **is correct** because p-value quantifies the probability of obtaining results at least as extreme as the observed data, assuming that the null hypothesis is true.
- 7.2. **is incorrect** because p-value does not provide the probability of the null hypothesis being true based on the observed data.
- 7.3. **is incorrect** because p-value does not provide the probability of the null hypothesis being false based on the observed data.
- 7.4. **is incorrect** because it describes the concept of statistical power, not p-value.

Problem 8. AB Testing Exercise:
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?

Solution. At 5% significance level, we can expect to have statistically significant results in 5% of the tests even if group A and B are exactly the same.

Therefore, for 4,000 tests, the expected number of statistically significant results is:

$$0.05 \times 4000 = \boxed{200}$$

Problem 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. (Power = 0.5)

Solution. From the control/test split, we have

$$m = \frac{\text{control}}{\text{test}} = \frac{80}{20} = 4$$

.
The conversion rate of control group is $p = 0.11$. Thus, the sample variance of control group, that can be assumed to be equal to the test group, is

$$\sigma^2 = p(1-p) = 0.11(1-0.11) = 0.0979$$

.

From the formula for minimum detectable effect (MDE),

$$Z_\alpha + Z_\beta = \frac{\text{MDE} - \mu}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{mn}\right)}}$$

While, the value of MDE is given as

$$\text{MDE} = 0.01$$

. And, we are testing at significance level of 15%, thus,

$$Z_\alpha = Z_{0.15,\text{one-tailed}} = 1.03643$$

Also, the power is given as 0.5, thus,

$$Z_\beta = Z_{0.5,\text{one-tailed}} = 0$$

Now, we can calculate the required sample size for the test group n as follows,

$$\begin{aligned}
Z_\alpha + Z_\beta &= \frac{\text{MDE} - \mu}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{mn}\right)}} \\
\frac{(m+1)\sigma^2}{mn} &= \left(\frac{\text{MDE}}{Z_\alpha + Z_\beta}\right)^2 \\
n &= \frac{m+1}{m}\left(\frac{(Z_\alpha + Z_\beta)\sigma}{\text{MDE}}\right)^2 \\
&= \frac{5}{4}\left(\frac{(Z_\alpha + Z_\beta)\sigma}{\text{MDE}}\right)^2 \\
&= \frac{5}{4}\left(\frac{(Z_\alpha + Z_\beta)}{\text{MDE}}\right)^2 \times \sigma^2 \\
&= \frac{5}{4}\left(\frac{(1.03643 + 0)}{0.01}\right)^2 \times (0.0979) \\
n &\approx 1314.55
\end{aligned}$$

Rounding up, we need at least **1,315 unique visitors** in the test group. From the control/test split, the required sample size for the control group is

$$mn = 4 \times 1,315 = 5,260 \text{ unique visitors}$$

Thus,

$$\boxed{\textbf{test group of 1,315 unique visitors}}$$

and

$$\boxed{\textbf{control group of 5,260 unique visitors}}$$

Problem 10. Hamster Inc. and His A/B Testing Experiment:
Let us say Hamster Inc. ran the experiment and got the following results.

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

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

Solution. To decide which variation to choose, we can perform a hypothesis test to compare the conversion rates of the Red and Gold packages.

The conversion rates for Red and Gold packages are given as follows:

$$p_{red} = 0.099170 \quad \text{and} \quad p_{gold} = 0.101995$$

The null hypothesis H_0 is that there is no difference in conversion rates:

$$H_0 : p_{red} = p_{gold}$$

The alternative hypothesis H_A is that the Gold package **has a higher conversion rate**:

$$H_A : p_{gold} > p_{red}$$

Since we are testing at a significance level of 7%, we need to find the critical value Z_α for a one-tailed test:

$$Z_\alpha = Z_{0.07,\text{one-tailed}} = 1.475791$$

And, we can consider the sample variances for both groups:

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

We can calculate the standard error (SE) for the difference in conversion rates:

$$\begin{aligned} SE &= \sqrt{\frac{p_{red}(1 - p_{red})}{n_{red}} + \frac{p_{gold}(1 - p_{gold})}{n_{gold}}} \\ &= \sqrt{\frac{0.099170(1 - 0.099170)}{59504} + \frac{0.101995(1 - 0.101995)}{58944}} \\ &= \sqrt{\frac{0.089253}{59504} + \frac{0.091584}{58944}} \\ SE &\approx 0.001748 \end{aligned}$$

Next, we calculate the test statistic Z:

$$\begin{aligned} Z &= \frac{p_{gold} - p_{red}}{SE} \\ &= \frac{0.101995 - 0.099170}{0.001748} \\ Z &\approx 1.616 > Z_\alpha = 1.475791 \end{aligned}$$

Since $Z > Z_\alpha$, we reject the null hypothesis. Therefore, we can conclude that the Gold package has a significantly higher conversion rate than the Red package at the 7% significance level.

Thus, Hamster Inc. should choose the $\boxed{\textbf{Gold package}}$ 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.

Solution. To calculate the confidence intervals for the conversion rates of Red and Gold packages at a 7% significance level, we can use the formula for the confidence interval of a proportion:

$$CI = p \pm \left(Z_{\alpha/2} \times SE\right)$$

At a 7% significance level, the critical value $Z_{\alpha/2}$ for a two-tailed test is:

$$Z_{\alpha/2} = Z_{0.035,\text{two-tailed}} = 1.811911$$

For each group, we can calculate the confidence intervals as follows:
Red Package:

$$\begin{aligned}SE_{red} &= \sqrt{\frac{p_{red}(1-p_{red})}{n_{red}}}\\ &= \sqrt{\frac{0.099170(1-0.099170)}{59504}}\\ SE_{red} &\approx 0.001225\\ \\ CI_{red} &= 0.099170 \pm (1.811911 \times 0.001225)\\ &= 0.099170 \pm 0.002220\\ CI_{red} &\approx (0.096950, 0.101390)\end{aligned}$$

Gold Package:

$$\begin{aligned}SE_{gold} &= \sqrt{\frac{p_{gold}(1-p_{gold})}{n_{gold}}}\\ &= \sqrt{\frac{0.101995(1-0.101995)}{58944}}\\ SE_{gold} &\approx 0.001247\\ \\ CI_{gold} &= 0.101995 \pm (1.811911 \times 0.001247)\\ &= 0.101995 \pm 0.002259\\ CI_{gold} &\approx (0.099736, 0.104254)\end{aligned}$$

In conclusion, the confidence intervals at 7% significance level are:

$$\boxed{CI_{red} \approx (0.096950, 0.101390)} \text{ and } \boxed{CI_{gold} \approx (0.099736, 0.104254)}$$

Problem 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.

Solution. True
Because, frequentist A/B tests primarily focus on determining whether there is a statistically significant difference between the control and test groups. They do not provide information about the size or practical significance of that difference.

11.2 We can never know when to stop the experiments.

Solution. False
Because, we can use techniques such as sequential testing or adaptive experimentation to determine when to stop experiments based on the data collected.

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

Solution. True
Because, we can never know for sure if the null hypothesis is true or false. We can only gather evidence to support or reject it based on the data we collect.

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

Solution. False
Because, while it is possible to run multiple experiments using the same significance level, doing so without proper adjustments (like Bonferroni correction) can increase the risk of Type I errors (false positives).

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

Solution. True
Because, excessively large sample sizes can lead to detecting statistically significant differences that are not practically meaningful, potentially leading to misguided decisions.

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.

Solution. True
Because, if the experiment is properly designed and conducted, statistical significance can provide strong evidence for determining which group is better. However, it is important to also consider practical significance and the context of the results.

11.7 We can only test difference between two proportions.

Solution. **False**

Because, frequentist A/B tests can be used to compare means, variances, and other statistical measures, not just proportions.

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

Solution. **False**

Because, while larger sample sizes can provide more accurate estimates and increase the power of a test, they can also lead to overfitting and the detection of statistically significant but practically meaningless effects.