Questions on A/B testing are being increasingly asked in interviews but reliable resources to prepare for these are still far and few. Let's say you completed a great course on A/B testing sometime back and felt confident in your understanding of A/B testing statistics. It's been a while now, you have an interview coming up and for the love of god, you can't seem to remember what statistical power means or what SUTVA stands for. OR you have been using experimentation in your current role and have automated most processes. Without the need for manual steps, you have become rusty and need a quick cheat sheet to recall important concepts and intuition behind them so that you can ace an upcoming interview.

Use this post as a quick resource for the most important concepts in A/B testing you need to know before interviews. You will find

a summary of the most important concepts along with examples to build your intuition.

Table of contents (click below to skip to specific topics)

- The Very Basics 📚

  Population, sample, sample mean, sample variability

- Experiment Design 🎭

  Null Hypothesis, Key Metrics, Overall Evaluation Criteria (OEC), Guardrail Metrics, Randomization Unit, Interference

- A/B Test Statistics & Sample Size Calculation 🧮

  Confidence Level, Margin of error, Confidence Interval, Type 1 Error, Type 2 Error, p-value, Statistical Significance, Statistical Power, Minimum Detectable Effect, Practical Significance, Sample Size & Duration

- Threats to Experiment Validity 🚧

  Novelty Effect, Primacy Effect, Seasonality, Day of Week Effect

Before we begin, let's establish an example of an A/B test we will utilize as we go over the concepts. Lets say you are running an A/B test on a web page with the goal to improve click-through rates (CTRs) — your original web page, which is the control offers a $25 saving using the text 'Save $25'; through this A/B test you are going to test a variation of this web page that presents the same offer (value is still $25) in a different way using the text 'Save 15%'.

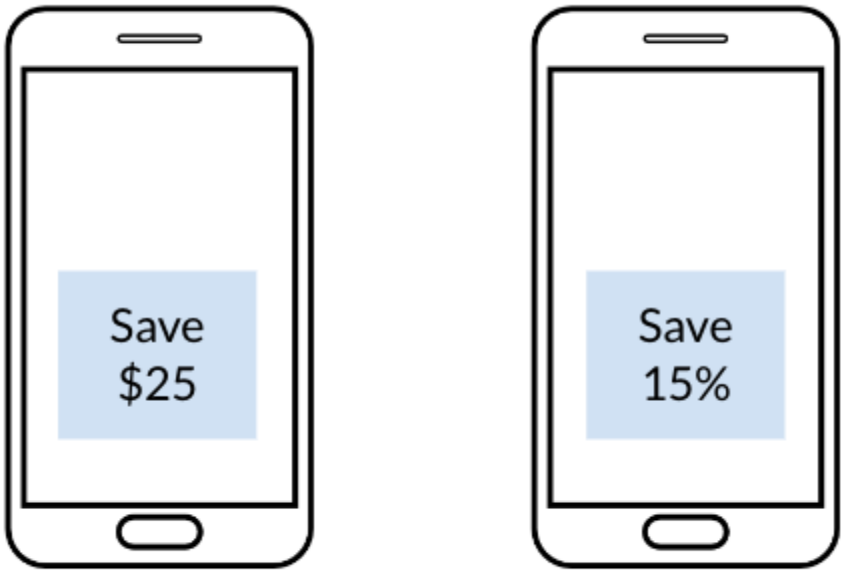


Image by Author

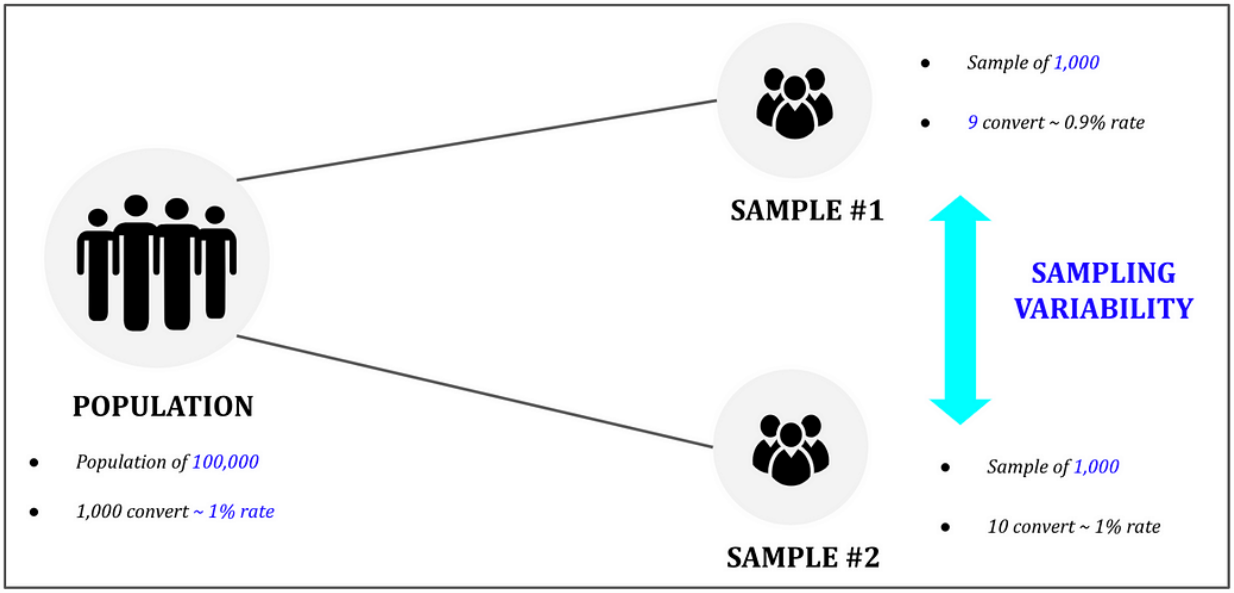
Now onto the concepts…

The Very Basics

1. **Population** — It is the entire group that you want to draw conclusions about. In statistics, a population is a set of similar items or events which is of interest for some question or experiment. *In our example above, the true population is every future individual who will visit the web page.*

2. **Sample** — A sample is a small portion of a population that is representative of the characteristics of the larger population. With statistical analysis, you can use data collected from samples to make estimates or test hypotheses about population. *In A/B testing, a sample is the randomly selected set of visitors we display each of our page variations to — control is exposed to one sample and treatment is exposed to another.*
3. **Sample Mean** — For a given metric, this is the mean or average based on data collected for the sample. *For our A/B test example that is aiming to optimize CTR (click-through rate), this is nothing but the average CTR for users in each sample.*
4. **Sample Variability** — Two randomly selected samples from a population may be different from each other. Whatever conclusions we are drawing about the population through a sample, because of sample

variability, there is likely to be error in our estimations. Sampling variability will decrease as sample size increases. *In A/B testing, the sampling variability affects the sample size we need in order to have a chance of deriving statistically significant results.*



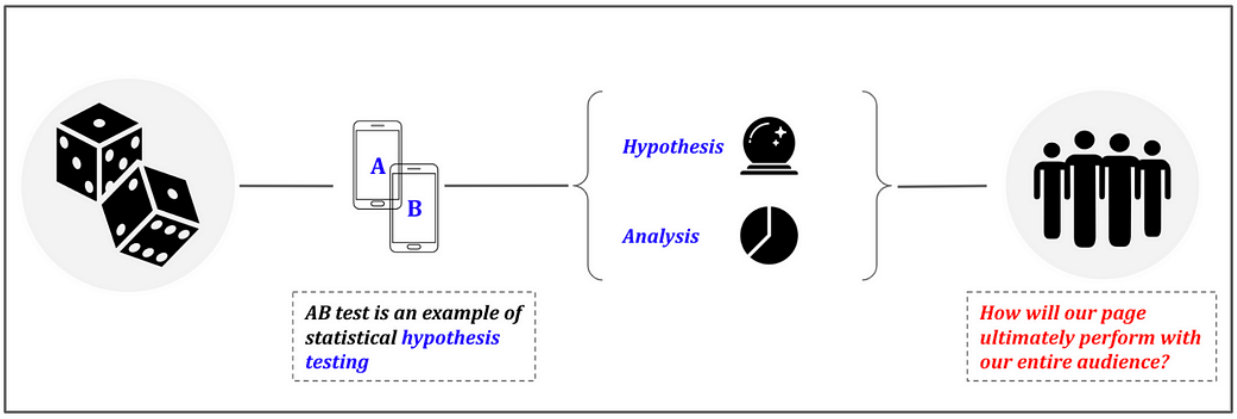


Image by Author

Experiment Design

5. Null Hypothesis — Inferential statistics is based on the premise that you cannot prove something to be true but you can disprove something by finding an exception. You decide what you are trying to provide evidence for — which is the alternate hypothesis, then you set up the opposite as the null hypothesis and find evidence to disprove that. *In our A/B test example, the null hypothesis is that the population CTR on the original page and the page variation are not different.*

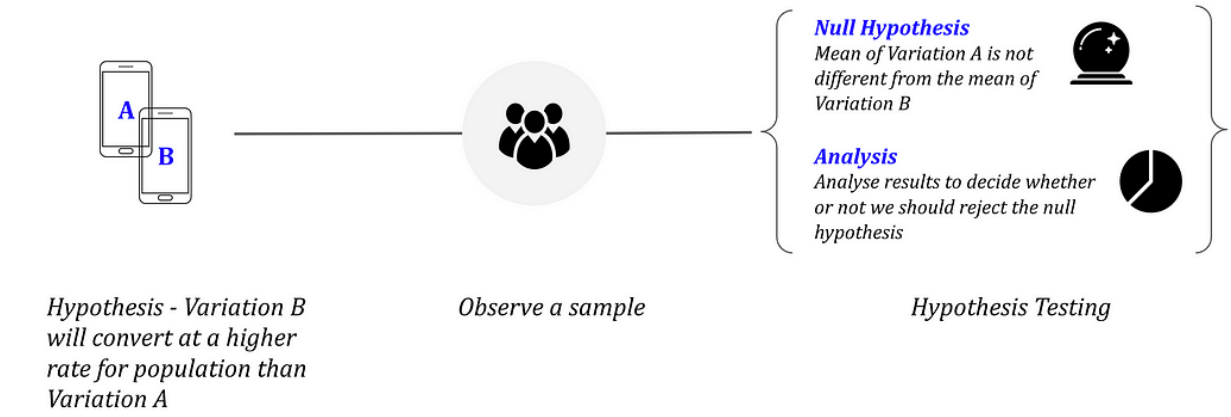


Image by Author

6. Key Metric/s — A set of metrics that you are trying to optimize through the experiment. *Some commonly used metrics are click through rates (CTR), sign up rate, engagement rate, average revenue per order, retention rates etc.* As you can imagine key metrics are going to relate to the priorities of the business, OKRs and goals. Many organizations examine multiple key metrics, and have a mental model of the trade offs they are willing to accept when they see a particular combination. *For e.g., they may have a good idea about how much they are willing to lose users (increase in churn) if the remaining users*

increase their engagement and revenue. This brings us to OEC explained below.

7. Overall Evaluation Criteria (OEC) — When there are multiple metrics to be optimized through the experiment, it is helpful to formulate trade-offs by devising a single metric called an Overall Evaluation Criteria (OEC) — which is essentially a weighted combination of such objectives. *One way to do this is to normalize each metric to a predefined range, say 0–1, and assign each a weight. Your OEC then is the weighted some of the normalized metrics. In the example above with needing to evaluate trade-off between churn and revenue, LTV can be used as an OEC.*

8. Guardrail Metrics — These are metrics that are important for the company and should not be negatively impacted by the experiment. *For e.g. our goal may be to get as many users as possible to register, but we don't want the per-user engagement*

level to drop drastically. Or we may want to increase app engagement but at the same time ensure that app uninstalls do not increase.

9. Randomization Unit — This is the unit e.g. users or pages that a randomization process is applied to map them to either control or treatment. Proper randomization is important to ensure that populations assigned to the different variants are similar statistically. Randomization unit should be chosen such that Stable unit treatment value assumptions (SUTVA) are satisfied. SUTVA states that experiment units do not interfere with one another i.e. the behavior of units in test and control is independent of each other. User-level randomization is the most common as it avoids inconsistent experience for the user and allows for long-term measurement such as user retention.

10. Interference — Sometimes also called spillover or leakage occurs when the behavior of the control group is influenced by

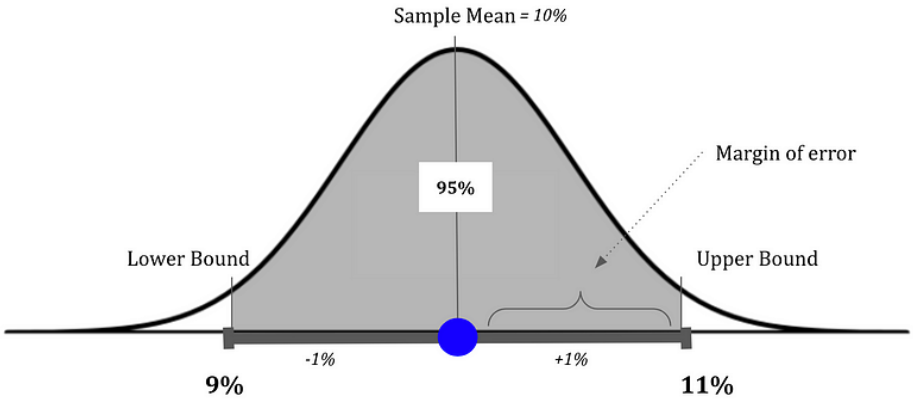
the treatment given to the test group. This leads to a violation of SUTVA assumption which results in potentially incorrect conclusions. There are two ways inference may arise —

- Direct — two units can be directly connected if they are friends on a social network or if they visited the same physical space at the same time. If one of these is assigned to Treatment and other to control, this will cause interference between variants
- Indirect — indirect connections are connections that exist because of certain shared resources. For e.g. If the Airbnb marketplace improved conversion flow for treatment users, resulting in more bookings, it would naturally lead to less inventory for Control users. Similarly marketplaces such as Lyft/Uber/Doordash where users in control and treatment might share the same pool of drivers/dashers will also face interference

Sample size calculation

11. **Confidence Level** — Confidence level refers to the percentage or probability or certainty, that the confidence interval would contain the true population parameter when you draw a random sample many times. In the tech world of online A/B Testing, a 95% confidence level is chosen most often but you can choose different levels depending on the situation. *A 95% confidence level means that the confidence interval around sample mean is expected to include the true mean value 95% of the time.*

Confidence Interval = [Mean - Margin of Error, Mean + Margin of Error]



12. **Margin of error** — As we noted earlier, due to sampling variability, it is possible that the conclusions your draw about the population based on samples is inaccurate. A margin of error tells you how many percentage points your results will differ from the real population value. *For example, a 95% confidence interval with a 4 percent margin of error means that your statistic will be within 4 percentage points of the real population value 95% of the time. The margin of error is added to and subtracted from the mean to determine the confidence interval (discussed below).*

13. **Confidence Interval** — In statistical inference, we aim to estimate population parameters using observed sample data. A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. *The width of confidence interval depends on 3 things — the*

variation within the population of interest, the size of the sample and the confidence level we are seeking.

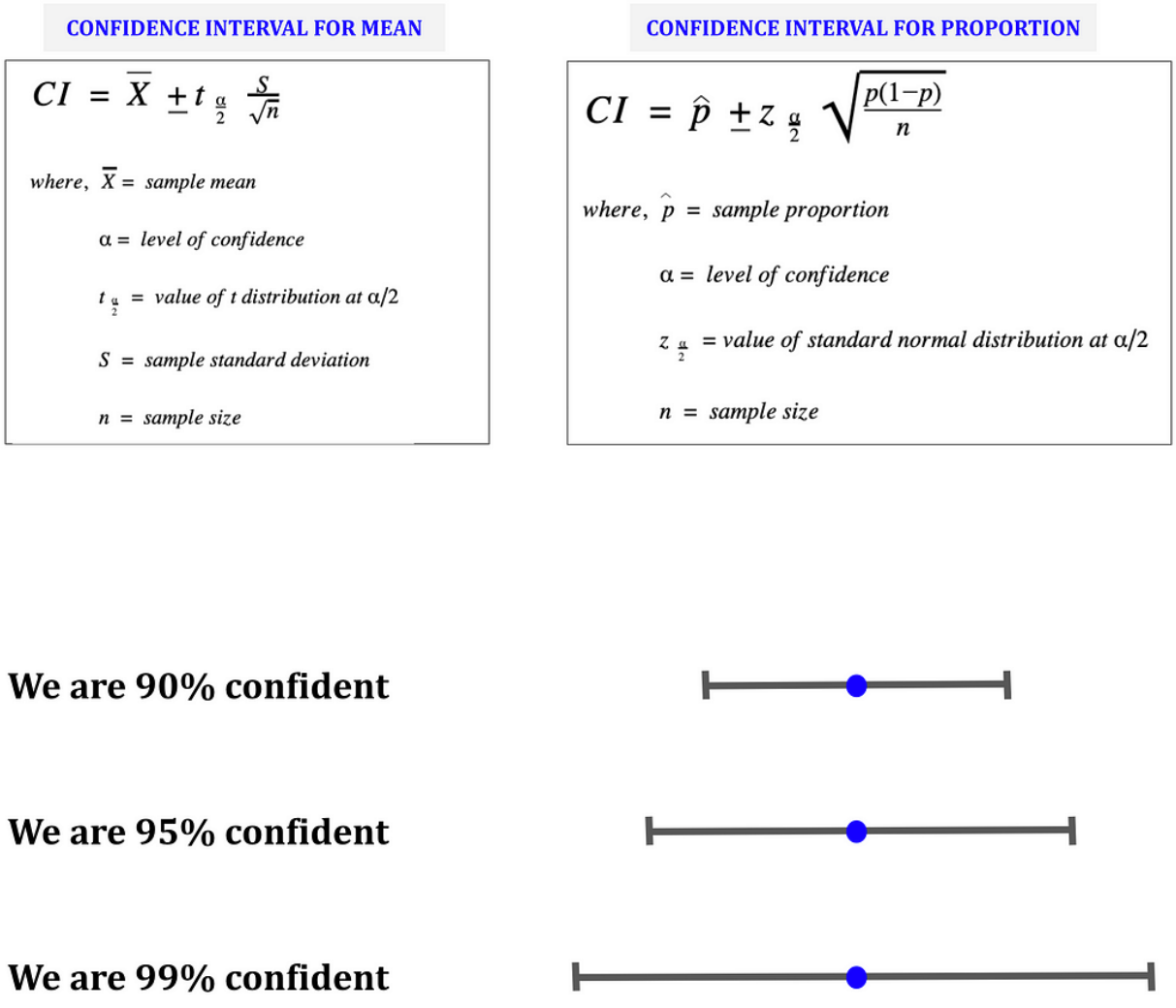


Image by Author

14. **Type 1 Error** — A type I error occurs when we incorrectly reject the null hypothesis. *In our A/B test example, a type I*

error would occur if we concluded that population mean of treatment is different from population mean of control when in reality they were the same. Type I error is avoided by achieving statistically significant results.

15. **Type 2 Error** — A type II error occurs when the null hypothesis is false, but we incorrectly fail to reject it. *In our A/B test example, a type II error would occur if we concluded that the population mean of Variation B is not different than the mean of Variation A when it actually was different*. These errors are avoided by running tests with a high statistical power.

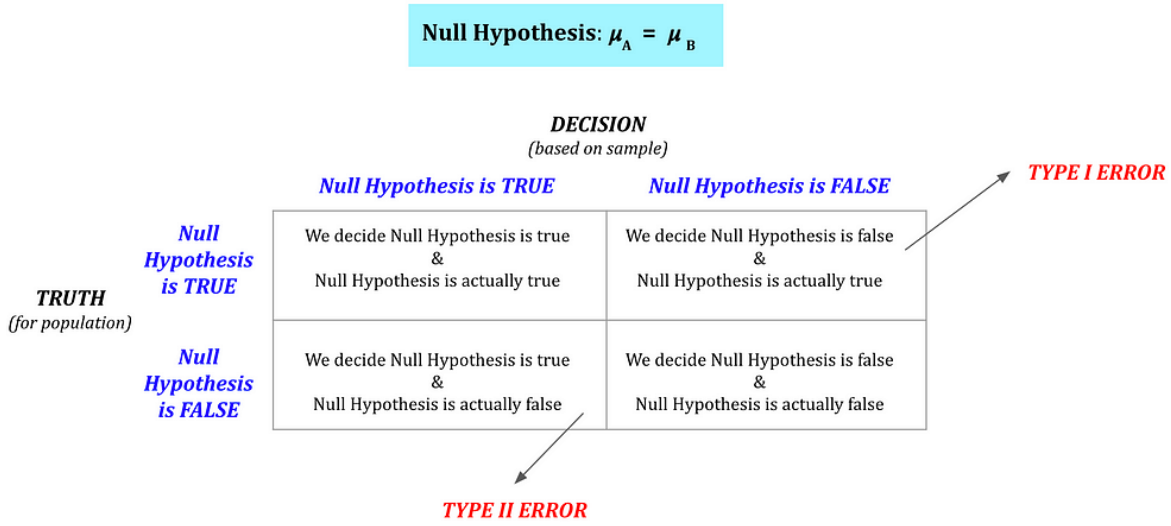


Image by Author

16. **P-value** — p-value is the probability of obtaining at least as extreme results as we are seeing if the null hypothesis of the test is true. p-value basically tells you whether your evidence makes your null hypothesis look ridiculous.

17. **Statistical Significance** — Statistical significance is attained when the p-value is less than the significance level. The significance level (**α**), is the threshold you want to use for the probability of making Type 1 error i.e. concluding that population mean of Control and Treatment are different when in fact they are the same. In other words, statistical significance is

another way of saying that the p-value of a statistical test is small enough to reject the null hypothesis. The scientific standard is to use a p-value < 0.05 i.e. $\alpha = 5\%$.

18. **Statistical Power** — Statistical Power, which as we know is the probability that a test correctly rejects the null hypothesis i.e. the percentage of time the minimal effect will be detected, if it exists.

19. **Minimum Detectable Effect (MDE)** — is the smallest change in conversion rate you are interested in detecting. In the example where you are optimizing for CTR, let's say the CTR for the control is 20%. And the smallest change you would like to detect at the minimum is 5% absolute lift relative to control i.e. if the CTR for the treatment is 25% or more. In this case 5% is the MDE.

20. **Practically significant —** It's possible for hypothesis tests to produce results that are statistically significant, despite having a small effect size. This usually happens due to 2 reasons — 1) Low sampling variance and 2) Large sample size. In both these cases, we may be able to detect even small differences between test and control with statistical significance. However, these may not be significant in the real world. *Let's take an example of an A/B test which shows a new module to users and is able to detect a difference of .05% in the CTR — however, is that cost of building that module justified for such a small lift. What is the practically significant lift that would be feasible in this case.*

21. **Sample size** — The number of units per variation needed to reach statistical significance, given the baseline conversion, minimum detectable difference (MDE), significance level & statistical power chosen. 'Duration' refers to how long you need to run the test to reach adequate sample size per variation.

Threats to experiment validity

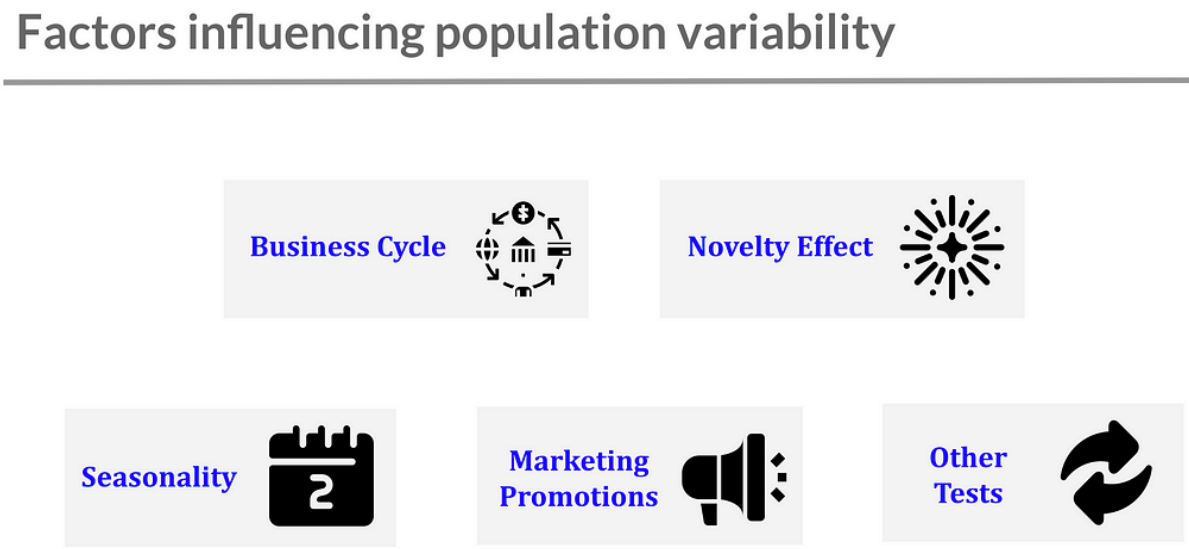


Image by Author

22. **Novelty Effect** — “Sometimes there’s a “novelty effect” at work. Any change you make to your website will cause your existing user base to pay more attention. Changing that big call-to-action button on your site from green to orange will make returning visitors more likely to see it, if only because they had tuned it out previously. This type of effect is not likely to last in the long run — but it can artificially impact your test results.

23. **Primacy effect** — When there's a change in the product, people react to it differently. Some users may be used to the way a product works and are reluctant to change. This is called primacy effect. The primacy effect is nothing but the tendency to remember the first piece of information we encounter better than information presented later on. This can be thought of as an ***opposite phenomenon to novelty effect***.

24. **Seasonality** — Businesses may have different user behavior say on 1st of a month and 15th of a month. For some eCommerce sites, their traffic and sales are not stable all over the year, they tend to peak on Black Friday and Cyber Mondays for example. Variability due to these factors could influence your test results.

25. **Day of week effect** — Similar to seasonality, a metric may have cyclicity based on day of week. For e.g. say conversion rates are much higher on Thursdays than they are on the weekends. In

this case, it is important to run tests for full-week increments, so you are including every day of the week.

Conclusion

A/B Testing is one of the most important and widely applied data science concepts with numerous applications in growth optimization — be it product experimentation *(optimizing onboarding flows, increasing CTRs, server side optimizations etc.)* OR for marketing acquisition *(creative testing, incrementality testing, geo-split testing)* and many more. With so many potential applications, it is very likely that you will be gauged on your knowledge of A/B testing in interviews — specifically for product data science/analytics or marketing data science/analytics roles.

If you would like to continue learning A/B testing concepts and applications, check out Complete Course on A/B Testing with Interview Guide.