**Diagnose a metrics problem:**

* **Clarify definition of metrics**
* **Decompose the metric**

i.e. daily active user (DAU) = existing users + new users + resurrected users – churned users

i.e. activation rate = activation / signup

i.e. ETA increases: riders increase or driver decreases

* **Time frame** – sudden or progressive change?
  + Sudden:
    - Outliers: set a quantile threshold
    - Data collection issue: data pipeline change
    - Technical issues: bug
    - Algorithm change
    - Bot
    - Marketing campaign
  + Gradual or Progressive (over a few weeks or over a year?):
    - Seasonality: compare with trends from last week / last year; is this drop large historically? (time series plot)
    - Other metrics also declining?
    - Competitor, Industry trends
    - Pricing and packaging: pricing concerns?
* **Product change**: Other experiment or product / feature change within company?
* **Segmentation**: Is the drop for all users or only some? Mix shift across segments? Targeted users who are less engaged
  + Product SKUs
  + Geographical region: country, region
  + Demographics (age group, gender, tenure (new users vs. existing users), language, interests)
  + Platforms (web vs mobile web vs mobile)
  + Device (desktop vs android vs IOS)
  + Channel (paid vs non-paid marketing channel)
  + Post types (friends, public content, ads)
    - Post volume by post types to see whether mix shift
    - Metrics by post type to see any type have higher or lower metric
* **User journey/funnel**: which steps in the funnel cause the drop?
* **Substitution** for other behaviors/product features? i.e. comment or reactions but overall time spent on app not dropping; time spent on other activities or other product feature increase?

**Product feature case:**

**How would you decide whether to build this feature/product? Is this a valuable product? How would you evaluate the feature performance / product success with metrics?**

**What’s the business goal?** Acquisition, Activation, Engagement, Retention, Monetization?

1. **Acquisition:**
   * User/business: new signup / signin
2. **Activation:**
   * Activation metrics: users with 1+ transaction or users who have done certain actions xx days post signup
   * Set up; aha; habit
3. **Engagement:**
   * User:
   * DAU or MAU (%DAU, %MAU)
   * User time spent per day; session time per user per day
   * Adoption / engagement: (#% user, average action per user): Clicks, save, edit, comment, share, reply, messages sent
   * Business: monthly active business
4. **Monetization**
   * Revenue (Ads revenue)
   * NNARR = GNARR + Upsell ARR – Downsell ARR – churn ARR
   * Direct purchase
   * Trial start, trial convert, TPCR
   * LTV: average 3-year revenue per user
   * Free to paid conversion
   * New paid licenses
5. **Retention**
   * User retention rate
   * Churn rate: involuntary churn vs. voluntary churn; mobile app churn rate

**Product /feature success metrics**: by user vs. business

* Reach, impression: view
* Adoption / engagement: (#% user, #% business, average action per user)
  + Clicks, save, edit, comment, share, reply
  + Sent messages
  + Store visits / sales
* Product funnel conversion
* Retention of product usage (1d, 7d, 30d)

**Counter metrics / trade off metrics**:

* Spam & fraud;
* Messages with no responses
* Cannibalization

**Guardrail metrics:**

* Revenue metrics
* User retention
* Web browse activity

**Growth accounting**:

* Over longer run, net new users? i.e. do people make net new friends or does FB attract net new users?
* Retention improvement?

**Experimentation:**

**How would you decide whether to launch this feature and measure impact?**

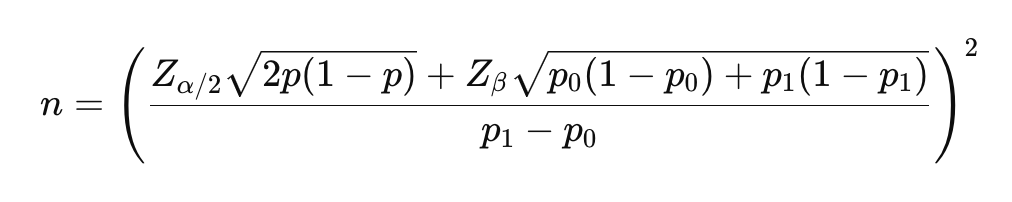
Hypothesis, metrics 🡪 power analysis to determine sample size 🡪 randomization, trigger point 🡪 launch test 🡪 data validation i.e. smoke test 1 day post launch 🡪 early results to check for large decrease in key metrics 🡪 run test for time duration and conduct analysis when we reach runtime

1. Hypothesis & goal?
2. Metrics? Primary, secondary and guardrail metrics. Set GA criteria
   1. Sum and count: volume
   2. Ratio: click through rate, conversion rate
   3. Distributional metrics: mean, medium, 25%, 75%, 90% percentile
      1. Mean is sensitive to outliers but not robust
      2. Median is robust but it may not change at all
3. Experiment design?
   1. Randomization unit:
      1. user id (concerns due to network effect)
      2. session id (include logged out users): group of webpages viewed on a single visit
         * cookie id: when a user visits, website writes a cookie containing an identifier; can be erased or incognito browse session used, leading to undercounting or overcounting
      3. device id (include logged out users and mobile only)
      4. cluster based: time block, time region block (address network effect)
   2. Target audience: eligible users
   3. Trigger point / unit of diversion: to be close to product experience change. Otherwise variants will dilute too much, baseline and variant value will be very close, because the majority of the users are not experiencing the change so harder to detect a stats sig change.
   4. Test group: # variants, test design
4. Event logging with eng: log events on key CTAs (client side & server side logging)
5. Runtime:

**Power analysis**: determine the smallest sample size required to detect the effect of a given test at the desired level of significance. We want to run the power analysis to make sure we get enough power to justify the investment into the test.

* 1. Statistical parameter:
     1. MDE: expected lift, effective size
     2. Significance level (alpha): probability of incorrectly flagging a stats sig change (0.05)
     3. Power (1-beta): probability of correctly flagging a real stats sig change
     4. Baseline conversion rate for proportion test, Baseline standard deviation for mean test

Sample size required for each group in two-sample proportion test:



Sample size required for each group in two-sample mean test:



Total required sample size = required sample size \* # variants

Runtime = Total required sample size / daily traffic

* 1. Non statistical parameter: seasonality, other test, product rollouts, full business cycle, countries, subpopulation

How to run test faster? Reduce sample size?

* + Increase significance level alpha
  + Reduce power (1-beta)
  + Increase MDE
  + Increase traffic: fewer variants; add trigger point more up the funnel to increase traffic (caveat on dilution)
  + Prefer to choose a metrics with a relatively larger absolute baseline proportion value / rate
  + Lower variance:
    - Ratio metrics have smaller standard error than mean metrics, meaning we will not need to expose the experiment to as many users to achieve the same sensitivity
    - Transform a metric through capping, binarization or log transformation. Heavy long-tailed metrics, consider log transformation; binary metrics to indicate whether user streamed more than x hours in a time period
    - Stratification: divide the sampling into strata, sample within each stratum separately, then combine results from individual strata for the overall estimate, which usually has smaller variance than estimating without stratification (post stratification, which applies stratification during analysis).
    - Control-variates or CUPED: it uses covariates as regression variables. CUPED utilizes pre-experiment data.
    - Randomize at a more granular unit i.e. each page view. Disadvantages: bad user experience; not able to measure any user level impact over time
    - Paired experiment: if you can show the same user both treatment and control in a paired design, you can remove between user variability and achieve a smaller variance
  + Sometimes we may decide to ship an experiment as long as there's no regression on primary metric + guardrails. We'd usually run experiments like this as a one-sided experiment.

How to increase power?

* + Increase sample size (Large samples 🡪 variance small 🡪 curve narrow 🡪 more power)
  + Decrease standard error
  + Increase the difference between sample statistic and hypothesized parameter or effect size

(When we increase the sample size, decrease the standard error, or increase the difference between the sample statistic and hypothesized parameter, the p value decreases, thus making it more likely that we reject the null hypothesis.)

* + Increase significance level (When we increase the alpha level, there is a larger range of p values for which we would reject the null hypothesis.)
  + Use a directional test as opposed to two tailed test (Going from a two-tailed to a one-tailed test cuts the p value in half.)

Note:

* + Small samples 🡪 variance large 🡪 curve flatten 🡪 less power
  + Large samples 🡪 variance small 🡪 curve narrow 🡪 more power
    - Large samples 🡪 variance small 🡪 t-statistic (z-score) larger and p-value smaller 🡪 more power to reject null

Underpowered test:

**Small sample size**

**Small effect size**

**High variance**

Why do you use ratio metrics instead of mean metrics as primary metrics?

* + Ratio metrics:
    - Account for differences in group sizes or variations in the data
    - A more stable measure of performance because they focus on relative relationship between the two groups
    - Normalize the data
    - Results are less sensitive to fluctuations and outliers
    - Offer a fair and standardized comparison regardless of absolute values or sample size differences;
  + Mean metrics:
    - have higher variance, especially when data includes outliers or extreme values. High variance in mean metrics can make it more difficult to detect statistically significant differences, requiring a larger sample size.
    - Mean metrics could be misleading when groups differ significantly in size, as the mean can be disproportionately affected by the larger group.

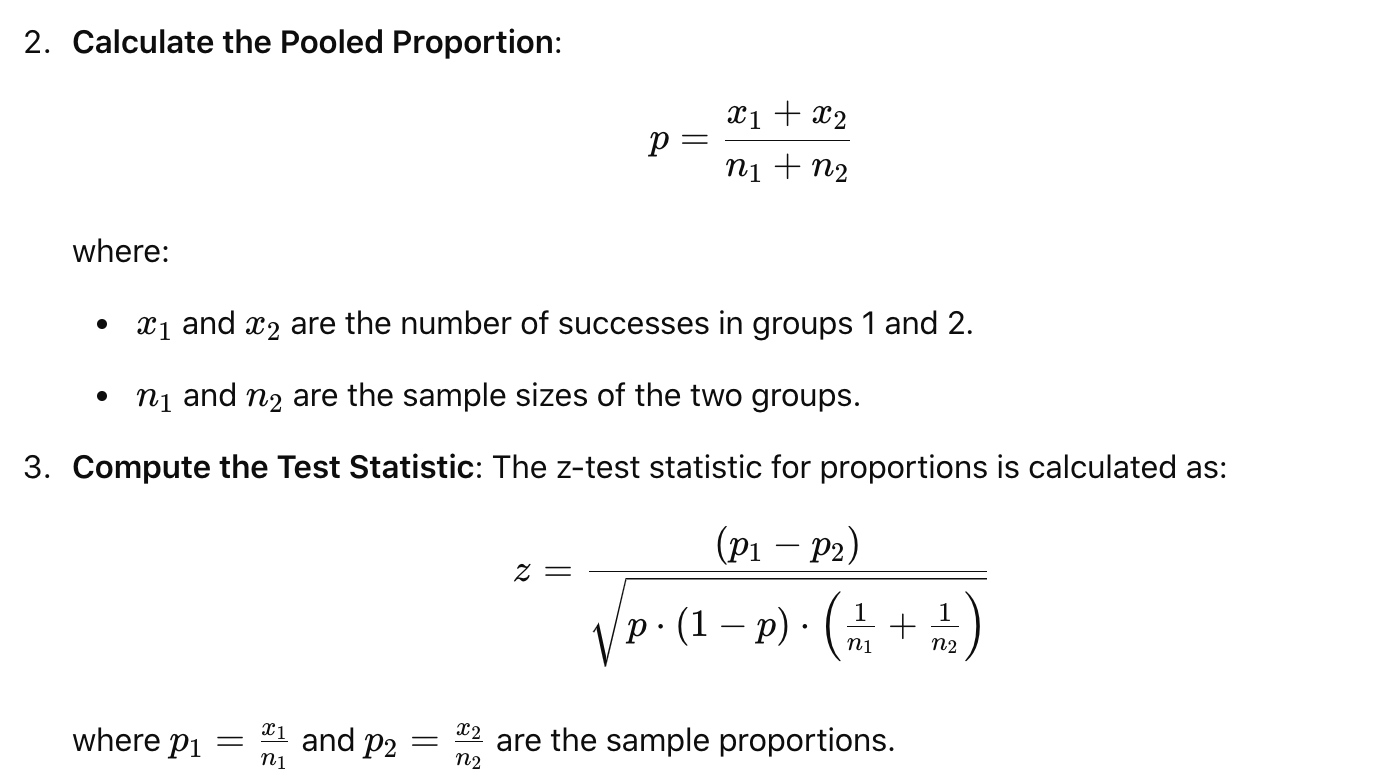
Issues with randomizing on session id?

* Cross-Session Contamination (Same User in Multiple Groups)
  + Since users can have multiple sessions, they might be assigned to different experiment groups across sessions.
  + Problem: The user experiences both conditions, making it hard to isolate the experiment’s true effect.
* Inflated Sample Size (Pseudoreplication)
  + A single user may generate multiple session IDs, leading to an overestimation of the number of independent observations.
  + Problem: The experiment may appear to have a large sample, but many sessions belong to the same users, violating the assumption of independence in statistical tests.
* Bias Toward High-Frequency Users
  + Users who have more sessions are overrepresented in the experiment.
  + Problem: Results become skewed toward behaviors of frequent users, not representing the overall user base.
* Short-Term Behavior Doesn’t Reflect Long-Term Impact
  + Randomizing by session ID means that a user’s experience can change each time they log in. If the experiment aims to measure long-term effects (e.g., retention, social connections), session-based randomization won’t reflect real behavior changes over time.
  + Problem: You won’t accurately measure true user-level outcomes like D30 retention or lifetime value (LTV).

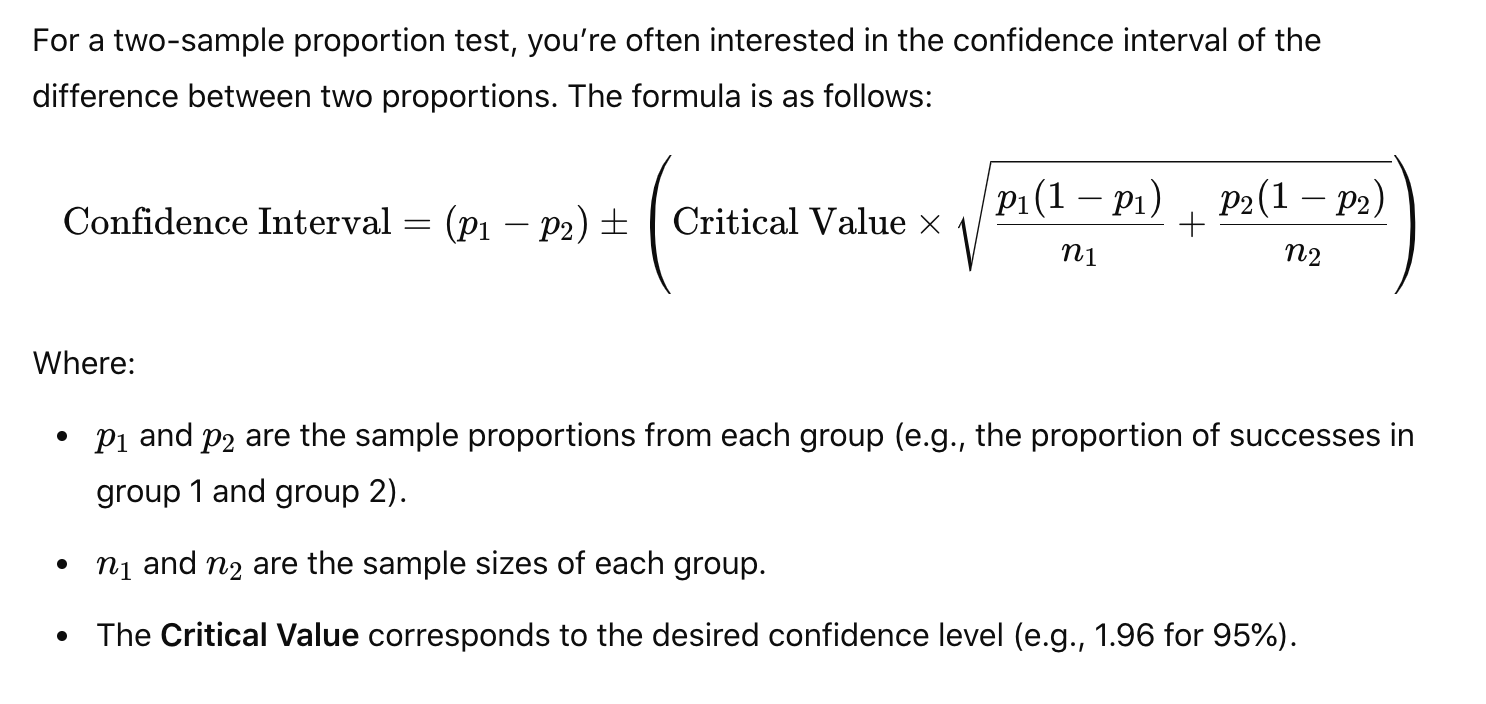
How to choose MDE?

* Check past experiments and use historical lift
* Assess whether the change is meaningful and valuable
* Show sensitivity analysis with a range on MDE and corresponding runtime for reference
  + If sample size is too large 🡪 increase MDE
  + If MDE is too high 🡪 test a more significant change
* Consider sequential test

1. Ramp up plan and launch test
2. Smoke test: control vs. treatment balanced?
3. Analysis: metrics stats sig (calculate p-value) and confidence interval
   1. Ratio metrics: two sample proportion z-test or chi-square test (if underlying population follows binomial distribution, which can be approximated by normal distribution when n is large; known population standard deviation)

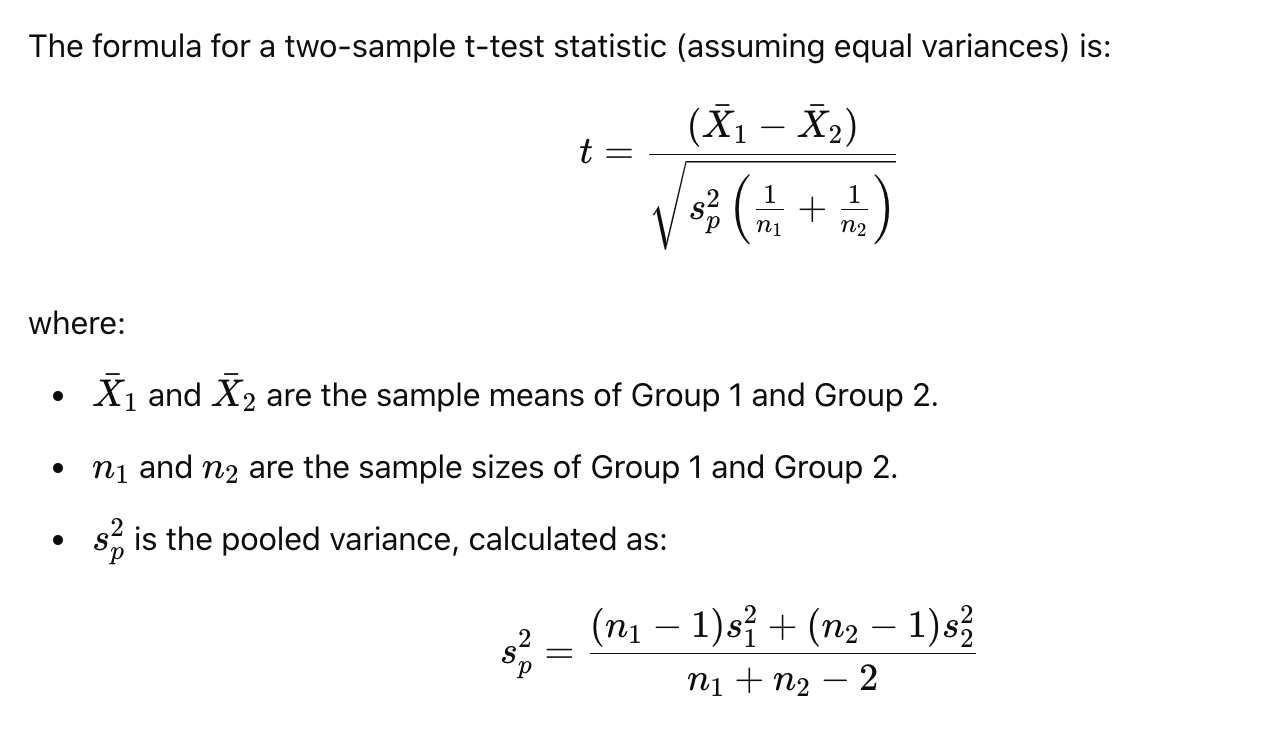


Confidence interval:

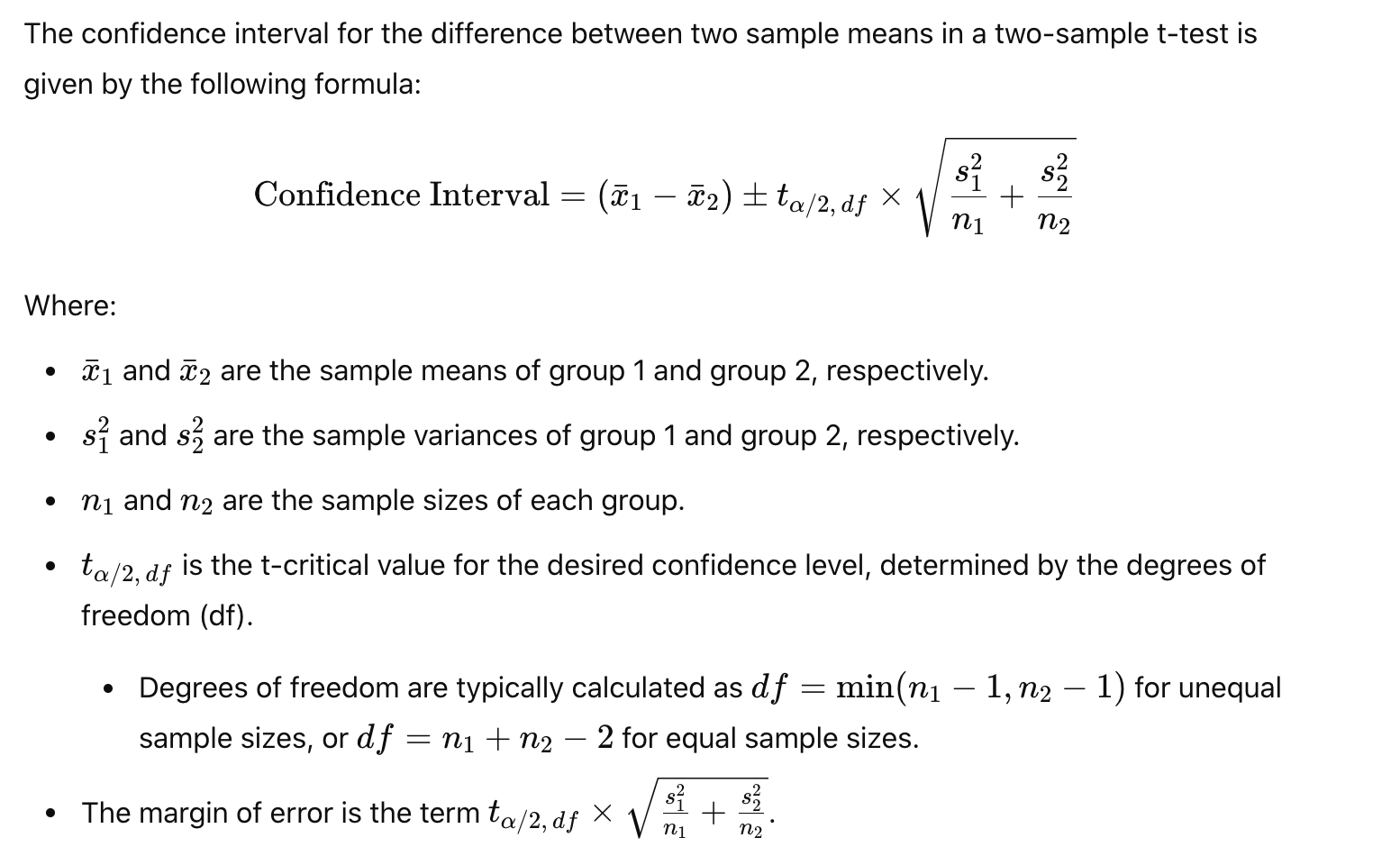


* 1. Mean metrics: two sample mean t-test

Test statistics follow a Student’s t distribution (if underlying population are normally distributed; population standard deviation is unknown)



Confidence interval:



1. Make recommendations:
   1. If meet GA criteria, then GA
   2. If not meet GA criteria: need to dig further
      1. Further slice and dice to locate segmentation with good performance; compare results for each segment and then decide which one to GA
      2. Funnel analysis to pinpoint any frictions to remove or any iterations for further optimization
      3. Short term vs long term effect: global holdout
      4. Regress on certain metrics (revenue): set threshold

**Mixed results:**

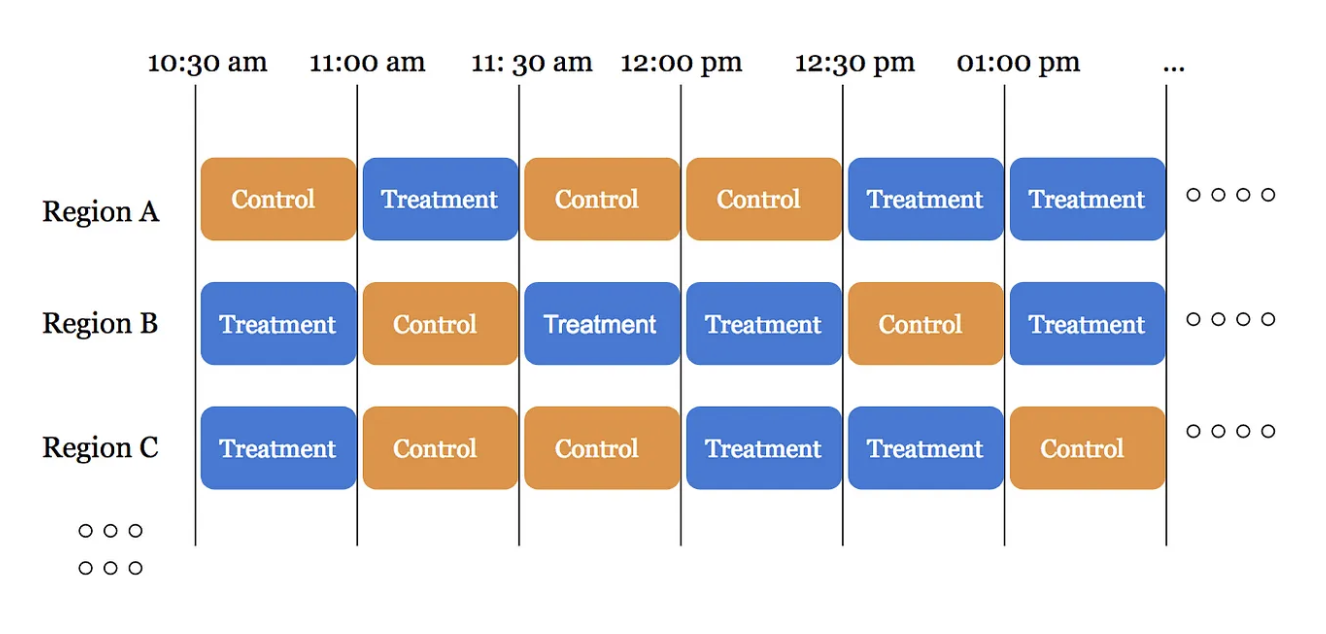
* Statistically significant vs. practically significant:
  + Balance with cost of building the feature and ongoing engineering maintenance after launch
* Deep dive on metrics to see why it is dropping.
  + Decompose the metrics
* Segmentation: if result for certain cohort is good, then launch to that cohort
* Trade-off between metrics: estimate total loss in metric and see whether this is within tolerance; need global holdout to monitor trends; check the over time performance for first week cohort
  + We calculate the impact: creating an impact table, for example, if primary metric is % user sharing is up but recipient / sharer is slightly down, then we look at usually revenue metrics i.e. GNARR impact. If leadership is okay with the impact, then proceed with launch. (This is after carefully assessing that experiment is strategically important to move forward and there are no biases.)
  + Short term vs. long term trade off: trade short term (revenue) loss for long term retention gain, which drives long term monetization gain
* Product decision

**Product recommendation:**

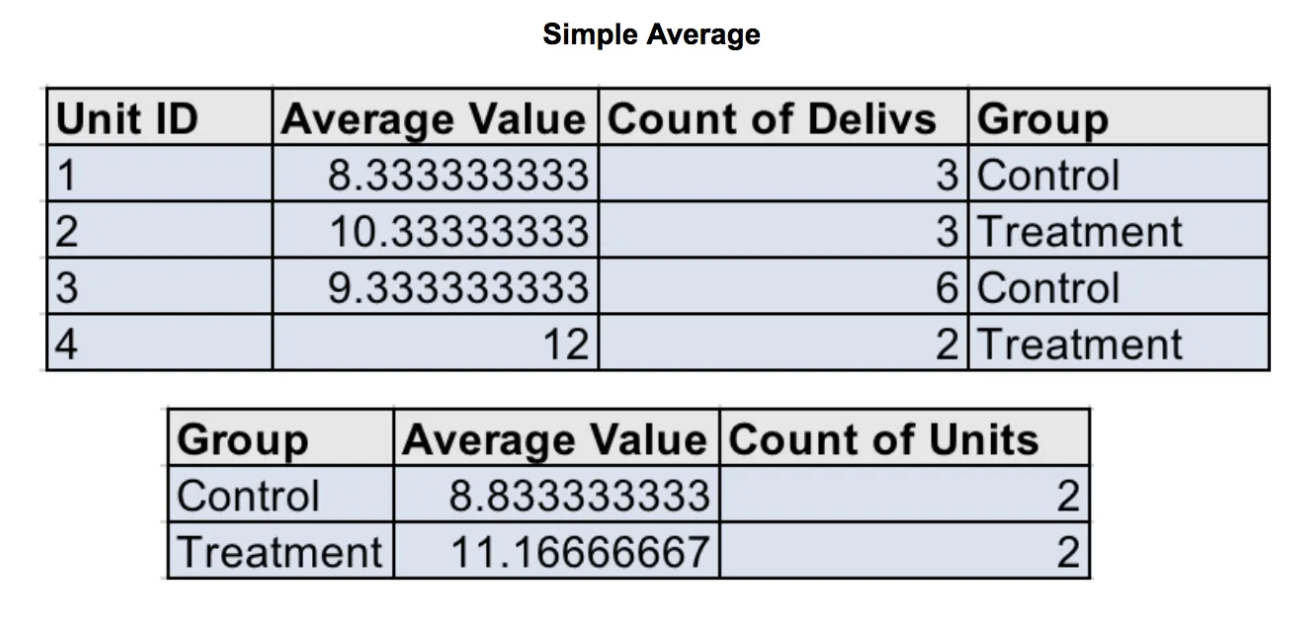
* + **Discoverability / entry points**: make product features easier to find and accessible to users
* **Increase visibility**: place features in highly visible areas of app or website i.e. homepage, header or navigation bars etc
* **Multiple entry points**: offer different ways for users to access key features through menu, search bars, pop-ups, quick links on homepage
  + **Lifecycle journey**: engage users at every stage of their journey, from onboarding to regular use
* **Emails and Notifications**: Implement personalized emails, push notifications or in app messages for key moments in the customer journey.
* **Deep linking**: bring users directly to relevant page of the app or website. This reduces friction by eliminating extra steps
* **In product marketing campaigns**: utilize banners, prompt, or notifications to prompt features;
  + On-screen tutorials or tooltips for onboarding
  + Contextual prompt: show prompts based on user activity and behavior. i.e. show prompt when user is about to complete an action
  + **Optimize UX flow**: reduce frictions and make it easier for users to take actions
* **Simplify funnel steps**: If you notice a high drop-off at a particular step, assess whether it's essential. Consider merging steps or eliminating unnecessary actions
* **Simplify interactions**: one click actions for key tasks i.e. purchase, subscribe, activating new features
* **UI design change**: make CTA more noticeable i.e. increase font size, use contrasting color or place in more prominent location to attract more clicks; set default that benefit the user experience or drive higher revenue; copy change to increase conversion i.e. use action driven words
  + **User segment targeting**: tailor the experience to each user based on their behavior and needs
* **Personalization**: personalize product experience based on user behavior and preferences and demographics. i.e. personalize onboarding experience with profiling response; personalized product recommendations based on customer demographics, life stage or lifestyle
* **Cross sell and upsell**: offer relevant cross sell or upsell suggestions for frequent or high value / loyal buyers
* **Loyalty based recommendations**: Reward users who have a history of frequent purchases with exclusive offers or early access to new features.
  + **User education**: engaging onboarding process that walks users through key features with tooltips (explain features when users hover over them), checklist (guide users to complete important actions) or short interactive in app tutorials
  + **Expand eligibility or scale wins**:
* **Broaden feature access**: expand successful features to other regions, user groups, product SKU or platform/device; i.e. expand to existing users, to mobile / desktop signups, remove eligibility requirements to increase exposure
* **Increase funnel exposure**: place important features earlier in the user journey to capture attention sooner
  + **Pricing strategy improvement**
* **Flexible pricing** models: tiered pricing or subscription models that align with different user needs
* **Discounts and promotions**: offers or personalized discounts to incentivize conversion
* **Trials**: introduce free trials to attract new users
  + **Retention improvement**:
* **Retention program**: Implement a rewards or loyalty program where users can earn points or benefits for using the product frequently, which can be redeemed for discounts, premium features, or exclusive content.
* **Reactivation campaigns**: target dormant users with re-engagement campaigns through email, message or notifications, offering incentives such as discounts, exclusive content or reminders of product’s value

**Risks and limitation:**

1. Cannibalization: into other features, cannibalize advertising
2. Revenue trade off:
   1. Trade short term revenue loss for long term retention gain, which drives long term monetization gain
   2. More ads means short term revenue gain but ads fatigue and user experience may cause user churn and leads to long term revenue loss
3. Data maturity: TPCR (daily plot), conversion rate, churn rate takes longer time to mature
4. Seasonality effect: User behaviors change in holiday seasons.
   1. Global holdout group to measure real effect
   2. Calculating runtime, I can scale the traffic depending on seasonality. For example, to run a test in Dec, instead of using Nov traffic directly, but scale Nov data in the same ratio as last year Nov/last year Dec
5. Novelty effect:
   1. Global holdout and measure impact over time to see if effect is diluted.
   2. Control vs. treatment: daily plot over time to see effect decreases
6. Network effect: Treatment effect in treatment group may spill over to control group. The difference would underestimate effect
   1. Market-based or geo-based test
      1. Test the feature in specific geographic regions or markets to limit exposure and cross-interaction. This approach can help approximate the impact without directly affecting users outside the test group.
      2. We choose two locations with similar demographics where we make sure key metrics such as DAU, MAU, time spent and ads revenue are roughly similar across the two locations pre-test.
      3. Use Diff in Diff:
      * Find two countries with similar historical trends
      * Treatment and synthetic control: one country is set as synthetic control and one country set as treatment
      * Calculate diff = post – pre for each country
      * Compare the diff of two countries; the difference of diffs would be actual effect of treatment
      1. Used commonly in marketing as Geo lift test
      2. Test at country level: pick certain countries first to lower risk. Prefer Australia and New Zealand as they are good proxy of US (geographically isolated, decent population, English speaking country, similar culture/user behavior to US); Canada is too close to US; UK needs to comply to GDPR in Europe
      3. Limitations: test performed in specific locations, which may not fully represent the larger population. Results may not generalize universally to all users; parallel trends assumption: the trend in outcomes for control units would be the same as the trend for treatment units
   2. Cluster based randomization:
      1. Instead of randomizing at the individual level, this approach groups users into clusters (such as friend groups, geographic regions, or communities) and assigns the whole cluster to either the treatment or control group. By grouping users who are likely to interact, you minimize the spillover between the two groups.
      2. Per community random assignment: a common mathematical approach to defining user communities is to model the relationships between users with a [social graph](https://en.wikipedia.org/wiki/Social_graph), and then apply graph partitioning algorithms to find isolated, non-interacting groups.
      3. Limitations: Difficult to detect stats sig result as cluster-based randomization is usually under powered. We have smaller sample size and bigger variance for the cluster-based test.
   3. Switchback test:
      1. Randomize spatial units ranging from city or region blocks to time intervals
      2. In switchback testing, the core concept is that we switch back and forth between control and treatment algorithms in a certain region at alternating time periods. For example, in the SOS pricing example, we switch back and forth every 30 minutes between having SOS pricing and not having SOS pricing. We then compare the customer experience and marketplace efficiency between the control time bucket and treatment time bucket metrics
      3. The coarser these experimental units, the stronger the protection against interference bias in your effect estimates.
      4. However, the cost is increased variance and smaller sample size in your estimators, which would cause test to be lower in power; violate independence: sandwich estimator of variance



* + 1. Alternating time intervals between global control and global treatment was a successful strategy for Lyft: Alternating time interval experiments perform well if the effect size is very large or we just happen to get lucky with temporal fluctuations
    2. How to decide on granularity: 30min time block at city level:
    - Don’t want time region units that are too small 🡪 risk introducing bias (more short rides)
    - Don’t want time region units that are too large 🡪 risk having large margin of error (variance) and sample size small to run test for long
    1. This means we’re conducting our statistical tests on the average valuesof control and treatment time-region units**,** rather than the individual values of control and treatment deliveries.



**Measure impact**:

1. Correlation: Pearson correlation, regression model, lift analysis, causal pseudo analysis (propensity score matching)
2. A/B test
3. Pre post analysis (monitor key metrics pre and post intervention)
4. Global holdouts (1% or 5% users as global holdout)

we already launch this feature for 9 months, How to measure success for this feature?

* Global holdouts
* Pre post analysis to monitor key metrics pre and post intervention
* Propensity score matching

**Improve product:**

Brainstorm ideas

1. Analyze user journey map (PM)
   1. User awareness: how often do people see it? Is it large enough? Do people hover over it? 🡪 increase size of component, use prompt / pop up window, send emails and push notifications; discoverability / entry points not easy to find: make it more visible, add more entry points, promote or market it in product
   2. Reduce friction: users not posting or drop off 🡪 fill in template, pre-generate content; remind users to post with emails or notifications
2. Segment users into groups and understand their behavior: (DS)
   1. New users who is not aware of feature 🡪 increase size of UI component and provide detailed instructions
   2. Existing users and never used it: send survey and address needs
   3. Existing users and not posted in a while: not get enough attention and lost motivation 🡪 send reminders, remove negative comments
   4. Users have intention but abandoned the post: worry about controversy
   5. Users post frequently: study active users to understand what makes them more engaged and compare with inactive users.
      1. ML models: characteristics and browsing behaviors with logistic regression or tree-based model for feature importance. If we found certain user group is less active, then focus on that group and identify problems i.e. awareness, competitors 🡪 launch marketing campaign to turn inactive users to active

Prioritize ideas: proportion of users impacted by each idea, select the most cost-effective ideas (low hanging fruit)

Measure success: success metrics for experiment design

**------------------------------------------------------------------------------------------------------------------------------------------------------------**

**Metrics for notification**

Reach: notification delivery rate (a high delivery rate ensures that notifications are getting through, whether through push, email or in-app channels)

Engagement:

* Open rate (view rate) (How compelling the notification’s content or subject line is)
* Click through rate = click CTA / open (whether content is relevant)
* Conversion rate = % users who complete desired action after clicking on notification (make purchase, complete signup; measure the ultimate success of notification in driving meaningful actions i.e. revenue, account creation or engagement)
* Bounce rate: % users who dismiss or ignore notification without taking any actions
* Unsubscribe rate: % users who opt out or unsubscribe from receiving future notifications.

Reasons for notification CTR drop:

1. Seasonality:
   * Users might have distracted by holidays, vacations, or other events in June
2. Product change:
   * Change content in messaging
   * Timing or frequency of notifications: less optimal frequency. Sending them too frequently that users might have started ignoring them, leading to decline in CTR.
   * New design could make CTA less visible or harder to interact with
3. Segmentation:
   * Notifications were sent to a broader or less engaged user base in June compared to May;
   * Total notification sent is high, spammy
   * Notifications increased heavily on notification type with lowest CTR: time sensitive as notification for live or story?

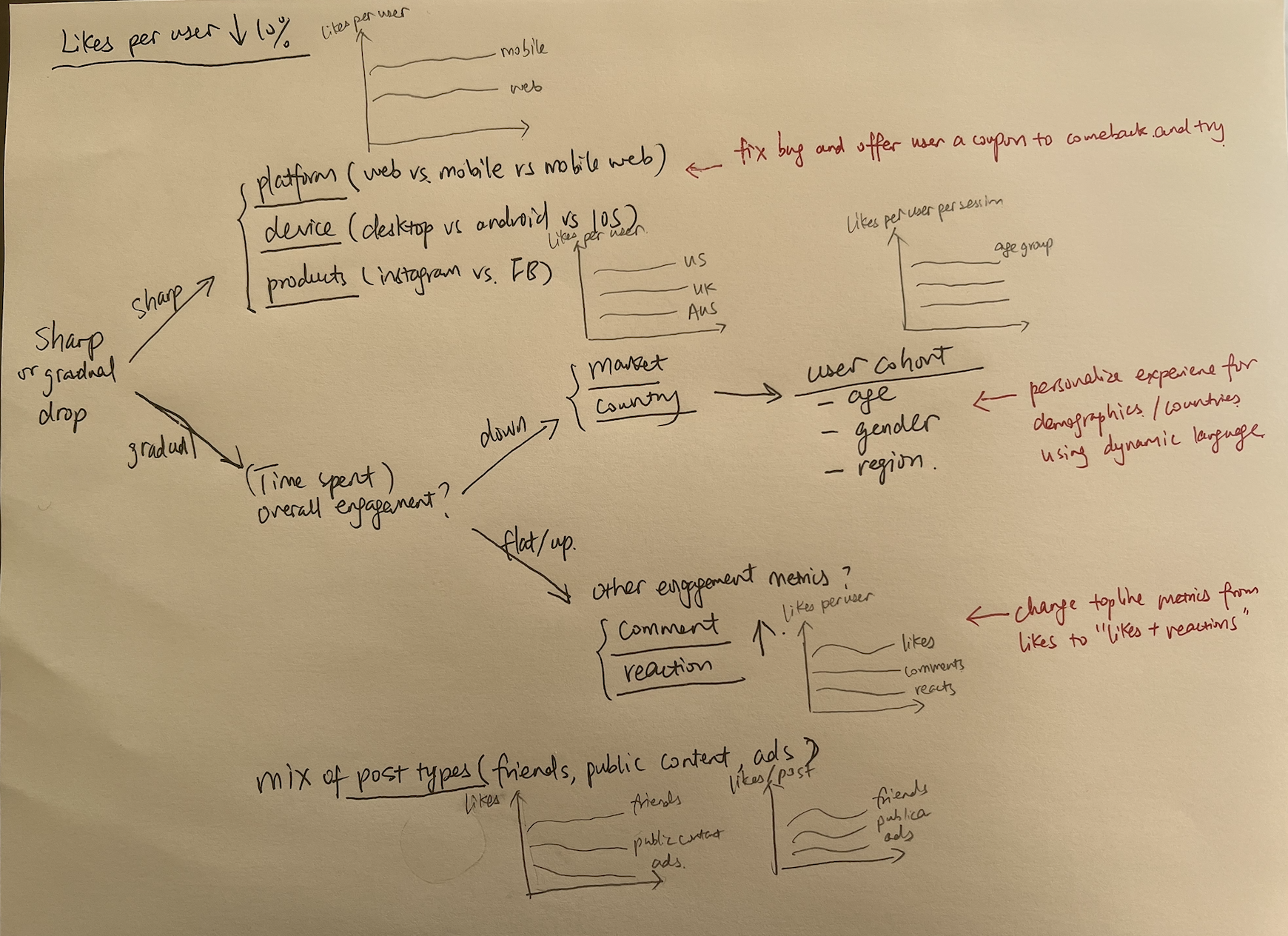
**Comment distribution**

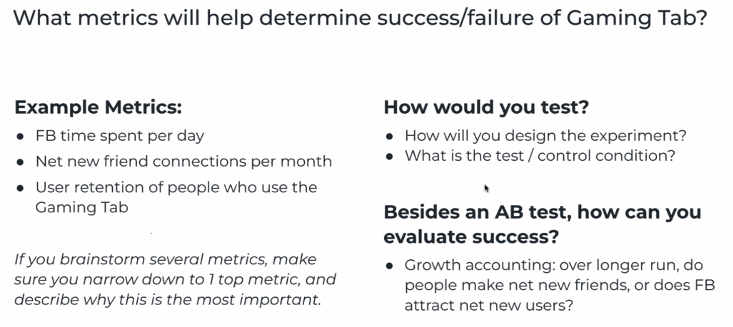
* **Mean**: Typically, between **1–5 comments per user** (since the distribution is usually skewed).
* **Median**: Often around **0–1 comments per user** (because many users are passive).
* **95th Percentile**: Typically **20–200** or more comments, reflecting the highly active users who contribute the majority of comments on the platform.

Average # comments per user increase from 2 to 3:

* **Changes to the content** (more relevant or controversial stories).
* **Platform feature changes** (e.g., a better commenting system: simplify the process, make comment box more visible, or added upvoting / downvoting features, notifications).
* **Change in user base a shift in user demographic could increase comment volume. (younger users join recently;** more active users recently**)**
* **Incentivization or gamification** that encourages commenting.

Example: Likes per user on app drop?



****

[How Meta tests products with strong network effects](https://medium.com/@AnalyticsAtMeta/how-meta-tests-products-with-strong-network-effects-96003a056c2c#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6IjczZTI1Zjk3ODkxMTljNzg3NWQ1ODA4N2E3OGFjMjNmNWVmMmVkYTMiLCJ0eXAiOiJKV1QifQ..XXAQIO6wu0kJh63j3R1TGZ0uOUXzIdAmDrikG7686g1gM_DteIQkKIkp5DJATAeIZSarh0vrkrrC2KrBN9nLSJ54vadPcOIXT_VWM_BS119sKHX04SFO2Luda6Hdz3_0wc06QJqK2f2FU8IN44LJxTcZnv5AcvsrsFrNkU7GhvkysBNfJWD0PmbkR4fzaOY_6bfI4BCERPjoNRXLfkMhV9XijbHaAPK_TqNg7wxMcSSO73F72IWSdgGXJs0nftCM7nJc8lSSZdKaILWyrKvE8TOZZxmBCoXPN5Zi4e6hCn38CXuMUMb7tXZu0xjYAGGoO0mc1I7iov9nvbGNeoQAuA)

[Why spillover effects bias your AB testing results and ways to overcome them](https://medium.com/@weonhyeok.chung/why-spillover-effects-bias-your-ab-testing-and-ways-to-overcome-them-e7f06efd0b56)

**Value to Users / Business?**

**What’s the business goal?** Acquisition, Activation, Engagement, Retention, Monetization?

**Pros:**

* Business:
  + improve product experience to make it easier to sell products
  + more sales
  + increase ads spend
* User:
  + improve customer experience to make users more willing to purchase products or store visit
  + more purchases / store visits
  + improve overall user engagement and retention
  + meaningfully connect users, increase friend connections

**Cons:**

* Business: cannibalize advertising
* User: Spam & fraud, cannibalize time spent on FB app

**Considerations:**

* Survey or market research?
* Gap in the market or competitor?
* Opportunity sizing
* Align with company goals