**Spotify Personalized Playlist Algorithm**

**Problem :**  
The product team believes the new algorithm will increase user engagement by making people listen to more music.

**Null and Alternative Hypotheses:**

* **H₀ (Null Hypothesis):** The new algorithm does **not** increase weekly listening time.
* **H₁ (Alternative Hypothesis):** The new algorithm **increases** weekly listening time.

**Reflection Question:**

1. What makes this initial request problematic for hypothesis testing?

The request is unclear because it doesn’t specify exactly how much of an increase matters as meaningful, like what does meaningful increase in listening looks like in mathematically form 5% or 20% increase

1. How would you refine this into a specific, measurable research question?

Does the new playlist algorithm increase average weekly listening time per user by at least X minutes compared to the current system

1. What additional information would you need from the product team?

Target user segments, and the timeframe for measurement

1. Would a one-sided test or two-sided test be best in this scenario? Why?

Appropriate because we only care if listening time increases.

1. What are the potential risks of choosing a one-sided test?

If a **one-sided test** only looks for an **increase** in listening time, but the algorithm reduces listening time, then the test **fails to detect a real effect in the opposite direction.** This is called a **Type II error**, also known as a **false negative.**

**Key Consideration:**

The platform has 100 million active users worldwide across different subscription types (free and premium). Given here is Spotify's user segmentation:

* **Premium subscribers:** 40% of user base
* **Free tier users:** 60% of user base
* **Geographic distribution:** 30% US, 40% Europe, 30% Rest of World
* **Device usage:**70% mobile, 20% desktop, 10% other

**Reflection Question:**

1. What factors might affect listening time that need to be controlled for?

* User subscription types (free vs premium).
* Geographic location (US, Europe, Rest of World).
* Device usage (mobile, desktop, other).

1. How would you ensure representative sampling across user segments?

Randomly assign users to control (current algorithm)

1. What potential biases could emerge from improper sampling?

Non-random assignment and Over-representing a region could skew results.

Setting Decision Criteria:

The product team says implementing the new algorithm will be costs so the change must be significant meaningful.

**Reflection Question:**

1. How would the implementation cost affect your choice of:
   * Significance level (α): 0.05
   * Minimum effect size: decide what increasing cost as significant meaningful in term of math.
   * Required statistical power: 80-90% detect meaningful changes matter.
2. What specific business metrics would you need to define "significant improvement"?

Average monthly or weekly listening time

1. What secondary metrics could be monitored for potential negative impacts?

User fulfillments, User playlist.

Data Collection After Experiment:

Data gather after running experiment for a week on control group and treatment group.

* Weekly listening time (primary metric)
* **Control group (n=500,000):**
  + Mean: 118.5 minutes
  + Standard deviation: 44.2 minutes
  + 95% Confidence Interval: [117.8, 119.2] minutes
* **Treatment group (n=500,000):**
  + Mean: 122.3 minutes
  + Standard deviation: 43.8 minutes
  + 95% Confidence Interval: [121.6, 123.0] minutes
* **Reflection Question:**

1. How would you interpret these confidence intervals in non-technical language

 Control group: 118.5 minutes/week (CI: 117.8–119.2)

 Treatment group: 122.3 minutes/week (CI: 121.6–123.0)

1. What does the overlap (or lack thereof) between confidence intervals tell us?

This shows treatment users listen about 4 minutes more on average. The small overlap in confidence intervals suggests the difference is likely real.

Analysis and Interpretation

**Scenario 1: Your test comparing the new and old playlist algorithms returns the following results:**

* p-value = 0.047
* Average listening time increase: 8 minutes per week
* Sample size: 500,000 users per group
* Test duration: 2 weeks

**Reflection Question:**

1. How would you explain this p-value to non-technical stakeholders? What common misinterpretations should you help them avoid? What other factors might you include? -> p-value = 0.047, effect = +8 minutes/week

There is statistical evidence that the new algorithm increases listening time. A p-value of 0.047 means there’s about a 4.7% chance the observed difference happened randomly if the algorithm had no effect. This does not mean there’s a 95% chance the algorithm works—just that the result is unlikely under the null hypothesis.

1. If running the same test with 5 million users per group returned a p-value of 0.001, but showed only a 2-minute increase in listening time, how would this change your interpretation? -> sample size is very large, but effect is small

A p-value could be extremely low even if the increase is just 2 minutes/week. Decisions should balance **statistical significance** and **business impact**.

**Scenario 2: Your test results show:**

* p-value = 0.08 (above traditional 0.05 threshold)
* Effect size: 12 minutes increased listening time per week
* High variance in the data
* Strong positive trends in user satisfaction metrics
* Summary: p-value = 0.08, effect = +12 minutes/week, high variance, positive user satisfaction trends

**Reflection Question:**

1. How would you approach making a recommendation in this case? What factors beyond the p-value would influence your decision?

Don’t rely on p-value alone. Consider the large practical effect and positive secondary metrics. explain that while statistical evidence is weaker, the practical benefit could be meaningful.

1. How would you handle stakeholder pressure to make a definitive "yes/no" decision based solely on statistical significance?

Emphasize difference between statistical significance and business relevance. Suggest further testing or a staged rollout.

1. What additional data or analyses might help provide more context for decision-making?

Longer experiments, segment analysis, engagement metrics