Tracksuit Technical Take-Home

Optimising Survey Structure to Minimise Respondent Cost

Problem Overview:

At Tracksuit, each customer category must receive approximately 200 qualified respondents per month. However:

- Categories have different incidence rates (probability a respondent qualifies).
- Categories have different survey lengths.
- Each respondent can only complete up to 8 minutes (480 seconds) of survey content.
- The demographic exposure must remain nationally representative.
- The objective is to minimise total respondents surveyed (cost).

This is fundamentally an allocation and optimisation problem under:

- Total survey length ($\leq$ 480 seconds)
- Probabilistic qualification
- Service-level guarantees ($\geq$ 200 qualified per category)

Baseline Analysis (Naive Strategy):

If we run each category independently, the total required number of respondents (naive cost):

$$\text{Required respondents} = \sum_{i=1}^{k} 200/p_i$$

Where K is the total number of categories and p_i is the incidence rate of the *i-th* category. And $200/p_i$ represents the expected number of respondents required to obtain 200 qualified respondents for category i.

Based on the dataset, the naive total is approximately 40,879 respondents. This approach is clearly inefficient because it ignores the fact that respondents can qualify for and complete multiple categories within the 8-minute time constraint.

Key Insight:

The key insight is that the true cost driver is low-incidence categories. For example, a category with a 10% incidence rate requires approximately 2,000 respondents to achieve 200 qualified completes, whereas a category with an 80% incidence rate requires only about 250 respondents. This means that total survey cost is largely dominated by the hardest (lowest incidence) category within each survey structure. Mathematically, the total cost can be expressed as $Total\ cost = \sum_{group} \max(200/p_i)$, since each group must collect enough respondents to satisfy its most difficult category. This principle forms the core insight behind the algorithm.

Methodology:

I model the problem as a bin-packing optimisation where each survey has a maximum capacity of 480 seconds, and each category represents an item with a weight equal to its category_length_seconds. To prioritise efficiency, categories are first sorted by difficulty, measured as $200/p_i$ meaning categories with lower incidence rates (harder to fill) are placed first. Using a Greedy First-Fit strategy, each category is sequentially assigned to the first survey group whose total allocated time does not exceed 480 seconds; if no existing group can accommodate it, a new survey group is created. This approach ensures that harder categories are secured early while maintaining the interview length constraint.

The resulting allocation produced 19 survey groups. A detailed summary of each group — including the number of categories, total survey time, and deterministic respondent requirement — is provided in **Table 1 (see survey_groups_summary.csv)**.

Then, we perform a deterministic cost estimate. For each survey group, the required number of respondents is determined by the most difficult category within that group, calculated as $N_{group} = \max\left(\frac{200}{p_i}\right)$. This ensures that all categories in the group reach at least 200 qualified respondents. Summing across all groups yields a total estimated cost of approximately 11,069 respondents, representing a 73% reduction compared to the naive baseline.

Probabilistic Reality:

In practice, qualification is random. If we set $N = \frac{200}{p_i}$, then 200 represents only the expected number of qualified respondents, not a guaranteed outcome. To account for this uncertainty, the solution must be validated probabilistically rather than relying solely on deterministic estimates. I therefore implemented a Monte Carlo validation approach. For each survey group, I simulated N respondents and, for each category, generated qualified counts using $q_i \sim \text{Binomial}(N, p_i)$. I then checked whether all categories achieved at least 200 qualified respondents. This process was repeated 500 times per group. To further reduce the risk of under-delivery, I added a 20% buffer, setting $N = 1.20 \times \max\left(\frac{200}{p_i}\right)$.

Validation Metrics:

For each survey group, I evaluated the success rate (the probability that all categories reach at least 200 qualified respondents), the mean minimum qualified count across simulations, and the 5th percentile of the minimum qualified count (**see validation_results.csv**). These metrics ensure statistical robustness, provide service-level reliability, and significantly reduce the operational risk of failing to meet contractual respondent targets.

As shown in **Table 1**, the optimised approach substantially reduces total respondent cost compared to the naive baseline, while maintaining high validation performance. Monte Carlo simulations demonstrated a strong probability of success across all survey groups. Additionally, the provided dataset does not include demographic structure, which may influence real-world implementation considerations.

Table 1: Results Summary	
Metric	**Value**
Naive cost	~40,879
Optimised deterministic cost	~11,069
With 20% buffer	~13,293
Reduction vs naive	~67.5–73%

Trade-Off Discussion

The solution involves several key trade-offs: a lower buffer reduces total respondent cost but increases the risk of under-delivery, while a higher buffer improves reliability at the expense of higher cost. Similarly, creating more survey groups may allow tighter time packing but can increase total cost due to duplication of high-incidence maxima across groups. The combined greedy construction and Monte Carlo validation approach strikes a balance between operational simplicity, computational efficiency, and statistical robustness, delivering a practical and reliable solution within real-world constraints.

Conclusion:

I designed a survey allocation policy that guarantees at least 200 qualified respondents per category, respects the 8-minute interview constraint, minimises the total number of respondents required, and validates performance probabilistically using Monte Carlo simulation. The solution achieves approximately a 70% cost reduction compared to the naive baseline while maintaining high reliability. The approach is transparent in its logic, reproducible in its methodology, extensible to more complex real-world constraints, and strongly aligned with business objectives around cost efficiency and service-level delivery.