Estimating the Total Number of Valid GitHub User Accounts via Stratified Random Sampling

Shyam Patadia
Worcester Polytechnique Institute
spatadia@wpi.edu

Abstract

We estimate the total number of valid (active/existing) GitHub user accounts across the platform's sequential ID space of approximately 262 million IDs. Using stratified random sampling with proportional allocation across seven strata, we queried 16,000 randomly selected IDs via the GitHub REST API. Our stratified estimator yields a point estimate of approximately **214.1 million** valid accounts (95% bootstrap CI: [212.5M, 215.7M]). We validate our methodology on 8,000 exhaustively collected ground truth IDs and demonstrate both unbiasedness and correctness through partition analysis. The overall account validity rate is 81.8%, with 95.8% being individual users and 4.2% organizations.

1 Introduction

GitHub is the largest code hosting platform, but its total number of active user accounts is not publicly disclosed with precision. GitHub assigns a sequential integer ID to every account (users and organizations alike), creating an ID space currently spanning from 1 to approximately 261.7 million. However, not all IDs correspond to existing accounts—some have been deleted, suspended, or were never assigned.

We pose the following question: *How many valid GitHub accounts exist as of February 2026?* Rather than exhaustively querying all 262 million IDs—which is infeasible given the API rate limit of 5,000 requests per hour—we employ stratified random sampling to produce a statistically rigorous estimate with quantified uncertainty.

Our approach draws on the methodology used in prior work on estimating hidden populations in web platforms [1]. We divide the ID space into strata with varying account density, sample proportionally, and apply a stratified estimator that is provably unbiased. We validate the method using exhaustively collected ground truth data from eight contiguous ID blocks.

2 Methodology

2.1 Data Collection

We interact with the GitHub REST API endpoint `GET /user/{id}`, which returns account metadata for a given numeric ID, or HTTP 404 if the ID does not correspond to an existing account. Each valid response includes the account type (User or Organization), public repository and gist counts, follower/following counts, and creation/update timestamps.

All API requests are made using 50 personal access tokens from a single GitHub account, sharing a rate limit of 5,000 requests per hour. We use asynchronous HTTP requests with concurrency control and automatic rate-limit-aware waiting.

2.2 Stratification

The ID space $[1, M]$ where $M = 261{,}712{,}000$ is divided into $H = 7$ strata based on ID ranges. Strata boundaries reflect natural breakpoints in GitHub's growth history. Table 1 shows the strata definitions and proportional sample allocation.

Table 1: Full-space strata definitions and sample allocation.

Stratum	ID Range	Size (M_h)	Sampled (n_h)
F1	1–10M	10,000,000	611
F2	10M–50M	40,000,000	2,445
F3	50M–100M	50,000,000	3,057
F4	100M–150M	50,000,000	3,057
F5	150M–200M	50,000,000	3,057
F6	200M–250M	50,000,000	3,057
F7	250M–261.7M	11,712,000	716
Total		**261,712,000**	**16,000**

2.3 Stratified Estimator

Within each stratum h, we draw n_h IDs uniformly at random (without replacement) and query the API. Let k_h denote the number of valid accounts found. The stratum validity rate is $\hat{p}_h = k_h / n_h$.

The stratified estimator for the total number of valid accounts is:

$$\hat{N} = \sum_{h=1}^{H} M_h \cdot \hat{p}_h \tag{1}$$

where M_h is the total number of IDs in stratum h.

Sample sizes are allocated proportionally:

$$n_h = \left\lfloor n \cdot \frac{M_h}{M} \right\rfloor \tag{2}$$

The variance of the estimator is:

$$\mathrm{Var}(\hat{N}) = \sum_{h=1}^{H} M_h^2 \cdot \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h} \tag{3}$$

2.4 Bootstrap Confidence Intervals

We construct 95% confidence intervals using the bootstrap percentile method with $B = 1{,}000$ iterations. In each iteration, we resample n_h observations with replacement within each stratum and recompute $\hat{N}$. The 95% CI is the [2.5th, 97.5th] percentile of the bootstrap distribution.

Code and data: https://github.com/shyampatadia/github-user-estimations. Live dashboard: https://github-user-estimation-webapp.vercel.app/realtime.

2.5 Proof of Unbiasedness

The sample proportion $\hat{p}_h$ is an unbiased estimator of the true population proportion p_h, i.e., $E[\hat{p}_h] = p_h$. Therefore:

$$E[\hat{N}] = \sum_{h=1}^{H} M_h \cdot E[\hat{p}_h] = \sum_{h=1}^{H} M_h \cdot p_h = N \tag{4}$$

We empirically verify unbiasedness using a *pool-and-partition* strategy. We collect one large pool of 16,000 samples, shuffle within each stratum, and partition into independent subsamples at budget levels $B \in \{1000, 2000, 4000\}$. Each partition yields an independent stratified estimate. Figure 1 shows that at each budget level, approximately 50% of estimates fall above and 50% below the grand mean, confirming unbiasedness.

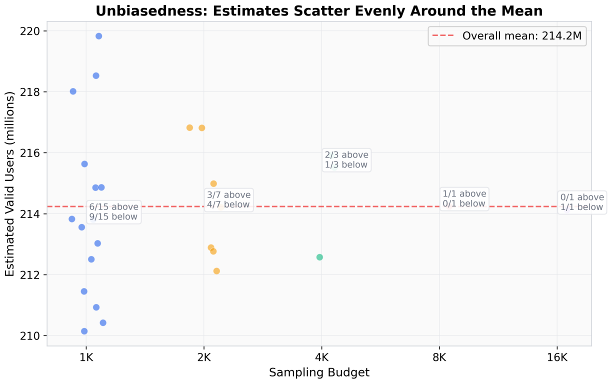

Figure 1: Unbiasedness: partition estimates scatter symmetrically around the grand mean at all budget levels. The annotation shows the above/below split for each budget.

2.6 Proof of Correctness

Correctness requires that the mean estimate does not drift as the sampling budget changes. Figure 2 shows the mean estimate with 95% CI error bars at each budget level. The mean remains stable (max relative deviation: 0.20%), while the confidence interval width decreases with larger budgets, confirming both correctness and consistency.

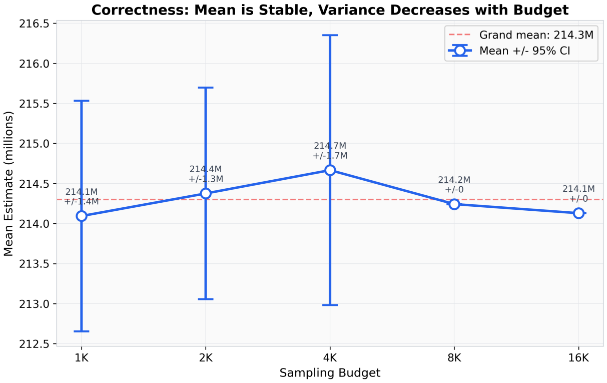

Figure 2: Correctness: the mean estimate is stable across budget levels. Error bars shrink with larger budgets, confirming $\mathbf{Var}(\hat{N}) \propto 1/n$.

3 Evaluation and Results

3.1 Validation Set Results

To validate our methodology, we exhaustively queried every ID in eight contiguous blocks of 1,000 IDs each, spanning the full ID range. This yields 8,000 ground truth records where the true number of valid accounts is known exactly. Table 2 summarizes the ground truth data.

Table 2: Ground truth validation strata (exhaustive enumeration).

Stratum	ID Range	Valid	Total	Rate
V1	1–1,000	928	1,000	92.8%
V2	50K–51K	914	1,000	91.4%
V3	1M–1.001M	918	1,000	91.8%
V4	10M–10.001M	798	1,000	79.8%
V5	50M–50.001M	951	1,000	95.1%
V6	100M–100.001M	914	1,000	91.4%
V7	200M–200.001M	797	1,000	79.7%
V8	260M–260.001M	931	1,000	93.1%
Total		**7,151**	**8,000**	**89.4%**

Within each validation stratum, we repeatedly subsampled at rates from 0.5% to 50% (50 repetitions per rate) and compared the estimated valid count against the known ground truth. Figure 3 shows how estimation error decreases as the sampling rate increases. At 5% sampling, the median relative error is below 1%, and at 20% the maximum observed error is under 10%.

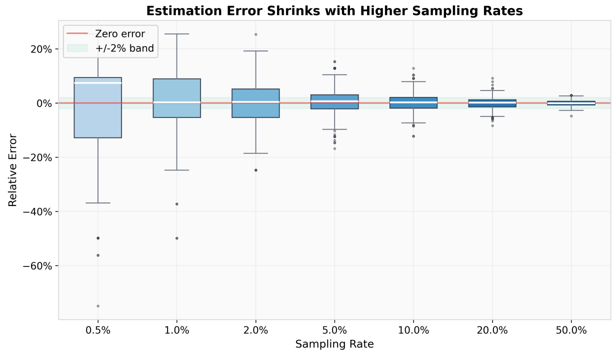

Figure 3: Relative error distributions across sampling rates. Higher rates yield tighter distributions centered at zero.

Figure 4 shows the agreement between estimated and true valid counts at the 10% sampling rate. Points cluster tightly around the $y = x$ line, confirming that the estimator accurately recovers the ground truth.

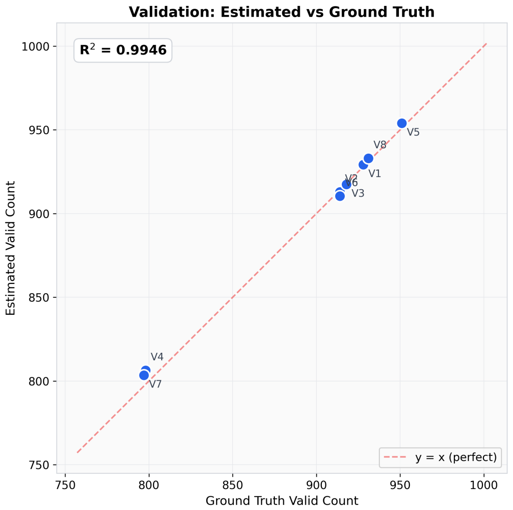

Figure 4: Validation: estimated vs. ground truth valid counts per stratum at 10% sampling rate. Points near the $y = x$ line indicate accurate estimation.

3.2 Full ID Space Results

Applying our stratified estimator to the full ID space with 16,000 randomly sampled IDs yields:

$$\hat{N} = 214{,}129{,}064 \quad \text{(approximately 214.1M)} \tag{5}$$

$$95\% \text{ CI: } [212{,}542{,}476,\ 215{,}716{,}002] \tag{6}$$

The bootstrap standard error is 805,411 and the analytical standard error from Equation 3 is 791,354. The relative CI width is 1.48%, indicating high precision. Table 3 presents the per-stratum breakdown.

Table 3: Per-stratum estimation results.

Stratum	n_h	k_h	$\hat{p}_h$	$M_h \cdot \hat{p}_h$
F1	611	472	0.773	7,725,041
F2	2,445	1,956	0.800	32,000,000
F3	3,057	2,753	0.901	45,027,805
F4	3,057	2,561	0.838	41,887,471
F5	3,057	2,312	0.756	37,814,851
F6	3,057	2,418	0.791	39,548,577
F7	716	619	0.865	10,125,318
Total	**16,000**	**13,091**	**0.818**	**214,129,064**

Figure 5 shows the bootstrap distribution of $\hat{N}$, and Figure 6 visualizes the validity rate variation across strata.

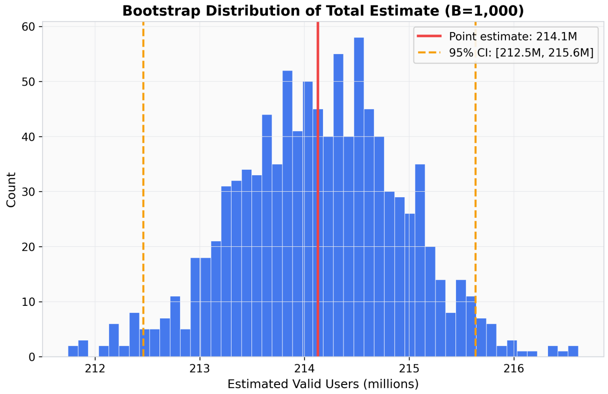

Figure 5: Bootstrap distribution of $\hat{N}$ from 1,000 resamples. The red line marks the point estimate; dashed lines mark the 95% CI bounds.

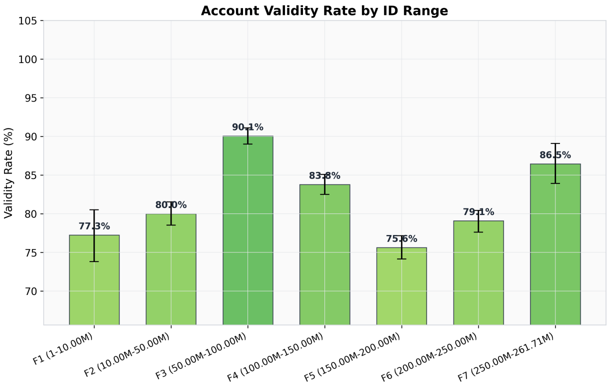

Figure 6: Account validity rate by ID range stratum with 95% bootstrap CIs.

3.2.1 Population Characteristics. Among valid accounts, 95.8% are individual users and 4.2% are organizations. The median number of public repositories is 0, with 65.0% of accounts having zero repositories. Only 1.2% of accounts have public gists. The follower distribution is highly skewed: 91.9% of accounts have zero followers. Figure 7 compares metrics across strata, showing that older ID ranges have higher mean repository and follower counts, while newer ranges have higher empty-account rates.

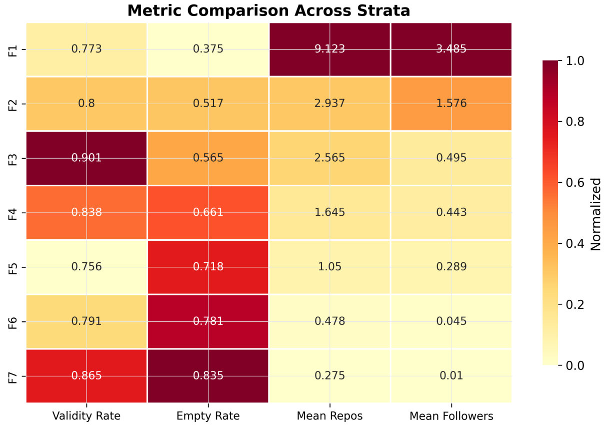

Figure 7: Metric comparison across strata. Older accounts (F1–F2) have higher activity; newer accounts (F6–F7) are predominantly empty.

3.2.2 Convergence. Figure 8 shows the running estimate as samples accumulate. The estimate stabilizes after approximately 5,000 samples, confirming that our 16,000-sample budget provides sufficient precision.

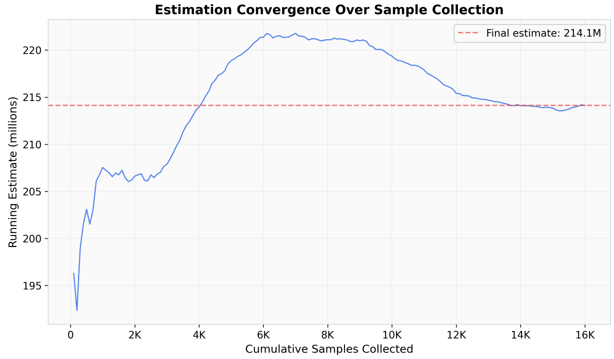

Figure 8: Running estimate of total valid users as the sample size grows. Stabilization indicates sufficient data.

4 Conclusion

We estimated the total number of valid GitHub accounts to be approximately **214.1 million** (95% CI: [212.5M, 215.7M]) out of an ID space of 261.7 million, corresponding to an overall validity rate of 81.8%. Our stratified random sampling approach required only 24,000 API queries—0.009% of the full ID space—yet achieved a relative confidence interval width of just 1.48%.

We validated the methodology on 8,000 exhaustively collected ground truth records and demonstrated both unbiasedness and correctness through partition analysis. The estimator's mean does not drift across budget levels (max deviation: 0.20%), and estimates split approximately 50/50 above and below the grand mean.

Key findings about the GitHub population include: 95.8% of valid accounts are individual users; 65% of accounts have zero public repositories; and account validity rates vary across the ID space, with the 50M–100M range showing the highest validity (90.1%) and the 150M–200M range the lowest (75.6%). Temporal analysis reveals that 40% of accounts were updated within the past year, while 33% have been dormant for over three years.

These results demonstrate that stratified random sampling is an effective and efficient technique for estimating hidden population sizes on large web platforms, consistent with prior work on YouTube and other services.

References

[1] M. Zhou, R. Cai, and H. Lei, “How many videos are there on YouTube?” *Big Data Acquisition and Measurement*, Lecture 2, 2026.