YouTube’s Violative View Rate Methodology

A Statistical Assessment

Arnold Barnett

Massachusetts Institute of Technology

November 2021

This evaluation was commissioned and funded by Google. The conclusions and opinions expressed are exclusively those of the author.

I. Introduction

Hundreds of hours of content are uploaded to YouTube every minute.[1] YouTube thrives by tapping into its ecosystem of creators, viewers, advertisers, but it considers some videos objectionable and wants them not to appear at its website. The company strives to limit the videos available to viewers to those that meet the standards for safety articulated in its "Community Guidelines."[2]

YouTube has developed specific rules about the types of content not permitted on its platform. These rules aim to exclude videos with unacceptable levels of violence, harassment, threats to children, and hate speech.[3] The company has also developed automated methods which aim to detect videos that violate these policies so that they can be removed. According to its latest "Community Guidelines Enforcement Report," YouTube removed nearly 9.6 million videos between January and March 2021.[4]

YouTube has set up procedures continuously to monitor the presence of violative content on its platform. Although initially automated, these procedures send potentially violative videos for human evaluation. But this process takes time, and during that time users might be exposed to violative content. YouTube is very much interested in estimating the extent of such exposure of such content to improve further its detection systems. It estimates this frequency through a metric called the **"Violative View Rate" (VVR)**. The VVR is the percentage of views of videos on the YouTube platform that violate its Community Guidelines.

Google asked me to review and report on the sampling and statistical methodologies behind the YouTube calculation of the VVR metric.[5] To facilitate that review, YouTube

[1] https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#detecting-violations.

[2] https://www.youtube.com/howyoutubeworks/policies/community-guidelines/.

[3] https://www.youtube.com/howyoutubeworks/policies/community-guidelines/.

[4] Between January and March 2021, the latest period of report, YouTube removed 2.2 million "channels," or collections of videos, which included more than 59 million individual videos; and another 9.6 million individual videos. https://transparencyreport.google.com/youtube-policy/removals.

[5] I have not evaluated certain aspects of YouTube's VVR methodology, including the details of their machine learning-based classifier and the quality of the human rater reviews. I did not have access to YouTube's underlying data,

provided me access to both publicly available external documents and internal documents describing the VVR approach. Statisticians, data scientists, and product managers at the company spoke with me about their modeling efforts and answered my questions. In pursuing my assignment, I considered the relationship between what YouTube does and what is possible using sophisticated and well-established statistical methods. I also performed some calculations that involved modifications of the YouTube methodology, to see whether any “tweaks” to YouTube’s methodology might yield a meaningful increase in accuracy.

Based on my review, I conclude that YouTube’s methodology for estimating the VVR is thoroughly sensible and statistically sound. Moreover, its approach has several salutary features that aid its efforts to strengthen and enforce its Community Guidelines. In the remainder of this paper, I provide the basis for my conclusion.

I begin the discussion in Section II with a brief “primer” on statistical sampling. Then in Section III, I outline YouTube’s approach to estimating the VVR. I discuss my main reactions to YouTube’s methodology in Section IV and, in Section V, offer a calculation pertaining to an alternative to the YouTube model. I offer a few final comments in Section VI.

II. A Very Brief Overview of Statistical Sampling

Through random sampling, one can use information about some members of the population to draw inferences about all of them. There is not one single way to conduct random sampling but instead there is a family of possibilities. When the aim of the sampling is to estimate a particular number (a parameter), it is generally desired that the estimate be *unbiased*, meaning that it is on average correct and not systematically too high or too low. Even if unbiased, however, the estimate could well differ from the actual parameter value because of the luck of the draw, in which case it suffers *sampling error*. The *margin of error*

code, or programs for selecting their samples and calculating the VVR metrics, although I did review extensive descriptions of their approach and calculations.

characterizes the level of imprecision in a given parameter estimate caused by random sampling fluctuations.

Simple Random Sampling

In the simplest form of random sampling, all members of the population have the same chance of being selected into the random sample. (Not surprisingly, this method is called *simple random sampling*). Such sampling always yields unbiased parameter estimates. However, it can also yield a sizable margin of error in the parameter estimate.

For example, suppose that a highly controversial policy is supported by 9.5% of the population (and opposed by the remaining 90.5%), and a poll is performed to estimate the support for the policy based on a simple random sample of 200 people. (These numbers are chosen to make the discussion easier to follow, and do not reflect real violation rates among YouTube views.) The random poll will, on average, uncover 19 supporters out of 200 people canvassed and thus generate a correct estimate of 9.5%. However, it would not be unusual for the randomly-drawn sample to contain considerably more or fewer than 19 supporters. The margin of error in this setting would be 4.1 percentage points, meaning that there is about a 95% chance that the support estimate based on the sample will be within 4.1 percentage points of the correct value of 9.5% (i.e., between 5.4% and 13.6%). Might a lower level of statistical uncertainty be possible at the same sample size? In some cases, the answer is yes.

Suppose, for example, that the population in question is sharply polarized on the policy question, and breaks into two distinct groups in terms of support: 90% of the population is in group A where support is 5%, while the remaining 10% is in group B where support is 50%. A random sample of 200 people would on average contain about 20 from group B (10% of 200), but the actual number from that group could easily be 16 or 25. And the percentage of supporters in that small sample from B could well differ appreciably from 50%. This is the main reason that the overall support estimate can oscillate around 9.5% by several percentage points. But, if the polarization is recognized at least in general terms, might that information point the way to an estimate that is unbiased but less vulnerable to sampling error?

Stratified Random Sampling

A different sampling strategy might be beneficial here. If it is correctly surmised that the main uncertainty about overall support for the policy relates to group B, then why not arrange that exactly 10% of the sample of 200 is allocated to group B (i.e., 20 from group B and 180 from group A)? That strategy would eliminate the component of sampling error that arises because, by the luck of the draw, group B might be overrepresented or underrepresented in the sample. Such a “proportional representation” scheme—which breaks the original population into two strata and takes separate random samples from each—is called *proportional stratified random sampling*. Once the sampling results are at hand, one would take a weighted average of the results for a population-wide estimate, with A getting nine times the weight of B because group A is nine times as large as group B. This stratified sampling would again be unbiased, yielding on average an overall support level of 9.5%. But it would cut the margin of error from 4.1 percentage points to 3.6, a reduction of about 12%.

However, this stratified scheme still estimates the support percentage within group B based on a sample size of only 20. That estimate is somewhat volatile: at a support level of 50%, the number of supporters among 20 randomly-chosen members of group B would bounce around the mean of 10 the way the number of heads would vary around 10 over 20 tosses of a fair coin. Suppose instead that one took a sample of size 40 from group B and 160 from group A. Then the support percentage within group B would be estimated based on 40 individuals rather than 20, and thus would be less vulnerable to sampling error. Again, one would weight the group A results nine times as much as those from group B to estimate the population-wide support level.

This more general form of stratified random sampling is again unbiased, and on average yields a support estimate of 9.5%. But the margin of error drops to 3.4 percentage points. Compared to simple random sampling at the same sample size, the margin of error has fallen 17% (i.e., from 4.1 to 3.4). And this reduction involves no increase in sampling effort, just some redirection of sampling resources towards the places where they can be maximally informative.

There is a further issue. If one is monitoring support for the policy over time, one might revise the sampling split between A and B if there is evidence of changes in one or both groups, perhaps based on results in the most recent poll. That would be an example of *dynamic stratified sampling*. Sampling theory offers approximations of the optimal division of sampling resources between two groups (or among several), based on existing estimates of parameter values within the groups. The aim remains to obtain an unbiased estimate of the population parameter with the least amount of sampling error.

As we will discuss, YouTube performs a sophisticated form of dynamic stratified sampling to estimate the proportion of views of objectionable videos. YouTube's approach is firmly rooted in reliable and fully developed statistical methodology.

III. YouTube's Calculation of the VVR

As noted, YouTube's goal is to estimate the fraction of views on its platform that violate its Community Guidelines. (In statistical parlance, the population of interest is all YouTube views). A viewing arises when one person watches a video for any length of time. The emphasis on views accommodates the commonsensical notion that a violative video that is widely watched is more harmful than another such video that is rarely seen.

Using a machine-learning classifier, YouTube has devised a score for each video tied to its characteristics that relates to the likelihood that the video is impermissible.[6] While the score itself is not the probability that the video violates the Guidelines, it arises from the premise that, the higher the score, the greater the chance of a violation. The scores can vary from 0 to 1.

YouTube then moved to an exercise in stratified sampling, based on creating non-overlapping ranges for the video scores. The aim was to devise a set of strata such that the probability of violation would not vary much within a given stratum but would vary appreciably

[6] Based on my conversations with YouTube, I understand that YouTube has a number of classifiers, each with different goals. The one that it uses to generate the scores in its VVR methodologies is designed to cast a wide net in searching for violative videos. The details about how classifier scores are determined are outside the scope of this evaluation. However, I recommend that YouTube periodically evaluate whether revisions to its scoring rules for videos might yield lesser uncertainty in its estimates of the VVR.

across strata. YouTube decided that it would create five strata, namely, lowest risk, 2nd lowest, 2nd highest, highest, and "no score available."[7] Random sampling would then take place among the views in each stratum, meaning that a given video's probability of selection would be proportional to the number of people who viewed it. Each sampled video would be evaluated by a trained person for its adherence to the Community Guidelines. Ultimately, the VVRs calculated from the five strata would be combined for an estimate of the overall VVR.

Given the decision to create five strata among videos that received scores, two questions arise:

- How should the range from 0 to 1 be divided into four distinct strata? For example, the lowest-risk stratum would include scores from 0 to X, but what is X?
- Given a total sample size for the number of views from the five strata, what size sample should be drawn from each?

A full answer to these questions would require much immersion into statistical sampling theory.[8] In general terms, one might start with an "educated guess" of how the probability of violation varies with score, and then use that approximation and a computer to create four score ranges plus a "no score" stratum, coupled with five sample sizes. (The ranges and sample sizes chosen would be those that yield the lowest overall margin of sampling error in the VVR estimate, given the initial guesses.) As indicated in our statistical primer, strata in which the expected VVR is especially low would receive a lesser share of the sampling than their share of the population (e.g., 30% of the population might get 20% of the amount of sampling). This makes sense because even small sample sizes are highly accurate when the VVR is very low. (Indeed, if there were a stratum with a VVR of zero, even a sample of size one would correctly reflect that VVR.) Instead, sampling resources

[7] I understand that videos for which the classifier did not return a score are, for example, videos uploaded very close to the time that sampling was done.

[8] A good place to begin that immersion would be the classic book *Sampling Techniques* by Cochran (Wiley, 1997), especially Chapter 5.

gravitate towards strata with higher VVRs as long as the VVR is less than 50% (which, fortunately, does not remotely happen).

Once the initial scheme based on educated guesses is in place, actual VVRs would be estimated for the various strata. Those numbers could then be used to reallocate future sampling resources to the strata where they would do the most to reduce sampling error.[9] Less emphasis would be placed on changing the boundaries among the four strata for scored videos.

The investigation with sampled views takes place daily at YouTube. In accordance with the principles of dynamic stratified sampling, the sample sizes in the five ranges are revised each day based on actual VVR rates in those ranges over the 90 preceding days. Such revisions could be important should there be, say, an increase in violative videos over time in low-score ranges. Even if the initial scoring rules became less predictive over time, the dynamic sampling procedure would be expected to blunt or even prevent any increases in overall sampling error.

I would suggest, however, that YouTube consider whether using VVRs averaged over the past 90 days is preferable to doing so for, say, the past 30 days. In situations in which VVR’s are changing rapidly, a 90-day average might not be especially illuminating about what is happening now. Indeed, YouTube might do well routinely to calculate the VVR averages for recent periods of varying length: if they are all similar, that would indicate that the VVR is stable over time.

YouTube recently released to the public its estimates of the VVR for recent calendar quarters. **Table 1** presents estimates of several recent quarterly VVRs, coupled with the margins of error for those estimates. The quarterly VVR is an average of 91 daily VVR estimates, weighted by the number of views on the platform each day (which can vary).[10]

[9] Indeed, it is theoretically possible that the observed violative rate would be higher in the low-risk strata than in the higher-risk ones. In that event, future sampling would tilt towards the allegedly low-risk strata and the descriptions of the strata would be changed. It is important to recognize that, under YouTube’s sampling approach, any initial misconceptions about risk as a function of video characteristics would not cause increased sampling error in future sampling. That is a strength of YouTube’s approach.

[10] If the number of views of videos on the YouTube platform were the same each day, the quarterly VVR would be a simple average of the daily VVRs.

As shown in Table 1, the VVR rates have been stable over the last five quarters, while the margins of error for the quarterly estimates have been only a small fraction of the estimates themselves (only 0.01 percentage points, which is about 1/18 of the estimate in the last two quarters). These results make clear that there is very little uncertainty around the estimated level of policy violations on YouTube. That outcome is prima facie evidence that YouTube's sampling methods are highly successful.

Table 1. YouTube's VVR by Quarter, Q1 2020 - Q1 2021

Calendar Quarter	Estimated VVR	Margin of Error (at 95% Confidence Level)
Q1 2020	0.19%	± 0.015 percentage points
Q2 2020	0.20%	± 0.015 percentage points
Q3 2020	0.17%	± 0.010 percentage points
Q4 2020	0.18%	± 0.010 percentage points
Q1 2021	0.18%	± 0.010 percentage points

Source: https://transparencyreport.google.com/youtube-policy/views

IV. General Assessment of YouTube's Approach to Estimating VVR

For several reasons, I see much merit in what YouTube is doing to reach an accurate VVR estimate. First and foremost, YouTube's analysts explored the statistical literature and recognized that the estimation problem they faced could fruitfully be handled by a powerful existing methodology, namely, dynamic stratified sampling. All too often, analysts try to reinvent the wheel and wind up with something that is anything but round. That did not happen at YouTube.

Moreover, YouTube realized that getting reliable results about VVR would require large sample sizes. Because violative videos are rare, even a stratified sampling effort targeted towards videos with high scores would yield noisy results at the moderate sample sizes that would suffice in other contexts. Despite the need for human involvement, YouTube samples thousands of video views per day, which works out to hundreds of thousands of sampled views per quarter. Given such sample sizes and an adaptive methodology, YouTube can pick up on changes over time in both overall violation rates and the kinds of videos where violations are increasing or decreasing.

Having human reviewers for sampled videos is an excellent idea. Evaluating videos for objectionable content cannot easily be automated, especially with “adversarial actors” working to evade any predictable guidelines produced by algorithms. Furthermore, YouTube goes to great lengths to prevent inconsistent or unreliable assessments by its reviewers, which include:

- Allowing raters to consult more experienced colleagues and quality assurance staff for assistance when they are unsure of the proper decision;
- Waiting 14 days after the close of a given quarter to incorporate the human evaluation into the calculation of the VVR metric (a process YouTube calls “windowing”), which also allows time for errors to be corrected before being included in calculations;
- Encouraging raters to seek help from specific language experts to assist with videos in a particular language; and
- Providing feedback to raters based on ongoing testing, audits, and evaluation of raters’ decisions.

Because of these activities, it is unlikely that either of the two kinds of potential errors – violative videos being classified as acceptable or acceptable videos being classified as violative – are common. But I believe YouTube would do well to monitor the process regularly, to be sure that its reliability does not diminish over time.

YouTube is well aware that VVR measures the size of a problem rather than cures the problem. The sampling exercise cannot identify which particular views outside the sample are violative; it can simply indicate approximately how many such views there are and whether that number is diminishing over time. That information, however, is highly valuable in its own right.

In one respect, I thought it worthwhile to investigate an assumption in the VVR analysis. YouTube's analysts have divided views of videos with scores into four strata (with an additional stratum for views of videos for which the classifier did not provide a score), believing that working with a larger number of strata would only minimally reduce sampling error. Though I considered that conclusion plausible and consistent with sampling theory, I thought it should be subject to an empirical test. Such a test is discussed in the next section.

V. A Test of YouTube's Stratification Methodology

To assess whether five strata might be too few, I worked with a hypothetical population of views that is similar to the one YouTube actually faces. The question was whether using additional strata would appreciably reduce the level of sampling error in the VVR estimate. As shown in Table 2A, the actual fraction of violative views in our hypothetical population is 0.2%, but the rate varies somewhat across the score ranges. Most views are in the lowest-risk category, and the VVR increases steadily as the category becomes riskier.

Consistent with YouTube's actual approach, I used a total sample of 4,000 random views for a hypothetical day,[11] and I followed YouTube's method for allocating the sampling across the five strata. The results appear in Table 2B. Note that the lowest-risk stratum contains 80% of all views but gets only about half the sampled views, while the very small highest-risk stratum gets a disproportionate share of those views. The outcomes in the individual strata will all be subject to sampling error, the level of which depends on both their sample sizes and their actual VVRs. As noted at the bottom of the table, the overall VVR estimate is on average correct at 0.2%, with a standard deviation of 0.054 percentage points

[11] A sample of 4,000 views a day means that YouTube samples nearly 1.5 million views each year. This is a substantial sample size.

Table 2A: Characteristics of a Hypothetical Population of Views Broken into Five Strata

	Share of Population	Violative Viewing Rate
Lowest Risk	80%	0.05%
Low Risk	10%	0.50%
Middle Risk	5%	1.0%
High Risk	1%	5.0%
No score available	4%	0.25%
Total	100%	0.20%

Table 2B. YouTube's Sampling Allocations Across the Five Strata

	Share of Population	YouTube Optimal Sample Sizes	
	%	#	%
Lowest Risk	80%	2,098	52.5%
Low Risk	10%	828	20.7%
Middle Risk	5%	584	14.6%
High Risk	1%	256	6.4%
No score available	4%	234	5.9%
Total	100%	4,000	100%

Expected Estimate of VVR	0.20%
Expected Standard Deviation	0.054 percentage points

But suppose that the population of views with known scores had been broken into eight strata rather than four. For example, suppose the lowest-risk stratum with an average VVR of 0.05% was divided into two halves, with an average VVR of 0.025% in the new “lowest lowest risk” stratum and of 0.075% in the other half (see Table 3A). Again, using a daily sample size of 4,000, the YouTube approach yields the allocations across strata shown in Table 3B. In each former category, a greater sample size is assigned to the higher-risk substratum than to the lower-risk one. This improved "targeting" would tend to reduce sampling error. But by how much?

As shown, the sampling will again on average yield an estimated VVR of 0.20%, while the expected standard deviation drops to 0.052 percentage points. But that decline is only 4% of the expected standard deviation based on five strata (0.054 percentage points). I view this outcome as supportive of YouTube's judgment that five strata offer sufficient accuracy. One can always reduce sampling error by increasing the number of strata: further dividing this population into sixteen strata would cut the expected standard deviation even more. But the analysis gets more unwieldy as strata proliferate, so it makes sense to limit their number. In my judgment, YouTube has done so sensibly.

Table 3A. Characteristics of the Hypothetical Population of Views Broken into Nine Strata

	Share of Population	Violative View Rate
Lowest Risk (Part 1)	40.0%	0.025%
Lowest Risk (Part 2)	40.0%	0.075%
Low Risk (Part 1)	5.0%	0.30%
Low Risk (Part 2)	5.0%	0.70%
Middle Risk (Part 1)	2.5%	0.95%
Middle Risk (Part 2)	2.5%	1.05%
High Risk (Part 1)	0.5%	3.04%
High Risk (Part 2)	0.5%	6.96%
No score available	4.0%	0.25%
Total	100%	0.20%

Table 3B. YouTube's Sampling Allocation Across the Nine Strata

	Share of Population	YouTube Optimal Sample Size	
		#	%
Lowest Risk (Part 1)	40.0%	760	19.0%
Lowest Risk (Part 2)	40.0%	1,316	32.9%
Low Risk (Part 1)	5.0%	329	8.2%
Low Risk (Part 2)	5.0%	501	12.5%
Middle Risk (Part 1)	2.5%	291	7.3%
Middle Risk (Part 2)	2.5%	307	7.7%
High Risk (Part 1)	0.5%	103	2.6%
High Risk (Part 2)	0.5%	153	3.8%
No score available	4.0%	240	6.0%
Total	100%	4,000	100%

Expected Estimate of VVR	0.20%
Expected Standard Error	0.052 percentage points

I also considered whether, assuming five strata, one might achieve a lower expected standard deviation with different boundaries between strata. For example, what if the lowest risk category contained 70% of all views rather than 80%? To put it briefly, I obtained no improvement over the boundaries that YouTube would have set. In all, I saw no way to modify YouTube's sampling procedure that would improve on its combination of accuracy and transparency.

VI. Conclusion

As Table 1 shows, only about one ten-thousandth of YouTube views violate its Community Guidelines. But that estimate is highly accurate because YouTube does extensive sampling on a daily basis, and does so using a highly sophisticated methodology that combines careful human evaluation with advanced statistical theory. I wholeheartedly endorse what YouTube is doing.

As noted, knowing that the VVR is about 0.01% does not reveal which particular views outside the sample violate the Guidelines. One cannot eliminate violative views without immediately examining every single one. But the VVR offers a realistic estimate of the magnitude of the problem and how that magnitude might be changing over time. And that knowledge is of great value to both YouTube and the huge community served by YouTube.

About the author

Arnold Barnett is the George Eastman Professor of Management Science and Professor of Statistics, MIT Sloan School of Management. Professor Barnett is an expert in applied statistics and mathematical modeling. He has written numerous articles and chapters on a broad range of topics, including statistical analysis, new directions for statistics and computing, and statistical misconceptions in empirical studies. He is the author of the textbooks *Applied Probability* and *Applied Statistics*, and his papers have been published in peer-reviewed journals such as the *Journal of the American Statistical Association*, *Operations Research*, and *Risk Analysis.*

Professor Barnett’s early work on homicide was presented to President Ford at the White House, and his analysis of U.S. casualties in Vietnam was the subject of a column by William F. Buckley. He has received the President’s Award and the Expository Writing Award from The Institute for Operations Research and the Management Sciences (INFORMS) and the President’s Citation from the Flight Safety Foundation for “truly outstanding contributions on behalf of safety.” He has written op-ed pieces for *The New York Times*, *The Wall Street Journal*, *The Boston Globe*, and *USA Today*. He has been honored frequently for outstanding teaching by students at the MIT Sloan School of Management, and he was described as the school’s best faculty member by *BusinessWeek*.

Professor Barnett holds a B.A. in physics from Columbia University and a Ph.D. in mathematics from MIT.