Estimating the Initial Mean Number of Views for Videos to be on YouTube's Trending List

Yao Yao
MSDS 6370 Section 403

127 days of metadata for **top trending videos** from U.S., Canada, UK, Germany, and France are collected (Nov 14, 2017 to Mar 20, 2018) [Kaggle/YouTube API]

The **trending algorithm** is derived from internet social interactions [Google]

Metadata for video views, shares, comments, and likes are all **correlated**

Publishers could disable embedding, comments, and likes, which **skews the dataset**

Views is the true indicator to where all subsequent user interaction could result

Videos can stay on the trending list for **multiple days**: only concerned about **initial view count** where duplicate subsequent observations are removed

Dataset: a video to start trending a **few hours since publish** or from as **old as 2010**

Data Cleaning

Limitations of data collection time frame: cannot capture when exactly a video starts trending

Older videos may have been trending before the data collection resulting in **inflated view count**

Data is collected **once a day**: time is relative to video **publish** and when a video starts trending

Remove outliers 1.5 times the inter-quartile range (Figure 1)

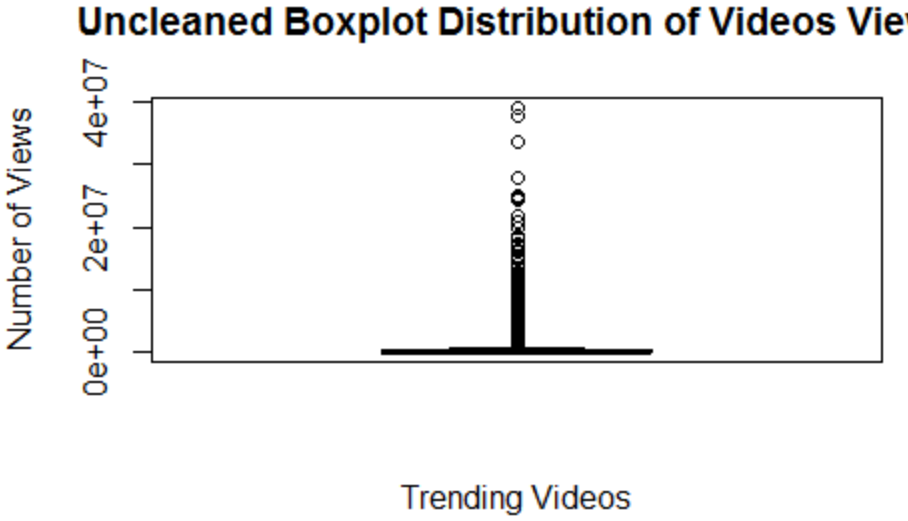

Figure 1a: Uncleaned Box Plot Distribution of Video Views (N = 49841)

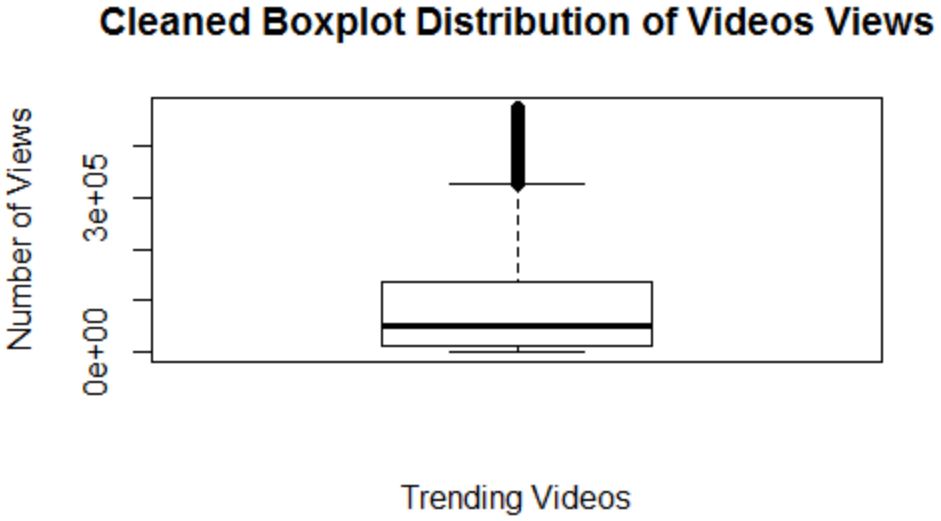

Figure 1b: Cleaned Box Plot Distribution of Video Views (N = 44506)

The **FPC adjustment is ignored** because the cleaned dataset is less than 10% of the original population without duplicate observations (89.29%)

Simple Random Sample

The MEANS Procedure

Analysis Variable : views

Minimum	Lower Quartile	Mean	Median	Upper Quartile	Maximum	Std Error	Std Dev
223.0000000	14840.00	224403.96	61786.00	198996.00	39118664.00	3275.13	731176.26

Figure 2a: Uncleaned Quartile Distribution of Video Views (N = 49841)

The MEANS Procedure

Analysis Variable : views

Minimum	Lower Quartile	Mean	Median	Upper Quartile	Maximum	Std Error	Std Dev
223.0000000	12030.00	93891.70	48041.50	137346.00	475127.00	519.5546202	109607.56

Figure 2b: Cleaned Quartile Distribution of Video Views (N = 44506)

Using the 95% confidence interval threshold, where margin of error is 5000 views:

$$n_{0,srs} = \frac{(Z_{\alpha/2}S)^2}{(moe)^2} = \frac{(1.96 * 109607.56)^2}{5000^2} = 1846.09 \approx 1847 \text{ samples}$$

Simple random sample where sample size is 1847:

Mean estimate: **91,932.82 views**
Standard error: **2,490.11 views**

True mean of 93,891.70 views is within standard error range

Strata Exploration

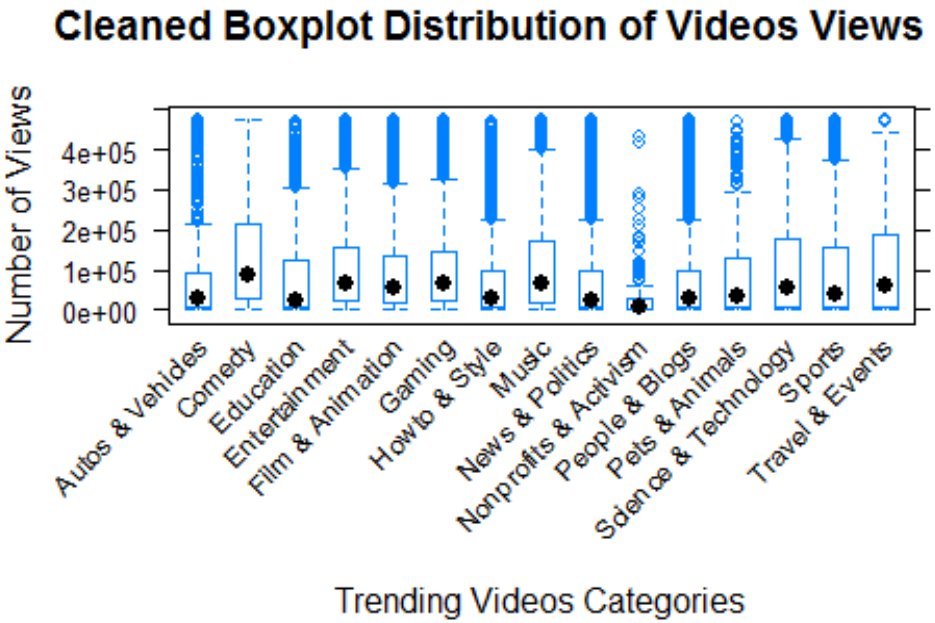

Figure 3a: Stratified YouTube Views Box Plot Distribution by Category

Most Obs:'Entertainment', 'People & Blogs'
Least Obs: 'Nonprofits & Activism', 'Travel & Events'

Most Mean: 'Music', 'Comedy'
Least Mean: 'Nonprofits & Activism','News & Politics'

Most Stdev: 'Science and Tech', 'Comedy'
Least Stdev: 'Nonprofits & Activism','News & Politics'

Category	N	Mean	Median	Std Dev
Autos & Vehicles	898	76605	26300	114529
Comedy	2732	134681	87519	128480
Education	1165	79569	22477	105764
Entertainment	13136	106526	65077	109982
Film & Animation	2307	94314	55111	104482
Gaming	1806	99460	64837	100712
Howto & Style	2824	71889	27639	96261
Music	2467	114123	68182	118843
News & Politics	4991	69045	25610	91645
Nonprofits & Activism	201	32268	9450	61492
People & Blogs	6522	72869	28452	98196
Pets & Animals	399	82816	32043	105063
Science & Technology	1043	113838	55731	133568
Sports	3740	96655	40939	118894
Travel & Events	275	104031	59758	108814

Figure 3b: Stratified YouTube **Views** Quartile Distribution by Category

Stratified Proportional Allocation / Summary

Stratum	Observ	N_h/N	$1847*N_h/N$	Sample Size	$1893*N_h/N$	Sample Size
Autos & Vehicles	898	0.020177	37.26702	37	38.19516	38
Comedy	2732	0.061385	113.3781	113	116.2018	116
Education	1165	0.026176	48.34753	48	49.55163	50
Entertainment	13136	0.295151	545.1443	545	558.7213	559
Film & Animation	2307	0.051836	95.74055	96	98.12499	98
Gaming	1806	0.040579	74.94904	75	76.81567	77
Howto & Style	2824	0.063452	117.1961	117	120.1149	120
Music	2467	0.055431	102.3806	102	104.9304	105
News & Politics	4991	0.112142	207.1266	207	212.2852	212
Nonprofits & Activism	201	0.004516	8.341505	9	8.549252	9
People & Blogs	6522	0.146542	270.6631	271	277.4041	277
Pets & Animals	399	0.008965	16.55851	17	16.9709	17
Science & Technology	1043	0.023435	43.28452	43	44.36254	44
Sports	3740	0.084034	155.2101	155	159.0756	159
Travel & Events	275	0.006179	11.41251	12	11.69674	12
Total	**44506**			**1847**		**1893**

Figure 4: Proportional Allocation of Stratified Views by Category for Sample Size 1847 and 1893 After Design Effect

Figure 5: Comparisons of Mean Estimate for Task 1 Sampling Procedures for True Mean of 93891.7

$$n_{0,complex} = n_{0,srs} * \frac{V(\bar{y}_{complex})}{V(\bar{y}_{srs})} = 1847 * \frac{2551.78}{2490.11} = 1892.74 \approx 1893 \text{ samples}$$

Sample Procedure	Sample Size	Mean Estimate	Standard Error	Lower CI	Upper CI	True Mean Within CI?	Abs Diff From True Mean
Simple Random	1847	91932.82	2490.11	89442.71	94422.93	Yes	1958.88
Proportional Allocation	1847	96607.95	2551.78	94056.17	99159.73	No	2716.25
Proportional Allocation After Design Effect	1893	95055.87	2484.97	92570.9	97540.84	Yes	1164.17

Comparisons of Sampling Procedures x5

Sample Procedure	Sample Size	Mean Estimate	Standard Error	Lower CI	Upper CI	True Mean Within CI?	Abs Diff From True Mean
Simple Random	1847	91494.1	2511.972	88982.13	94006.07	Yes	2397.6
Simple Random	1847	95733.22	2562.7	93170.52	98295.92	Yes	1841.52
Simple Random	1847	94591.21	2575.722	92015.49	97166.93	Yes	699.51
Simple Random	1847	93210.51	2568.106	90642.4	95778.62	Yes	681.19
Simple Random	1847	96893.45	2609.723	94283.73	99503.17	No	3001.75
Proportional Allocation	1847	98122.43	2600.976	95521.45	100723.4	No	4230.73
Proportional Allocation	1847	92762.59	2533.739	90228.85	95296.33	Yes	1129.11
Proportional Allocation	1847	95200.77	2577.481	92623.29	97778.25	Yes	1309.07
Proportional Allocation	1847	91645.58	2441.759	89203.82	94087.34	Yes	2246.12
Proportional Allocation	1847	96482.08	2465.507	94016.57	98947.59	No	2590.38
Proportional Allocation After Design Effect	1893	97117.92	2547.008	94570.91	99664.93	No	3226.22
Proportional Allocation After Design Effect	1893	92751.74	2451.187	90300.55	95202.93	Yes	1139.96
Proportional Allocation After Design Effect	1893	91844.2	2463.036	89381.16	94307.24	Yes	2047.5
Proportional Allocation After Design Effect	1893	93422.21	2506.443	90915.77	95928.65	Yes	469.49
Proportional Allocation After Design Effect	1893	95806.36	2549.604	93256.76	98355.96	Yes	1914.66

Figure 6a: Comparisons of Mean Estimate for Task 2 Sampling Procedures for True Mean of 93891.7 (Different Seed Values)

Conclusions of Sampling Procedures / Summary

Average Sample Procedure	Sample Size	Mean Estimate	Standard Error	Lower CI	Upper CI	%True Mean Within CI	Abs Diff From True Mean
Simple Random	1847	94384.5	2565.645	91818.86	96950.15	80%	492.8
Proportional Allocation	1847	94842.69	2523.892	92318.8	97366.58	60%	950.99
Proportional Allocation After Design Effect	1893	94188.49	2503.456	91685.03	96691.95	80%	296.79

Figure 6b: Average Comparisons of Mean Estimate for Task 2 Sampling Procedures for True Mean of 93891.7

Proportional allocation after design effect is the best method to estimate the mean views -- lowest average absolute difference from true mean and 80% that the true mean is within CI

- 'Music' and 'Comedy' may be harder categories to get on trending
- 'Nonprofits & Activism' and 'News & Politics' may be easier categories to get on trending
- Use other social platforms and increase social interactions to get on trending

References

[Kaggle] "Trending YouTube Video Statistics: Daily statistics for trending YouTube videos," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/datasnaek/youtube-new [Accessed 23-Mar-2018]

[Google] "Trending on YouTube," Google, 2018. [Online]. Available: https://support.google.com/youtube/answer/7239739 [Accessed 23-Mar-2018]