International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"

Bollywood Box Office Revenue Distributions (2005–2017): A Categorical and Temporal Econophysics Analysis

Anhad Singh

Inventure Academy, Bengaluru, India

DOI: 10.46609/IJSSER.2026.v11i07.047 URL: https://doi.org/10.46609/IJSSER.2026.v11i07.047

Received: 15 July 2026 / Accepted: 26 July 2026 / Published: 31 July 2026

ABSTRACT

Box office revenue is famously unequal: a few films earn enormous amounts while most earn very little. Econophysics has studied this kind of skew, and earlier work found that Indian film revenue follows a log-normal distribution rather than a power law. This study extends that result to a larger and more recent sample and asks a question earlier work did not: does the distribution depend on what kind of film it is? Using 1,694 Hindi-language films released between 2005 and 2017 from the Premkumar (2020) dataset, I fitted log-normal and power-law distributions by maximum likelihood and compared them with the Vuong test, across three categorical splits (release period, remake status, and franchise status) and two eras defined by the arrival of OTT streaming. I also measured concentration with the Gini coefficient and fitted a budget–revenue scaling law. Log-normal was favoured over power-law in every large subgroup and in both eras. Concentration rose over time, with the annual Gini increasing from 0.575 to 0.693 (Spearman rho = +0.93) and the top-decile revenue share crossing 50 percent, which matches Elberse's blockbuster prediction rather than Anderson's long tail. The budget–revenue exponent was 0.519 and stayed close to that value across every film type.

Keywords: Econophysics; box-office revenue; log-normal distribution; power law; revenue concentration; Gini coefficient; Bollywood; OTT streaming

1. Introduction

Markets for cultural products, like films, music, and books, tend to produce very unequal outcomes. A small number of titles become huge hits while the rest earn modest amounts or lose money. De Vany and Walls [1,2] studied this pattern in the Hollywood film industry and found that revenue does not follow a normal, bell-shaped distribution at all. Instead it is heavy-tailed, which means that extreme outcomes happen far more often than a bell curve would predict. Sinha and Raghavendra [3] brought this idea into econophysics, the field that applies methods

ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"

from physics to economic data, and argued that movie revenue follows a power law. Pan and Sinha [4] tested this more carefully for 500 Indian films and reached a different conclusion: the data fit a log-normal distribution better than a power law. A log-normal and a power law are both heavy-tailed and look alike on a graph, so telling them apart needs a formal statistical test rather than the eye.

Most of this earlier work treats all films as one group. But films are not all the same. A film released during a holiday competes in a different market from an ordinary weekend release, and a sequel or a remake arrives with an audience that already exists. If these categories behave differently, then pooling them together could hide part of the picture. This study asks a first question that earlier work has not: do Bollywood revenue distributions differ across categories of film, and does the log-normal keep beating the power law once the data are split this way?

A second question is about time. Between 2015 and 2016 three streaming platforms launched in India: Hotstar in 2015, and Netflix India and Amazon Prime Video in 2016. Streaming changes how audiences find and watch films, so it might change how revenue is spread out.

Two theories predict opposite things. Anderson [5], in The Long Tail, argues that digital platforms remove the limits of shelf space and screen count, so audiences spread out and revenue becomes less concentrated. Elberse [6], in Blockbusters, argues the reverse: when people face too many choices they lean on big names and franchises, so revenue concentrates even more at the top. This study tests which prediction matches Bollywood by measuring whether revenue concentration changed across the point where OTT arrived.

To answer both questions I use a public dataset of 1,694 films, fit distributions by maximum likelihood, and compare them with the Vuong test. I measure concentration with the Gini coefficient and fit a scaling relationship between budget and revenue. The Methods section describes the data and each test, the Results section reports what I found, and the Discussion explains the findings and the limits of what the data can show.

2. Methods

2.1 Dataset

The data come from the Bollywood movies dataset compiled by Premkumar [7] and hosted on Mendeley Data under a Creative Commons licence (CC BY 4.0). It contains 1,698 Hindi-language films released between 2005 and 2017, with 14 columns and no missing values. The columns used most here are Revenue and Budget, both in Indian rupees, together with three categorical columns: Release Period (Normal or Holiday), Whether Remake (Yes or No), and Whether Franchise (Yes or No). I removed films with revenue below Rs. 10 lakh (Rs.

ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"

1,000,000), because very small entries are likely to be incomplete records, which left 1,694 films for the categorical analysis.

The dataset does not record each film's release year, only the overall 2005 to 2017 span. Because the era comparison needs years, I recovered them by matching film titles against TIMDB, a public database of about 7,400 Indian films that includes release years. The matching used exact titles first and then fuzzy matching (the rapidfuzz library) for titles that differ in spelling or punctuation, accepting a match only when the similarity score was at least 85 out of 100. This recovered years for 1,439 of the 1,698 films (84.7 percent). The 15.3 percent that did not match were used only in the categorical analysis, not the era analysis.

Two cautions apply to the data. Budget figures are self-reported and not audited, so any result involving return on investment should be read as a direction rather than an exact number.

Also, the matched subset contains 253 films dated to 2017, which is much higher than nearby years and is probably caused by date errors in TIMDB, so 2017 values are treated with care throughout.

2.2 Distribution fitting

For each group of films I fitted two distributions to the revenue values by maximum likelihood estimation (MLE), which chooses the parameter values that make the observed data most probable. The fitting used the powerlaw Python package [8], which follows the methods of Clauset, Shalizi, and Newman [9]. Because revenue is stored as whole rupees, the fit was run in discrete mode. A log-normal distribution is described by two parameters, mu and sigma, and a power law by two parameters, xmin and alpha.

2.3 Vuong likelihood-ratio test

To decide which of the two distributions fits better, I used the Vuong [10] likelihood-ratio test, which is built to compare two models where neither is a special case of the other. The test reports a statistic R and a p-value. A positive R favours the log-normal and a negative R favours the power law. A p-value below 0.05 means the preference is statistically significant, while a p-value above 0.10 means the test cannot separate the two models, usually because the group is small.

2.4 Two-sample Kolmogorov–Smirnov test

To test whether two groups of films have different revenue distributions, I used the two-sample Kolmogorov–Smirnov (KS) test. It reports a statistic D, the largest gap between the two groups' cumulative distributions, and a p-value. A large D with a small p-value means the two groups come from clearly different distributions.

ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"

2.5 Gini coefficient and concentration

To measure how concentrated revenue was in each year, I calculated the Gini coefficient of revenue, a standard measure of inequality that runs from 0 (every film earns the same) to 1 (one film earns everything). I also drew Lorenz curves for each era, calculated the share of revenue earned by the top 10 percent of films, and used the Spearman rank correlation to test whether the yearly Gini rose steadily over time.

2.6 Budget–revenue regression and ROI

Finally, I looked at the relationship between budget and revenue with an ordinary least squares regression of log(Revenue) on log(Budget). The slope of this line, beta, is an elasticity: it shows how revenue scales as budget grows. I also calculated return on investment (ROI) as Revenue divided by Budget, and compared ROI between groups with the Mann–Whitney U test, a rank-based test that does not assume the data are symmetric, which suits ROI because it is strongly right-skewed.

3. Results

3.1 Categorical distribution analysis

Table 1 reports the distribution comparisons. In the three largest subgroups the log-normal was favoured over the power law at the 0.05 level: Normal releases ($R = +6.723$, $p = 0.031$), Holiday releases ($R = +8.019$, $p = 0.012$), and non-franchise films ($R = +8.699$, $p = 0.013$). For remakes the result pointed the same way but was only marginal ($R = +3.443$, $p = 0.090$, $n= 71$), and for non-remakes and franchise films the test was inconclusive, which reflects their sample sizes rather than support for the power law. The two-sample KS tests showed that all three categorical splits separate into clearly different distributions, with the largest gaps for franchise ($D = 0.575$) and remake ($D = 0.531$) and a smaller gap for release period ($D = 0.100$). The median revenue figures point the same way: Holiday films earned a median of Rs. 7.5 crore against Rs. 4.75 crore for Normal films, remakes Rs. 27 crore against Rs. 5 crore, and franchise films Rs. 29 crore against Rs. 5 crore. Figure 1 shows these splits as log-revenue histograms and complementary CDFs.

Table 1. Log-normal vs. power-law comparison (Vuong test) by subgroup and era

Split	Subgroup	n	Vuong R	Vuong p	Verdict
Release Period	Normal	1061	+6.723	0.031	Log-normal (significant)
Release Period	Holiday	633	+8.019	0.012	Log-normal (significant)

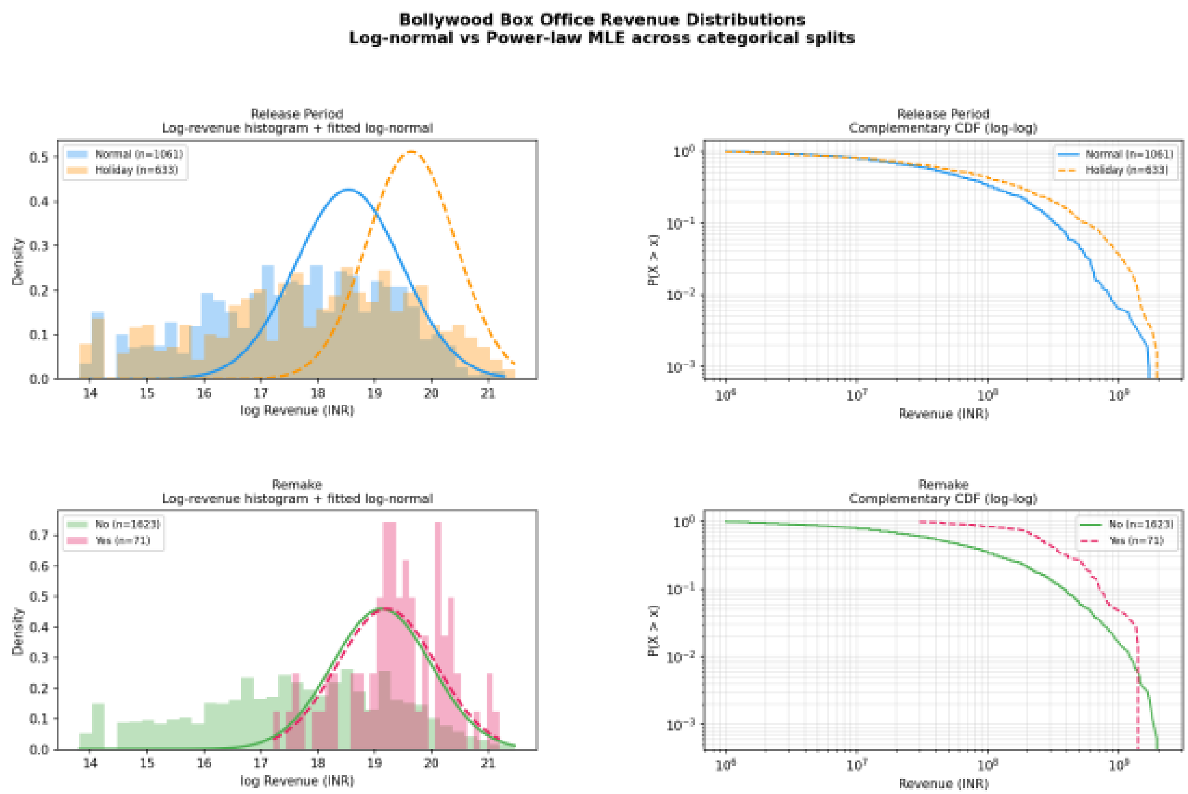
ISSN: 2455-8834

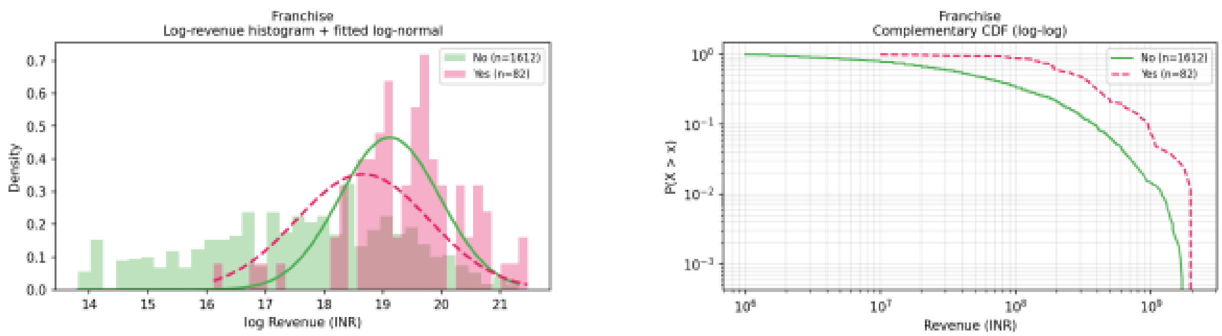
Volume:11, Issue:07 "July 2026"

Remake	No	1623	—	>0.10	Inconclusive (low power)
Remake	Yes	71	+3.443	0.090	Log-normal (marginal)
Franchise	No	1612	+8.699	0.013	Log-normal (significant)
Franchise	Yes	82	+0.875	0.371	Inconclusive (low power)
Era	Pre-OTT (2005–2012)	705	+9.458	0.006	Log-normal (significant)
Era	Early-OTT (2013–2017)	733	+5.161	0.033	Log-normal (significant)

Note. Between-group Kolmogorov–Smirnov distances: Release Period D = 0.100; Remake D = 0.531; Franchise D = 0.575; Era D = 0.177. All three categorical splits are statistically distinct (p < 0.05). Verdicts labelled "inconclusive" reflect limited sample size in the Yes cells, not evidence against log-normal.

Fig. 1. Log-revenue histograms (left) and complementary CDFs on log–log axes (right) for the Release Period, Remake, and Franchise splits, with fitted log-normal curves overlaid. Holiday, Remake=Yes, and Franchise=Yes subgroups show a clear rightward shift



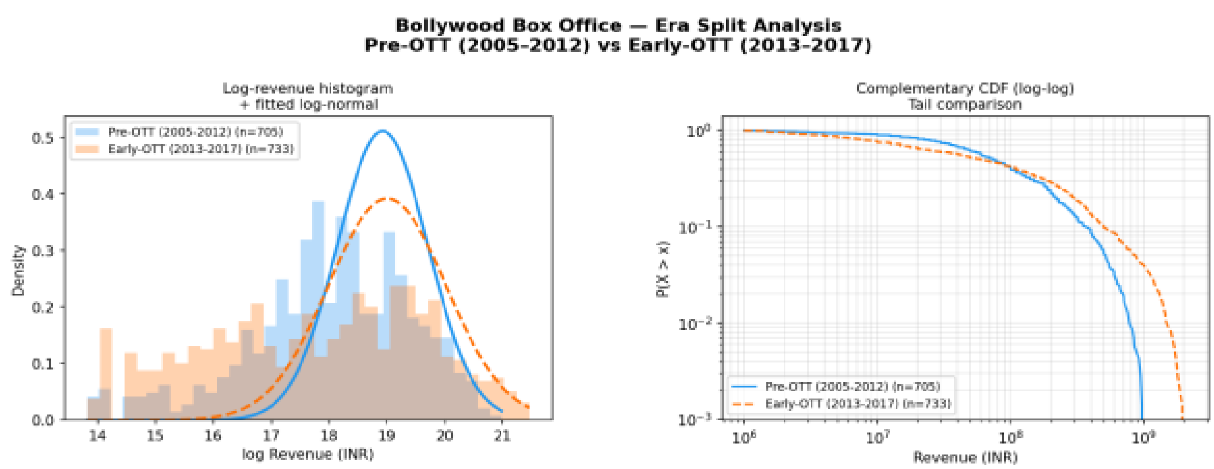


Source: author's analysis of the Premkumar (2020) Bollywood dataset [7].

3.2 Era split analysis

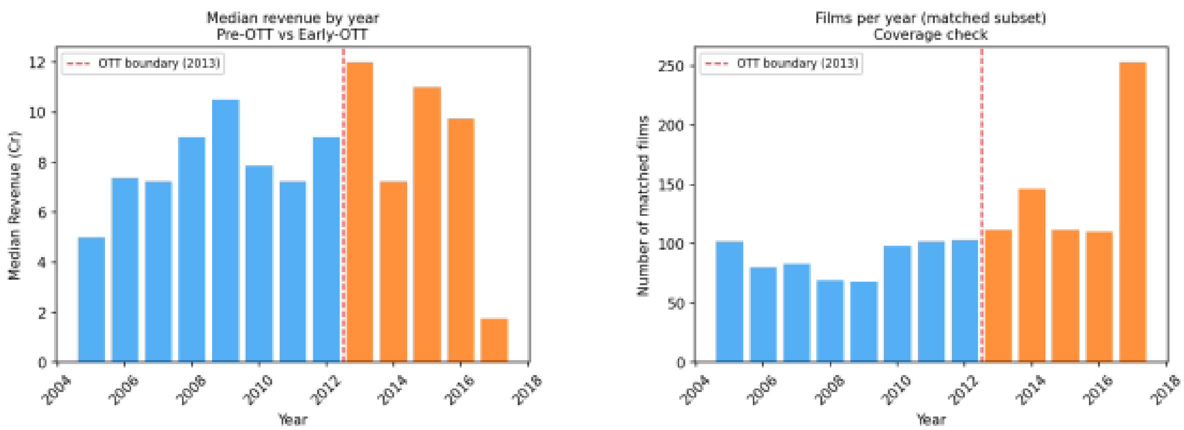
When the films with recovered years were split into a pre-OTT era (2005 to 2012, n = 705) and an early-OTT era (2013 to 2017, n = 733), the log-normal was again favoured in both (pre-OTT R = +9.458, p = 0.006; early-OTT R = +5.161, p = 0.033). The fitted parameters show how the distribution changed. The centre barely moved, with mu rising by only 0.079, but the spread grew, with sigma rising from 0.780 to 1.018. Over the same period the median revenue fell from Rs. 7.5 crore to Rs. 6.0 crore while the mean rose from Rs. 14.1 crore to Rs.19.2 crore. The KS distance between the two eras was D = 0.177, smaller than any of the categorical splits. Figure 2 shows the era comparison.

Fig. 2. Era split comparing pre-OTT (2005–2012, n=705) and early-OTT (2013–2017, n=733) films. Top left: log-revenue histograms with fitted log-normal curves. Top right: complementary CDFs. Bottom left: median revenue by year. Bottom right: film count by year (note the elevated 2017 count discussed in Section 2.1)



ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"

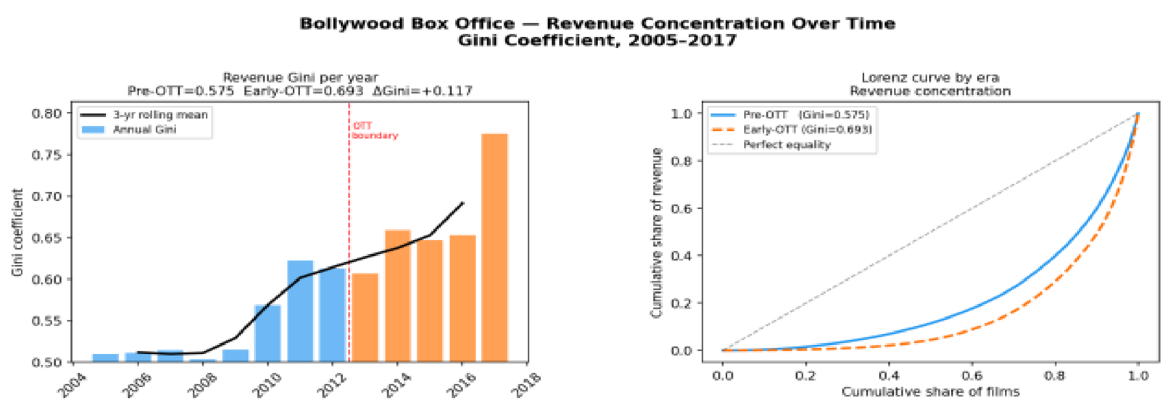


Source: author's analysis of the Premkumar (2020) Bollywood dataset [7], years recovered via TIMDB.

3.3 Revenue concentration over time

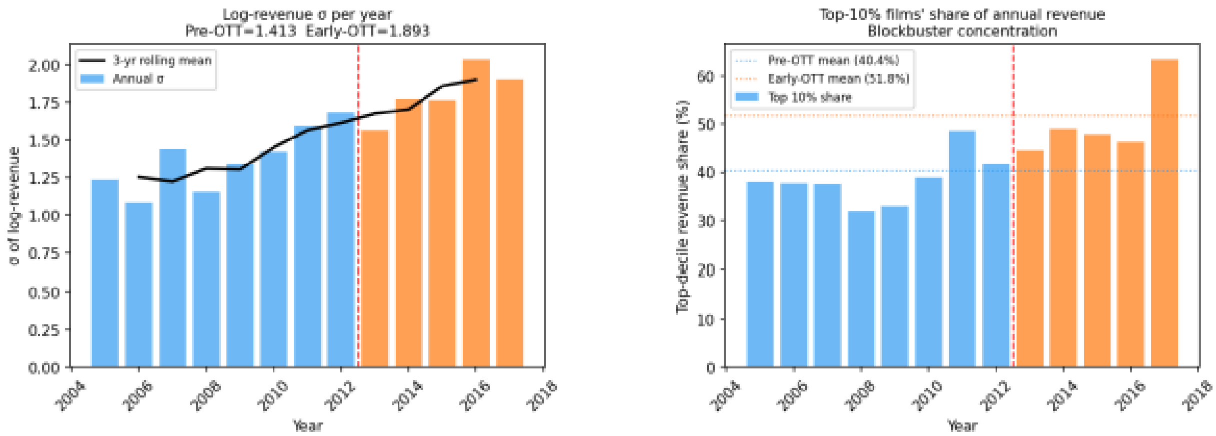
The yearly Gini coefficient rose from 0.575 in the pre-OTT era to 0.693 in the early-OTT era, an increase of 0.117. The Spearman rank correlation between year and Gini was +0.929 ($p < 0.001$), which means the rise was almost perfectly steady from year to year. The share of revenue earned by the top 10 percent of films rose from 40.4 percent to 51.8 percent, crossing the halfway mark. A second, separate measure agreed with the Gini: the standard deviation of log-revenue within each year rose from 1.413 to 1.893. Because 2017 is over-represented in the matched data, its values are read with the caution noted in the Methods. Figure 3 shows all four concentration measures.

Fig. 3. Revenue concentration, 2005–2017. Top left: annual Gini with 3-year rolling mean and the 2013 OTT boundary. Top right: Lorenz curves by era. Bottom left: annual log-revenue sigma. Bottom right: top-decile revenue share. All four panels show increasing concentration



ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"

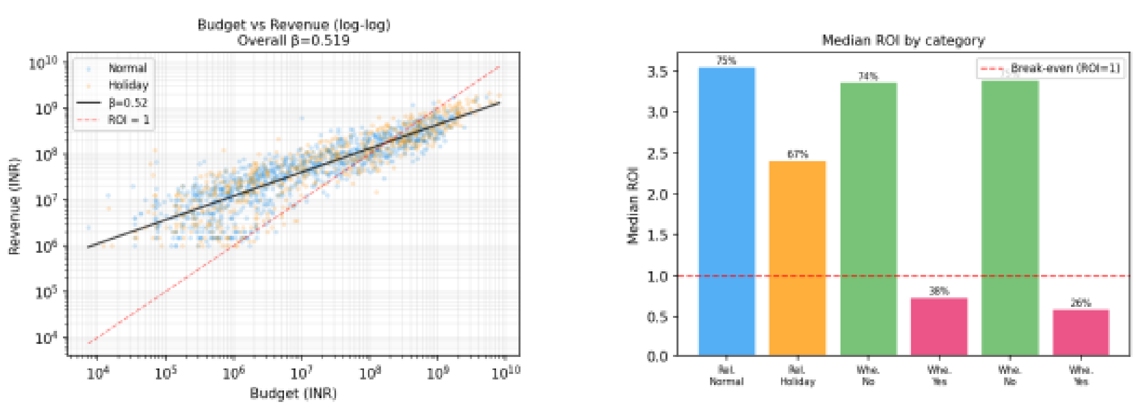


Source: author's analysis of the Premkumar (2020) Bollywood dataset [7].

3.4 Budget–revenue scaling and ROI

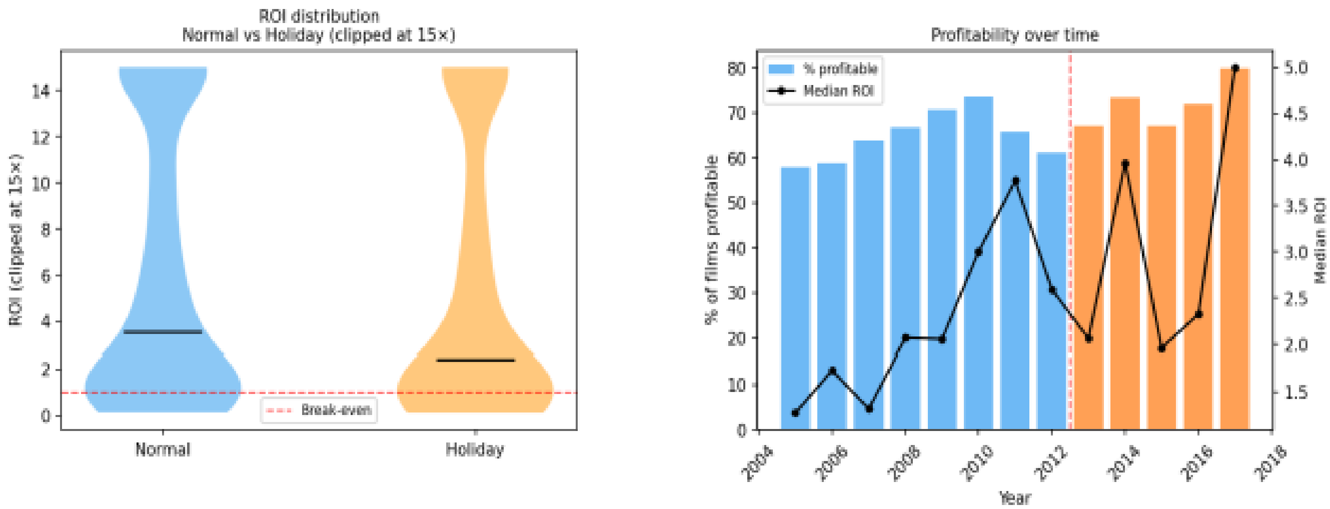
The regression of log(Revenue) on log(Budget) gave a slope of beta = 0.519 (R-squared = 0.784, p < 0.001), so doubling a film's budget was associated with about 2^0.519 = 1.43 times the revenue, less than double. This slope stayed close to 0.519 across every categorical split, ranging only from 0.487 to 0.535. The ROI results were more surprising. Franchise films were profitable only 25.6 percent of the time, with a median ROI of 0.59 (Mann–Whitney p < 0.001), and remakes were profitable 38 percent of the time, with a median ROI of 0.73. Holiday films had a lower median ROI (2.40) than Normal films (3.55, p = 0.001) even though their absolute revenue was higher. Figure 4 shows the budget and ROI results.

Fig. 4. Budget and ROI analysis. Top left: budget vs revenue (log–log) with the regression line (beta=0.519) and the break-even line. Top right: median ROI by category. Bottom left: ROI violin plots for Normal vs Holiday
Bottom right: profitability rate and median ROI by year



ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"



Source: author's analysis of the Premkumar (2020) Bollywood dataset [7]; budget figures are self-reported (Section 2.1).

4. Discussion

The log-normal distribution has a natural explanation. A film's revenue depends on many factors multiplied together: the size of the release, the marketing spend, the appeal of the cast, word of mouth, and timing. When many independent factors multiply, the result tends toward a log-normal distribution, an idea known as Gibrat's law. This is why the log-normal, rather than the power law, keeps winning the comparison, and it lines up with what Pan and Sinha [4] found for Indian films. The two subgroups where the Vuong test was inconclusive, non-remakes and franchise films, are best read as cases where the sample was too small for the test to decide, not as support for the power law. In every group large enough to give the test real power, the log-normal won.

Release period gives the clearest structural result. Both the Normal and Holiday groups are large, and both favour the log-normal significantly. Holiday films sit to the right of Normal films, with a higher median (Rs. 7.5 crore against Rs. 4.75 crore) and a heavier upper tail, as Figure 1 shows. Remakes and franchise films have the largest KS distances of all (0.531 and 0.575), so their distributions are the most different from the rest. But because only 71 films are remakes and 82 are franchise films, the distribution fits for these two groups are less certain, and a larger dataset would be needed to describe their shape with confidence. What can be said now is that these categories sit clearly higher in the revenue distribution than originals do.

The concentration results speak to the debate between Anderson [5] and Elberse [6]. Revenue

ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"

became more concentrated over the study period, not less: the Gini rose from 0.575 to 0.693, the rise was almost perfectly steady (rho = +0.93), and the top 10 percent of films moved from earning about 40 percent of revenue to earning more than half. This is the direction Elberse predicts, not Anderson. Cause is harder to pin down, though. Concentration was already climbing before the OTT platforms launched, and other things changed at the same time, such as bigger marketing budgets and more multiplex screens. So the data show that concentration rose alongside the arrival of OTT, but they cannot show that OTT caused it.

What does add confidence is that two different measures, the Gini (which assumes nothing about the distribution) and the fitted sigma (which does), moved together, which rules out the rise being an artefact of the fitting method.

The scaling law is the most physics-like result. The exponent beta stayed near 0.519 no matter how the films were split, and an exponent that holds steady across categories is the kind of invariance econophysics looks for; De Vany and Walls [11] reported a similar pattern for Hollywood. Because beta is below 1, revenue grows more slowly than budget, so bigger films face diminishing returns. This also explains the ROI paradox. Franchise and remake films earn more money in total, but studios pay more to make them, and since revenue does not keep up with budget, the extra spending does not pay off at the median. The same logic fits the holiday result: holiday films earn more in absolute terms but return less per rupee spent, which suggests the holiday slot carries a premium that the revenue does not fully repay.

5. Conclusion

This study set out to answer two questions about Bollywood box office revenue. The first was whether revenue distributions differ across film types and whether the log-normal beats the power law once the data are split. The log-normal was favoured in four of the six categorical subgroups and in both eras, and it lost in none; the two undecided cases were simply too small to call. Release period, remake status, and franchise status each separated films into clearly different distributions. The second question was whether concentration changed as OTT streaming arrived. Concentration rose steadily, with the Gini climbing from 0.575 to 0.693 and the top decile passing 50 percent of revenue, which matches Elberse's blockbuster prediction rather than Anderson's long tail. These findings come with limits. The data cannot establish cause, 15.3 percent of films were left out of the time analysis, the 2017 figures are uncertain, and only Hindi-language films were studied. Future work could extend the data past 2017, add screen counts as a further variable, and separate streaming releases from theatrical ones.

Acknowledgement

I thank my research mentor for guidance throughout this project, and Prashant Premkumar for

ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"

making the Bollywood movies dataset publicly available under a Creative Commons licence. Funding: this research received no external funding. Conflicts of interest: the author declares no conflict of interest.

References

[1] De Vany, A. S., & Walls, W. D. (1996). Bose-Einstein dynamics and adaptive contracting in the motion picture industry. Economic Journal, 106(439), 1493–1514. DOI: 10.2307/2235197.

[2] De Vany, A. S., & Walls, W. D. (1999). Uncertainty in the movie industry: Does star power reduce the terror of the box office? Journal of Cultural Economics, 23(4), 285–318. DOI: 10.1023/A:1007608125988.

[3] Sinha, S., & Raghavendra, S. (2004). Hollywood blockbusters and long-tailed distributions: An empirical study of the popularity of movies. European Physical Journal B, 42(2), 293–296. DOI: 10.1140/epjb/e2004-00382-7.

[4] Pan, R. K., & Sinha, S. (2010). The statistical laws of popularity: Universal properties of the box-office dynamics of motion pictures. New Journal of Physics, 12(11), 115004. DOI:10.1088/1367-2630/12/11/115004. https://doi.org/10.1088/1367-2630/12/11/115004

[5] Anderson, C. (2006). The Long Tail: Why the Future of Business Is Selling Less of More. New York: Hyperion. ISBN 978-1-4013-0237-5.

[6] Elberse, A. (2013). Blockbusters: Hit-Making, Risk-Taking, and the Big Business of Entertainment. New York: Henry Holt and Company. ISBN 978-0-8050-9433-6.

[7] Premkumar, P. (2020). Bollywood Movies Data. Mendeley Data, V1. DOI: 10.17632/3c57btcxy9.1. https://data.mendeley.com/datasets/3c57btcxy9/1

[8] Alstott, J., Bullmore, E., & Plenz, D. (2014). powerlaw: A Python package for analysis of heavy-tailed distributions. PLoS ONE, 9(1), e85777. DOI: 10.1371/journal.pone.0085777.

[9] Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. SIAM Review, 51(4), 661–703. DOI: 10.1137/070710111. https://doi.org/10.1137/070710111

[10] Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57(2), 307–333. DOI: 10.2307/1912557.

ISSN: 2455-8834

Volume:11, Issue:07 "July 2026"

[11] De Vany, A. S., & Walls, W. D. (2004). Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar. Journal of Economic Dynamics and Control, 28(6), 1035–1057. DOI:10.1016/S0165-1889(03)00065-4.

