Goodreads-10k data narrative

Pranav Joshi

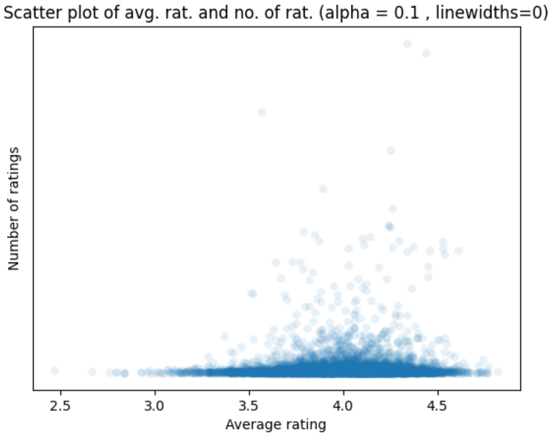
C.S.E. at IITGN ,22110197

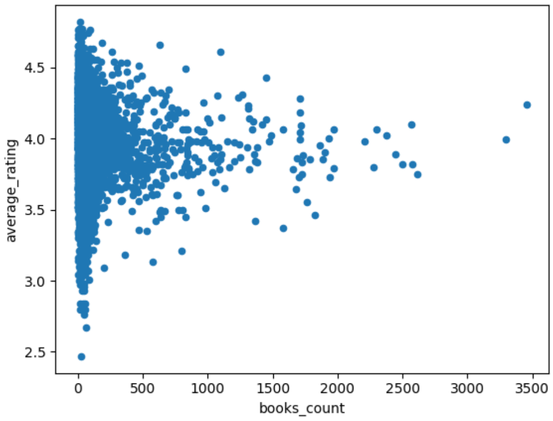
Overview

The goodreads-10k data-set is publicly available on github. It consists of data like number of ratings, average ratings, 1 star ratings, 2 star ratings … , original publication year, etc., for 10,000 books.

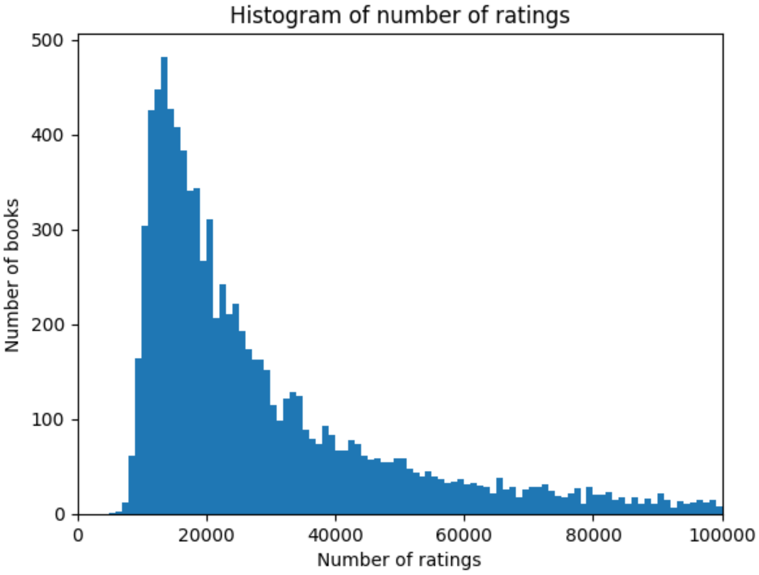
Observations :

1) The distribution of books according to average ratings moves very slowly towards higher average_rating with increasing books_count or ratings_count (no. of ratings) .

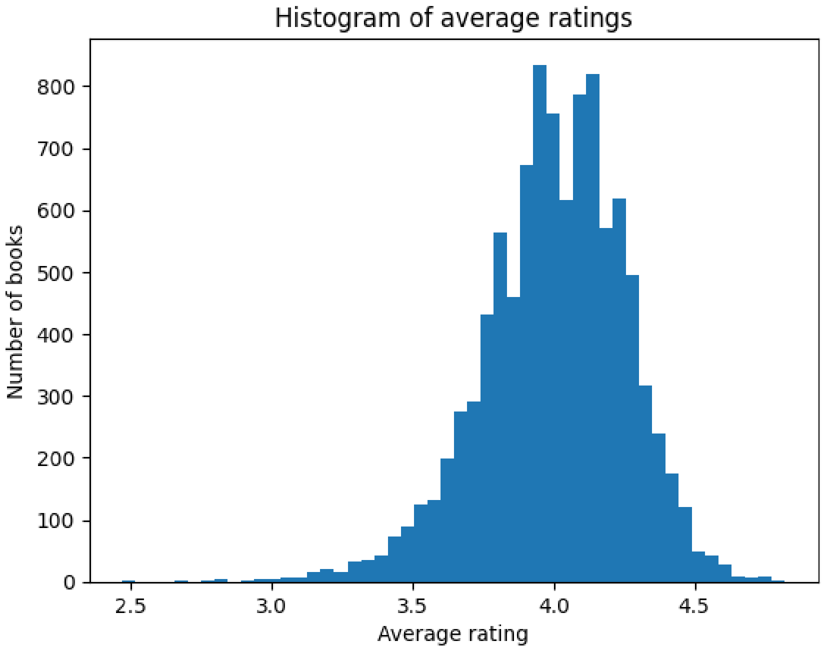




2) The histogram of ratings_count first increases, then peaks at ~20000 and then decreases



3) Most common average rating is ~4 stars.

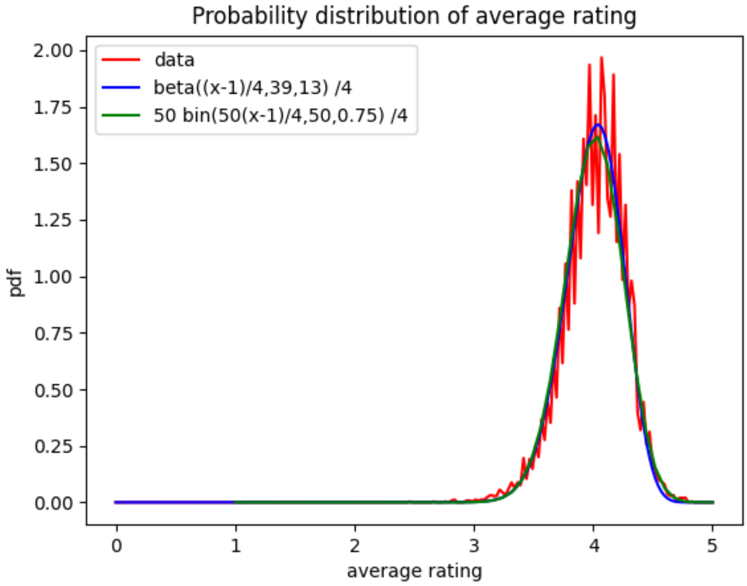


Hypotheses

1. Popularity is a consequence, not a factor

For a certain book, the probability that a new reader gives it a certain rating is independent of the book's current average rating and number of ratings… i.e. people who rate the book do it considering their own experience with the book **only**. So the fraction of ratings which are 1 star doesn't change as the popularity of book grows, since this probability is a quality of only the book and the kind of crowd reading the book. Same happens for 2,3,4,5 star ratings. I am **not** saying that the current average rating doesn't affect the decision of reading the book. What I am saying is it doesn't affect the decision of **rating** the book. In fact, we can see that there are very less or no books below 2.5 star average rating, which must be because users don't read books with bad average rating or too few ratings. Thus good books (ones with high true average ratings, as rated by a professional) are more popular, which explains the 1^{st} observation.

I think that the true average rating of the book (which is achieved at infinite rating count) is all that matters, not the popularity of the book. More ratings do make us more confident that the book's real rating is close to its rating on the site though. That's why I filtered out books with less than 50 ratings before making the plots. But considering that there aren't many books like that, the plots were not changed all that much. Now, knowing that the average ratings on the site are close to the true average ratings, we can treat the histogram of average rating as a probability distribution (after normalising it) . This is the result:



Here the red curve is just a histogram of the data with 200 bins , divided by the number of observations , and multiplied by 200 , to get the probability distribution.

I read an article[1] where the histogram is estimated by a beta function, like the one in above picture (blue curve) .

I developed a theory which says that the **binomial** distribution (green) is a good estimate, as visible in this picture.

2. Books are a collection of traits :

A good book has certain traits like good cover, good content, correct tone, suitable graphics, etc. Let's consider A such traits for every book. The probability of a certain book having one of these traits is independent of all the other traits. For simplicity, let's assume that this probability is the same for every trait , namely q . Then the probability that the book has xA of these traits is :

$$P_x(x) \;=\; {}^{A}C_{Ax}\ q^{Ax}\ (1-q)^{A(1-x)}$$

Whether a book has a certain trait is decided by its audience, and not an unbiased professional. Since most people who read the book, read it because they believed **i**t has desired traits (after maybe reading its description online) they will most likely not judge the book critically, as it was their choice to read it. Thus we expect the probability q to be high. Another reason for most books being 'good' as viewed by their audience is that really bad books don't get popular and maybe aren't on the site.

While rating a book, a person considers whether a trait he considered desirable is present or not. Let's say every user looks for 4 desired traits. If all 4 are present, the user gives the book a rating of 5 stars. For 3 of 4 traits, he gives 4 stars …… For 0 of these 4 traits, he gives 1 star (the minimum rating) . Then the probability that N users give the book a total of S stars is :

$$P_S(S|N,x) \;=\; {}^{4N}C_{S-N}\ x^{(S-N)}\ (1-x)^{4N-(S-N)}$$

Here I am thinking of every user presenting 4 Bernoulli trials for the book. In each trial we check a randomly selected trait of the book. Thus, to get a total of S stars, the book must succeed in $(S-N)$ out of $4N$ of these trials. As N get bigger, this curve gets narrower, with peak at $S \approx N(1+4x)$. Thus the book's average rating would be roughly $(1+4x)$. So the probability that a book has rating R is the same as probability that it has $\frac{R-1}{4}A$ 'good book traits', which we know is $P_x(\frac{R-1}{4})$. Now, in a range of $[R, R+\Delta R]$, you will have around $\frac{A\Delta R}{4}$ of numbers which we can put in our expression to get a probability out … because non integer number of good book traits are impossible (probability = 0). All these numbers will give almost the same output (around $P_x(\frac{R-1}{4})$) if ΔR is small enough.

Thus, the probability **density** function for R is :

$$f(R) = \frac{A}{4}\, P_x(\frac{R-1}{4}) = \frac{A}{4}\, {}^{A}C_{A\,\frac{R-1}{4}}\ q^{A\,\frac{R-1}{4}}\ (1-q)^{A(\frac{5-R}{4})}$$

With some hit and trial, i found that $q=$ 0.75 and $A=$ 50 fit the data good enough.

Binomial to Beta distribution:

The normal approximation for binomial p.d.f is well known [2] :

$$Bin(x|A,q)\ =\ A^{A}C_{Ax}\ q^{Ax}\ (1-q)^{A(1-x)} \approx \sqrt{\frac{A}{2\pi q(1-q)}}exp(-\frac{(x-q)^2 A}{2q(1-q)})$$

Replacing x by q and vice versa, and replacing A by $A+1$, we get a beta p.d.f.

$$Beta(x\,|\,Aq+1, A(1-q)+1)\ =\ (A+1)^{A}C_{Aq}\ x^{Aq}\ (1-x)^{A(1-q)} \approx \sqrt{\frac{A}{2\pi x(1-x)}}exp(-\frac{(x-q)^2 A}{2x(1-x)})$$

So when $x \approx q$ or $x \approx 1-q$, the two functions are very close.

Also, when $q \approx 1/2$, the beta p.d.f. can be approximated as a normal distribution [3] of mean $\approx q$ and variance $\approx q(1-q)/A$, which is :

$$\sqrt{\frac{A}{2\pi q(1-q)}}exp(-\frac{(x-q)^2 A}{2q(1-q)})$$

Thus in this case, for any x value, the two curves are very close.

Fraction of s-star ratings

Given a book has xA good traits, the probability that a user would rate it s stars is:

$$f_{(x,s)} =^{4} C_{(s-1)} x^{(s-1)} (1-x)^{(5-s)}$$

This is the fraction of N ratings that are s star ratings when N gets very big.

Now, we can calculate the expected values of fraction of s - star ratings.

$$E(f(x)) = \sum_{Ax=0}^{Ax=A} f(x) P_x(x) =$$

$$\sum_{k=0}^{A} \mathrm{nCr}\,(4, s-1) \left(\frac{k}{A}\right)^{(s-1)} \left(1-\frac{k}{A}\right)^{(5-s)} \mathrm{nCr}\,(A,k)\, q^{k}\, (1-q)^{(A-k)}$$

Evaluation of this sum can be done using a computer in this case as $A = 50$ is small enough. This gives us this table :

s	E (f (x,s))
1	0.005392
2	0.052332
3	0.208152

s	E (f (x,s))
4	0.405132
5	0.328992

Now, we can try to approximate this sum as an integral

$$\sum_{Ax=0}^{Ax=A} f(x)P_x(x) \approx A\int_0^1 f(x)P_x(x)dx$$

Here, we can approximate $A\,P_x(x)$, a binomial p.d.f, as a beta p.d.f to get:

$$E(f(x)) \approx A^A C_{Aq}\,{}^4C_{(s-1)} \int_0^1 x^{(qA+s-1)}(1-x)^{((1-q)A+5-s)}dx$$

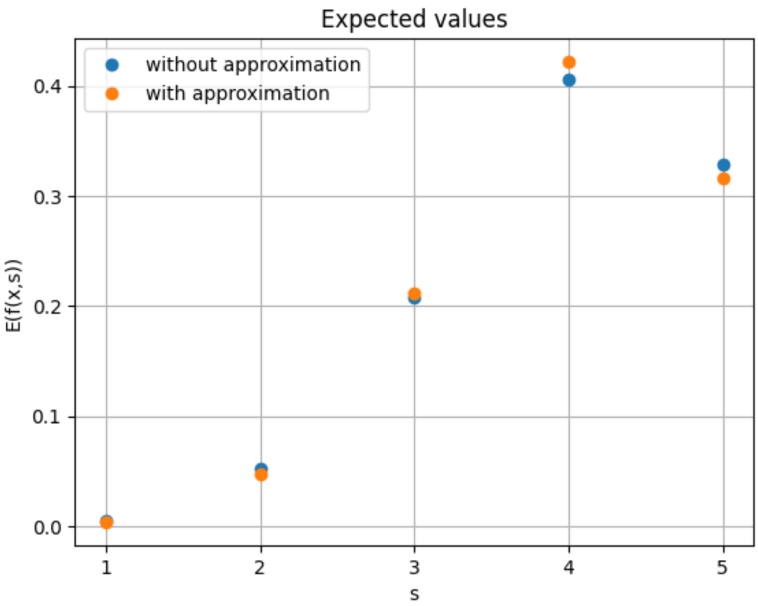
The integral evaluates to $B(qA+s,(1-q)A+6-s) = [(A+5)(\ {}^{(A+4)}C_{(qA+s-1)}\ )]^{-1}$

Approximating further, we get : $E(f(x,s)) \approx {}^4C_{(s-1)}q^{(s-1)}(1-q)^{(5-s)}$

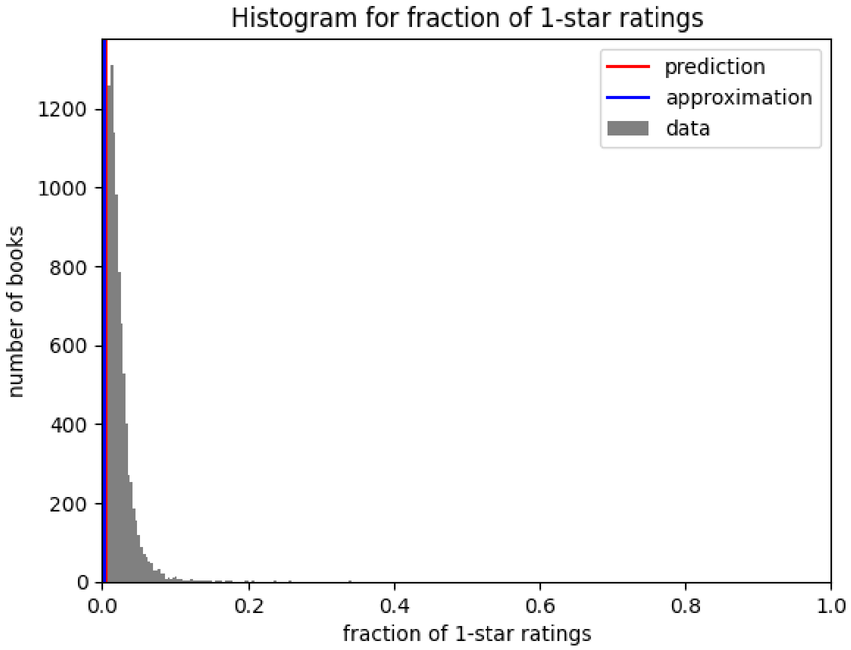
Using this, we get these values:

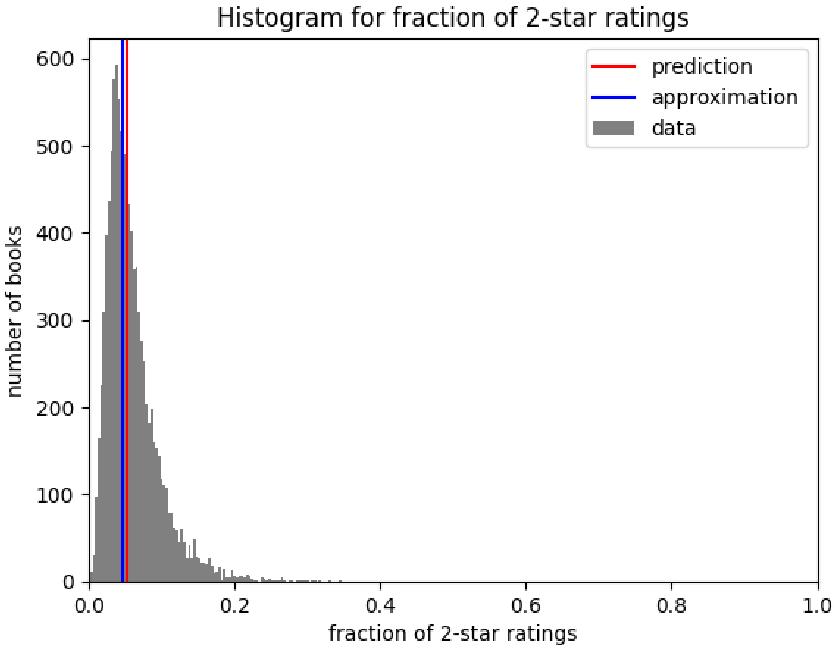
s	E (f (x,s))
1	0.00390625
2	0.046875
3	0.2109375
4	0.421875
5	0.31640625

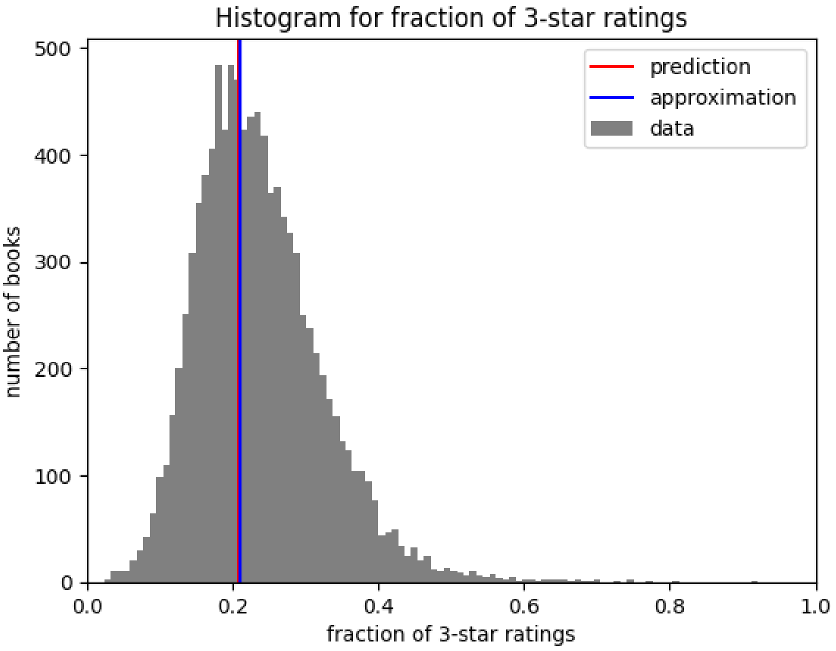
Graphically , this is how these values compare:

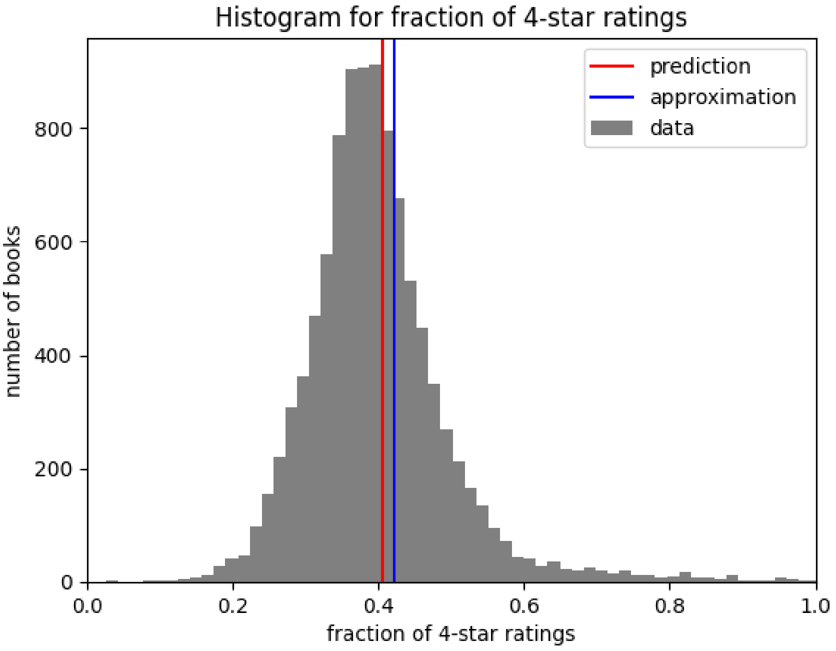


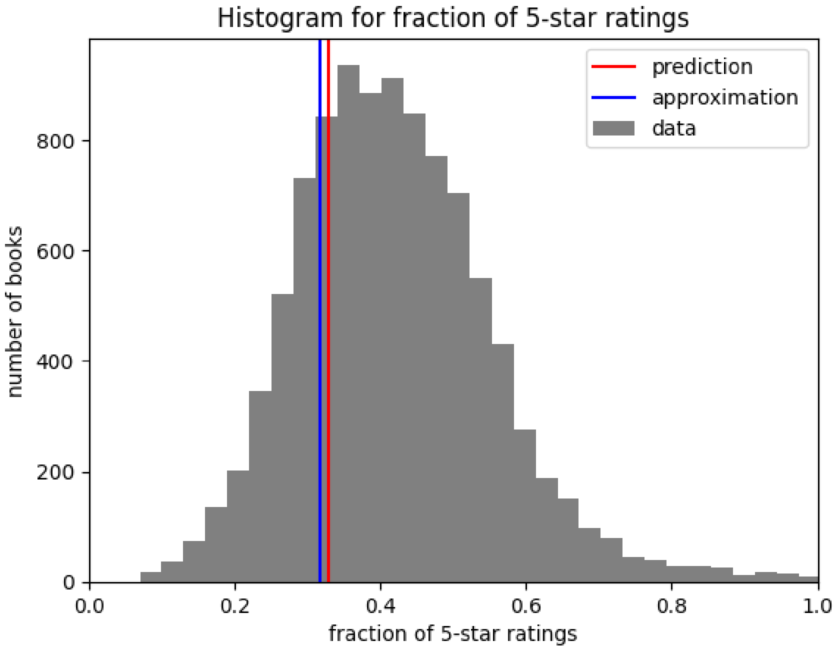
Comparison with data :


Histogram for fraction of 1-star ratings
prediction
approximation
data
number of books
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
fraction of 1-star ratings


Histogram for fraction of 2-star ratings
prediction
approximation
data
number of books
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0
fraction of 2-star ratings


Histogram for fraction of 3-star ratings
prediction
approximation
data
number of books
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
fraction of 3-star ratings


Histogram for fraction of 4-star ratings
prediction
approximation
data
number of books
0
200
400
600
800
0.0
0.2
0.4
0.6
0.8
1.0
fraction of 4-star ratings



Unanswerable questions

I couldn’t prove my hypothesis about average rating approaching the true rating as ratings_count increases since we do not have time related data for individual books.

Libraries and Functions used:

- numpy python library
- matplotlib python library
- pandas python library
- comb(·) function from math python library
- beta.pdf(·) function from scipy.stats python sub-module

References

[1]: Looking at the distribution of ratings on Goodreads, *Adam Fontenot*, 27 February 2022

[URL]: https://adamfontenot.com/post/looking_at_the_distribution_of_ratings_on_goodreads

[2]: Binomial Distribution, Wikipedia

[URL]: https://en.wikipedia.org/wiki/Binomial_distribution#Normal_approximation

[3]: Beta Distribution, Wikipedia

[URL]: https://en.wikipedia.org/wiki/Beta_distribution#Normal_approximation_to_the_Beta_distribution

Acknowledgements

I want to thank professor Shanmuga R. for providing necessary practice and exposing me (and my batch-mates) to this topic and helping me in proving the approximate equality between Binomial and Beta distribution. I also want to acknowledge these online resources that helped me teach myself basic Bayesian statistics and thus understand Adam Fontenot's article. : Beta distribution , Bayesian Inference