Ratings – Average is not the Middle.

By Eric Oulster

I’ve often read internet-based movie reviews like the following:

“This movie/film/album was average. It took no risks and was completely middle of the road,  
70/100.”

They have always struck me as odd. A 5/10 or 50% would be the middle-ground between an abysmal 0/10 and a perfect 10/10, yet I’ve noticed people seem to stand around a 7/10 or 70% when rating a piece of media that is ‘so-so’ but also ‘not bad’.

Is 70% really ‘average’, or have I just been imagining things?

I decided to download several different datasets of internet reviews to find out.

Rather than just sticking to a single website or medium, I decided to look at several different sources. That way, I could be sure that the results were just a by-product of a single website’s scoring system, and instead represented a larger trend in review-scores.

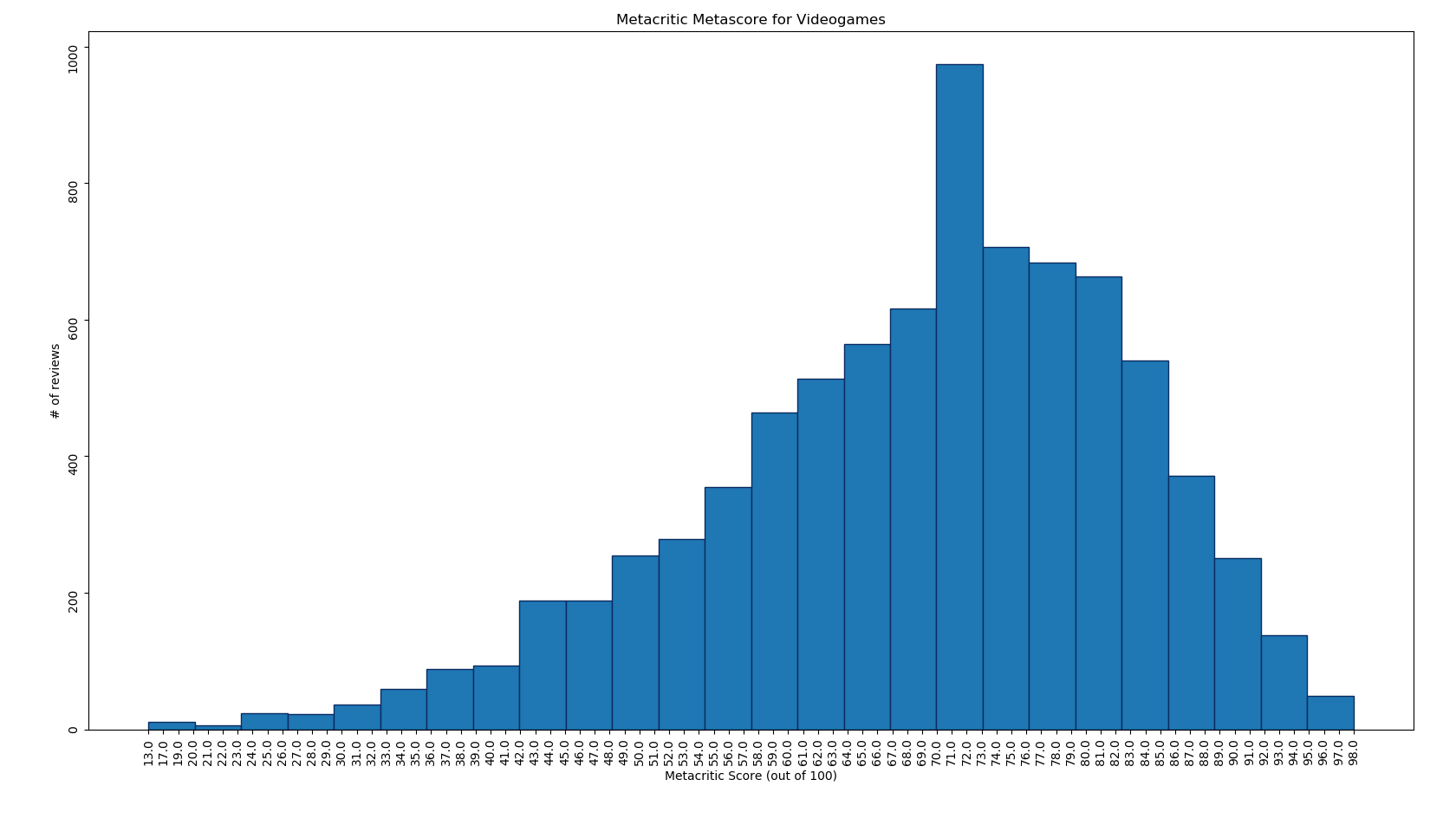
Before looking at all of the different reviews normalized and smooshed into one large dataset (which you can see near the bottom) I decided to take a quick look at the datasets individually- just to see if they stood out differently from one another, or if one of them was completely unlike the rest.

Metacritic: Metascore (Videogames)

Metacritic has two different types of scores it uses: Metascores and User scores.

According to the [Metacritic website](https://www.metacritic.com/about-metascores), their ‘Metascore’ is a weighted average of individual critic scores. They take a bunch of reviews from official critics, assign a ‘weight’ to each of the critics, ‘normalize’ all the scores (so a 9/10 becomes a 90/100, a B+ becomes an 83, etc… Then they base their Metascore off of this.

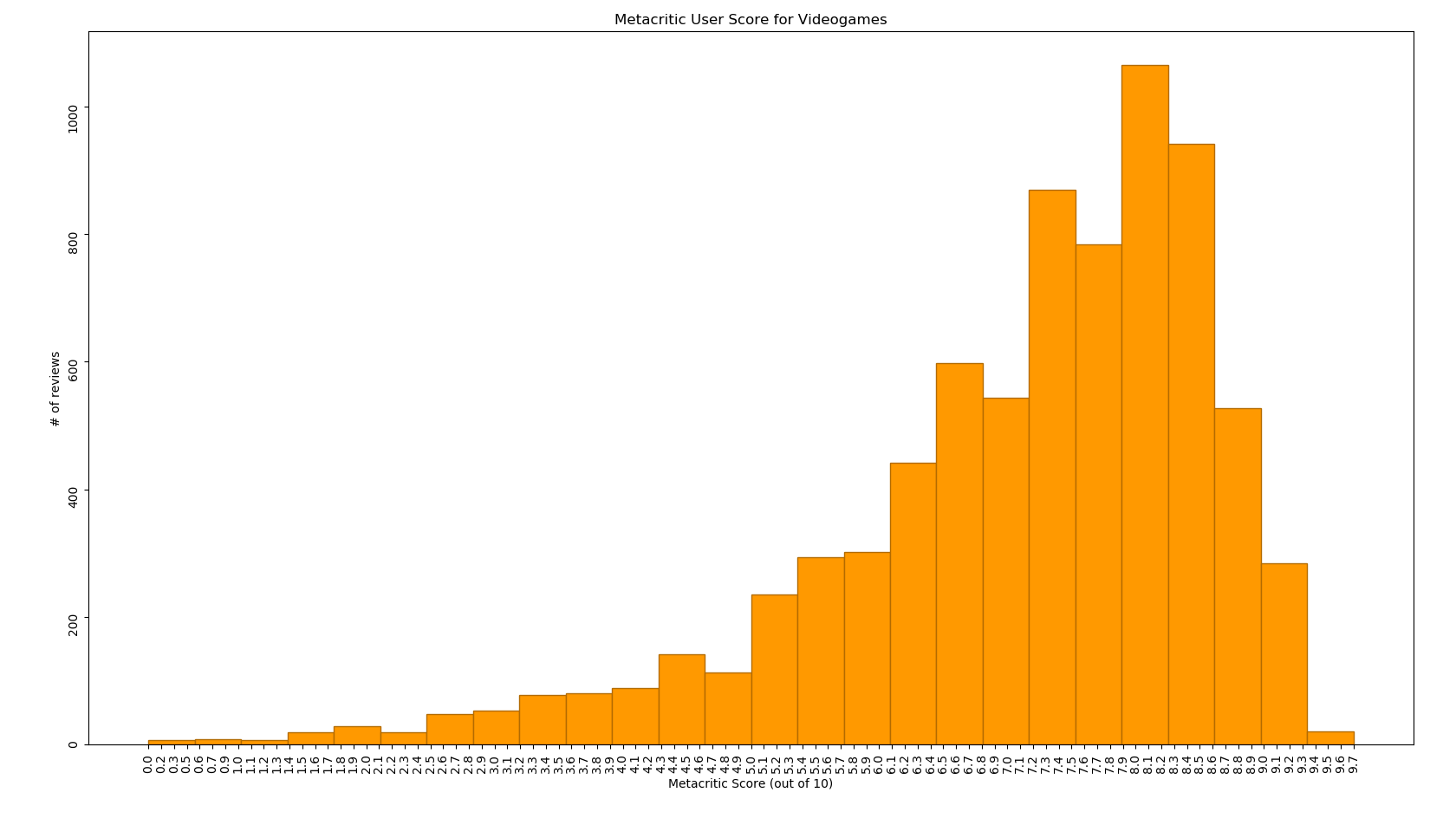
Reverse engineering how Metacritic gets their Metascores would be worth an entire article on its own- [one which has already been done, no less](http://shashank.ramaprasad.com/2015/06/14/reverse-engineering-the-metacritic-movie-ratings). The ‘meta’ in Metacritic is because their scores are a combined total of many others.



We see an obvious peak around the mid-70’s, but other than that, a fairly normal & tapering off in both directions. Consulting with the [Metacritic website](https://www.metacritic.com/about-metascores), ‘Mixed or Average Reviews’ for videogames fall between 50-74%. Based on that, I would have suspected that the peak of the bell-curve would’ve fallen in the middle of the ‘Average Reviews’ score range (62%). Instead, it peaks around the 70-73% range- illustrating that a 70/100 score is, indeed, average. It’s also nice to know that the average Metascore is actually within the boundaries of ‘Mixed or Average Reviews’, even if it’s touching the upper-ceiling of that classification.

Metacritic: Userscore (Videogames)

Userscores on Metacritic follow a far simpler process: Metacritic users submit a personal score between 1 and 10, Metacritic averages the submitted scores from all users, and voila, the user score is created.



Here, we’re seeing a peak around the 8/10 range. This would be an ~80% when compared to the ~70% of Metascores, showing a clear disparity between user scores and critic scores.

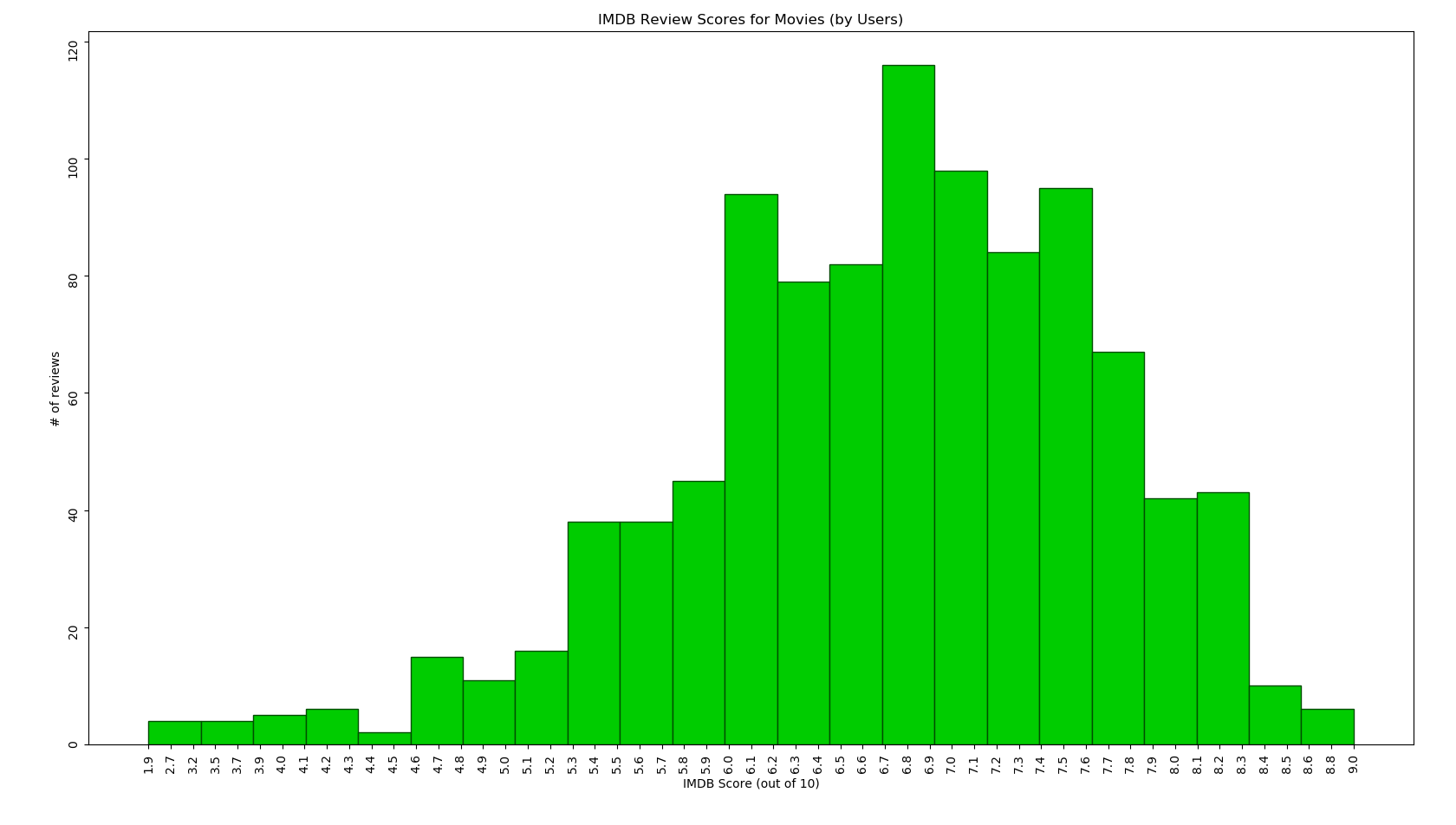
I suspect this is because of a self-selection bias: a user can pick and choose what he wants to play, while a critic often cannot.

I plan to look at this disparity on its own soon.

IMDB Film Ratings:

IMDB’s scoring system is also determined by popular vote among users. [IMDB mentions](https://help.imdb.com/article/imdb/track-movies-tv/faq-for-imdb-ratings/G67Y87TFYYP6TWAV?ref_=helpsect_pro_2_5) that its rating system is a weighted average designed to “eliminate and reduce attempts at vote stuffing by people more interested in changing the current rating of a movie than giving their true opinion of it.”

Suffice to say, these scores are largely determined by users, though some users (who presumably have proven themselves in some way to be reputable at voting) have more of an impact than others.



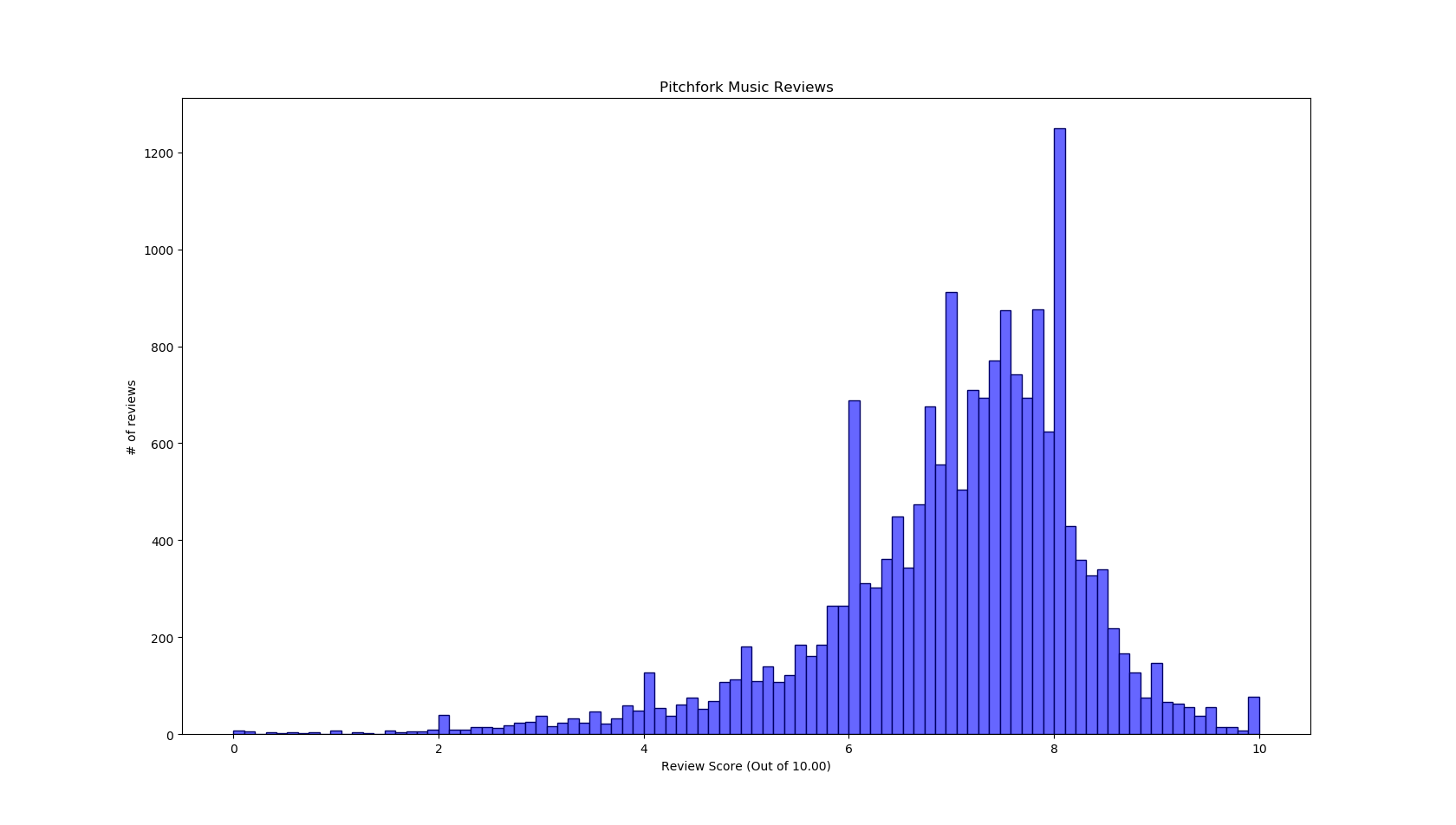
The average IMDB score is definitely lower than the average Metacritic score. The middle of the bell-curve here is clocking in around 6.8, and tapers off aggressively below 6.0 and above 7.6.

This is a little lower than a ‘70%’ average, but it’s certainly higher than a ‘50%’ one, too.

Pitchfork Music Ratings:

Pitchfork’s rating system is far less ‘by the numbers’ than Metacritic or IMDB. A single writer for Pitchfork will review a specific album, giving it both a write-up explaining what was good & bad about it, and also providing a numerical score out of 10 for it. The write-up is ultimately meant to help inform & give rationale for the 0-10 rating, and also provides a degree of accountability for specific scores.

If Pitchfork has a “formula” for their reviews, they don’t have it explicitly posted. [Though given their entire team of reviewers can be viewed on a single page](https://pitchfork.com/masthead/), I would imagine their reviews are more of a practiced art than a science.



Of all the individual datasets, this one is easily the most interesting. Pitchfork has reviews out of 10, but they also give decimal values (rather than just giving an 8/10, they may give an 8.2/10.0). In essence, this means their reviews function similarly to a review out of 100. However, if you look at the above graph, you may notice that, relative to nearby bars, the tallest peaks are on whole numbers. It seems critics are still more apt to assign a ‘6.0’ than they are a 6.1 or 5.9. If you ignore the x-axis for a moment, you can probably make an educated guess as to where 7.0 would be in relation to the peaks for 6.0 and 8.0.

Another trend worth noting is the sheer number of 8.0 reviews. The Median (which, in many ways, signifies the “middle” of our hill of ratings) of this dataset is 7.2, yet 8.0 is where a majority of reviews sit.

The likely explanation for this phenomenon has to do with Pitchfork’s review categories. The site has a special category for ‘8.0+ reviews’(link), which only gives you search results which scored equal to or above that threshold. I suspect there is a similar thought process for many reviewers: “I’d give this a 7.8 or 7.9, but I think it deserves to be seen on the high-score list”. Perhaps you also end up with reviewers who feel something is “the worst of the best”, and that it deserves to be on the high score list, but only at the very lower threshold of it.

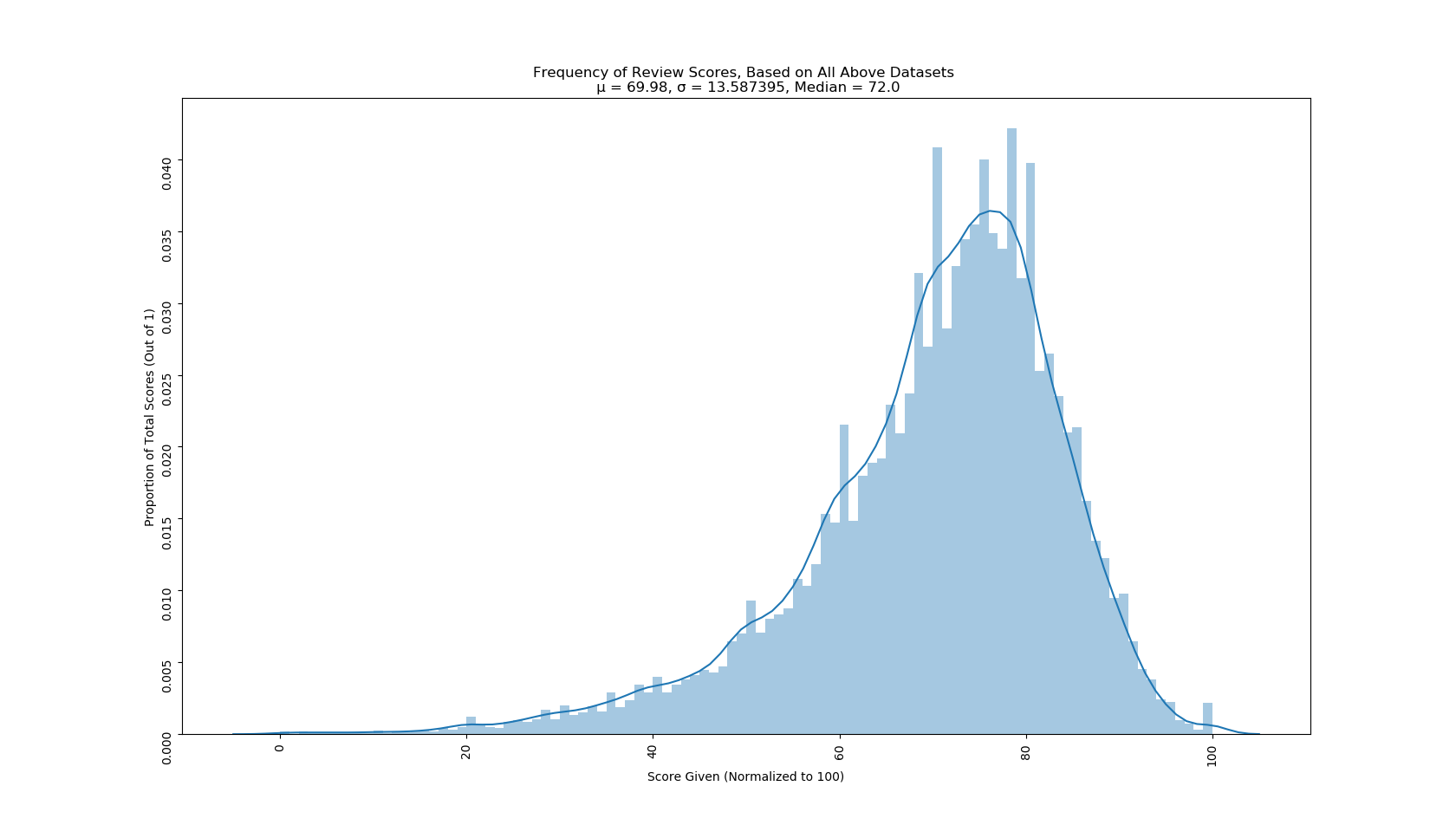
Whatever the reason may be, we see many ratings of around 80%, and a large pool of reviews around the 70% range, too. Once again, it looks as if 5.0/10 or 50% is far from the average score.

All together:

We’ve seen some pretty compelling evidence that the average review for something lies higher than 50/100. But from the looks of it, there’s still a bit of variation from review-system to review-system as to what exactly IS a “middle-score”.

The following is a combination of all four datasets. Since Metacritic User Scores & Pitchfork reviews were only out of 10, they were multiplied by 10, normalizing them to a 100-point scale.

Since we’re looking for the average review (across different websites and creative mediums), I chose to give equal weight to every entry from every dataset here. The IMDB dataset is far smaller than the others, and therefore contributes less to the aggregated average below. However, since we’re trying to find ‘the average review’ independently from whatever rating system or site it is from, this is acceptable.



You may notice that, unlike the previous bar-graphs, this one doesn’t measure the # of entries for each bar, but instead looks at what proportion of reviews that fall into each bar. Rather than simply showing ‘how many’ reviews were a specific score, this helps suggest the standard distribution of these scores. Though there is some fluctuation in the frequency of particular scores, these values are mostly centered around the mid-70-to-low-80 area.

The results here appear to cluster strongly around the mid-70’s range. Both the mean (69.98) and median (72.0) are fairly close to a score of 70/100.

While looking at things from a purely scientific point of view, there are a number of factors to consider which prevent me from generalizing these results to talk about EVERYONE. Mainly, all of these reviews are collected from the internet and are operated/designed with an English-Speaking audience in mind.

That being said, the implications of these results are fairly clear: For internet reviews, at least, a low-to-mid 70/100 review can be interpreted as ‘average’.

I suspect the reason for this is the North American schooling system and its propensity for viewing 70-75% as an ‘average’ grade. Perhaps this grading system has “primed” people to evaluate and assess using the same scale. I will leave testing this theory to another analysis, however.

There is a particular grading system common to [America](https://nces.ed.gov/nationsreportcard/hsts/howgpa.aspx), [Canada](www2.gov.bc.ca/gov/DownloadAsset?assetId=8CE36B223FE640A6A6DF748388381677), and [Mexico](https://www.classbase.com/countries/Mexico/Grading-System) that uses a percentage grade out of 100. In this system, 100% is a flawless grade, and anything below 50% is a failing grade. If a <50% is a failure, and 100% is perfect, 75% falls somewhere in the “middle” of unsatisfactory and perfect. The best way to reinforce this hypothesis would be to compare these datasets to ratings from a culture which uses a separate academic grading system. I currently have no such datasets to test if this effect is cross-cultural, but will investigate into it in future analyses.