Received: 3 July 2019 | Revised: 4 February 2020 | Accepted: 10 February 2020
DOI: 10.1111/test.12220

ORIGINAL ARTICLE

WILEY

Bayesian statistics in the classroom: Introducing shrinkage with basketball statistics and the internet movie database

Arthur Berg

Division of Biostatistics & Bioinformatics, Penn State University, State College, Pennsylvania

Correspondence
Arthur Berg, Division of Biostatistics & Bioinformatics, Penn State University, State College, PA.
Email: berg@psu.edu

Summary
This article promotes Bayesian intuition by introducing the concept of shrinkage with some motivational examples involving basketball statistics of Rudy Gobert and movie rankings from the Internet Movie Database. Supplementary and fully automated R code is provided to allow students to explore shrinkage in this dataset on their own. The content is targeted at the first-year undergraduate level. Students with prior exposure to R would most benefit from the supplementary R code.

KEYWORDS
Bayesian statistics, posterior distribution, R, ranking

1 | INTRODUCTION TO SHRINKAGE

Shrinkage (noun): the process, fact, or amount of shrinking.

Shrinking (adjective): becoming smaller in size or amount.

Shrinkage is an important concept in statistics, but it is usually not directly discussed in the classroom. We start by providing a mathematical formulation of shrinkage, but the reader is encouraged to first think about formulating their own definition of shrinkage first.

After thinking about shrinkage for a moment, one would likely conclude that shrinkage is in fact a relative concept—values are shrunk toward a reference value. A common reference value is taken to be zero, so if we say a number x (positive or negative) is shrunk toward zero, we could write the shrunken number as $y = \rho x$, where $0 \leq \rho \leq 1$. In particular, if we are shrinking a number x toward 0, then y should be somewhere on the interval between 0 and x. More generally, we can describe the shrinking of a number x toward another number x_0 as a convex combination of x_0 and x; that is, we let

$$y = \rho x + (1 - \rho) x_0, \tag{1}$$

where $\rho \in [0, 1]$. Therefore, y lies on the interval from x_0 to x and more shrinkage toward x_0 is represented by a smaller value of ρ.

To understand how shrinkage relates to statistics, we introduce a simple example relating to Basketball statistics and apply a Bayesian inferential approach. Readers unfamiliar with Bayesian Statistics would still benefit from the general ideas presented by this example.

2 | MODELING RUDY GOBERT'S FIELD GOAL PERCENTAGE

Suppose we were interested in estimating the field goal percentage of a particular NBA player after he returns to the game following an injury such as an Achilles tear. We will represent the unknown shooting percentage as θ. Unlike classical statistics, a Bayesian approach assumes that θ is not fixed but an unknown parameter, and beliefs on its possible values are expressed through a probability distribution. This might be a little difficult to grasp as first, but intuitively we can understand this as the variability associated with the NBA player's performance. But if we wanted a *single number*—not a distribution of values—to represent the player's shooting percentage, a natural approach would be to use the mean of the distribution of θ, which would typically be written as $\hat{\theta} = E[\theta]$.

There happens to be ample performance data for that player prior to the injury, but after the injury the player has only played a few games, providing limited data to estimate the shooting percentage postinjury. Therefore, we would intuitively estimate the postinjury shooting percentage as some weighted average of the preinjury shooting percentage and postinjury shooting percentage; that is, we will shrink the postinjury shooting percentage by some amount toward the preinjury shooting percentage. This is precisely how our Bayesian estimate $\hat{\theta}$ will be ultimately produced, however the Bayesian inferential approach provides more clarity as to how much we would let the pre-injury data influence the inference of the postinjury performance.

To make things a little more concrete, we will look at a specific NBA player—Rudy Gobert—a 2.16 m (7 ft 1 in) French player playing for the Utah Jazz. Among his many accomplishments as an NBA star, he had the highest field goal percentage (66.9%) in 2019. The focus of this example is on predicting Rudy's field goal percentage for the 2015 to 2016 season after he sustained a serious injury (medial collateral ligament sprain on his left knee) on December 2, 2015. After missing 18 games, Rudy returned to action on January 7, 2016, but his court time, at first, was limited to 15 minutes. Prior to his injury, Rudy hit 47 field goals out of 84 yielding a 56% field goal percentage. In his first three games after returning, he scored 9 field goals out of 10 yielding an impressive (though most certainly nonsustainable) field goal percentage of 90%. We are now interested in predicting Rudy's field goal percentage for that season after he returned from his injury just from the limited provided data.

To construct our Bayesian model, we need to determine a prior distribution on θ; that is, we must come up with a best guess for Rudy's field goal percentage just based on prior-injury data while also accounting for the uncertainty in that guess. A natural probability distribution to model θ is to use the two-parameter beta distribution. It is common to use the beta distribution to model proportions, like the field goal percentage, as it is a versatile distribution with support on the interval [0,1]. The two parameters of the beta distribution—α and β—govern the shape of the distribution. The mean of the beta distribution is $\frac{\alpha}{\alpha+\beta}$, so we would select the parameters α and β to have a mean close to 0.56, but there is still the issue of the variance to work out. If you simply had no idea as to what a reasonable value for θ would be upon recovering from the injury, a natural choice of prior distribution would be to take a Beta(47,37) distribution, representing 47 successful field goals out of $47 + 37 = 84$ attempts. This distribution has a mean of 56%, matching the preinjury field goal percentage, and has 95% of the distribution within the interval between 45% and 66%. However, we may be inclined to select a prior distribution with greater variability due to the uncertainty associated with how his performance might change. In particular, if we thought that upon returning to the game Rudy's field goal percentage could be anywhere in the range of 40% to 70%, then we would find parameters α and β so that most of the distribution (say eg, 95% of the distribution) lies inside the interval [0.4,0.7] with a mean that is around 0.56. After a little bit of experimentation, one can find that a Beta(23,18) distribution has the property of having a mean of 0.56% and 95% of its distribution in the interval [0.4,0.7] (see Figure 1A).

Note that this prior distribution attempts to model our best guess at Rudy's field goal rate *after* he returns to the game from his injury, but the only data we used to construct this distribution was the data available prior to the injury. We can incorporate our uncertainty in this rate by selecting a distribution that ranged from 40% to 70% with concentration around 56%, which leads to the Beta(23,18) distribution depicted in Figure 1A. Alternatively, we might consider a prior distribution that is even more diffuse, say with a Beta(8,6) distribution, which can be interpreted as representing prior data of eight successful field goals out of $8 + 6 = 14$ attempts. We see that this prior, as depicted in Figure 1B, is much more uncertain about Rudy's field goal percentage upon returning to the game.

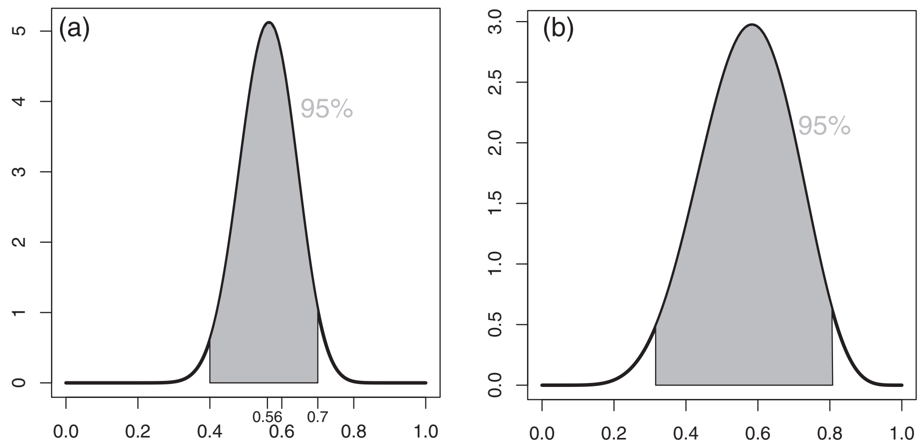
Now we incorporate the new data provided by the first three games after Rudy's return from injury (hitting 9 field goals out of 10 attempts). The simplest approach to model the field goal success rate is to use a binomial distribution model. More specifically, we assume the number of successful field goals out of the 10 field goal attempts follows a binomial distribution. We use Bayes' Theorem to meld together the prior assumption (distribution) with the available data. Using the notation $\pi(\theta)$ to represent the prior distribution and $f(x|\theta)$ to represent binomial likelihood, the resulting melded distribution, formally called the posterior distribution in Bayesian Statistics, has the following mathematical form.

$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{\int \pi(\theta) f(x \mid \theta) d\theta}.$$

This can be interpreted as updating the prior distribution $\pi(\theta)$ with the data x (through the likelihood $f(x|\theta)$) to yield the posterior distribution $\pi(\theta|x)$. We will not delve any further into the details of calculating the posterior distribution for this example except to say that the posterior distribution in this case also turns out to be a beta distribution. In particular, for a prior with Beta(α,β) distribution and a binomial likelihood with x successes out of n attempts, the posterior distribution is a Beta($\alpha + x$, $\beta + n - x$) distribution. In this Rudy Gobert example, having observed $x = 9$ successes out of $n = 10$ attempts, the posterior distributions corresponding to the two prior distributions in Figure 1 are Beta($23 + 9 = 32, 18 + 10 - 9 = 19$) and Beta($8 + 9 = 17, 6 + 10 - 9 = 7$), and these distributions have means of $\frac{32}{51} \approx 63\%$ and $\frac{17}{24} \approx 71\%$, respectively.

In Figure 2, we display how the postinjury mean of 90% is shrunk toward the prior mean to provide a more robust and comprehensive estimate of the field goal percentage that integrates the limited postinjury data with the prior data. Furthermore, we see that the amount of shrinkage toward the prior decreases with decreased certainty of the prior information. Similarly, when more data is available postinjury, less shrinkage will occur toward the prior mean. For those curious as to Rudy's performance in the rest of the season, he ended up playing 61 games that season after his injury making 151 field goals out of 270 attempts yielding a final postinjury percentage that precisely matched his preinjury percentage of 56%. So, in this example, the less conservative prior in Figure 1A proved more accurate compared to the more uncertain prior in Figure 1B.

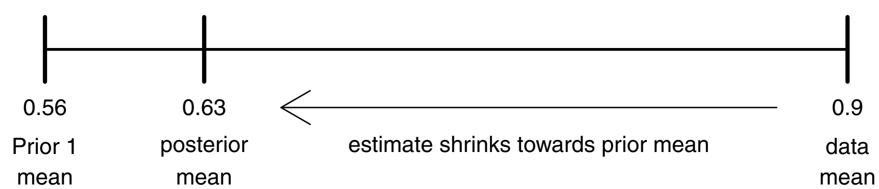
More generally, following a Beta(α,β) prior for θ, the posterior mean can be written as a convex combination



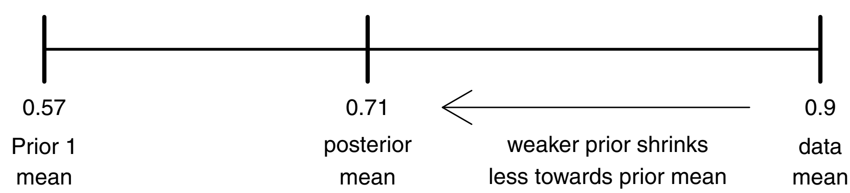
Beta(23,18) distribution with mean 0.56 and 95% of the distribution in the interval [0.4,0.7].

Beta(8,6) distribution with mean 0.57 and 95% of the distribution in the interval [0.32,0.81].

FIGURE 1 Possible prior distributions to model Rudy Gobert's postinjury field goal percentage. A, Beta (23,18) distribution with mean 0.56 and 95% of the distribution in the interval [0.4,0.7]. B, Beta(8,6) distribution with mean 0.57 and 95% of the distribution in the interval [0.32,0.81]



(a) Showing how the data estimate of 90% is shrunk towards the prior mean of 56% when using the prior in Figure 1a.



(b) Showing how the more uncertain prior represented in Figure 1b will result in less shrinkage of the data-based estimate.

FIGURE 2 Visualizing shrinkage in the Rudy Gobert field goal percentage Bayesian model corresponding to the priors in Figure 1. A, Showing how the data estimate of 90% is shrunk toward the prior mean of 56% when using the prior in Figure 1A. B, Showing how the more uncertain prior represented in Figure 1B will result in less shrinkage of the data-based estimate of the data-based estimate $\frac{x}{n}$ (also referred to as the maximum likelihood estimator) and the prior mean $\frac{\alpha}{\alpha+\beta}$ in a similar structure to Equation (1). More specifically, we can write

$$\underbrace{\frac{x+\alpha}{n+\alpha+\beta}}_{\text{posterior mean}} = \underbrace{\frac{x}{n}}_{\text{data mean}} \underbrace{\left(\frac{n}{n+\alpha+\beta}\right)}_{\rho} + \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{prior mean}} \underbrace{\left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)}_{1-\rho},$$

where $\rho \in [0, 1]$.

We will now look at another example of shrinkage in statistics and provide statistical code for students to further explore the effects of shrinkage on their own.

3 | SHRINKAGE IN MOVIE RATINGS

Say it is movie night tonight, but you do not know what to watch, so you open up the Internet Movie Database (IMDb) to find a movie from their latest top 250 list (imdb.com/chart/top). However, after opening up the website you notice this list really has not changed much since the last time you checked it. In fact, the highest ranked movie—"Shawshank Redemption"—has been in the top position since 2008 when it overtook "The Godfather".[1] This may lead you to question the approach IMDb uses to ranks its movies. It turns out that *shrinkage* is at the heart of its ranking methodology, but first, we will take a closer look at IMDb and its data.

IMDb is owned and operated by IMDb.com, which is currently a subsidiary of Amazon. As of May 2019, IMDb has approximately 6 million titles, including television episodes, and 83 million registered users. IMDb permits any movie that was released and/or screened to the public at least once to be included in the database and voted on. Therefore, many small budget and esoteric movies are included in the database, but, of course, YouTube videos are not included. Any registered user can rate the movies with rating on a scale of 1 star (worst) to 10 stars (best). IMDb makes various undisclosed efforts to filter out inappropriate voting such as vote stuffing from the database by users more interested in changing the current rating of a movie than giving their true opinion of it.

Now consider you wanted to construct your own all-time top 250 movie list from the IMDb rating data. If you include all movies in your search, including movies with only a few votes cast, you would likely end up with a list of perfect 10-star rated movies that no one's heard of before. For example, at the time of writing this article, the movie "The Do's & Don'ts of Sharing an Apartment" is a comedy with a perfect 10-star rating, yet it was only voted on by five users. It is quite clear that such a naïve approach to finding the best-ever movies does not yield a reliable or believable top-ranked list. In order to mitigate the influence of movies with only a small number of votes, the IMDb requires at least 25 000 user votes in order to be considered in their top-ranked list. The threshold of 25 000 is somewhat arbitrarily chosen and is explored later on in this article, but it is intuitively clear that top-rated list becomes more stable as this minimum threshold of votes increases. We intuitively understand that the average rating of a movie that just made it over the 25 000 vote threshold (eg, the Indian movie Drishyam) will still be much less reliable than a movie with over 2 million votes (such as the movie Shawshank Redemption). So how might we adjust the ratings to incorporate information on the number of votes cast? This brings us now to the notion of shrinkage.

When we are less sure about the rating of a film due to its low number of votes, a reasonable approach is to adjust (or shrink) the rating that is based on only a few votes toward the average rating across all movies in the database. This approach is quite understandable when we consider the extremes. Indeed, the best guess for a new movie with no votes would be the overall database average, or if the movie just had a few votes, then we would still use the overall database average but with a very slight adjustment toward the average user vote. On the other extreme, a movie with over a million votes does not need to be adjusted or shrunk much to the overall mean.

More generally, we described the weighted rating (WR) as a convex combination (weighted average) of the movie's original rating (R) and the average rating across all included movies (C). Mathematically, we can express this in the same form as Equation (1) as

$$\mathrm{WR} = \rho_v R + (1-\rho_v) C, \tag{2}$$

where R represents the movie's original rating, C represents the average rating across all of the movies, and ρ_v is a value between zero and one that determines the weights in the average. The subscript on ρ_v reinforces that notion that the weights should vary with the number of votes, v, for the movie. Essentially, with an extremely large number of votes, we would trust the movies rating and set ρ_v equal to one. Similarly, a smaller number of votes would correspond with a smaller ρ_v.

One key question is to determine a good choice of ρ_v. In order to determine an optimal value of ρ_v, we would need some notion of a "gold standard rating" for the movies. For example, it might be of interest to use a different rating system, such as the Rotten Tomatoes rating, to serve as the gold standard rating, but the substantial differences between the different rating methods limit the feasibility of such an approach. Therefore, for this example we will not seek to identify a statistically optimal value of ρ_v, but, instead, we will somewhat arbitrarily select a value of ρ_v that seems reasonable. This leads us to the method invoked by IMDb for their Top 250 Movie Chart. In particular, IMDb uses the following choice of ρ_v:

$$\rho_v = \frac{v}{v+m},$$

where m represents the minimal number of votes required to make it into the top 250 list (so at present $m = 25000$).

This example showcases how shrinkage is seen to be at the heart of the IMDb Top Rated Movie Chart. In particular, movie ratings are shrunk toward the average movie rating, and movies with more votes cast shrink less compared to movies that have fewer votes cast. Next, we introduce the reader to implementing these ideas with the latest datasets provided by IDMb.

4 | SHRINKAGE IMPLEMENTATION WITH R MARKDOWN

The Internet Movie Database provides a substantial amount of their data as seven publicly available compressed tab-delimited datafiles on the IMDb website (https://www.imdb.com/interfaces/). We will just use two of those files for analyzing the rankings - title.basics.tsv and title.ratings.tsv—which are detailed below.

- title.basics.tsv
- tconst (string)—alphanumeric unique identifier of the title
- titleType (string)—the type/format of the title (eg, movie, short, tvseries, tvepisode, video, etc)
- primaryTitle (string)—the more popular title/the title used by the filmmakers on promotional materials at the point of release
- startYear (YYYY)—represents the release year of a title. In the case of TV series, it is the series start year
- genres (string array)—includes up to three genres associated with the title
- Other variables will not be used
- title.ratings.tsv
- tconst (string)—alphanumeric unique identifier of the title
- averageRating—weighted average of all the individual user ratings
- numVotes—number of votes the title has received

These two datasets are linked together with the `tconst` variable. Only movies—no tv shows—will be included by filtering the data with the variable `titleType`.

The freely available and open source statistics program R ([2]) is used to analyze the data, and R Markdown tools [3,4] are used to prepare convenient html documentation of the output. The R code also utilizes some additional functions from user-generated libraries. These libraries include `R.utils` [5] for decompressing the downloaded files and `kableExtra` [6] for a more elegant display of the tables in R Markdown. The code is most conveniently run within the freely available software program R Studio.[7]

The provided software should be fully automatic in that it contains code to download the required datasets, then load the data, and then finally analyze the data and produce the same tables presented in this article below. The software also utilizes the cache feature in R Markdown, which allows the somewhat time-consuming steps of downloading and importing of the data to be cashed. Hence, after the software is run through once, subsequent attempts to run the software will only require a small fraction of the original amount of time to complete. It should be noted, however, that care should be used with the caching feature activated as subsequent modifications may be affected by earlier cashed code. An easy resolution to this is to simply delete the entire cache from the working directory.

In Table 1, we list out the top 10 rated movies among all movies with at least 25 000 votes cast.

TABLE 1 Top 10 ranked movies based on the average voter ratings

Rank	Top ranked titles (no shrinkage)	# of votes (v)	Original rating (R)
1	The Chaos Class	34 329	9.4
2	The Shawshank Redemption	2 142 788	9.3
3	The Godfather	1 470 882	9.2
4	CM101MMXI Fundamentals	41 973	9.2
5	The Dark Knight	2 108 361	9.0
6	The Godfather: Part II	1 023 498	9.0
7	Joker	183 751	9.0
8	The Mountain II	101 502	9.0
9	Pulp Fiction	1 681 729	8.9
10	The Lord of the Rings: The Return of the King	1 523 053	8.9

Note: Only movies with at least 25 000 votes are considered. No shrinkage to the ratings is applied here.

Rank	Top titles with $m = 25000$	# of votes (v)	R	WR
1	The Shawshank Redemption	2 142 788	9.3	9.3
2	The Godfather	1 470 882	9.2	9.2
3	The Dark Knight	2 108 361	9.0	9.0
4	The Godfather: Part II	1 023 498	9.0	9.0
5	Pulp Fiction	1 681 729	8.9	8.9
6	The Lord of the Rings: The Return of the King	1 523 053	8.9	8.9
7	Schindler's List	1 111 836	8.9	8.9
8	12 Angry Men	611 569	8.9	8.8
9	Inception	1 878 747	8.8	8.8
10	Fight Club	1 712 479	8.8	8.8

TABLE 2 Top 10 ranked movies based on the weighted ratings (WR) with $m = 25000$. Only movies with at least 25 000 votes are considered

Rank	Top titles with $m = 2500$	# of votes (v)	R	WR
1	The Shawshank Redemption	2 142 788	9.3	9.3
2	The Chaos Class	34 329	9.4	9.2
3	The Godfather	1 470 882	9.2	9.2
4	CM101MMXI Fundamentals	41 973	9.2	9.1
5	Wheels	17 379	9.3	9
6	The Dark Knight	2 108 361	9	9
7	The Godfather: Part II	1 023 498	9	9
8	Joker	183 751	9	9
9	The Mountain II	101 502	9	9
10	Aynabaji	18 693	9.2	8.9

TABLE 3 Top 10 ranked movies based on the weighted ratings (WR) with $m = 2500$. Only movies with at least 2500 votes are considered

In Table 2, we utilize the IMDb published formula to produce a WRs and extract the top 10 movies based on the WRs.

Comparing Table 2 to Table 1, we see that when a greater degree of shrinkage is applied as in Table 2, only movies with a large number of votes (over a half-million votes as it happens) made the top 10 list. In other words, movies that are vying for the top 10 positions on the IMDb Top Ranked Movie Chart practically need a 9+ rating with at least a half-million votes. Clearly this list is catering to a populist perspective. By reducing the amount of shrinkage as governed by the parameter m and reducing the minimum number of votes required for consideration in the top ranked list, we allow for somewhat less popular movies to be considered among the top 10 ranked movies. Table 3 lists the top 10 movies based on WRs with $m = 2500$ and considering all movies with at least 2500 votes.

With this smaller amount of shrinkage applied, we see more esoteric movies in the top 10 list. In particular, the very recently released movie "Joker" has already made it onto this list. On the other hand, we also see some likely unfamiliar Turkish movies in this list—"The Chaos Class," "M101MMXI Fundamentals," and "The Mountain II." Evidently, IMDb is very popular in Turkey. An exercise for the aspiring student would be to use R and the available data from IMDb to filter out non-English language films prior to constructing the top ranked lists.

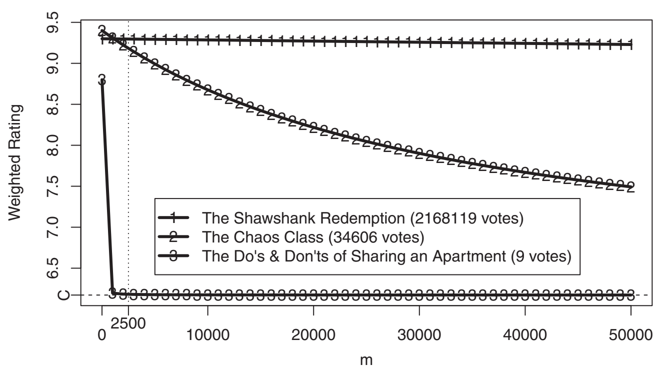


FIGURE 3 The effect of shrinkage by varying m

We see that this last list in Table 3, representing say moderate shrinkage, can be thought of as a compromise between Table 1 (no shrinkage, $m = 0$) and Table 2 (heavy shrinkage, $m = 25000$). Of course, one could apply even more shrinkage than what was depicted in Table 2, and as the number of votes cast on IMDb continues to climb, IMDb at some point will presumably increase the shrinkage parameter m when calculating the top 250 list.

We also point out that although some movies with a relatively few number of votes are effected, there is a handful of movies with a very high number of votes and a very high rating which therefore are found across all of the tables. An exercise for the reader is to study the effect of shrinkage on the lowest rated movies for movies receiving at least 25 000 votes.

Finally, we illustrate the effects of the shrinkage parameter m on three highly rated movies but with wildly varying number of votes. As shown in Figure 3, the rating for "The Shawshank Redemption," which has over 2 million votes, is hardly changed for values of m less than 50 000. However, "The Do's & Don'ts of Sharing an Apartment," with only nine votes, very quickly shoots down to the overall mean. A movie with a moderate amount of votes like "The Chaos Class" with 34 606 votes has a more gradual decline toward the overall mean.

5 | CONCLUSION

Shrinkage is an important concept in Statistics, which can be intuitively described as "learning from the experiences of others" in which individual data values are shrunk toward the grand mean. The optimal amount of shrinkage is a key question to be worked out. In certain contexts, an appropriate amount of shrinkage can be determined from the data alone; however, in other contexts, such as the IMDb example explored here, the amount of shrinkage is somewhat arbitrarily chosen. Applications of shrinkage are seen in many diverse contexts, and in particular it is a fundamental concept in Bayesian statistics. Finally, fully detailed R code accompanies this article allowing students to delve into the details of the analysis and apply their own shrinkage to the IMDb data.

ACKNOWLEDGEMENTS

The author acknowledges the tremendous effort and encouragement provided by the editor, Dr. Helen MacGillivray, and two anonymous reviewers.

ORCID

Arthur Berg https://orcid.org/0000-0002-4097-7348

REFERENCES

1. M. Heiderny, The little-known story of how the shawshank redemption became one of the most beloved films of all time, *Vanity Fair*, 2014, available at https://www.vanityfair.com/hollywood/2014/09/shawshank-redemption-anniversary-story.
2. R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019 available at https://www.R-project.org/.
3. J. Allaire et al. *rmarkdown: Dynamic Documents for R*, 2019, available at https://github.com/rstudio/rmarkdown.
4. Y. Xie, J. Allaire, and G. Grolemund, *R Markdown: The definitive guide*, Chapman and Hall, Boca Raton, FL, 2018 available at https://bookdown.org/yihui/rmarkdown.
5. H. Bengtsson, *R.utils: Various programming utilities*, 2019, available at https://CRAN.R-project.org/package=R.utils.
6. H. Zhu, *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*, 2019, https://CRAN.R-project.org/package=kableExtra.
7. RStudio Team, *RStudio: Integrated development environment for R*, RStudio, Boston, MA, 2015 available at http://www.rstudio.com/.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Berg A. Bayesian statistics in the classroom: Introducing shrinkage with basketball statistics and the internet movie database. *Teaching Statistics*. 2020;1–7. https://doi.org/10.1111/test.12220