Application of Bayesian Credibility Theory in Movie Rankings to Reduce Financial Risk of Production Houses

Palash Ranjan Das
Dept of Applied Mathematics, University of Calcutta, Kolkata
palash.iitm@gmail.com

Tripti Chakrabarti
Dept of Applied Mathematics, University of Calcutta, Kolkata
triptichakrabarti@gmail.com

ABSTRACT

Credibility theory is a branch of actuarial science devoted to quantify how unique a particular outcome will be compared to an outcome deemed as typical.

In this paper, we will examine the application of the principles of Bayesian Credibility Theory in rating and ranking movies by a premier online movie database based on user's votes. Although the Bayesian credibility theory was developed originally as a method to calculate the risk premium by combining the individual risk experience with the class risk experience, it is generic enough to deal with a wide range of practical applications quite different from the classical application mentioned above. One such diverse application of the theory in an unlikely domain will be discussed in this paper.

Keywords: *Credibility Theory; Prior distribution; Likelihood function; Posterior distribution; Loss function; Bayesian approach*

Introduction

Undoubtedly, the financial risk is a primary characteristic of the motion picture industry-prediction of demand is notoriously difficult and almost all costs are incurred before any demand is realized. Thus to suggest an appropriate financial strategy is the key variable that shapes the film industry. It thus became necessary for production houses to consult a reliable database of movie rankings so that they can co-finance a motion picture by considering its ranking on the basis of user's votes. Simultaneously, it became imperative for movie databases to rank the movies using robust statistical techniques so as to avoid any controversy regarding these rankings, the whole process being statistically sensitive in nature.

In this paper, we will analyse the statistical approach adopted by a well-known online movie database in rating and ranking movies based on user's votes and will argue on the superiority of its approach over other traditional unsophisticated straightforward approaches known so far, thereby reducing the risk of haphazard and unscientific rating of movies. This makes the database more reliable among the production houses which can now consult it freely to finance a movie thereby laying more importance on the public's opinion about box office hits through user's votes.

This minimizes their risk of financially backing unproductive movies whose ranking based on user's votes do not seem satisfactory. Thus it is almost certain that such movies will flop in the movie business market and producing them is financially hazardous. So production houses can accept or decline to financially back a movie on the basis of its ranking by some reliable movie database. As mentioned earlier, this substantially reduces the financial risk associated with the production in movie business industry.

The paper is organized as follows: Literature review on credibility theory and objectives of this study are followed by an overview of IMDb, a well-known online movie database and the process by which it rates and ranks movies is discussed. It is followed by explaining the statistical formula adopted by IMDb for calculating the Top Rated 250 titles and the similarity in the formula adopted with the Bayesian credibility theory formula used by actuaries of insurance companies to calculate premiums. Thereafter, we discuss Bayesian estimation in general and its use in determining credibility estimates. Subsequently, application of credibility estimates in movie rating problem is presented and analysed in details. Further a comparison between the classical premium pricing insurance risk problem and the movie rating problem is shown in the form of a table. Numerical results and findings are then presented. Finally the conclusion is offered along with scope for further research.

Literature Review

In actuarial parlance the term 'credibility' was originally attached to experience rating formulae that were convex combinations (weighted averages) of individual and class estimates of the individual risk premium. Credibility theory thus was the branch of insurance mathematics that explored model-based principles for construction of such formulae. The development of the theory brought it far beyond the scope so that in today's usage credibility covers more broadly linear estimation and prediction in latent variable models.

The origin and advent of credibility theory dates back to Whitney[11] who in 1918 addressed the problem of assessing the risk premium 'm', defined as the expected claim expenses per unit of risk exposed for an individual risk selected from a portfolio (class of similar risks). Incorporating and advocating the combined use of individual risk experience and class risk experience, he proposed that the premium rate be a weighted average of the form:

$$\bar{m} = z^*\hat{m} + (1\text{-}z)^* \text{ì}$$

Where, '$\bar{\hat{m}}$' is the observed mean claim amount per unit of risk exposed for the individual contract and 'ì' is the overall mean in the insurance portfolio.

Whitney viewed the risk premium as a random variable. In terms of modern credibility theory, it is a function of m(è) of a random element è representing the unobservable characteristics of the individual risk. The random nature of è signifies and expresses the notion of heterogeneity, the individual risk is a random selection from a portfolio of similar but not identical risks and the distribution of è describes the variation of individual risk characteristics across the portfolio. It is to be noted that the weighted 'z' in the above formula was defined as the credibility factor since it measures the amount of credence attached to the individual experience and '$\bar{m}$' was called the credibility premium.

T.Bauwelinckx.*et.al* (1991) [6] in their study on loaded credibility premium introduced a new technique for estimating credibility premium risks, containing a fraction of the variance of the risk as loading on the n et insurance premium. This method provides us with another approach to the known results for credibility loaded premiums, not having the drawback of estimating an approximation of the so-called fluctuation part. It also provides us with an elegant extension to loaded premiums in the hierarchical credibility model. The results are obtained in the semi linear hierarchical credibility theory.

E. Gomez-Deniz (2007) [7] considered an alternative to the usual credibility premium that arises for weighted balance loss function. He generalizes the credibility theory by balance loss function and it includes as a particular case the weighted quadratic loss function traditionally used in actuarial science. This function is used to derive credibility premiums under approximate likelihood and priors. Further generalized credibility premiums are obtained that contain as particular cases other credibility premiums.

Jean-Philippe Boucher and Michel Denuit (2007) [9] explored and compared the credibility premiums in zero-inflated Poisson models for panel data. They derived predictive premiums based on quadratic loss and exponential loss. They showed that the credibility premiums of the zero-inflated model allow for more flexibility in the prediction and argued that the future premiums not only depend on the number of past claims but also on the number of insured period with at least one claim. Their model also analysed in another way the hunger for bonus phenomenon.

Harald Dornheim and Vytaras Brazauskas (2010) [8] embedded the classical credibility theory models within the framework of mixed linear models with the objective to develop robust and efficient methods of credibility when heavy tailed claims are approximately log-local-scale distributed. To accomplish that, they expressed additive credibility models as mixed linear models with symmetric or asymmetric errors. They adjusted adaptive truncated likelihood methods and compute highly robust credibility estimates for heavy-tailed claims.

Joseph H.T. Kim and Yongho Jeon (2013) [10] in their study proposed a credibility theory which is based on truncation of loss data, or the trimmed mean. Their proposed framework addresses the classical credibility theory as a special case and is developed on the idea of varying the trimming threshold to investigate the sensitivity of the credibility premium. They showed that the trimmed mean is not a coherent risk measure and investigated some related asymptotic properties of the structural parameters in credibility. They finally showed that the proposed credibility models can successfully capture the tail risk of the underlying loss model, thus providing a better landscape of the overall risk that insurers assume.

This widespread research on credibility theory has opened new avenues of application of the theory hitherto unknown. Thus now, its applications were not only confined to the classical premium risk problem but also to other diverse fields.

Prasham M. Rambhia (2015) [4] hinted at the application of the theory by IMDb to rank and rate movies based on user's votes. Although detailed calculations and statistical theory is missing in the study, he gave an overall picture of the unusual application of this usual theory known so far. His article was an attempt to examine a diverse application of the credibility theory actuaries encounter in their curriculum to a field that is as different from actuarial science as chalk is from cheese.

With further passage of time we will certainly observe more interesting, myriad and diverse applications of such known statistical theories.

Objectives

The present study focuses primarily on the approach adopted by IMDb for ranking the Top Rated 250 titles. The ultimate purpose of this research is to study the diverse application of Bayesian credibility theory in movie rating problem thereby exploring new avenues of application of known statistical theories.

The paper primarily aims to:-

1. Examine the financial risk of production involved in movie business industry and how production houses can minimise financial risk by consulting movie ranking database.
2. Study the application of Bayesian credibility theory in movie rankings.

3. Exhibit the similarity in approach of the movie rating problem and the premium calculation problem of insurance companies.
4. Finding the unknown parameter and credibility estimates of the movie rating problem.
5. Demonstrating the formula of WR (the weighted rating) in the form of the credibility estimate. The credibility factor is also shown as the function of number of user's votes (v) and minimum number of required votes (m).

IMDb and Movie Ratings:

Internet movie database (abbreviated as IMDb) [1] is a premier movie database which rates and ranks movies based on cinephile votes. Most cinema lovers use it to know movie ratings and collect other ancillary information about movies. Each registered user is eligible to rate each movie. The rating of a movie is done by assigning a positive integer score of 10, where 10 is regarded as the highest score possible. Each such rating is regarded as a 'vote' by an individual registered user. For each movie, the average rating from various individual users (say, R) is computed and displayed.

Incorporating the special popular feature 'Top 250' chart of IMDb [2], we are mainly concerned about the theory adopted by IMDb in determining these rankings of movies given that the average ratings of each movies(and other ancillary data like number of votes of each movie, etc) is known.

Approach adopted by IMDb

A screenshot taken from IMDb [2] reveals the formula adopted by them.

The formula for calculating the Top Rated 250 Titles gives a true Bayesian estimate:

$$\text{Weighted Rating (WR)} = \left(\frac{v}{v+m}\right) * R + \left(\frac{m}{v+m}\right) * C$$

Where:-

R= average for the movie (mean)=(Rating)

v= number of votes for the movie= (votes)
m = minimum number of votes required to be listed in the Top 250 (at present 25000)

C= the mean vote across the whole report (at present 7.0)

Note:- For the Top 250, only votes from regular voters are considered.

It is interesting to notice that the weighted rating (WR) used for ranking movies is a weighted average between R (the movie's own average rating based on user's votes on it) and C(average rating of all movies). Clearly WR will lie between R and C. Also it is worth mentioning that R and C are averaged with weights in the ratio v: m. Hence, a higher value of v implies more weightage being given to R, than C. This is intuitively reasonable as R is more relevant if v is large. When v tends to infinity, the weighted rating (WR) approaches R which is also intuitively obvious.

The formula of WR can be rearranged and written as:

WR $= (\frac{v}{v+m}) * R + [1-(\frac{v}{v+m})] * C$

$= Z*R + (1-Z)*C$, where $Z= (\frac{v}{v+m})$

Thus we see that the weighted rating adopted by IMDb fits exactly to the Bayesian Credibility Theory adopted by insurance companies to calculate risk premiums. Here the credibility factor Z= $(\frac{v}{v+m})$ which is a real number lying between 0 and 1 consistent with the usual theoretical development.

However the discussion appears vague and intuitive without a strong and rigorous statistical foundation which we will deal in the subsequent sections.

Bayesian Estimation and its use in determining credibility estimates

The Bayesian approach to credibility is discussed in this section.

Under the Bayesian framework, the unknown parameter (say, è) is estimated on the basis of some observed data (say) by involving the following steps:-

Prior parameter distribution

A prior parameter distribution is adopted to describe the possible values of the unknown parameter under consideration. The form of the prior distribution is derived from the collateral data.

The unknown parameter è is regarded as a random variable which has a specific distribution. Some idea about is it known beforehand without considering the observed data ($\underline{X}$). We call that the prior distribution of è.

Likelihood function

For any given value of the parameter, there is a certain probability of incurring the particular pattern observed in the direct data. This determines the likelihood of a given pattern as a function of the unknown parameter.

On the basis of the observed data ($\underline{X}$) and the probability density function (PDF) of /è, we construct the likelihood function L(/è).

Posterior parameter distribution

The prior parameter distribution is combined with the likelihood function using Bayes' formula to determine a posterior parameter distribution for the parameter.

The Bayes' formula enables us to determine the posterior distribution of è using the observed data X_ by the following relationship:

Posterior PDF Prior PDF* Likelihood

Loss Function

The loss function quantifies the difference between the true value of the parameter and it's estimated one. It shows how serious misjudging the parameter value would be.

We find the Bayesian estimate of the unknown parameter è on the basis of the chosen loss function and the posterior distribution. For the mostly applied 'quadratic error loss', the mean of the posterior distribution is the required optimal Bayesian estimate.

It is to be noted that this Bayesian estimate is regarded as a credibility estimate if it can (after rearrangement) be expressed in the form:-

Z*Mean based on Sample data + (1-Z)*Mean of prior distribution

Thus in the movie rating problem, we will make appropriate choices for the prior distribution, the likelihood function and the loss function to obtain the credibility estimate of the above form in the next section.

Application of Credibility Estimates in movie rating problem

We will use the Binomial/Beta model to realistically approach the movie rating problem in contrary to the more common Bayesian credibility models such as the Poisson/Gamma model and the Normal/Normal model.

Suppose R_j denote the individual rating score by the j^{th} user.

Likelihood function:

Let us assume that for j=1,2,...v, R_j/è are independent and identically distributed as Binomial(10, p). This distribution is realistic and consistent with the demand of the problem since users can only assign integer scores on a scale of 10. The likelihood function will be dependent on 'v' and 'R' (used above) and 'p' where p is the unknown parameter to be estimated.

The likelihood function is thus written as:

$$L(p) = \prod_{j=1}^{v} \binom{10}{j} p^{rj} (1-p)^{10-rj}$$

$$= (\text{constant}).\ p^{\sum_{j=1}^{v} rj} * (1-p)^{10v-\sum_{j=1}^{v} rj}$$

Thus we see that:

$$L(p) \propto p^{\sum_{j=1}^{v} rj} * (1-p)^{10v-\sum_{j=1}^{v} rj}$$

Prior distribution:

'p' denoted the probability which can take all real values between 0 and 1. Hence we consider a Beta(á, â) distribution as a prior. This is because the Beta distribution is the conjugate prior of the Binomial distribution. Particular values of 'á' and 'â' would be functions of 'm' and 'C', denoted before.

Thus if f(p) is the pdf of the prior then:

f(p)=[{Ã(á).Ã(â)}/Ã(á+â)] $p^{\alpha-1} * (1-p)^{\beta-1}$

Thus we see that:

Prior distribution $\propto p^{\alpha-1} * (1-p)^{\beta-1}$

Posterior distribution

The posterior distribution is obtained by multiplying the prior probability distribution function with the likelihood function.

Thus we have:

Posterior distribution

$$\propto p^{\alpha-1} * (1-p)^{\beta-1} * p^{\sum_{j=1}^{v} rj} * (1-p)^{10v-\sum_{j=1}^{v} rj}$$

$$= p^{\sum_{j=1}^{v} rj+\alpha-1} * (1-p)^{10v-\sum_{j=1}^{v} rj+\beta-1}$$

Hence, $\propto p^{\sum_{j=1}^{v} rj+\alpha-1} * (1-p)^{10v-\sum_{j=1}^{v} rj+\beta-1}$

Posterior distribution

That is, ***Posterior distribution***

$$\propto p^{\sum_{j=1}^{v} rj+\alpha-1} * (1-p)^{10v-\sum_{j=1}^{v} rj+\beta-1}$$

Loss function

Using the quadratic loss function, the mean of the above posterior beta distribution will be the Bayesian estimate for 'p'. Let us regard it as $\hat{p}$.

Thus we have: $\hat{p} = \frac{\sum_{j=1}^{v} rj + \alpha}{10v + \alpha + \beta}$

The expected weighted rating for a Binomial (10, p) distribution would be = $\widehat{WR}$=10*$\hat{p}$..

Therefore: $\widehat{WR} = \frac{10 * \sum_{j=1}^{v} rj + 10\alpha}{10v + \alpha + \beta}$

Calculations

Now, we have from definition of 'R' that $\frac{\sum_{j=1}^{v} rj}{v} = R$

Thus: $\sum_{j=1}^{v} rj = v * R$........(1)

Substituting (1) in the expression of $\sum_{j=1}^{v} rj = v * R$........(1), we get:

$$\widehat{WR} = \frac{10vR + 10\alpha}{10v + \alpha + \beta}$$

$$= \left(\frac{10v}{10v + \alpha + \beta}\right)*R + \left(\frac{10\alpha}{10v + \alpha + \beta}\right)$$

$$= \left\{\frac{v}{v + \left(\frac{\alpha+\beta}{10}\right)}\right\}*R + \left\{\frac{\alpha}{v + \left(\frac{\alpha+\beta}{10}\right)}\right\}$$

$$= \left\{\frac{v}{v + \left(\frac{\alpha+\beta}{10}\right)}\right\}*R + \left\{\frac{\left(\frac{\alpha+\beta}{10}\right)}{v + \left(\frac{\alpha+\beta}{10}\right)}\right\}*\frac{\alpha}{\left(\frac{\alpha+\beta}{10}\right)}$$

$$= \left\{\frac{v}{v + \left(\frac{\alpha+\beta}{10}\right)}\right\}*R + [1 - \{\ \ \}]*$$

Thus we have:

$\widehat{WR} = \frac{v}{v+m}*R + \{1 - (\frac{v}{v+m})\}*C$

where, $m = \frac{\alpha+\beta}{10}$ $C = \frac{10\alpha}{\alpha + \beta}$

Thus the expression of $\widehat{WR}$ can be rearranged to yield the formula in the form of the credibility estimate as:

= $\widehat{WR}$ = 10*{Z* Mean based on Sample data + (1-Z)*Mean of prior distribution}

(where + (1- $\acute{Z}$)

Thus, here we have the credibility factor

$Z = \frac{v}{v+m} = \frac{10v}{10v + \alpha + \beta}$

Also we have the following relations:

á + â = 10m, $C = \frac{10\alpha}{m}$

Hence the parameters 'á' and 'â' are expressed in terms of 'm' and 'C' as:

$\acute{a} = \frac{mC}{10}$

â = 10m-á

$= 10m - \frac{mC}{10}$

$= \frac{100m - mC}{10}$

Thus we have: $\hat{a} = \frac{m(100-C)}{10}$

Comparison

We provide a comparison between the classical premium pricing of insurance risk problem and the movie rating problem to highlight the similarity of the Bayesian credibility theory approach to two completely different problems in Table 1.

Table 1

	Rating movies	Premium pricing of insurance risk	Fitting in the Bayesian Framework
Direct data	Average rating for the particular movie(R)	Average cost of claims for a particular insurance risk(X)	Sample data, its mean
Collateral data	Average ratings for all movies(C)	Average cost of claims for all insurance risks(μ)	Prior distribution, its mean
Overall rating/price	Weighted average of R and C	Weighted average of X and μ	Posterior mean, credibility estimate
Weights	$\frac{v}{v+m}$ and $\frac{m}{v+m}$	Z and (1-Z)	Credibility factor

Results and Findings

The primary source of data is the Internet Movie Database (IMDb), which is accessible on the internet at www.imdb.com. The study period is for the last three months during which IMDb updates its database daily.

The top rated movie for the last three months is 'The Shawshank Redemption' which gathers an IMDb ranking of 9.2. In this section, we will calculate how this figure is achieved.

The sample size (number of votes cast in favour of the movie along with breakdown details) is already mentioned in the IMDb webpage [4] and given below as it is.

1646089 IMDb users have given a weighted average vote of 9.3 / 10

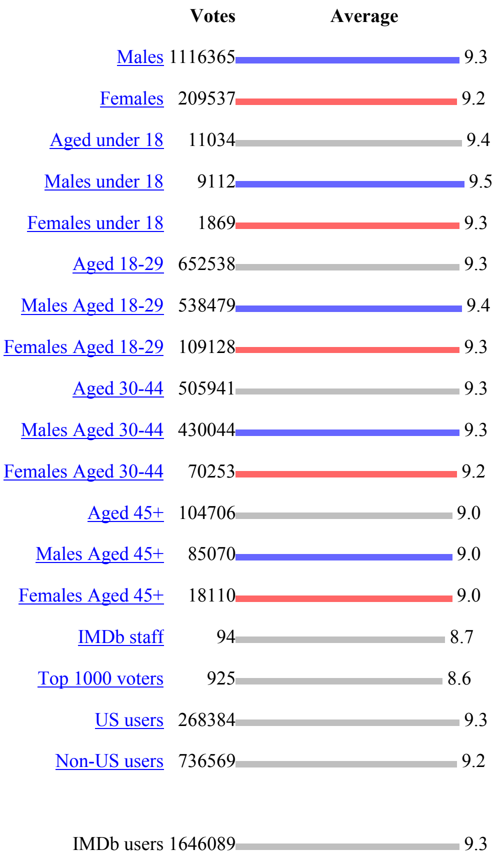
Demographic breakdowns are shown below.

Votes	Percentage	Rating
923158	56.1%	10
410536	24.9%	9
184472	11.2%	8
60339	3.7%	7
18291	1.1%	6
9196	0.6%	5
4285	0.3%	4
3358	0.2%	3
3244	0.2%	2
29210	1.8%	1

Arithmetic mean = 9.1. Median = 10

Ranked #1 in the Top 250 Movies

This page is updated daily.



So, here using the credibility formula used by IMDb.we have:

v = 1646089, m =25000, R = 9.3, C=7.0

Hence,

$$\widetilde{WR} = \left(\frac{v}{v+m}\right)*R + \left\{1-\left(\frac{v}{v+m}\right)\right\}*C$$

$$= \left(\frac{1646089}{1646089+25000}\right)*9.1 + \left\{1-\left(\frac{1646089}{1646089+25000}\right)\right\}*7.0$$

$= 8.9639 + 0.1047$

=9.0686, which is approximately rated by IMDb as **9.3.** This approximation is due to the consideration of manipulative voting and trimmed means.

It is to be noted that although the actual ranking approach followed by IMDb is more complicated and only an approximation of the actual ranking is followed presently, the Bayesian framework provides a necessary tool to tackle the basic problem of actual ranking. The complexity arises from the fact that IMDb now treats votes from different users differently (for tackling manipulative voting) and uses trimmed means[9](to minimise influence of outliers).[3]

Conclusion

The approach adopted by IMDb is more robust than any other statistical approach known so far in rating and ranking movies as it considers both the likelihood function of the individual rating scores and the prior distribution of the unknown parameter 'p', thereby reducing the risk exposure due to anomalous rating using straightforward and unsophisticated approaches.

Thus for calculating the Top Rated 250 titles based on user's votes, it considers the weighted rating which is actually a true Bayesian estimate calculated using the credibility approach by applying it on both the direct data (average rating based on user's votes) and the collateral data (mean number of votes across the whole report) and finally finding a suitable credibility factor. The robustness of this approach is best understood when we compare it with the following two layman's approach of ranking movies.

Layman's approach-1:

Consider the approach of rating all the movies in descending order of their average rating. This approach is simple and obvious, but there exists shortcomings in this approach. Let there be a little-known movie with just a few votes but of a high score. Consider an extreme example of just one vote of score 10. The average being 10, the movie would be on the top of the list which is quite undeserving considering the fact that very few people have seen it and thus voted for it.

Layman's approach-2:

Let us now refine the above approach in an attempt to cover the shortcomings of the previous approach. We modify it by stipulating that a minimum number of votes, say 'm' should be cast for movie before it becomes eligible to be considered for such a listing. Among the movies that meet the cut-off, a simple sorting is done in descending order according to the average rating as before. However, this approach is also far from perfect. Let us consider two movies M_1 and M_2 each with an average rating of 8.5 in a scale of 10, but having a widely different number of votes, say 40,000 and 6,00,000 respectively. If the stipulated minimum number of votes, 'm' is less than 40,000 (say 30,000) then both these movies will be ranked equal under this approach. Intuitively, though, we know that there is a significant difference in number of user votes. The rating of M_2 having been voted by a much larger number of users is more reliable than that of M_1. Thus M_2's rating is more credible than that of M_1.

This notion of credibility is made more precise, robust and rigorous by IMDb in its approach. IMDb's approach is more statistically robust in the sense that its calculations of ratings not just depend on 'R' and 'm', but also on 'v (the number of votes received for a movie).

Thus we see significant difference in weighted rating of M_1 and M_2 according to the approach adopted by IMDb. The calculations to justify it are as follows:

Weighted rating of M_1 (according to IMDb):

Here: v= 40,000, m=30,000, R =8.5, C = 7.0

Therefore, $\widetilde{WR} = (\frac{40000}{40000+30000})*8.5+ (\frac{30000}{40000+30000})*7.0$

= 4.857 + 3

= 7.857

Weighted rating of M_2 (according to IMDb):

Here: v= 6,00,000, m=30,000, R =8.5, C = 7.0

Therefore, = *8.5+ ()*7.0

= 8.095+ 0.333

= 8.428

This matches with our intuition that rating of M_2 is more credible and closer to the average rating 8.5 than that of M_1 as it received significantly larger number of user votes. Thus we find that IMDb's approach is not only statistically robust and rigorous but also intuitively sound and matches with general perception.

Scope for further research

Based on the facts and findings, it is suggested that future studies may focus on incorporating manipulative voting (treating votes from different users differently) and using trimmed means of truncated data to reduce effect of outliers. Also in line with loaded credibility premium, studies can be taken up to include loaded user's votes where votes of regular viewers were given more weightage to present a more realistic and accurate rating list.

References

http://www.imdb.com/

http://www.imdb.com/chart/top

http://www.imdb.com/help/showleaf?votes

http://www.imdb.com/title/tt0111161/ratings?ref_=tt_ov_rt

Prasham M.Rambhia, "Usual theory, unusual applications: Credibility Theory and Movie Rankings", *The Actuary India, Vol VIII, Issue 9*(2015) 18-20.

T. Bauwelinckx, E. Labie and M.J. Goovaerts, "A new approach for loaded credibility premiums" *Journal of Computational and Applied Mathematics* 37(1991) 301-314.

E. Gomez-Deniz, "A generalization of the credibility theory obtained by using the weighted balance loss function", *Insurance: Mathematics and Economics* 42(2008) 850–854.

Harald Dornheim, Vytaras Brazauskas, "Robust-efficient credibility models with heavy-tailed claims: A mixed linear models perspective", *Insurance: Mathematics and Economics* 48(2011) 72-84.

Jean-Philippe Boucher, Michel Denuit, "Credibility premiums for the zero-inflated Poisson model and new hunger for bonus interpretation", *Insurance: Mathematics and Economics* 42(2008) 727-735.

Joseph H.T. Kim, Yongho Jeon, "Credibility theory based on trimming", *Insurance: Mathematics and Economics* 53(2013) 36-47.

Whitney, A.W, "The Theory of Experience Rating", *Proceedings of Casualty Actuarial Society* 4(1918) 274-292.