Implementation of a Comment Rating System Based on Wilson's Score Interval for Binomial Proportions

Arun J Nair[1], Abhilash R Namboodiri[1], Pratibha Kantanavar [2]

[1]*Department of Electronics and Communication Engineering, RV College of Engineering, Bengaluru -59*
[2]*Assistant Professor, Department of Electronics and Communication Engineering, RV College of Engineering, Bengaluru -59*

---***---

Abstract - *Comparing the relevance and quality of comments and posts on social networking sites and ranking them on this basis has long been a goal of software developers. Different approaches have been tested for this function including rating systems based on total positive ratings, average ratings etc. with unsatisfactory results. The use of statistical tools, particularly modelling each post or comment as a Bernoulli trial allows them to be quantitatively compared. This paper seeks to apply computational techniques in the form of binomial proportion confidence intervals to estimate quality. This is the normal approximation to the binomial distribution, with the assumption that the likely distribution of error about an observation is normally distributed and draws upon improved techniques by Wilson (1927) and is applied to a web application use case.*

Key Words: Bayesian Average Rating, Binomial Distribution, Comment Rating, Confidence Intervals, Wilson's Score

1. INTRODUCTION

Rating an ordered list based on merit has been an ever-increasing desire for web developers and designers. Examples include categorizing entities such as posts on a social networking site, products on an e-commerce page and answers on an online forum. Quantitatively answering this need requires the re-framing of the question to a statistical equivalent. Comments in an online forum are rated with two variables: positive and negative votes.

Multiple approaches have been suggested which include:

Rating = Positive – Negative

Ordering in this way causes a comment with a higher rating but lower percentage of positive votes to rated higher as elucidated by this example:

Table - 1: Ratings calculated using the difference rating method

Ranking	Positive	Negative	Rating	Positive %
1	5500	4500	1000	55%
2	600	400	200	60%

Another approach is to sort each entity by an average rating:

Ranking = Average rating = Positive ratings/Total ratings

Sorting by average causes newer comments to be rated higher than older comments with more ratings which ultimately leads to incorrect results. Quantitatively arranging these comments requires re-framing of the requirement to thus: Given a finite number of ratings, extrapolate the real ratings with a fixed probabilistic confidence.

2. CONFIDENCE INTERVALS

Binomial proportion confidence interval [2-10] is a confidence interval for the probability of success calculated from the outcome of a series of success–failure experiments. These experiments can be modelled as a Bernoulli trial with a random variable X. The distribution of X is entirely determined by $P(X = 1) = p$, since $P(X = 0) = 1-p$. Suppose $X_1, X_2, ... X_n$ are a sample of size *n* from a Bernoulli distribution with parameter *p*. The sum $Y = X_1 + ... + X_n = \Sigma X_j$ is a random variable with sample space *{0, 1, ..., n}* which is also a binomial distribution. The expected value of *Y* is *E[Y] = np* and the variance is *var[Y] = np(1-p).* We can then estimate *p* using

$p_0 = (X_1 + ... + X_n)/n = Y/n$

Variance is $var[p_0] = p(1-p)/n$, and the standard deviation is $SD[p_0] = \sqrt{p(1-p)/n}$. This is an unbiased estimator since $E[p_0] = p$ which is also consistent due to law of large numbers.

Upon standardizing p_0 by subtracting mean and dividing by the standard deviation, we get

$$(p_0 - p)/\sqrt{p(1-p)/n} \sim [0,1]$$

Taking a 95% standard deviation, this converts to:

$$P(-1.96 \leq (p\_0 - p)/\surd\{p(1-p)/n\} \leq +1.96)$$

which on simplification leads to a 95% Wald type confidence interval:

$$p_0 \pm 1.96\sqrt{p(1-p)/n} \approx p_0 \pm 2\sqrt{p(1-p)/n}$$

3. WILSON'S SCORE CONFIDENCE INTERVALS

The usual statement of probable inference states that if a certain frequency or rate (p_0) is observed in a population of (n), the true value of the rate (p) can be calculated within the limits of $p_0 \pm \lambda\sigma_0$ (where σ_0 is the standard deviation) with a probability of P_λ. But this statement depends on the fallibility of a particular observation p_0 as opposed to a general solution [1].

Thus, a better way of inference is by assuming a rate *p* with a standard deviation of σ. Then the probability of an observation p_0 lying within the limits $p_0 \pm \lambda\sigma_0$ is less than or equal to P_λ.

This new method can be translated into an equation as $(p_0 - p) = \lambda\sigma$, or alternately as $(p_0 - p)^2 = \lambda^2 p(1-p)/n$. If $\lambda^2/n = t$, then the solution to the equation is

$$p = (p_0 + t/2)/(1+t) \pm \sqrt{p_0(1-p_0)t + t^2/4}/(1+t)$$

Thus, if the true value of probability *p* for an observed p_0 lies outside the range

$$p = \left(p_0 + \lambda^2/2n \pm \lambda\sqrt{(p_0(1-p_0) + \lambda^2/4n)/n}\right)/(1+\lambda^2/n)$$

then the observation p_0 occurs with a probability less than or equal to P_λ.

4. IMPLEMENTATION

The Wilson's score confidence interval is used to rate user-generated comments in a web application where users can give positive or negative votes to each comment. The intention is to deduce the "true rating" of a comment from the sample set of positive and negative votes and sort it on that basis. Thus, the lower bound of the Wilson's score confidence interval is used.

$$Score = \left[\hat{p} + (\lambda^2/2n) - \lambda\sqrt{(\hat{p}(1-\hat{p}) + \lambda^2/4n)/n}\right]/(1+\lambda^2/n)$$

where $\hat{p}$ is the fraction of positive votes, *n* is the total number of votes and λ =1.96 for a 95% confidence interval.

This can be converted into a corresponding SQL query which can be used in the web application:

```
SELECT comment_id, ((positive + 1.9208) / (positive + negative) -
1.96 * SQRT((positive * negative) / (positive + negative) + 0.9604)
 / (positive + negative)) / (1 + 3.8416 / (positive + negative))
AS ci_lower_bound FROM comments WHERE positive + negative > 0
ORDER BY ci_lower_bound DESC;
```

Fig - 1: SQL query for rating comments based on Wilson's score

5. BAYESIAN AVERAGE RATINGS

An alternative approach to the frequentist model is the Bayesian approach [11]. Sorting comments using a Bayesian model helps to initialize a prior belief and update the belief iteratively using the posterior.

Questions which could be addressed are:

- Time: Old ratings count less than new.
- Prior beliefs: Unrated items from a popular author could appear above an unrated item from an unknown author.
- Attitudes toward risk: Set risk averseness of the algorithm

The prior as before will be a Bernoulli trial and as such to ensure that the likelihood function has a conjugate prior for ease of analysis, a beta distribution is chosen. The two parameters taken by the distribution are positive and negative votes. The posterior belief will then be a beta distribution with the new numbers after a set interval. Sorting these trials requires the creation of a loss function which will dictate the decision (sorting criterion).

Assuming a linear loss function with a scalar k

$$L_k(x, X) = k(X - x) \; for \; x < X$$

$$L_k(x, X) = x - X \;\; for \;\; x \geq X$$

The expected loss is:

$$E[L_k(x, X)|U, D] = k\int_0^X (X - x) f(x; U + 1, D + 1)dx + \int_X^1 (x - X) f(x; U + 1, D + 1)dx$$

where f is the beta distribution.

Integrating by parts

$$E[L_k(x, X)|U, D] = k\int_0^X I_y(U + 1, D + 1)dy + \int_X^1 \left(1 - I_y(U + 1, D + 1)\right) dy$$

Minimizing this expression with respect to X

$$I_X(U + 1, D + 1) = 1/(1 + k)$$

I_X is the incomplete beta function and inverting that would lead us to the required random variable X. Precise beliefs would now be closer to average beliefs and imprecise ones would be marked lower.

Another parameter to include would be the time of the rating, a newer rating should hold more value than an old rating. Beliefs are scaled by an exponential decay function where t is the time elapsed since the last update and H is a scalar called the half-life modified based on the use case.

$$new = old \times 2^{-t/H} + 1$$

The loss function is then updated accordingly by adding decay

$$I_X\left(U \times 2^{-t_U/H}, D \times 2^{-t_D/H}\right) = 1/(1+L)$$

where t_U is the time since the most recent positive vote (upvote) and t_D is the time since the most recent negative vote (downvote).

6. RESULTS

Table - 2: Observations sorted by Wilson's score

Positive (p)	**Negative (n-p)**	**Total (n)**	$\hat{p}$**=p/n**	**Score**
118	25	143	0.825174825	0.754647868
209	50	259	0.806949807	0.754536732
128	34	162	0.790123457	0.721061488
50	10	60	0.833333333	0.719681485
10	2	12	0.833333333	0.551963643

Wilson's score allows a set of different experiments with finite set of Bernoulli trials to be compared with each other and rank them according with a 95% certainty.

The first two rows of Table 2 depict observations which differ in the percentage of positive ratings (82.5% and 80.69% respectively). A score calculated based on the difference between negative and positive ratings would rate the second row higher. This is not the case with a Wilson's score confidence interval-based rating.

The last two rows of Table 2 depict observations which have the same average ratings (0.83). In this case, the score based on Wilson's confidence interval gives priority to the observation with a higher number of ratings.

Bayesian average ratings allow us to formally define a belief or an intuition and update it according to a loss function and chosen parameters. A rigorous analysis of the choice of these parameters is beyond the scope of this paper.

7. CONCLUSION

A rating system based on the Wilson's score confidence interval overcomes drawbacks which might arise due to a small sample size or incorrect ordering inherent to methods like the difference method. Such techniques can also be applied to any success-failure based experiments and ascertain from user responses and action to detect spam or abuse. Quality assurance can also be monitored and maintained from people who purchase and their return percentages. Confidence intervals have also been heavily applied in the field of biomedicine for drug testing to predict from a finite set of clinical trials. Bayesian techniques are particularly helpful in cases where flexibility to define the prior and update accordingly with a specific loss function is required and it can be an improved alternative to a standard frequentist approach.

REFERENCES

[1] E. B. Wilson, "Probable Inference, the Law of Succession, and Statistical Inference," Journal of the American Statistical Association, 22, 209-212, 1927.

[2] A. Agresti and B. A. Coull, “Approximate is Better than ‘Exact’ for Interval Estimation of Binomial Proportions,” The American Statistician, 1998.
[3] B. K. Ghosh, “A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter”, Journal of the American Statistical Association, 1979.
[4] J. Neyman, “Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability”, Philosophical Transactions of the Royal Society of London A, 1936.
[5] C. R. Blyth and H. A. Still, H. A., "Binomial Confidence Intervals," Journal of the American Statistical Association, 78, 108-116, 1983.
[6] D. Bohning, "Better Approximate Confidence Intervals for a Binomial Parameter," Canadian Journal of Statistics, 22, 207-218, 1994.
[7] H. Chen, "The Accuracy of Approximate Intervals for a Binomial Parameter," Journal of the American Statistical Association, 85, 514518, 1990.
[8] C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," Biometrika, 26, 404-413, 1934.
[9] D. E. Duffy and T. J. Santner, "Confidence Intervals for a Binomial Parameter Based on Multistage Tests," Biometrics, 43, 81-93, 1987.
[10] L. M. Leemis and K. S. Trivedi, "A Comparison of Approximate Interval Estimators for the Bernoulli Parameter," The American Statistician, 50, 63-68, 1996.
[11] L. Sahabo and Seongbaek Yi, “Normally approximated Bayesian credible interval of binomial proportion,” Journal of the Korean Data & Information Science Society, 30(1), 233-244, 2019.