Confidence interval for user scores

Frost

1 Introduction

There are many websites that allow users to rate certain items, like songs and products, based on their perceived quality. Each item gets a user score, an average of all the ratings. It is useful to rank these items using some sort of criterion, since consumers want to know what the best or worst item is.

Having just the user score as our criterion is usually not a good idea, since items with a few ratings will jump to the top of the rankings. Thus we would like something that takes into account the user score and the number of ratings.

In this paper, we achieve this by calculating a lower bound on the user score, which takes into account the number of ratings, and then sorting the items using this lower bound as a criterion.

2 A lower confidence bound for the user score

2.1 The usual way to generate a lower bound

Suppose we have a set of ratings $x_1, x_2, \ldots, x_n$ that are independent and identically distributed. There are many ways to place a lower confidence bound on the mean μ of the population. One common way is to calculate the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

the sample standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

and supposedly, we can be $100(1-\alpha)\%$ confident that

$$\mu \geq \bar{x} - \frac{s}{\sqrt{n}} t_{\alpha,n-1},$$

where $t_{\alpha,n-1}$ is the critical value of the t-distribution with $n-1$ degrees of freedom. As $n \to \infty$, the t-distribution approaches the standard normal distribution. So for large n, we can instead have

$$\mu \geq \bar{x} - \frac{s}{\sqrt{n}} z_{\alpha},$$

where z_α is the critical value of the standard normal distribution.

The problem with this method is that it produces terrible results when n is small, or when the number of ratings doesn't deviate far from the mean. This is especially true when the x_i come from a discrete distribution, which is the case for systems where a user gives star ratings (0 to 5 stars) or when the user rates on a similar scale (from 0 to 10, or from 0 to 100).

For example, if $n = 1$, we can't place any meaningful bound on μ; we can only say that $\mu > -\infty$.

Additionally, if each x_i is identical (say an item has 10 ratings, each of them being 5 stars), then we have that $s = 0$, so our lower bound for the mean is

$$\mu \geq \bar{x},$$

which is clearly wrong.

2.2 Our approach

If we're going to place a bound on μ, we need to take a different approach. Let's go back to the definition of μ, which for any distribution $\mathcal{D}$ with support $[a, b]$, is given by

$$\mu = \int_a^b xf(x)\mathrm{d}x, \tag{1}$$

where f is the density function of $\mathcal{D}$. To place a bound on μ, we need to place a bound on $xf(x)$. This seems tedious because we have two things to worry about, x and $f(x)$. But there is an equivalent definition for μ which is much nicer for our purposes:

$$\mu = b - \int_a^b F(x)\,\mathrm{d}x, \tag{2}$$

where F is the cumulative distribution function of $\mathcal{D}$. We can show that this is equivalent to (1) using integration by parts. Letting $u = F(x)$ and $\mathrm{d}v = \mathrm{d}x$,

$$\begin{aligned}
\mu &= b - \int_a^b F(x)\,\mathrm{d}x, \\
&= b - \left(xF(x)|_a^b - \int_a^b xf(x)\,\mathrm{d}x \right), \\
&= b - \left(bF(b) - aF(a) - \int_a^b xf(x)\,\mathrm{d}x \right), \\
&= b - b + \int_a^b xf(x)\,\mathrm{d}x, \\
&= \int_a^b xf(x)\,\mathrm{d}x.
\end{aligned}$$

So now, if we want a bound for μ, we just need a bound for $F(x)$. It turns out that it is possible to bound $F(x)$ exactly. Consider the following estimator for F, the empirical cumulative distribution function $\hat{F}$, defined as

$$\hat{F}(x; n) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}(x_i \leq x),$$

where

$$\mathbf{I}(x_i \leq x) = \begin{cases} 1 & \text{if } x_i \leq x, \\ 0 & \text{otherwise} \end{cases}$$

is the indicator function. The value $\hat{F}(x; n)$ is the proportion of samples that are less than or equal to x, and the value $n\hat{F}(x; n)$ is the number of samples that are less than or equal to x. Then

$$n\hat{F}(x; n) \sim \mathrm{Bin}(n, F(x)).$$

To see why this is true, recall that $\mathbb{P}(x_i \leq x) = F(x)$. Since we have n ratings and each x_i is independent, the number of ratings below x follows a binomial distribution, where the number of trials is n and the success probability is $F(x)$.

We don't know what $F(x)$ is, but since it appears in a binomially-distributed random variable, we can find an exact confidence interval for it.

We are only interested in a lower bound for μ. This implies that we should find an upper bound for $F(x)$. This is because the faster F approaches 1, the smaller the mean; on the other hand, the slower F approaches 1, the larger the mean.

Thus, we should find an upper bound $F_U(x; n, \alpha)$ such that

$$\mathbb{P}\left(\mathrm{Bin}(n, F_U(x; n, \alpha)) \leq n\hat{F}(x; n) \right) = \alpha,$$

where $\alpha \in [0, 1]$. This upper bound depends on n, the number of samples, and α, which is the significance level.

The CDF of a binomial distribution with n trials and success probability p is

$$x \mapsto \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}.$$

So $F_U(x; n, \alpha)$ is the solution to the equation

$$\sum_{k=0}^{x} \binom{n}{k} F_U(x; n, \alpha)^k (1 - F_U(x; n, \alpha))^{n-k} = \alpha.$$

Computing the numerical value of $F_U(x; n, \alpha)$ exactly typically requires some root-finding algorithm. There is an explicit solution, given by

$$F_U(x; n, \alpha) = I^{-1}_{1-\alpha}\left(1 + n\hat{F}(x; n), n\left(1 - \hat{F}(x; n)\right)\right),$$

where I^{-1} is the inverse regularized incomplete beta function, but computing I^{-1} once again requires root-finding.

We can now finally solve our problem. If we let μ_L be the lower bound of μ, then

$$\mu_L = b - \int_a^b F_U(x; n, \alpha)\,\mathrm{d}x.$$

The function F_U only changes in value at every unique value of x_i, and at the endpoints a and b. Define $y_{(1)}, y_{(2)}, \ldots, y_{(m)}$ as the *distinct* order statistics of the set of user ratings, with $m \leq n$.

For our convenience, also define $y_{(0)} := a$ and $y_{(m+1)} := b$. These aren't actual order statistics, but they will make our equations look less messy. Then we have that

$$\int_a^b F_U(x; n, \alpha)\,\mathrm{d}x = \sum_{i=0}^{m} \left(y_{(i+1)} - y_{(i)}\right) F_U\left(y_{(i)}; n, \alpha\right).$$

Thus the lower confidence bound for μ is

$$\mu_L = b - \sum_{i=0}^{m} \left(y_{(i+1)} - y_{(i)}\right) F_U\left(y_{(i)}; n, \alpha\right).$$

Thus, with $100(1-\alpha)\%$ confidence, the user score is at least μ_L, and we are done.

2.3 Confidence Interval for the Weighted User Score

Suppose we have fixed weights $w_1, w_2, \ldots, w_n$, where $w_i \in [0, 1]$ for $i = 1, 2, \ldots, n$, that multiply each user score $x_1, x_2, \ldots, x_n$, and we define the weighted user score μ' as

$$\mu' = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

Then a confidence interval for μ' can be derived in a very similar way. Begin by defining a new random variable $v_i = w_i x_i$ for $i = 1, 2, \ldots, n$, and write the identity

$$\frac{\mu'}{n} \sum_{i=1}^{n} w_i = \frac{1}{n} \sum_{i=1}^{n} v_i.$$

We can then consider a confidence interval for the mean of $v_1, v_2, \ldots, v_n$. Notice that the right hand side of the above equation is the sample mean of each v_i. If we define the empirical distribution function of the sample as $\hat{G}$, we have

$$\frac{\mu'}{n} \sum_{i=1}^{n} w_i = b - \int_a^b \hat{G}(x; n)\mathrm{d}x.$$

Then we can compute lower and upper confidence bands $\hat{G}_L$ and $\hat{G}_U$ for G to obtain a confidence interval for the mean of v, which is just

$$\left[b - \int_a^b \hat{G}_U(x; n, \alpha) \mathrm{d}x, b - \int_a^b \hat{G}_L(x; n, \alpha) \mathrm{d}x \right]$$

However, this would also be the confidence interval for $\frac{\mu'}{n} \sum_{i=1}^n w_i$. To get a confidence interval for μ', we just multiply the lower and upper limits by $\frac{n}{\sum_{i=1}^n w_i}$. So the confidence interval for μ' is just

$$\left[\frac{n}{\sum_{i=1}^n w_i} \left(b - \int_a^b \hat{G}_U(x; n, \alpha) \mathrm{d}x \right), \frac{n}{\sum_{i=1}^n w_i} \left(b - \int_a^b \hat{G}_L(x; n, \alpha) \mathrm{d}x \right) \right],$$

where once again $v_i = w_i x_i$ for $i = 1, 2, \ldots, n$.