Confidence interval for user scores

Frost

1 Introduction

There are many websites that allow users to rate certain items, like songs and products. Usually, each item gets a user score, an average of all the ratings. But when it comes to ranking a set of items from best to worst, how should it be done? Here we propose one way to obtain a confidence interval for the mean using the empirical distribution function of the dataset.

2 Confidence interval for μ using the eCDF

Suppose we have a set of i.i.d. user ratings $x_1, x_2, \ldots, x_n$ that come from some unknown probability distribution $\mathcal{D}$ with unknown mean μ, unknown cumulative distribution function F, and **known** support $[a, b]$. Consider the empirical cumulative distribution function $\hat{F}_n$ given by

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

where

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

The value $\hat{F}_n(x)$ is the proportion of samples that are less than x, and the value $n\hat{F}_n(x)$ is the number of samples that are less than x. Then

$$n\hat{F}_n(x) \sim \text{Binomial}(n, F(x)). \tag{2}$$

To see why (2) is true, use the fact that $\mathbf{1}(x_i \leq x)$ for $i = 1, 2, \ldots, n$ equals 1 if $x_i \leq x$ and 0 otherwise. Since each $x_i \sim \mathcal{D}$, the probability that $x_i \leq x$ is $F(x)$, so $\mathbf{1}(x_i \leq x)$ follows a Bernoulli distribution with probability of success $F(x)$. This means that $n\hat{F}_n(x) = \sum_{i=1}^{n} \mathbf{1}(x_i \leq x)$ is the sum of n i.i.d. Bernoulli random variables with success probability $F(x)$, which implies that $n\hat{F}_n(x) \sim \text{Binomial}(n, F(x))$.

We would like to find upper and lower bounds $U(x)$ and $L(x)$ such that

$$\mathbb{P}\left(\text{Binomial}(n, U(x)) \leq n\hat{F}_n(x)\right) = \frac{\alpha}{2}, \tag{3}$$

$$\mathbb{P}\left(\text{Binomial}(n, L(x)) \geq n\hat{F}_n(x)\right) = \frac{\alpha}{2} \tag{4}$$

for $x = 0, 1, \ldots, n$. These bounds form what is called a **Clopper-Pearson confidence interval** of $F(x)$, and they can be used to form what is called a **pointwise confidence band** of F.

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

$$k \mapsto I_{1-p}(n-k, 1+k),$$

where I is the regularized incomplete beta function given by

$$I_x(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$$

and Γ is the Gamma function given by

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t.$$

We can rewrite (3) and (4) in terms of I to get

$$I_{1-U(x)}\left(n - n\hat{F}_n(x), 1 + n\hat{F}_n(x)\right) = \frac{\alpha}{2},$$

$$1 - I_{1-L(x)}\left(n - n\hat{F}_n(x) + 1, n\hat{F}_n(x)\right) = \frac{\alpha}{2}.$$

Using the fact that $I_{1-x}(q,r) = 1 - I_x(r,q)$ for all $q, r \in \mathbb{R}$, we get

$$\begin{aligned}
1 - I_{U(x)}\left(1 + n\hat{F}_n(x), n - n\hat{F}_n(x)\right) &= \frac{\alpha}{2}, \\
I_{L(x)}\left(n\hat{F}_n(x), n - n\hat{F}_n(x) + 1\right) &= \frac{\alpha}{2}.
\end{aligned}$$

Simplifying,

$$\begin{aligned}
I_{U(x)}\left(1 + n\hat{F}_n(x), n\left(1 - \hat{F}_n(x)\right)\right) &= 1 - \frac{\alpha}{2}, \\
I_{L(x)}\left(n\hat{F}_n(x), n\left(1 - \hat{F}_n(x)\right) + 1\right) &= \frac{\alpha}{2}.
\end{aligned}$$

We can then use the inverse regularized incomplete beta function I^{-1} to solve for $U(x)$ and $L(x)$:

$$\begin{aligned}
U(x) &= I^{-1}_{1-\frac{\alpha}{2}}\left(1 + n\hat{F}_n(x), n\left(1 - \hat{F}_n(x)\right)\right), \\
L(x) &= I^{-1}_{\frac{\alpha}{2}}\left(n\hat{F}_n(x), n\left(1 - \hat{F}_n(x)\right) + 1\right).
\end{aligned}$$

It should be noted that $U(x)$ is undefined for $x \geq x_{(n)}$ and $L(x)$ is undefined for $x < x_{(1)}$. We can get around this by noting that

$$\begin{aligned}
\lim_{r\to 0} I^{-1}_p(q,r) &= 1, \\
\lim_{q\to 0} I^{-1}_p(q,r) &= 0
\end{aligned}$$

for any $p \in [0,1]$ and $q, r > 1$. So we can redefine U as

$$U(x) := \begin{cases} I^{-1}_{1-\frac{\alpha}{2}}\left(1 + n\hat{F}_n(x), n\left(1 - \hat{F}_n(x)\right)\right) & \text{if } x < x_{(n)}, \\ 1 & \text{otherwise.} \end{cases} \tag{5}$$

and redefine L as

$$L(x) := \begin{cases} I^{-1}_{\frac{\alpha}{2}}\left(n\hat{F}_n(x), n\left(1 - \hat{F}_n(x)\right) + 1\right) & \text{if } x \geq x_{(1)}, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

We can now finally derive a confidence interval of μ. First, note that μ can be expressed in terms of the cumulative distribution function F using the identity

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

Let μ_L and μ_U be the lower and upper confidence bounds of μ, respectively. If we consider all possible distribution functions on $[a,b]$ that are between L and U, then the CDF that minimizes the mean is U and the CDF that maximizes the mean is L. Therefore, the upper limit of a $100(1-\alpha)\%$ confidence interval of μ is

$$\mu_U = b - \int_a^b L(x)\mathrm{d}x,$$

and the lower limit is

$$\mu_L = b - \int_a^b U(x)\mathrm{d}x.$$

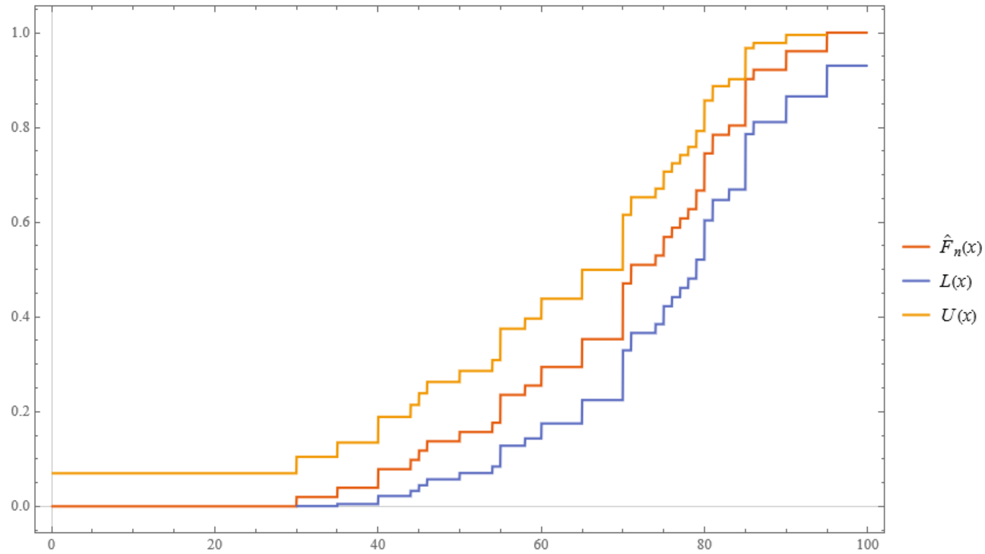
Using properties of U and L,

$$\begin{aligned}
\mu_U &= b - \int_a^b L(x)\mathrm{d}x \\
&= b - \left(\int_a^{x_{(1)}} 0\mathrm{d}x + \int_{x_{(1)}}^b L(x)\mathrm{d}x\right) \\
&= b - \int_{x_{(1)}}^b L(x)\mathrm{d}x
\end{aligned}$$

and

$$\begin{aligned}
\mu_L &= b - \int_a^b U(x)\mathrm{d}x \\
&= b - \left(\int_a^{x_{(n)}} U(x)\mathrm{d}x + \int_{x_{(n)}}^b 1\mathrm{d}x\right) \\
&= x_{(n)} - \int_a^{x_{(n)}} U(x)\mathrm{d}x.
\end{aligned}$$

For ease of computation, we can replace the integral with a sum by realizing that L and U are both step functions that jump at every unique value of x_i for $i = 1, 2, \ldots, n$:



The eCDF $\hat{F}_n(x)$ along with its pointwise 95% confidence bands for some set of data ($n = 51$) with support $\{0, 1, \dots, 100\}$.

Let $\tilde{x}_{(1)}, \tilde{x}_{(2)}, \dots, \tilde{x}_{(m)}$ be the *distinct* order statistics of $x_1, x_2, \dots, x_n$, with $\tilde{x}_{(1)} = x_{(1)}, \tilde{x}_{(m)} = x_{(n)}$, and $m \leq n$. Then since L and U only jump at each $\tilde{x}_{(i)}$ for $i = 1, 2, \dots, m$, we can get that

$$\mu_L = \int_{x_{(1)}}^{b} L(x)\mathrm{d}x = \left(b - x_{(n)}\right) L(x_{(n)}) + \sum_{i=1}^{m-1} \left(\tilde{x}_{(i+1)} - \tilde{x}_{(i)}\right) L(\tilde{x}_{(i)})$$

and

$$\mu_U = \int_{a}^{x_{(n)}} U(x)\mathrm{d}x = \left(x_{(1)} - a\right) U(a) + \sum_{i=1}^{m-1} \left(\tilde{x}_{(i+1)} - \tilde{x}_{(i)}\right) U(\tilde{x}_{(i)}).$$

With these formulas, we finally get that the $100(1-\alpha)\%$ confidence interval of μ is

$$[\mu_L, \mu_U]$$

where

$$\mu_L = x_{(n)} - \left(x_{(1)} - a\right) U(a) - \sum_{i=1}^{m-1} \left(\tilde{x}_{(i+1)} - \tilde{x}_{(i)}\right) U(\tilde{x}_{(i)}),$$
$$\mu_U = b - \left(b - x_{(n)}\right) L(x_{(n)}) - \sum_{i=1}^{m-1} \left(\tilde{x}_{(i+1)} - \tilde{x}_{(i)}\right) L(\tilde{x}_{(i)}).$$

3 Code

Here's a program in Rust that outputs the $100(1-\alpha)\%$ confidence interval of the mean for a dataset with support $[a, b]$.

Install git and run the command `git clone https://github.com/Fr0stium/mean_ci.git` to download the program from GitHub.

To run the program, enter the command `cargo run -- path alpha min_support max_support` inside the program's directory, where `path` is the path of a text file containing a set of numbers separated by commas, `alpha` is α, `min_support` is a, and `max_support` is b. Note that the path to the text file should not contain any spaces.

An example of a valid text file generated from a distribution with support $\{0, 0.5, \dots, 5\}$ is

```
4.5,2.5,4,3.5,4.5,5,4.5,3.5,4.5,4,3,5,4.5,4,5,3,4,4,3,4,3,3,3.5,4,4,4,4,4.5,5,3.5,4,4.5,4,3.5,3,1,4.5
```

With this file, the program outputs the following:

```
Number of ratings: 37
Mean: 3.864864864864865
90% confidence interval: (3.3494896400115475, 4.177131394268508)
```