Data Science Example Social Ratings

Last time

Model social ratings (thumbs up/down) as n 0,1 (Bernoulli) variables

Rating $\bar{X} \equiv \hat{p} = \frac{k}{n}$ "well" approximated by normal distribution

Use Normal's Confidence Intervals to determine likely/unlikely values of p (item's true rating) given $\hat{p}, n$.

$$\hat{P}_L = P - 1.96\sqrt{\tfrac{1}{n}P(1-P)} \qquad \text{mean} - 1.96(\text{stdv})$$

$$\hat{P}_R = P + 1.96\sqrt{\tfrac{1}{n}P(1-P)} \qquad \text{mean} + 1.96(\text{stdv})$$

$\rightarrow 1.96 \Rightarrow 95\%$ C.I.

* Lower Confidence Bound (LCB) sort

- Sort items using $\hat{P}_L$ (a "worst-case" estimate of p)
- statistically well-motivated way to combine our sort objectives $\Rightarrow$ Rank items incorporating uncertainty

Remaining Limitation

$\hat{P}_L = p - 1.96\sqrt{\frac{1}{n}P(1-P)}$ depends on p, p unknown!

How to compute $\hat{P}_L$?

Here's two solutions.

1. Use sample statistics $\rightarrow$ replace p w/ $\hat{p}$ in $\hat{P}_L$.

$$\hat{P}_L = \underset{\substack{\uparrow \\ \text{sample} \\ \text{mean} = \hat{p}}}{\hat{p}} - 1.96\sqrt{\underset{\substack{\uparrow \\ \text{sample} \\ \text{variance} = S_x^2}}{S_x^2}}$$

* sometimes called "Wald Approximation" Can be OK to use but not always accurate.

2. Wilson Score - Let's study this for some nice insights. $\rightarrow$

Wilson Score (W.S.)

Let $\pm z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ 95% C.I on z is $[-1.96, 1.96]$

W.S. → solve this for p:

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\dfrac{\hat{p}(1-\hat{p}) + \frac{z^2}{4n}}{n}}}{1 + \frac{z^2}{n}} = p$$

← Get P_L, P_R by plugging in $\hat{p}, n, z$

That's the answer but let's dig deeper:

Let's rewrite this to understand it better.

Here p is of the form $A \pm B$, let's focus on A (stuff left of $\pm$):

$$\frac{\hat{p} + \frac{z^2}{2n}}{1 + \frac{z^2}{n}}$$

- recall $\hat{p} = \frac{k}{n}$ k t.u.'s of n ratings
  → plug in
- Also, Lets plug in $z = 2 \approx z_L = 1.96$ close enough!

$$\approx \frac{\frac{k}{n} + \frac{4}{2n}}{1 + \frac{4}{n}} = \frac{\frac{k+2}{n}}{\frac{n+4}{n}} = \frac{k+2}{n+4}$$

⇒ Wilson score is a <u>smoothed</u> approximation!

add 2 $\underbrace{\text{successes}}_{\text{t.u.}}$ and 2 $\underbrace{\text{failures}}_{\text{t.d.}}$ $k \to k+2$, $n \to n+4$

⇒ This idea of "smoothing" low count data is very common. Can appear <u>ad hoc</u> but in many situations is statistically well principled (of course, here we only looked at term to left of $\pm$).