Data Science Example - Social Ratings

Goal Link our statistical model to the problem of rating items w/ varying amounts of data.

Recall

Rate products A $R = 5/5$ stars but $n = 1$
B $R = 4.5/5$ stars but $n = 30$

want to show products to a shopper using these social ratings, but naive to just use $R_A > R_B$ b/c R_A is very uncertain

Let's build a statistical model to capture rating uncertainty.

1. simplify: stars $\rightarrow$ thumbs $\Rightarrow$ user j rating X_j is a Bernoulli R.V. $X_j = 1$ w/ prob p. 0 otherwise

   Statistics $E[X_j] = p$, $Var(X_j) = p(1-p)$

2. n users independently rate a product, $\{X_j\}$ are iid (independently and identically distributed)

3. Product's observed rating is $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$

   Let $k \equiv n\bar{x} = \sum_{j=1}^{n} x_j$ $\qquad$ $k = \#$ thumbs ups.

   To understand the rating of a product, need a model $\Rightarrow Pr(k; n, p)$. Knowing $Pr(k)$ we can also study $Pr(\bar{X})$

4. Since $\{X_j\}$ are iid, $k = \sum_{j=1}^{n} X_j$ is a Binomial R.V.:

$$Pr(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

5. Relate statistics of X_j to statistics of k and, what we really want, statistics of $\bar{X}$

$$E[k] = np \qquad Var(k) = np(1-p)$$

$$\hookrightarrow E[\bar{x}] = p \qquad Var(\bar{x}) = \frac{1}{n} p(1-p)$$

That variance of $\bar{x}$ decreases w/ n for fixed p is important: the mean of random variables will "fluctuate" less than the RVs themselves, and these fluctuations decrease as n increases!

→ Let's use this to our advantage!

Outline

- ✓ 1. Problem Formulation
- ✓ 2. Modeling a user's rating
- ✓ 3. Modeling a product's rating
- _ 4. Connecting models to sorting products ←

4. Connecting models to sorting products

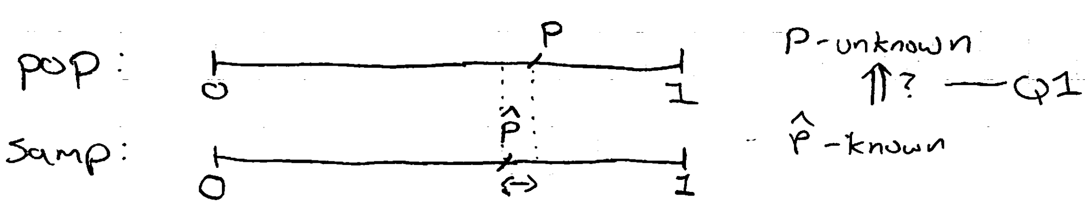
Our derivation shows how much a rating can vary given n → understand better how certain we are about the population rating p given the observed (sample) rating $\bar{x}$.

⇒ We need to address a remaining limitation*, but modeling this uncertainty can be a powerful solution to our sorting problem. Let's see how

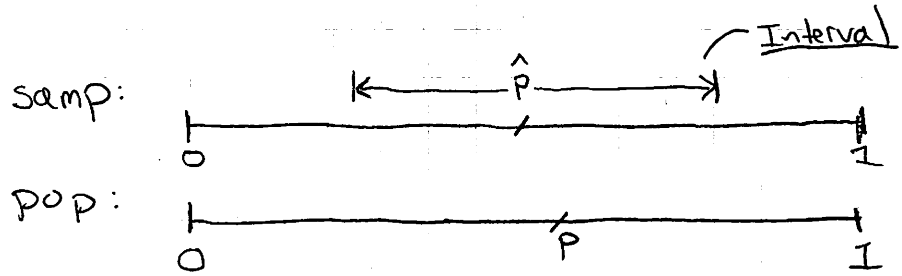
* described shortly

Variance to Confidence (Intervals)

Let $\bar{x} \equiv \hat{p}$ (common notation). How well does $\hat{p} \approx p$? (Q1)
Another question: Given $\hat{p}$ and n, what are (un)likely values of p? (Q2)



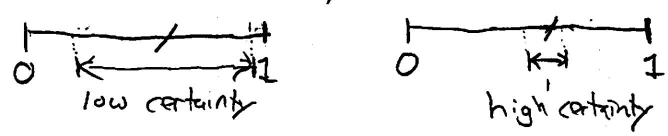
Hard to get at Q1 w/o knowing p. Let's flip this around to look at Q2



Suppose we somehow define an interval around $\hat{p}$: $[\hat{p}_L, \hat{p}_R]$ such that values of $p \in [\hat{p}_L, \hat{p}_R]$ are likely and values outside are unlikely.

If we can do this - from the data - then we can rule out values of p and understand better our uncertainty of p given the data.

→ The width of the interval relates to our uncertainty:



Def 95% Confidence Interval (CI)

The range of values of p (in this case) such that there is a 95% probability the true (population) value falls w/in this range

Find $\hat{p}_L, \hat{p}_R$ s.t. $Pr(\hat{p}_L < p < \hat{p}_R) = 0.95$

Ex CI = [0,1] not just a 95% chance p is in this range, but a 100% chance! Not very helpful though...

How to calculate/estimate a C.I. on p using $\hat{p}, n$?

[Notebook]

Ah, normal approximation!

1.96 is related to .95

Normal distribution: 95% CI is mean ± 1.96(stdv).

$E[\bar{x}]$ $\qquad \sqrt{Var(\bar{X})} = \sigma$

So that's our C.I. $\quad \hat{p}_L = p - 1.96\sqrt{\frac{p(1-p)}{n}}$

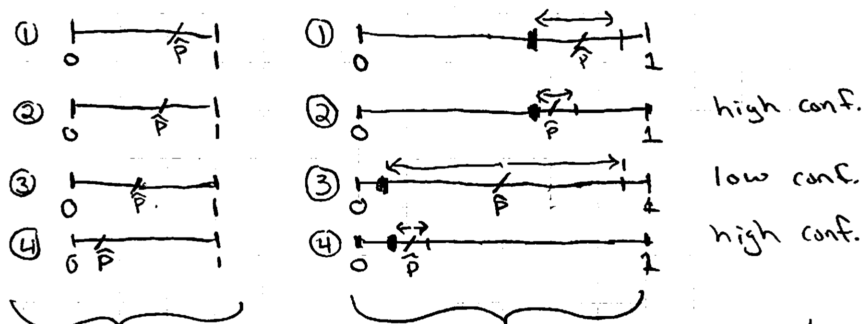
$$\hat{p}_U = p + 1.96\sqrt{\frac{p(1-p)}{n}}$$

Great! We've got it. Just two small things:

1. So what? $\qquad$ 2. $[\hat{p}_L, \hat{p}_U]$ depend on p — unknown $\Rightarrow$ we've got nothing!

Let's tackle these in turn.

1. <u>So what</u> $\quad$ How to use C.I. to sort products?



sort on $\hat{p}$ $\qquad$ sort on $\hat{p}_L$ $\rightarrow$ <u>Lower confidence bound!</u>

1, 2, 3, 4 $\qquad$ 1, 2, 4, 3

Use "LCB sort" to incorporate uncertainty!

Q: Useful when normal approx fails (such as $p \approx 0$, $p \approx 1$)?
Maybe! Even if we can't trust the C.I. it
might still give good sorting in practice $\Rightarrow$
unusual, practical perspective!

2. Depends on p - how to compute LCB? (* Remaining limitation)

Let's tackle this next time!

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

$\hat{P}_L = p - 1.96\sqrt{\tfrac{1}{n}P(1-P)}$ 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 $\hat{P}_L$.

$$\hat{P}_L = \underbrace{\hat{p}}_{\text{sample mean} = \hat{p}} - 1.96\sqrt{\underbrace{S_x^2}_{\text{sample variance} = S_x^2}}$$

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

2. <u>Wilson Score</u> - 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:

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

← 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 $\pm$).