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 RV $X_j = 1$ 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 $k \equiv n\bar{x} = \sum_{j=1}^{n} x_j$ $\quad$ $k = \#$ 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!

<u>Outline</u>

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

4. <u>Connecting models to sorting products</u>

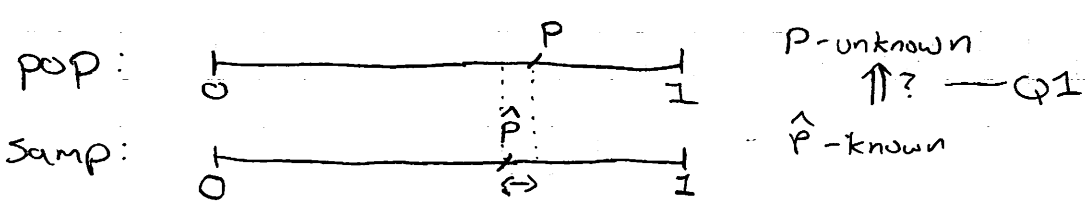
Our derivation shows how much a rating can vary given n → understand better how certain we are about the population rating p given the observed (sample) rating $\bar{x}$.

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

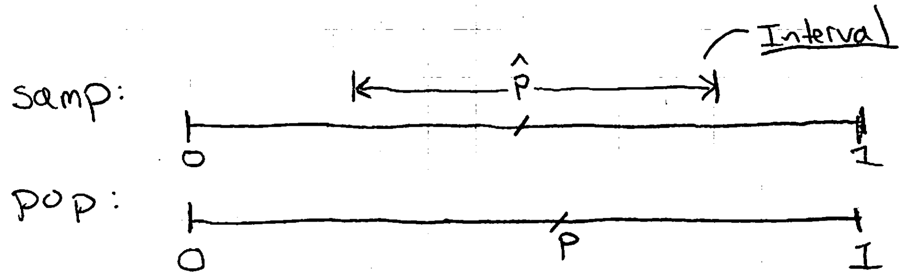
* described shortly

<u>Variance to Confidence (Intervals)</u>

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)



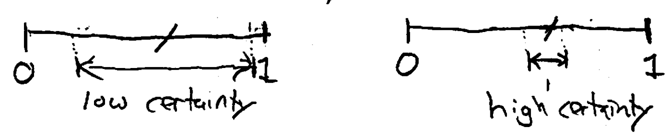
Hard to get at Q1 w/o knowing p. Let's <u>flip</u> this around to look at Q2



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 <u>likely</u> and values outside are <u>unlikely</u>.

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

→ The width of the interval relates to our uncertainty:



<u>Def</u> 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. $Pr(\hat{p}_L < p < \hat{p}_R) = 0.95$

<u>Ex</u> $CI = [0, 1]$ 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 $\pm$ 1.96(stdv).

So that's our C.I.

$$\hat{p}_L = \overset{E[\bar{x}]}{p} - 1.96 \overbrace{\sqrt{\frac{p(1-p)}{n}}}^{\sqrt{Var(\bar{X})} = \sigma}$$

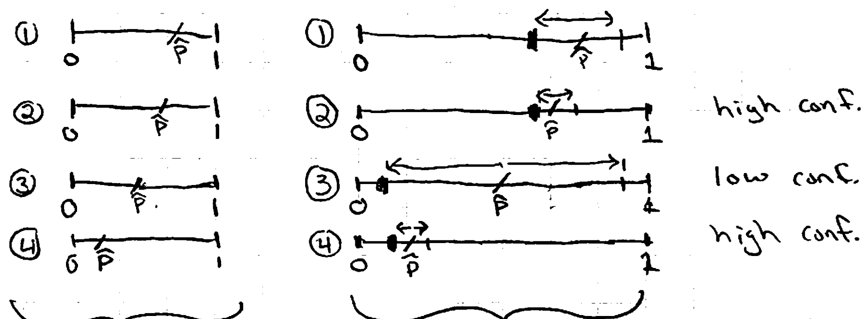
$$\hat{p}_U = p + 1.96\sqrt{\frac{p(1-p)}{n}}$$

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

1. So what?
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> How to use C.I. to sort products?



sort on $\hat{p}$ ①,②,③,④

sort on $\hat{p}_L$ ①,②,④,③ $\rightarrow$ <u>Lower confidence bound!</u>

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!