Data Science Example - social ratings

Goal: Learn how statistics and probability can inform a data-driven service

Start w/ projector:

1. Reminder of expectation (mean) and variance, properties [if omitted from previous lecture]

2. Motivating slide - Amazon.com star ratings

3. Switch to board

Question: How to rank products using ratings of users when some products have fewer ratings than others.

Ex

Product A	Product B
5 out of 5 stars ($R_A = 5$)	4.5 out of 5 stars ($R_B = 4.5$)
1 rating ($n_A = 1$)	30 ratings ($n_B = 30$)

→ naive to sort based on R, b/c A has so few ratings.

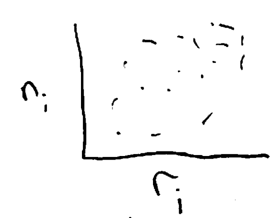
Brainstorm solutions: Gen Q: How to sort using two objectives

Here's some of my ideas:

- Ignore it, sort using R → bad idea

- Define cutoff, only show products w/ $n > n_{cutoff}$.
  → bad idea: biased to popular products, how do new products get any traction?

- Show shoppers a fancy 2D visualization
  → bad idea: way too complicated!



- Scalarize the objectives to sort using one number
  → data-driven scalarization?
  → design scalarization using statistical inference ⊛

Outline

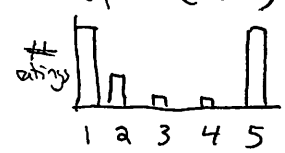
1. Problem formulation
2. Modeling a user rating
3. Modeling a product's rating
4. Connecting models to sorting problem (next time)

I. Problem formulation

Product i has n_i ratings

Design choice: 1-5 stars is complicated, simplify to thumbs up/down

- rating is now proportion of thumbs ups (t.u.)
- Real ratings often support this →



easy to transform

→ 10 people rate product, 6 give t.u. $\Rightarrow R_i = \frac{6}{n_i} = \frac{6}{10} = 0.6$

- This simplification is plausible for many data and makes calculations much easier, but going from a 1-5 scale to a 0-1 scale does not solve our problem of how to sort products.

Q: $R_i = \# t.u./n_i$ is the observed (sample) rating. What will the rating be if everyone (population) rated product i?

→ Asking this is critical: If we somehow knew the population rating we would know the true way to sort products!

But we don't know pop. rating. What information do we know about pop. rating given sample?

2. Modeling a user rating.

The process by which a user picks thumbs up vs. down is <u>complicated</u> and will depend on many details specific to the user and unknown to us.

A statistical model replaces these unknown factors w/ a reasonable approximation using randomness

Introduce a random variable (R.V.):

$$X_j = \begin{cases} 1 & \text{if } j^{\text{th}} \text{ person gives t.u.} \\ 0 & \text{otherwise} \end{cases}$$

(suppressing index of product b/c we are only considering a single product here.)

RV needs an associated prob. distribution:

$$Pr(X_j = 1) = p, \quad Pr(X_j = 0) = 1 - p$$

$$\text{or} \quad Pr(X_j = x_j) = \begin{cases} p & \text{if } x_j = 1 \\ 1-p & \text{if } x_j = 0 \end{cases}$$

- This is called a <u>Bernoulli</u> R.V. Essentially, a "coin flip"
- We have introduced a <u>parameter</u> p, if $p = \frac{1}{2}$ the user is equally likely to vote t.u. or t.d.

<u>Statistics of Bernoulli</u>:

$$E[X_j] = \sum_x x\, Pr(X=x) = 1 \cdot Pr(X=1) + 0 \cdot Pr(X=0)$$
$$= 1 \cdot p + 0(1-p)$$
$$= p$$

$$Var(X_j) = E[X_j^2] - \underbrace{E[X_j]}_{=p}{}^2 = \sum_x x^2 Pr(X=x) - p^2$$
$$= 1^2 \cdot p + 0^2(1-p) - p^2 = p - p^2$$
$$= p(1-p)$$

<u>Sample mean</u>: $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ = fraction of 1's in sample

<u>Sample variance</u>: $s_x^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$

be careful n vs. $n-1$

3. Modeling a product's rating

A product will receive n ratings (supressing index i), so we need to deal w/ multiple Bernoulli R.V.'s

Let's make some simplifying assumptions (that may or may not be plausible)

1. Each user j rating the product follows the same prob. distribution
2. The parameter p is constant for the product (of course, different products can and will have different p's)
3. Each user rates the product independent of any other user's rating.

   → may not be reasonable in practice. Network effects, for example, may induce a dependency ("Everyone else loves this, so it must be great!")

Taken together, these assumptions tell us that the set of n ratings given to the product are independent and identically distributed (iid). Denoted

- The X_j's are iid
- The $\{X_j\}$ are iid.

A product's rating is the average of the n t.u./t.d. user ratings it receives.

$$\text{rating} = \bar{X} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

This is proportional to a sum:

$$n\bar{x} = \sum_{j=1}^{n} x_j \equiv k \quad \text{(call the sum } k\text{)}$$

Q we need to ask: What is the probability that the sum of n "coin flips" is k?

A: under our assumptions, summing Bernoulli variables gives a binomial random variable

$$\underbrace{\Pr(\overset{\text{R.V.}}{k}\overset{\text{semi colon}}{;}\ \underbrace{n, p}_{\text{parameters}})}_{} = \binom{n}{k} p^k (1-p)^{n-k}$$

idea: k 1's occur w/ prob p^k

$n-k$ 0's occur w/ prob $(1-p)^{n-k}$

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ arrangements of the k 1's and $n-k$ 0's.

Understanding k lets us see how much the collective rating $\bar{X}$ can vary if/when different sets of users rate the product

Statistics of Binomial RV.

- $E[k] = \sum_k k \Pr(k; n, p) = [\text{lot of work}] = np$

-or- $E[k] = E\left[\sum_{j=1}^n X_j\right] \underset{\substack{\uparrow \\ \text{b/c } E[\cdot] \\ \text{is linear.}}}{=} \sum_{j=1}^n E[X_j] = \sum_{j=1}^n p = np$

- $Var(k) = Var\left(\sum_{j=1}^n X_j\right) = \sum_{j=1}^n Var(X_j)$ (b/c X_j are iid → indep)

$= \sum_{j=1}^n p(1-p) = np(1-p)$

But we really want statistics of $\bar{X}$ not k. $\left(\bar{x} = \frac{k}{n}\right)$

- $E[\bar{X}] = E\left[\frac{k}{n}\right] \underset{\substack{\uparrow \\ \text{b/c } n \\ \text{const.}}}{=} \frac{1}{n} \underbrace{E[k]}_{np} = \frac{1}{n} np = p$ makes sense!

- $Var(\bar{X}) = Var\left(\frac{k}{n}\right) = \underset{\substack{\uparrow \\ \text{careful}}}{\frac{1}{n^2}} Var(k) = \frac{1}{n^2} \cdot np(1-p) = \underset{\substack{\uparrow \\ \text{smaller*}}}{\frac{1}{n}} p(1-p)$

→ $Var(\bar{X}) < Var(X_j)$ and $Var(\bar{X})$ decreases w/ n! makes sense: averaging over more data (higher n) gives less fluctuation

⇒ Recall standard deviation vs. standard error (or S.E.M.