Example: Amazon.com star ratings.

Product A
5 out of 5 stars (R=5)
1 rating

Product B
4.5 out of 5 stars (R=4.5)
30 ratings.

⇒ naive to sort based on ratings, \( R_A > R_B \), but...

\( b/c \) A has few ratings.

⇒ instead, incorporate uncertainty in \( R \) into sort. But how?

Let's first simplify ⇒ use thumbs up/down instead of star rating (common on Youtube, Reddit, ...) and very similar to stars b/c most people use either 1 or 5 stars.

⇒ rating is now the proportion of thumbs ups (t.u.)

10 people rate the product, 6 give it t.u. ⇒ rating = \( \frac{6}{10} = 0.6 \)

⇒ but this is the observed (sample) rating. What will it be if everyone rated it? (population).

Model the rating process with a random variable:

\[
X_i = \begin{cases} 
1 & \text{if the } i\text{th person gave a thumbs up} \\
0 & \text{otherwise}
\end{cases}
\]

\[
\Pr(X_i = 1) = p, \quad \Pr(X_i = 0) = 1 - p \quad \text{Bernoulli R.V.}
\]

or \( \Pr(X_i = x_i) = \begin{cases} 
p & \text{if } x_i = 1, \\
1-p & \text{if } x_i = 0.
\end{cases} \)

The modeling assumption is that each person's rating obeys the same distribution, \( p \) is a constant, and the \( X_i \)'s are independent and identically distributed (iid).

⇒ these assumptions are reasonable, but not guaranteed to be accurate; for example, network effects may induce a dependence among different users.
P is unknown though we can approximate it w/ pop of 1's in the sample.

Pop. and Sam. Statistics:

\[
E[X_i] = \sum_x x \cdot Pr(X=x) = 1 \cdot Pr(X=1) + 0 \cdot Pr(X=0) = 1 \cdot p + 0(1-p) = p.
\]

\[
\text{Var}(X_i) = E[X_i^2] - (E[X_i])^2 = \sum_x x^2 Pr(X=x) - p^2
\]

\[
= 1^2 Pr(X=1) + 0^2 Pr(X=0) - p^2 = 1^2 p + 0^2 p - p^2 = p - p^2 = p(1-p)
\]

Sample mean \[\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i\] for bernoulli variables, this is the fraction of 1's in the sample.

Sample variance \[S_X^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2\] w/ \(\bar{X}\) from above. \((\text{be careful w/ n})\)

\[S_X^2 \text{ vs. } \sigma_X^2, \text{ m vs. } \mu\]

Uncertainty? If I collected another sample, how different would it be?

I don't want \(Pr(X_i = 1)\), I want \(Pr(\bar{X})\). sampling distr.

\[E_{\text{sample }} [\bar{X}] = \sum_x x \cdot Pr(\bar{X})\]

recall that \(\bar{X}\) is proportional to a sum: \(\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i\)

\(n \bar{X} = \sum_{i=1}^{n} X_i \cong k\)

We need to ask: what is prob. that the sum of n "coin flips" is k?

\[\text{bernoulli variable } \Rightarrow \text{ binomial variable}\]

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

\(k\)'s are 0's and \(n-k\)'s and \(\binom{n}{k}\) is arranged of 1s and 0's.
Mean of binomial \( E[K] = np \)

\[
E[K] = \sum_k k P(k; n, p) = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np
\]

Likewise, variance of binomial \( \text{Var}(k) = np(1-p) \)

\[
\text{Var}(k) = \text{Var}\left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{Var}(X_i) \quad \text{since iid.}
\]

\[
= \sum_{i=1}^{n} p(1-p) = np(1-p)
\]

Mean of \( K \Rightarrow \text{mean of } X \)

\[
E[k] = np \quad \Rightarrow \quad E[X] = E\left[\frac{K}{n}\right] = \frac{1}{n} E[K] = \frac{1}{n} np = \frac{p}{n}
\]

Var of \( k \Rightarrow \text{var of } X \)

\[
\text{Var}(k) = np(1-p) \quad \Rightarrow \quad \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{Var}(k) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}
\]