Last time

Amazon, user ratings
Stars → thumbs = bernoulli variables X_i
E(X_i) = p, Var(X_i) = p(1-p)
Assume each X_i is independent
Sample mean \( \frac{1}{n} \sum x_i = \bar{x} \)
Sample var \( \frac{1}{n} \sum (x_i - \bar{x})^2 \)
Uncertainty →
Mean of binomial $E[K] = np$
$E[K] = \sum_k k P(k; n, p) = \left[ \ldots a lot of work \right] = np$

or
$E[K] = E\left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} P \cdot np$

Likewise, Variance of binomial $\text{Var}(k) = np(1-p)$
$\text{Var}(k) = \text{Var}\left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{Var}(X_i) \quad \text{since iid.}$
$= \sum_{i=1}^{n} p(1-p) = np(1-p)$

But what about the original ratings?

mean of $K \rightarrow$ mean of $X$ (fraction of "thumbs ups")
$E[K] = np \quad \Rightarrow \quad E[X] = \frac{1}{n} E[K] = \frac{1}{n} np = p$

var of $K \rightarrow$ var of $X$
$\text{Var}(K) = np(1-p) \quad \Rightarrow \quad \text{Var}(X) = \text{Var}\left( \frac{K}{n} \right) = \frac{1}{n^2} \text{Var}(k) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$

How can we relate this to our data? We don't know $p$

\[ \hat{p} \]
\[ \text{ standard error of } \hat{p} \]
\[ \text{ distribution of } \hat{p} \]

we really want a range of values that we are confident contains the true proportion. \Rightarrow confidence interval

E.g., we are 100% confident that $p$ is in $[0, 1]$.

Useless though, can we do better?

Python
Q: why bother w/ normal distribution?
CI: confidence intervals.

\[ \hat{p} \text{ into } z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \]

approx.

this has normal distr. w/ mean 0 and s.d. 1.

from normal distr. 95% probability that \( z \) lands in \((-1.96, 1.96)\)

so there's our confidence interval!

\[ L_p = \hat{p} - 1.96 \frac{\sqrt{p(1-p)}}{n} \]

\[ U_p = \hat{p} + 1.96 \frac{\sqrt{p(1-p)}}{n} \]

Specifically, the probability a sample from this distribution falls between \( L_p \) and \( U_p \) is 95%:

\[ P( L_p \leq \hat{p} \leq U_p ) = 0.95 \]

1. Wald approx \( \Rightarrow \) replace \( \sigma \) w/ s sample variance. (crazy!)
2. Wilson score \( \Rightarrow \)

but \( \hat{p} \) is inside \( \sigma \)!

How to take care of this unknown?

Start here
plug in \( \hat{p} \) and \( z \) and solve for \( p \):

\[
\text{solve } \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} = z \text{ for } p:
\]

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

\[
1 + \frac{z^2}{n}
\]

rewrite this to understand it better:

\[
(\ldots) = A \pm B \quad \text{let's look at just } A
\]

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

remember \( \hat{p} = \frac{k}{n} \) \( k \) successes out of \( n \) trials

\[
z = 1.96 \times 2.
\]

\[
\approx \frac{k/n + 4/2n}{1 + 4/n} = \frac{k+2}{n+4} = \frac{k+2}{n+4}.
\]

Wilson score is a smoothed wald approximation!

add 2 successes and 2 failures \( k \to k+2 \)

\( n \to n+4+2 \)

this idea (smoothing) is very common. can appear ad hoc but can be well motivated!