MATH 180A: INTRO TO PROBABILITY
(FOR DATA SCIENCE)

www.math.ucsd.edu/~tkemp/180A

Today: § 4.3 - 4.4

Next: § 4.5

HW4 due TONIGHT by 11:59pm

Lab 4 due next Wednesday (Nov 6) by 11:59pm

Example

Flip a fair coin n times. How does

$$\mathbb{P}\left(\frac{\#\text{Heads}}{n} \geq 50.01\%\right)$$

behave as $n \to \infty$?

Suppose after 10,000 flips, there are 5,001 Heads. Should we doubt that the coin is really fair?

What if, after 1,000,000 flips, there are 500,100 Heads. Now how confident should we be that the coin is really fair?

Confidence

Suppose we have a coin that is biased by some unknown amount;

$$X \sim \mathrm{Ber}(p)$$ unknown p !

How can we figure out what p is?

Use the law of large numbers: $p = \lim_{n\to\infty} \frac{S_n}{n}$

We can't actually wait around for $n \to \infty$. Instead, we estimate

$$p \approx \hat{p} := \frac{S_n}{n}$$ for some large n.

The question is: how good an estimate is this for given n?
Or, turning it around: how big must you take n to get an estimate of a certain accuracy?

A Maximum Likelihood Estimate

Want to find n large enough that (with $\hat{p} = S_n/n$)

$$\mathbb{P}(|\hat{p} - p| < \varepsilon) \geq \text{(high probability)}$$

$\uparrow$ chosen tolerance

$$\mathbb{P}(|\hat{p} - p| < \varepsilon) \approx 2\,\Phi\left(\varepsilon\sqrt{n}/\sqrt{p(1-p)}\right) - 1 .$$

Conclusion: $\mathbb{P}(|\hat{p} - p| < \varepsilon) \underset{(\approx)}{\geq} 2\,\Phi(\qquad) - 1 .$

Example: (of the Beast) How many times should we flip a coin, biased an unknown amount p, so that the estimate $\hat{p} = S_n/n$ is within a tolerance of 0.05 of the true value p, with probability $\geq 99\%$?

Confidence Intervals

Turning this around: if we can't control n, we would like to say how accurate the sample mean is as an estimate of the true mean, for a given number n of samples.

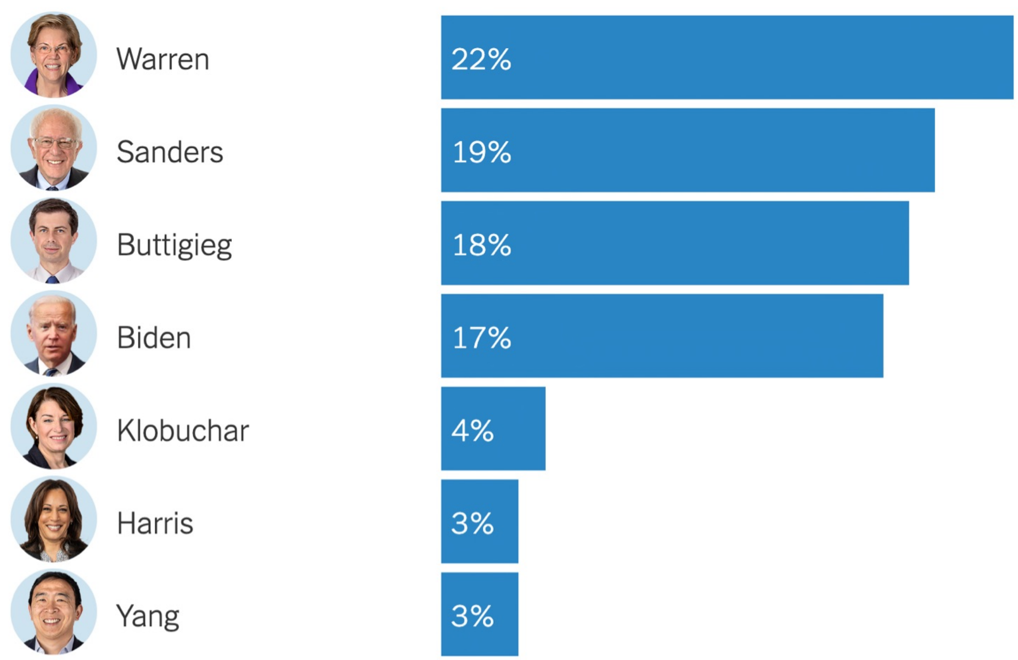
Eg. A coin (of unknown bias p) is tossed 1000 times. 450 Heads come up. Within what tolerance can we say we know the true value of p with probability $\geq 95\%$?

If an experiment is repeated in many independent trials, and the preceding (normal approximation) estimates yield

$$P(|\hat{p}-p|<\varepsilon) \geq 95\%$$

we say $[\hat{p}-\varepsilon, \hat{p}+\varepsilon]$ is the 95% confidence interval for p.

The same statement might be given as "$p=\hat{p}$ with margin of error ε (95 times out of 100)".



Source: New York Times Upshot/Siena College poll conducted Oct. 25-30.

Poll conducted Oct 25-30 of 439 Iowa Democratic caucusgoers.

Margin of error:

Poisson Approximation

4.4

$S_n \sim \mathrm{Bin}(n, \lambda/n)$: $\lim_{n\to\infty} \mathbb{P}(S_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}$

Quantitative Bound:

Theorem: If $X \sim \mathrm{Bin}(n,p)$ and $Y \sim \mathrm{Poisson}(np)$, for any subset $A \subseteq \mathbb{N}$

$$|\mathbb{P}(X\in A) - \mathbb{P}(Y \in A)| \le np^2$$

Upshot: if np^2 is small, use Poisson Approximation.
if $np(1-p)$ is big, use Normal Approximation.