Math 180A: Intro to Probability
(for Data Science)

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

Today: § 4.3 - 4.4

Next: § 4.4 - 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

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\#\text{Heads}}{n} \geq 50.01\%\right) = 0$$

$\uparrow$ $\frac{1}{2} + 0.0001$

behave as $n \to \infty$?

50.01%

$= 0.5001$

$\boxed{n = 10^{10} \quad \sqrt{n} = 10^5 \quad 2\sqrt{n}(0.0001) = 200 \quad 1 - \Phi(200) = 0}$

$\sqrt{n} = 100$
$n = 10{,}000$
$\varepsilon = 0.0001$

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

$\Big\}$ $1 - \Phi(0.02)$
$\geq 40\%$
(49%)

$\sqrt{n} = 1000$

$\Phi(0.2)$
$1-$
$\geq 40\%$

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

$S_n = \#\text{Heads} \sim \text{Bin}(n, \tfrac{1}{2})$

$$\mathbb{P}\left(\frac{S_n}{n} \geq \frac{1}{2} + \varepsilon\right) = \mathbb{P}\left(\frac{S_n - \frac{1}{2}n}{n} \geq \varepsilon\right) = \mathbb{P}\left(\frac{S_n - \frac{1}{2}n}{\sqrt{n}/2} \geq 2\varepsilon\sqrt{n}\right) \approx \mathbb{P}(X \geq 2\sqrt{n}\,\varepsilon)$$

Normal: $\mathcal{N}(0,1)$

replace n w/ $\sqrt{\text{Var}(S_n)} = \sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{\sqrt{n}}{2}$

$= 1 - \mathbb{P}(X < 2\varepsilon\sqrt{n})$

$= 1 - \Phi(2\varepsilon\sqrt{n})$

Confidence

4.3

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?

$$|\hat{p} - p| < \varepsilon \ (= 0.01)$$

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

" $\hat{p}$ is within margin of error ε of p, w probability 95%."

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

$X \sim \mathcal{N}(0,1)$

$$\mathbb{P}(|\hat{p}-p| < \varepsilon) = \mathbb{P}\left(\frac{|S_n - np|}{n} < \varepsilon\right) = \mathbb{P}\left(\frac{|S_n-np|}{\sqrt{np(1-p)}} < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \approx \mathbb{P}\left(|X| < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

$\hat{p} = \frac{S_n}{n}$ $\quad$ $\sqrt{np(1-p)}$

$$= \Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi(- \text{" "})$$

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

$0 \leq p \leq 1$

$p(1-p) \leq \frac{1}{4}$

max @ $p = \frac{1}{2}$

$$\frac{1}{\sqrt{p(1-p)}} \geq 2$$

$\Phi \uparrow$

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

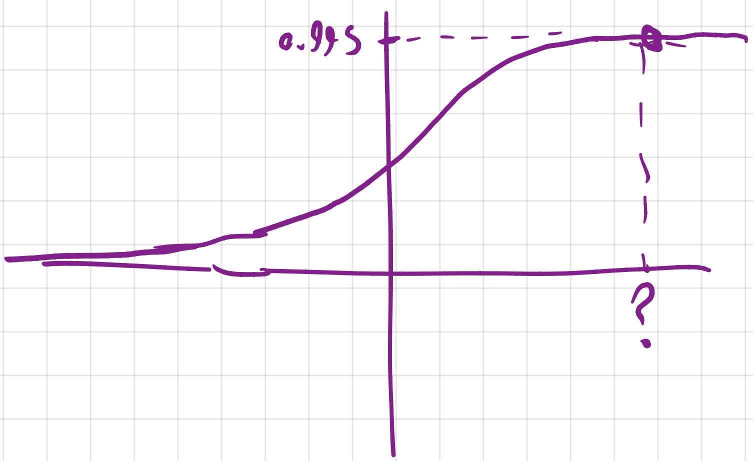
Example (of the Beast): How many times should we flip a coin, biased an unknown amount p, so that the estimate $\hat{p} = S_n/n$ is within a tolerance of 0.05 of the true value p, with probability $\geq 99\%$?

Want n large enough that

$$\mathbb{P}(|\hat{p} - p| < 0.05) \gtrsim 99\%$$

make sure

We know $\mathbb{P}(|\hat{p} - p| < 0.05) \geq (\approx) \; 2\Phi(2(0.05)\sqrt{n}) - 1 \geq 99\%$



$$\Phi(\underbrace{2(0.05)\sqrt{n}}) \geq 0.995$$

$$\therefore \; 2(0.05)\sqrt{n} \gtrsim 2.58$$

$$\sqrt{n} \gtrsim 25.8$$

$$n \gtrsim 665.64$$

666

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 $\geq 95\%$?

Estimate $p \approx \hat{p} = \frac{S_{1000}}{1000} = 0.45$

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

Know: $\mathbb{P}(|p-\hat{p}|<\varepsilon) \gtrapprox 2\Phi(2\varepsilon\sqrt{1000})-1 \geq 0.95$

$\Phi(2\varepsilon\sqrt{1000}) \geq 0.975$

$2\varepsilon\sqrt{1000} \geq 1.96 \leadsto \varepsilon \geq \frac{1.96}{2\sqrt{1000}} \doteq 0.031$

I.e. $|p-0.45| < 0.031$ w $\mathbb{P} \geq 95\%$

$0.45 - 0.031 < p < 0.45 + 0.031$

$p \in [0.419, 0.481]$ w $\mathbb{P} \geq 95\%$

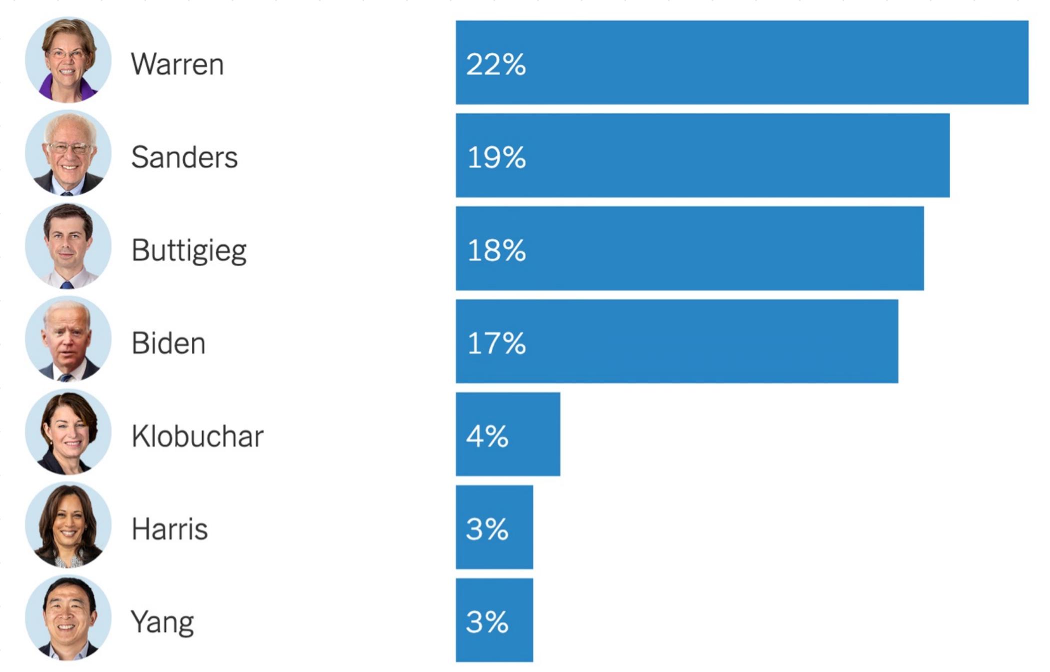
95% confidence interval

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

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

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

$$\mathbb{P}(|\underset{0.22}{p-\hat{p}}| < \varepsilon) \geq 2\Phi(2\varepsilon\sqrt{439})-1$$

$$(\approx) \quad \geq 0.95$$

i.e. $2\varepsilon\sqrt{439} \geq 1.96$

$\varepsilon \geq 4.68\%$

Margin of error: 4.7%

Poisson Approximation

4.4

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

Quantitative Bound:

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

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

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