CSE 312

Foundations of Computing II

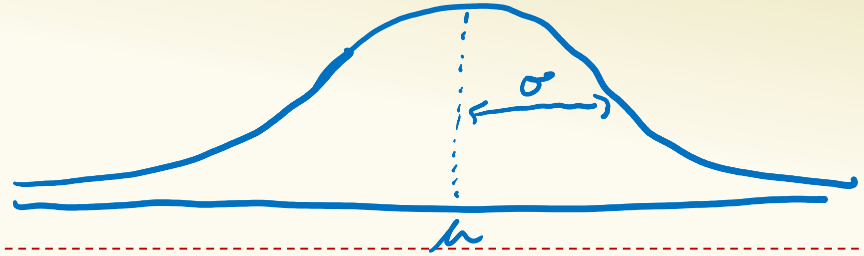
Lecture 18: CLT & Polling

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Rachel Lin, Hunter Schafer

Slide Credit: Based on Stefano Tessaro's slides for 312 19au
incorporating ideas from Alex Tsun's and Anna Karlin's slides for 312 20su and 20au

The Normal Distribution



Definition. A **Gaussian (or normal) random variable** with parameters $\mu \in \mathbb{R}$ and $\sigma \geq 0$ has density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

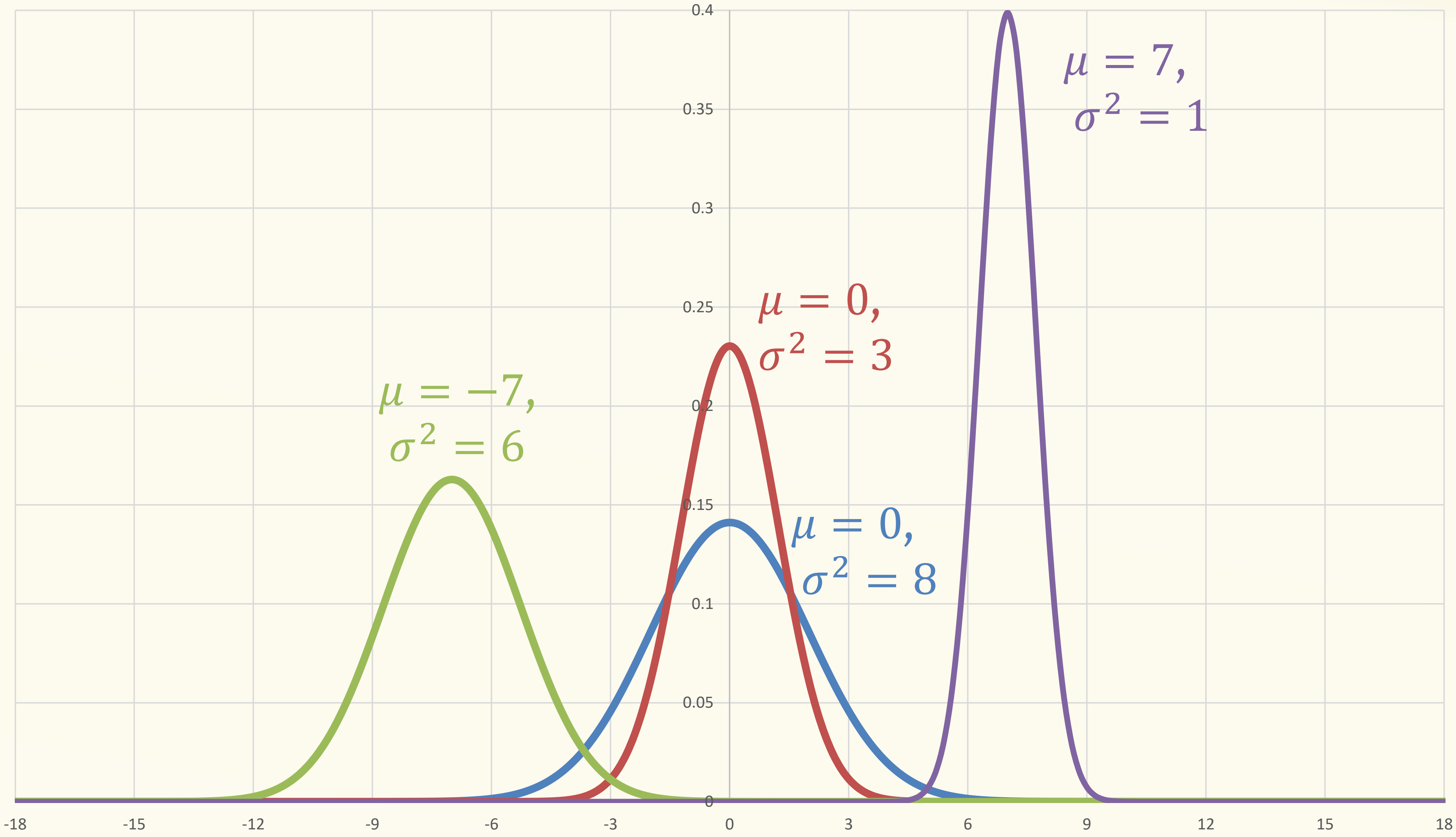
(We say that X follows the Normal Distribution, and write $X \sim \mathcal{N}(\mu, \sigma^2)$)

Fact. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}(X) = \mu$, and $\mathrm{Var}(X) = \sigma^2$

Proof is easy because density curve is symmetric around μ, $f_X(\mu - x) = f_X(\mu + x)$

The Normal Distribution

Aka a “Bell Curve” (imprecise name)



What about Non-standard normal?

$Var(aX) = a^2 Var(X)$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

← "standardize"

Therefore,

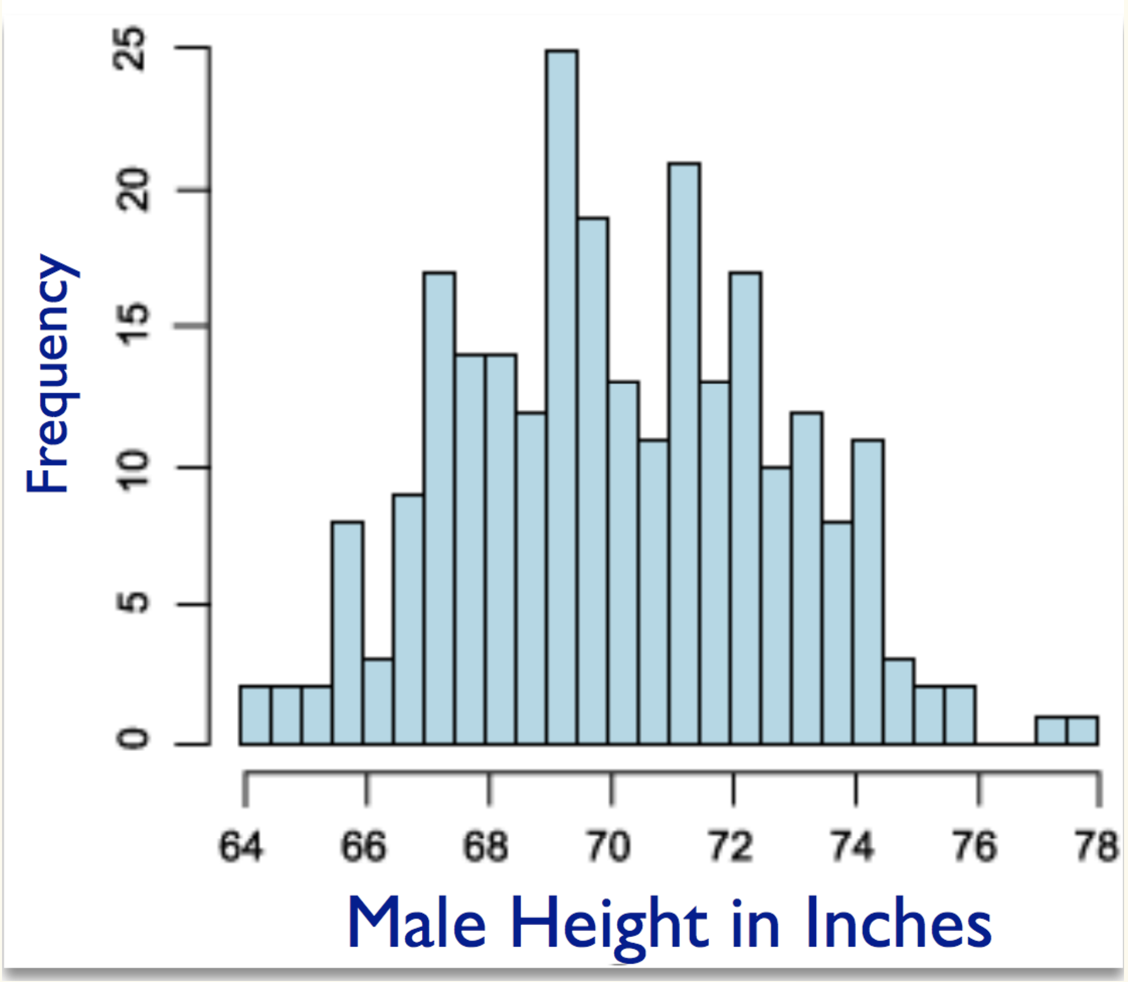
$$F_X(z) = \mathbb{P}(X \le z) = \mathbb{P}\left(\frac{X-\mu}{\sigma} \le \frac{z-\mu}{\sigma}\right) = \Phi\left(\frac{z-\mu}{\sigma}\right)$$

Agenda

- Central Limit Theorem (CLT)
- Polling

Gaussian in Nature

Empirical distribution of collected data often resembles a Gaussian …



e.g. Height distribution resembles Gaussian.

R.A.Fisher (1918) observed that the height is likely the outcome of the sum of many independent random parameters, i.e., can written as

$$X = X_1 + \cdots + X_n$$

Sum of Independent RVs

i.i.d. = independent and identically distributed

$X_1, \dots, X_n$ i.i.d. with expectation μ and variance σ^2

Define

$$S_n = X_1 + \cdots + X_n$$

$$\mathbb{E}(S_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = n\mu$$

$$\mathrm{Var}(S_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\sigma^2$$

Not surprising
- Linearity of expectation
- Variance of sum of indep. r.v.s

Empirical observation: S_n looks like a normal RV as n grows.

very surprising!

Central Limit Theorem

$E[S_n] = n\mu$ $\quad$ $Var(S_n) = \sigma^2 n$

$X_1, \ldots, X_n$ i.i.d., each with expectation μ and variance σ^2

Define $S_n = X_1 + \cdots + X_n$ and

$$Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Standardized S_n

$$\mathbb{E}(Y_n) = \frac{1}{\sigma\sqrt{n}}(\mathbb{E}(S_n) - n\mu) = \frac{1}{\sigma\sqrt{n}}(n\mu - n\mu) = 0$$

$Var(aX) = a^2 Var(X)$

$$\mathrm{Var}(Y_n) = \frac{1}{\sigma^2 n}\left(\mathrm{Var}(S_n - n\mu)\right) = \frac{\mathrm{Var}(S_n)}{\sigma^2 n} = \frac{\sigma^2 n}{\sigma^2 n} = 1$$

$$Var\left(\frac{1}{\sigma\sqrt{n}} S_n\right) = \frac{1}{\sigma^2 n} Var(S_n) = \frac{\sigma^2 n}{\sigma^2 n} = 1$$

Central Limit Theorem

$$Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

Theorem. (Central Limit Theorem) The CDF of Y_n converges to the CDF of the standard normal $\mathcal{N}(0,1)$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}(Y_n \leq y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-x^2/2} \mathrm{d}x$$

Also stated as:

- $\lim_{n\to\infty} Y_n \to \mathcal{N}(0,1)$
- $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ where $\mu = E[X_i]$ and $\sigma^2 = Var(X_i)$

Agenda

- Central Limit Theorem (CLT)
- Polling

Magic Mushrooms

Last fall, Oregonians voted on whether to legalize the therapeutic use of “magic mushrooms”.

Poll to determine the fraction of the population that will vote in favor.

- Call up a random sample of n people to ask their opinion
- Report the empirical fraction

Questions

- Is this a good estimate?
- How to choose n?

Polling Accuracy

Often see claims that say

> "Our poll found 80% support. This poll is accurate to within 5% with 98% probability"

Will unpack what this and how they sample enough people to know this is true.

Formalizing Polls

Population size N, true fraction of voting in favor p, sample size n.

Problem: We don't know p

Polling Procedure

for i = 1 … n:

1. Pick uniformly random person to call (prob: $1/N$)
2. Ask them how they will vote

$$X_i = \begin{cases} 1, & \text{voting in favor} \\ 0, & \text{otherwise} \end{cases}$$

Example:

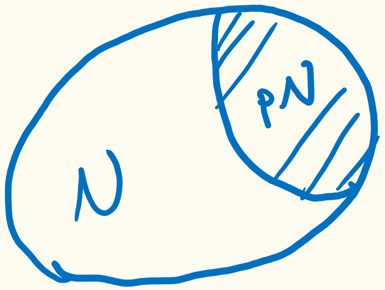
$$\frac{271}{500} \approx 0.542$$

Report our estimate of p: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Random Variables

What type of r.v. is X_i?

$X_i \sim Ber(p)$



Poll: pollev.com/hunter312

	Type	$E[X_i]$	$Var(X_i)$
✓ a.	Bernoulli	p	$p(1-p)$
b.	Bernoulli	p	p^2
c.	Geometric	p	$\frac{1-p}{p^2}$
d.	Binomial	np	$np(1-p)$

What type of r.v. is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$?

$E[\bar{X}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{n\mu}{n} = \mu$

$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$

$\mu = p, \quad \sigma^2 = p(1-p)$

Poll: pollev.com/hunter312

	$E[\bar{X}]$	$Var(\bar{X})$
a.	np	$np(1-p)$
b.	p	$p(1-p)$
✓ c.	p	$p(1-p)/n$
d.	p/n	$p(1-p)/n$

Central Limit Theorem

With i.i.d random variables $X_1, X_2, \ldots, X_n$ where $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$

As $n \to \infty$,

$$\frac{X_1 + X_2 + \cdots X_n - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$$

Restated: As $n \to \infty$,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Roadmap: Bounding Error

Goal: Find the value of n such that 98% of the time, the estimate $\bar{X}$ is within 5% of the true p

1. Define probability of a "bad event"
2. Apply CLT
3. Convert to a standard normal
4. Solve for n

See notes for walk through

Idealized Polling

So far, we have been discussing “idealized polling”. Real life is normally not so nice ☹

Assumed we can sample people uniformly at random, not really possible in practice

- Not everyone responds
- Response rates might differ in different groups
- Will people respond truthfully?

Makes polling in real life much more complex than this idealized model!