ENGG 2430 / ESTR 2004: Probability and Statistics

Spring 2019

9. Limit Theorems

Andrej Bogdanov

Many times we do not need to calculate probabilities exactly

An approximate or qualitative estimate often suffices

$$\mathbf{P}(\text{magnitude 7+ earthquake within 10 years}) = ?$$

I toss a coin 1000 times. The probability that I get a streak of 3 consecutive heads is

A $< 10\%$

B $\approx 50\%$

C $> 90\%$

1000 TOSSES

$\underbrace{H\,H\,T}_{1/8}\ \underbrace{H\,T\,H}_{1/8}\ \cdots\ \underbrace{\cdot\ \cdot\ \cdot}_{1/8}\ H\,T$

$E_1 \quad E_2 \quad \cdots \quad E_{333}$

$P(\text{NONE OF } E_1, \ldots, E_{333} \text{ OCCUR}) = \left(\frac{7}{8}\right)^{333} = 0.\underbrace{00 \cdots 0}_{20}4$

$P(3 \text{ CONSECUTIVE HEADS}) \geq 1 - \left(\frac{7}{8}\right)^{333}$

I toss a coin 1000 times. The probability that I get a streak of 14 consecutive heads is

A	B	C X
$< 10\%$	$\approx 50\%$	$> 90\%$

$P(14\ H) \geq 1 - (1 - 2^{-14})^{\lfloor 1000/14 \rfloor} \approx 0$

N = NUMBER OF 14 H STREAKS

H T H . . . . .

$\underbrace{\quad}_{N_1}$ $\underbrace{\quad}_{N_2}$ . . .

$N = N_1 + \dots + N_{987}$

$N_i = \begin{cases} 1 & \underbrace{H \cdots H}_{14} \text{ STARTING AT } i \\ 0 & \text{IF NOT} \end{cases}$

$E[N] = \underbrace{E[N_1]}_{2^{-14}} + \dots + \underbrace{E[N_{987}]}_{2^{-14}} = 987 \cdot 2^{-14} \approx 0.06$

$P(N \geq 1) \leq E[N]/1 \approx 0.06.$

Markov's inequality

For every non-negative random variable X and every value a:

$$\mathbf{P}(X \geq a) \leq \mathbf{E}[X] \;/\; a.$$

Proof

$$E[X] = \underbrace{E[X \mid X \geq a]}_{\geq a} P(X \geq a) + \underbrace{E[X \mid X < a]}_{\geq 0} \underbrace{P(X < a)}_{\geq 0}$$

$$\geq a \cdot P(X \geq a)$$

1000 people throw their hats in the air. What is the probability at least 100 people get their hat back?

N = NUMBER OF HATS RETURNED

$E[N] = 1$

$$P(N \geq 100) \leq \frac{E[N]}{100} = \frac{1}{100} = 1\%$$

$X = \text{Uniform}(0, 4)$. **How does $\mathbf{P}(X \geq x)$ compare with Markov's inequality?**

$E[X] = 2$

x	1	2	3	4
$P(X \geq x)$	$3/4$	$1/2$	$1/4$	0
$E[X]/x$	2	1	$2/3$	$1/2$

USELESS (for $x = 1, 2$)

FAR FROM REALITY (for $x = 3, 4$)

IN GENERAL MARKOV IS NOT VERY USEFUL WHEN PDF IS "SPREAD OUT" AROUND MEAN.

I toss a coin 1000 times. What is the probability I get 3 consecutive heads

(a) at least 700 times

(b) at most 50 times

N = # TIMES I GET HHH

$E[N] = E[N_1 + \cdots + N_{998}] = 998 \cdot \frac{1}{8} = 124.75$

$P(N \geq 700) = P(N \geq 5.61 \cdot E[N]) \leq \frac{1}{5.61} \approx 18\%$

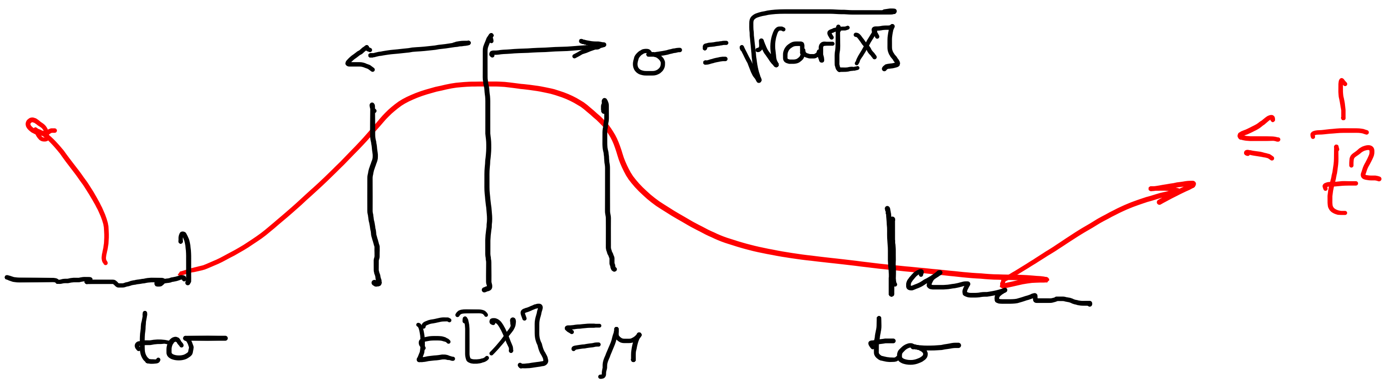
$P(N \leq 50)$ NO INFORMATION

Chebyshev's inequality

For every random variable X and every t:

$$P(|X - \mu| \geq t\sigma) \leq 1 / t^2.$$

where $\mu = E[X]$, $\sigma = \sqrt{Var[X]}$.



Chebyshev's inequality

For every random variable X and every t:

$$P(|X-\mu| \geq t\sigma) \leq 1 / t^2.$$

where $\mu = E[X]$, $\sigma = \sqrt{Var[X]}$.

Proof. $Y = (X - E[X])^2$

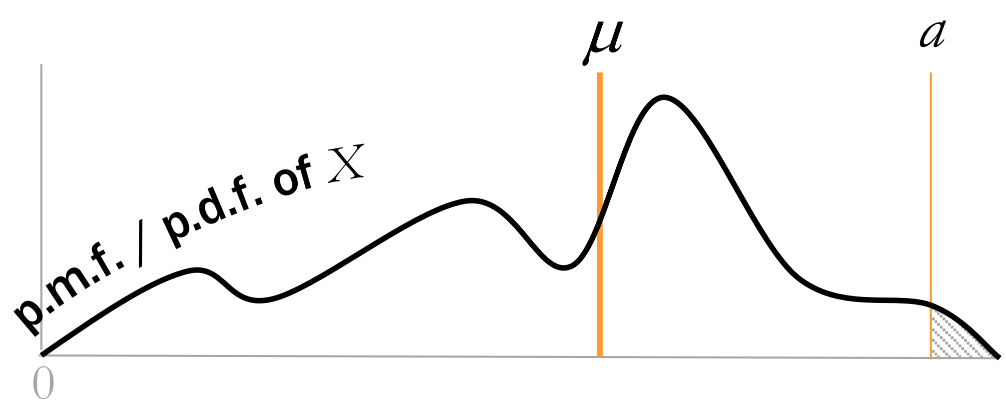
$Var[X] = E[Y]$ $\qquad Y \geq 0$

$Pr[Y \geq t^2 E[Y]] \leq \frac{1}{t^2}$ Markov

$Pr[(X - E[X])^2 \geq t^2 Var[X]] \leq \frac{1}{t^2}$

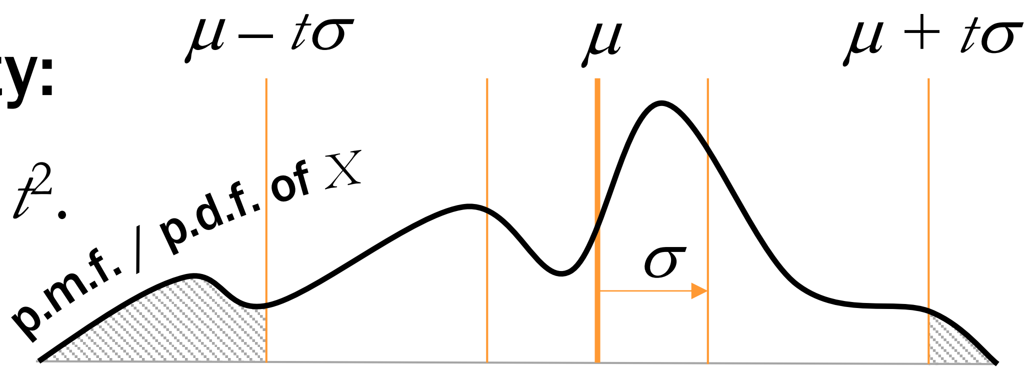
Markov's inequality:

$$P(X \geq a) \leq \mu / a.$$



Chebyshev's inequality:

$$P(|X - \mu| \geq t\sigma) \leq 1 / t^2.$$



I toss a coin 64 times. What is the probability I get at most 24 heads?

$X = \text{Binomial}(64, \tfrac{1}{2})$

$E[X] = 32$

$Var[X] = 64 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = 16$

$\sigma = 4$

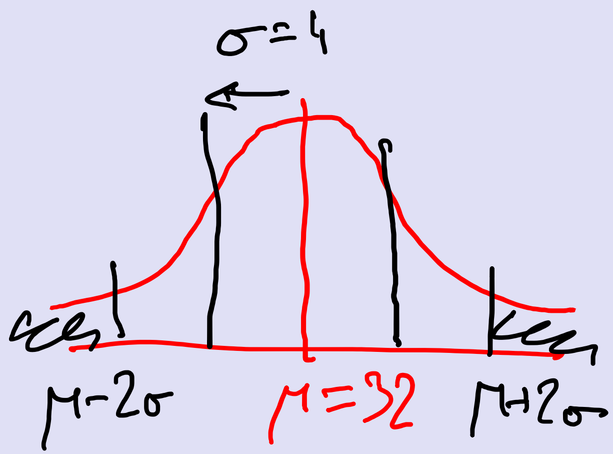
$P(X \le 24)$

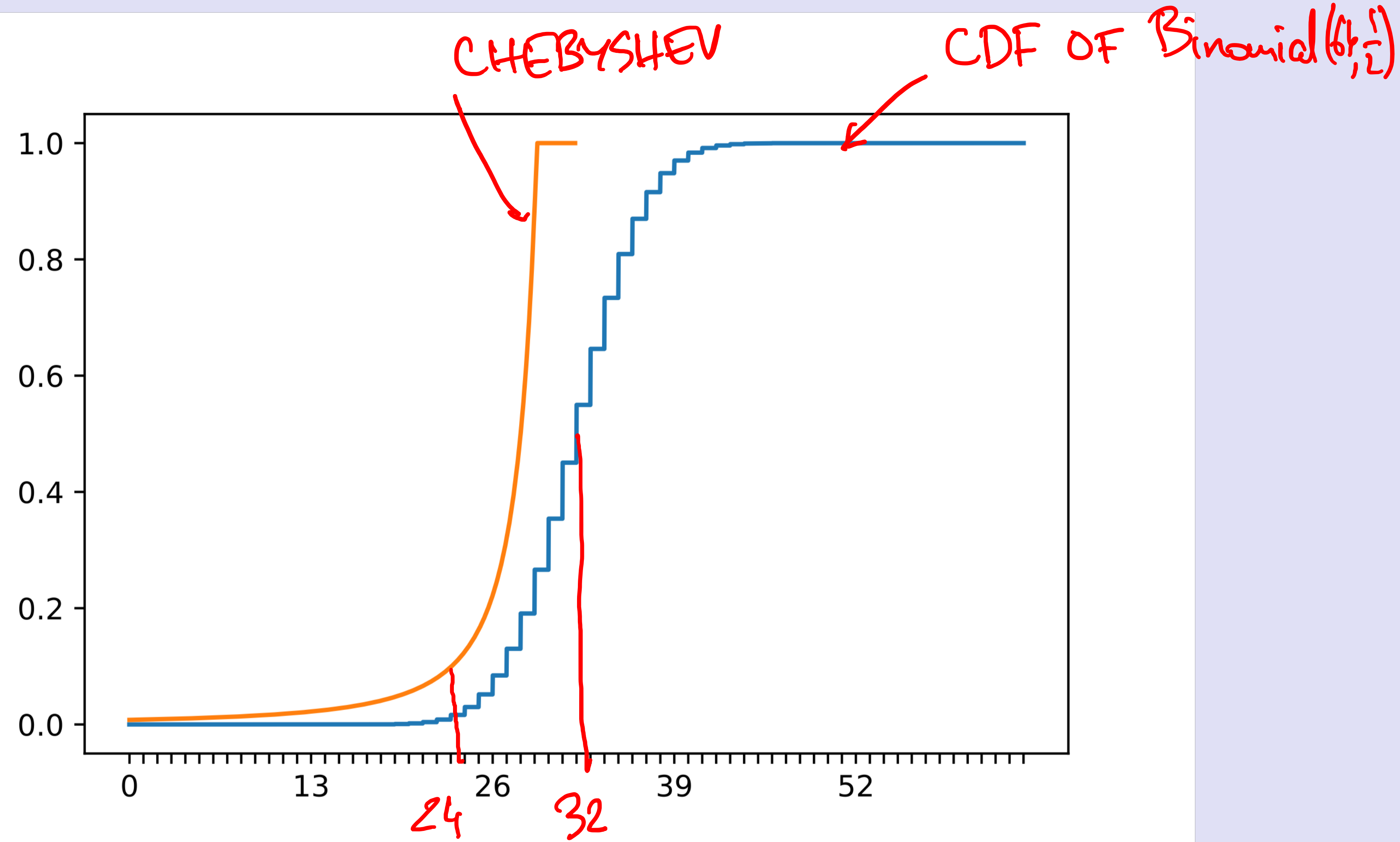
$= P(X \le \mu - 2 \cdot \sigma)$

$P(|X - \mu| \le 2\sigma) \le \frac{1}{2^2} = \frac{1}{4}$

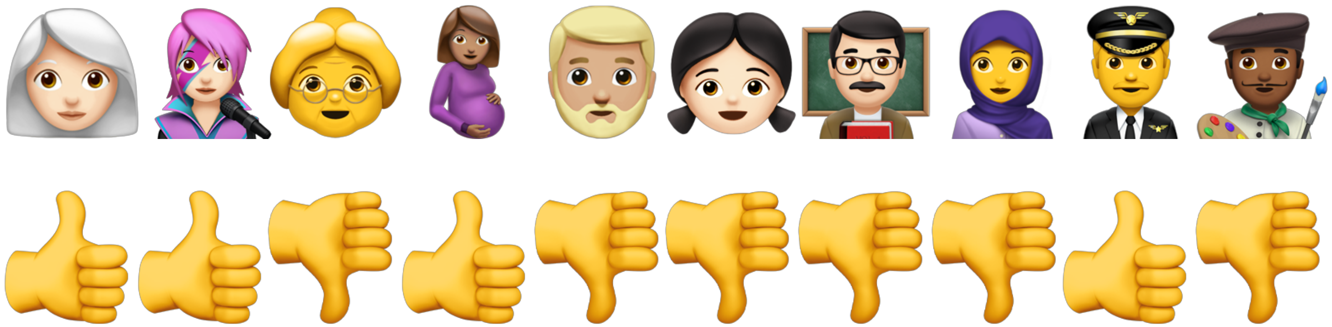
$P(X \le 24) \le \frac{1}{4}.$

$\le \frac{1}{8}$ BY SYMMETRY




CHEBYSHEV
CDF OF Binomial(64, 1/2)
1.0
0.8
0.6
0.4
0.2
0.0
0
13
26
39
52
24
32

Polling



μ = TRUE POPULARITY

$$X = X_1 + \ldots + X_n$$

NUMBER OF PEOPLE POLLED

1 WITH PROB μ

0 WITH PROB $1-\mu$

EST $\hat{X} - \frac{X}{n}$

Polling

How accurate is the pollster's estimate X/n?

$\mu = \mathbf{E}[X_i]$, $\sigma = \sqrt{\mathbf{Var}[X_i]}$ $= \sqrt{\mu(1-\mu)}$

$\mathbf{E}[X] =$ $E[X_1] + \ldots + E[X_n] = n\mu$

$\mathbf{Var}[X] =$ $Var[X_1] + \ldots + Var[X_n] = n\sigma^2$

$\sigma_X = n\sqrt{\sigma}$

Polling

$$
\begin{aligned}
\mathbf{P}(\,|X/n-\mu| \geq \varepsilon\,) &= P(|X-\mu n| \geq \varepsilon \cdot n) \\
&= P(|X-E[X]| \geq \varepsilon n) \\
&= P\Big(|X-E[X]| \geq \underbrace{\tfrac{\varepsilon}{\sigma}\sqrt{n}}_{t} \cdot \underbrace{\sqrt{n}\,\sigma}_{\sigma_X}\Big) \\
&\leq \frac{1}{t^2} \\
&= \underbrace{\frac{\sigma^2}{\varepsilon^2 \cdot n}}_{\delta}
\end{aligned}
$$

The weak law of large numbers

$X_1, \ldots, X_n$ are independent with same PMF/PDF

$\mu = E[X_i]$, $\sigma = \sqrt{Var[X_i]}$, $X = X_1 + \ldots + X_n$

For every $\varepsilon, \delta > 0$ and $n \geq \sigma^2/(\varepsilon^2\delta)$:

$$\mathbf{P}(|X/n - \mu| \geq \varepsilon) \leq \delta$$

SAMPLING ERROR

CONFIDENCE ERROR

We want confidence error $\delta = 10\%$ **and** sampling error $\varepsilon = 5\%$ **. How many people should we poll?**

$$n = \frac{\sigma^2}{\varepsilon^2 \delta} = \frac{\mu(1-\mu)}{\varepsilon^2 \delta} \leq \frac{1}{4\varepsilon^2 \delta} = \frac{1}{4\left(\frac{1}{20}\right)^2 \cdot \frac{1}{10}} = 1000$$

1000 IS <u>ENOUGH</u> (BUT MAYBE NOT NECESSARY)

1000 people throw their hats in the air. What is the probability at least 100 people get their hat back?

MARKOV $P(N \geq 100) \leq \frac{1}{100} = 0.01$

CHEBYSHEV $P(|N - \underbrace{E[N]}_{1}| \geq t\underbrace{\sigma}_{1}) \leq \frac{1}{t^2}$

$= P(|N-\mu| \geq 99\sigma)$

$\leq \frac{1}{99^2}$

≈ 0.0001

I toss a coin 1000 times. What is the probability I get 3 consecutive heads

(a) at least 250 times

(b) at most 50 times

$$N = N_1 + N_2 + \dots + N_{998} \qquad N_i = \begin{cases} 1 & \text{IF HHH AT } i \\ 0 & \text{IF NOT} \end{cases}$$

$$\mu = E[N] = 998 \cdot \frac{1}{8} = 124.75$$

$$\mathrm{Var}[N] = \sum \underline{\mathrm{Var}[N_i]} + \sum \underline{\mathrm{Cov}[N_i, N_j]}$$

$$\mathrm{Var}[N_i] = \frac{1}{8} - \frac{1}{64}$$

$$\mathrm{Cov}[N_i, N_{i+1}] = P(N_i = N_{i+1} = 1) - P(N_i = 1)P(N_{i+1} = 1) = \frac{1}{16} - \frac{1}{64}$$

$$\mathrm{Cov}[N_i, N_{i+2}] = P(N_i = N_{i+2} = 1) - P(N_i = 1)P(N_{i+2} = 1) = \frac{1}{32} - \frac{1}{64}$$

ALL OTHERS = 0 BY INDEPENDENCE

$$\mathrm{Var}[N] = 998 \cdot \left(\frac{1}{8} - \frac{1}{64}\right) + 2 \cdot 997 \cdot \left(\frac{1}{16} - \frac{1}{64}\right) + 2 \cdot 996 \cdot \left(\frac{1}{32} - \frac{1}{64}\right)$$

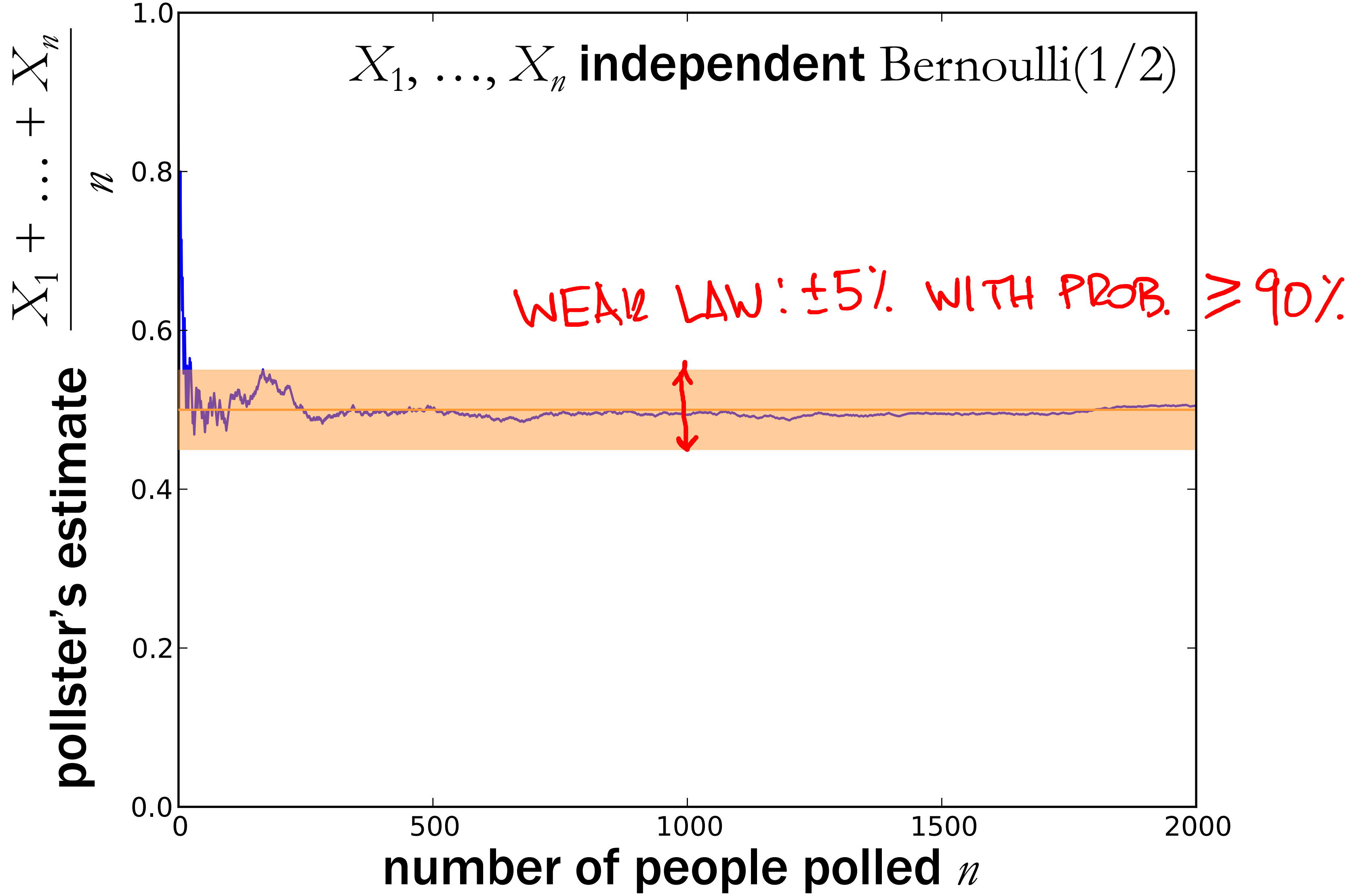
$$= 233.75$$

$$\sigma_N \approx 15.29$$

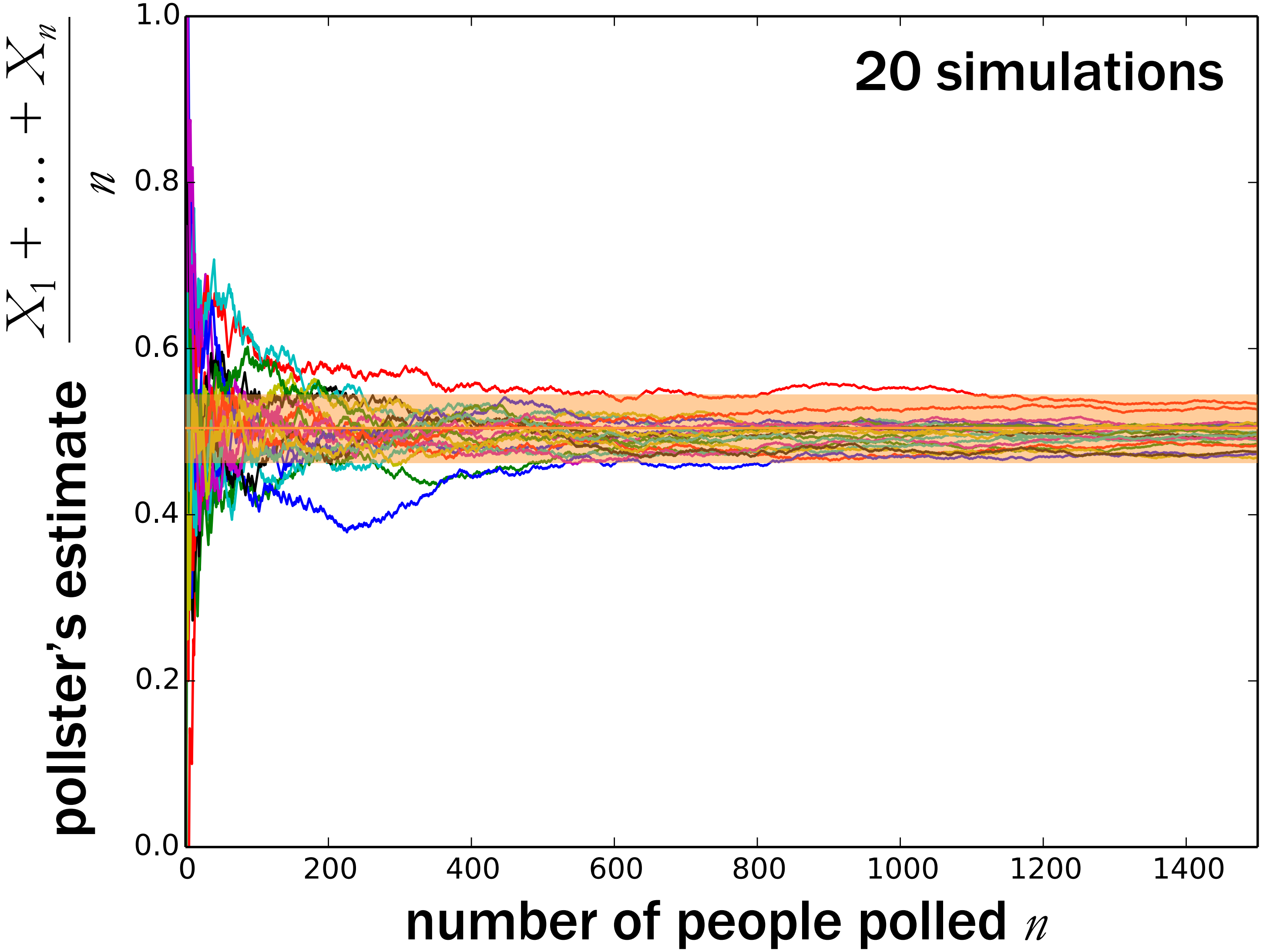
$$P(N \geq 250) \approx P(N \geq \mu + 8.19\sigma) \leq \frac{1}{8.19^2} \approx 0.015$$

$$P(N \leq 50) \approx P(N \leq \mu - 4.89\sigma) \leq \frac{1}{4.89^2} \approx 0.042$$

A polling simulation



A polling simulation



$X_1, \ldots, X_n$ are independent with same PMF/PDF

Let's assume n is large.

Weak law of large numbers:

$$X_1 + \ldots + X_n \approx \mu n \qquad \text{with high probability}$$

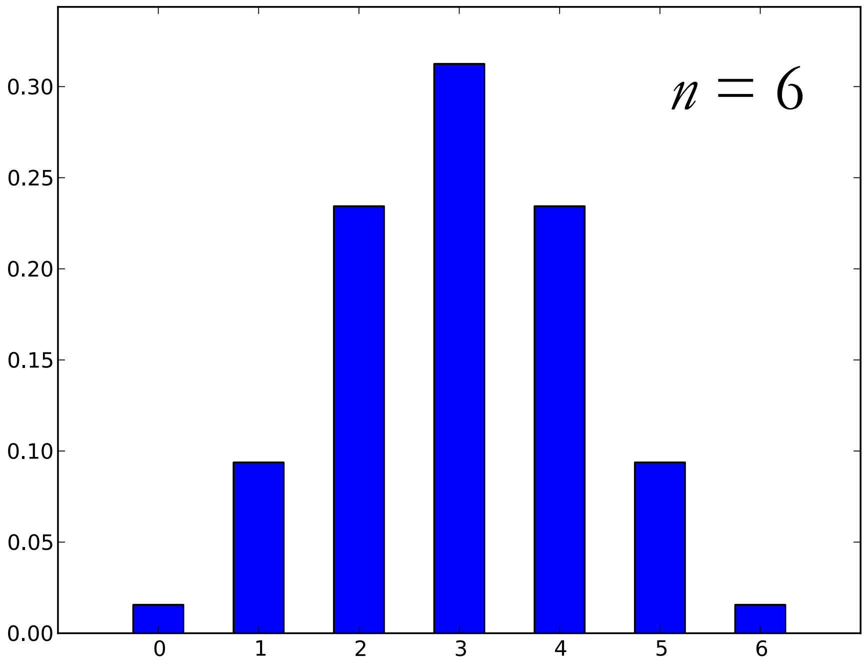
$$P(\,|X - \mu n| \geq t\sigma\sqrt{n}\,) \leq 1 / t^2.$$

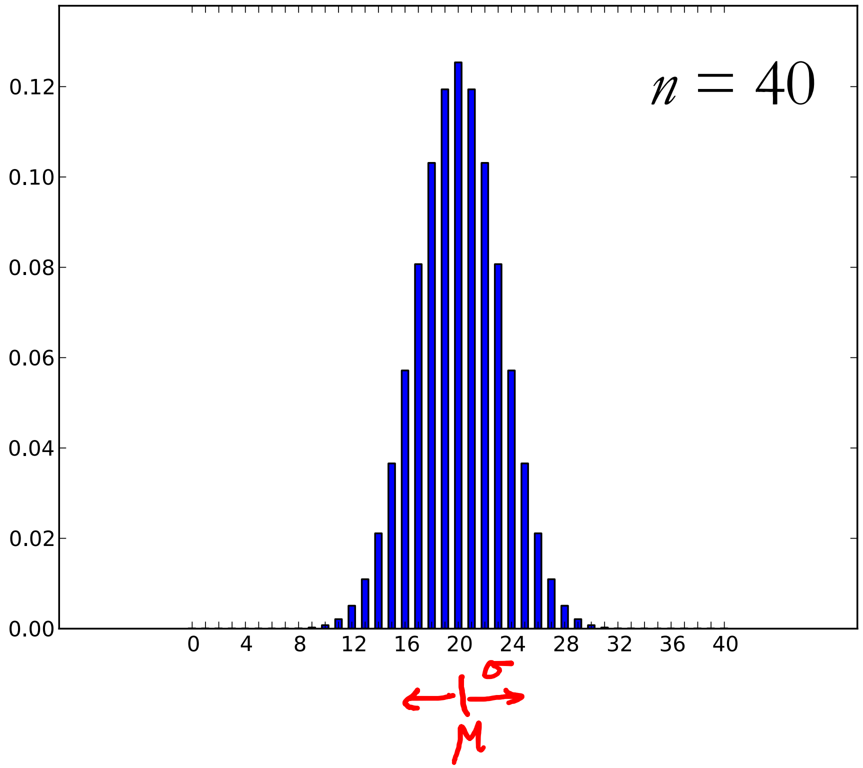
this suggests $X_1 + \ldots + X_n \approx \mu n + T\sigma\sqrt{n}$

RANDOM VARIABLE

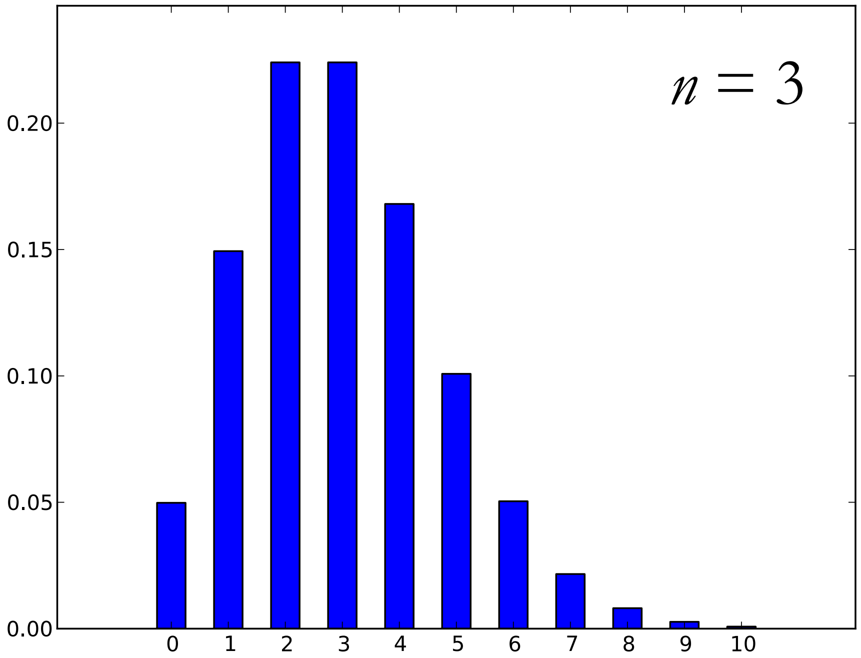
Some experiments

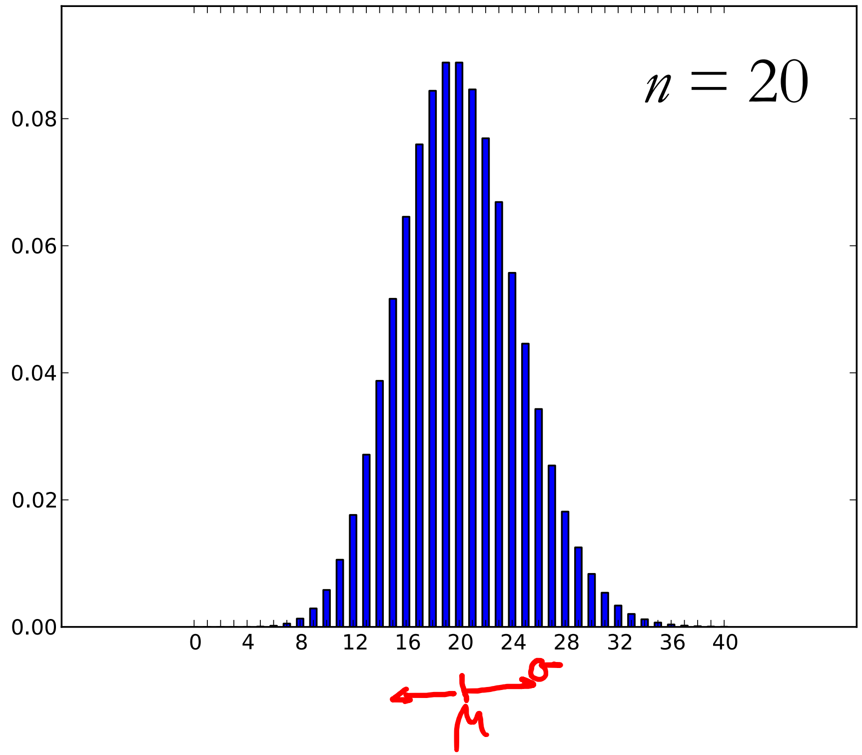
$X = X_1 + \ldots + X_n$ $\quad\quad$ X_i **independent** $\text{Bernoulli}(1/2)$



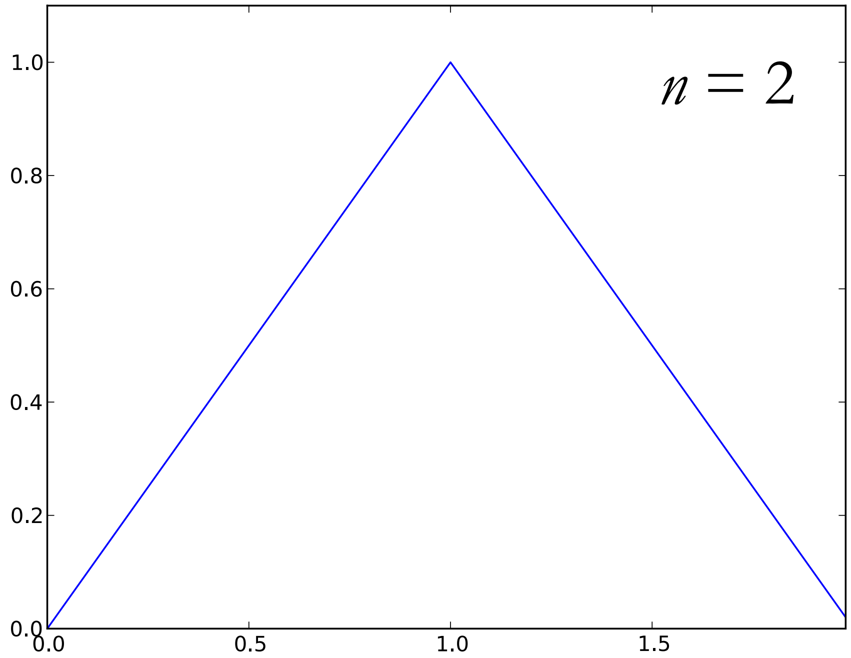


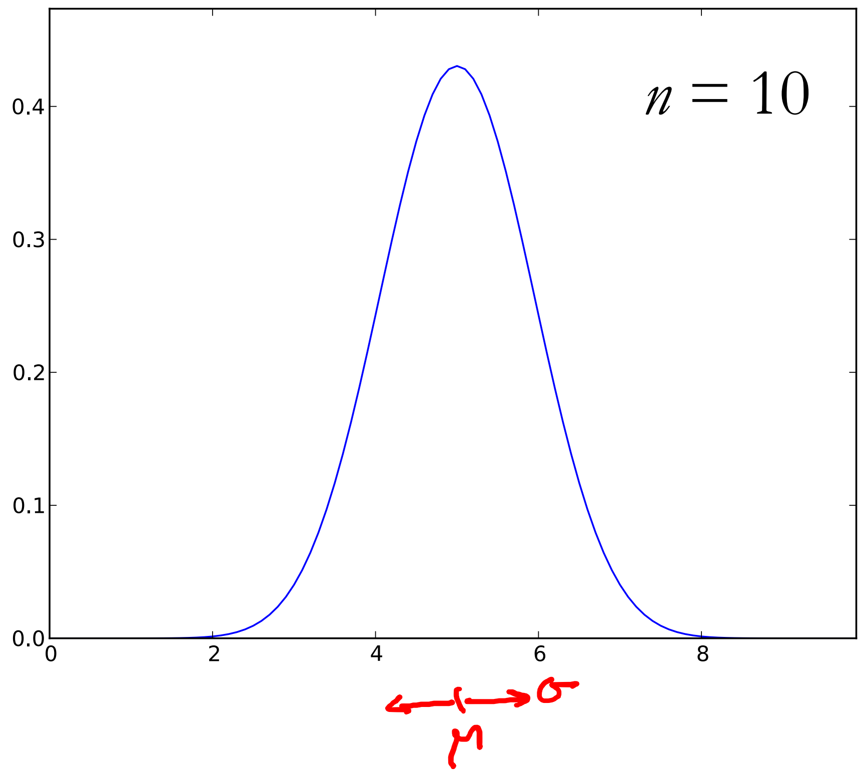
$$X = X_1 + \ldots + X_n \qquad X_i \text{ **independent** Poisson}(1)$$

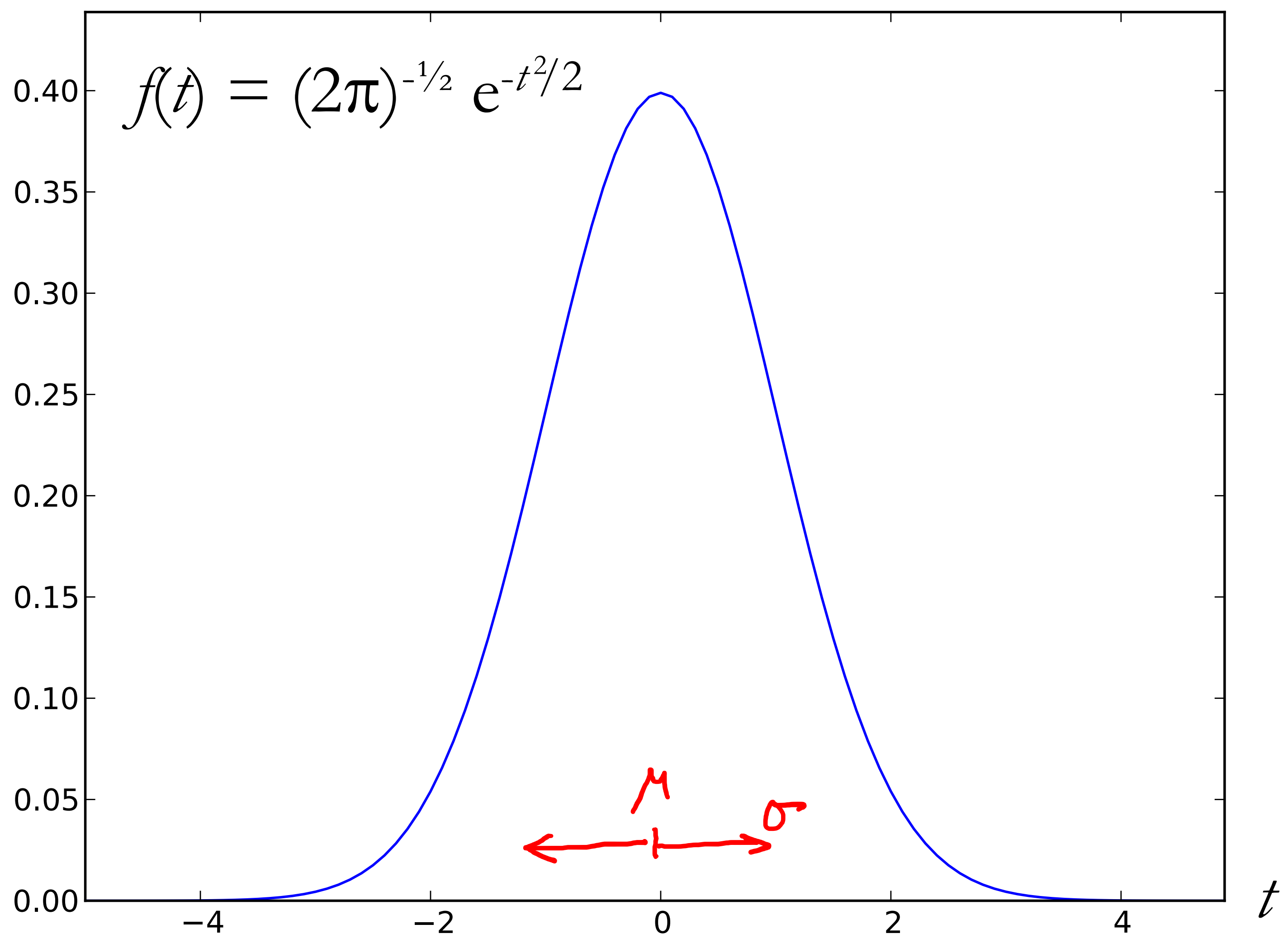




$X = X_1 + \ldots + X_n$ X_i **independent** Uniform(0, 1)







$f(t) = (2\pi)^{-1/2}\, e^{-t^2/2}$
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4
t
μ
σ

The central limit theorem

$X_1, \ldots, X_n$ are independent with same PMF/PDF

$\mu = E[X_i]$, $\sigma = \sqrt{Var[X_i]}$, $X = X_1 + \ldots + X_n$

For every t (positive or negative):

$$\lim_{n \to \infty} P(X \leq \mu n + t\sigma\sqrt{n}) = P(N \leq t)$$

where N is a normal random variable.

eventually,
everything
is normal

Toss a die 100 times. What is the probability that the sum of the outcomes exceeds 400?

$$X = X_1 + \dots + X_{100}$$

$$\mu = E[X] = 100 \cdot 3.5 = 350$$

$$Var[X] = 100 \cdot \left[\tfrac{1}{6}(1^2 + \dots + 6^2) - 3.5^2\right] \approx 291.67$$

$$\sigma = \sqrt{Var[X]} \approx 17.08$$

$$\left.\begin{aligned} P(X \geq 400) &\approx P(X \geq \mu + 2.92\sigma) \\ &\approx P(\mathrm{Normal}(0,1) \geq 2.92) \end{aligned}\right\} \text{CENTRAL LIMIT THEOREM}$$

$$\approx 0.0018.$$

We want confidence error $\delta = 1\%$ and sampling error $\varepsilon = 5\%$. How many people should we poll?

n PEOPLE $\quad X = X_1 + \dots + X_n \quad X_i = 0$ OR 1

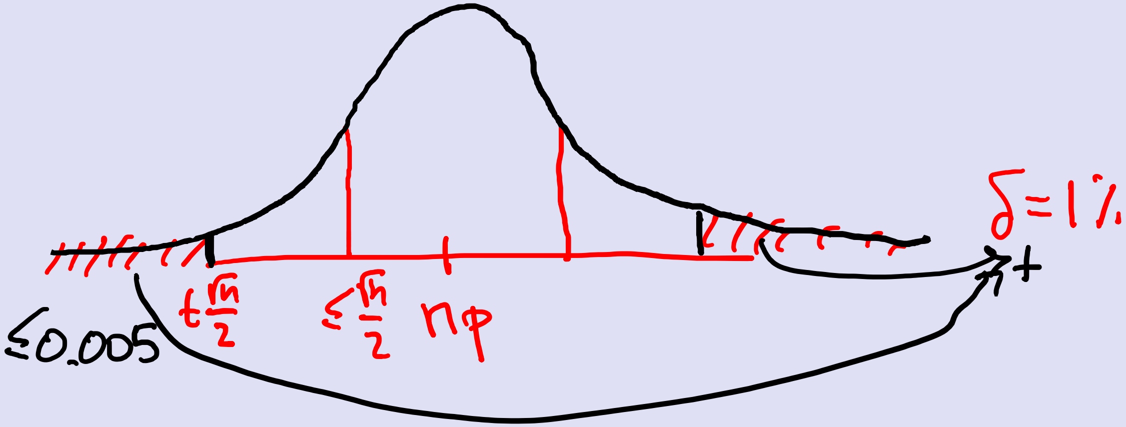
$E[X_1] = \dots = E[X_n] = p$

$\mathrm{Var}[X] = np(1-p)$

GOAL: ESTIMATE p.

$\sigma = \sqrt{np(1-p)}$

$\leq \sqrt{n}/2$



$t = 2.576$

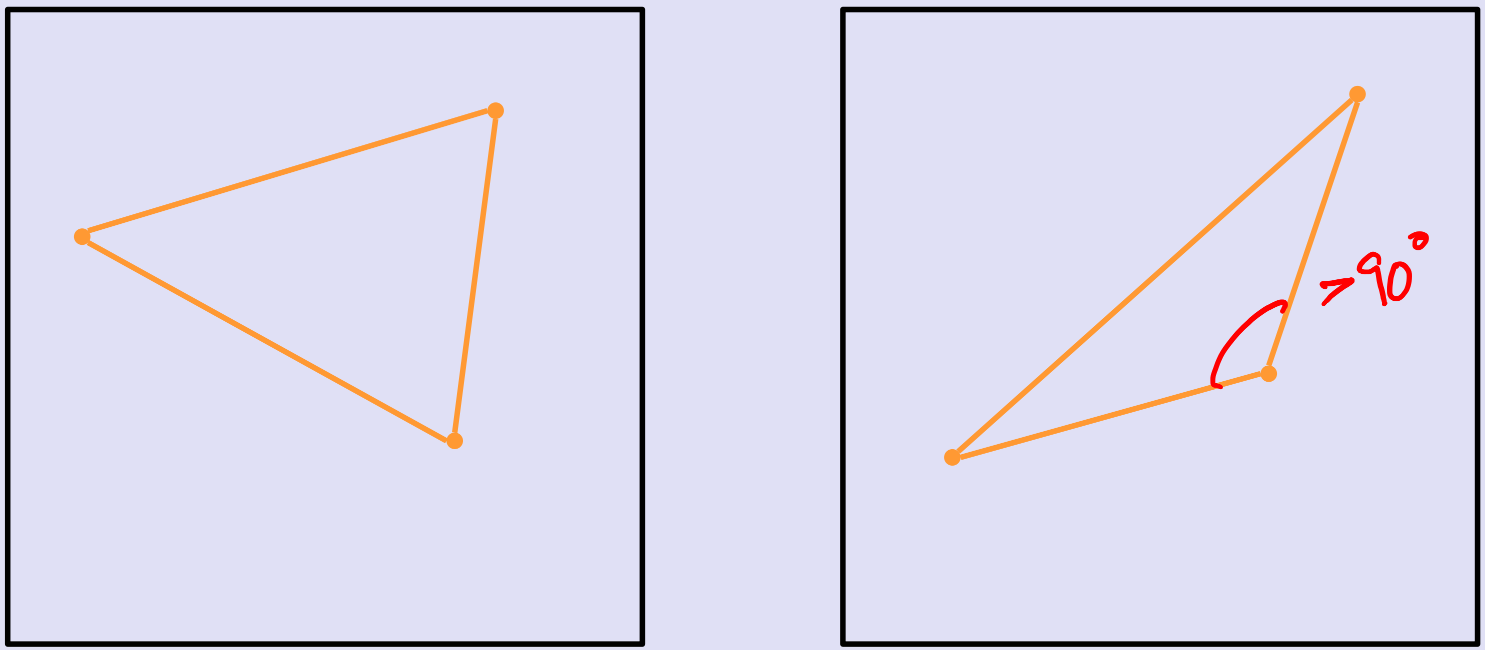
$t\frac{\sqrt{n}}{2} = 5\% \cdot n$

$\sqrt{n} = \frac{t}{2 \cdot 0.05}$

$n = \left(\frac{t}{2 \cdot 0.05}\right)^2 = \left(\frac{2.576}{2 \cdot 0.05}\right)^2 \approx 515$

NORMAL APPROXIMATION OF PMF OF X

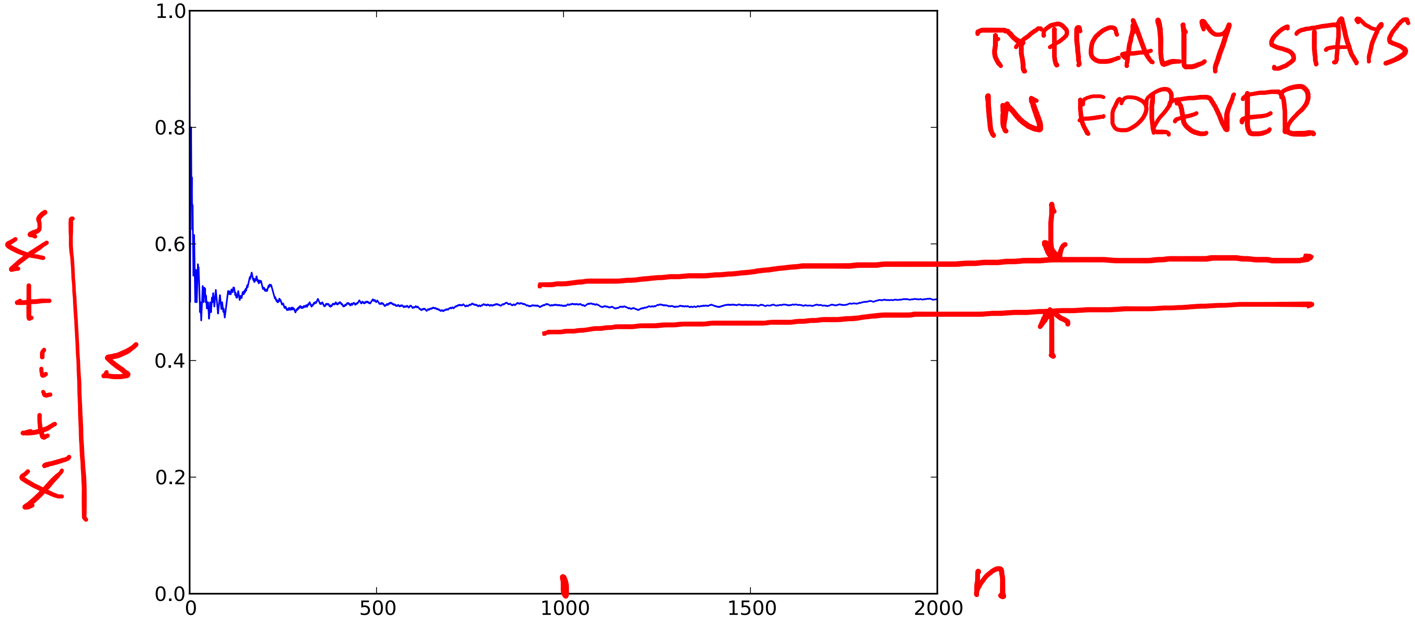
Drop three points at random on a square. What is the probability that they form an acute triangle?



$\varepsilon = 5\%$ $\delta = 1\%$ ENOUGH TO PICK 515

method	requirements	weakness
Markov's inequality	$\mathbf{E}[X]$ only	one-sided, often imprecise
Chebyshev's inequality	$\mathbf{E}[X]$ and $\mathbf{Var}[X]$	often imprecise
weak law of large numbers	pairwise independence	often imprecise
central limit theorem	independence of many samples	no rigorous bound

The strong law of large numbers



$$P\left(0.55 \le (X_1 + \dots + X_n)/n \le 0.65 \text{ FOR ALL } n \ge 1000\right)$$

The strong law of large numbers

$X_1, \ldots, X_n$ are independent with same PMF / PDF

$\mu = E[X_i]$, $X = X_1 + \ldots + X_n$

If $\mathbf{E}[X_i^4]$ is finite then

$$\mathbf{P}(\lim_{n \to \infty} X/n = \mu) = 1$$