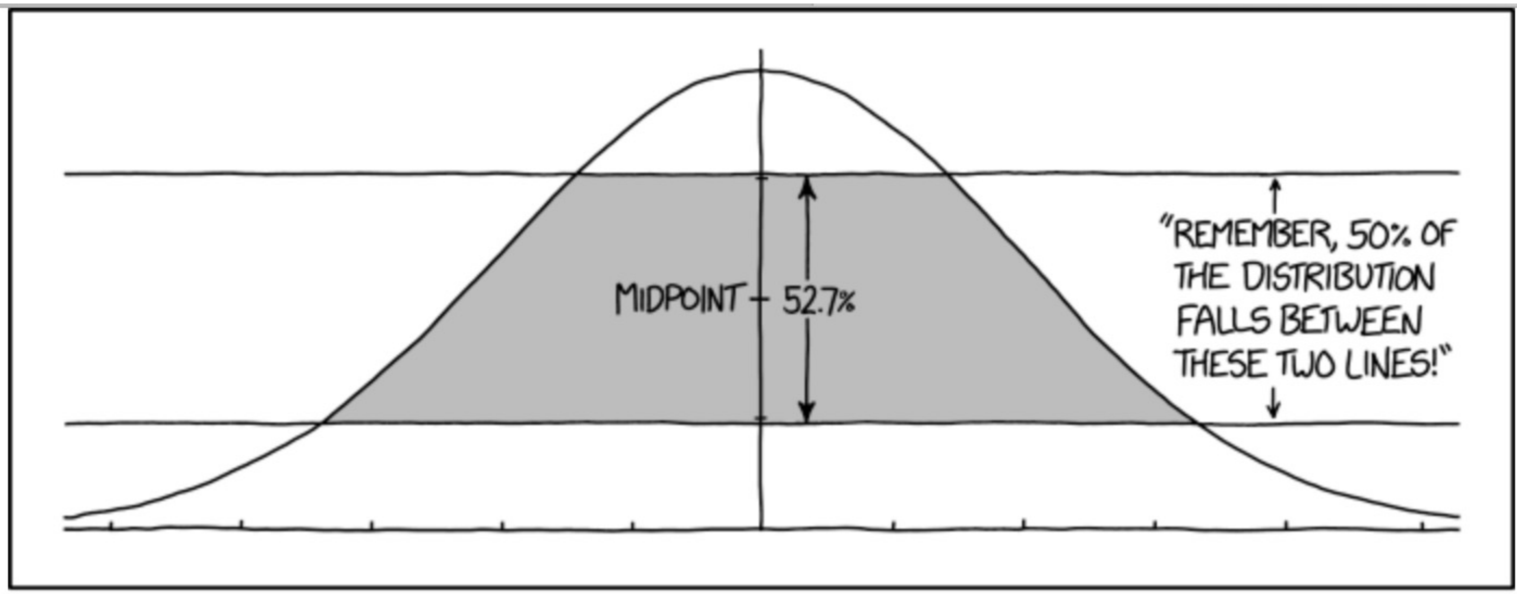
Stat 88: Prob. & Math. Statistics in Data Science



HOW TO ANNOY A STATISTICIAN

xkcd.com/2118

Lecture 29: 4/3/2024

The law of averages, distribution of a sample sum

7.3, 8.1, 8.2

Law of Averages

- Essentially a statement that you are already familiar with: If you toss a fair coin many times, roughly half the tosses will land heads.

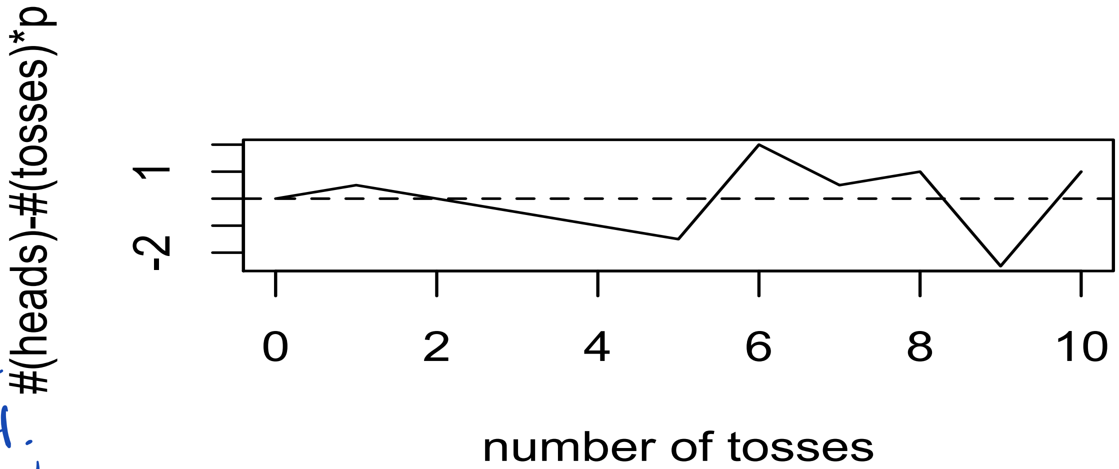
- We are going to consider sample sums and sample means of iid random variables $X_1, X_2, \dots, X_n$ where the mean of each X_k is μ and the variance of each X_k is σ^2.

- Recall the ***sample sum*** $S_n = X_1 + X_2 + \cdots + X_n$, with $E(S_n) = n\mu$, $Var(S_n) = n\sigma^2, SD(X_n) = \sqrt{n}\sigma$

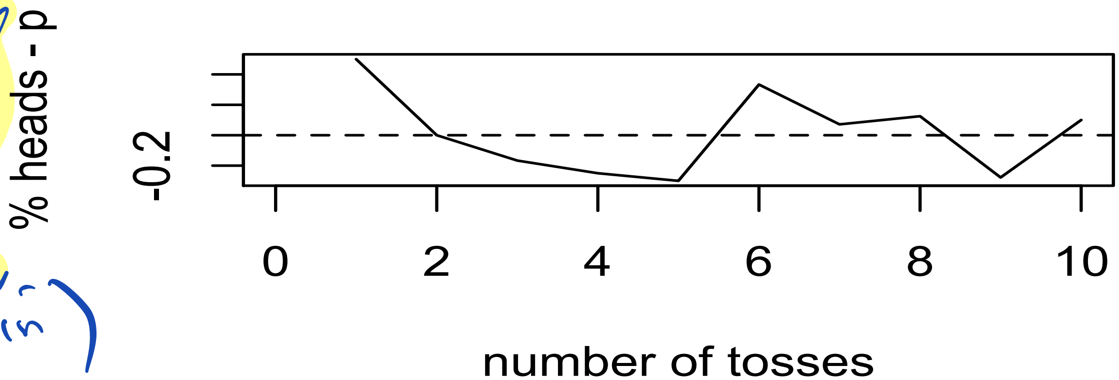
- We see here, as we take more and more draws, the variability of the sum keeps increasing, which means the values get more and more dispersed around the mean ($n\mu$).

Simulating coin tosses: 10 tosses (adapted from FPP)

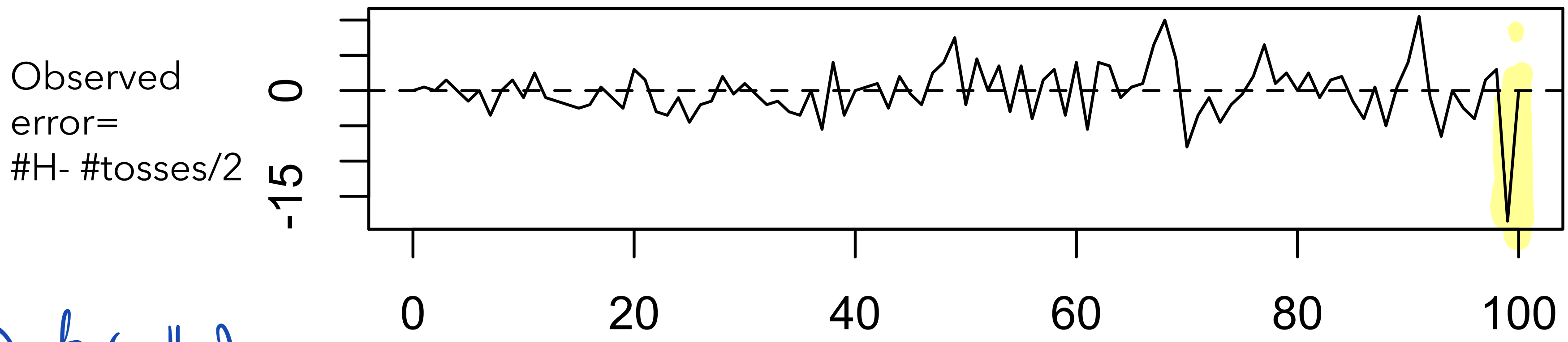
observed error = obs # of Heads − $\underbrace{\text{exp. \# of H.}}_{= \frac{n}{2}}$



% heads − 0.5 (for a fair coin)

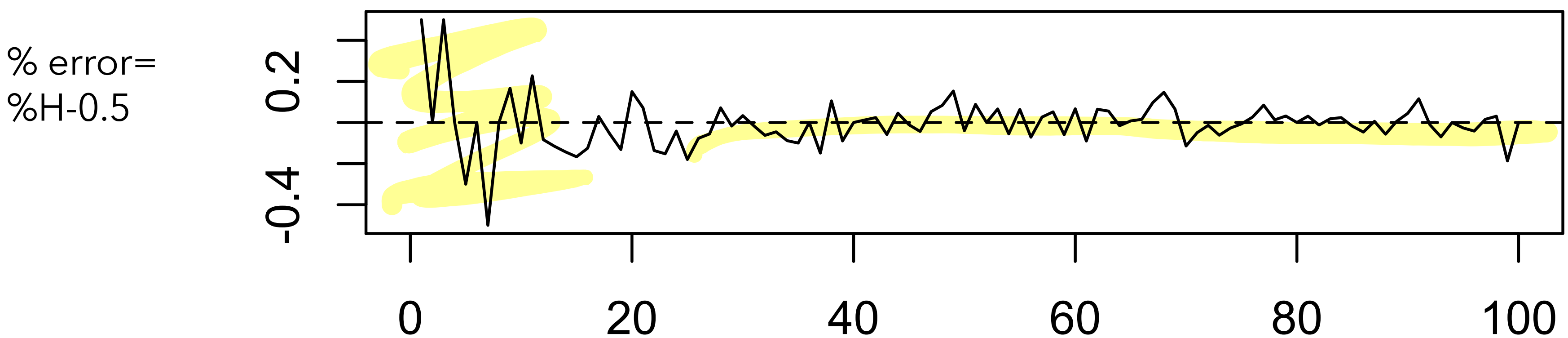


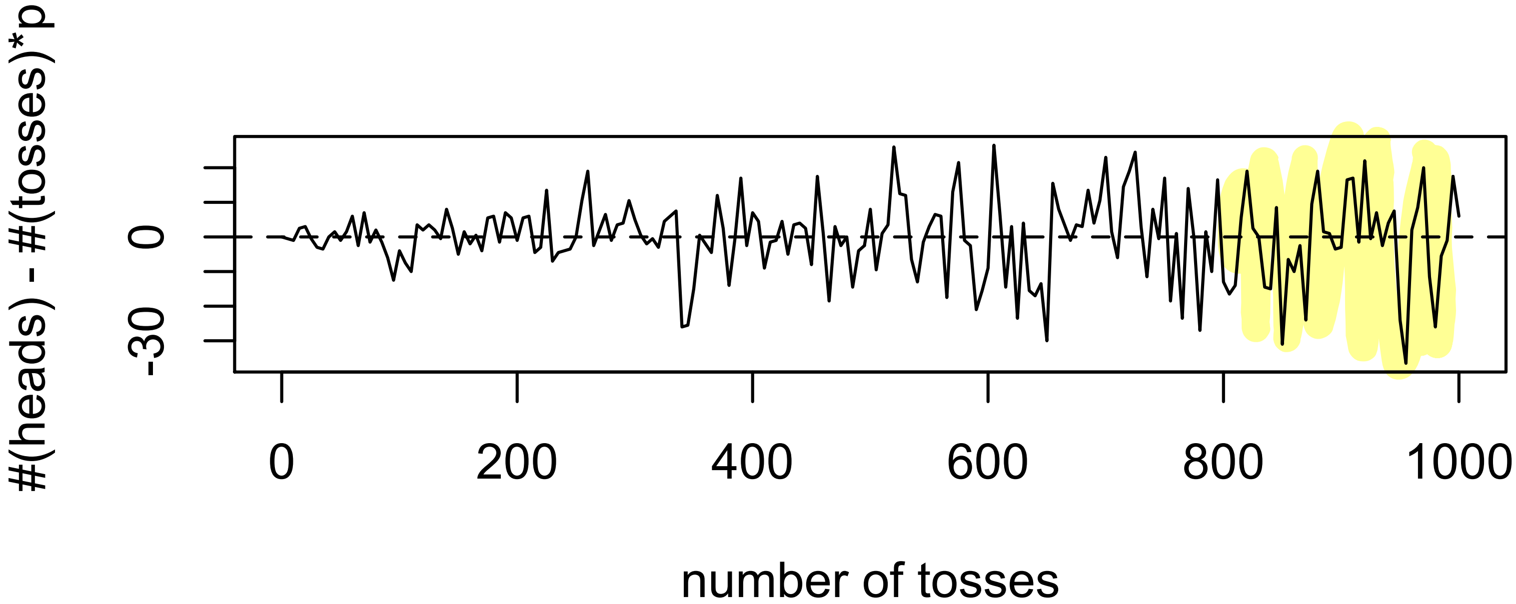
\# of H $= X \sim \text{Bin}(n, p)$ X is a sum of Bernoullis

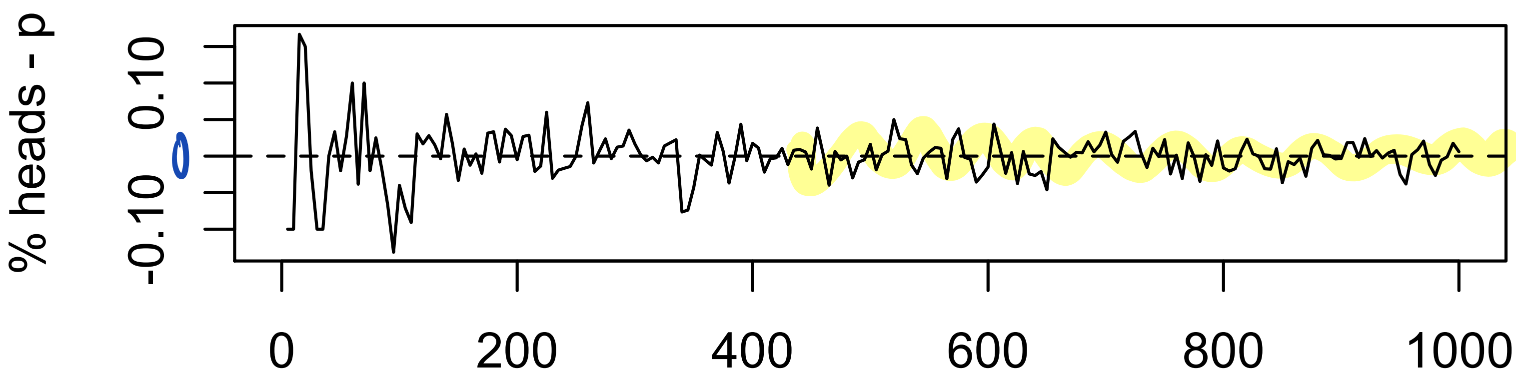


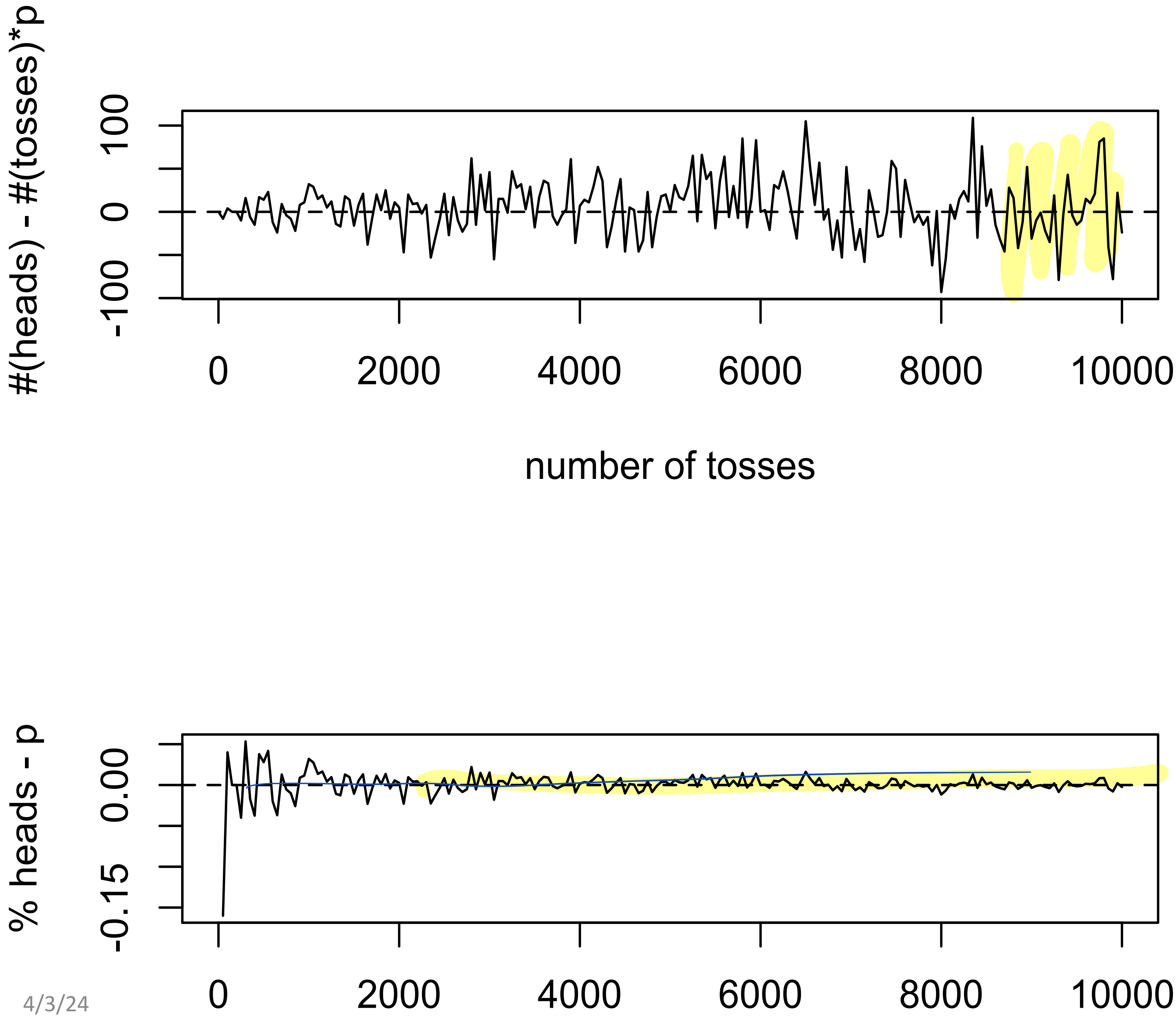
SD for # of Heads $= \sqrt{n}\sqrt{p(1-p)}$

% Heads $= \frac{X}{n}$ $SD\left(\frac{X}{n}\right) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$




#(heads) - #(tosses)*p
0
-30
0
200
400
600
800
1000
number of tosses


% heads - p
0.10
0
-0.10
0
200
400
600
800
1000



#(heads) - #(tosses)*p
100
0
-100
0
2000
4000
6000
8000
10000
number of tosses
% heads - p
0.00
-0.15
0
2000
4000
6000
8000
10000

Law of Averages for a fair coin

- Notice that as the number of tosses of a fair coin increases, the *observed error* (number of heads – half the number of tosses) increases. This is governed by the standard error.

$$\% \text{ error} = \% \text{ Heads} - 0.5 \longrightarrow 0$$

- The *percentage* of heads observed comes very close to 50%

- *Law of averages*: The long run *proportion* of heads is very close to 50%.

$$\frac{\# \text{ of } H}{n} = \frac{X}{n}$$

$S_{100} \sim Bin(100, \frac{1}{2})$ $S_{400} \sim Bin(400, \frac{1}{2})$

$\#\text{ of } H \sim Bin(n, p)$

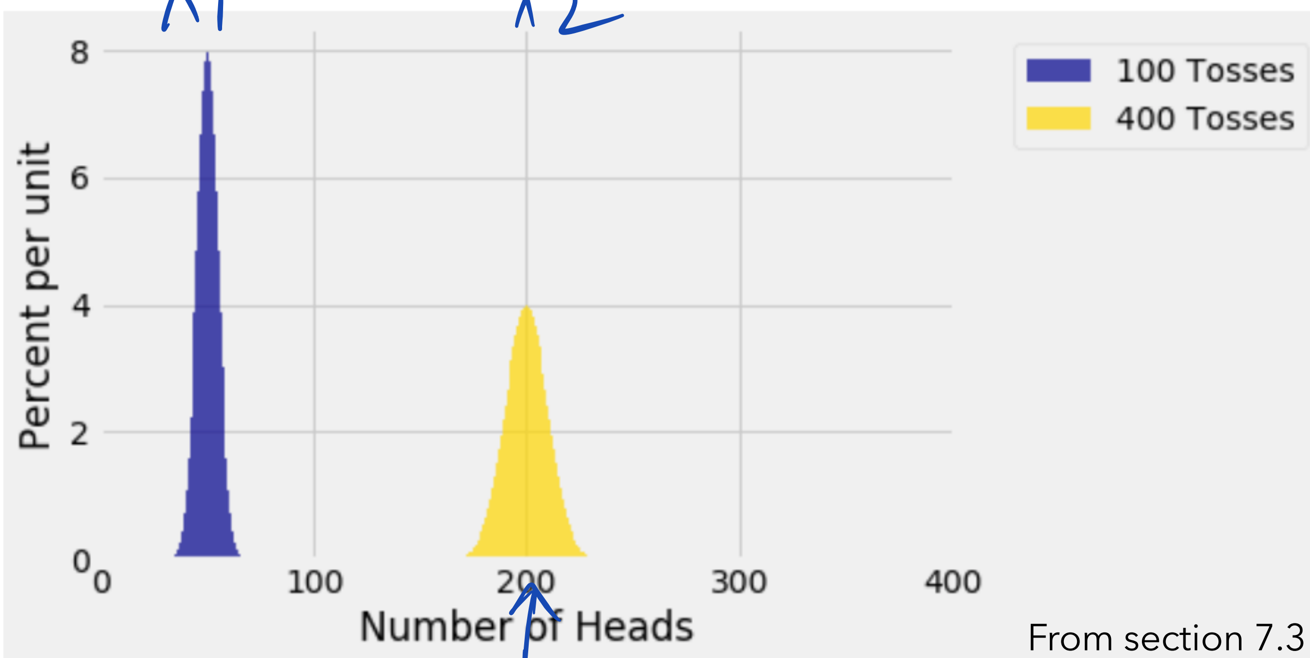
Coin tosses

- Consider a fair coin, toss it 100 times & 400 times, count the number of H
Expect in first case, roughly 50 H, and in second, roughly 200 H.

- So do you think chance of 50 H in 100 tosses and 200 H in 400 tosses should be the same?

Total area = 1 for both graphs.

X_1 X_2



From section 7.3

spread is much greater

$S_{100} \sim Bin(100, \frac{1}{2})$ $S_{400} \sim Bin(400, \frac{1}{2})$

Example: Coin toss

- $SD(S_{100}) =$ $\sqrt{np(1-p)} = \sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 5$
- $SD(S_{400}) =$ $\sqrt{400 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 20 \cdot \frac{1}{2} = 10$

square root law.

- P(200 H in 400 tosses)

$P(S_{400} = 200) = \binom{400}{200}\left(\frac{1}{2}\right)^{200}\left(\frac{1}{2}\right)^{200}$

≈ 0.04

- P(50 H in 100 tosses)

$P(S_{100} = 50) = \binom{100}{50}\left(\frac{1}{2}\right)^{50}\left(\frac{1}{2}\right)^{50} \approx 0.08$

Sample sum, sample average, and the square root law

- $S_n = X_1 + X_2 + \cdots + X_n$
- Let $A_n = {}^{S_n}/_{n}$, so A_n is the average of the sample (or sample mean).
- If the X_k are indicators, then A_n is a proportion (proportion of successes)
  (Bernoulli)
- Note that $E(A_n) = \mu$ and $SD(A_n) = \cancel{??}$ $\frac{SD(X_k)}{\sqrt{n}}$
- **The square root law:** the *accuracy* of an estimator is measured by its SD, the ***smaller*** the SD, the ***more accurate*** the estimator, but if you multiply the sample size by a factor, the accuracy only goes up by the square root of the factor.
- In our earlier example, we _____ the accuracy by quadrupling the size.

We doubled SD so halved the accuracy.

$SD(X_k) = \sigma$, $E(X_k) = \mu$ $SD(S_n) = \sqrt{n}\,\sigma$

$X_1, X_2, \ldots X_n$ $S_n = X_1 + X_2 + \cdots + X_n$, $A_n = \frac{S_n}{n}$, $SD(A_n) = SD(\bar{X}) = \frac{SD(X_k)}{\sqrt{n}} = \boxed{\frac{\sigma}{\sqrt{n}}}$

as $n \to \infty$

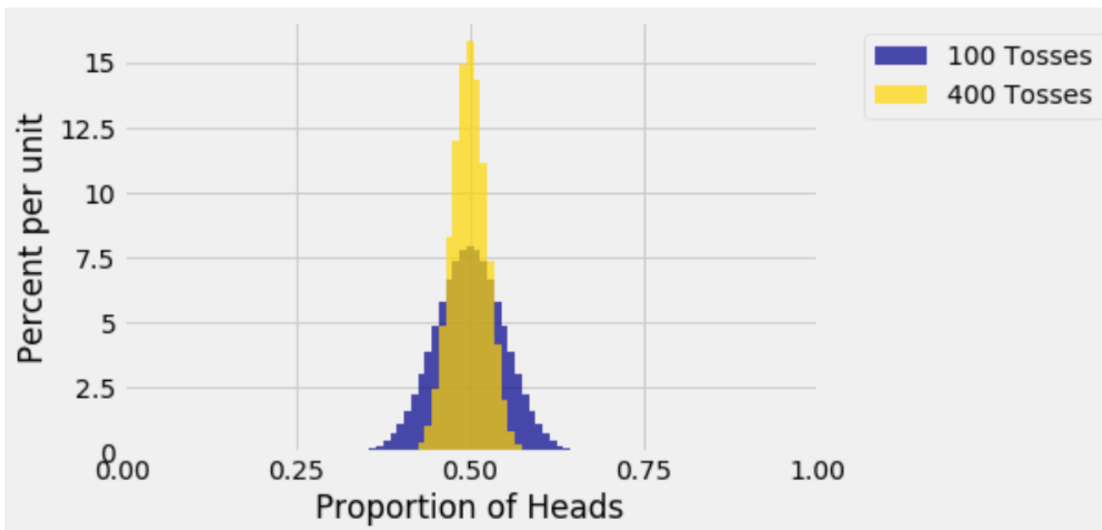
Concentration of probability

- This is when the SD decreases, so the probability mass accumulates around the mean, therefore, the larger the sample size, the more likely the values of the sample average $\bar{X}$ fall very close to the mean.
- **Weak Law of Large numbers:**

distance b/w the sample mean & exp. value

$$\boxed{\text{For } c > 0, P(|\, A_n - \mu \,| < c) \to 1 \text{ as } n \to \infty}$$

$|\, A_n - \mu \,|$ is the distance between the sample mean and its expectation.



From section 7.3

For any $c > 0$, HOWEVER TINY, as long as n is large enough, the chance that A_n is VERY CLOSE to μ is very high.

Law of averages

- The law of averages says that if you take enough samples, the proportion of times a particular event occurs is very close to its probability.
- In general, when we repeat a random experiment such as tossing a coin or rolling a die over and over again, the average of the observed values will come the expected value.

- The *percentage* of sixes, when rolling a fair die over and over, is very close to 1/6. True for any of the faces, so the *empirical* histogram of the results of rolling a die over and over again looks more and more like the *theoretical* probability histogram.

- *Law of averages*: The individual outcomes when averaged get very close to the theoretical weighted average aka expected value

Moving On

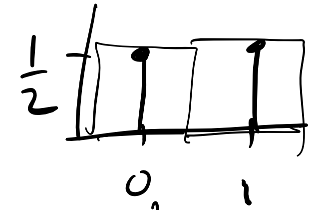
8.1: Distribution of a sample sum

$S_n = X_1 + X_2 + \cdots + X_n$

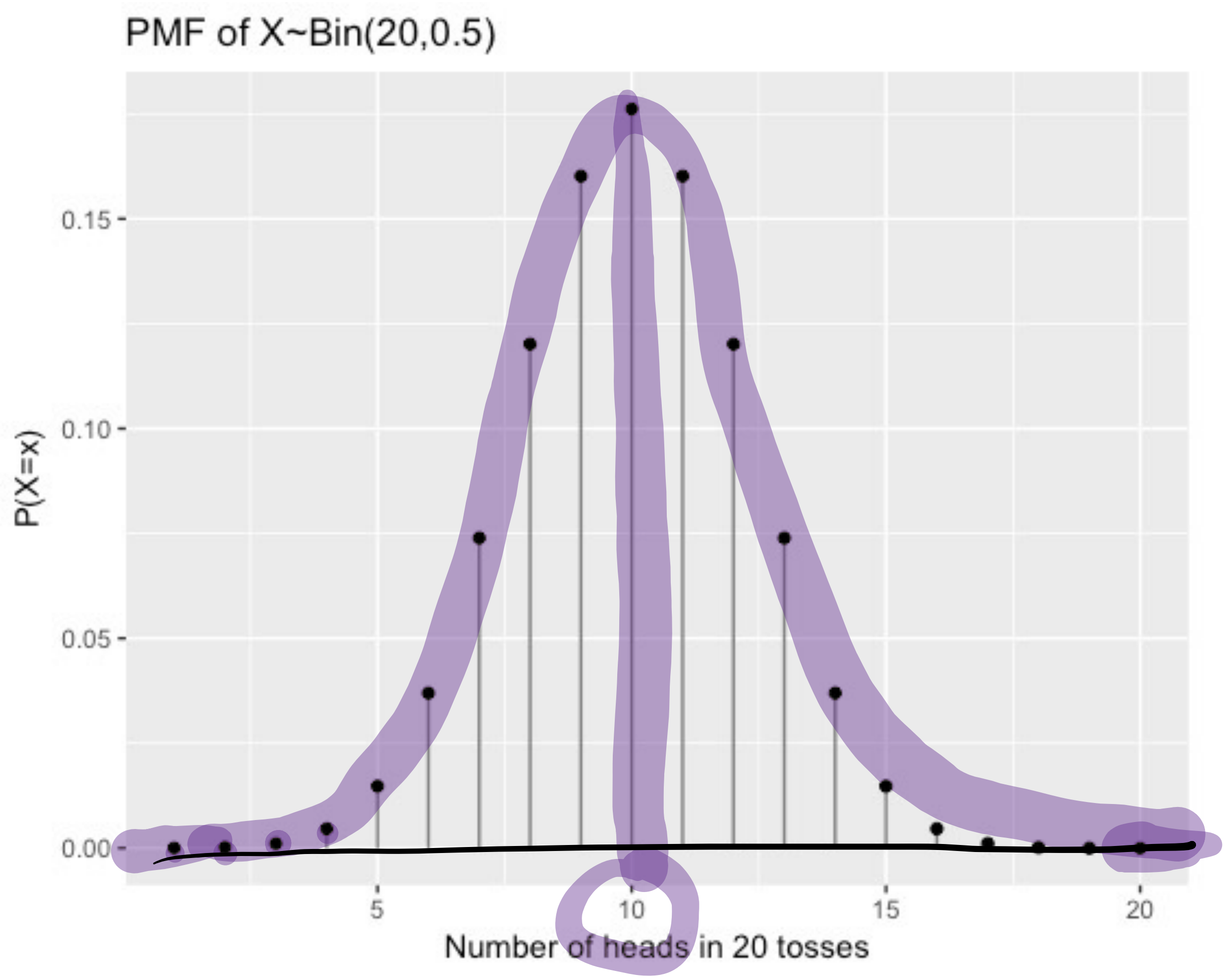
$X \sim \text{Bernoulli}(\frac{1}{2})$

- We can consider $X \sim Bin(20, 0.5)$ as the sum of 20 Bernoulli iid rvs.

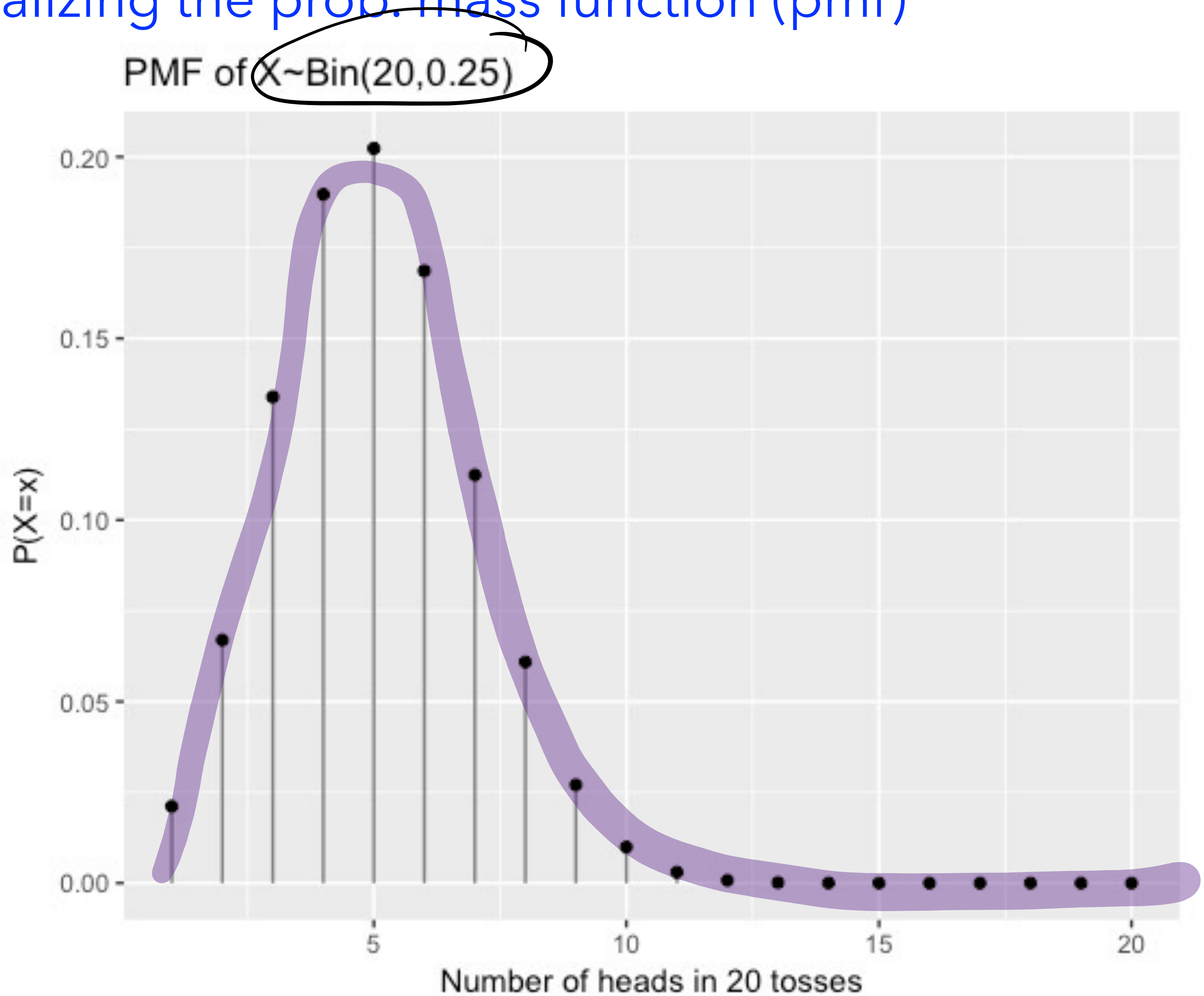
Visualizing the prob. mass function (pmf) of the binomial below:



single coin toss



Visualizing the prob. mass function (pmf)

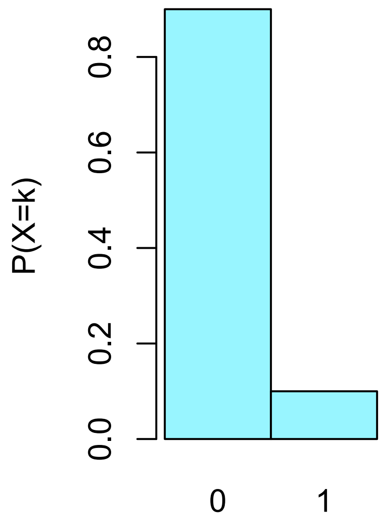


What if p is small?

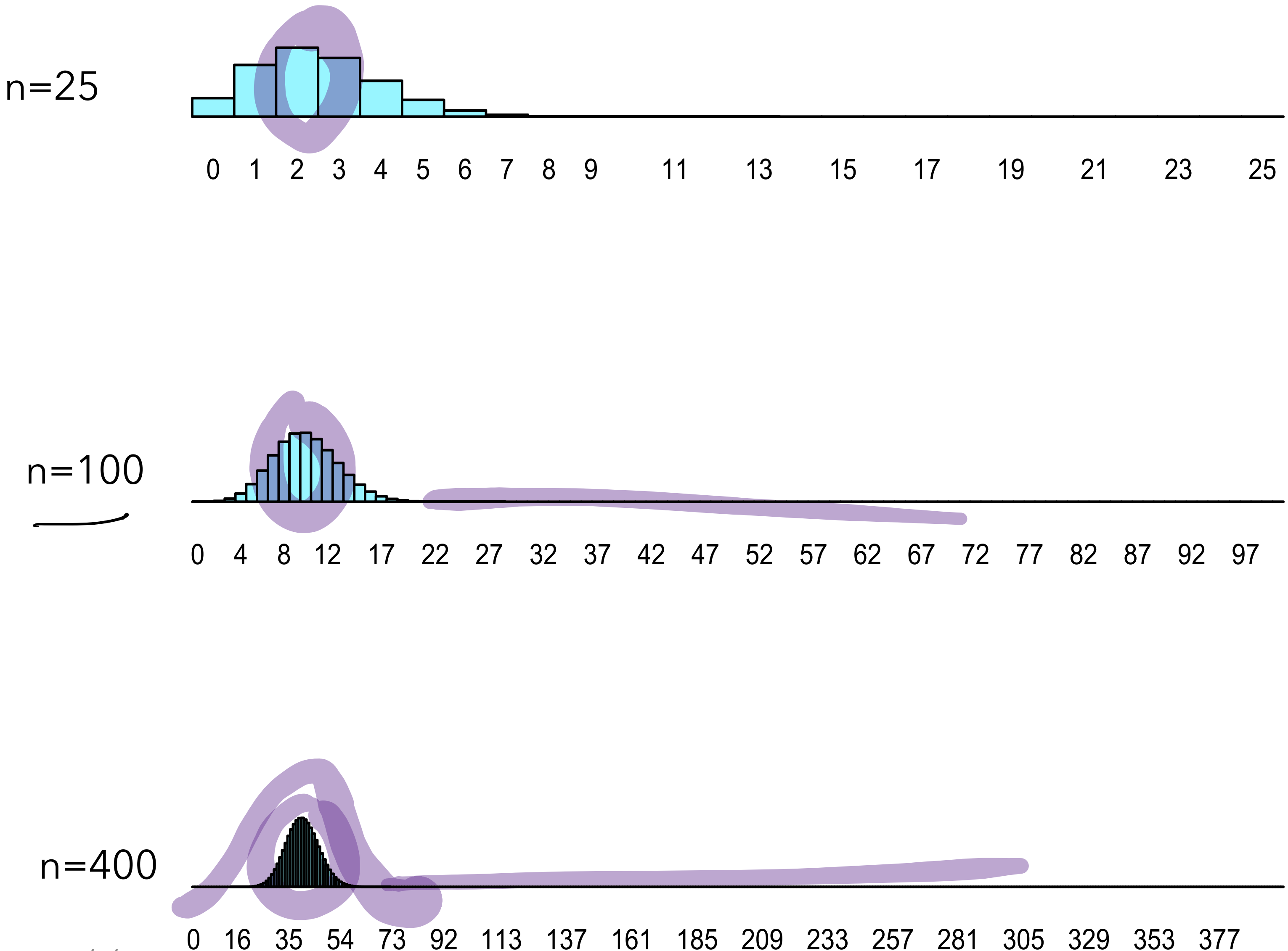
- Consider $X_k \sim Bernoulli\left(\frac{1}{10}\right), S_n = X_1 + X_2 + X_3 + \cdots + X_n, Sn \sim Bin(n, \frac{1}{10})$

- Draw the probability histogram for X_k:

$$X_k = \begin{cases} 0 & \text{w.p. } \frac{9}{10} \\ 1 & \text{w.p. } \frac{1}{10} \end{cases}$$



When p is small (picture from *Statistics* by Freedman, Pisani, and Purves)



Distribution of the sample sum

- More generally, let's consider $X_1, X_2, \dots, X_n$ iid with mean μ and SD σ
- Let $S_n = X_1 + X_2 + \cdots + X_n$
- We know that $E(S_n) = n\mu$ and $SD(S_n) = \sqrt{n}\sigma$

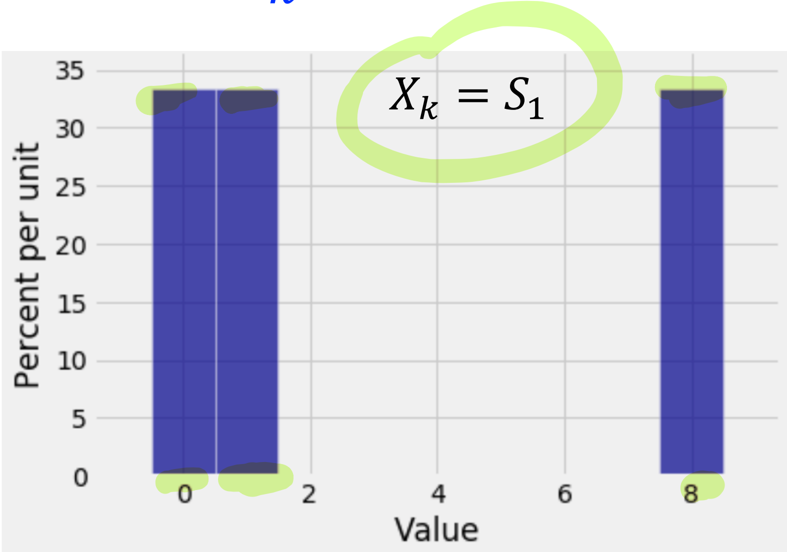
probabilities of sample sum.

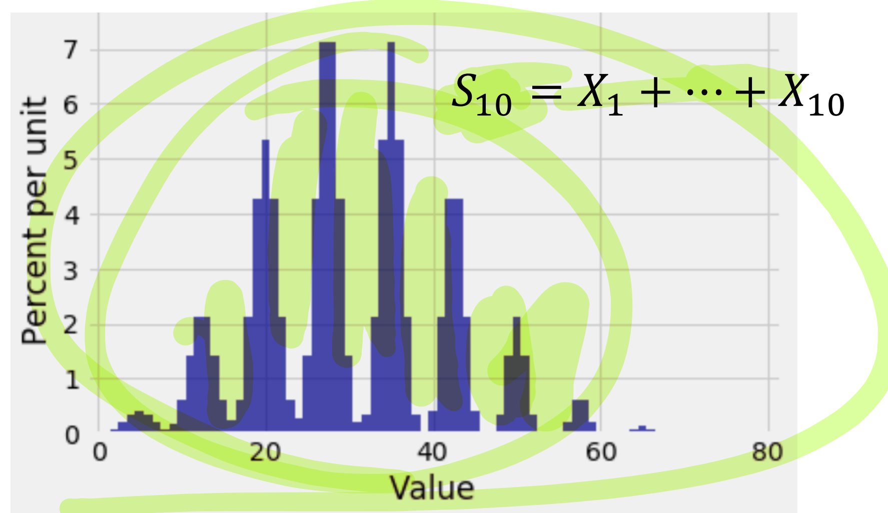
- We want to say something about the distribution of S_n, and while it may be possible to write it out analytically, if we know the distributions of the X_k, it may not be easy. And we may not even know anything beyond the fact that the X_k are iid, and we might be able to guess at their mean and SD.

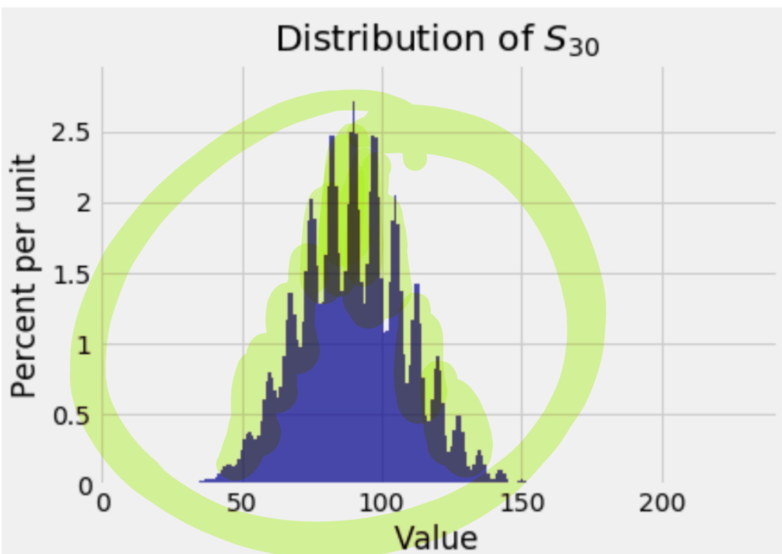
- We saw in the previous slides that even if the X_k are very far from symmetric, the distribution of the sum begins to look quite nice and bell shaped.

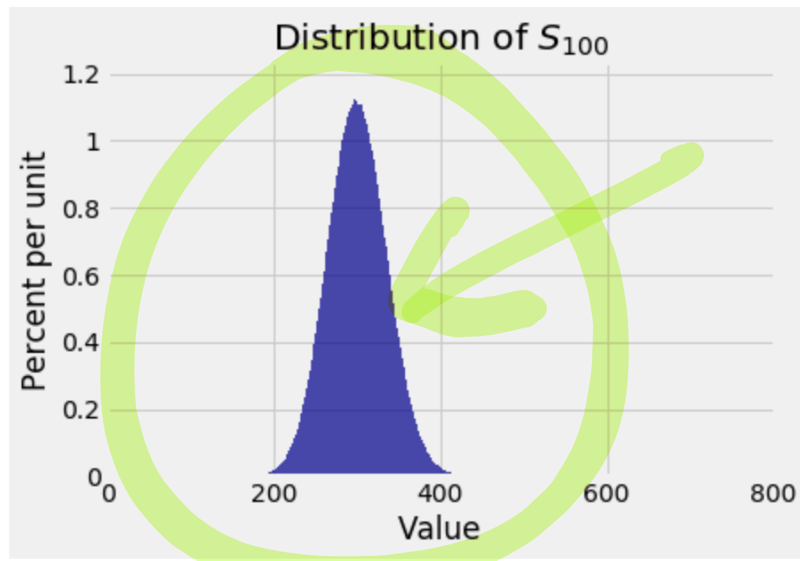
- What if the X_k are strange looking?

Weird X_k distributions – is the distribution of S_n different?





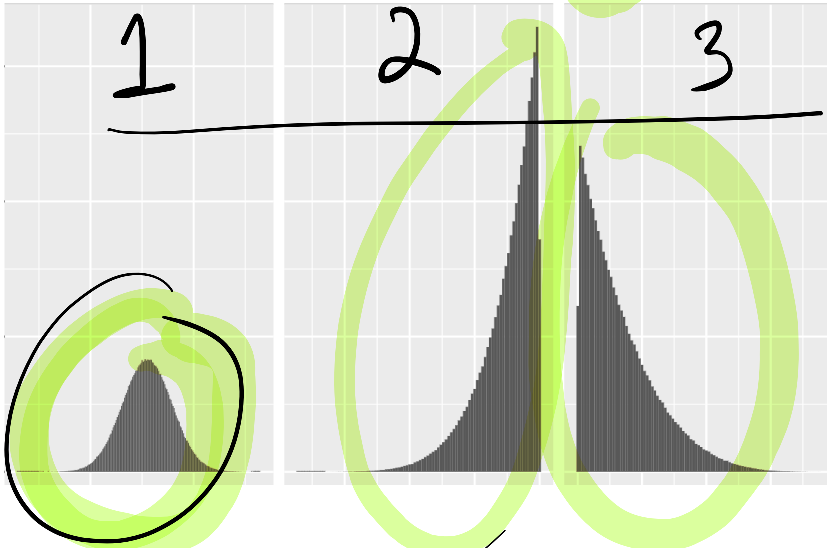




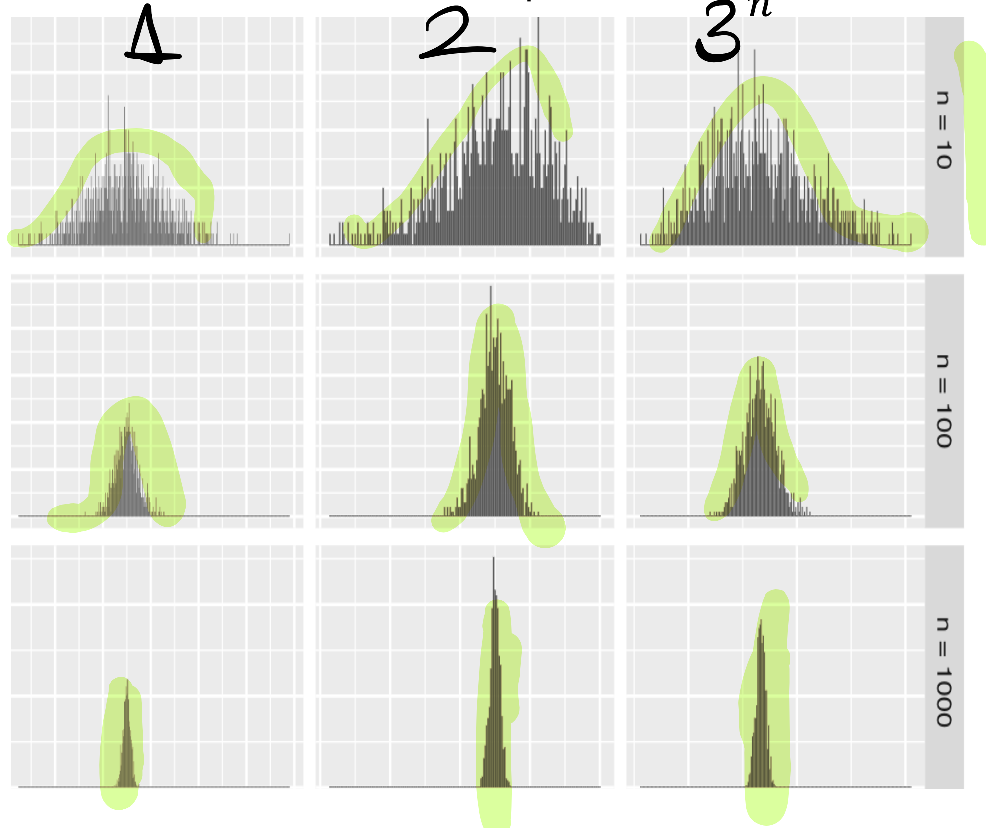
From section 8.1

Examples by picture

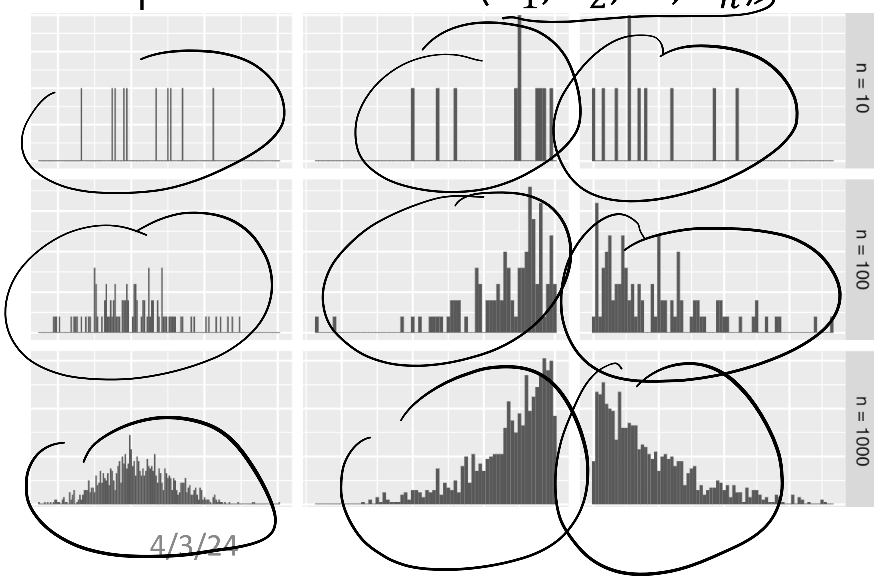
Probability distribution of X_k



Distribution of the sample mean $\frac{S_n}{n} = A_n$



Sample distribution $(X_1, X_2, \ldots, X_n)$



Graphs created by Sarah Johnson

The Central Limit Theorem

- The bell-shaped distribution is called a *normal curve*.

- What we saw was an illustration of the fact that if $X_1, X_2, \dots, X_n$ iid with mean μ and SD σ, and $S_n = X_1 + X_2 + \cdots + X_n$, then the distribution of S_n is approximately normal for large enough n.

- The distribution is approximately normal (bell-shaped) centered at $E(S_n) = n\mu$ and the width of this curve is defined by $SD(S_n) = \sqrt{n}\,\sigma$

Bell curve: the Standard Normal Curve

- Bell shaped, symmetric about 0
- Points of inflection at $z = \pm 1$
- Total area under the curve = 1, so can think of curve as approximation to a probability histogram
- Domain: whole real line
- Always above x-axis
- Even though the curve is defined over the entire number line, it is pretty close to 0 for $|z|>3$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, -\infty < z < \infty$$