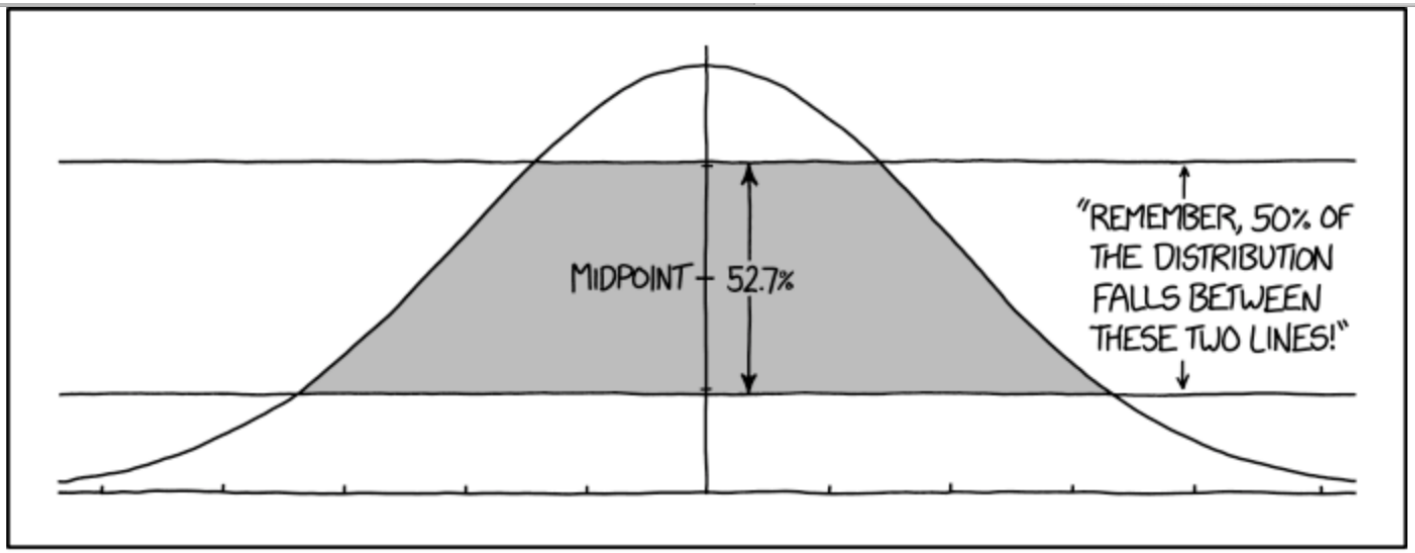
Stat 88: Probability & Mathematical Statistics in Data Science



HOW TO ANNOY A STATISTICIAN

xkcd.com/2118

Lecture 25: 3/19/2021

Sections 7.3

The Law of Averages & setting up Chapter 8

Warm up

Draw 6 cards and count the number of red cards you get. To increase your odds of getting 3 red cards should you draw with or without replacement?

X = # of red cards in draw of 6 cards. (26 R cards in 52 cards)

$N = 52$, $n = 6$

F.P.C. $= \sqrt{\frac{N-n}{N-1}} \leq 1$

$= \sqrt{\frac{52-6}{52-1}} \approx 0.95$

$SD(HG) = SD(Bin) \cdot FPC < SD(Bin)$

w/ repl. $X \sim Bin\left(6, \frac{1}{2}\right)$, $P(X=3) = \binom{6}{3}\left(\frac{1}{2}\right)^6 \approx 0.313$

w/o repl. $X \sim HG(N=52, G=26, n=6)$

$P(X=3) = \frac{\binom{26}{3}\binom{26}{3}}{\binom{52}{6}} \approx 0.332$ } variability less

Draw 8 cards

0 0 0 0 0 1 1 1

draw w/ repl

X = # of 1

$0 \leq X \leq 8$

draw w/o repl

$X = 3$

Example (from *Statistics*, by Freedman, Pisani, and Purves)

A survey organization wants to take an SRS in order to estimate the percentage of people who watched the 2021 Grammys. To keep costs down, they want to take as small a sample as possible, but their client will only tolerate a random error of 1 percentage point or so in the estimate. Should they use a sample size of 100, 2500, or 10000? The population is very large and the fpc is about 1.

Let $X_1, X_2, \ldots X_n$ be Bernoulli(p), p is the prob we are trying to estimate (% of popn that watched the Grammys)

$S_n = X_1 + X_2 + \cdots + X_n$, $S_n \sim Bin(n, p)$

$A_n = \frac{S_n}{n}$ ← Sample mean

If we say $Var(X_k) = \sigma^2$
$Var(A_n) = \sigma^2/n$
$SD(A_n) = \sigma/\sqrt{n}$

$Var(S_n) = n \cdot Var(X_k) = npq$

$SD(S_n) = \sqrt{npq}$

$Var(A_n) = Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2} Var(S_n) = \frac{1}{n^2} \cdot npq = \frac{pq}{n}$

$SD(A_n) = \frac{\sqrt{pq}}{\sqrt{n}}$

When S_n is sum of 0-1 r.v., A_n is a proportion

Last time saw that $pq \leq 0.25$

$\sqrt{pq} \leq 0.5$

Client specification

$$SD(\hat{p}_n) = \frac{\sqrt{pq}}{\sqrt{n}} \leq \frac{0.5}{\sqrt{n}} \leq 0.01$$

$$\frac{0.5}{0.01} \leq \sqrt{n} \quad \Rightarrow \quad n \geq 2500$$

So they should use a sample size of 2,500.

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 $X_1, X_2, \ldots, X_n$ where the mean of each X_k is μ and the variance of each X_k is σ^2.

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

$Var(A_n) = \frac{\sigma^2}{n}$, $SD(A_n) = \sigma/\sqrt{n}$

$SD(S_1) = \sigma$
$SD(S_2) = \sqrt{2}\sigma$
$SD(S_5) = \sqrt{5}\sigma$

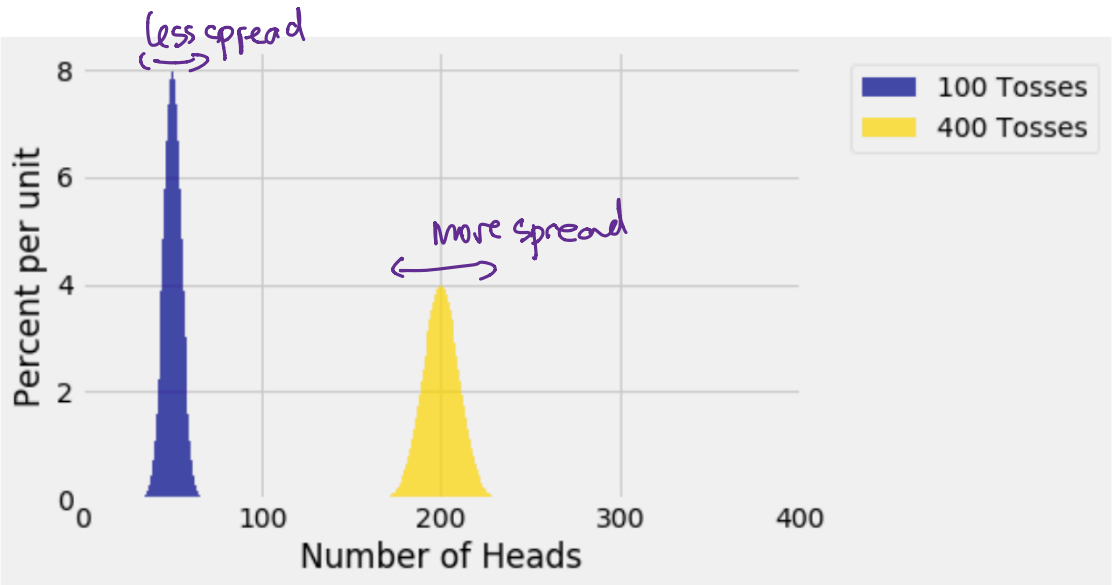
- We see here, as we take more and more draws, their sum's variability keeps increasing, which means the values get more and more dispersed around the mean ($n\mu$).

$SD(A_1) = \sigma$
$SD(A_2) = SD\left(\frac{X_1+X_2}{2}\right) = \frac{\sigma}{\sqrt{2}}$

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 in both cases. So in case of 400 tosses total prob of 1 spread b/w many more possible outcomes.

Example: Coin toss

$Var(X_k) = pq = \frac{1}{4} = \sigma^2$

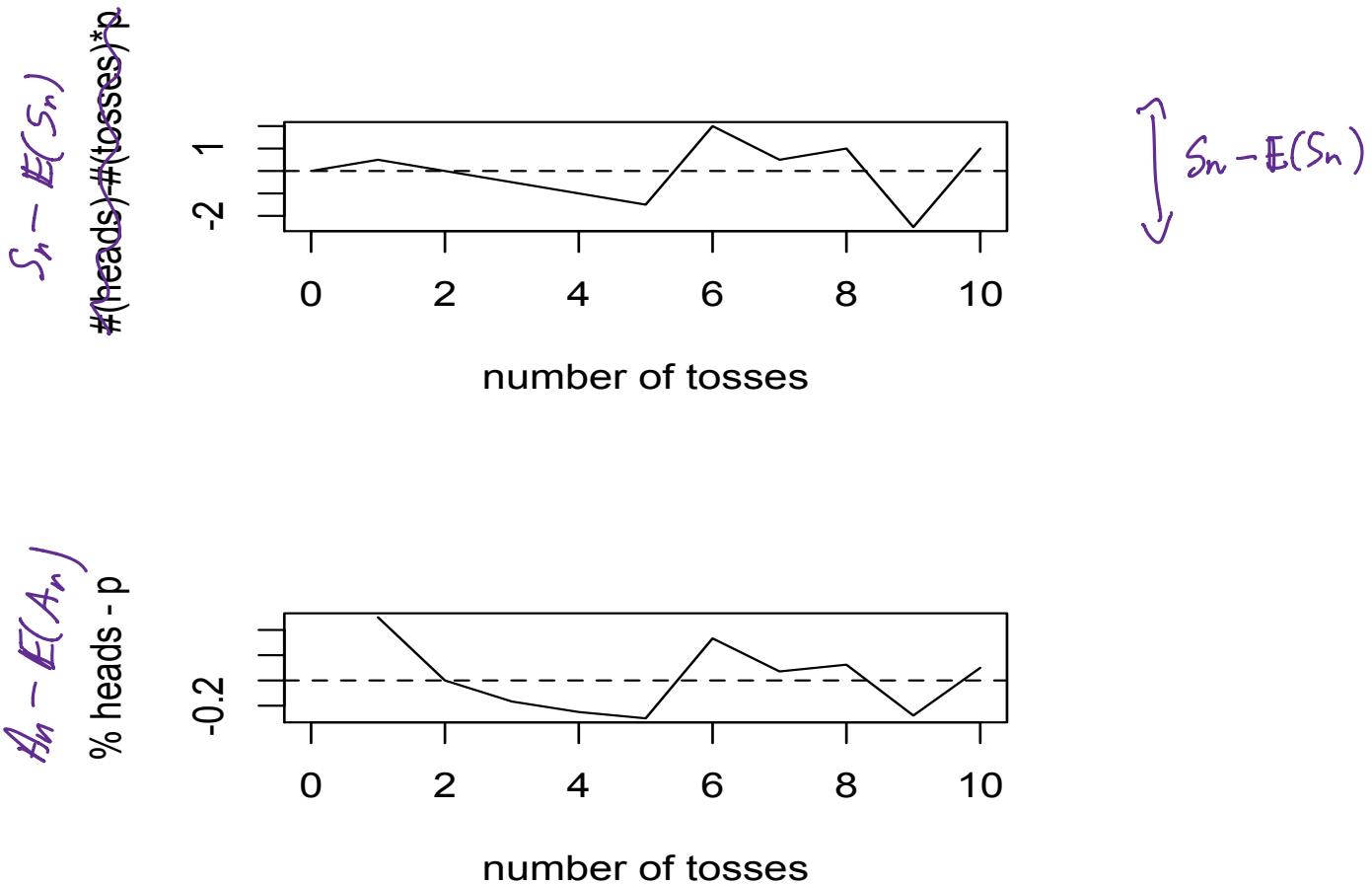
(a coin tossing is like drawing with replacement from [0] [1])

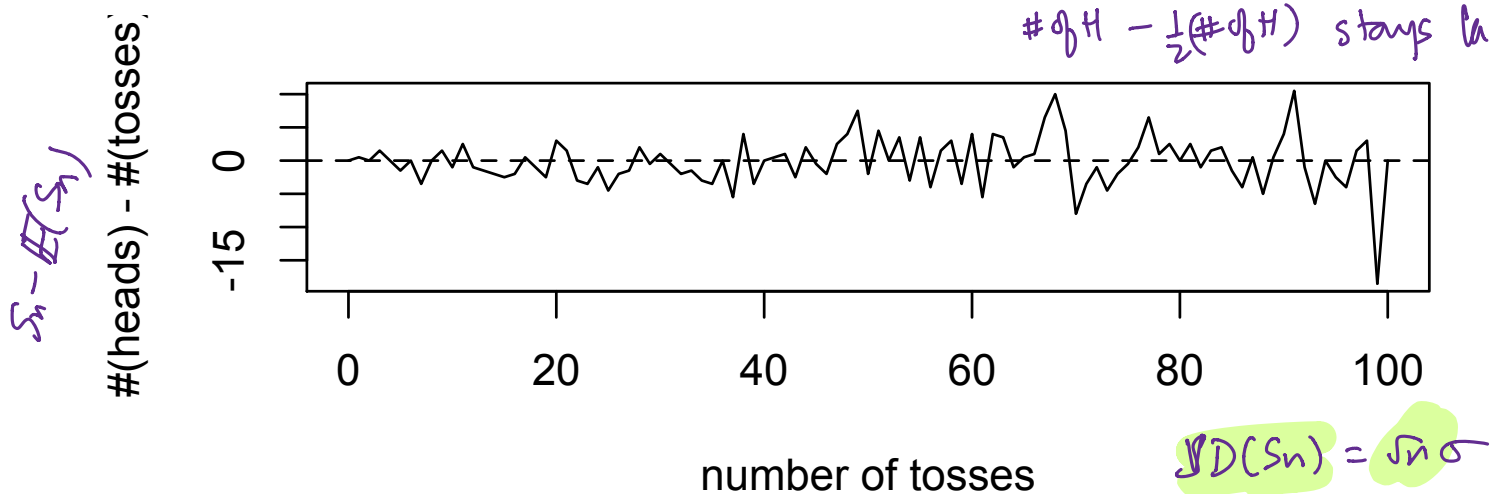
- $SD(S_{100}) = \sqrt{100} \cdot \frac{1}{2} = 10 \cdot \frac{1}{2} = 5$
- $SD(S_{400}) = \sqrt{400} \cdot \frac{1}{2} = 20 \cdot \frac{1}{2} = 10$

$SD(S_{100}) < SD(S_{400})$

- P(200 H in 400 tosses) $= P(S_{400} = 200) = \binom{400}{200}\left(\frac{1}{2}\right)^{200}\left(\frac{1}{2}\right)^{200} \approx 0.04$
- P(50 H in 100 tosses) $= P(S_{100} = 50) = \binom{100}{50}\left(\frac{1}{2}\right)^{50}\left(\frac{1}{2}\right)^{50} \approx 0.08$

Simulating coin tosses: 10 tosses

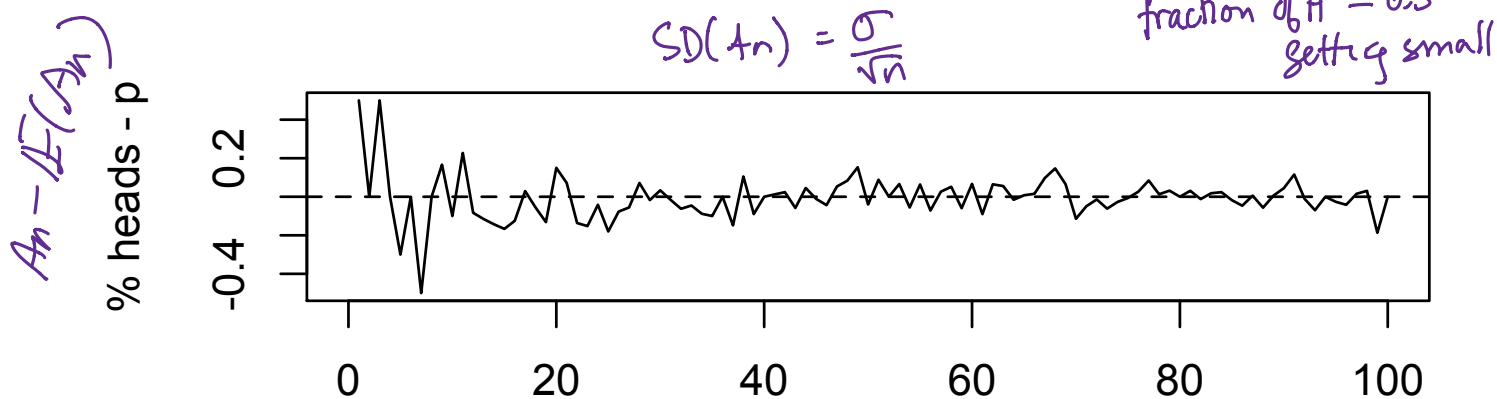




$S_n - E(S_n)$

# of H − ½(# of H) stays large.

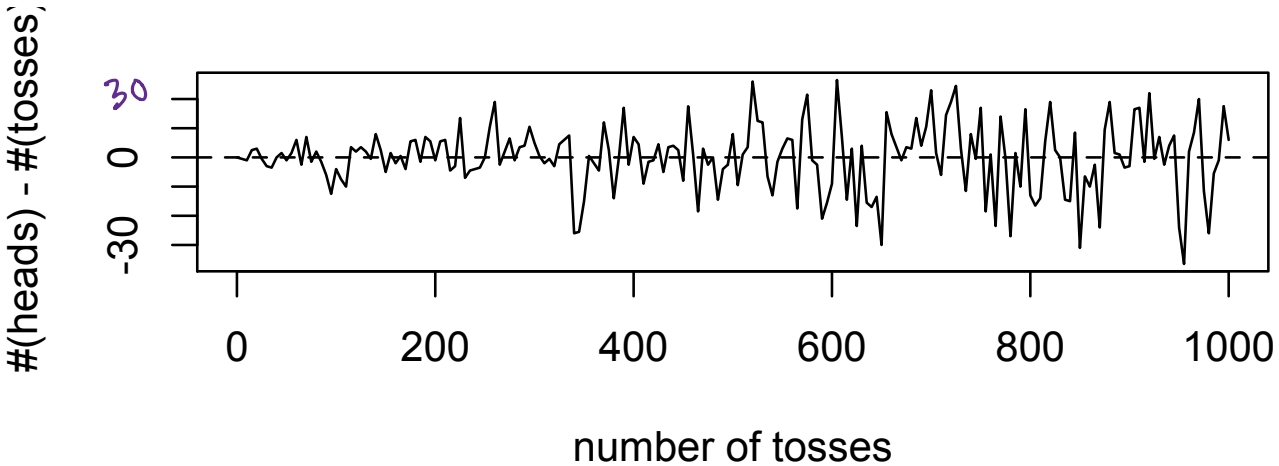
$SD(S_n) = \sqrt{n}\,\sigma$

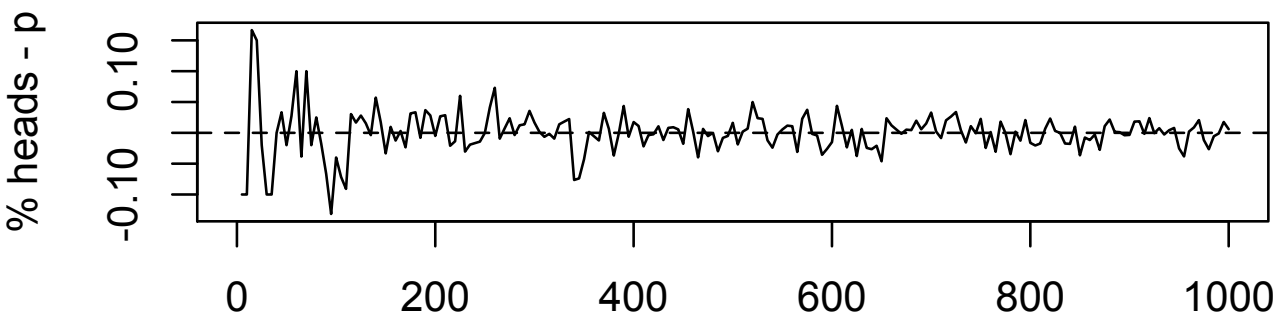


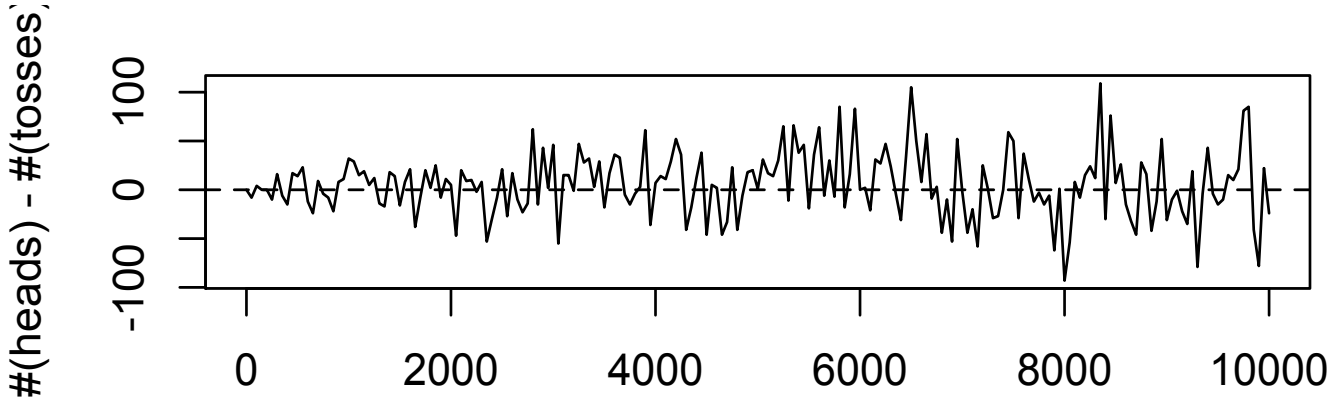
$A_n - E(A_n)$

$SD(A_n) = \frac{\sigma}{\sqrt{n}}$

fraction of H − 0.5 getting small

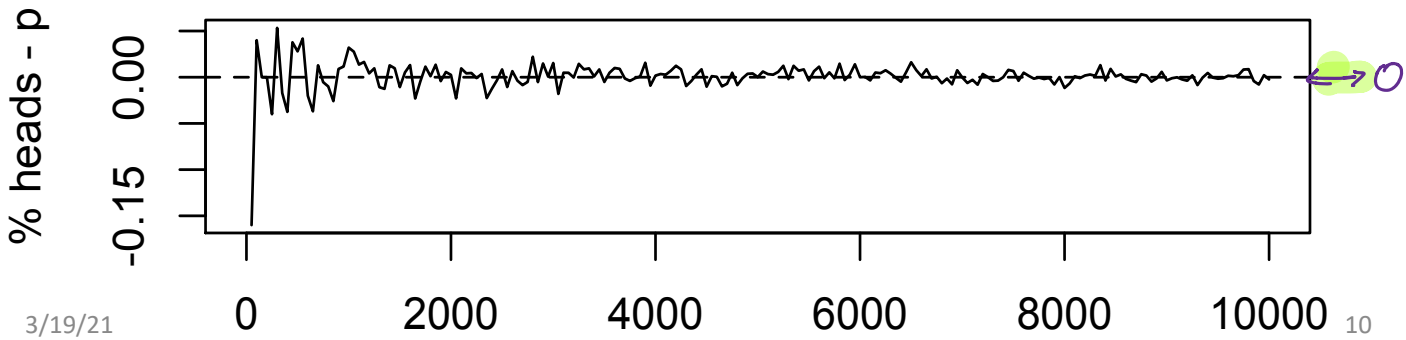

#(heads) - #(tosses)
30
0
-30
0
200
400
600
800
1000
number of tosses


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



$SD(S_n) = \sqrt{n}\,\sigma$ grows with n

$SD(A_n) = \frac{\sigma}{\sqrt{n}}$ gets smaller



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.

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

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

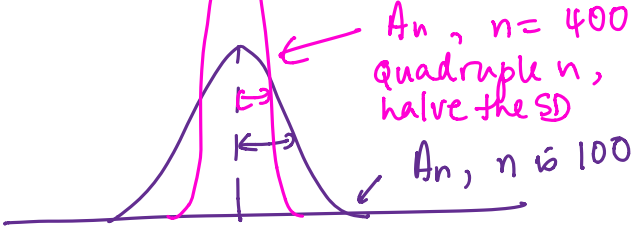
Sample sum, sample average, and the square root law

- $S_n = X_1 + X_2 + \cdots + X_n$
- 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)
- Note that $E(A_n) = \mu$ and $SD(A_n) = ??$ $\frac{\sigma}{\sqrt{n}}$
- **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 double the accuracy by quadrupling the size.

SD reduces by $\frac{1}{2}$

$SD(A_n) = \frac{\sigma}{\sqrt{n}}$ ←

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:**

usually c is small

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

sample mean

$\exists \delta > 0$

$\forall\ \varepsilon > 0$, there exists N s.t. for all $n \geq N$

$P(|A_n - \mu| < \varepsilon) = 1 - \delta$

$|\ A_n - \mu\ |$ is the distance between the sample mean and its expectation. So when your sample size is large, then the chance that the sample mean is VERY CLOSE to its expected value is super high.

How close? As close as you like. Just take a large enough sample. BUT the chance that it is *exactly* equal to the expected value is tiny.

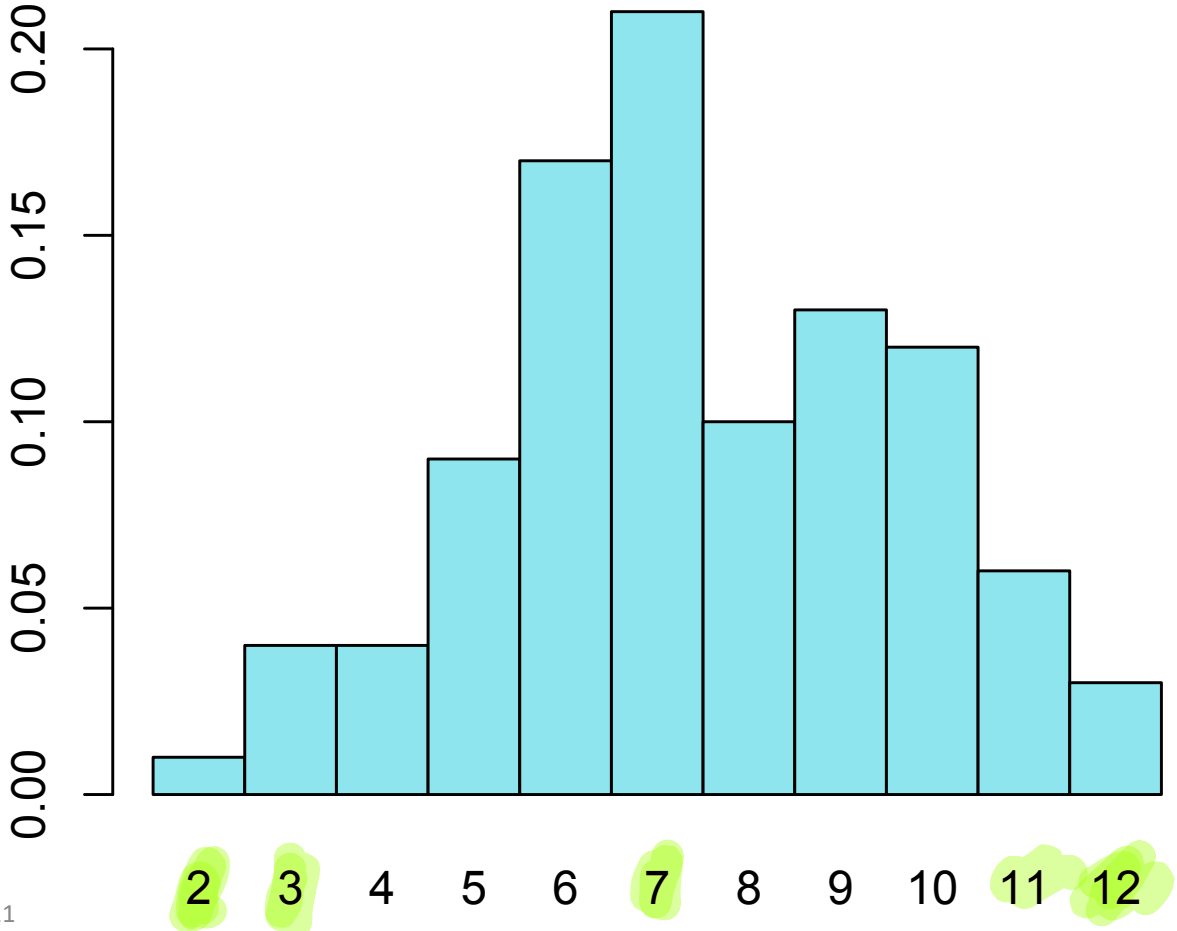
Law of averages

- 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.

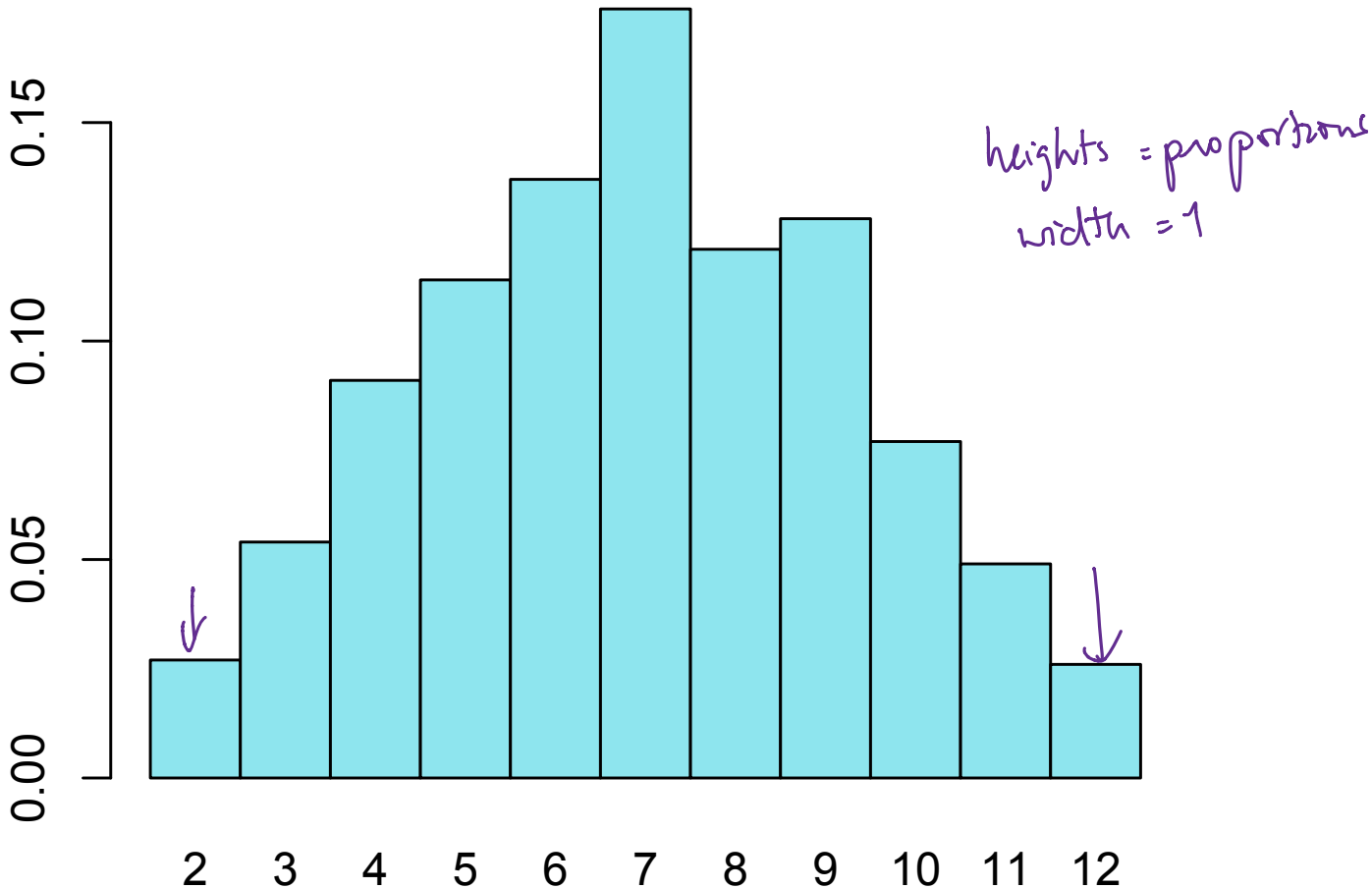
- The *percentage* of sixes, when rolling a fair die over and over, is very close to 1/6. True for any of the faces, so the *empirical* histogram of the results of rolling a die over and over again looks more and more like the *theoretical* probability histogram.

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

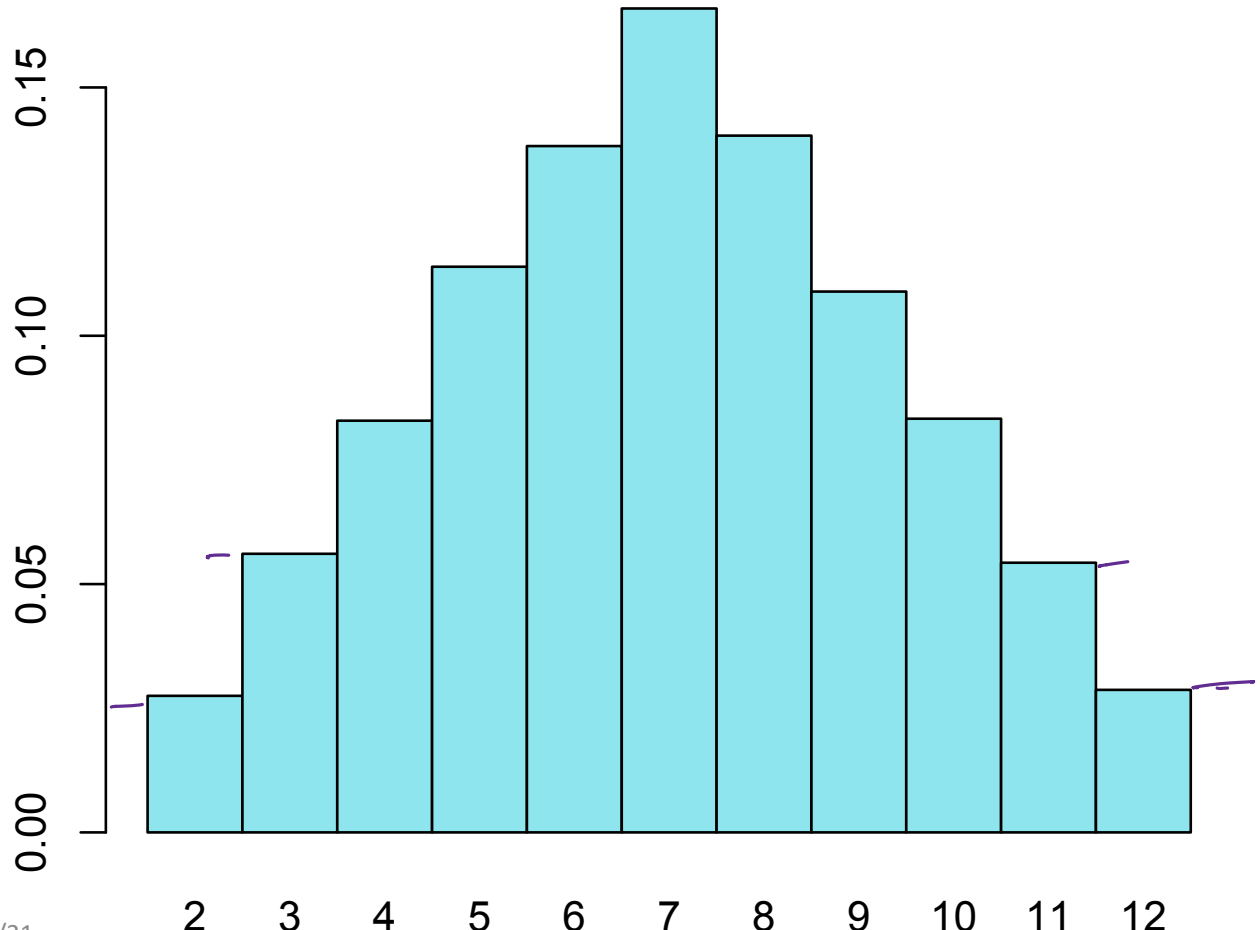
Empirical histogram for the sum of 2 dice (100 repetitions)



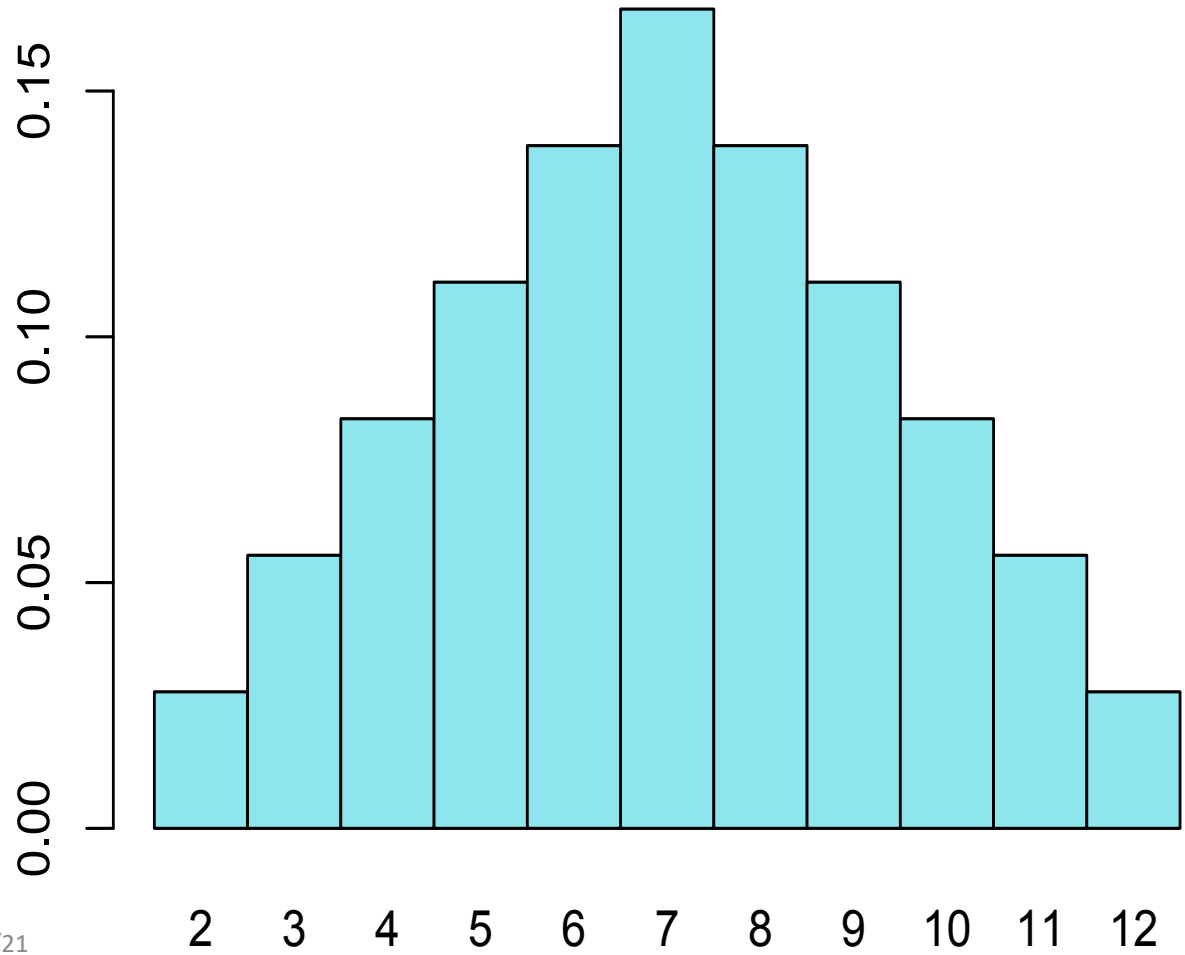
Empirical histogram for the sum of 2 dice (1,000 repetitions)



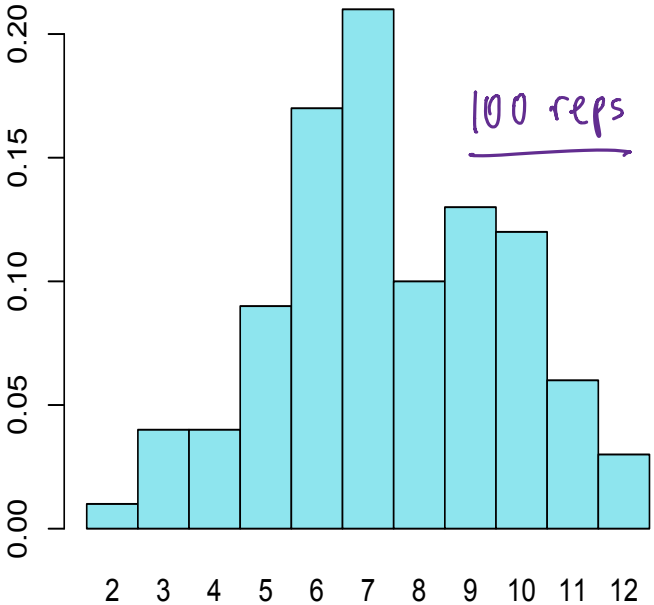
Empirical histogram for the sum of 2 dice (10,000 repetitions)

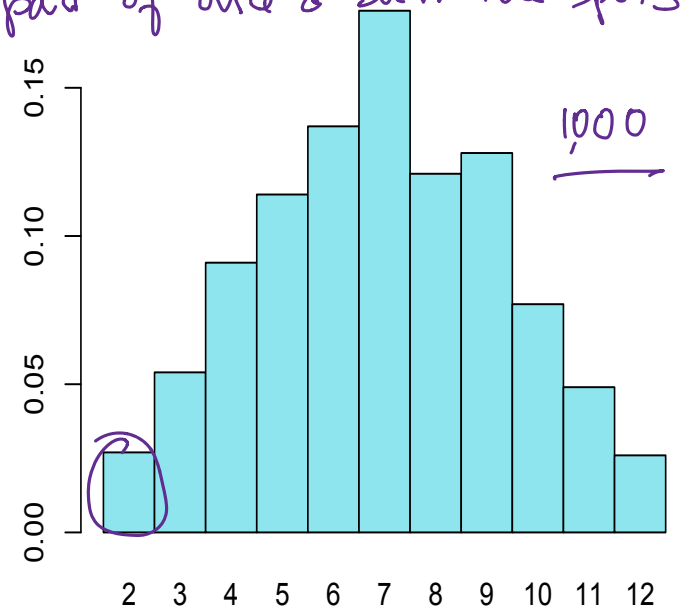


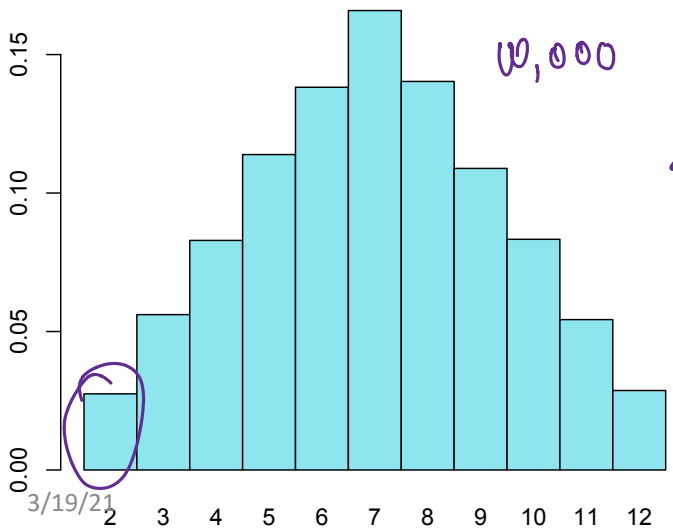
Probability histogram for the sum of 2 dice

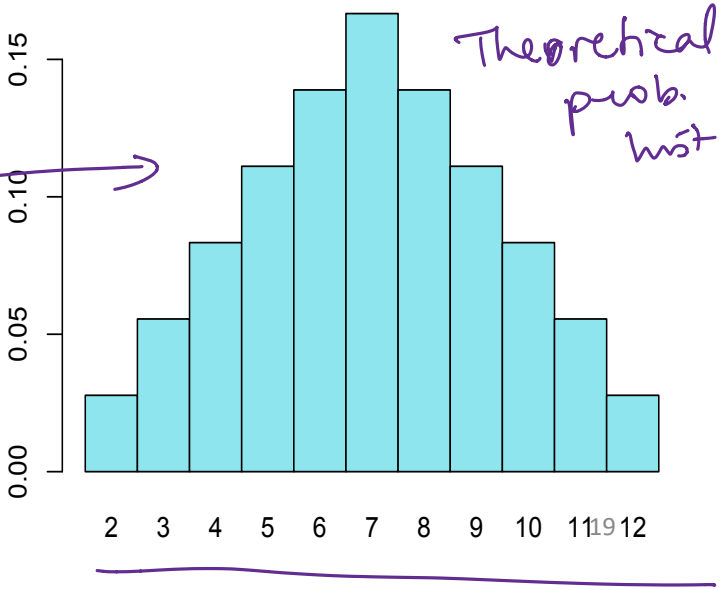


Roll a pair of dice & sum the spots

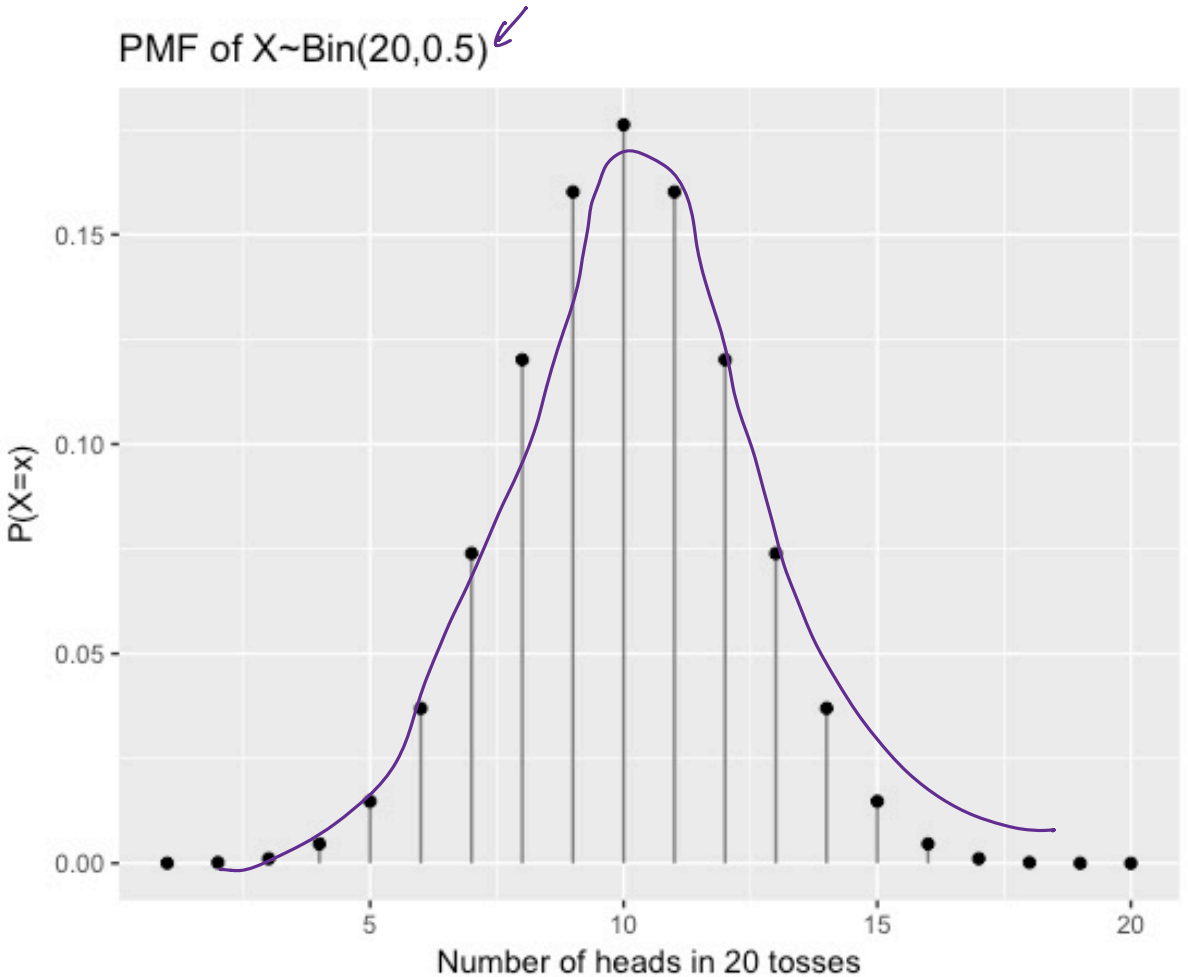




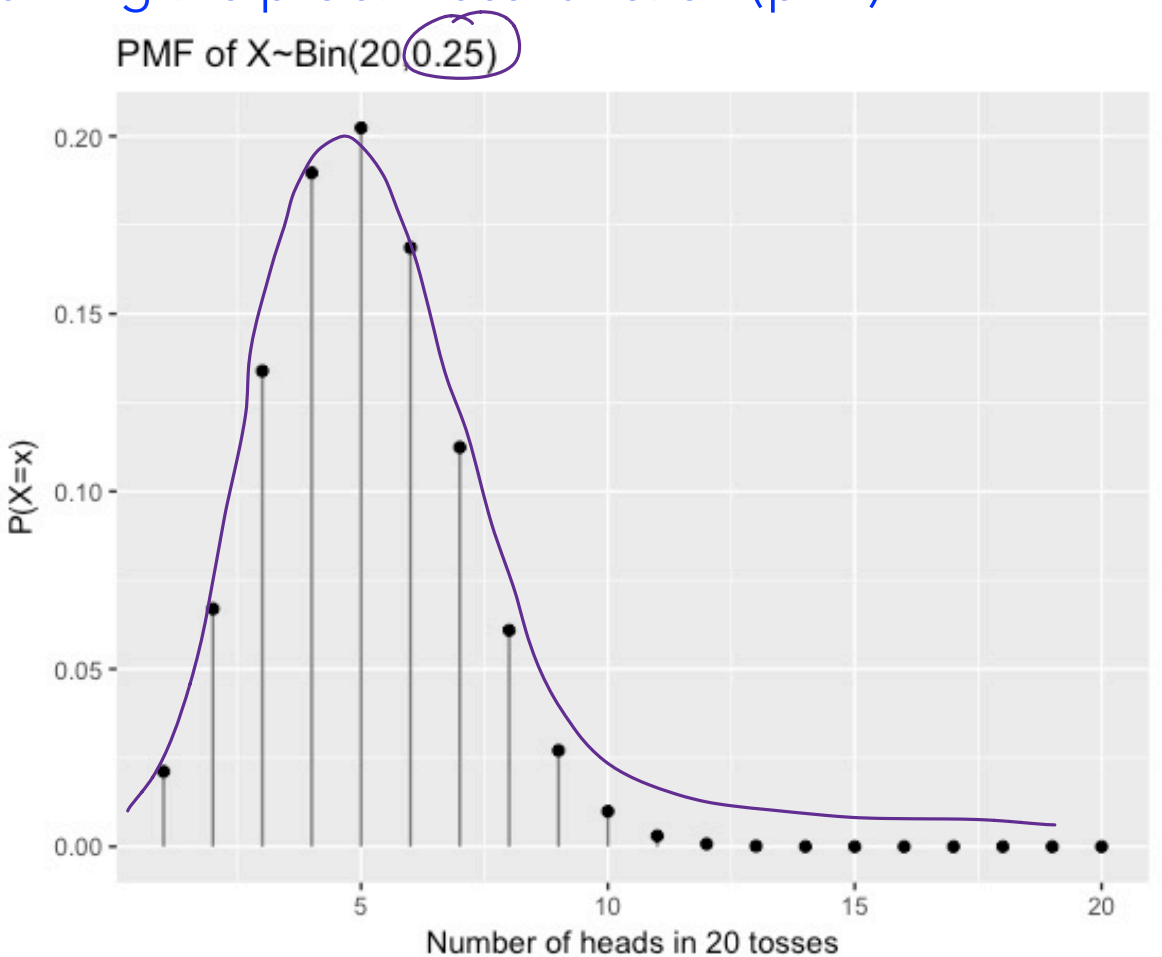




Visualizing the prob. mass function (pmf)



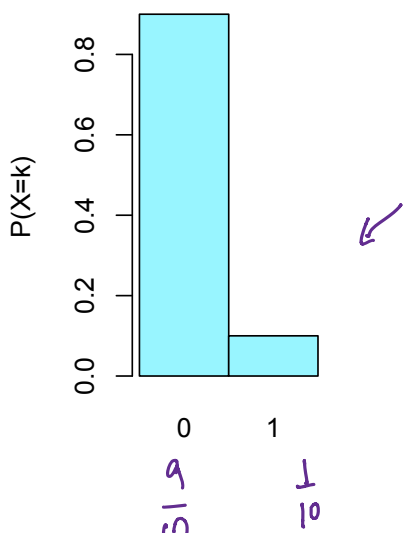
Visualizing the prob. mass function (pmf)



What if p is small?

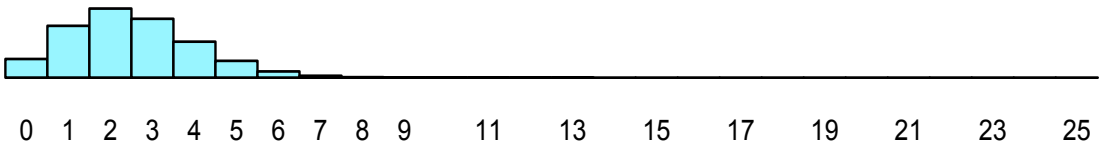
- Consider $X_k \sim Bernoulli\left(\frac{1}{10}\right), S_n = X_1 + X_2 + X_3 + \cdots + X_n, Sn \sim Bin(n, \frac{1}{10})$

- Draw the probability histogram for X_k:

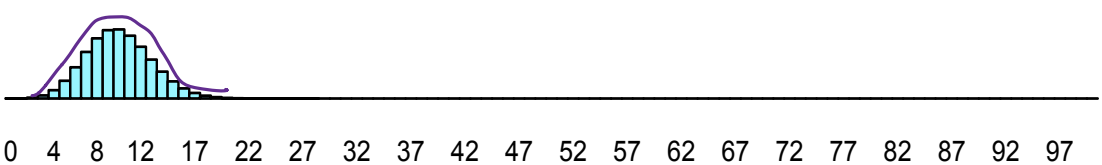


When p is small

n=25



n=100



n=400