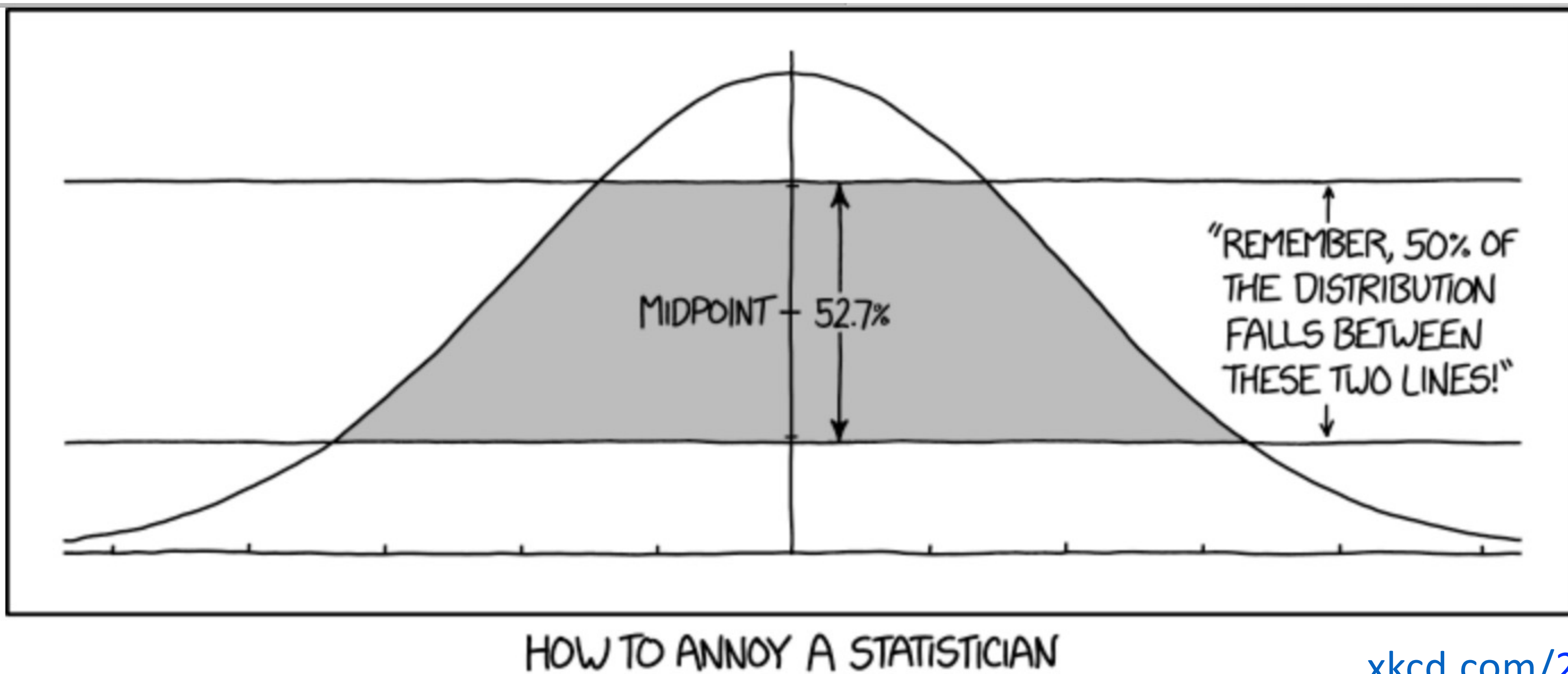
Stat 88: Prob. & Math. Statistics in Data Science



xkcd.com/2118

Lecture 19: 3/31/2022

The law of averages, distribution of a sample sum, the normal distribution, the Central Limit Theorem

7.3, 8.1, 8.2, 8.3, 8.4

Last lecture:

- The finite population correction or **fpc** $= \sqrt{\frac{N-n}{N-1}}$, and is the constant that we multiply the SD of sample sum computed WITH replacement by, to get the SD of the sample sum WITHOUT replacement.
- **SD of sum of an SRS = SD of sum WITH repl. × fpc**

- Let $S_n = X_1 + X_2 + \cdots + X_n$, then $SD(S_n) = \sqrt{n}\sigma$ and $SD\left(\frac{S_n}{n}\right) = \sigma/\sqrt{n}$
- The SD of the sample sum INCREASES with n
- The SD of the sample mean DECREASES with n

Accuracy of samples (depend on the SD of the sample mean/sum)

- Simple random samples of the same size of 625 people are taken in Berkeley (population: 121,485) and Los Angeles (population: 4 million). True or false, and explain your choice: The results from the Los Angeles poll will be substantially more accurate than those for Berkeley.

Fpc in case of Berkeley: 0.9974285

Fpc in case of LA: 0.999922

- A survey organization wants to take an SRS in order to estimate the percentage of people who watched the 2022 Oscars. 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.

What n to use? Note that the number of people who have watched the Oscars in the sample is a rv with the $HG(N, G, n)$ distiribution.

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

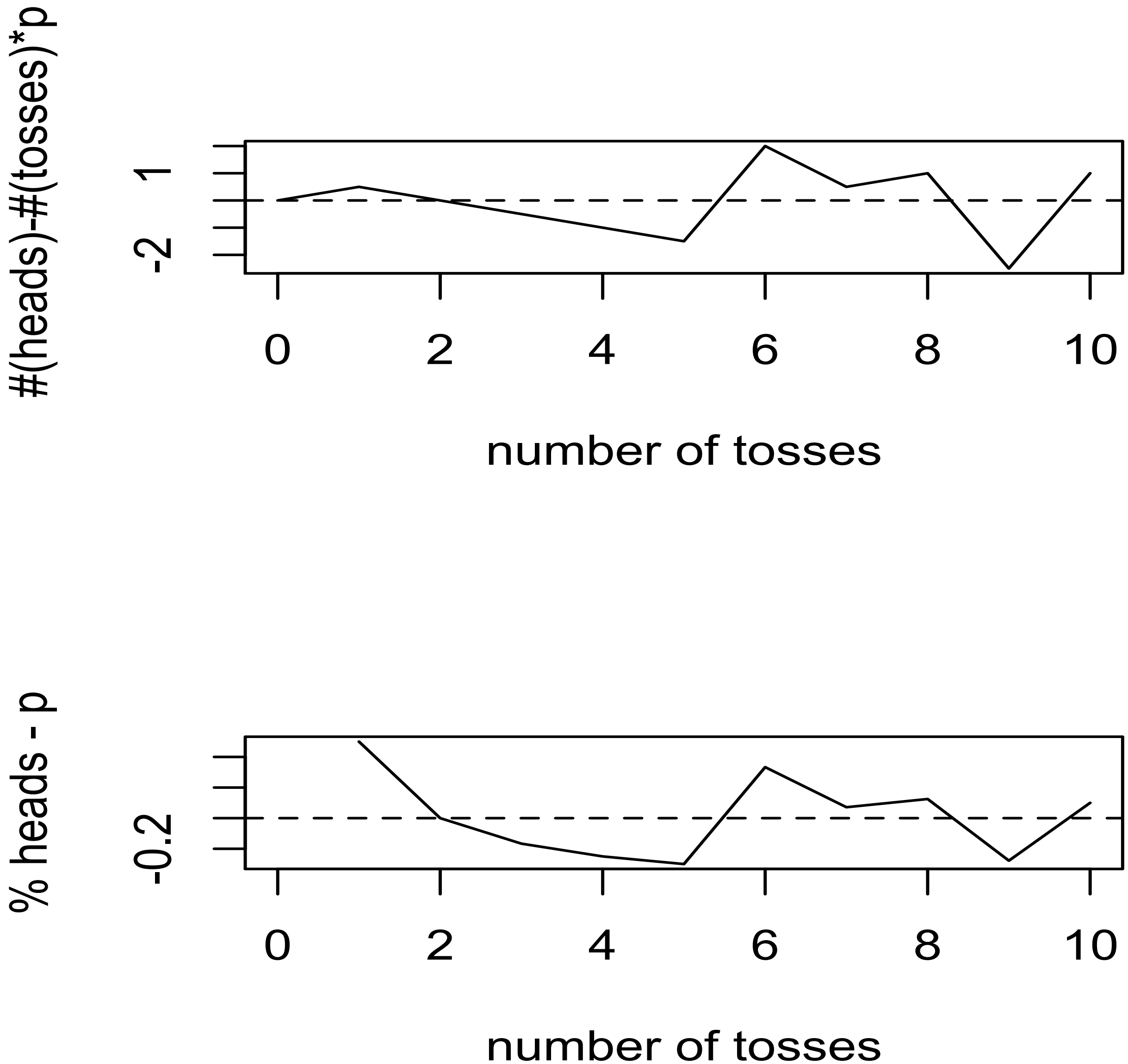
A survey organization wants to take an SRS in order to estimate the percentage of people who watched the 2022 Oscars. To keep costs down, they want to take as small a sample as possible, but their client will only tolerate a random error of 1 percentage point or so in the estimate. Should they use a sample size of 100, 2500, or 10000? The population is very large and the fpc is about 1.

- What n to use? Want to choose n such that percentage of people in the sample who have watched the Oscars is not more than 0.01.

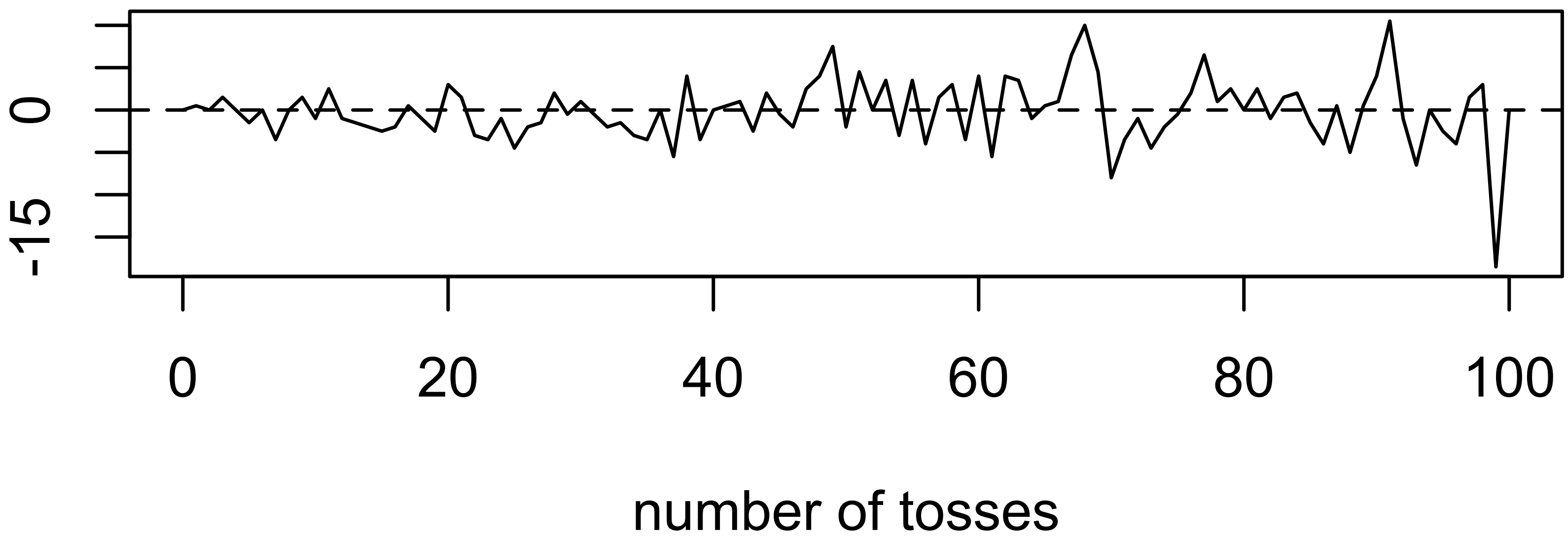
- Note that the number of people who have watched the Oscars in the sample is a rv with the $HG(N, G, n)$ distribution, but we are told that N is very large & $fpc \approx 1$, so we can approximate the prob. using the $Bin(n, p)$ distribution, where p is the percentage of people who watched the Oscars (which is what we are trying to estimate).

- $SD\left(\frac{S_n}{n}\right) = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{pq}}{\sqrt{n}} \leq \frac{0.5}{\sqrt{n}} \leq 0.01 \Rightarrow n \geq 2500$

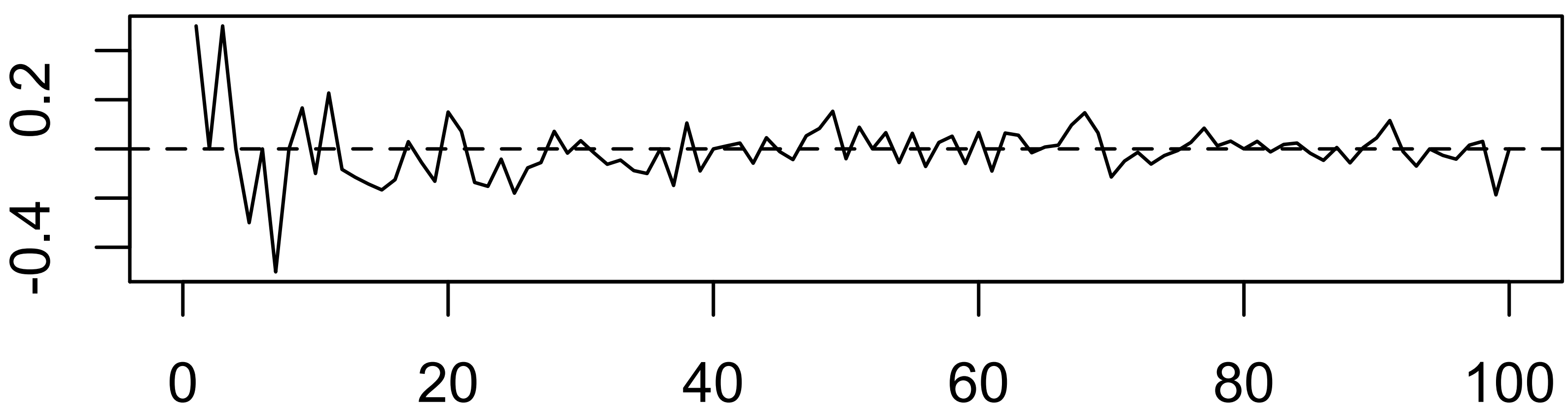
Simulating coin tosses: 10 tosses (adapted from FPP)

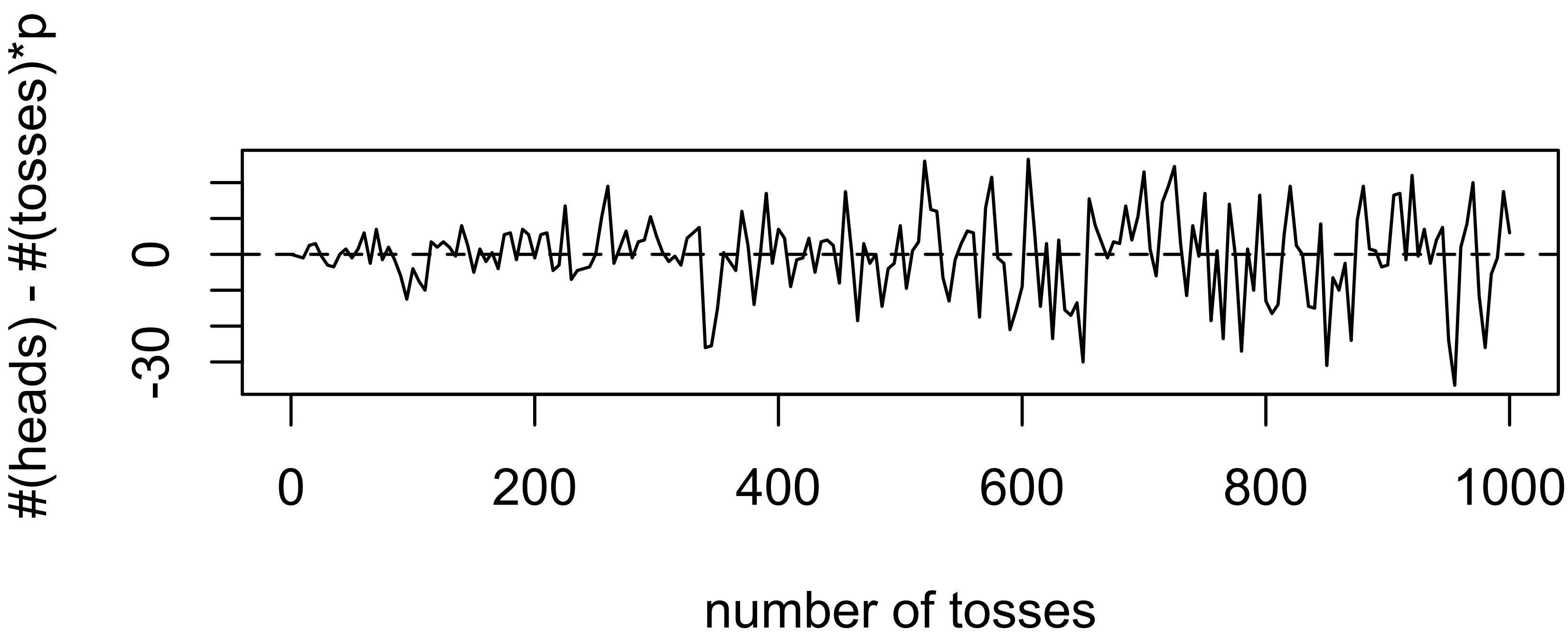


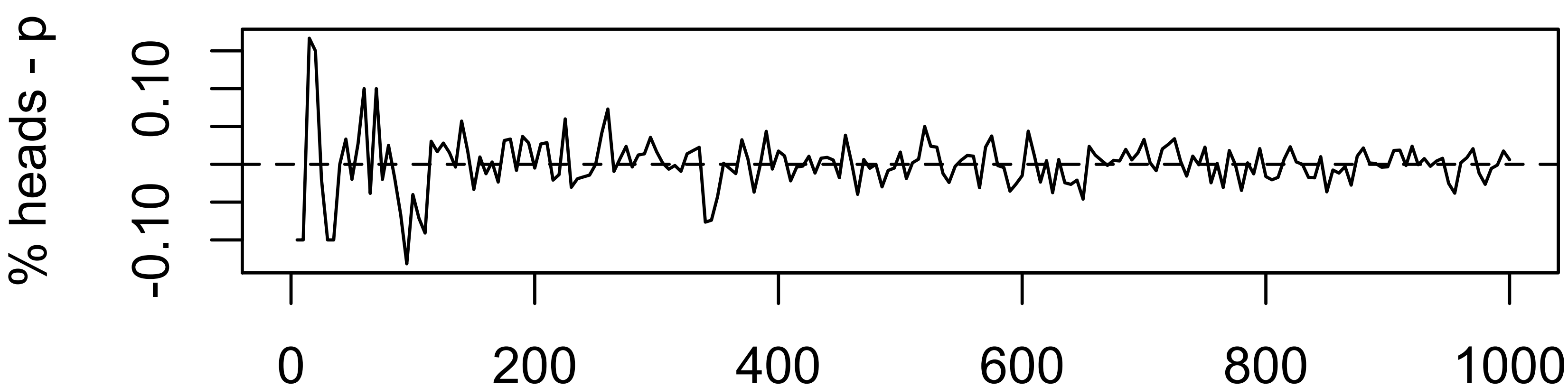
Observed
error=
#H- #tosses/2

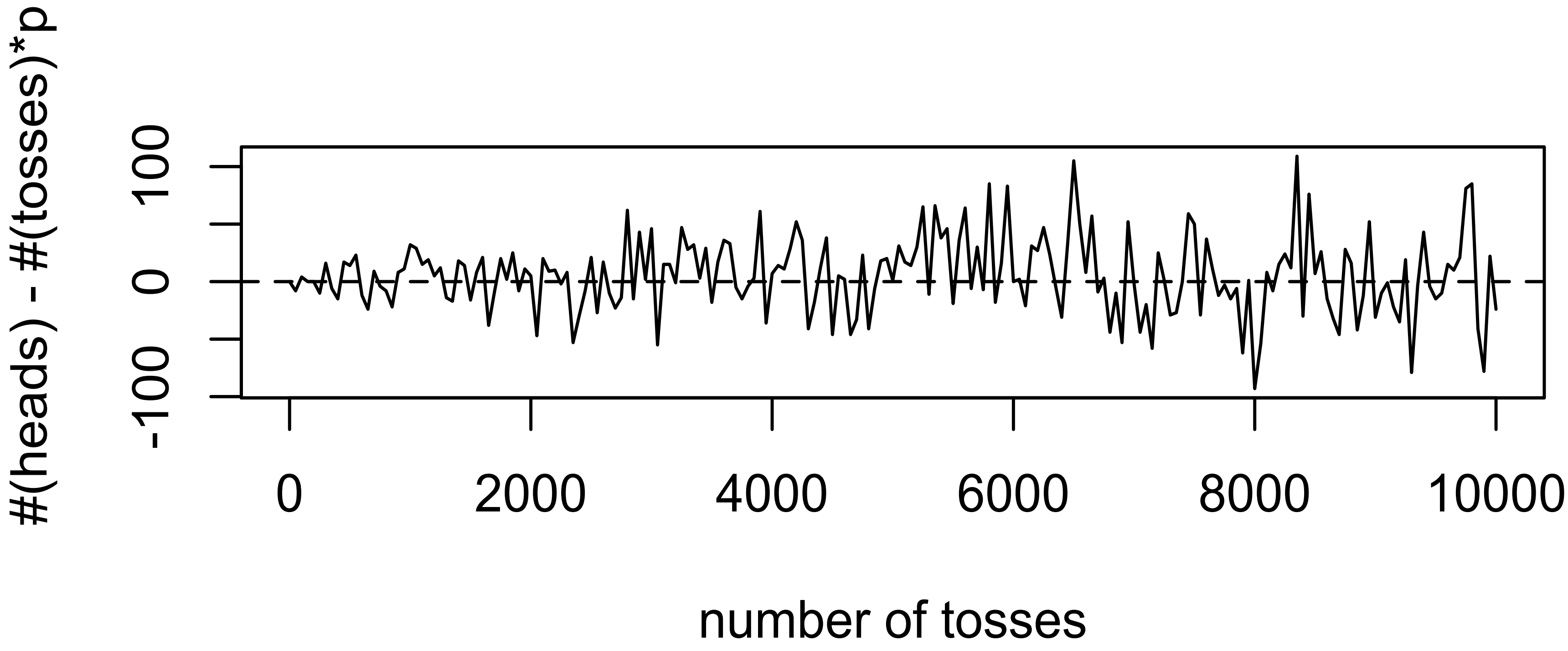


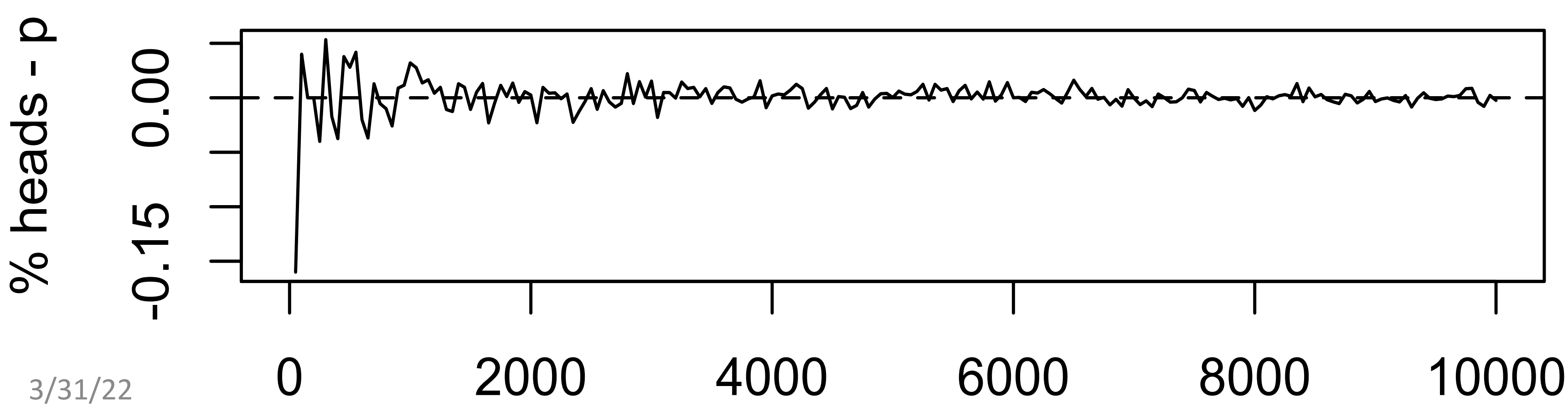
% error=
%H-0.5




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


% heads - p
0.10
-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
-0.15
0.00
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.

- 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 $SD(A_n) = ??$

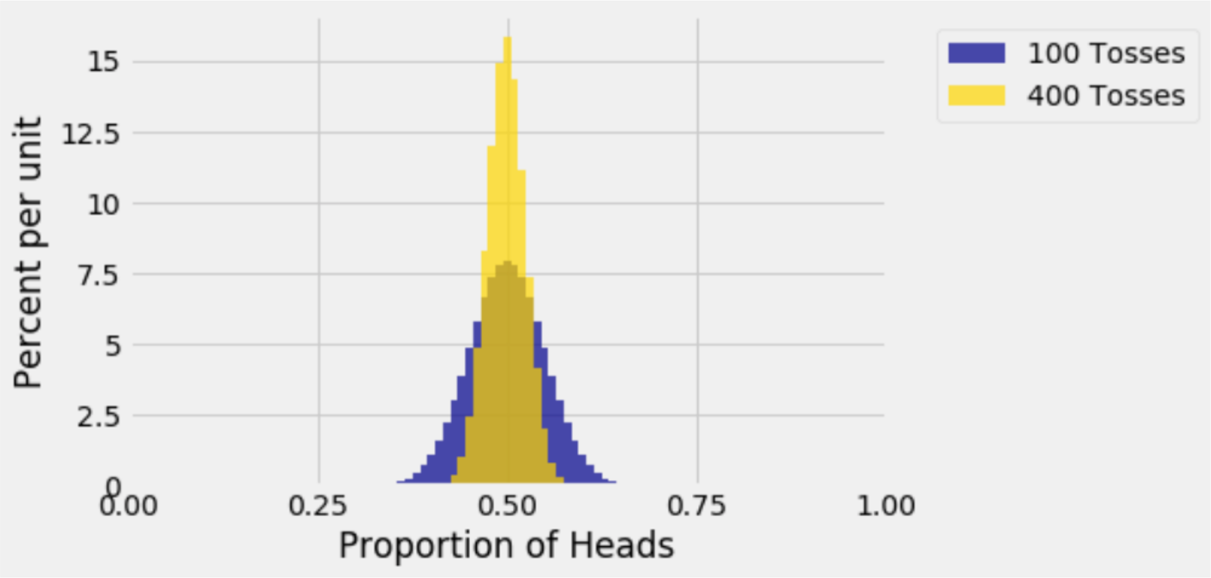
- **The square root law:** the *accuracy* of an estimator is measured by its SD, the ***smaller*** the SD, the ***more accurate*** the estimator, but if you multiply the sample size by a factor, the accuracy only goes up by the **square root** of the factor.
- In our earlier example, we _____ the accuracy by quadrupling the size.

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:**

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

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



From section 7.3

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)

Exercise 7.4.11

Each Data 8 student is asked to draw a random sample and estimate a parameter using a method that has chance 95% of resulting in a good estimate.

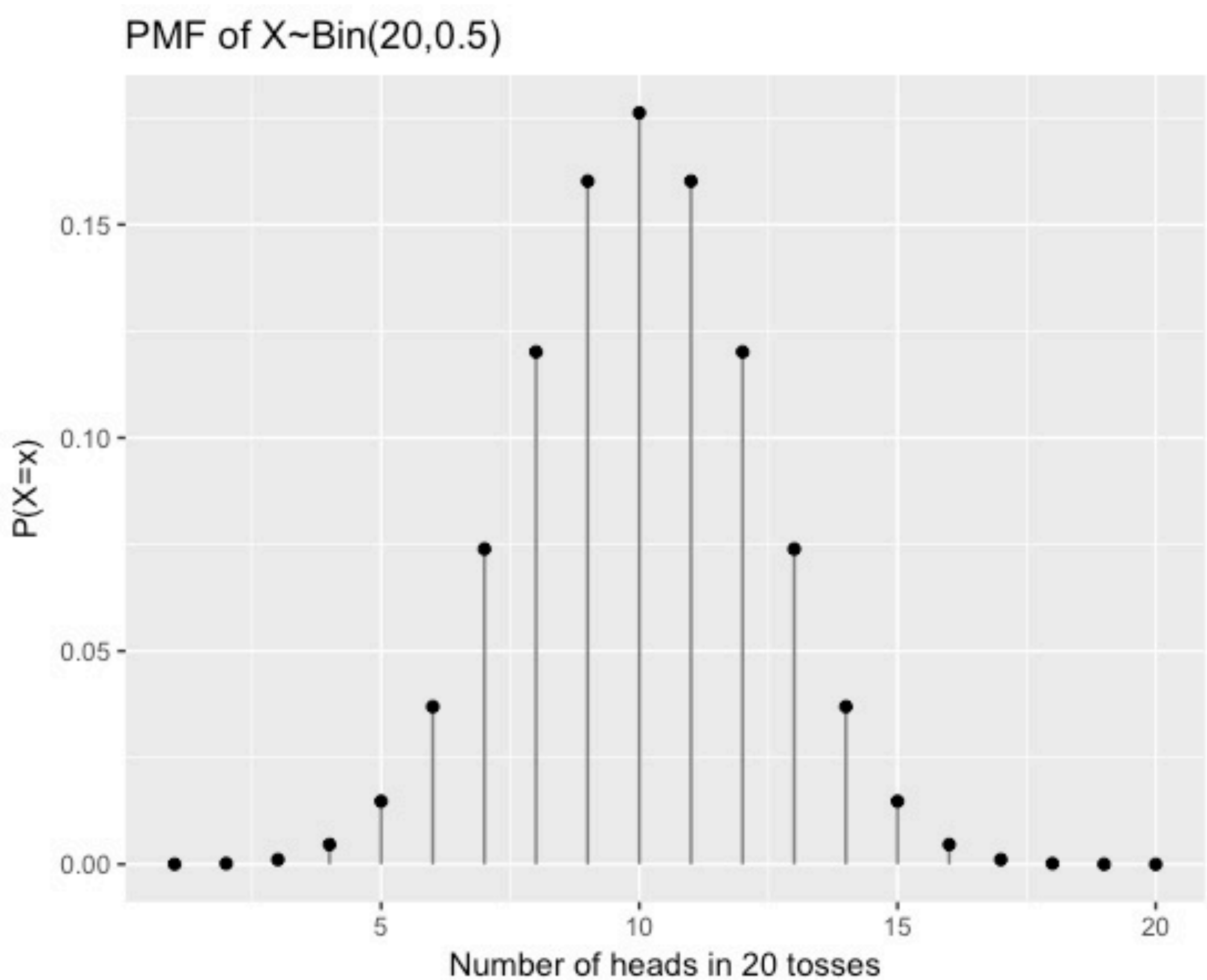
Suppose there are 1300 students in Data 8. Let X be the number of students who get a good estimate. Assume that all the students' samples are independent of each other.

- a) Find the distribution of X
- b) Find $E(X)$ and $SD(X)$.
- c) Find the chance that more than 1250 students get a good estimate.

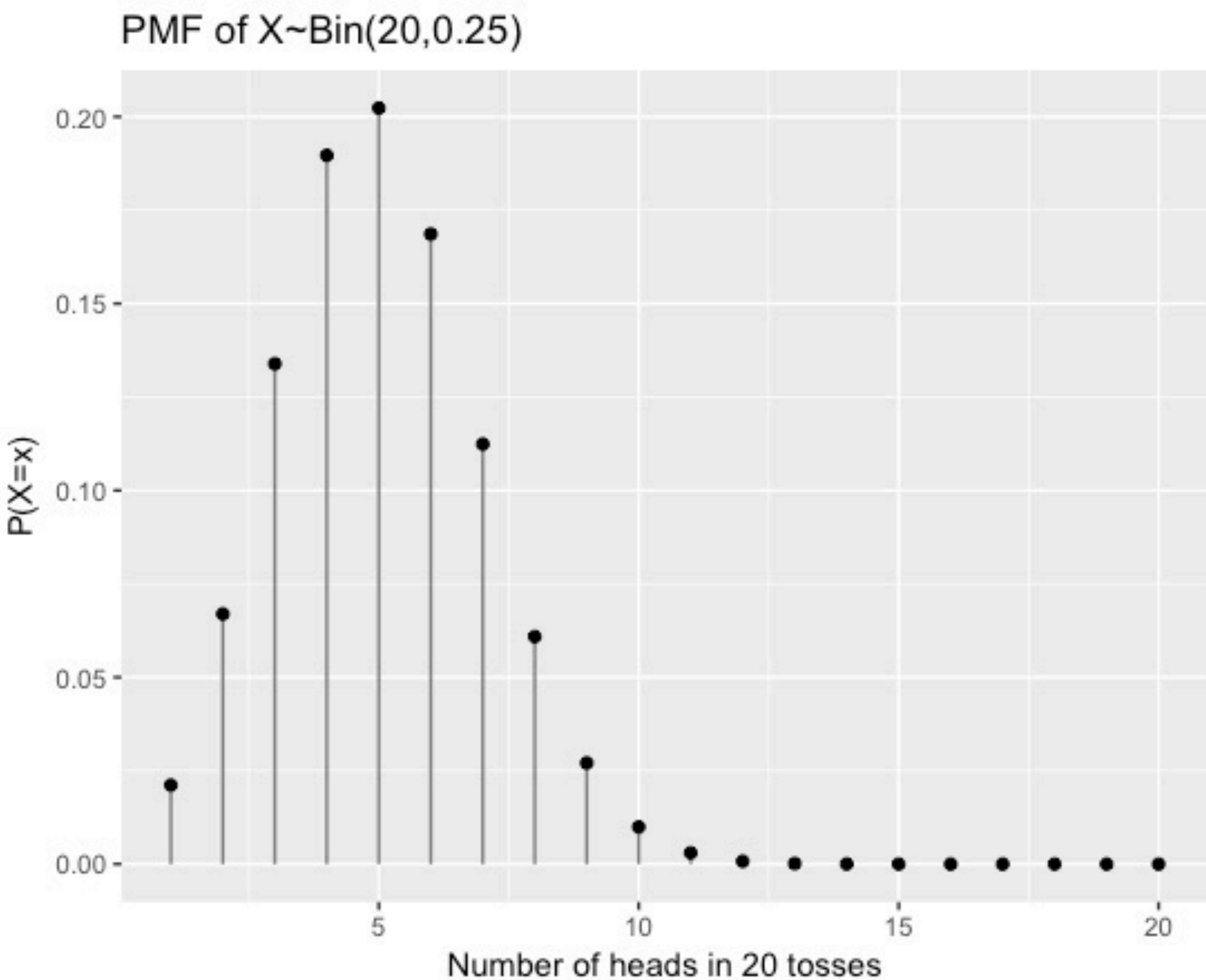
8.1: Distribution of a sample sum

- 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:



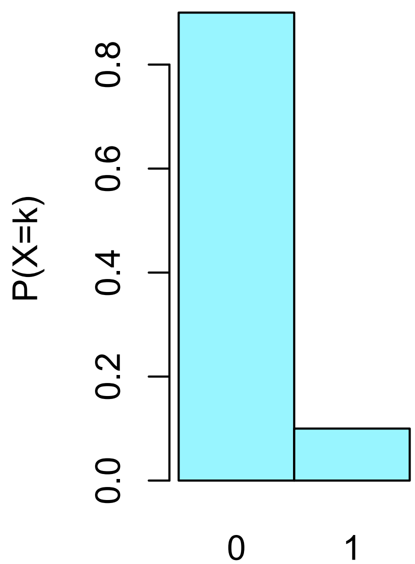
Visualizing the prob. mass function (pmf)



What if p is small?

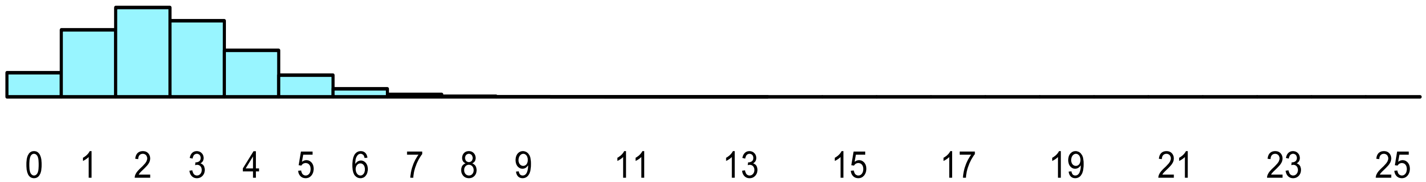
- Consider $X_k \sim Bernoulli\left(\frac{1}{10}\right), S_n = X_1 + X_2 + X_3 + \cdots + X_n, Sn \sim Bin(n, \frac{1}{10})$

- Draw the probability histogram for X_k:

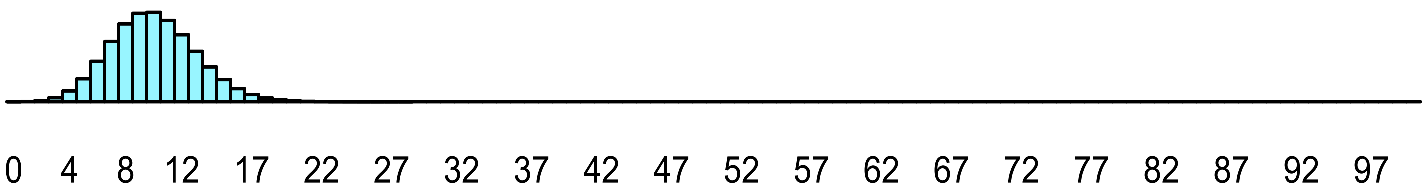


When p is small (picture adapted from *Statistics* by Freedman, Pisani, and Purves)

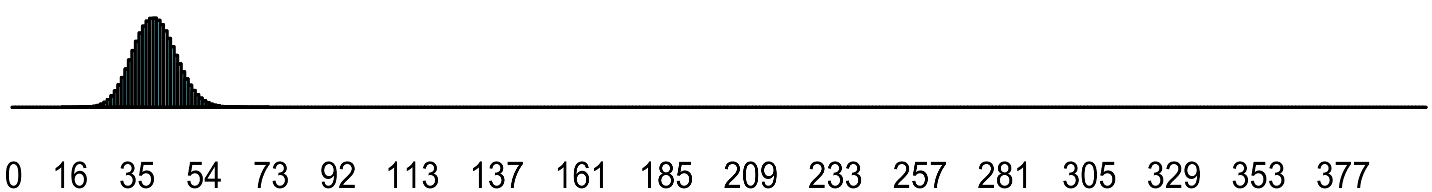
n=25



n=100



n=400



Distribution of the sample sum

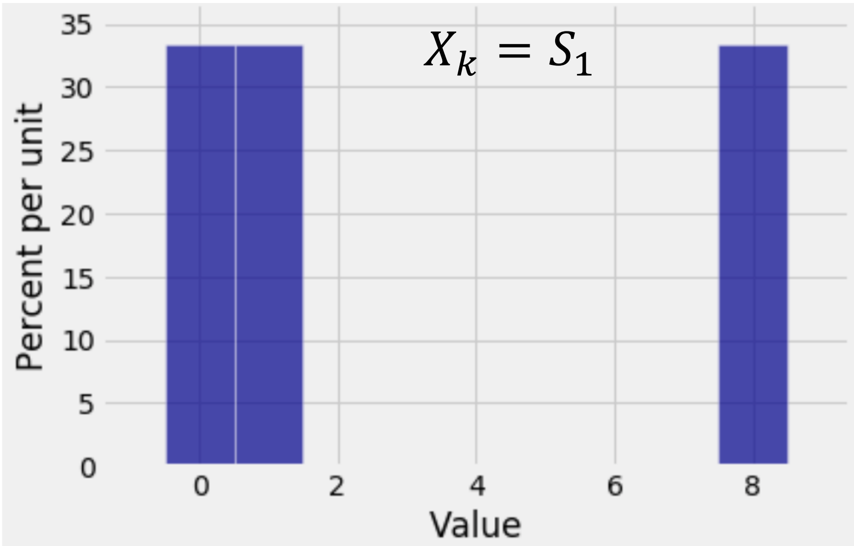
- More generally, let's consider $X_1, X_2, \ldots, X_n$ iid with mean μ and SD σ
- Let $S_n = X_1 + X_2 + \cdots + X_n$
- We know that $E(S_n) = n\mu$ and $SD(S_n) = \sqrt{n}\sigma$

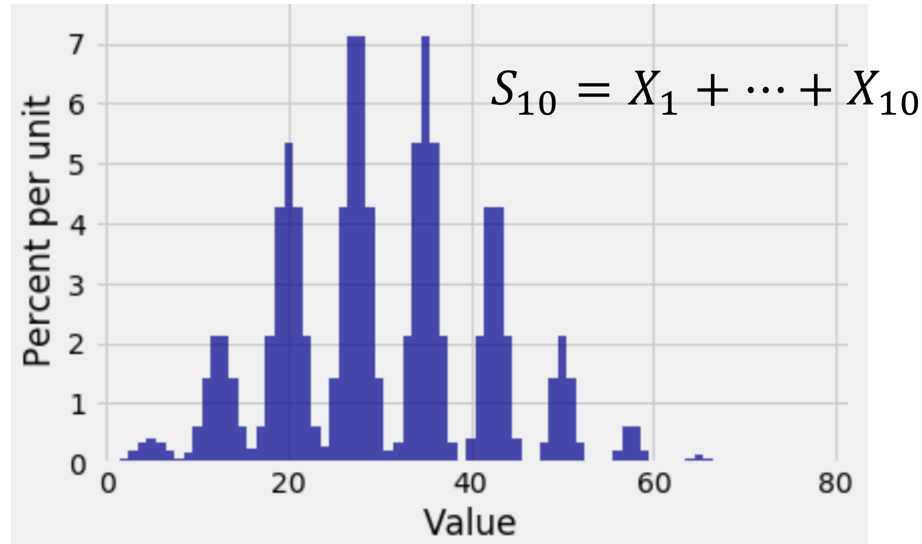
- We want to say something about the distribution of S_n, and while it may be possible to write it out analytically, if we know the distributions of the X_k, it may not be easy. And we may not even know anything beyond the fact that the X_k are iid, and we might be able to guess at their mean and SD.

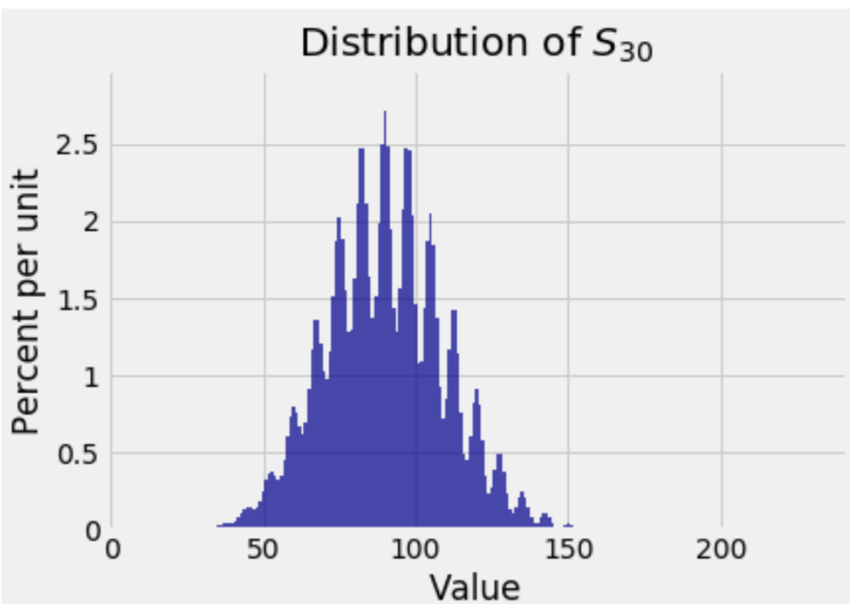
- We saw in the previous slides that even if the X_k are very far from symmetric, the distribution of the sum begins to look quite nice and bell shaped.

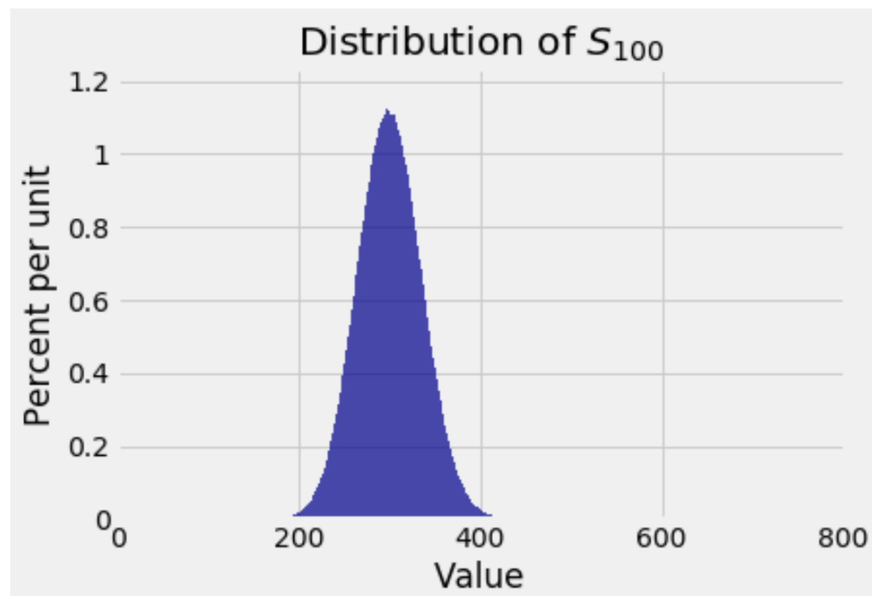
- What if the X_k are strange looking?

Weird X_k distributions – is the distribution of S_n different?





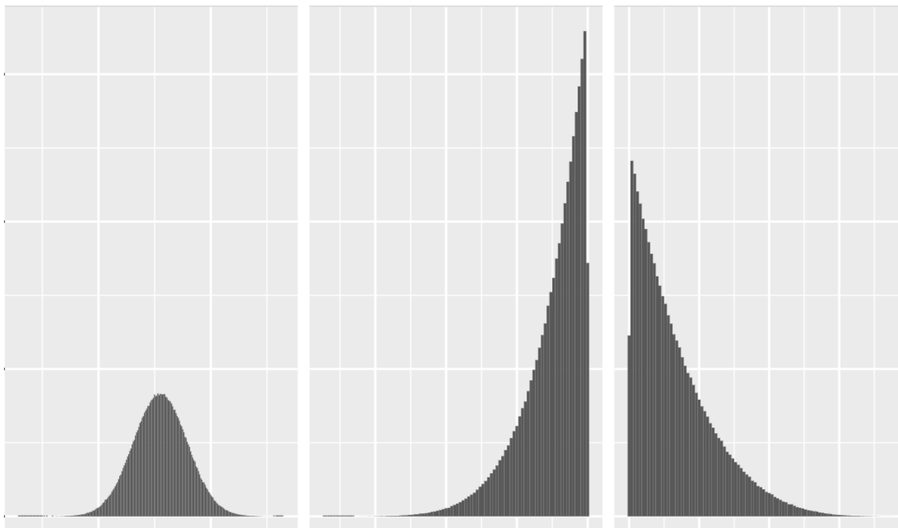




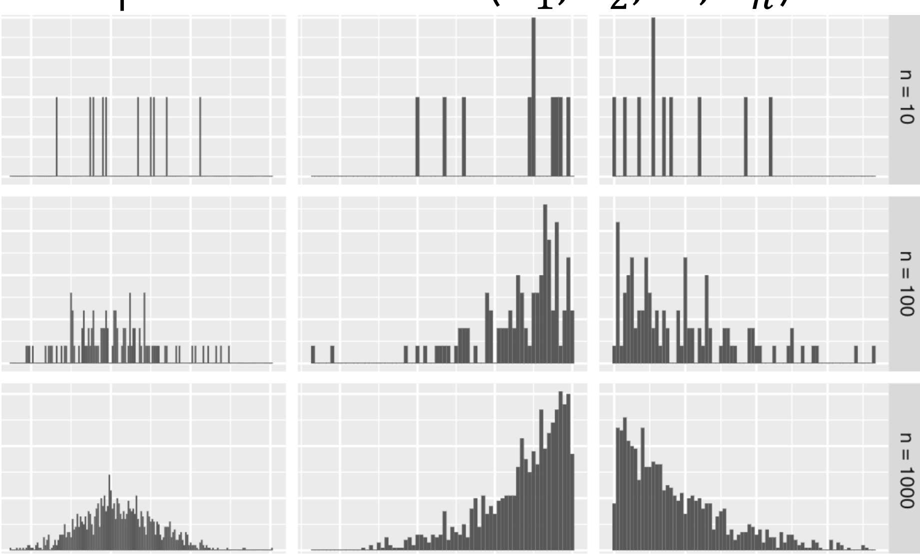
From section 8.1

Examples by picture

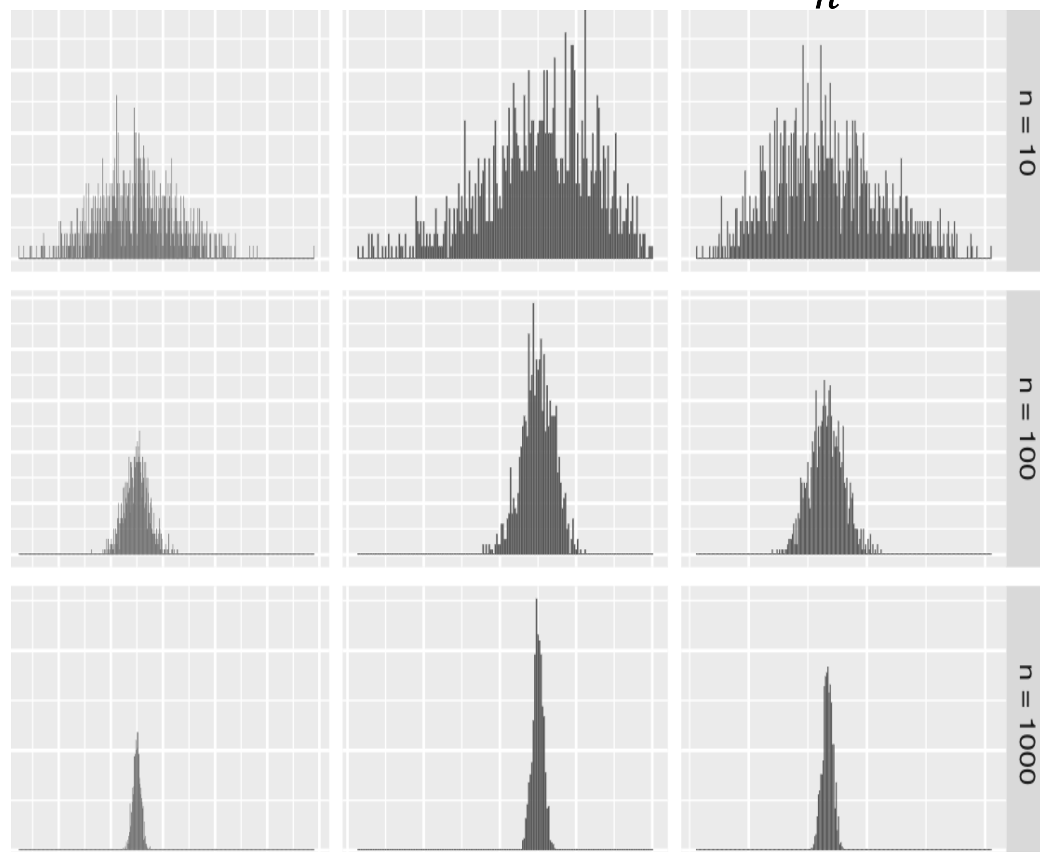
Probability distribution of X_k



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



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



Graphs created by Sarah Johnson for Stat 20

The Central Limit Theorem

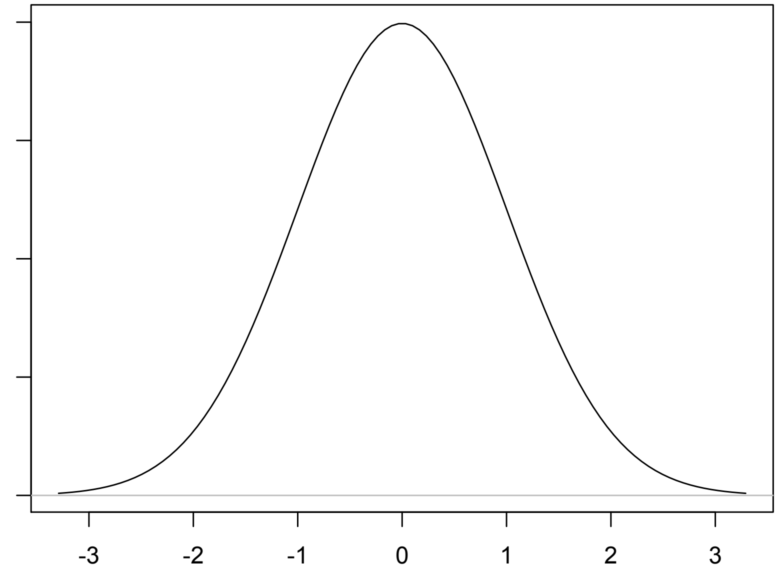
- The bell-shaped distribution is called a *normal curve*.

- 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.

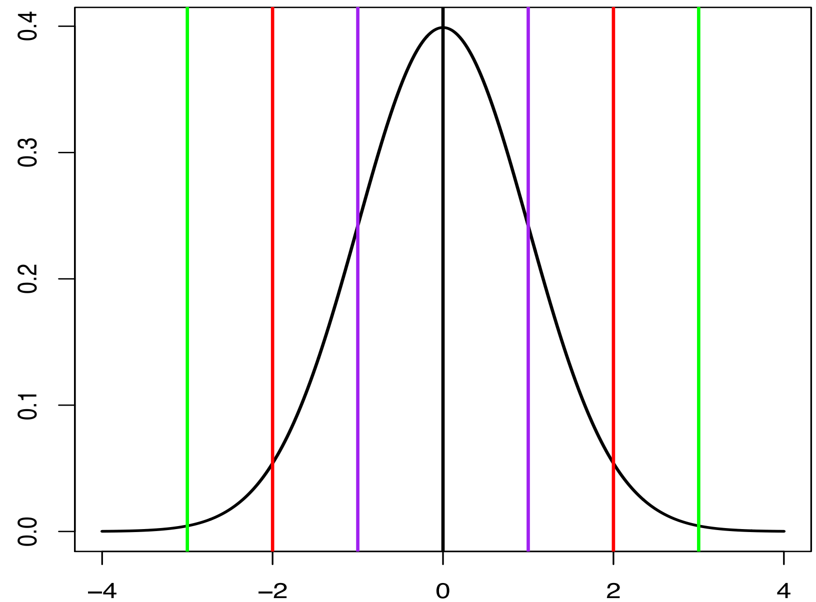
- The distribution is approximately normal (bell-shaped) centered at $E(S_n) = n\mu$ and the width of this curve is defined by $SD(S_n) = \sqrt{n}\,\sigma$

Bell curve: the Standard Normal Curve

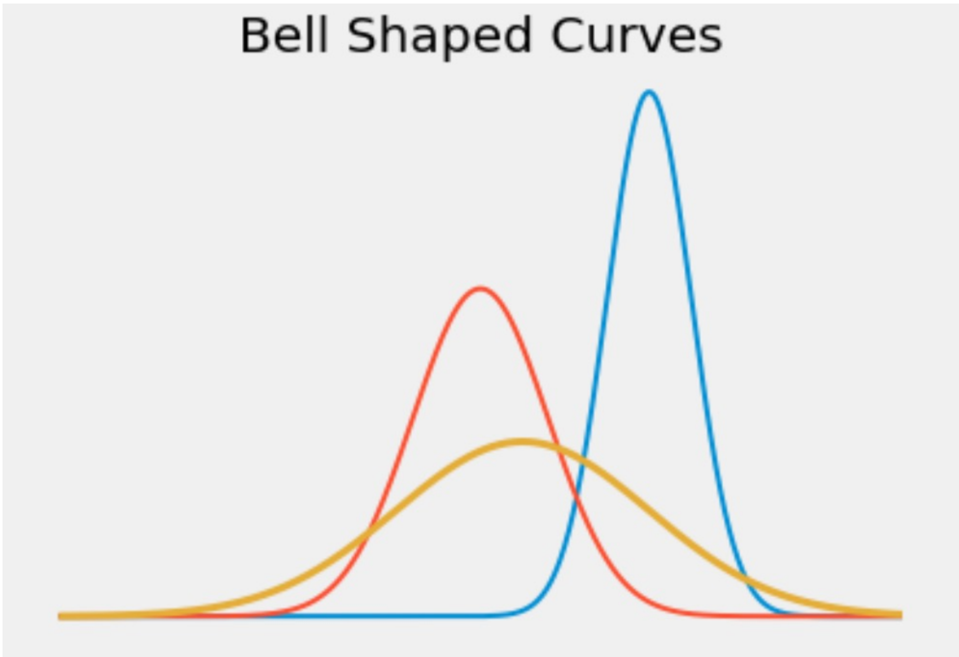
- Bell shaped, symmetric about 0
- Points of inflection at $z = \pm 1$
- Total area under the curve = 1, so can think of curve as approximation to a probability histogram
- Domain: whole real line
- Always above x-axis
- Even though the curve is defined over the entire number line, it is pretty close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



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



The many normal curves → the *standard normal* curve

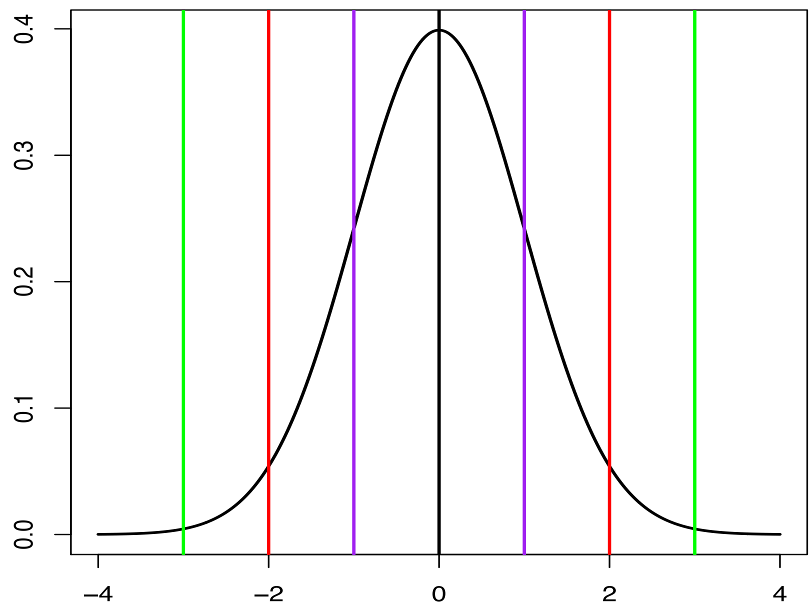


- Just one normal curve, standard normal, centered at 0. All the rest can be derived from this one.

Standard normal cdf

- $\Phi(z) = \int_{-\infty}^{z} \phi(x)dx$

How do we approximate the area of a range in a histogram?



Total area under the curve = 100%
Curve is symmetric about 0
The areas between 1, 2 and 3 SDs away:
Between -1 and 1 the area is 68.27%
Between -2 and 2 the area is 95.45%
Between -3 and 3 the area is 99.73%

- Many histograms bell-shaped, but not on the same scale, and not centered at 0
- Need to convert a value to ***standard units*** – see how many SDs it is above or below the average
- Then we can **approximate** the area of the histogram using the area under the standard normal curve (using for example, stats.norm.cdf for the actual numerical computation)

- 68%-95%-99.7% rule

  (**Empirical rule**)

Normal approximations : standard units

- Let X be any random variable, with expectation μ and SD σ, consider a new random variable that is a linear function of X, created by shifting X to be centered at 0, and dividing by the SD. If we call this new rv X^*, then X^* has expectation ______ and SD ______.
- $E(X^*) =$
- $SD(X^*) =$

- This new rv does not have units since it measures how far above or below the average a value is, in SD's. Now we can compare things that we may not have been able to compare.

- Because we can convert anything to standard units, ***every normal curve is the same.***

How to decide if a distribution could be normal

- Need enough SDs on both sides of the mean.

- In 2005, the murder rates (per 100,000 residents) for 50 states and D.C. had a mean of 8.7 and an SD of 10.7. Do these data follow a normal curve?

- If you have indicators, then you are approximating binomial probabilities. In this case, if *n* is very large, but *p* is small, so that *np* is close to 0, then you can't have many sds on the left of the mean. So need to increase *n,* stretching out the distribution and the n the normal curve begins to appear.

- If you are not dealing with indicators, then might bootstrap the distribution of the sample mean and see if it looks approximately normal.