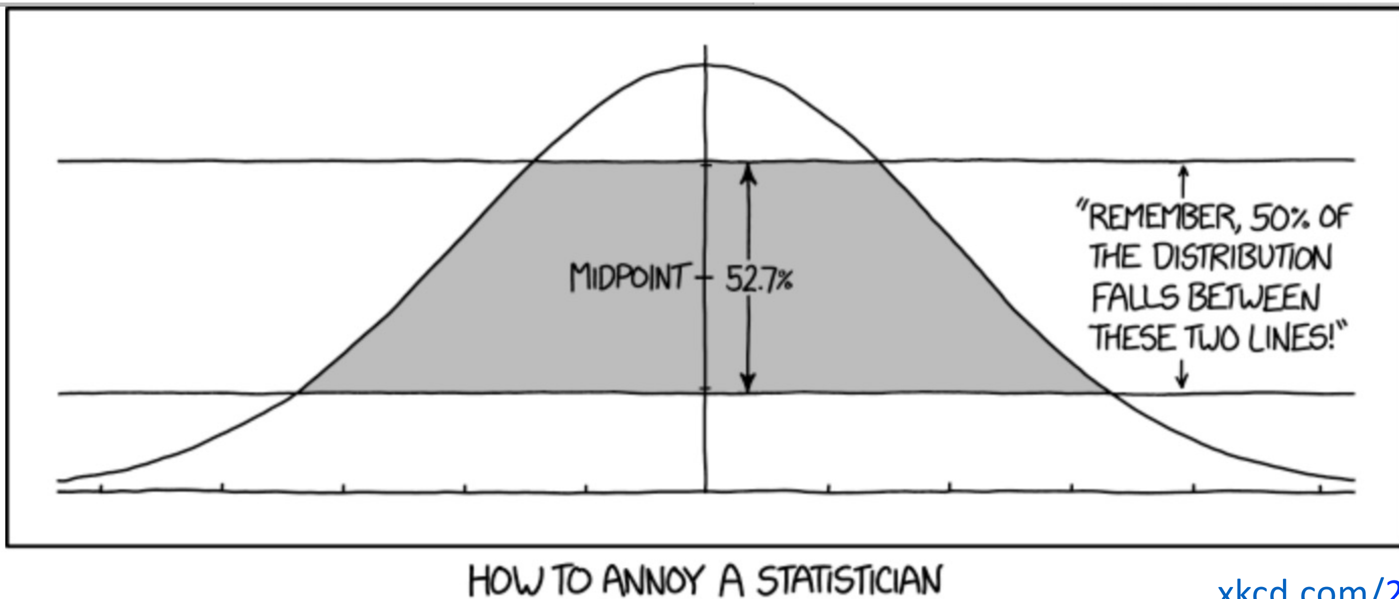
Stat 88: Prob. & Math. Statistics in Data Science



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:

$FPC = \sqrt{\frac{N-n}{N-1}} < 1, n > 1$

- 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 = $\underbrace{\text{SD of sum WITH repl.}}_{S_n}$ × fpc
- Let $S_n = X_1 + X_2 + \cdots + X_n$ (X_k are iid), then $\boxed{SD(S_n) = \sqrt{n}\sigma}$ 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

$= \frac{1}{n} SD(S_n) = \frac{\sqrt{n}\sigma}{n} = \sigma/\sqrt{n}$

Let $S_n = X_1 + X_2 + \cdots + X_n$, X_k are iid., $E(X_k) = \mu$, $Var(X_k) = \sigma^2$

$Var(S_n) = Var(X_1 + X_2 + \cdots + X_n)$

$= Var(X_1) + Var(X_2) + \cdots + Var(X_n)$

$= \sigma^2 + \sigma^2 + \cdots + \sigma^2 = n\sigma^2$

$SD(S_n) = \sqrt{n} \cdot \sigma \leftarrow$ SD of S_n, W/REPL

SD of S_n w/o repl $= \sqrt{n} \cdot \sigma \cdot \sqrt{\frac{N-n}{N-1}}$

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.

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

Fpc in case of Berkeley: 0.9974285

Fpc in case of LA: 0.999922

Example adapted from Statistics, by FPP

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

- 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.

But we will pretend that it is a Binomial(n, p) dsn b/c $fpc \approx 1$

Since N is very large. $X = \#$ of people in sample who watched

$$\boxed{\begin{aligned} &SD(aX) \\ &= |a|\,SD(X) \end{aligned}}$$

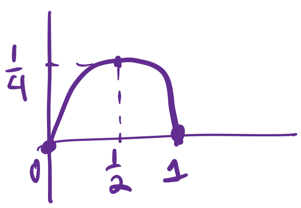
$X \sim Bin(n, p)$ $\quad E(X) = np$, $\quad SD(X) = \sqrt{npq}$

$q = 1-p$

$$SD(\underbrace{\text{sample percent}}_{(= A_n)}) = SD\left(\frac{X}{n}\right) = \frac{1}{n} SD(X) = \sqrt{\frac{pq}{n}}$$

$f(x) = x(1-x)$

$0 \le x \le 1$



$$p(1-p) \le \frac{1}{4} = \frac{1}{2}\cdot\frac{1}{2}$$

$$\sqrt{p(1-p)} = \sqrt{pq} \le \frac{1}{2}$$

$$SD\left(\frac{X}{n}\right) \le \frac{0.5}{\sqrt{n}} \le 0.01$$

??

$$\frac{0.5}{0.01} \le \sqrt{n}$$

$$\boxed{2500 \le n}$$

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

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

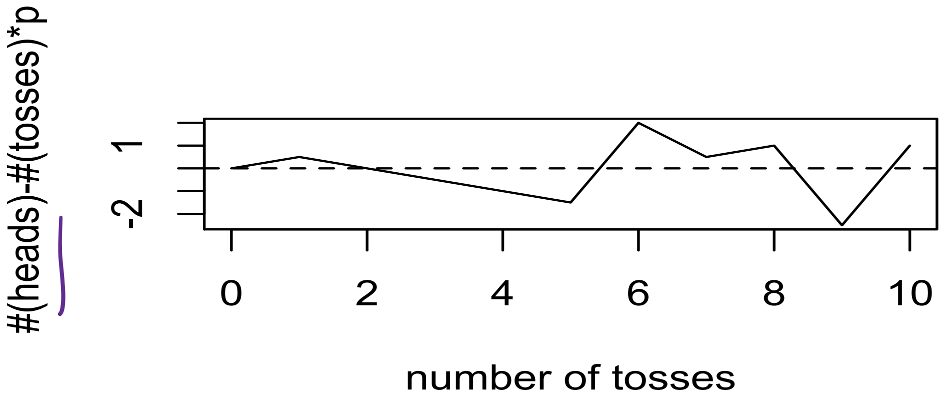
Simulating coin tosses: 10 tosses (adapted from FPP)

Statistics by Freedman, Pisani & Purves.

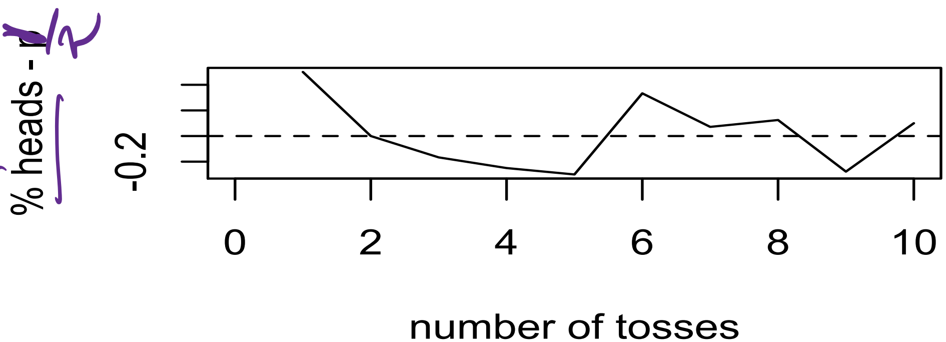
$p = \frac{1}{2}$

observed error

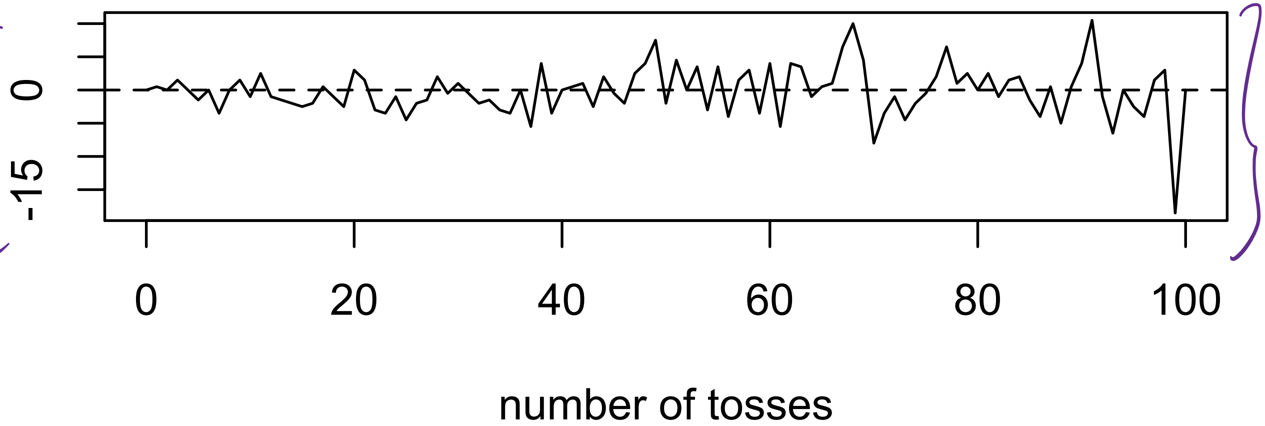
$\#H - \frac{n}{2}$



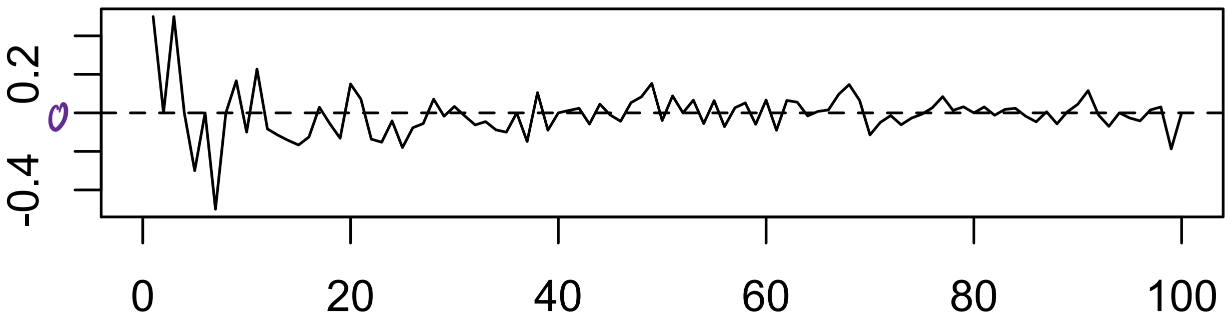
% error $= \frac{\text{obs. error}}{n}$

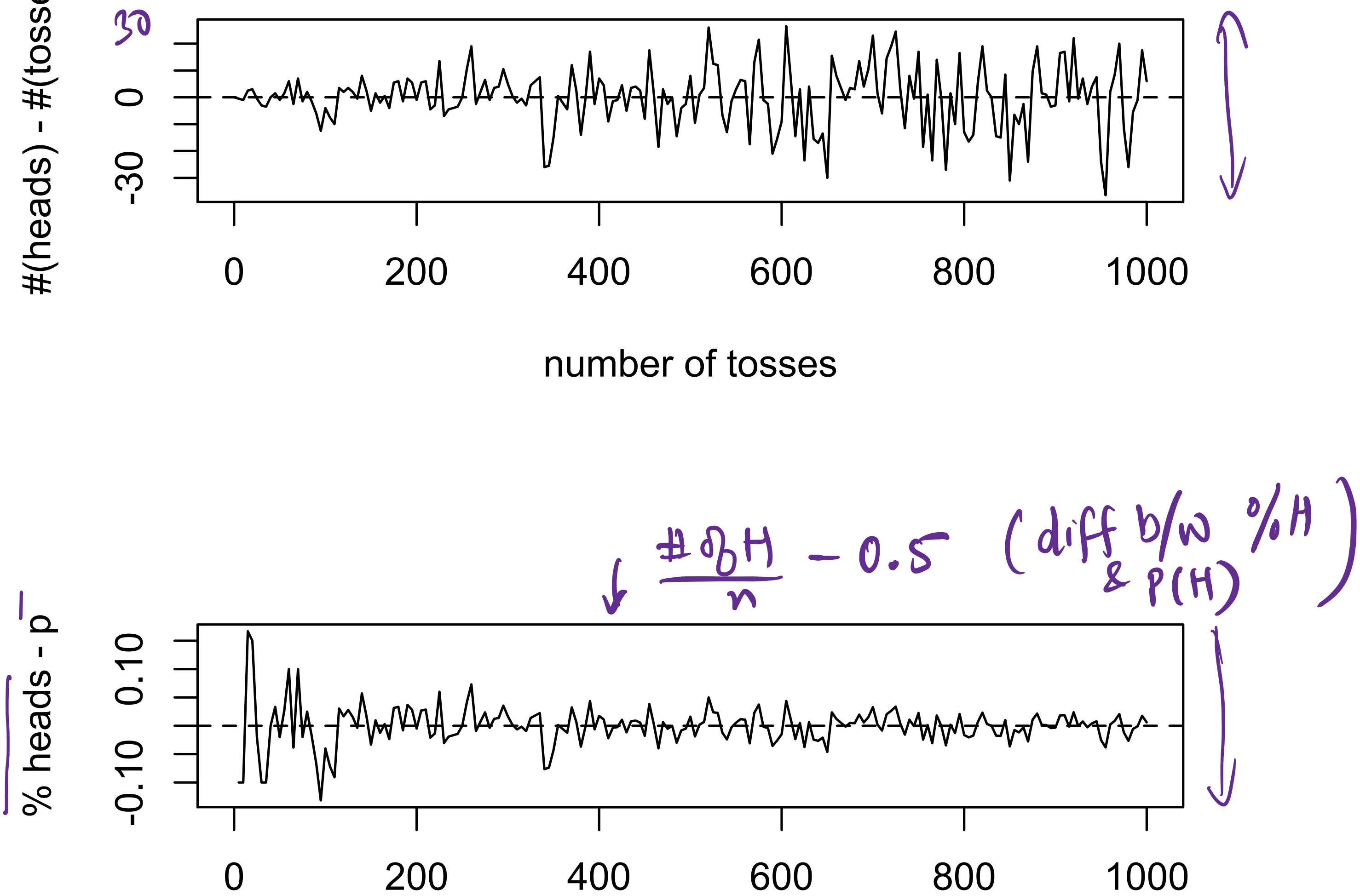


Observed error=
#H- #tosses/2

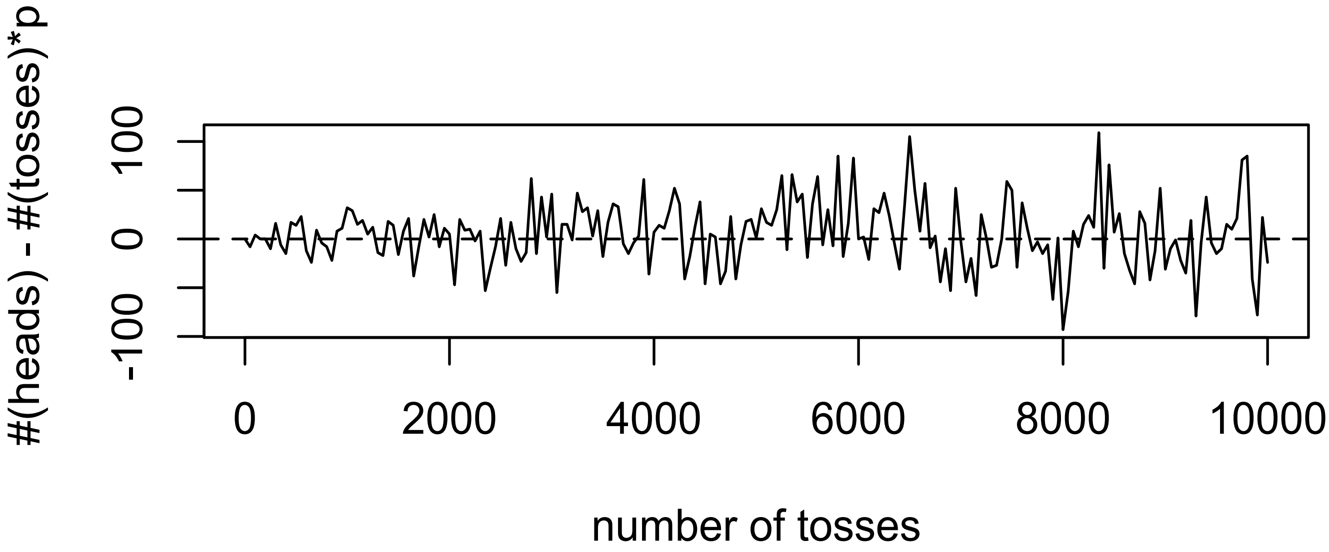


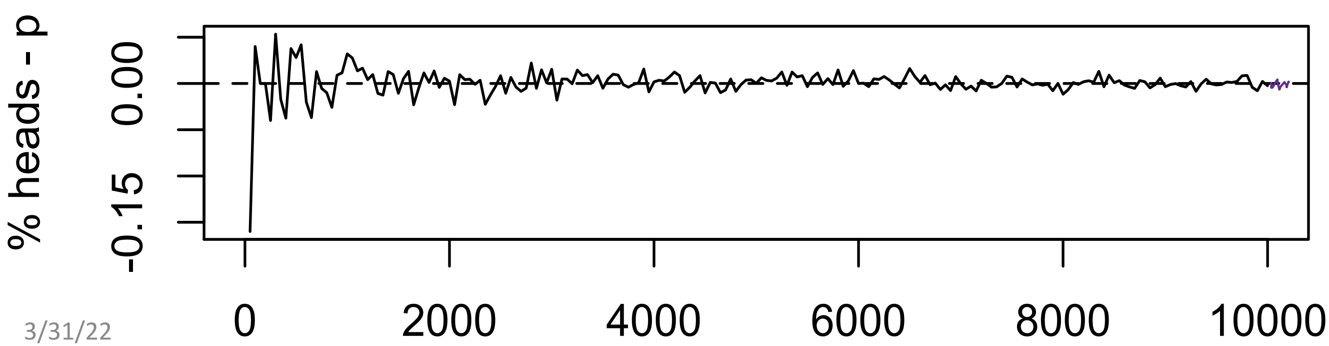
% error=
%H-0.5





30
#(heads) - #(tosses)*p
0
-30
0
200
400
600
800
1000
number of tosses
$\frac{\# \text{ of } H}{n} - 0.5$ (diff b/w %H & P(H))
% 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. $= \sqrt{n} \times \sigma$

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

  % error : $\sigma / \sqrt{n}$

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

$$\left(\frac{\text{sample count of H}}{n} - 0.5 \right) \to 0$$

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)

  0-1 r.v.

- Note that $E(A_n) = \mu$ and $SD(A_n) = \cancel{??}\ \frac{\sigma}{\sqrt{n}}$, $E(X_k) = \mu$, $SD(X_k) = \sigma$
- **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.
- For example, we double the accuracy by quadrupling the size.

using A_n as the estimator

$n \downarrow 4n$

New SD of A_n $= \frac{\sigma}{\sqrt{4n}} = \frac{\sigma}{\sqrt{n}} \cdot \frac{1}{2}$

Concentration of probability

$S_n = X_1 + X_2 + \cdots + X_n$, X_k are iid
$E(X_k) = \mu$
$SD(X_k) = \sigma$

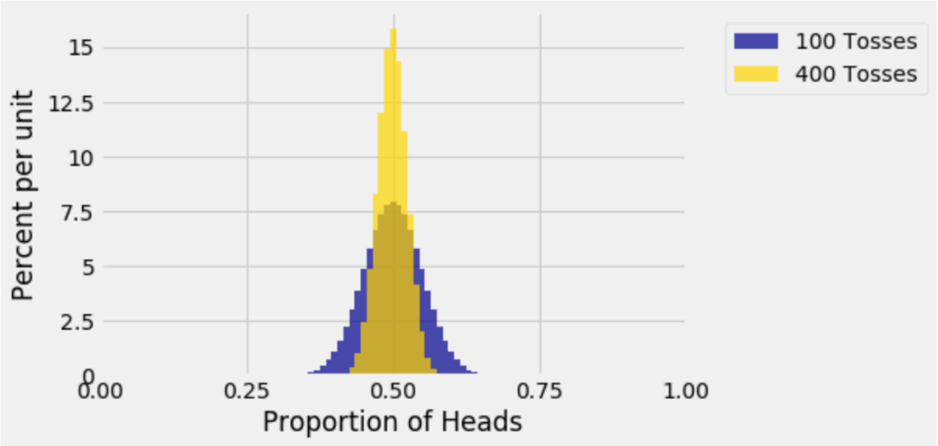
- This is when the SD decreases, so the probability mass accumulates around the mean, therefore, the larger the sample size, the more likely the values of the sample average $\bar{X}$ fall very close to the mean.

$= A_n = \frac{S_n}{n}$

- **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.

  get close to

  from data

- 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

$X \sim Bin(1300, 0.95)$

- b) Find $E(X)$ and $SD(X)$.

$E(X) = (1300)(0.95)$

$SD(X) = \sqrt{1300 \times 0.95 \times 0.05}$

- c) Find the chance that more than 1250 students get a good estimate.

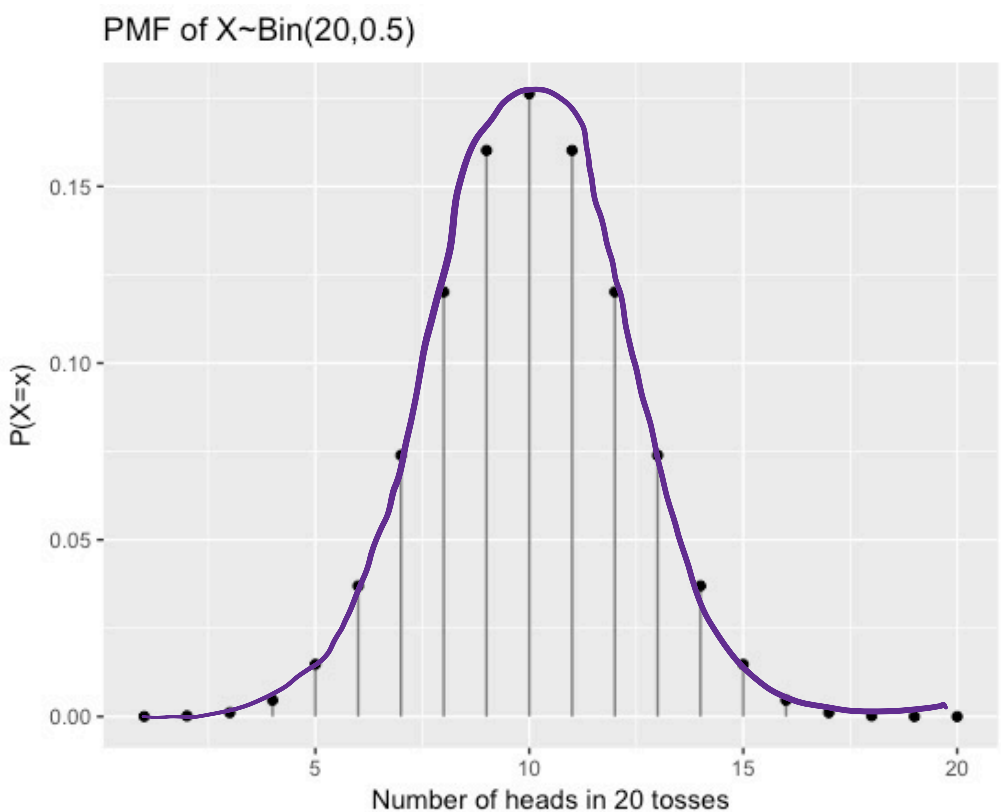
$$\sum_{k=1251}^{1300} f(k) = \sum_{k=1251}^{1300} \binom{1300}{k} (0.95)^k (0.05)^{1300-k}$$

8.1: Distribution of a sample sum

S_n

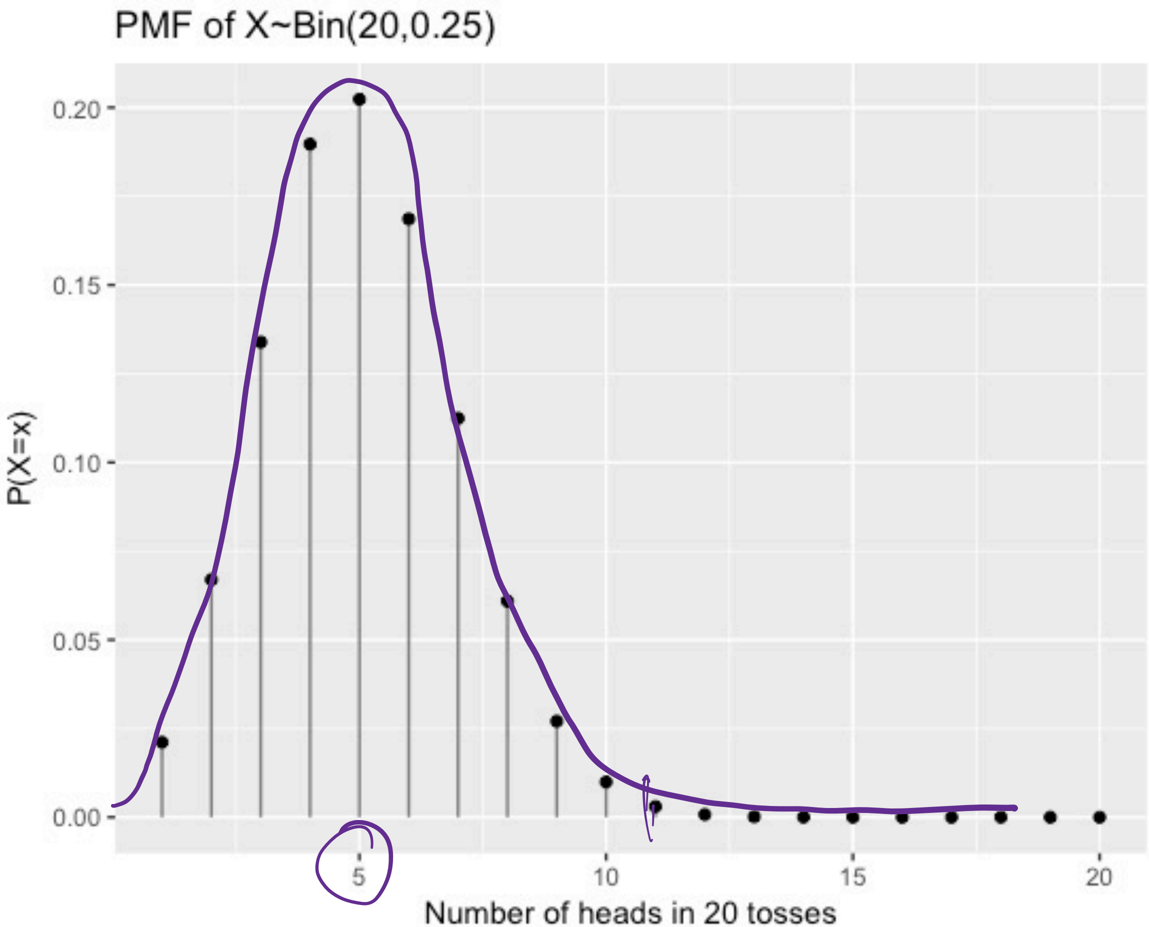
- We can consider $X \sim Bin(20, 0.5)$ as the sum of 20 Bernoulli iid rvs.

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



Visualizing the prob. mass function (pmf)

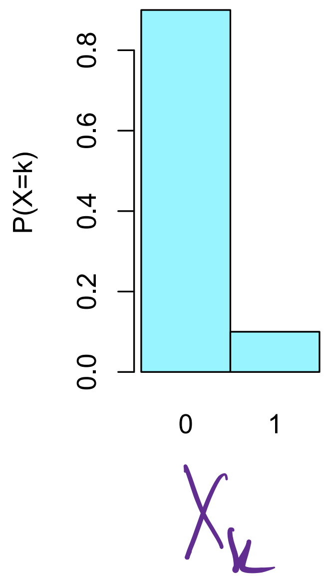
$p = 0.25$

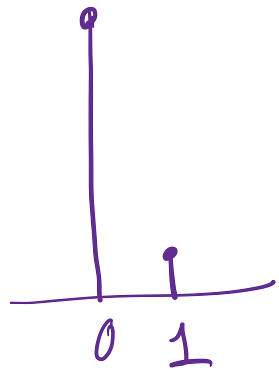


What if p is small?

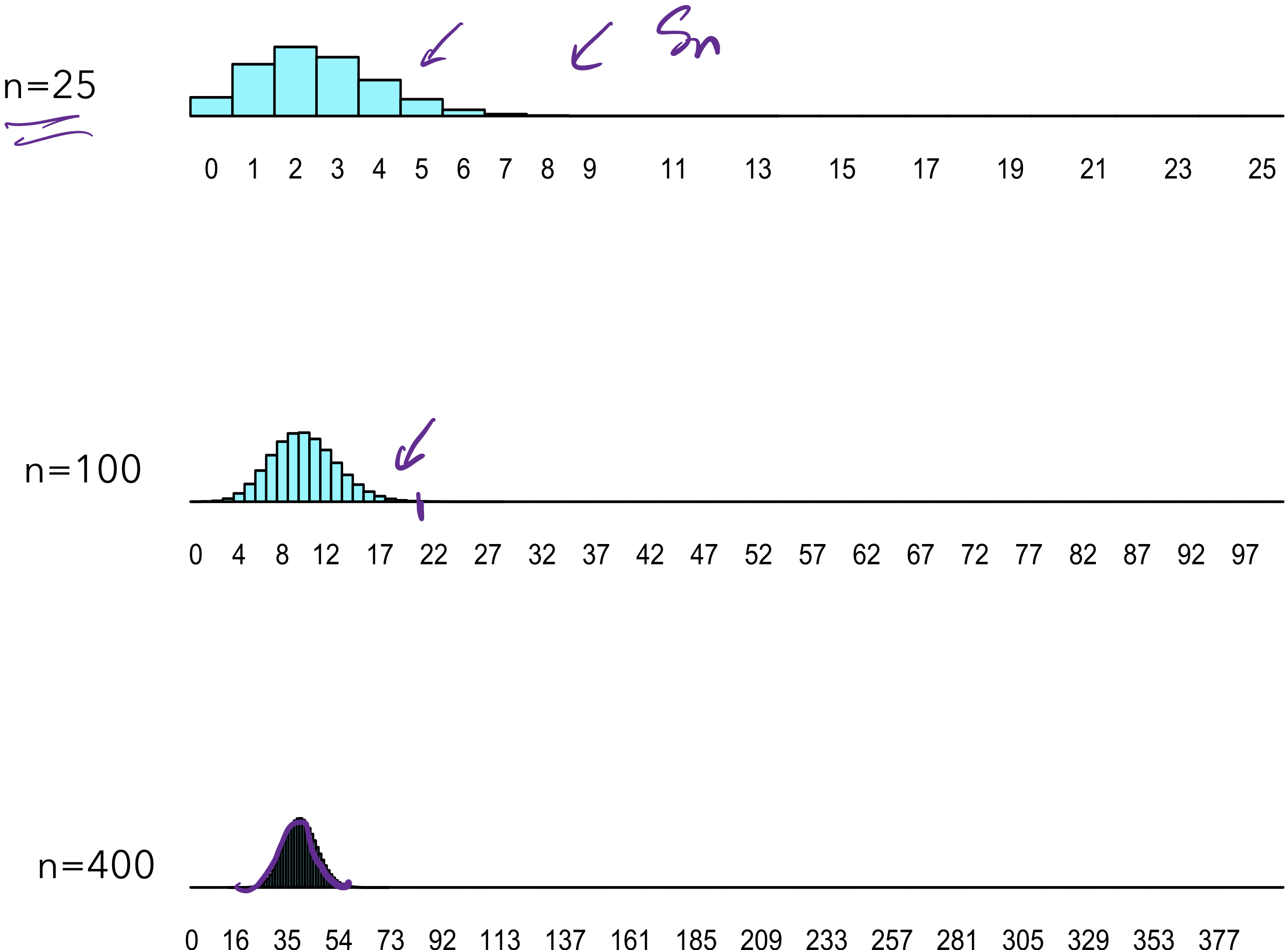
- Consider $X_k \sim Bernoulli\left(\frac{1}{10}\right), S_n = X_1 + X_2 + X_3 + \cdots + X_n, Sn \sim Bin(n, \frac{1}{10})$

- Draw the probability histogram for X_k:





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



Distribution of the sample sum

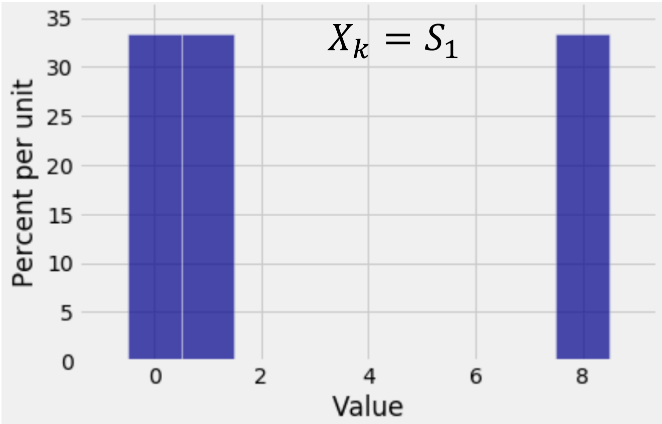
- More generally, let's consider $X_1, X_2, \dots, X_n$ iid with mean μ and SD σ
- Let $S_n = X_1 + X_2 + \cdots + X_n$
- We know that $E(S_n) = n\mu$ and $SD(S_n) = \sqrt{n}\sigma$

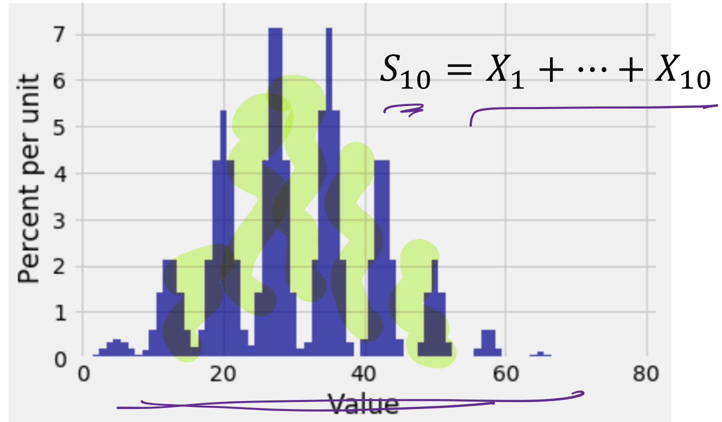
- We want to say something about the distribution of S_n, and while it may be possible to write it out analytically, if we know the distributions of the X_k, it may not be easy. And we may not even know anything beyond the fact that the X_k are iid, and we might be able to guess at their mean and SD.

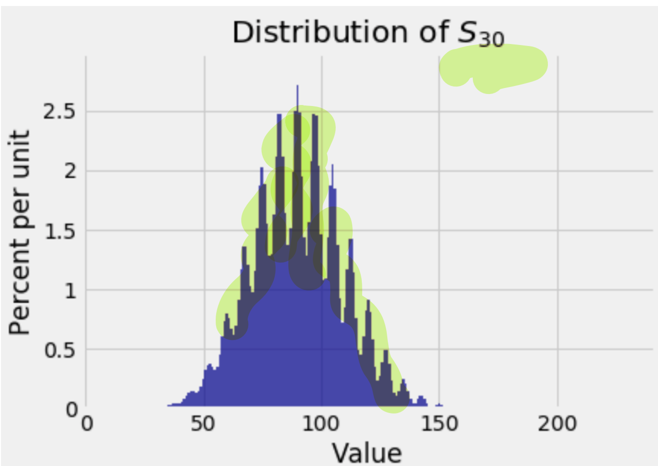
- We saw in the previous slides that even if the X_k are very far from symmetric, the distribution of the sum begins to look quite nice and bell shaped.

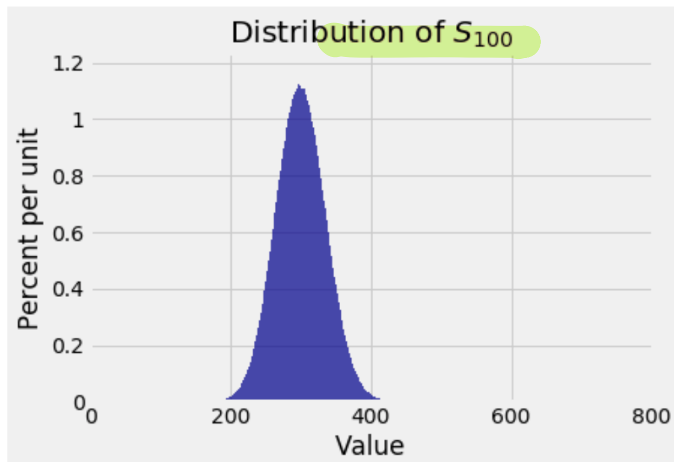
- What if the X_k are strange looking?

Weird X_k distributions – is the distribution of S_n different?









From section 8.1

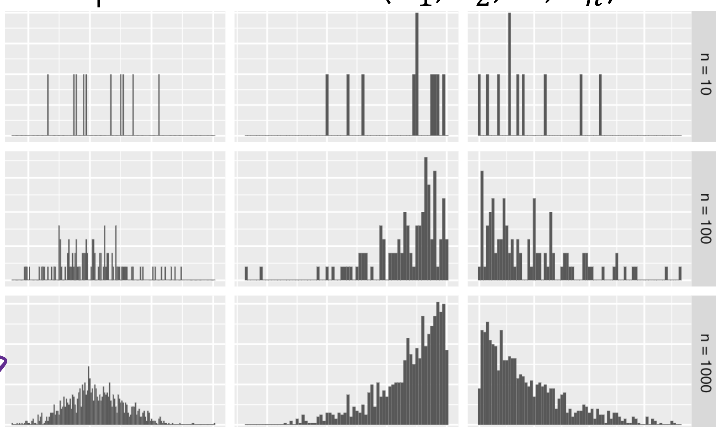
Examples by picture

Probability distribution of X_k

A B C

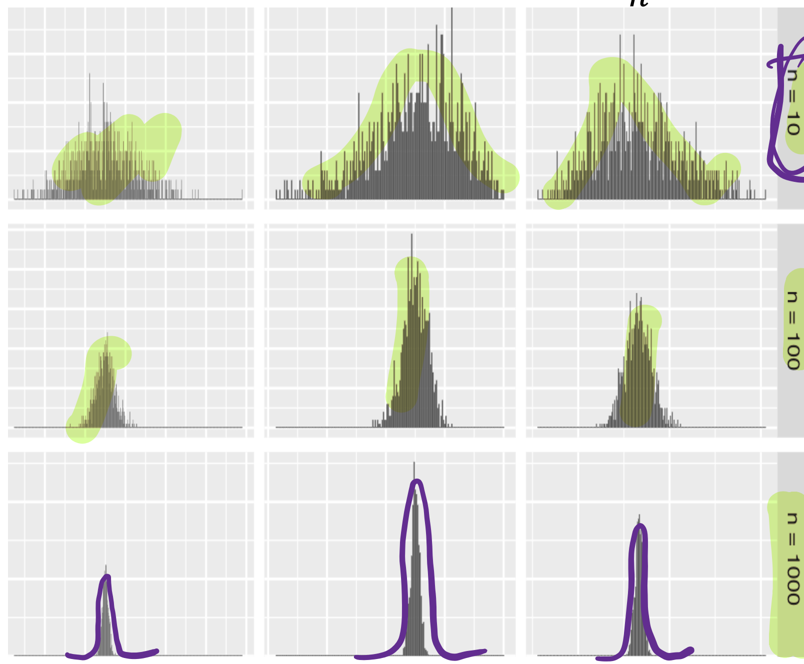
A B C

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



A B C

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



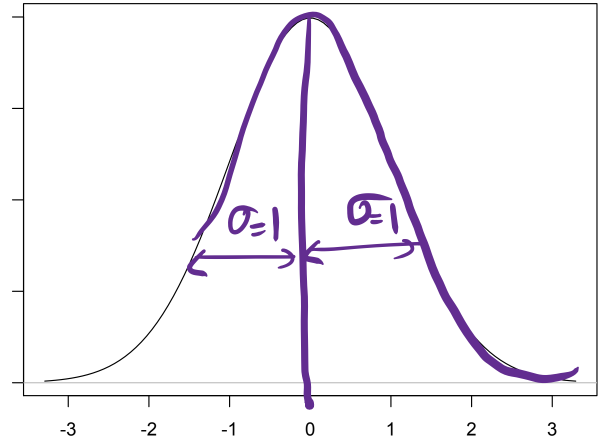
Graphs created by Sarah Johnson for Stat 20

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

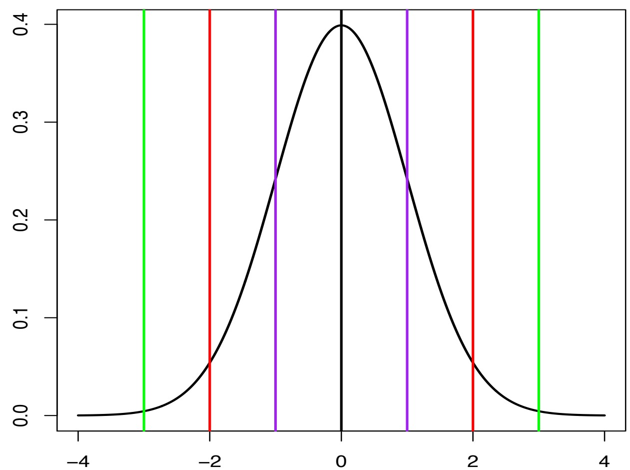
Bell curve: the Standard Normal Curve

- 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 $|z|>3$

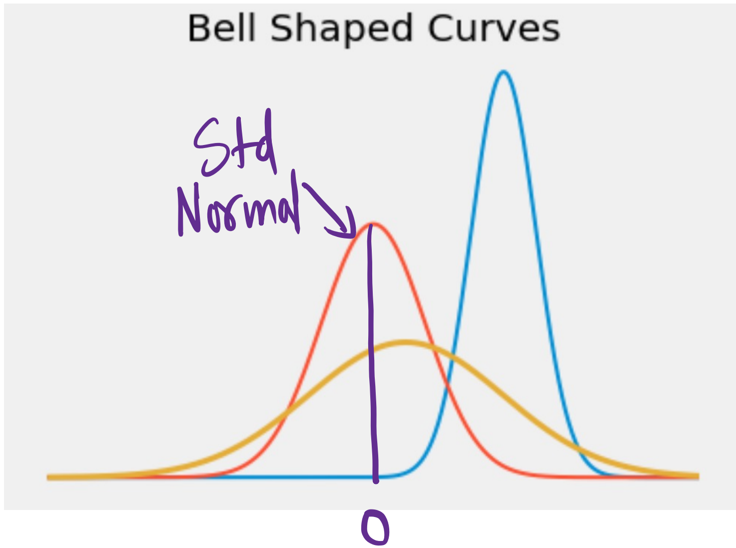


$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, -\infty < z < \infty$$

$$\int_{-\infty}^{\infty} \phi(z)\,dz = 1$$



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.