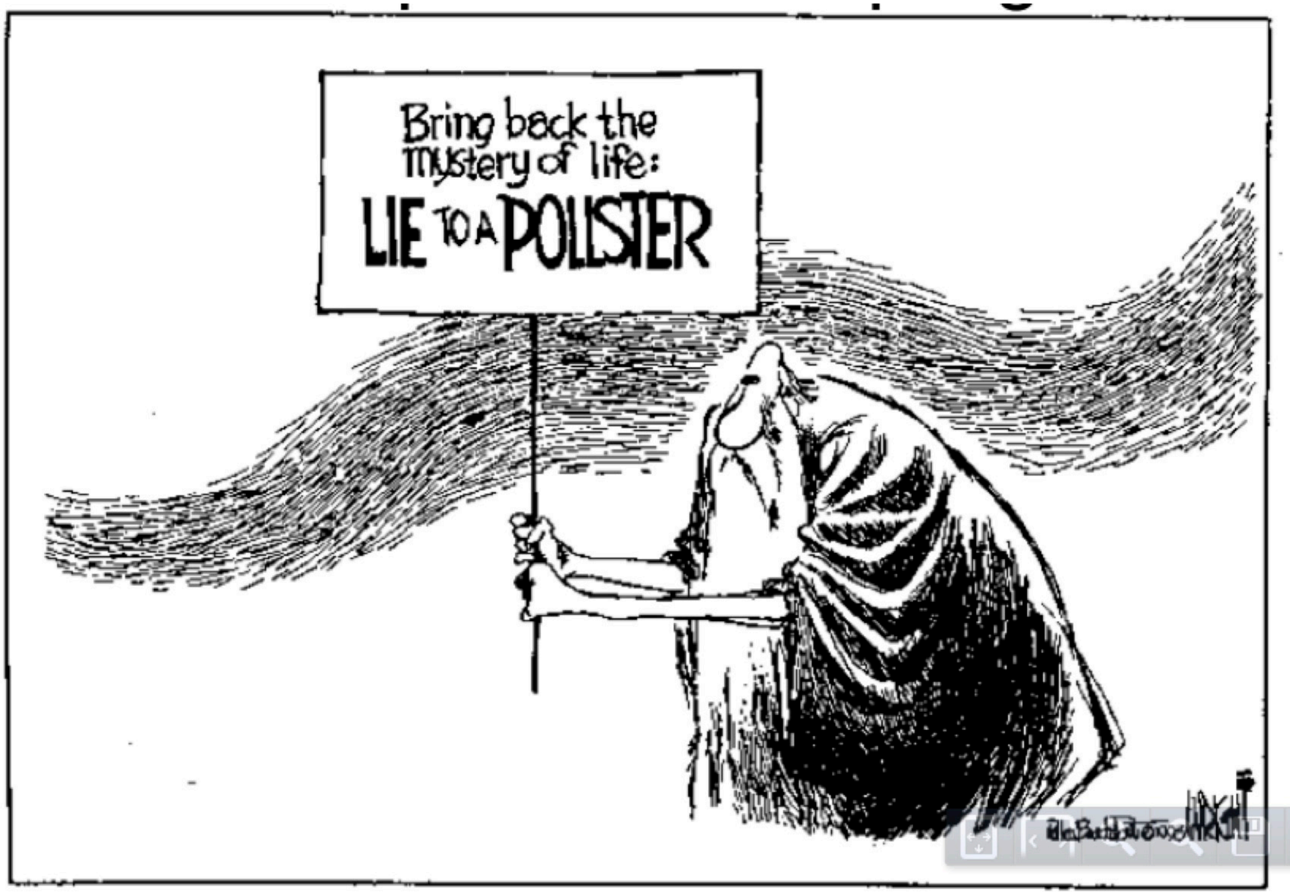
Stat 88: Prob. & Math. Statistics in Data Science



Lecture 18: 3/29/2022

Sampling without replacement, the law of averages, distribution of a sample sum

7.2, 7.3, 8.1

Review: variance, SD, inequalities when we don't know dsn

1. The variance of a rv X: $Var(X) = \sigma^2 = E[(X - E(X))^2] = E(X^2) - \left(E(X)\right)^2$
2. The SD or standard deviation is given by $SD(X) = \sigma = \sqrt{Var(X)} \geq 0$

If X and Y are ***independent***, then

3. $Var(X + Y) = Var(X) + Var(Y)$ & $SD(X + Y) = \sqrt{Var(X) + Var(Y)}$
4. **Markov's Inequality**: For a nonnegative rv X, and constant $c > 0$

$$P(X \geq c) \leq \frac{E(X)}{c}$$

5. **Chebyshev's inequality**: For any random variable X (not necessarily non-negative), with mean μ and standard deviation σ, for any positive constant $c > 0$, we have:

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2} = \frac{Var(X)}{c^2}$$

Example

A list of non negative numbers has an average of 1 and an SD of 2. Let p be the proportion of numbers over 4. To get an upper bound for p, you should:

a) Assume a binomial distribution

b) Use Markov’s inequality.

c) Use Chebyshev’s inequality

d) None of the above.

Recap: sums of iid random variables

- Let $X_1, X_2, \dots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Define $S_n = X_1 + X_2 + \cdots + X_n$
- $E(S_n) = \sum E(X_k) = n\mu$ & $Var(S_n) = Var(X_1) + \cdots + Var(X_n) = n\sigma^2$
- $SD(S_n) = \sqrt{n}\,\sigma$ (Square root law: sd grows by a factor of $\sqrt{n}$, **not** n)

- $X \sim Bin(n, p)$: $E(X) = np, Var(X) = np(1-p), SD(X) = \sqrt{np(1-p)}$

- $X\sim$ Poisson(μ): $E(X) = Var(X) = \mu, SD = \sqrt{\mu}$

- $X\sim$ Geometric(p) distribution: $E(X) = \frac{1}{p}$, $Var(X) = \frac{1-p}{p^2}$

7.3: Sampling without replacement

- When we have a simple random sample (SRS), the draws are without replacement (like drawing cards from a deck).
- The random variables are no longer independent
- So, how do we compute the variance of the sum of draws of a SRS?
- To begin with, let's look at the squares and products of indicators
- If I_A and I_B are indicator functions, what can we say about I_A^2 and $I_A I_B$?

Variance of a hypergeometric random variable

- Let $X \sim HG(N, G, n)$, then can write $X = I_1 + I_2 + \cdots + I_n$, where I_k is the indicator of the event that the kth draw is good.

- We can compute the expectation of X using symmetry: $E(X) = \frac{nG}{N}$
- But what about variance?
- Since the indicators are not independent, we can't just add the variances
- Let's just use the formula: $Var(X) = E(X^2) - \left(E(X)\right)^2 = E(X^2) - \left(\frac{nG}{N}\right)^2$

$$X^2 = (I_1 + I_2 + \cdots + I_n)^2 = \sum_{k=1}^{n} I_k^2 + \sum_{j} \sum_{k, k \neq j} I_j I_k$$

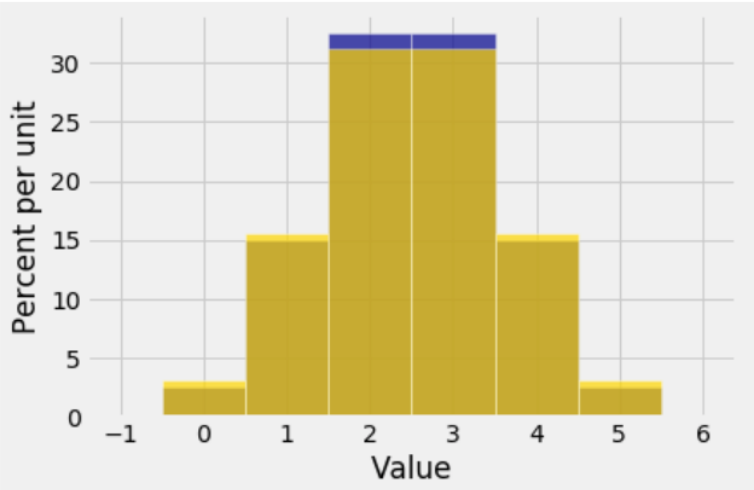
$$E(X^2) = nE\left(I_k^2\right) + n(n-1)E\left(I_j I_k\right)$$

$$= n\frac{G}{N} + n(n-1)P\left(I_j = 1\right)P\left(\,I_k = 1 \mid I_j = 1\,\right)$$

$$E(X^2) = n\frac{G}{N} + n(n-1)\,\frac{G}{N} \cdot \frac{G-1}{N-1}$$

Variance of a hypergeometric random variable

The finite population correction & the accuracy of SRS



From section 7.2

fpc = $\sqrt{\dfrac{N-n}{N-1}}$

Note that fpc ≤ 1

So $SD(HG) \leq SD(Bin)$

SD is less when we don't have independence

In general we have that the :

SD of sum of an SRS = SD of sum WITH repl. $\times$ fpc

Accuracy of samples

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.

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.

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.

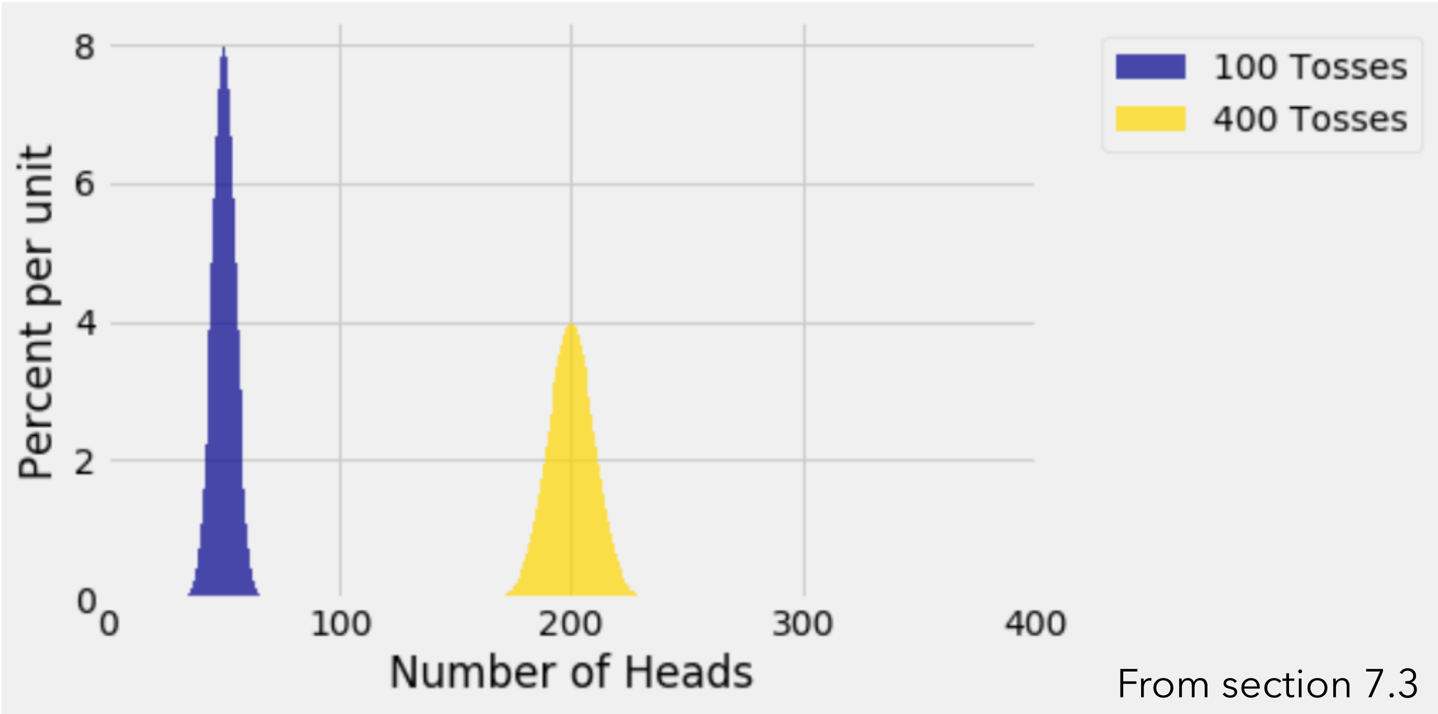
- Recall the ***sample sum*** $S_n = X_1 + X_2 + \cdots + X_n$, with $E(S_n) = n\mu$, $Var(S_n) = n\sigma^2, SD(X_n) = \sqrt{n}\sigma$

- 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 ($n\mu$).

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?



From section 7.3

Example: Coin toss

- $SD(S_{100}) =$
- $SD(S_{400}) =$

- P(200 H in 400 tosses)

- P(50 H in 100 tosses)

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 $SD(A_n) = ??$

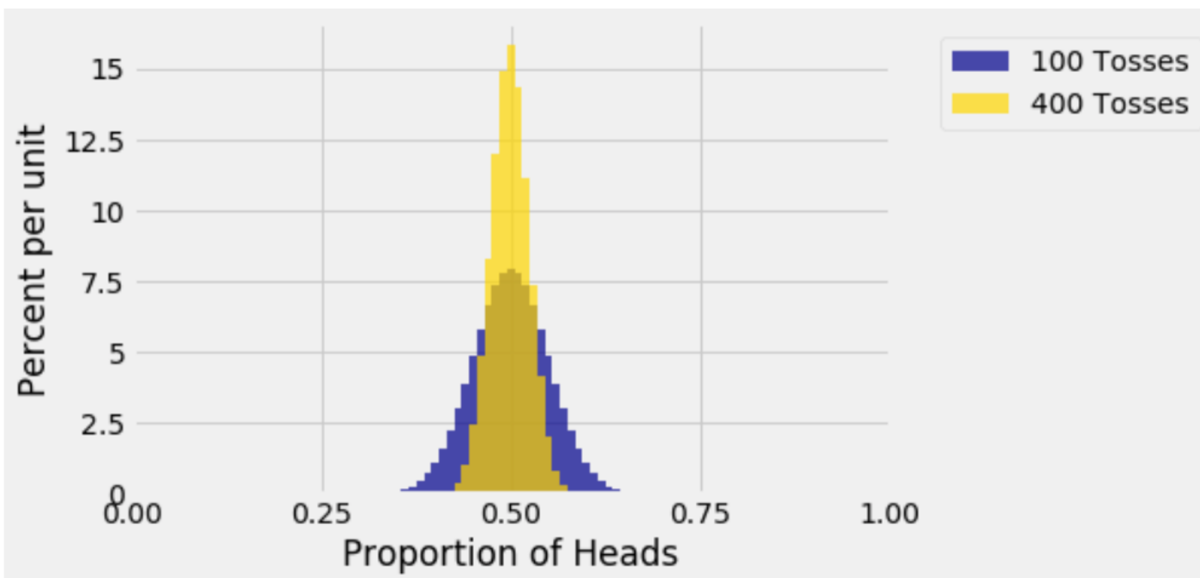
- **The square root law:** the *accuracy* of an estimator is measured by its SD, the ***smaller*** the SD, the ***more accurate*** the estimator, but if you multiply the sample size by a factor, the accuracy only goes up by the **square root** of the factor.
- In our earlier example, we _____ the accuracy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

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

$|\ A_n - \mu\ |$ 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.

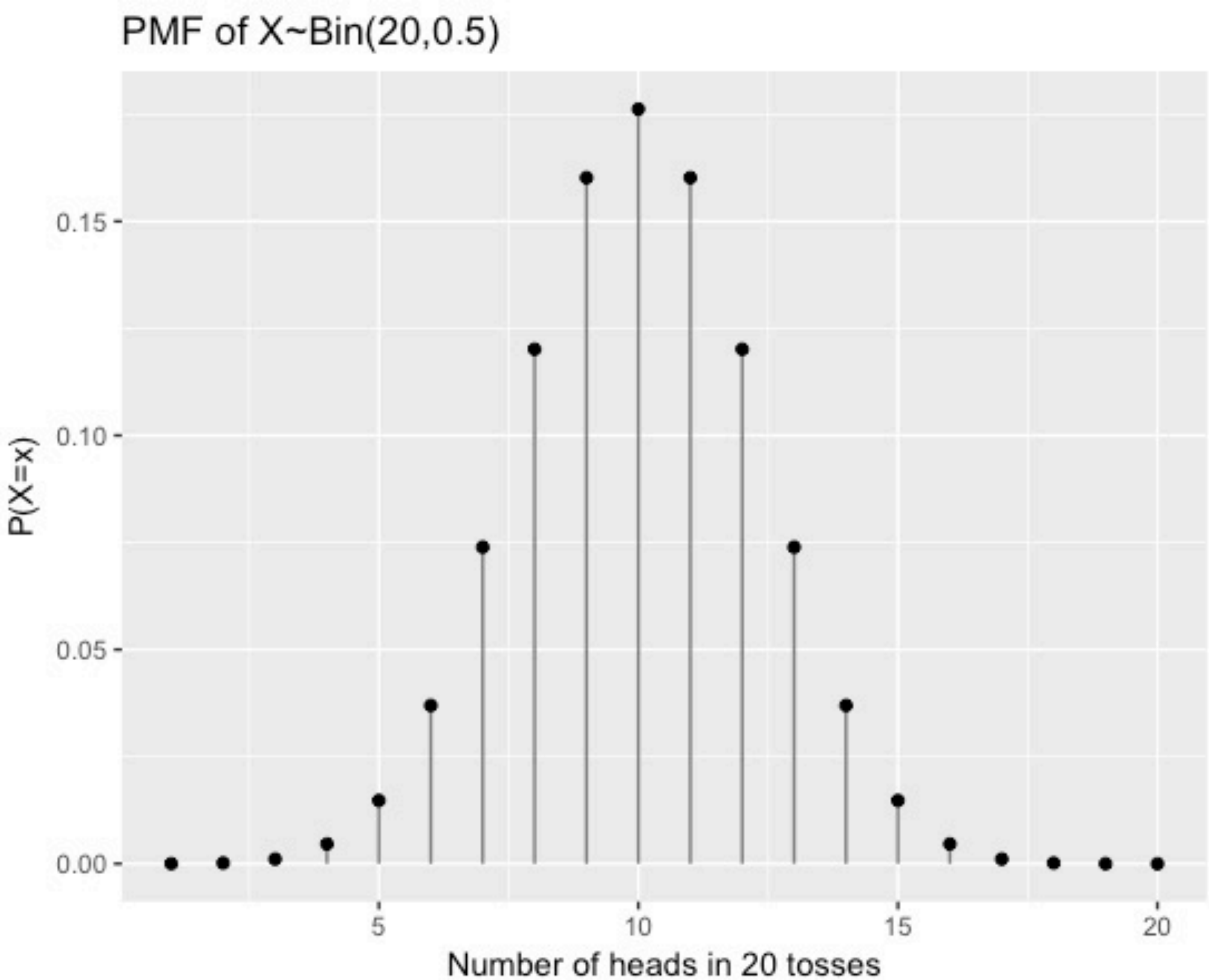
- The *percentage* of sixes, when rolling a fair die over and over, is very close to 1/6. True for any of the faces, so the *empirical* histogram of the results of rolling a die over and over again looks more and more like the *theoretical* probability histogram.

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

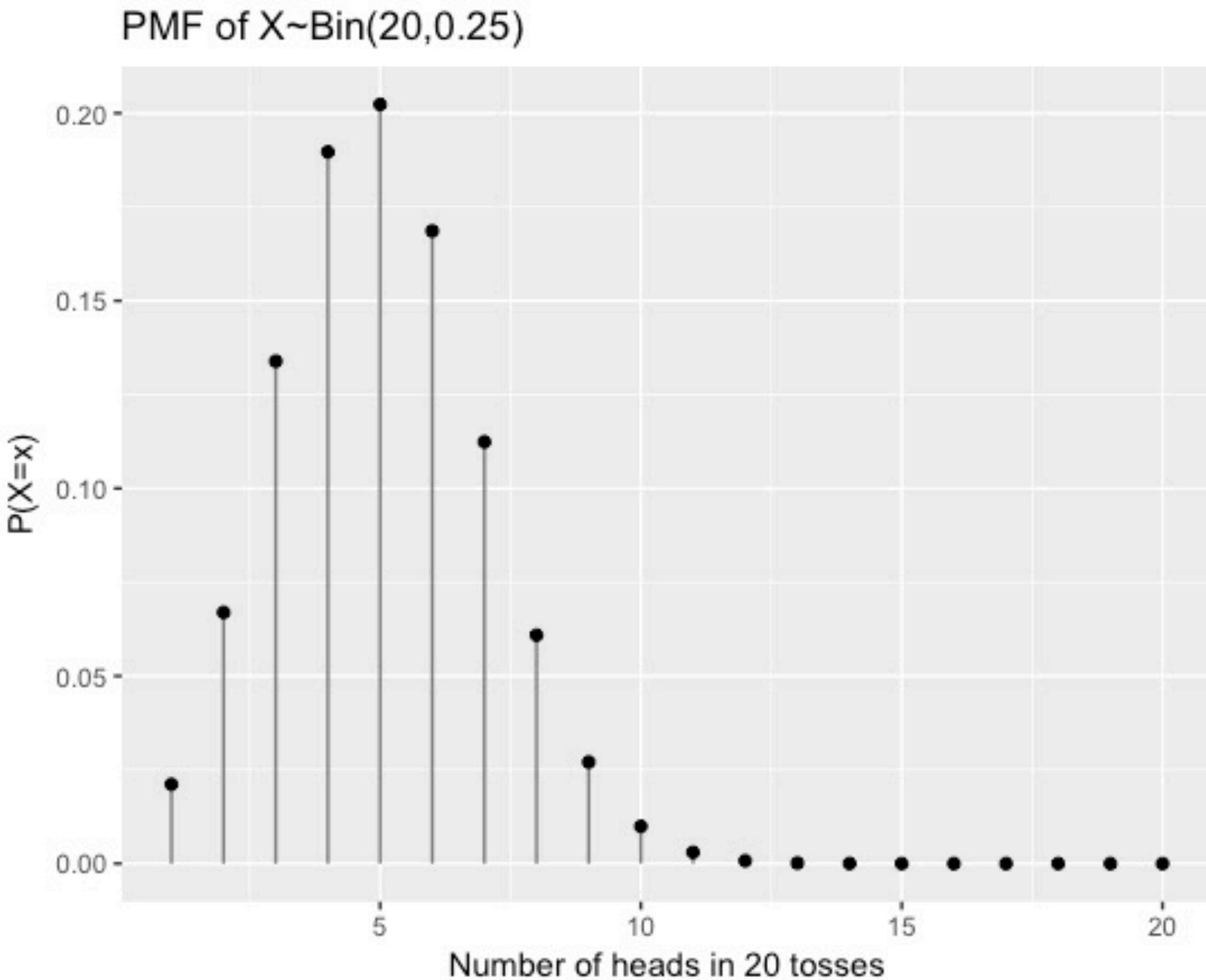
8.1: Distribution of a sample sum

- 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:



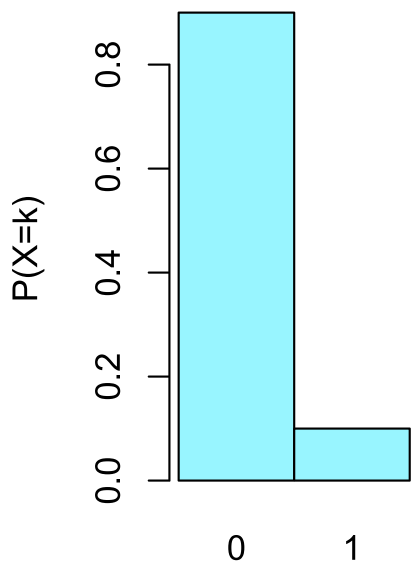
Visualizing the prob. mass function (pmf)



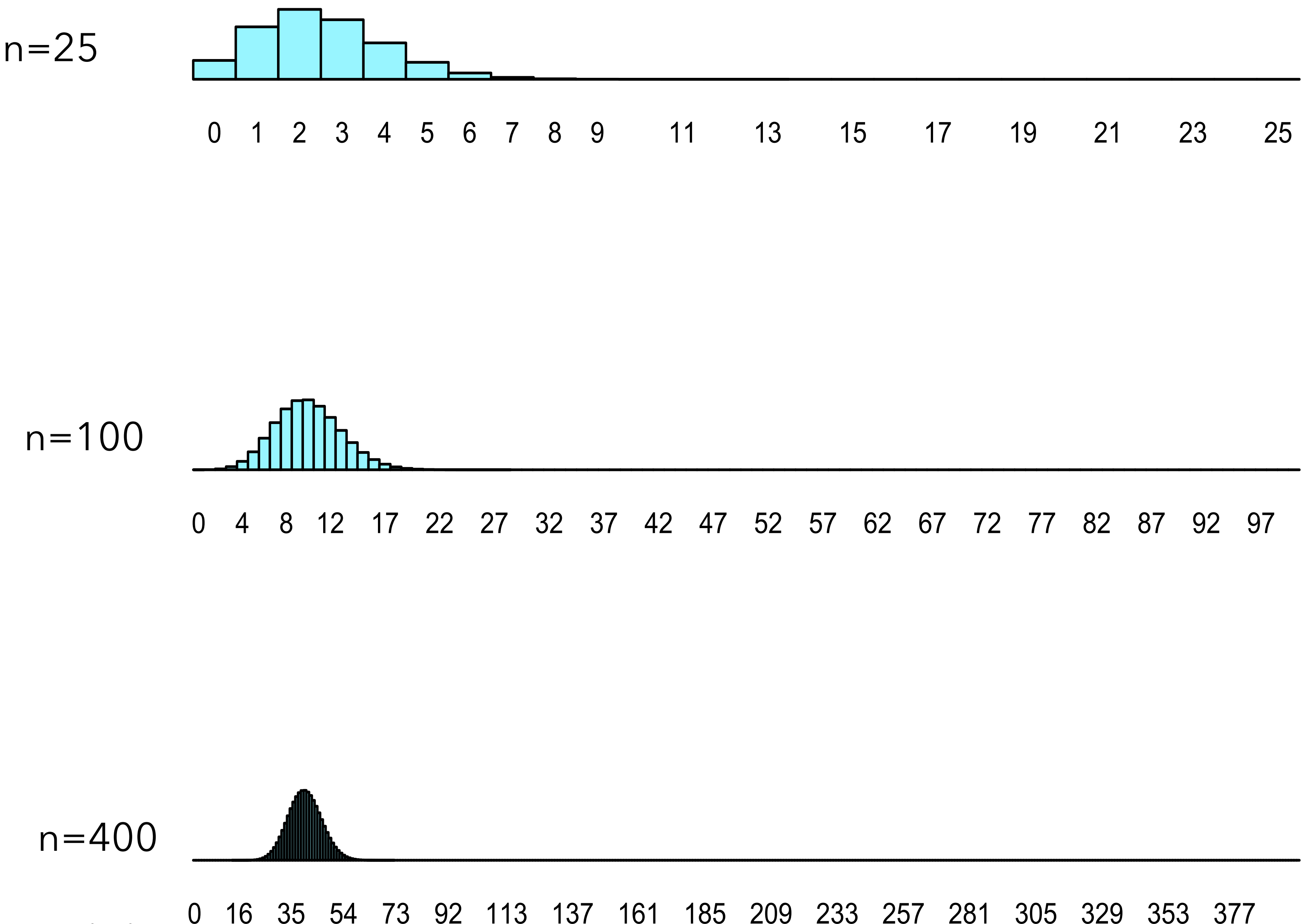
What if p is small?

- 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 from *Statistics* by Freedman, Pisani, and Purves)



Distribution of the sample sum

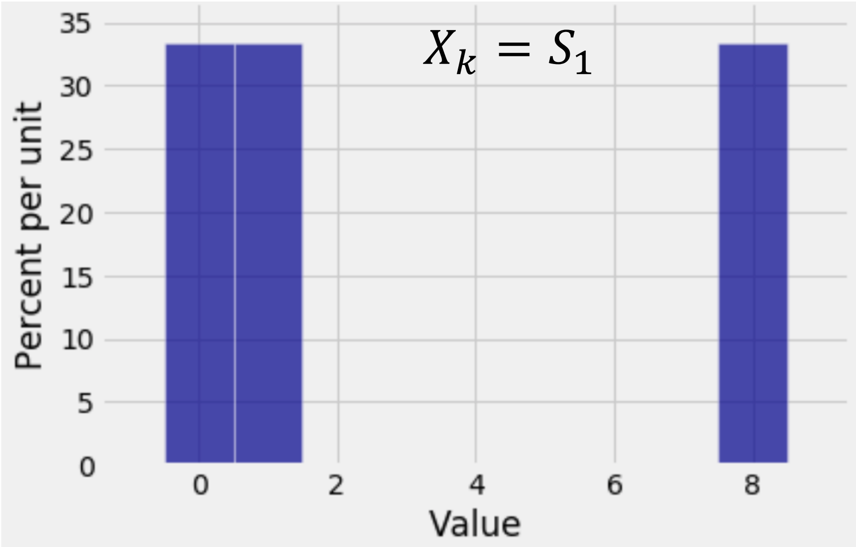
- More generally, let's consider $X_1, X_2, \ldots, X_n$ iid with mean μ and SD σ
- Let $S_n = X_1 + X_2 + \cdots + X_n$
- We know that $E(S_n) = n\mu$ and $SD(S_n) = \sqrt{n}\sigma$

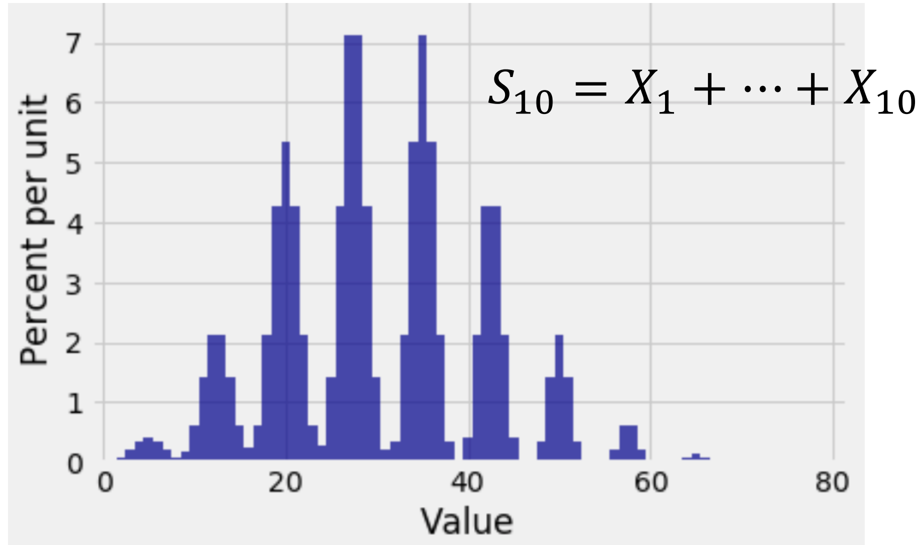
- We want to say something about the distribution of S_n, and while it may be possible to write it out analytically, if we know the distributions of the X_k, it may not be easy. And we may not even know anything beyond the fact that the X_k are iid, and we might be able to guess at their mean and SD.

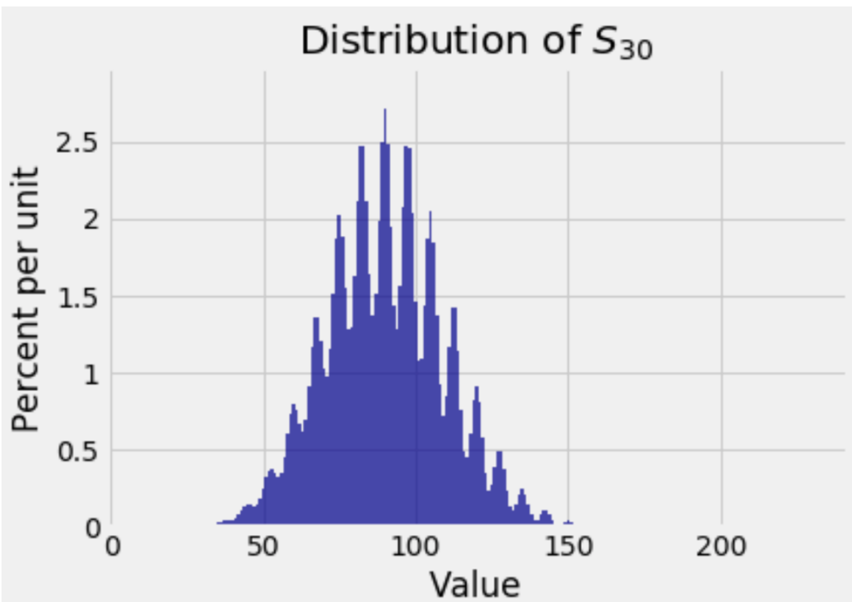
- We saw in the previous slides that even if the X_k are very far from symmetric, the distribution of the sum begins to look quite nice and bell shaped.

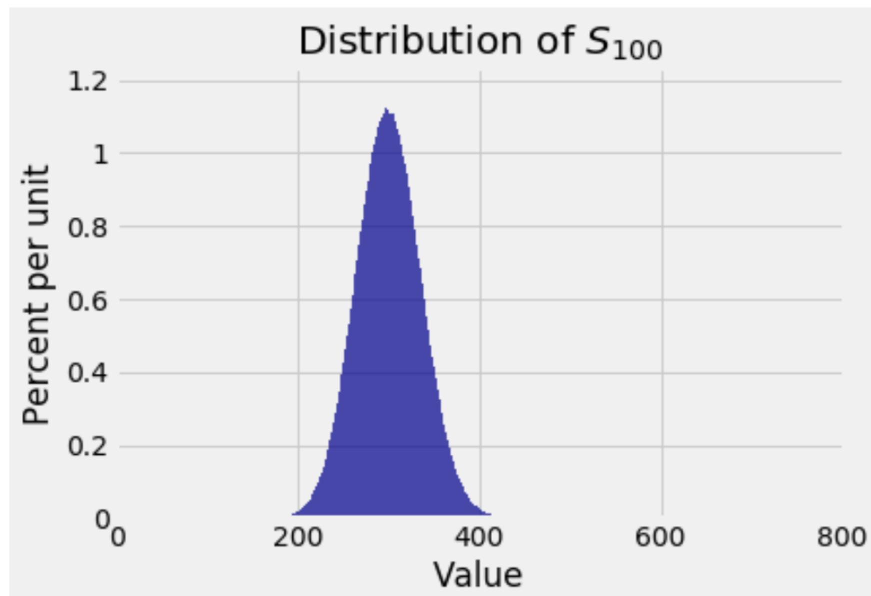
- What if the X_k are strange looking?

Weird X_k distributions – is the distribution of S_n different?





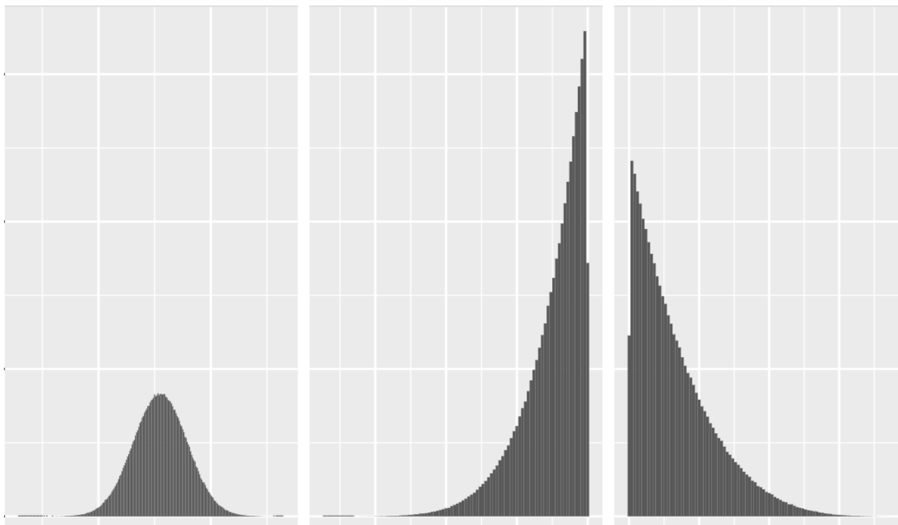




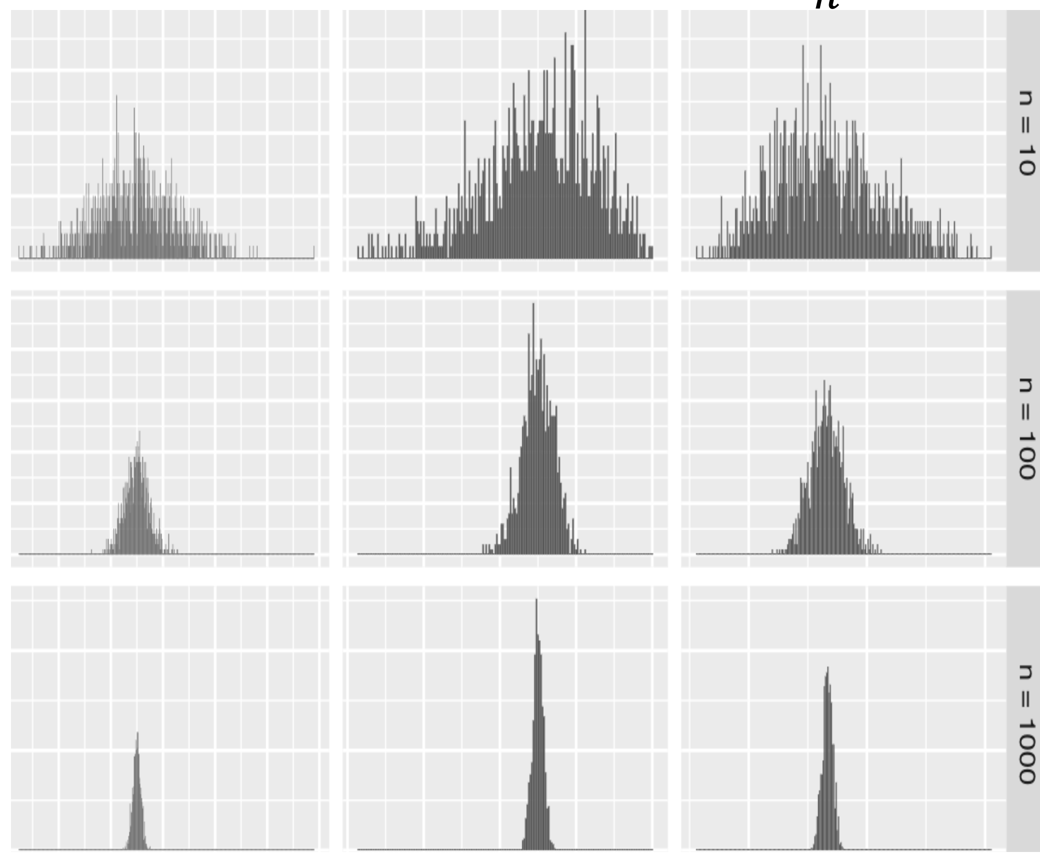
From section 8.1

Examples by picture

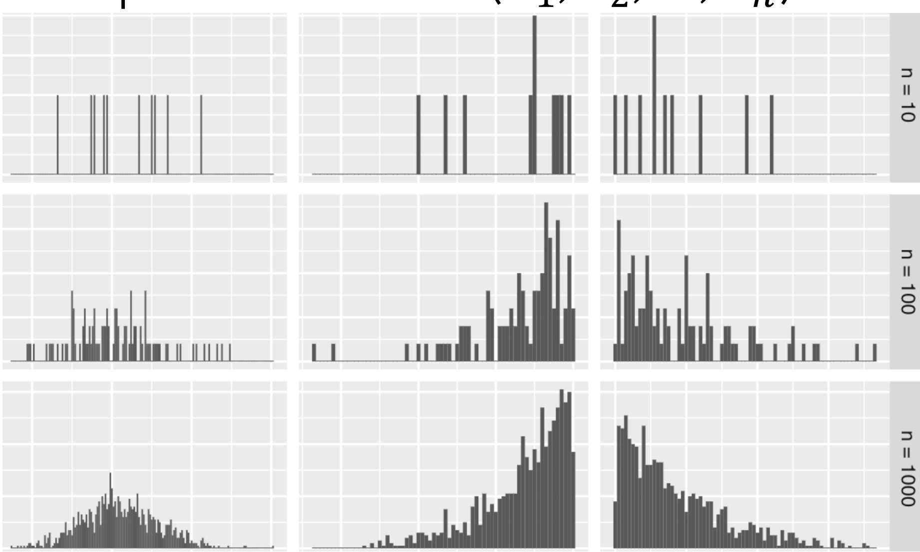
Probability distribution of X_k



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



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



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