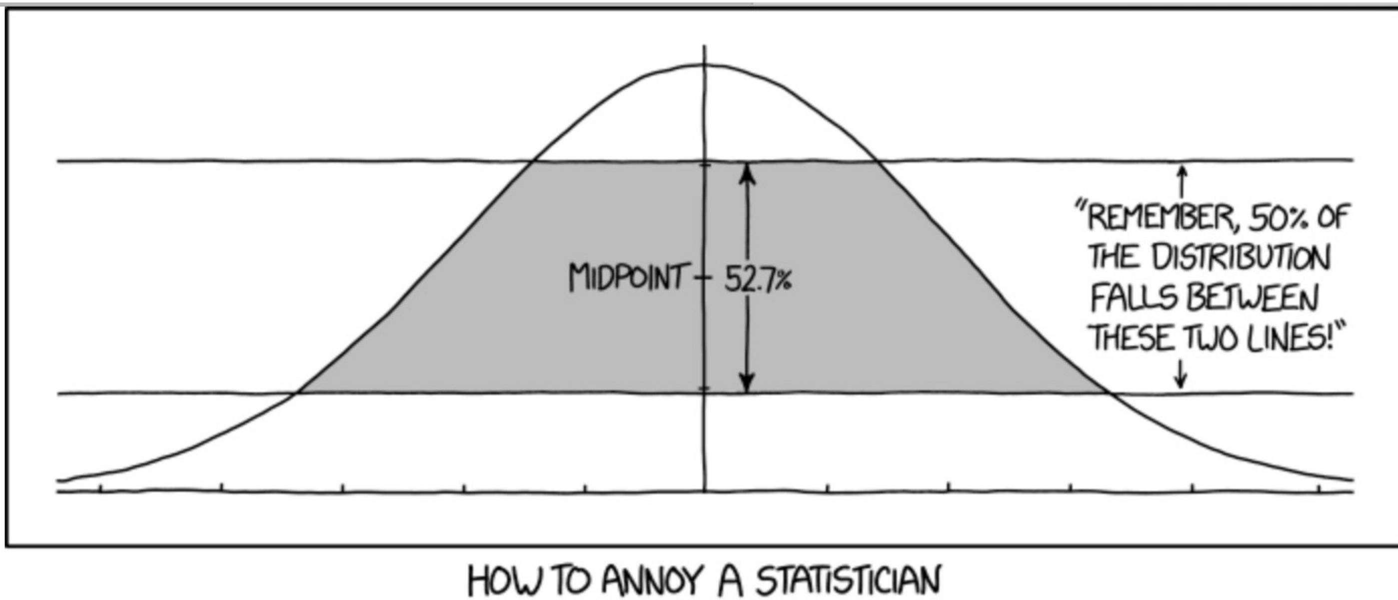
Stat 88: Probability & Mathematical Statistics in Data Science



xkcd.com/2118

Lecture 26: 3/29/2021

Sections 8.1, 8.2

The Central Limit Theorem

Recall: expected value and SD of the Sample sum, sample average, and the square root law

- $S_n = X_1 + X_2 + \cdots + X_n$ iid. $E(X_i) = \mu$, $Var(X_k) = \sigma^2$
- $\boldsymbol{E(S_n) = n\mu}$ and $\boldsymbol{SD(S_n) = \sqrt{n}\sigma}$
- Let $A_n = {}^{S_n}/_{n}$, so A_n is the average of the sample (or sample mean).

  why is the sample mean a random variable?
- If the X_k are indicators, then A_n is a proportion (proportion of successes)

  σ is a constant
- Note that $\boldsymbol{E(A_n) = \mu}$ and $\boldsymbol{SD(A_n) = \sigma/\sqrt{n}}$

  $\left.\frac{\sigma}{\sqrt{n}}\right\}$ want this to be halved, need to take new sample size $=4n$

  $\frac{\sigma}{\sqrt{4n}} = \frac{\sigma}{2\sqrt{n}}$
- **The square root law:** the *accuracy* of an estimator is measured by its SD.
- The ***smaller*** the SD, the ***more accurate*** the estimator (sample mean), but if you multiply the sample size by a factor, the accuracy only goes up by the **square root** of the factor.

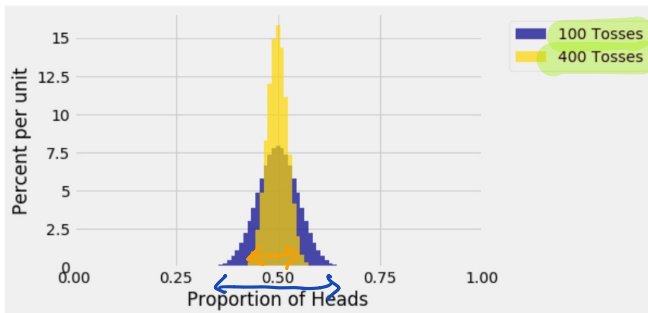
Concentration of probability & WLLN

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

$\frac{\sigma}{\sqrt{n}}$ ← increasing n decreases SD

- **Weak Law of Large numbers:**

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



- *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 $\quad X \sim Bin(n=1300,\ p=0.95)$
- b) Find $E(X)$ and $SD(X)$. $\quad E(X) = 1300(0.95)$, $\quad SD(X) = \sqrt{1300(0.95)(0.05)}$
- c) Find the chance that more than 1250 students get a good estimate.

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

$$1 - P(X \le 1250)$$

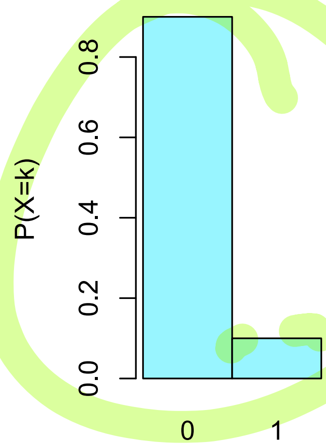
$$= 1 - F(1250) \quad (F \text{ is bin. cdf})$$

What about the *distribution* of the sample sum and mean?

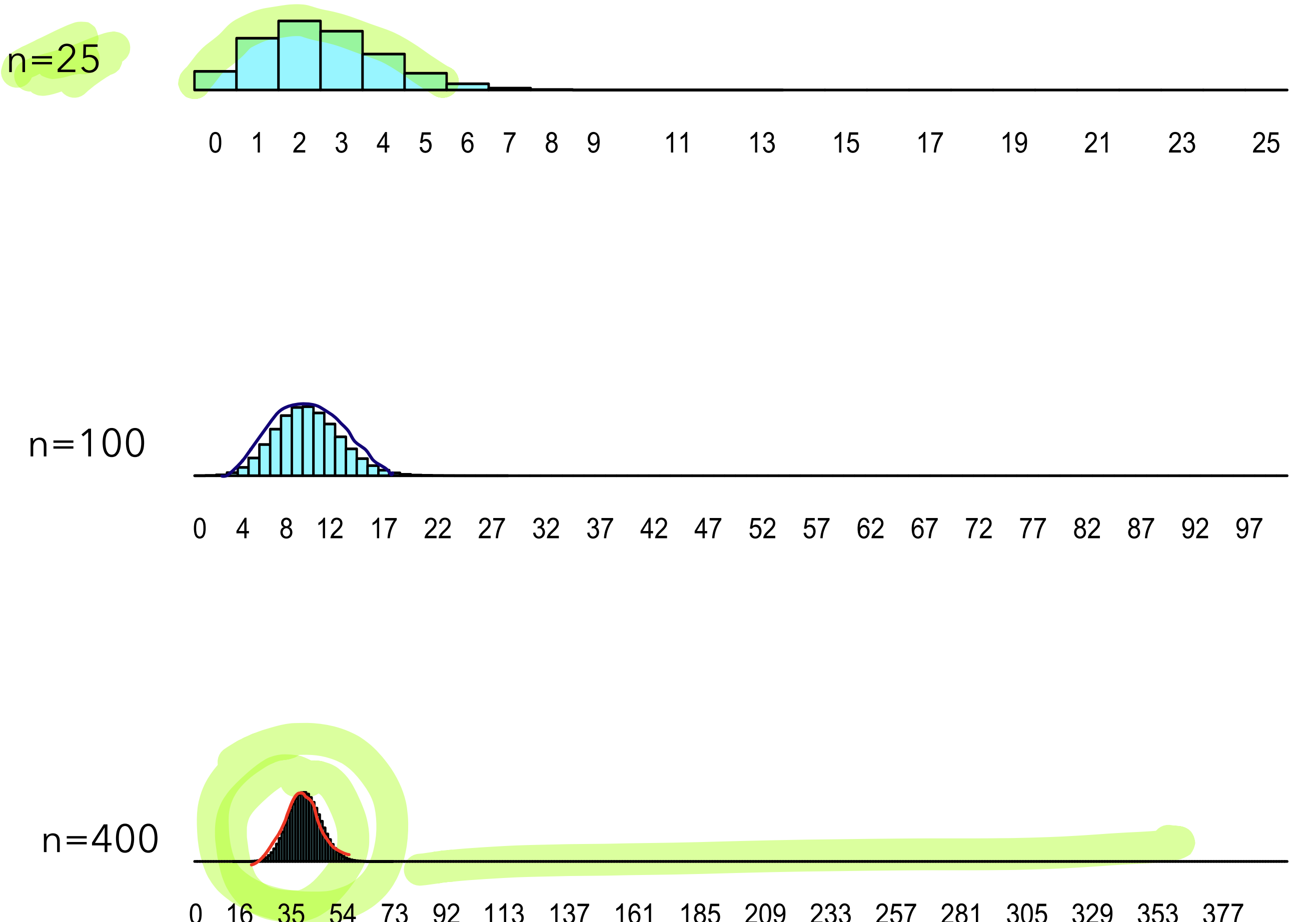
- Can we say something about the distribution of the sample sum and sample mean?
- Not just the expectation and standard deviation, but the probabilities themselves.

- 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:

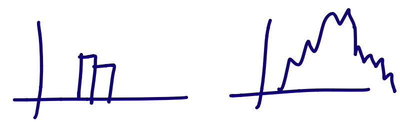


Probability histogram for binomial rv, p=0.1



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

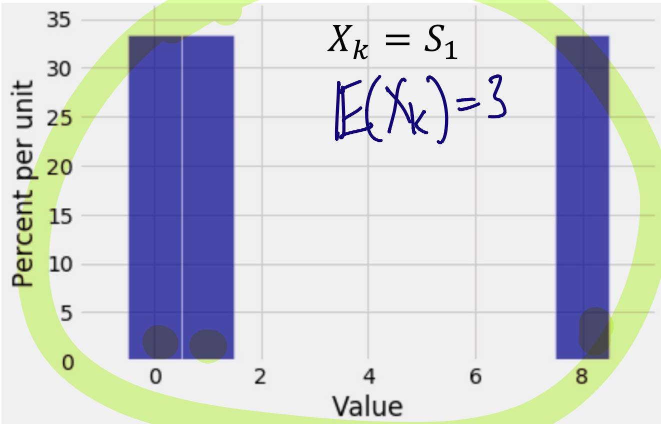


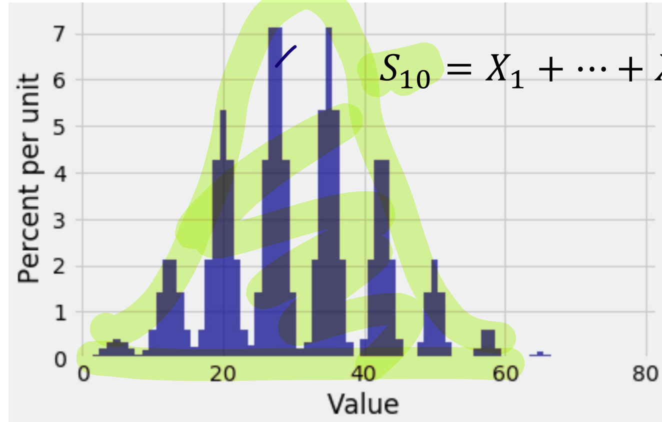
- We want to say something about the distribution of S_n, and while it may be possible to write it out analytically, if we know the distributions of the X_k, it may not be easy. And we may not even know anything beyond the fact that the X_k are iid, and we might be able to guess at their mean and SD.

- 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.

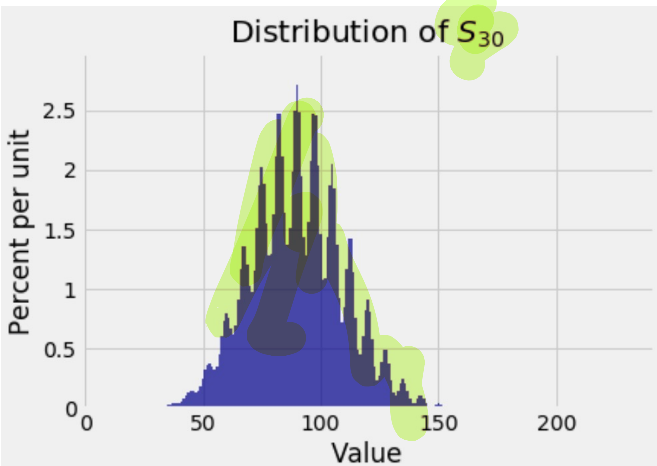
- What if the X_k are strange looking?

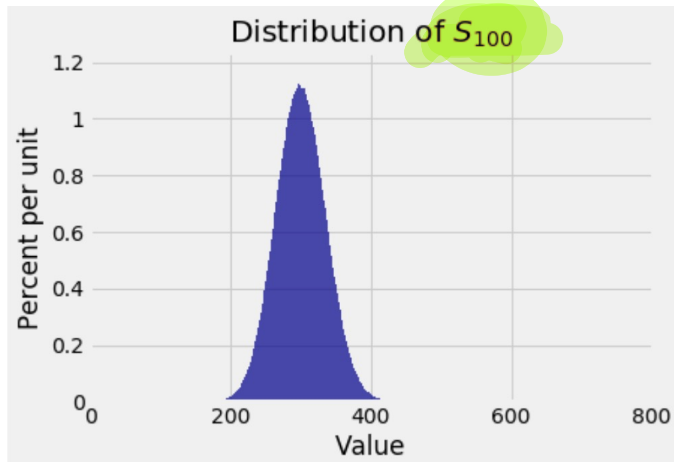
Weird X_k distributions – is the distribution of S_n different?





$S_{10} = X_1 + \cdots + X_{10}$





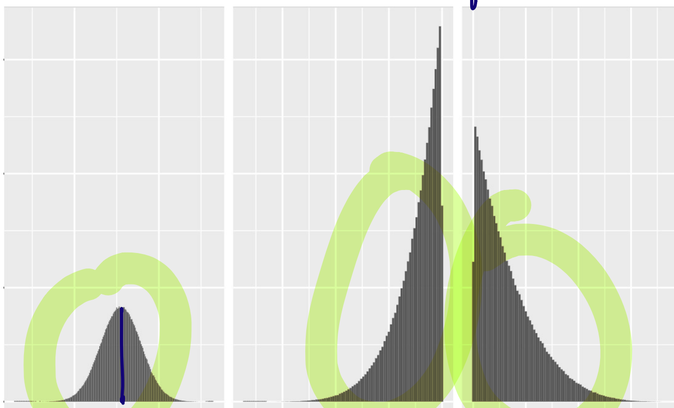
The Central Limit Theorem

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

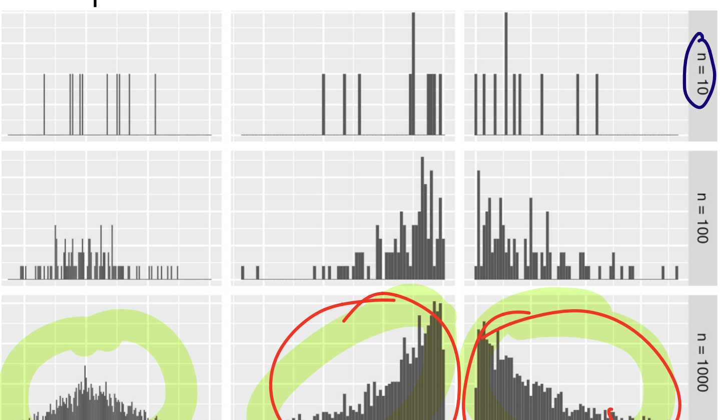
If sample sum is approx normal, so is the sample mean because $A_n = \frac{S_n}{n}$

Examples by picture

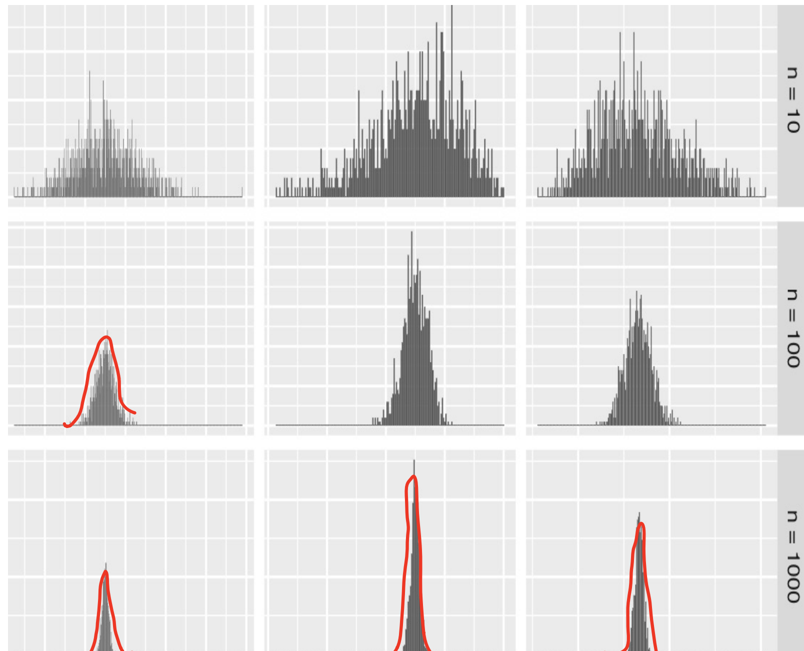
Probability distribution of X_k



Sample distribution $n=10, 100, 1000$

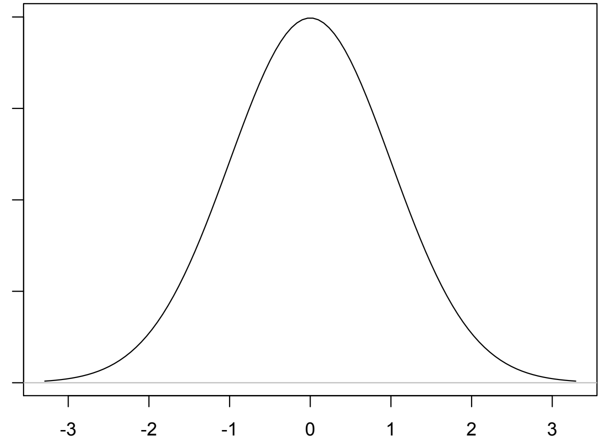


Distribution of sample mean $= \frac{S_n}{n}$



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 $|z|>3$



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