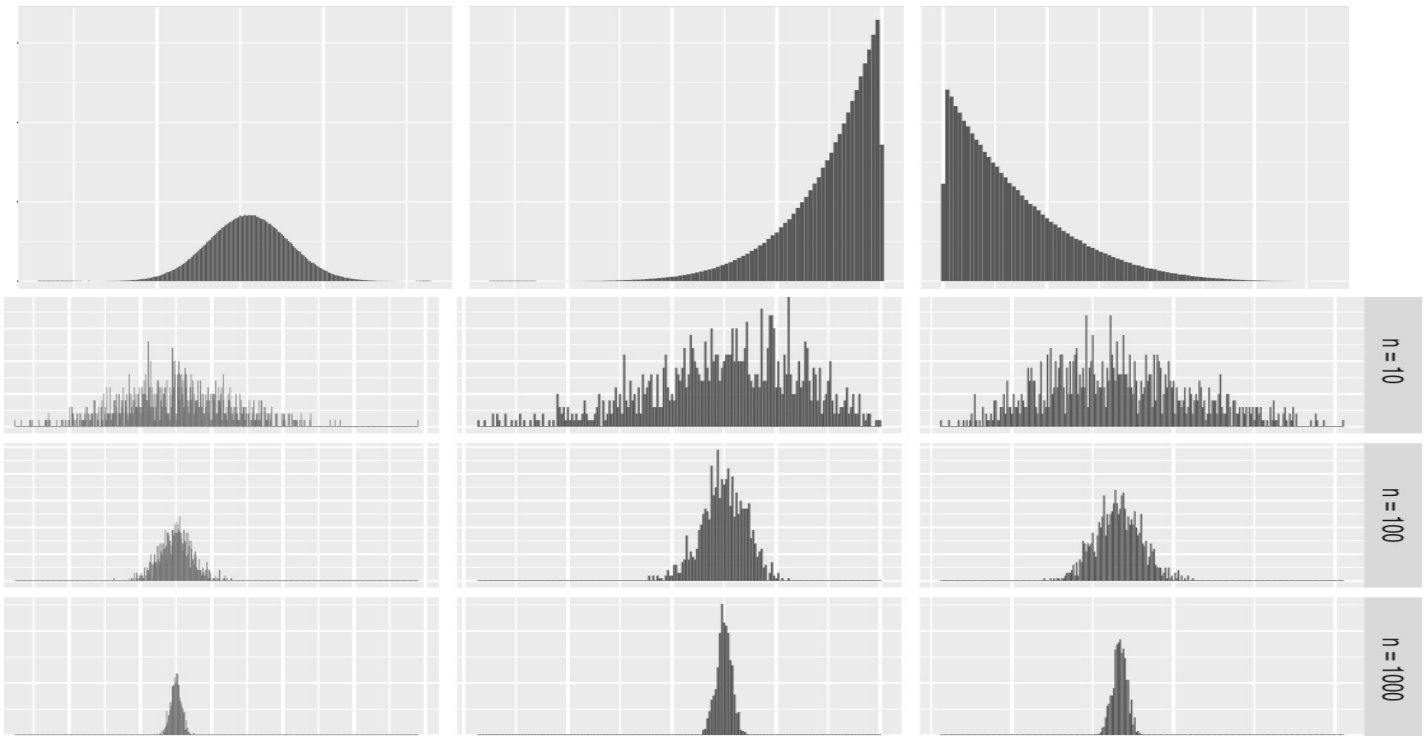
Stat 88: Probability & Mathematical Statistics in Data Science



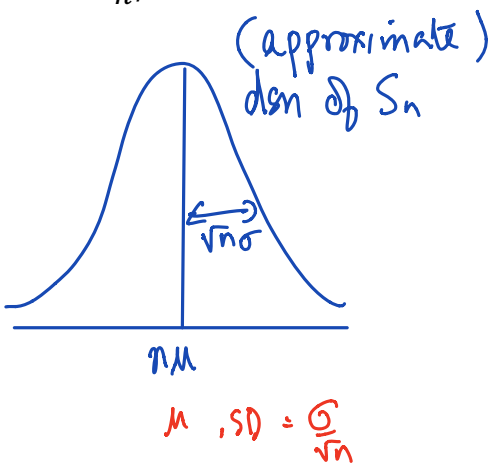
Lecture 28 PART 1: 4/2/2021

Sections 8.3, 8.4

Using the Central Limit Theorem

The Central Limit Theorem

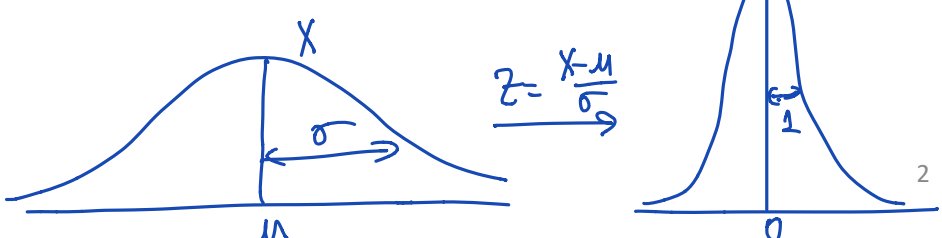
- Suppose that $X_1, X_2, \ldots, X_n$ are iid with mean μ and SD σ & $S_n = X_1 + \cdots + X_n$ is the sample sum
- Then the distribution of S_n (and A_n) is ***approximately normal*** for large enough n.
- The distribution is approximately normal (bell-shaped) centered at $\boldsymbol{E(S_n) = n\mu}$ and the width of this curve is defined by $\boldsymbol{SD(S_n) = \sqrt{n}\,\sigma}$.
- For A_n, the distribution is centered at $\boldsymbol{E(A_n) = \mu}$ with spread $\boldsymbol{SD(A_n) = \sigma/\sqrt{n}}$



Every normal curve is the same
X has a bell shaped dsn, mean μ, SD σ

(X*) $Z = \frac{X-\mu}{\sigma}$, $E(Z) = 0$, $SD(Z) = 1$

Normal curve of X → standard Normal Z



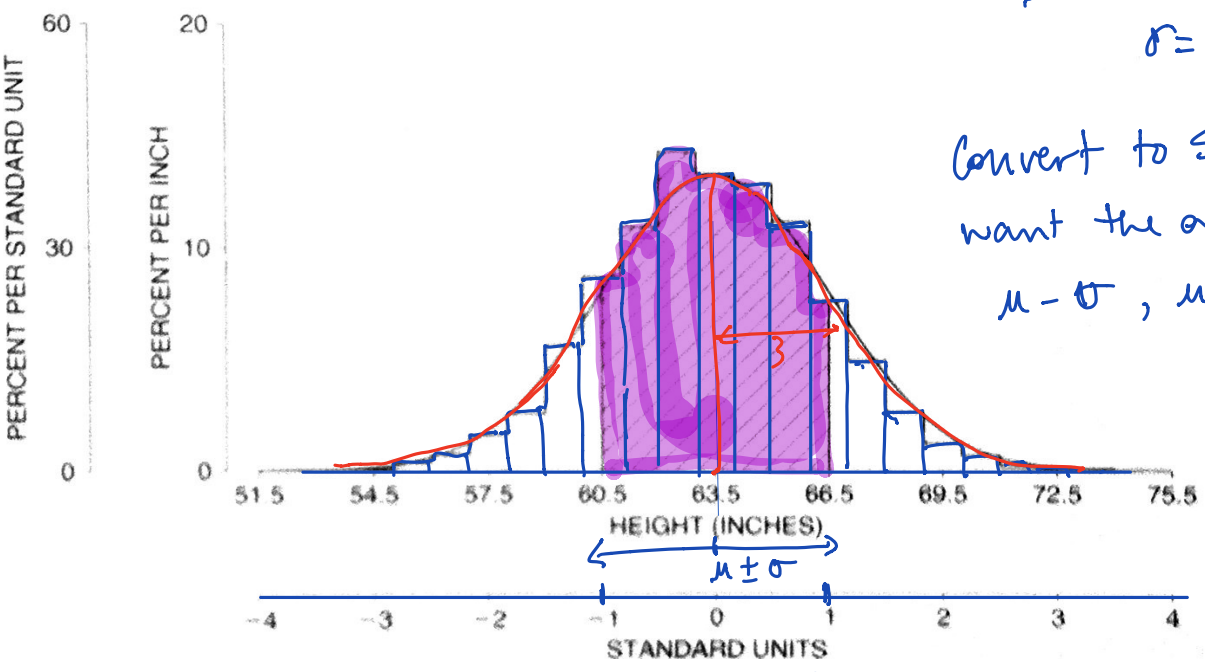
Example: Heights of women
Mean = 63.5 inches, SD =3 inches

stats.norm.cdf

Figure 2. A histogram for heights of women compared to the normal curve. The area under the histogram between 60.5 inches and 66.5 inches (the percentage of women within one SD of average with respect to height) is about equal to the area between −1 and +1 under the curve—68%.

μ = mean of data = 63.5"

σ = SD = 3"



Convert to std units

want the area b/w

μ − σ , μ + σ

Picture from *Statistics,* by Freedman, Pisani, & Purves

How large is "large"?

Suppose that $X_1, X_2, \ldots, X_n$ are iid with mean μ and SD σ & $S_n = X_1 + X_2 + \cdots + X_n$ is the sample sum, then the distribution of S_n is ***approximately normal*** for ***large*** enough n.

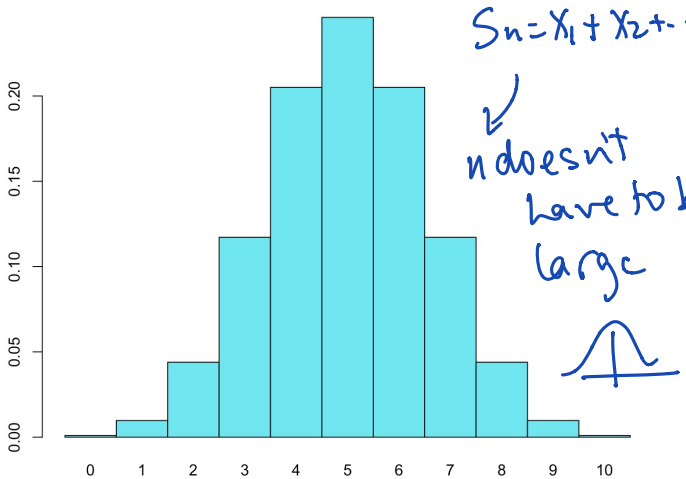
Question: How large is "large enough"

Answer: Well, it depends.

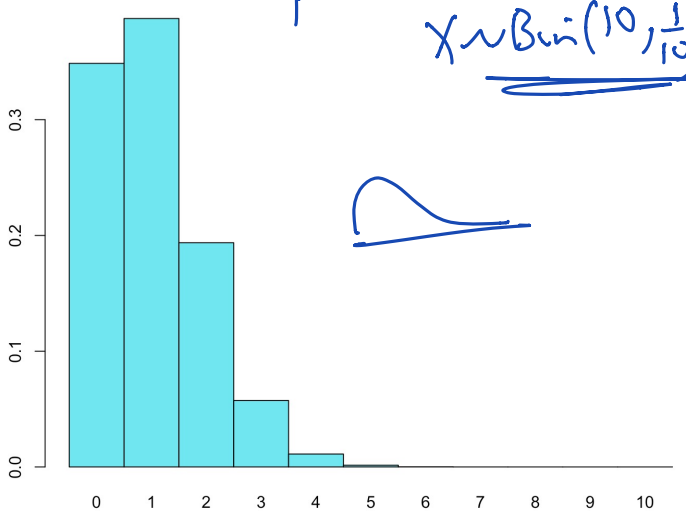
prob hist for $X \sim Bin(10, \frac{1}{2})$

$S_n = X_1 + X_2 + \cdots + X_n$

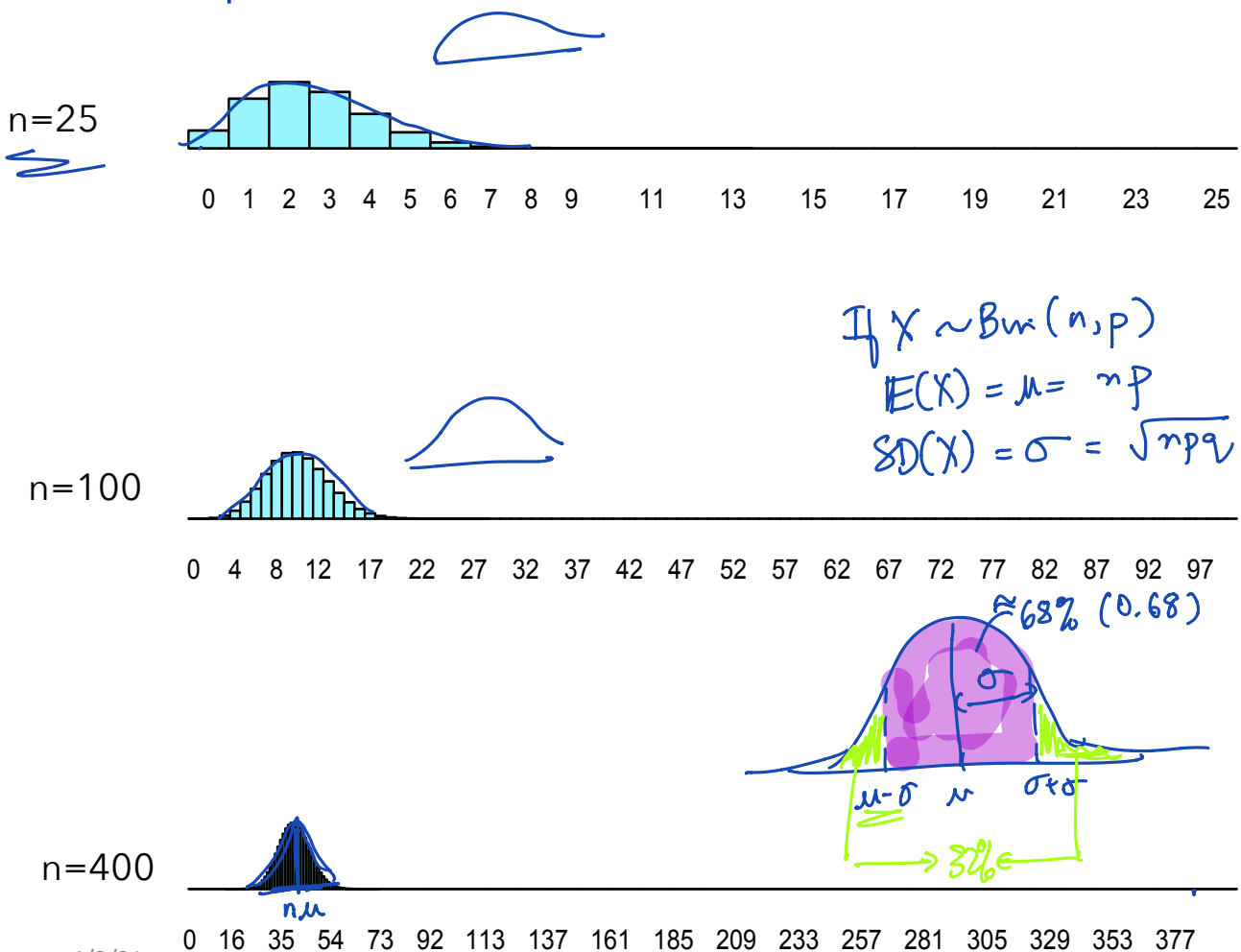
n doesn't have to be large



prob hist for $X \sim Bin(10, \frac{1}{10})$



When p is small



If $X \sim Bin(n,p)$

$E(X) = \mu = np$

$SD(X) = \sigma = \sqrt{npq}$, $q = 1-p$.

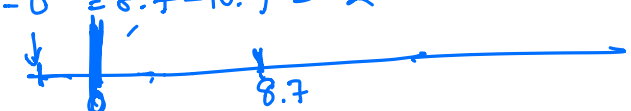
How to decide if a distribùtion could be normal

area to the left & right of $\mu - \sigma$, $\mu + \sigma$ respectively

- Need enough SDs on both sides of the mean.

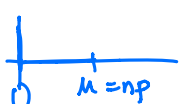
- In 2005, the murder rates (per 100,000 residents) for 50 states and D.C. had a mean of 8.7 and an SD of 10.7. Do these data follow a normal curve?

$\mu - \sigma = 8.7 - 10.7 = -2$



- 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.

$\sigma = \sqrt{npq}$ is larger than np.



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

Example

In the gambling game of Keno, there are 80 balls numbered 1 through 80, from which 20 balls are drawn at random. If you bet a dollar on a single number, and that number comes up, you get your dollar back, and win $2. If you lose, you lose your dollar (win = $-1). Your chance of winning is 0.25 each time.

Suppose you play 100 times, betting $1 on a single number each time, what is the chance that you come out ahead (win some positive amount of money)?

Each time we play, we win $2 with prob 0.25
& lose 1 dollar w/prob 0.75

So let X_k be our gain on the k^{th} round. $X_k = \begin{cases} 2 \text{ w.p. } 0.25 \\ -1 \text{ w.p. } 0.75 \end{cases}$ $E(X_k) = -0.25$

Net gain = sum of X_i after 100 rounds of play.

$S_{100} = X_1 + X_2 + \cdots + X_{100}$

$E(S_{100}) = 100 \cdot E(X_k) = 100\left(2(0.25) + (-1)(0.75)\right) = -25$ dollars

$SD(S_{100}) = 10\,(SD(X_k)) = 10 \cdot \frac{3\sqrt{3}}{4} \approx 12.99$

$$\mathrm{Var}(X_k) = \mathbb{E}(X_k^2) - (\mathbb{E}(X_k))^2$$

$$\mathbb{E}(X_k^2) = 4\left(\tfrac{1}{4}\right) + (1)\tfrac{3}{4} = \tfrac{7}{4}$$

$$\mathrm{Var}(X_k) = \tfrac{7}{4} - \tfrac{1}{16} = \tfrac{28-1}{16} = \tfrac{27}{16}, \quad SD(X_k) = \tfrac{3\sqrt{3}}{4}$$

Assuming that we can use the CLT (n is large enough)

$$Z = \frac{S_{100} - \mathbb{E}(S_{100})}{SD(S_{100})}, \quad P(S_{100} > 0) = \text{chance of coming out ahead.}$$

$$P(S_{100} > 0) = P\left(\frac{S_{100} - (-25)}{12.99} > \frac{0 - (-25)}{12.99}\right) = P(Z > 1.9245)$$

(Z: standard normal r.v.)

$$\approx 0.027$$

So chance of coming out ahead $\approx 2.7\% = 0.027$.